To my family

DEVELOPMENT OF A SHINY APPLICATION FOR COMPARATIVE TRANSCRIPTOMICS AND DIFFERENTIAL GENE EXPRESSION ANALYSIS

A THESIS SUBMITTED TO

THE GRADUATE SCHOOL OF ENGINEERING AND SCIENCE

OF BILKENT UNIVERSITY

IN PARTIAL FULLFILLMENT OF THE REQUIREMENTS FOR

THE DEGREE OF

MASTER OF SCIENCE

IN

MOLECULAR BIOLOGY AND GENETICS

By

Ronaldo Leka

September 2022

DEVELOPMENT OF A SHINY APPLICATION FOR COMPARATIVE TRANSCRIPTOMICS AND DIFFERENTIAL GENE EXPRESSION ANALYSIS

By Ronaldo Leka

September 2022

We certify that we have read this thesis and that in our opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

_____

Özlen Konu Karakayalı (Advisor)

_____

Bahar Değirmenci Uzun

_____

Ceren Sucularlı

Approved for the Graduate School of Engineering and Science:

_____

Orhan Arıkan

Director of the Graduate School

ABSTRACT

DEVELOPMENT OF A SHINY APPLICATION FOR COMPARATIVE TRANSCRIPTOMICS AND DIFFERENTIAL GENE EXPRESSION ANALYSIS

Ronaldo Leka

M.Sc. in Molecular Biology and Genetics

Advisor: Özlen Konu Karakayalı

September 2022

RNA sequencing has proven to be an effective technique for divulging information about the transcriptome in molecular biology research. Compared to microarrays and early methods of cDNA sequencing, high-throughput RNA sequencing has better resolution, low background noise, a higher range to quantify gene expression, and relatively lower cost. The development of sequencing technique has led to the development of tools for analyzing the high volume of data that is generated. Statistical methods for normalizing, filtering, performing exploratory and differential analysis, and other functional analyses based on RNA sequencing count data have made RNA sequencing one of the most popular techniques in genomics. To help facilitate the use of such statistical tools, web applications developed in R using the shiny package offer an advantageous environment where researchers can use a graphical interface to give inputs and instructions to the underlying server-side libraries that analyze and generate results in tables and plots. This thesis presents a new tool that has been developed for exploratory analysis, data normalization and filtering, differential gene expression analysis (DGEA), correlation analysis, principal component analysis, and functional analysis such as over-representation analysis and gene set enrichment analysis. When compared to other available applications, this new application offers the ability to run multiple methods for DGEA and compare results between them, leading to the output of gene sets that are discovered as DEGs in multiple tests. Here I present the features of this application in detail where I aim to improve upon the applications that are available in the literature. An example dataset from our lab was also investigated by this RNA-seq tool leading to a better understanding of Mineralocorticoid Receptor (MR) signaling in breast cancer.

Keywords: RNA sequencing, data analysis, data visualization, web application, differential gene expression analysis, functional analysis, breast cancer, aldosterone, spironolactone

ÖZET

KARŞILAŞTIRMALI TRANSKRİPTOMİK VE DİFERANSİYEL GEN EKSPRESYONU İÇİN BİR SHINY UYGULAMASININ GELİŞTİRİLMESİ

Ronaldo Leka

Moleküler Biyoloji ve Genetik, Yükses Lisans

Tez Danışmanı: Özlen Konu Karakayalı

Eylül 2022

RNA dizilemenin, moleküler biyoloji araştırmalarında transkriptom hakkında bilgi vermek için etkili bir teknik olduğu kanıtlanmıştır. Mikrodizinler ve cDNA dizilemenin ilk yöntemleriyle karşılaştırıldığında, yüksek verimli RNA dizileme, daha iyi çözünürlüğe, düşük arka plan gürültüsüne, gen ifadesini ölçmek için daha yüksek bir aralığa ve nispeten daha düşük maliyete sahiptir. Sekanslama tekniğinin gelişimi, üretilen yüksek hacimli verilerin analiz edilmesi için farkli araçların geliştirilmesine yol açmıştır. Normalleştirme, filtreleme, keşif ve diferansiyel analiz gerçekleştirme ve RNA dizileme sayım verilerine dayalı diğer işlevsel analizler için istatistiksel yöntemler, RNA dizilemeyi genomikteki en popüler tekniklerden biri haline getirmiştir. Bu tür istatistiksel araçların kullanımını kolaylaştırmaya yardımcı olmak için, Shiny paketi kullanılarak R'de geliştirilen web uygulamaları, araştırmacıların tablo ve grafiklerde sonuçları analiz eden ve üreten altta yatan sunucu tarafı kütüphanelerine girdi ve talimat vermek için grafik bir arayüz kullanabilecekleri avantajlı bir ortam sunar. Bu tez, keşif analizi, veri normalleştirme ve filtreleme, diferansiyel gen ekspresyon analizi (DGEA), korelasyon analizi, temel bileşen analizi ve aşırı temsil analizi ve gen seti zenginleştirme analizi gibi fonksiyonel analizler için geliştirilen yeni bir araç sunmaktadır. Mevcut diğer uygulamalarla karşılaştırıldığında, bu yeni uygulama DGEA için birden fazla yöntem çalıştırma ve bunlar arasındaki sonuçları karşılaştırma yeteneği sunarak, çoklu testlerde DEG olarak keşfedilen gen setlerinin çıktısına yol açmaktadır. Burada, literatürde mevcut olan uygulamaları ilerletmeyi amaçladığım bu uygulamanın özelliklerini ayrıntılı olarak sunuyorum. Laboratuvarımızdan örnek bir veri seti de bu RNA-seq aracı ile incelenerek meme kanserinde Mineralokortikoid Reseptör (MR) sinyalizasyonunun daha iyi anlaşılması sağlanmıştır.

Anahtar sözcükler: RNA dizileme, veri analizi, veri görselleştirme, web uygulaması, diferansiyel gen ekspresyon analizi, fonksiyonel analiz, meme kanseri, aldosteron, spironolakton

Acknowledgements

Firstly, I would like to express the highest level of gratitude towards my advisor, Özlen Konu Karakayalı. She opened the door to her laboratory and created an excellent working environment, instilling in me the importance of collaboration, research, and teaching others. She has been understanding, supportive, and incredibly patient with me through my graduate studies journey.

Secondly, I would like to thank Dr. Bahar Değirmenci Uzun at Bilkent University and Dr. Ceren Sucularlı at Hacettepe University for participating in my thesis defense committee, and for dedicating time to evaluate my thesis and helping me graduate.

I would like to acknowledge Bilkent University as an excellent establishment for education that has been my home for the last seven years. I acknowledge the fact that the university has provided me a scholarship for my undergraduate studies, and a scholarship and stipend during my graduate studies.

I would like to thank the department of Molecular Biology and Genetics at Bilkent University where I have very much enjoyed studying and now look fondly upon all these moments spent learning from the many amazing professors and instructors through the years. I thank all faculty and staff for creating a welcoming environment for us students.

I acknowledge that MR signaling work included in this thesis was funded by TÜBİTAK (Grant No. 114S226 to OK) and carried out by Seniye Targen, Bircan Çoban, and Vladimir Benes (EMBL, Germany). I thank Kübra Çalışır and Merve Vural for their contribution in the development and validation of the shiny application. I thank undergraduate students, namely, Ömer Fatih Konar, Cem Özdemir, Ömer Ergül, and Ezgi Şen who helped me with this app.

I would like to thank my collaborators, members of the Konu Laboratory starting with previous members: Tuğberk Kaya, Fatma Betül Dinçaslan, Ayşe Gökçe Keşküş, Seniye Targen, Murat Yaman, Damla Güneş, and Büşra Korkmaz. I would like to also thank current members of the lab (some of them already graduated): Melike Tombaz, Burçin İrem Arıcı, Merve Vural, Kübra Çalışır, Farid Ahadli, and the remaining members: Moiz, Güneş, Siber, Rüya, and Çağdaş. Many hours were spent learning, exchanging ideas, helping one another, and socializing which is why this laboratory group was an amazing place to be in.

I would like to thank my closest Bilkent graduate school friends, Filiz Şahin and Melike Tombaz. I have enjoyed our time together inside the bigger MBG family.

Lastly, I would like to thank my family: my mother Luljana Leka, my father Ramazan Leka, my siblings Migena, Indrit, and Osman, my nephews Marvin and Arbër, and my niece Agnesa. I have received more love and support than I can hope to ever give back, which has kept me going during this long part of my life in Turkey. I can only hope that I recoup this lost time.

Contents

List of figures

List of tables

Chapter 1. Introduction

Since the introduction of RNA sequencing as a technology and the mapping and quantifying of the transcriptomes of yeast, *Arabidopsis thaliana*, mouse, and human cells at first, it quickly became advantageous to researchers to use these high-throughput methods to divulge information about the transcriptome of their research subjects (Lister et al., 2008; Mortazavi et al., 2008; Nagalakshmi et al., 2008). Compared to microarrays and early methods of cDNA sequencing, high-throughput RNA sequencing has better resolution, low background noise, a higher range to quantify gene expression, and relatively lower cost (Z. Wang et al., 2009).

The development of high-throughput RNA sequencing methods was accompanied by the development of bioinformatics tools for mapping and preprocessing cDNA sequences to get read count data, and statistical methods for count data normalization, differential expression, and functional analyses. A flow chart of RNA sequencing data analysis is shown in **Figure 1** (D. Li, 2019). This thesis focuses on RNA-seq data normalization, differential gene expression analysis, and functional analyses including over-representation analysis and gene set enrichment analysis.

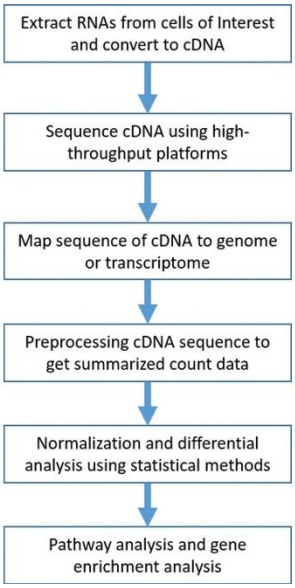


Figure 1. Flow chart of RNA sequencing data analysis. Figure adapted from the Computation Biology book, chapter 6 *(Husi, 2019; D. Li, 2019)*.

1.1 Statistical methods for RNA-seq count data normalization

After RNA-seq data goes through the steps of sequencing alignment and quantification, the read count matrix obtained consists of raw read counts. To get informative and unbiased values of gene expression, raw read counts need to be normalized to account for factors that affect the number of counts. These factors include, but are not limited to, transcript length, GC content, and sequencing depth (Oshlack & Wakefield, 2009; Risso et al., 2011; Robinson & Oshlack, 2010). Transcript length affects read counts because of the mRNA fragmentation, which is a sequencing approach for RNA samples to get sequence coverage of the transcript. Therefore, a long mRNA transcript will have more fragments and consequently more reads that map to it than a short mRNA transcript of similar expression. GC percentage introduces bias in high-throughput sequencing by confounding the read coverage and is unimodal; it has been shown that fragments that are GC-rich or GC-poor are underrepresented in sequencing data

(Benjamini & Speed, 2012). Both transcript length and GC content bias are within-sample biases, whereas sequencing depth is a between-samples bias.

The third bias mentioned here, sequencing depth is a factor that affects comparison of expression values across samples and prevents read data from correctly reflecting the differences in expression (Robinson & Oshlack, 2010). Sequencing depth bias is a result of technical effects (as opposed to the biological effects described previously). These technical effects result from experimental and technical variability, the main one being the variability in the total number of sequenced fragments, which leads to differences in total read counts, also known as differences in sequencing depth. Between-sample normalization is needed to reduce the effects of sequencing depth in differential analysis.

Many statistical methods for between-sample normalization have been developed and described in the literature, and Evans *et al*. have written an extensive review of these methods by taking their assumptions into account (Evans et al., 2018). Researchers have developed normalization methods that depend on library size, read count distribution, testing, or controls. In this thesis, four of these methods have been used, namely counts-per-million (CPM), log2-transformed CPM, voom, and DESeq normalization (Law et al., 2014; Love et al., 2014; Robinson & Oshlack, 2010). The web tool described here can produce normalized counts using all these methods while keeping the option to let the differential gene expression analysis methods use their internal normalization methods.

Counts-per-million (CPM) and log2-transformed CPM are methods of normalization by library size (Robinson & Oshlack, 2010). These two methods aim to remove differences in sequencing depth, using the assumption that the cells have the same total number of expression values under different experimental conditions. CPM normalization involves the division of each read count by the sum of reads in the sample and then multiplication of the result by one million.

The resulting effect is that all samples now have a library size of one million reads. log2-transformed CPM normalization is achieved similarly to CPM, ending with taking the binary logarithm of the CPM values.

voom is a read count normalization method that aims to estimate the trend of mean-variance of log-transformed read counts and to use the relationship between the mean and variance to predict the variance of the log2CPM values (Law et al., 2014). In voom normalization, variance at the gene level is modeled based on normalized log fold change values. The aim is to incorporate experimental treatment groups as factors. Voom can handle multiple factors ranging from replicates of biological treatments and covariates and generates gene-wise residual standard deviation. A predicted count is obtained upon fitting a robust trend to these standard deviation values. Moreover, standard deviation of predicted count also is calculated and weights are generated using the inverse squared predicted standard deviation.

The DESeq normalization method is a method based on the principle of normalization by distribution. The method a by which read counts are scaled with for a particular sample (Anders & Huber, 2010).

Other normalization methods include RPKM (Mortazavi et al., 2008) and FPKM (Trapnell et al., 2010) (variants of CPM that normalize for transcript length), upper quantile (Bullard et al., 2010), CuffDiff (Trapnell et al., 2012) (variant of DESeq), trimmed mean of M-values (Robinson & Oshlack, 2010) (TMM), and Poisson-Seq (J. Li et al., 2012). An extensive review of these methods has been authored by Evans *et al*. in 2018 (Evans et al., 2018).

1.2 Statistical methods for RNA-seq differential gene expression analysis

Similarly to methods for normalization, a lot of statistical methods have been described for the differential gene expression analysis of RNA-seq data. The most commonly used methods remain edgeR (Robinson et al., 2010) and DESEq2 (Love et al., 2014), both designed to use the negative binomial distribution as reference and provide their own internal normalization methods. There are other methods that are Bayesian, namely, baySeq (Hardcastle & Kelly, 2010), EBSeq (Leng et al., 2013) and limma (Ritchie et al., 2015). This latter uses linear mean-variance models and relative weighting of global and gene-wise variance estimators to improve statistical power and accuracy of finding differential expression. Another type of approach is the non-parametric approach executed by methods such as NOISeq (Tarazona et al., 2011) and SAMseq (J. Li & Tibshirani, 2013). These methods use the actual data to estimate the null distribution for statistical analysis while making minimal assumptions. Based on these different types of statistical methods, the ease of use and popularity of methods, edgeR, DESeq2, limma, and NOISeq were selected for the differential gene expression analysis part of this thesis and the web application.

edgeR (Robinson et al., 2010) models count data using an overdispersed Poisson model and an empirical Bayes procedure. Its strength comes from adjustment of overdispersion uses a negative binomial distribution to model the data, which reduces to Poisson distribution when the dispersion is zero. The method can separate biological and technical variation through the dispersion value, which in this model represents the coefficient of between-sample biological variation. An exact test is used for testing differential expression has and this has been adapted for over-dispersed data.

DESeq2 (Love et al., 2014) is an R package that, differently from the other three methods explained here, requires a matrix of integer count data that is modeled using a negative binomial distribution. The DESeq2 method estimates size factors as described in the first implementation of the DESeq method (Anders & Huber, 2010). Size factors are calculated considering each

sample as well as all samples using median ratios and geometric mean, respectively. DESeq2 uses generalized linear models (GLMs) that, in the simplest case of comparing two groups, returns values indicating the log2 fold change between treatment and control. DESeq2 makes use of empirical Bayes shrinkage for dispersion and fold change estimation. After GLMs are fitted for each gene, a Wald test and adjustment of p-values via methodology of Benjamini and Hochberg are used (Benjamini & Hochberg, 1995).

limma (Ritchie et al., 2015) is an R package that was previously restricted to microarray data but has been made compatible for RNA-seq data analysis because of the emergence of RNA-seq as a more powerful technique. limma is able to analyze RNA-seq read counts by estimating the mean-variance relationship in log-scale read counts. The voom normalization converts the mean-variance model to precision weights. Testing for differential expression is achieved by shrinking gene-wise variances towards the variance trend. This is important to reduces the false positive rate of stable genes and help detection of DEGs with larger variances. In this method t-statistics and their P-values are moderated using an empirical Bayes approach are used to assess the differential expression for each gene. This method gives comparable results to the packages that use the negative binomial distribution (edgeR and DESeq2) while being faster and more reliable for large datasets (Rapaport et al., 2013).

NOISeq (Tarazona et al., 2011) is an R package providing a non-parametric tests; it is based on a noise distribution of fold change (M) versus expression (D, absolute) values and comparisons with that of reference, to compare the M and D values of a given gene are likely to be differential gene expression or noise. NOISeq implements this method in two variants: NOISeq-real which uses biological and/or technical replicates for DGE analysis, and NOISeq-sim which uses a simulation for replicates when replication is absent. The method internally accounts for library size bias by correcting the counts by a factor closely related to the CPM method and based on the standard deviation of the read counts. The differential expression

statistics include the aforementioned M and D values (log2-transformed fold change and absolute expression difference, respectively), and the probability q value. One of the biggest issues in RNAseq analysis is multiple tests generated on the same samples and this increases the probability of finding genes as differentially expressed when they are not. The q value is calculated by the odds ratio of being DE over not being DE, e.g., an odds ratio 4:1. In this example q-value equals to 0.8 and in fact this indicates that odd of being differentially expressed is four times greater than being non-differentially expressed. NOISeq method recomments q>0.8 as a threshold (Tarazona et al., 2011).

1.3 Overcoming the single replicate issue

It is a truth universally acknowledged that scientific research is heavily dependent on replicated measurements. Replicates are important in yielding more accurate and reliable statistical analysis results and all experiments, including RNA sequencing experiments need to have biological or technical replicates. In RNA-seq data analysis, replicates can be used to assess sources of variation and limit their effect on the results of differential expression analysis. However, replicates are not always possible, and researchers have to find ways to make statistical inferences using less-than-ideal conditions. In the case of RNA-sequencing data, the technology can be cost-prohibitive, technical replication might have been erroneous, data can be of low quality, or any of a myriad of reasons can cause the lack of replicate data. Therefore, some work arounds have been devised to work with data that lacks replicates.

Of the four DGE analysis methods used throughout this thesis and available in the application, only DESeq2 (Love et al., 2014) requires at least one of the groups compared in an analysis to have replicates. This package is designed to use the negative binomial distribution and to give significant statistical results the implementation requires at least one replicate. Therefore, it is

halfway useful when dealing with groups with no replicates. Of the other three packages, edgeR (Robinson et al., 2010), and limma (Ritchie et al., 2015), the two most commonly used pipelines for differential gene expression can be used to analyze groups with one sample each. In the case of edgeR, in case the data has no replicates the exactTest() function can be run using the dispersion argument with a pre-defined value of 0.16 for human data, or 0.01 for model organisms (Chen et al., 2022). limma similarly can take as input two groups with one sample without need additional input as long as the dataset count matrix has at least one group with replicates in order to be able to calculate variance through the voom normalization. NOISeq is the method that is most compatible with data that have no replicate. A variant of the NOISeq method is the NOISeq-sim function, which uses data from groups with no replicates to simulate reads. This simulation makes the assumption that read counts follow a multinomial distribution. Given the standard deviation of the real sample, the simulated samples have an SD value generated randomly from a uniform distribution. Compared to NOISeq-real, which uses real replicates, NOISeq-sim has been tested and found to perform similarly but with a slightly higher detection rate of DEGs and more dependency on SD values (Tarazona et al., 2011).

1.4 Shiny applications used for RNA-seq data analysis and visualization

Researchers need to analyze and visualize RNA-seq data in a streamlined manner without the need for learning to code in R or Python and get results in a customizable user interface. The {shiny} framework is an R package that can be used to build web applications in R (Chang et al., 2022). Using {shiny} and other complementary packages such as {shinydashboard} (Chang et al., 2022), {shinybrowser} (Attali [aut & cre, 2022), {shinyWidgets} (Perrier et al., 2022), etc., bioinformaticians have been able to develop web applications that can be used for RNA-seq data analysis and visualization. Shiny web applications have the advantage of having

modern interfaces with the server-side part being well-integrated and compatible with all R packages that offer statistical and visualization methods. These apps can offer full functionality of any underlying R package while being user-friendly, and the results are easily reproducible. There are numerous web tools in the literature to implement statistical and visualization methods for RNA-seq data analysis, each of them unique in graphical user interface, features, and options that are available to the user. A summary of features found across these Shiny applications can be found in **Tables 1-4**.

DEBrowser (Kucukural et al., 2019) is a web application developed as an R Bioconductor project, meaning that it is also available as an R package. Out of all the shiny web applications described in this thesis, it is the most versatile as it offers features from many of the categories listed in **Tables 1-4**. To use this application, users need to upload two files, a count matrix file, and a sample metadata file. It offers many count data normalization methods, namely: MNR, TMM, RLE, and upper quartile normalization, some of which are unique compared to the other apps in literature. Another unique feature is batch-effect normalization which is achieved through Combat, CombatSeq, and Harman normalization methods. This application is similar to the one described in this thesis in the category of DGE analysis, and functional analyses. However, the normalization and DGE analysis methods cannot be applied simultaneously and thus cannot be compared in app.

GENAVi (Reyes et al., 2019) is another application that aims to provide a user-friendly experience in a stable online tool that uses high quality analysis tools and generates downloadable plots and tables. The app achieves its aims by having interactive plots for exploratory analysis and DGE analysis, multiple ORA and GSEA options and multiple plots to visualize these results. Overall, it still lacks some visualization methods for exploratory analysis and does not offer many options for DGE analysis, however it is one of the only two

apps in **Tables 1-4** that have an option to save a report of the analyses performed in the application.

GREIN (Mahi et al., 2019) is a web application used for manipulating and re-analyzing RNA-seq data from the Gene Expression Omnibus (GEO). The user can download and analyze processed data, construct differential gene expression (DGE) signatures and perform functional characterization. This application contains >28,000 datasets with than ~600,000 samples in total but does not have any data upload functionality. While it offers a lot in terms of features related to DGE analysis and offers the unique feature of DGE signature creation, it is not very useful for newly generated data unless this data is first uploaded to GEO and processed by the app servers.

IRIS-EDA (Monier et al., 2019) is a web application that implements a lot of the main features of an application designed for RNA-seq exploratory data analysis and differential gene expression analysis. The app offers most of the core functionality and data exploratory analysis visualization methods available. IRIS-EDA is one of the apps that offers multiple methods for data normalization and DGE analysis, however, like DEBrowser it cannot use these methods at the same time and does not aim to compare their results. This application does not have capabilities to perform functional analyses that others listed here can do and does not have any unique features.

TCC-GUI (Su et al., 2019) is GUI that belongs to TCC R package (Sun et al., 2013), which normalization methods. TCC also provides an interface to perform analyses using a combination of methods such as edgeR, DESeq, and baySeq. The TCC-GUI app adds SAMSeq and limma to this list of methods and provides multiple visualizations for exploratory data analysis and DGE analysis, and writes out a report with the plots and tables generated in the app. Similarly to other apps, it lacks functional analyses and does not offer any unique features.

The application described in this thesis has the most used features from the categories in the comparison tables shown below. Firstly, it is easier to start using with a carefully prepared count matrix file. The input file must follow pre-determined rules for the application to function correctly, while avoiding the requirement of a separate metadata file like other applications described previously. While the application lacks some features such as batch-effect normalization, and databases for gene set enrichment analysis, it has the unique capability of applying multiple normalization and DGE analysis methods. This enables the application to have the feature of comparing DGEA results across multiple methods to find genes that are commonly discovered as DEGs. This serves as a comparison between methods, as well as emphasizes statistical significance of the different statistical tests that are implemented by the DESeq2, edgeR, limma, and NOISeq packages.

Table 1. Availability and core functionality features of the available shiny applications compared to the application described in this thesis.

Feature \ App		NEW RNA-SEQ APP (this thesis)	BISR-RNAseq (Gadepalli et al., 2019)	DEBrowser (Kucukural et al., 2019)	DEGenR (Choudhary et al., 2021)	GENAVi (Reyes et al., 2019)	GREIN (Mahi et al., 2019)	ideal (Marini et al., 2020)	IRIS-EDA (Monier et al., 2019)	MicroScope (Khomtchouk et al., 2016)	TCC-GUI (Su et al., 2019)
Availability	Online server	Yes	Yes (example demonstration)	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes
	Online repository	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
	Available in Bioconductor as R package	No	No	Yes	No	No	No	Yes	No	No	Yes
Core functionality	Upload capability	Yes (one count file)	No	Yes (count file and metadata file)	Yes	Yes (count file, metadata file, results file)	No	Yes (count file, metadata file)	Yes	Yes	Yes (count file and group info)
	Demo data	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
	Import from online data repositories	No	No	No	Yes	No	Yes (~28k GEO datasets)	No	No	No	No
	Plot interactivity	Yes	Yes	Yes	No	Yes	Yes	No	Yes	No	Yes
	Download options	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	No	Yes
	Introduction page Help page	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Table 2. Data wrangling and exploratory data analysis features of the available shiny applications compared to the application described in this thesis.

App / Feature		NEW RNA-SEQ APP (this thesis)	BISR-RNAseq	DEBrowser	DEGenR	GENAVi	GREIN	ideal	IRIS-EDA	MicroScope	TCC-GUI
Data wrangling	Data filtering step	Yes	No	Yes	Yes	No	Yes	Yes	Yes	No	Yes
	Data normalization	Yes	No	Yes	No	No	No	No	Yes	No	Yes
	Normalization methods	CPM Log2CPM limma-voom transformation DESeq	N/A	MRN TMM RLE upperquartile	N/A	N/A	N/A	N/A	log2 transformation rlog vst	N/A	TMM DESeq
	Batch effect correction	No	No	Yes	No	No	No	No	No	No	No
	Correction methods	N/A	N/A	Combat CombatSeq Harman	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Exploratory data analysis	Scatter plot	Yes	No	Yes	No	No	No	Yes	Yes	No	No
	PCA	Yes	No	Yes	No	Yes	Yes	No	Yes	Yes	Yes
	MultiQC plots	No	Yes	No	No	No	Yes	No	No	No	No
	Alignment statistics	No	Yes	No	No	No	No	No	No	No	No
	MDS plot	Yes	Yes	No	Yes	No	No	No	Yes	No	Yes
	IQR plot/boxplot	Yes	No	Yes	Yes	No	No	No	Yes	No	Yes
	Density plot/histogram	Yes	No	Yes	Yes	No	Yes	No	Yes	No	Yes
	Heatmap	Yes	No	No	No	Yes	Yes	Yes	Yes	Yes	No
	Clustering plot	Yes	No	No	No	Yes	No	No	Yes	No	Yes
	t-SNE plot	No	No	No	No	No	Yes	No	No	No	No

Table 3. Differential gene expression analysis and comparison between DGEA methods features of the available shiny applications compared to the application described in this thesis.

App / Feature		NEW RNA-SEQ APP (this thesis)	BISR-RNAseq	DEBrowser	DEGenR	GENAVi	GREIN	ideal	IRIS-EDA	MicroScope	TCC-GUI
Differential gene expression analysis	Differential gene expression analysis	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
	DGEA methods	DESeq2 edgeR limma NOISeq	Limma	DESeq2 edgeR limma	eBayes TREAT topConfects	DESeq2	edgeR Power analysis	DESeq2	DESeq2 edgeR limma	edgeR	edgeR DESeq2 baySeq SAMSeq limma
	Covariate selection	No	No	Yes	No	Yes	Yes	No	No	No	No
	Volcano plot	Yes	Yes	Yes	No	Yes	No	Yes	Yes	No	Yes
	MA plot	No	Yes	Yes	No	No	Yes	Yes	Yes	No	Yes
	Heatmap	Yes	Yes	Yes	No	No	Yes	Yes	No	No	Yes
	Gene detectability plot	No	No	No	No	No	Yes	No	No	No	No
	Power curve	No	No	No	No	No	Yes	No	No	No	No
	MD plot	No	No	No	Yes	No	No	No	No	No	No
Comparison of DGEA methods	DGEA method comparison capability	Yes	No	No	Yes, limited	No	No	No	No	No	No
	Venn diagram	Yes	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	UpSet plot	Yes	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	Gene set tables	Yes	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

Table 4. Functional analyses and extra features of the available shiny applications compared to the application described in this thesis.

Feature \ App		NEW RNA-SEQ APP (this thesis)	BISR-RNAseq	DEBrowser	DEGenR	GENAVi	GREIN	ideal	IRIS-EDA	MicroScope	TCC-GUI
Over representation analysis	ORA	Yes	No	No	No	Yes	No	No	No	No	No
	Databases available for ORA	Gene Ontology KEGG	N/A	N/A	N/A	WikiPathways MSigDb Gene Ontology KEGG Disease Ontology	N/A	N/A	N/A	N/A	N/A
	Dot plot	Yes	N/A	N/A	N/A	Yes	N/A	N/A	N/A	N/A	N/A
	Network	Yes	N/A	N/A	N/A	Yes	N/A	N/A	N/A	N/A	N/A
	Tree plot	Yes	N/A	N/A	N/A	No	N/A	N/A	N/A	N/A	N/A
Gene set enrichment analysis	Gene set enrichment analysis	Yes	Yes	Yes	Yes	Yes	No	Yes	No	No	No
	Databases available for GSEA	Gene Ontology KEGG	Gene Ontology KEGG	Gene Ontology KEGG Disease Ontology	camera Gene Ontology Human Phenotype Ontology Reactome MSigDB Biocarta KEGG PID WikiPathways	WikiPathways MSigDb Gene Ontology KEGG Disease Ontology	N/A	Gene Ontology	N/A	N/A	N/A
	Functions	enrichGO enrichKEGG gseGO gseKEGG fgsea		enrichGO enrichDO enrichKEGG	camera fgsea CERNO Enrichr Hypergeometric	enricher, enrichGO, gseGO, enrichKEGG, gseKEGG, enrichDO, gseDO	N/A	goseq topGO	N/A	N/A	N/A
	Bar plot	No	No	Yes	Yes	No	N/A	No	N/A	N/A	N/A
	Dot plot	Yes	No	No	No	Yes	N/A	No	N/A	N/A	N/A
	Ridgeline plot	No	No	No	No	Yes	N/A	No	N/A	N/A	N/A
	Running score plot	Yes	No	No	No	Yes	N/A	No	N/A	N/A	N/A
	Ranked list of genes	Yes	No	No	No	Yes	N/A	No	N/A	N/A	N/A
	Network	Yes	No	No	No	Yes	N/A	No	N/A	N/A	N/A
	Barcode plot	Yes	No	No	Yes	No	N/A	No	N/A	N/A	N/A

	Panel plot	No	No	No	Yes	No	N/A	No	N/A	N/A	N/A
Extra features					TF enrichment analysis		Signature creation	Schweder-Spjotoll plot, p-value/ log2fc histograms, gene list comparison, signature explorer		Network analysis	
Output	Download report	No	No	No	No	Yes (html file)	No	Yes (html file)	No	No	Yes

1.5 Breast cancer

The World Cancer Research Fund International reported that breast cancer is the most common cancer worldwide, with more than 2.26 million new cases and more than 680,000 deaths in women in 2020 (WCRF, 2022). In the US in 2020, the National Cancer Institute appropriated more than $500 million of its $6.4 billion budget, or around 8% to breast cancer research (National Cancer Institute, 2022). Efforts in cancer research have resulted in the understanding that breast cancer, and cancer in general, are diverse in response to treatment. To address this diversity, Perou *et al*. reasoned that the phenotypic diversity of breast cancer must be accompanied by transcriptomic diversity and used cDNA microarray data to classify breast cancer into subtypes (Perou et al., 2000). Initially, four subtypes were defined: normal breast-like, basal, ERBB2-positive, and luminal epithelial/ER-positive. Later work defined five subtypes by dividing the luminal epithelial/ER-positive subtype in luminal A and luminal B subtypes (Sørlie et al., 2001, 2003). These subtypes have been repeatedly tested, redefined, and validated (Kapp et al., 2006).

1.6 The MCF-7 breast cancer cell line

The human breast cancer MCF-7 cell line was first grown and characterized in 1973 (Soule et al., 1973). The MCF-7 cell line is estrogen receptor (ER)- and progesterone receptor (PR)-positive, and therefore is categorized in the luminal A subtype (Comşa et al., 2015; Neve et al., 2006). The mineralocorticoid receptor (MR) was found to have no expression or very low expression in breast cancer cell lines in an expression profiling study in a panel of cancer cells (Holbeck et al., 2010). This makes the MCF-7 cell line an ideal experimental subject for MR overexpression studies.

1.7 The mineralocorticoid receptor signaling pathway

The mineralocorticoid receptor (MR) is a transcription factor (TF) that regulates gene expression in a ligand-dependent manner. **Figure 2** shows that MR interacts with chaperone proteins that are found in the cytoplasm like the heat shock protein 90 (hsp90) and FK506-binding protein 51 (FKBP51). MR remains inactive but in a conformation that is ready for ligand binding. When the ligand binding happens, the receptor changes its conformation to replace FKBP51 with FKBP52 and translocate into the nucleus. Further conformation changes allow the separation from the chaperone proteins and enables dimerization and interaction with DNA or other TFs (Griekspoor et al., 2007; Yang & Fuller, 2012).

The MR's physiological ligands are aldosterone and cortisol; it also has two antagonists, eplerenone and spironolactone (Fuller et al., 2019; Sica, 2015). Since the discovery of aldosterone in the 1950s there has been a lot of progress in the understanding of the relationship between aldosterone and MR (Messaoudi & Jaisser, 2011; Simpson et al., 1953). The current understanding of MR signaling upon aldosterone binding is shown in **Figure 2**. Early experiments with MR and its ligands show that aldosterone-bound MR is more stable MR bound by spironolactone due to reduced proteolysis(Trapp & Holsboer, 1995). Spironolactone interacts with MR in a "passive" version of antagonism: this type of antagonism involves a conformational change that prevents receptor dimerization and DNA binding ability of the receptor (Fuller et al., 2019).

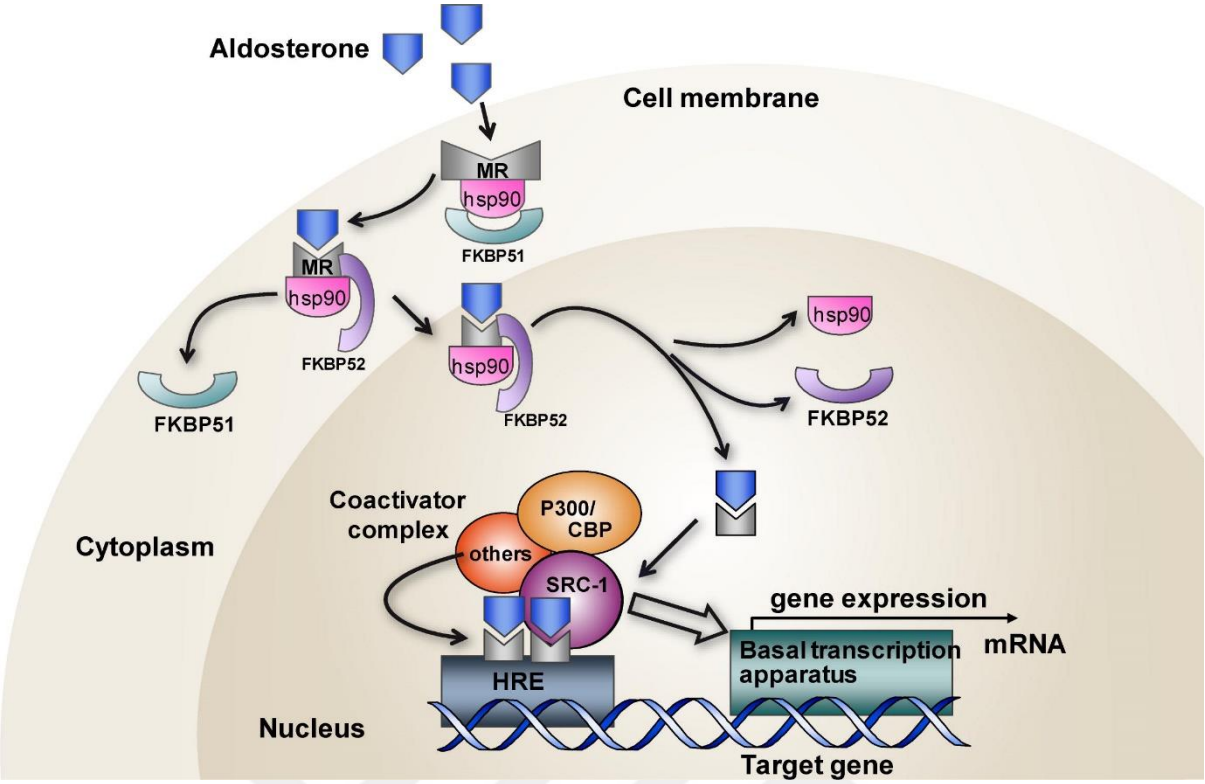


Figure 2. A schematic representation of mineralocorticoid receptor (MR)-mediated transactivation. Reproduced with permission (Yang & Fuller, 2012).

1.8 MR signaling in breast cancer

As stated previously, breast cancer cell lines do not express or express the mineralocorticoid receptor at a very low level. Studies regarding MR signaling in breast cancer show that inducing the expression of MR and its co-activation with the retinoic acid receptor (RAR) with their respective ligands is correlated with tumor growth inhibition and a reverse Warburg effect, meaning that cancer cells induce neighboring cells to produce energy-rich metabolites for them (Doan et al., 2020; Pavlides et al., 2009). This effect is consistent with findings in other types of cancer (Nie et al., 2015). Co-activation of MR and the other nuclear receptors glucocorticoid (GR) and progesterone receptor (PR) has also been shown to have a inhibitory effect on breast cancer cell growth (Leo et al., 2004). Given this effect, spironolactone was thought to be linked

with higher risk of breast cancer and other gynecologic cancers because of its antagonistic role in the MR signaling pathway, but a retrospective study of 2.3 million women showed that spironolactone treatment has some positive association with increased risk, but the increase itself was weak (Biggar et al., 2013; C. Wei et al., 2020).

Chapter 2. Aims

RNA sequencing techniques generate high-throughput data and have become mainstream in biological sciences. The development of these techniques has led to the usage of large datasets to conduct biological research where handling and analyzing these datasets can be demanding in statistics knowledge and programming skills. These skills can be difficult to develop for biology-oriented researchers, therefore assistance is needed to help analyze and get insights from large amounts of high-throughput sequencing data.

Previously in our laboratory, Huma Shehwana developed a shiny web tool called CompariZome for comparative statistics between public human and zebrafish expression datasets (Shehwana, 2017). More recently, Tuğberk Kaya developed an application called CoEX-PM that helps users analyze available microarray data from the GEO database (Kaya, 2018). Other useful Shiny applications developed in our laboratory include ZenoFishDb, a database application for xenotransplantation zebrafish studies; CoVrimer, an alignment tool for SARS-CoV-2 primer sequences and selection of conserved or degenerate primers; and SmulTCan, a web tool that can analyze and visualize gene multivariable survival and hazard ratio profiles across available TCGA-PANCAN datasets (Ozhan et al., 2021; Targen et al., 2020; Vural-Ozdeniz et al., 2021).

Following these projects, the specific purposes of this thesis are to:

1. Design and develop a Shiny application using the R programming language and the {shiny} package that can summarize, normalize, perform differential gene expression (DGE) analysis, and visualize results for our MR in breast cancer study and for user-provided RNA-seq read count data. To do this, the application includes visualization and analysis modules for exploratory RNA-seq data analysis, data

normalization and filtering, differential gene expression analysis, other analyses based on expression data, and downstream functional analyses.

2. Compare the DGE analysis results of four R libraries including {DESeq2}, {edgeR}, {limma}, and {NOISeq}. To do this comparison, Venn diagrams, UpSet plots, and tabular results are utilized to see the difference in statistical results such as mean expression values, fold-change, p-values, and adjusted p-values. Differential gene expression analysis is the focus of the application and because there are a variety of R libraries and statistical methods for analyzing RNA sequencing data, it is important to compare results from different methods and choose a combination of results to confirm differential expression and come to conclusions based on RNA-seq expression data.
3. Demonstrate a case study utilizing the Shiny application to perform DGE analysis and visualize the RNA sequencing data from a study of MR overexpression in breast cancer cell lines after treatment with the hormone aldosterone and its antagonist spironolactone. Using the different DGEA methods in the application and accounting for the small number of samples for each group in our data, the aim of this part is to optimize the application to be able to analyze data with no replicates to gain some statistically significant results. This can be achieved by using simulation (NOISeq-sim method) or linear models (edgeR and limma) to get insights about the differential expression of features in the dataset.

Chapter 3. Materials and Methods

3.1 RNA-seq data acquisition

Cell culture studies using the MCF-7 and ZR-75-1 cell line, their transfection, overexpression with mineralocorticoid receptor (OV) or empty vector (EV), hormone treatment, and RNA extraction were performed by Dr. Seniye Targen (Targen, 2020). RNA sequencing service was purchased from Genomics Core Facility EMBL, Germany. The RNA samples were run on the Illumina NextSeq 500 System using 75bp long single end reads and raw FASTQ files were provided with the service.

3.2 Quality control, alignment, and quantification of RNA-seq data

FASTQ files were uploaded to the Seven Bridges Cancer Genomics Cloud, a cloud platform that facilitates analysis, storage, and computation of high-throughput sequencing data (Lau et al., 2017). Quality control was performed using the FASTQC tool (version 0.11.4) developed by the Babraham Institute (Andrews, 2015). Reports for every FASTQ file for 14 samples were generated by FASTQC, and MultiQC tool (version 1.12) was used to summarize FASTQC statistics and plots (Ewels et al., 2016).

FASTQ files were aligned using the STAR RNA-seq aligner tool (version 2.5.4b) using an index file pre-created by STAR aligner using GRCh37.75 human genome reference, the respective splice junction gene transfer format (GTF) file, and the default settings for single end reads (Dobin et al., 2013).

BAM files created from the alignment step were used as input files for HTSeq-count (version 0.6.1) to create a read count matrix with samples (columns) and features (rows) (Anders et al., 2015).

3.3 Statistical analyses and shiny application development in R

Analyses presented in this thesis were performed using the R programming language, version 4.2.1 and the RStudio IDE version 2022.07.1+554 (R Core Team, 2020; RStudio Team, 2022). The development of the application described in this thesis has been done using the {shiny} package, a web application framework for R (version 1.7.2) (Chang et al., 2022). Additional packages used in enhancing the UI/UX part and all functionality of the application were listed in **Table 5**. The analyses applied to the case study datasets in terms of functions and arguments have been provided in more detail in the following sections.

3.4 Exploratory data analysis of RNA-seq data

The first part of the exploratory data analysis was performed using the generated RNA-seq read count matrix as input and included a library size bar plot, log2-transformed counts boxplot, and a multidimensional scaling plot using the barplot() and boxplot() functions from base R packages, and the plotMDS() function from {limma} package (Ritchie et al., 2015).

The second part of the exploratory data analysis was accomplished using functions from the {NOISeq} package. As a prerequisite, the {biomaRt} package was used to get annotation information on the features in the dataset such as gene symbol, transcript length, GC content percentage, and gene biotype (Durinck et al., 2009). The functions in the {NOISeq} package

make use of the annotation data to plot biodetection plots, counts distribution plots for samples and biotypes, sensitivity plot, and an RNA composition plot (Tarazona et al., 2015).

Table 5. Software and package resources table.

Resource name	Reference	Identifier
RNA-seq quality control, alignment, quantification		
Seven Bridges Cancer Genomics Cloud	Lau et al. 2017	https://www.cancergenomicscloud.org/
FASTQC (0.11.4)	Andrews, 2015	https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
MultiQC (1.12)	Ewels et al., 2016	https://multiqc.info/
STAR (2.5.4b)	Dobin et al., 2013	https://github.com/alexdobin/STAR
GRCh37.75 human genome reference	Genome Reference Consortium, 2009	https://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.13/
GRCh37.75 human genome splice junction GTF file	Genome Reference Consortium, 2009	https://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.13/
MR RNA-seq data	Targen, 2020	http://repository.bilkent.edu.tr/handle/11693/54266 This thesis
Software and packages		
R (4.2.1)	R Core Team, 2020	https://www.r-project.org/
RStudio IDE (2022.07.1+554)	RStudio Team, 2022	https://www.rstudio.com/
Python (3.10.6)	Van Rossum and Drake, 2009	https://www.python.org
Shiny application user interface		
shiny (1.7.2)	Chang et al., 2022	https://shiny.rstudio.com/
shinydashboard (0.7.2)	Chang et al., 2021	https://CRAN.R-project.org/package=shinydashboard
shinybusy (0.3.1)	Meyer et al., 2022	https://cran.r-project.org/web/packages/shinybusy/index.html
shinybrowser (1.0.0)	Attali, 2022	https://cran.r-project.org/web/packages/shinybrowser/index.html
shinyWidgets (0.7.2)	Perrier et al., 2022	https://cran.r-project.org/web/packages/shinyWidgets/index.html
Exploratory data analysis		
biomaRt (2.52.0)	Durinck et al., 2009	https://bioconductor.org/packages/release/bioc/html/biomaRt.html
limma (3.52.2)	Ritchie et al., 2015	https://bioconductor.org/packages/release/bioc/html/limma.html
NOISeq (2.40.0)	Tarazona et al., 2015	https://bioconductor.org/packages/release/bioc/html/NOISeq.html
Data normalization, filtering, differential gene expression analysis		
edgeR (3.38.4)	Robinson et al., 2010	https://bioconductor.org/packages/release/bioc/html/edgeR.html
limma (3.52.2)	Ritchie et al., 2015	https://bioconductor.org/packages/release/bioc/html/limma.html
DESeq2 (1.36.0)	Love et al., 2014	https://bioconductor.org/packages/release/bioc/html/DESeq2.html
NOISeq (2.40.0)	Tarazona et al., 2015	https://bioconductor.org/packages/release/bioc/html/NOISeq.html
affy (1.74.0)	Gautier et al., 2004	https://www.bioconductor.org/packages/release/bioc/html/affy.html
plotly (4.10.0)	Sievert et al. 2021	https://cran.r-project.org/web/packages/plotly/index.html
Comparison of DGEA methods		
VennDetail (1.12.0)	Guo and McGregor, 2022	https://www.bioconductor.org/packages/release/bioc/html/VennDetail.html
Other analyses		
heatmaply (1.3.0)	Galili et al., 2018	https://github.com/talgalili/heatmaply
corrplot (0.92)	Wei and Simko, 2021	https://github.com/taiyun/corrplot
visNetwork (2.1.0)	Almende B. V. and Contributors et al., 2021	https://cran.r-project.org/web/packages/visNetwork/index.html
Downstream analyses		
clusterProfiler (4.4.4)	Wu et al., 2021	https://bioconductor.org/packages/release/bioc/html/clusterProfiler.html
enrichplot (1.16.1)	Yu, 2022	https://bioconductor.org/packages/release/bioc/html/enrichplot.html
visNetwork (2.1.0)	Almende B. V. and Contributors et al., 2021	https://cran.r-project.org/web/packages/visNetwork/index.html

3.5 Data normalization and filtering

Data normalization was done using four available methods: counts per million (CPM), log2-transformed CPM, voom transformation, and DESeq normalization (Law et al., 2014; Love et al., 2014; Robinson & Oshlack, 2010). CPM, and log2-CPM normalization were done using the cpm() function from the {edgeR} package, voom transformation was done using the voom() function from the {limma} package, and the DESeq normalization was done using the estimateSizeFactors(), sizeFactors(), and counts() functions from the {DESeq2} package (Love et al., 2014; Ritchie et al., 2015; Robinson et al., 2010). Normalized read counts were visualized in a density plot and a counts boxplot using plotDensity() function from the {affy} package and the boxplot() function from base R (Gautier et al., 2004).

Data filtering was done after data normalization using the a modified version of the filterByExpr() function in the {edgeR} package. The function follows the filtering strategy described by Chen *et al*. (Chen et al., 2016). In simplified terms, the strategy keeps genes that have CPM values above *k* in *n* samples, where *k* is the count threshold determined by the user (10 by default) and *n* is the smallest group sample size in the dataset. Additionally, the total sum of CPM values for each gene must be above a threshold (15 by default).

3.6 Differential gene expression analysis (DGEA)

Differential gene expression analysis was performed using four available methods using functions included in the respective R packages: {DESeq2}, {edgeR}, {limma}, and {NOISeq}. Because of two prerequisites that make DESeq2 different from the other packages, comparison of groups that do not have replicates was not possible, therefore only the edgeR, limma, and NOISeq results were presented at times. The filtered raw read counts were used as input because all four DGEA methods have internal functions that normalize read counts and

perform statistical tests. The results included statistics on average normalized counts, log-transformed fold-change values, and p-values among other method-specific statistics. The p.adjust() function from base R with the Bonferroni-Hochberg method was used to adjust p-values for independent multiple tests (Benjamini & Hochberg, 1995). The results were visualized in a volcano plot using a {plotly} scatterplot function (Sievert et al., 2021).

3.7 Comparison of DGEA methods

DGEA methods were compared in two parts, i.e., using a Venn diagram and an UpSet plot. The Venn diagram and UpSet gene sets and visualizations were created using functions from the {VennDetail} package (Guo & McGregor, 2018/2022). Results from the differential gene expression analysis were filtered with adjusted p-values <= 0.05 and |log2FC values| > 1 and input as gene sets grouped based on log2FC values (upregulated or downregulated gene sets). The Venn diagram was used to compare methods one-to-one, whereas the UpSet plot was used to compare all available methods simultaneously.

3.8 Other analyses and visualizations

Results from DGE analyses (log2FC or p-values) were filtered and selected from the results table (adjusted p-values <= 0.05) before being plotted as a heatmap using the heatmaply() function from the {heatmaply} package (Galili et al., 2018). The heatmap function was used with the following arguments: distance method = “euclidean”, distance function = “pearson”, hierarchical clustering function = “complete”.

The principal component analysis (PCA) was performed using DESeq-normalized expression data. The data were selected using the top 75% of most highly expressed genes (ranked by

average expression across all samples). The analysis was performed using the prcomp() function from base R. A grid of scatterplots plotting the four first principal components was graphed using the plot_ly() function from the {plotly} package with a grid layout (Sievert et al., 2021).

The correlation analysis was done using DESeq-normalized expression data and the selection was done either by importing a filtered table or manually inputting genes from the dataset. The cor() function from base R was used to calculate correlation coefficients and cor.mtest() function from the {corrplot} function to calculate p-values (T. Wei & Simko, 2011/2021). A heatmap of correlation coefficients was drawn using the heatmaply_cor() function from the {heatmaply} package. Scatter points in the heatmap were colored by correlation coefficient and the size was based on -log10(p-values). A gene network was drawn using the visNetwork() function from the {visNetwork} package (Almende B. V. and Contributors et al., 2021). The network nodes were named after the selected genes, the node size was based on average expression value of the gene, the edge thickness was based on the absolute value of correlation coefficient, whereas the color reflected the correlation coefficient.

3.9 Downstream functional analyses

The over-representation analysis (ORA) was based on a selection of genes filtered from the results table (adjusted p-values <= 0.05). The ORA enrichment algorithm was run using the enrichGO() and enrichKEGG() functions from the {clusterProfiler} package (Wu et al., 2021). The ORA was run for three sub-ontologies (MF: molecular function, BP: biological process, and CC: cellular component) using annotation data from the Gene Ontology Resource and Kyoto Encyclopedia of Genes and Genomes Database (Ashburner et al., 2000; Kanehisa & Goto, 2000; The Gene Ontology Consortium, 2021). The results were plot into the dot plot,

tree plot, terms network, and heatmap using functions from the {enrichplot}, {visNetwork}, and {heatmaply} packages (Almende B. V. and Contributors et al., 2021; Galili et al., 2018; Yu, 2022).

The gene set enrichment analysis (GSEA) was based on a selection of genes filtered from the results table (adjusted p-values <= 0.05). The GSEA enrichment was performed using the gseGO() and gseMKEGG() functions from the {clusterProfiler} package (Wu et al., 2021). The GSEA was run for three sub-ontologies: MF, BP, and CC using annotation data from GO and KEGG databases. The results were plot into the dot plot, tree plot, terms network, ridgeline plot, and running score plot using functions from the {enrichplot} package (Yu, 2022).

Chapter 4. Results

4.1 Shiny application for RNA-seq data analysis

The shiny application described in this thesis comprises five main modules (tabs): “Data summary”, “Normalization step”, Filtering step”, “DGE analysis”, and “Visualization”. A constant sidebar panel where the control inputs are found is used for the initialization of the application. The “Data summary” tab serves as the start page of the application, where upon the initialization of the application, several plots, summary statistics, and a table make up the RNA-seq exploratory data analysis part of the tool. The “Normalization step” tab contains other sub-tabs where each of the four methods for RNA-seq data normalization are displayed in a results table and plots. The “Filtering step” tab contains more input related to filtering the dataset based on expression values thresholds. The filtered dataset is summarized similarly to the exploratory data analysis part and acts as input for the “DGE analysis” tab. The “DGE analysis” tab contains up to four sub-tabs, each containing the results and visualization of differential gene expression analysis done using four different methods and R libraries. The “Visualization” tab contains sub-tabs that include functionality for plotting heatmaps, PCA plots, correlation analysis, and two downstream functional analyses: over-representation analysis and gene set enrichment analysis. All functions of the application are described in detail throughout this section of the thesis. A simplified flow chart of the application is shown in **Figure 1**.

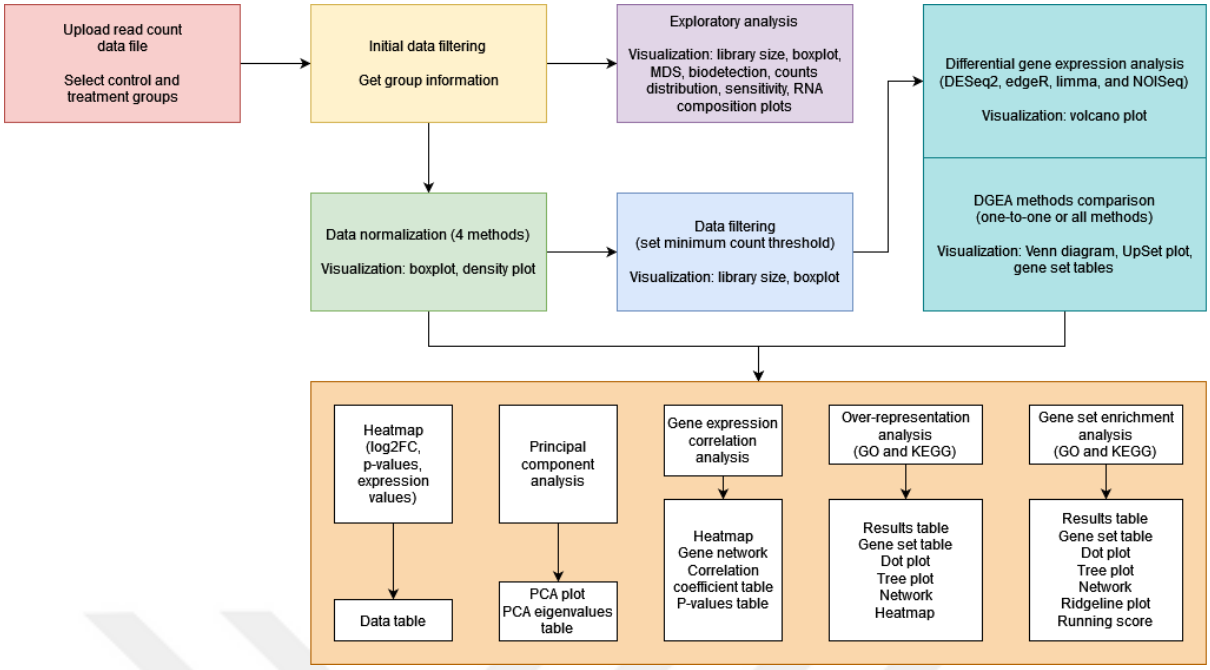


Figure 3. Simplified application flow chart.

4.1.1 Application initialization and control panel

The application starts in the "Data summary" tab view. This view includes the sidebar control panel shown in **Figure 2**. To initialize the application, the user can either choose an available dataset (**Fig. 2A**) or upload a comma-separated values (CSV) file that contains RNA-seq count data (**Fig. 2B**). After choosing a dataset (or uploading a complying file), the application displays drop-down menus for a control group and a treatment group. In **Figure 2A**, the control group is "MCF_7_EV" (empty vector, control), and the treatment group is "MCF_7_OV" (cells transfected that overexpress the MR gene). Pressing the "Analyze data" button starts the application, firstly showing results of the exploratory data analysis (**Fig. 3**), then data normalization (**Fig. 6**) and DGE analysis (**Fig. 7**).

4.1.2 Input data file format and requirements

For the application to function correctly, the input file must meet a few requirements as listed below. These requirements are met by the available dataset file (MR_in_BC_data.csv); the first few rows of this file are shown in **Figure 2C**.

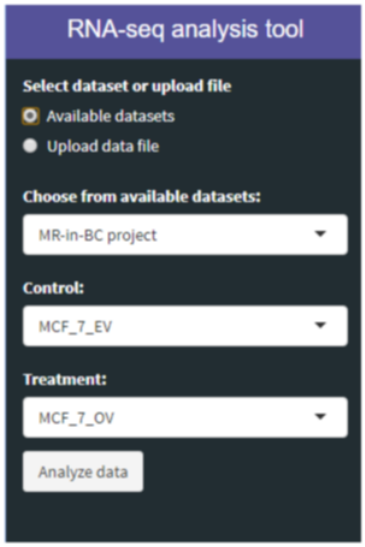
The input data file requirements are the following:

1. The file must be in the comma-separated file (CSV) format.
2. The file must have at least three columns and three rows.
3. Of the three columns, one of them must contain information on feature IDs (examples: ENSEMBL_GENE_STABLE_ID, HGNC_GENE_SYMBOL). One of them must be the Ensembl Gene stable ID of the feature. There can be more than one feature ID column and all IDs must be unique.
4. The other two columns must contain raw (unnormalized) read counts for two samples, in integer format. The samples must have unique IDs.
5. Of the three rows, the first one must contain the column names (feature ID type(s) and sample name for each remaining column).
6. The second row must contain metadata information on the columns (the metadata string for the feature ID must start with "ID_", and the sample metadata strings should describe the group the sample is part of, and therefore there must be two different group names).
7. The third required row must include the RNA-seq feature and raw read count for each sample, in integer format.

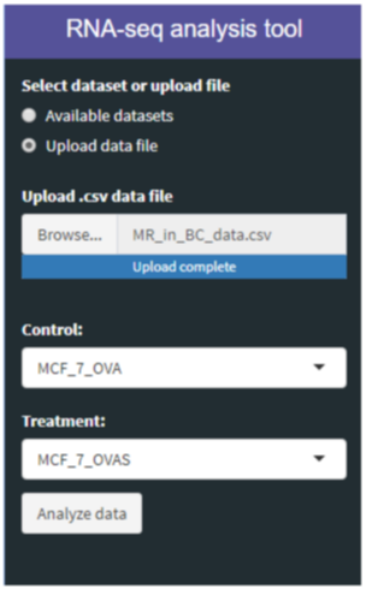
This ruleset is strict and determined in a way such that it requires the user to have only one input file instead of having multiple files including a separate metadata file. In particular, the Ensembl Gene stable IDs are required because of the functions that use these IDs to query other

annotation data that is used in the exploratory data analysis part ({NOISeq} package functionality) and in the downstream functional analyses (ORA and GSEA tabs). The application is optimized in a way that having a second column with HGNC gene symbols would save time and make results human-readable, and such as column is added if missing. The sample metadata strings are important and there must be at least two different ones describing two different groups/treatments, otherwise the analyses cannot run correctly. Lastly, the expression values must be unnormalized and integers, because there is already normalization functionality in the application, and because the DESeq2 method requires integer values.

A



B



C

Ensembl_ID	HGNC_symbol	MCF_7_EV	MCF_7_OV	MCF_7_EVA	MCF_7_EVAS	MCF_7_OVA1	MCF_7_OVA2	MCF_7_OVAS1	MCF_7_OVAS2	ZR_75_1_EV	ZR_75_1_OV	ZR_75_1_EVA	ZR_75_1_OVA
ID_1	ID_2	MCF_7_EV	MCF_7_OV	MCF_7_EVA	MCF_7_EVAS	MCF_7_OVA	MCF_7_OVA	MCF_7_OVAS	MCF_7_OVAS	ZR_75_1_EV	ZR_75_1_OV	ZR_75_1_EVA	ZR_75_1_OVA
ENSG00000000003	TSPAN6	0	0	0	1	0	0	0	0	4	3	1	0
ENSG00000000005	TNMD	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000000419	DPM1	6	16	12	15	13	17	12	8	46	25	28	26
ENSG00000000457	SCYL3	568	628	410	330	350	445	386	400	436	514	584	428
ENSG00000000460	C1orf112	416	391	314	434	228	353	375	343	616	653	868	683
ENSG00000000938	FGR	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000000971	CFH	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000001036	FUCA2	16	51	18	15	20	20	17	23	39	31	52	9
ENSG00000001084	GCLC	17	20	19	10	14	22	15	16	16	14	17	5
ENSG00000001167	NFYA	27	27	17	16	13	23	17	9	25	31	22	11
ENSG00000001460	STPG1	46	56	40	36	37	32	43	37	60	56	91	60
ENSG00000001461	NIPAL3	6	6	4	3	3	1	7	0	43	40	60	13
ENSG00000001497	LAS1L	30	41	12	10	18	26	21	20	93	80	93	24
ENSG00000001561	ENPP4	0	2	0	0	0	1	1	0	4	3	6	1
ENSG00000001617	SEMA3F	11	54	5	15	11	30	28	15	113	83	99	16
ENSG00000001626	CFTR	2	0	0	1	0	0	0	0	3	2	0	0
ENSG00000001629	ANKIB1	5	30	2	8	2	9	20	4	43	29	32	11
ENSG00000001630	CYP51A1	3	9	10	3	8	9	7	6	2	4	1	0
ENSG00000001631	KRIT1	12	22	15	8	13	7	12	7	51	38	39	10
ENSG00000002016	RAD52	3	6	5	5	6	12	10	10	53	43	48	38
ENSG00000002079	MYH16	130	113	77	56	40	72	92	84	23	22	18	15
ENSG00000002330	BAD	27	71	17	26	26	30	21	19	16	34	35	7
ENSG00000002549	LAP3	3	20	5	5	4	11	7	6	33	13	19	3
ENSG00000002586	CD99	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000002587	HS3ST1	0	0	0	0	0	0	0	0	0	2	2	6
ENSG00000002726	AOC1	1	1	1	0	0	1	0	0	1	4	3	2
ENSG00000002745	WNT16	0	0	0	0	0	0	0	3	7	1	6	4
ENSG00000002746	HECW1	1	0	0	0	0	0	1	0	0	0	0	0
ENSG00000002822	MAD1L1	14	63	10	27	11	28	34	7	22	19	24	8
ENSG00000002834	LASP1	13	62	5	17	16	27	32	15	226	171	183	49

Figure 4. Application sidebar panel and input data file format. **A-B.** Application sidebar panel showing a group of inputs that are used for selecting (**A**) or uploading datasets (**B**), selecting groups for analysis, and the "Analyze data" button that initializes visualization and data analysis in the application. **C.** An example of the input data file (MR_in_BC_data.csv file) that complies with the requirements listed in this thesis.

4.1.3 RNA-seq exploratory data analysis

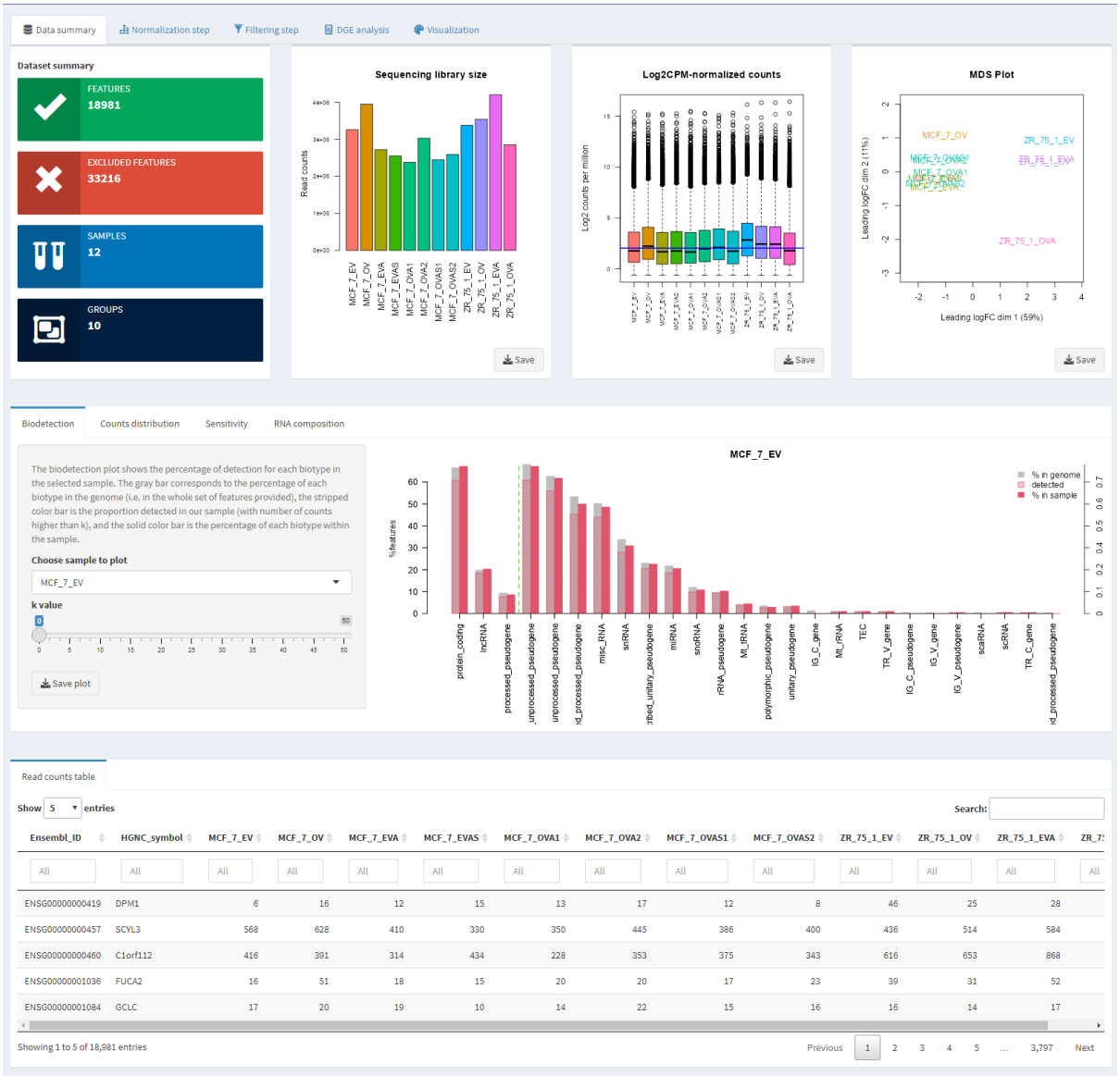


Figure 5. The "Data summary" tab. This tab contains a dataset summary part, three summarizing plots (library size bar plot, log2CPM-normalized counts boxplot, multidimensional scaling plot), and the NOISeq part that includes a biodetection plot, counts distribution plot, sensitivity plot, and RNA composition plot. A read counts table is printed out for the user to check whether the data is read in correctly.

After initialization, the application firstly discards “zero genes”, defined here as features that are not expressed in any of the samples in the dataset (all values are 0). An example of a “zero

gene" is TNMD shown in **Figure 2C**. These genes are discarded because they do not contain any valuable information and slow down calculations. The dataset summary statistics shows the number of genes that are left ("Features") and the number of discarded zero genes ("Excluded features") in **Figure 3**.

The metadata row that is required in the input file is processed and as a result the summary part shows the number of samples and groups in the dataset ("Samples" and "Groups" in **Fig. 3**). Internally, the application creates a group_information object that contains information on samples IDs, group names, group name for each sample, assigns a number to each group and sample, and checks whether groups have replicates or not. An example of the group_information list object is shown in **Figure 4**.

```
> init_sample_names <- c("ENSEMBL_GENE_ID", "GENE_SYMBOL", "Control_1", "Control_2", "Control_3", "Treatment_1", "Treatment_2", "Treatment_3")
> init_group_names <- c("ID_1", "ID_2", "Control", "Control", "Control", "Treatment", "Treatment", "Treatment")
> get_group_info(init_sample_names, init_group_names)
$sample_ids
[1] "Control_1"   "Control_2"   "Control_3"   "Treatment_1" "Treatment_2" "Treatment_3"

$group_names
[1] "Control"   "Treatment"

$sample_group_names
[1] "Control"   "Control"   "Control"   "Treatment" "Treatment" "Treatment"

$group_numbers
[1] 1 2

$sample_group_numbers
[1] 1 1 1 2 2 2

$group_has_replicates
[1] TRUE TRUE
```

Figure 6. group_information list object. The get_group_info() function takes the first row (init_sample_names) and second row (init_group_names) of the input file and returns a list of information on samples and groups that is later used in the normalization and DGE analysis tabs.

The sequencing library plot, normalized counts boxplot, and MDS plot provide a visual summary of the dataset. The sequencing library plot gives insight on the depth of sequencing and quality of the dataset, the normalized counts boxplot shows the distribution of the expression values compared to the overall mean value (shown by the blue horizontal line), and

the MDS plot visualizes the similarities of individual samples of the dataset. These three plots can help the user determine the quality of the dataset. In general, a good dataset has large library sizes, equally distributed counts, and samples of the same group cluster in the MDS plot.

The {NOISeq} package provides functions that are used for data quality checks. The functions dat() and expo.plot() are used in the second part of the "Data summary" tab to calculate and draw a biodetection plot, a counts distribution plot, a sensitivity plot, and an RNA composition plot (**Fig. 5**). These functions use information obtained from Ensembl/Biomart regarding gene symbols, gene biotypes, and transcript length. This information is saved in a file called "metadata_information.csv" in order to avoid slowing down the process by querying the Ensembl database on-the-fly.

The biodetection plot (**Fig. 5A**) shows the percentage of each biotype in the genome (gray bar), the proportion of each biotype detected with more than *k* counts (stripped color), and the percentage of each biotype in the sample (solid color). The most abundant biotypes are on the left side of the green dotted line. Increasing the *k* value will lower the proportion of detected biotypes, and higher proportions of detected biotypes with higher *k* value means that certain biotypes are highly expressed in the selected sample.

The counts distribution box plot (**Fig. 5B**) shows how the expression values are distributed within the sample. The values plotted are CPM-normalized, and the numbers on top of the plot indicate the number of features in each biotype. Another type of counts distribution box plot is possible where the box plots show the distribution of expression values of one biotype across all samples in the dataset.

The sensitivity plot (**Fig. 5C**) shows the proportion of features that would remain if the user filtered out features using a cut-off value for each sample. In the plot, pre-determined threshold values of 0, 1, 2, 5, and 10 CPM are applied, and the bars indicate the percentage of features

that remain after each filtering threshold is used. Numbers on top of the bars indicate the library size in millions of reads. This plot can help to decide the correct threshold to use in the "Filtering step" tab.

The RNA composition plot (**Fig. 5D**) serves to check the bias in the distribution of sequencing reads across features. The dat() function with argument type = "cd" calculates an M value as $\log2(counts_{sample} / counts_{reference})$. The M values are plotted as a distribution for all samples. If the bias in sequence reads distribution is not present the median M value (colored vertical line) for the comparison must be 0. In most cases, this is not so and therefore data normalization is necessary before analyzing differential expression.

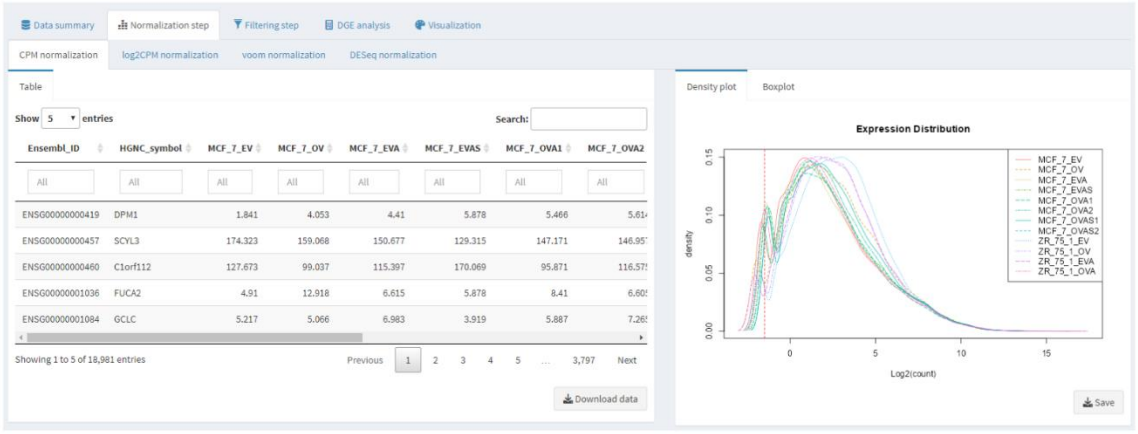
4.1.4 Data normalization

The "Normalization step" tab (**Fig. 6**) contains four sub-tabs, each named after one of the available normalization methods. The four methods available are counts-per-million (CPM), log2CPM, voom transformation, and DESeq normalization method. All four methods can be used simultaneously to produce normalized counts. The distribution of the normalized counts across all samples is visualized in the density plot (**Fig. 6A**), and in box plots (**Fig. 6B**). While CPM and log2CPM are very basic normalization methods that only consider library sizes, voom transformation and DESeq normalization use normalization/size factors to turn raw library sizes into effective library sizes. voom transformation is also log2-transformed. Both voom transformation and DESeq normalization stabilize counts distribution close to the overall dataset mean expression value (shown in blue horizontal line in **Fig. 6B**). The normalized data is further used as input in the "Filtering step" tab, but also in "Heatmap", "PCA", and "Correlation analysis" sub-tabs.

Figure 7. The four plots drawn using functions from the {NOISeq} package. **A.** A biodetection plot showing the percentage of detected features separated by biotype. **B.** A counts distribution box plot showing expression values distribution for each biotype in one sample (MCF_7_EV). **C.** A sensitivity plot showing the proportion of features within each sample that have more than 0, 1, 2, 5, and 10 CPM. **D.** An RNA composition plot showing the distribution of

sequencing reads across features for all MCF-7 samples compared to a reference sample (MCF_7_EV).

A



B

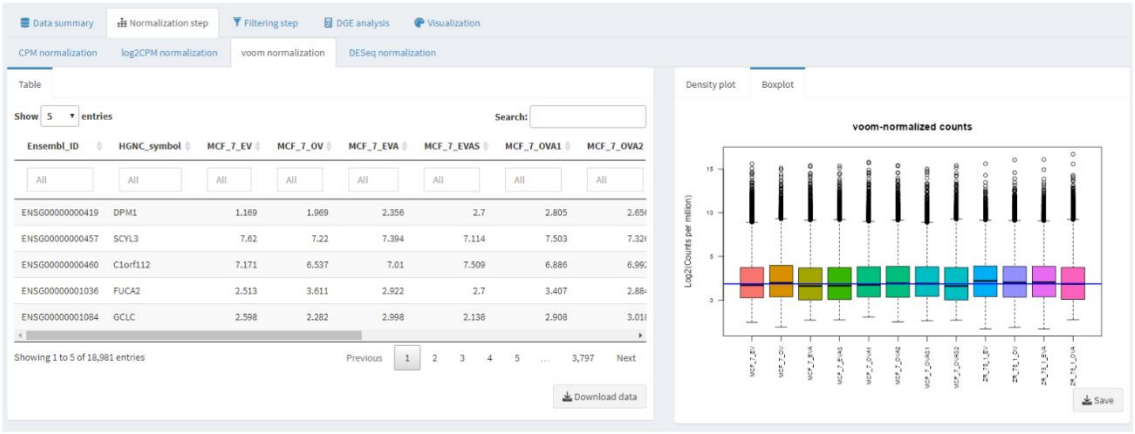


Figure 8. The "Normalization step" tab and two sub-tabs. **A.** The “CPM normalization” sub-tab showing the normalized counts in the results table and the density plot for all samples in the dataset. **B.** The “voom normalization” sub-tab showing the normalized counts in the results table and the counts distribution box plot for all samples in the dataset.

4.1.5 Data filtering

The “Filtering step” tab (**Fig. 7**) is the part of the application where the user can import read counts and use count thresholds to remove lowly expressed features in the dataset. The tab

contains a control panel with inputs, where first the user must select to import either raw read counts (as inputted during initialization) or normalized counts from any of the four methods in the normalization step. Two filtering thresholds are available and the filtering function is an implementation of filterExpression(), a modified version of the filterByExpr() function from the {edgeR} package. This function uses a design matrix created internally using the group_information object created before and two inputs from the user: min.count, which is defined as the minimum count required for at least some of the samples, and min.total.count, which is defined as the minimum total count required across all samples. This is a strategy first described by Chen *et al*. (Chen et al., 2016), which keeps the genes that:

1. have counts above min.count in *n* number of samples, where *n* is the smallest group sample size, and
2. have a total sum of counts across all samples above min.total.count.

Figure 7 shows a demonstration of this filtering step where the min.count has been set to 1, min.total.count to 12 and imports voom normalized count data, which results in the removal of 3,002 genes out of 18,981 (15.8%) that can be considered lowly expressed genes. Summary statistics on the filtering step are displayed and include number of remaining and removed features, the average removed raw read count, and average reduction in library size. The remaining features in the dataset are visualized in the library size plot and filtered counts distribution box plot. Tables showing the filtered dataset (left) and removed features (right) give the user the ability to explore the dataset before analyzing it in the next tabs.

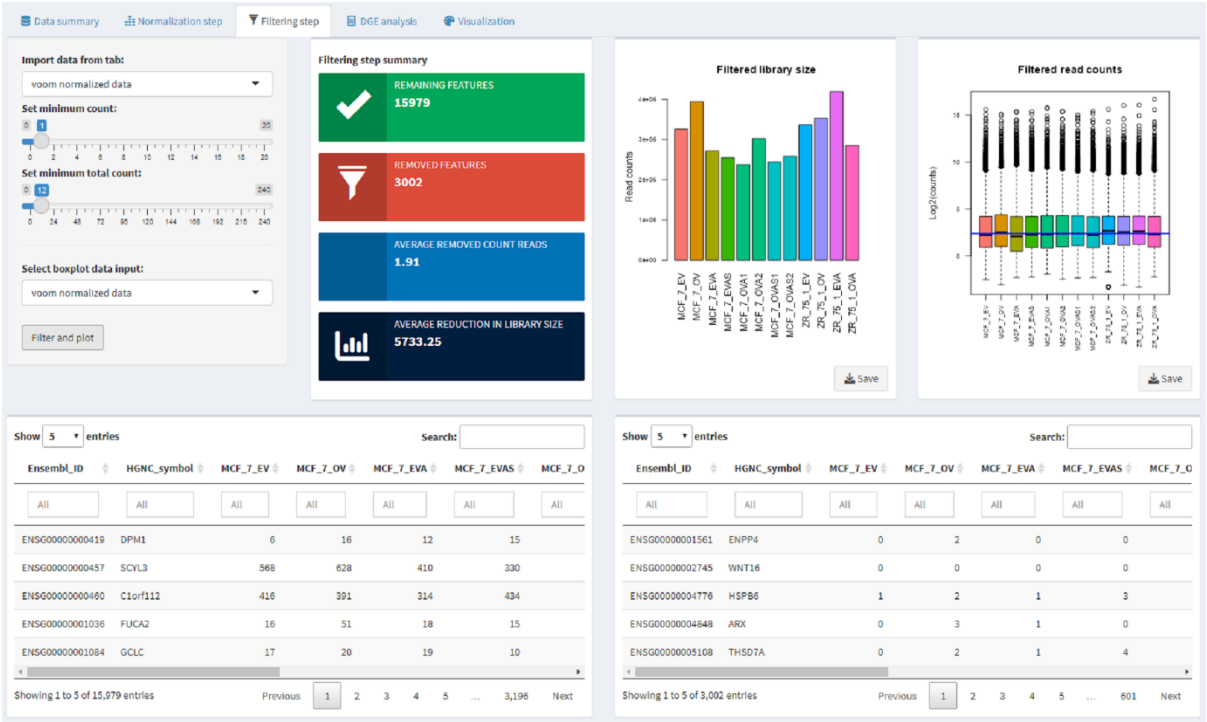


Figure 9. The "Filtering step" tab. The tab contains a sidebar panel with filtering options, a summary of the filtering process, two summarizing plots (library size plot, and counts distribution boxplot) and two tables showing the remaining and removed features.

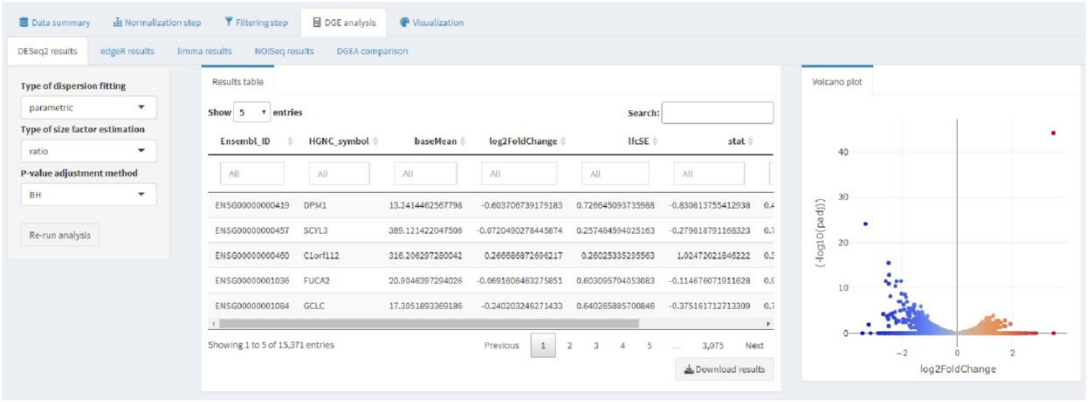
4.1.6 Differential gene expression analysis

4.1.6.1 DGE analysis results tabs

There are four available methods for DGE analysis: DESeq2, edgeR, limma, and NOISeq. The “DGE analysis” tab contains four sub-tabs where the results and visualization of results from all four methods are presented. **Figure 8** shows two of these sub-tabs for each DESeq2 (**Fig. 8A**) and edgeR (**Fig. 8B**). Each method is different from one another and therefore there are different options on the sidebar panel shown in each sub-tab. For DESeq2, users can choose options regarding the type of dispersion fitting (parametric, local, or mean), the type of size factor estimation (ratio, poscounts, or iterate), and the p-value adjustment method (Benjamini

& Hochberg, Holm, Hochberg, Hommel, Bonferroni, Benjamini & Yuketieli, or none). For the other three methods, users can choose option for trend method for dispersion estimation (locfit, none, movingave, loess, locfit.mixed), and the p-value adjustment method (same options as in DESeq2 tab). The analyses are run when pressing the “Filter and plot” button on the “Filtering step” tab, but there is an option to re-run analyses if the options are changed. The results of the DGE analyses are shown in separate tables that can be filtered using the numeric value columns (log2 fold change, p-values, adjusted p-values and other method-specific statistics) or searched using the available feature IDs e.g., Ensembl Gene stable IDs or HGNC gene symbols. A volcano plot is drawn to visualize the distribution of log2 fold change values and -log10-tranformed adjusted p-values. Options to download the tables and plots are available.

A



B

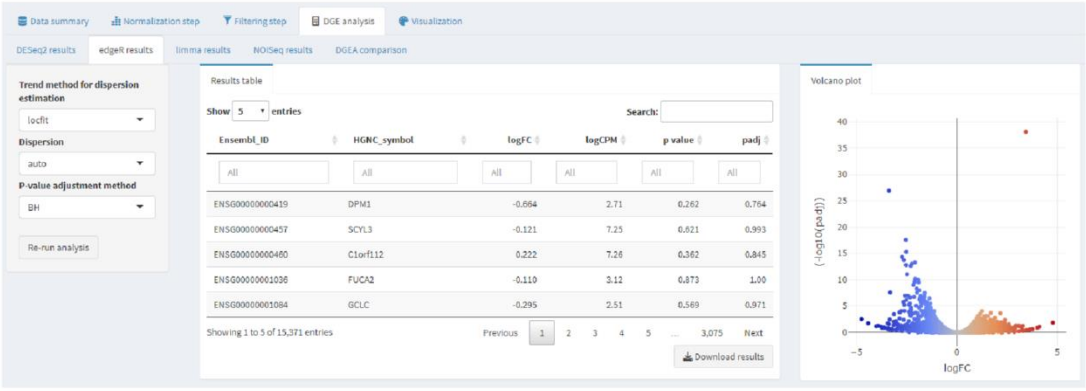


Figure 10. The "Differential gene expression analysis" tab. Two different sub-tabs are shown. **A.** DESeq2 results sub-tab showing the options, results and volcano plot visualization. **B.** edgeR results sub-tab showing the options, results and volcano plot visualization.

4.1.6.2 DGE analysis comparison tab

The “DGE analysis” tab contains a fifth sub-tab where the unique feature of the application is found. The “DGEA comparison” tab (**Figure 9**) enables the user to compare the results of DGE analysis methods in a one-to-one or all-to-all manner. The one-to-one comparison is more specific and allows for setting different thresholds (absolute log2 fold change and p-value) for each method. The result of the comparison is shown in a Venn diagram and the user is able to explore gene sets that are found in the comparison. Specifically, when comparing the results

of DESeq2 and edgeR methods with absolute log2FC threshold of 1 and p-value threshold of 0.05, 168 genes are discovered as upregulated DEGs by both methods, 10 genes are discovered only by DESeq2, and 235 genes only by edgeR. For downregulated genes, 290 are discovered by both methods, and an additional 357 genes only by edgeR. The user can explore which genes are in each of these sets in the table below the Venn diagram where a summary of log2FC and p-values are displayed.

The all-to-all comparison is a more general method of comparing DGEA methods. The user can only set a single log2FC threshold and p-value threshold. The number of genes in each subset is indicated by the barplot on the UpSet plot, whereas the intersections between methods are indicated by the black dots below each bar. As with the one-to-one comparison, the user can explore each gene set and display the DGEA statistics (log2FC and p-values).

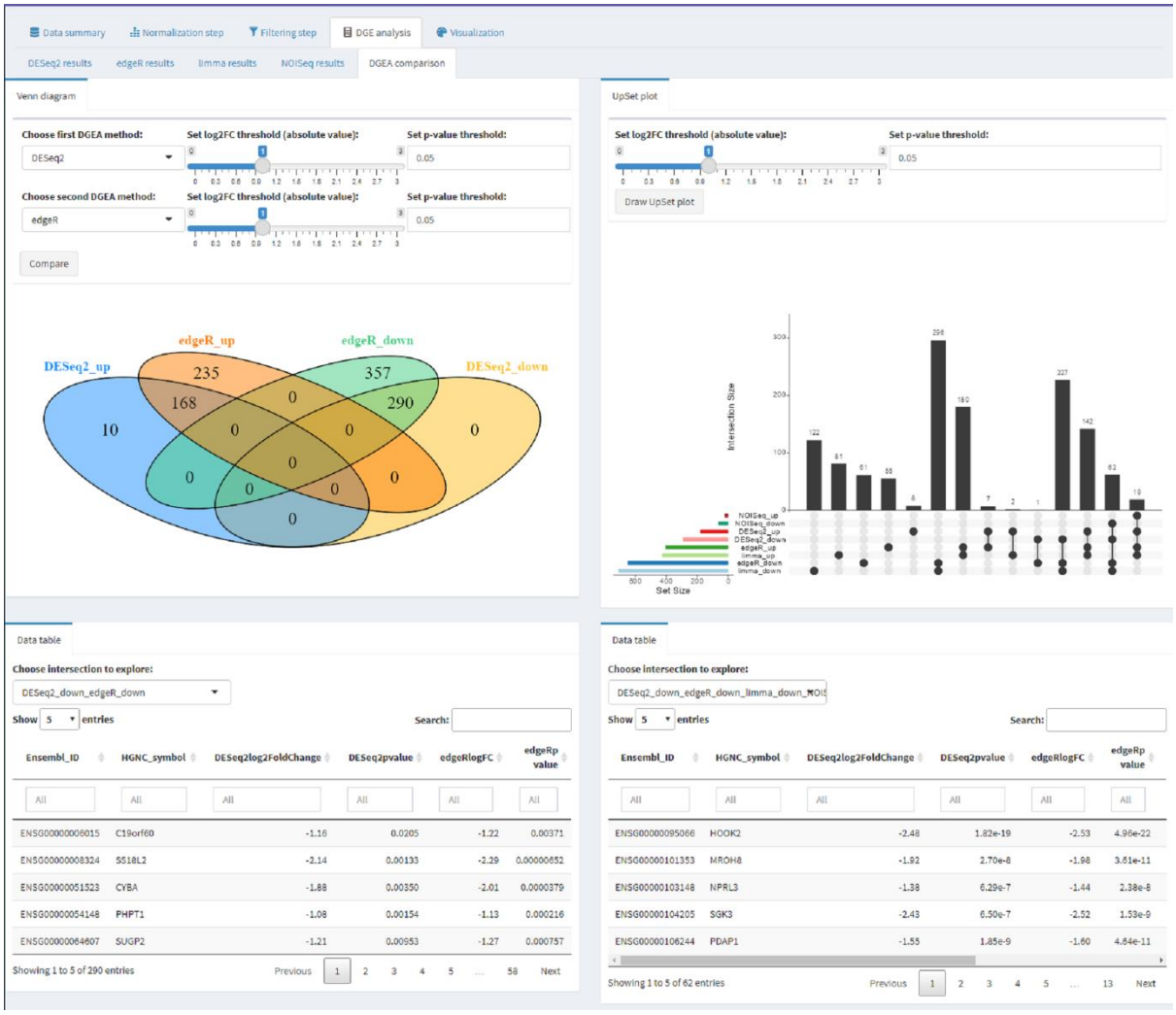
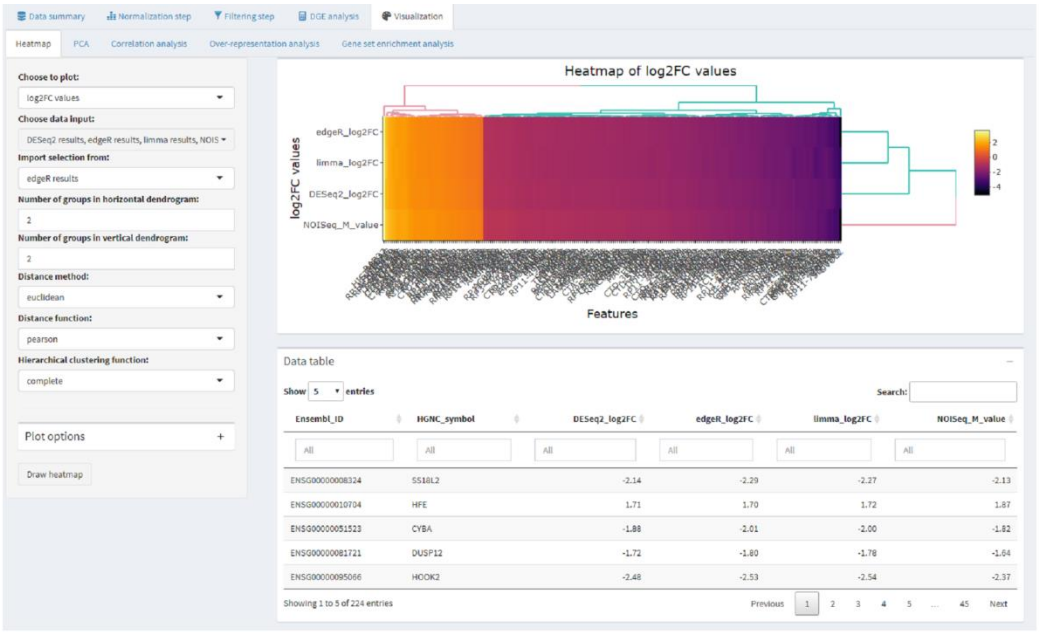


Figure 11. The "DGEA comparison" sub-tab. The tab contains two methods for comparing results of different DGEA methods. Left side: a Venn diagram comparing DESeq2 and edgeR results with log2FC thresholds of 1 (absolute value) and p-value threshold of 0.05. Right side: an UpSet plot where the results of all-to-all comparison of DGEA methods is shown after setting a log2FC threshold of 1 (absolute value) and p-value threshold of 0.05.

4.1.7 Heatmap

The “Heatmap” sub-tab is one of the visualization sub-tabs that is available in the application and can be used to plot log2FC values, p-values from the “DGEA analysis” results tabs, or expression values (raw or normalized). Two views of the “Heatmap” sub-tab are shown in **Figure 10**. The heatmap uses input data depending on whether log2FC/p-values or expression values are chosen and any of the filtered tabs (such as the edgeR results table) can be used as input. In **Figure 10**, the edgeR results were filtered using adjusted p-values <= 0.01 (216 genes, **Fig. 10A**) or adjusted p-values <= 0.05 (411 genes, **Fig. 10B**), and this selection of genes is used to obtain the values that are plotted in the heatmap. Options related to hierarchical clustering (number of groups of clusters, distance method and function, clustering function) as well as plot options (color palette, labels and title) are available in the sidebar panel. The heatmap is interactive, meaning that it is zoomable, and a text box is displayed when hovering over the heatmap that displays gene name, value title and value. The values used to draw the heatmap are also displayed in a table below the heatmap.

A



B

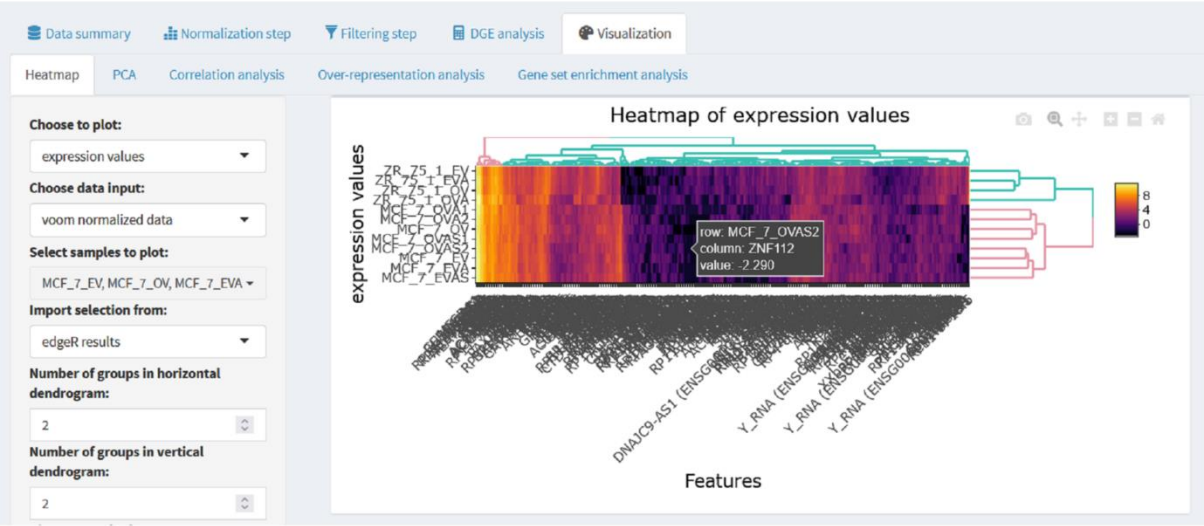
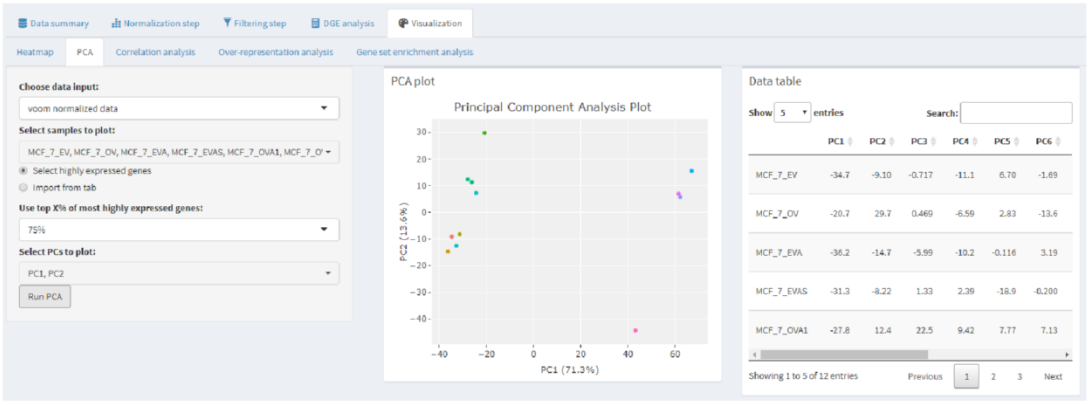


Figure 12. The "heatmap" sub-tab. **A.** A heatmap of log2FC values across all four methods is drawn using hierarchical clustering for features and methods. **B.** A heatmap of voom normalized expression values is drawn using data from all samples in the dataset. Hierarchical clustering is applied to features and expression values.

4. 1. 8 Principal component analysis

The “principal component analysis” sub-tab contains the functionality of running a PCA step using a selection of genes based on previous tables (e.g., DGEA results tables) or a percentage of most highly expressed genes and visualizing the principal components for all samples in the dataset. PCA based on the expression value of filtered genes (from a previous DGEA results table) uses information from differentially expressed genes to reduce dimensionality of the data. The option to use a percentage of the most highly expressed genes is executed by first calculating row means of the dataset (i.e., the average expression value of the features across the dataset), ranking them from highest to lowest average, and subsetting the dataset based on a set percentage. **Figure 11** shows two variants of the PCA sub-tab where first a scatterplot of PC1 and PC2 for all samples is plotted, and then a grid of scatterplots with all combinations of PC1, PC2, PC3, and PC4 are plotted. On the side of the plot, a table with eigenvalues for principal components is displayed. The scatterplots are interactive, meaning that information about sample name shows up when hovering over the plot.

A



B

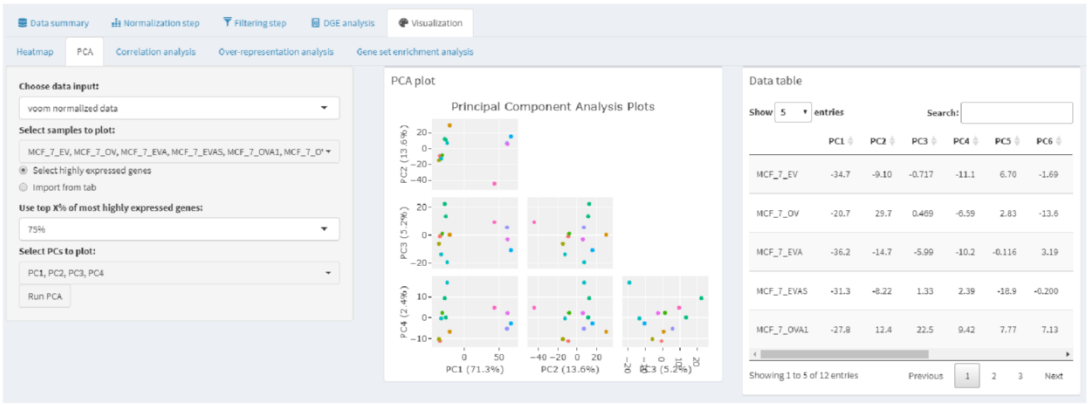


Figure 13. The "PCA" sub-tab that contains 2-D PCA plots and a table with principal component eigenvalues. **A.** A principal component analysis scatterplot of PC1 and PC2 for all samples in the dataset using the top 75% most highly expressed genes. **B.** A grid of PCA plots with all combinations of PC1, PC2, PC3, PC4 plotted for all samples in the dataset using the top 75% most highly expressed genes.

4.1.9 Correlation analysis

The “Correlation analysis” sub-tab has a visualization and exploratory analysis functionality related to the expression values of queried genes. The sub-tab has a sidebar panel where the user can choose which data to plot (raw or normalized expression data), whether to input genes of interest manually or import a selection from a previous table (e.g., a DGEA results table)

and a correlation coefficient cutoff value for the gene network plot. **Figure 12** is a demonstration of the correlation analysis using a selection of genes from the edgeR results table (adjusted p-value <= 0.0001, 74 genes) using voom normalized expression data and a coefficient of correlation cutoff value of 0.5 for the gene network plot. In the dot plot, the color of the dot represents the coefficient of correlation, and the dot size is scaled based on -log10-transformed p-values. The gene network has nodes named after the genes, the node size is based on the average expression value across the dataset, the edge thickness is based on the absolute value of the coefficient of correlation and the color reflects the coefficient. The dot plot is interactive and can show information on p-value and coefficient of correlation when hovering. The gene network is more restrictive, although the user can zoom in like in the dot plot and move the nodes. Two tables that display the coefficient of correlation and p-values matrices are included alongside the plots.

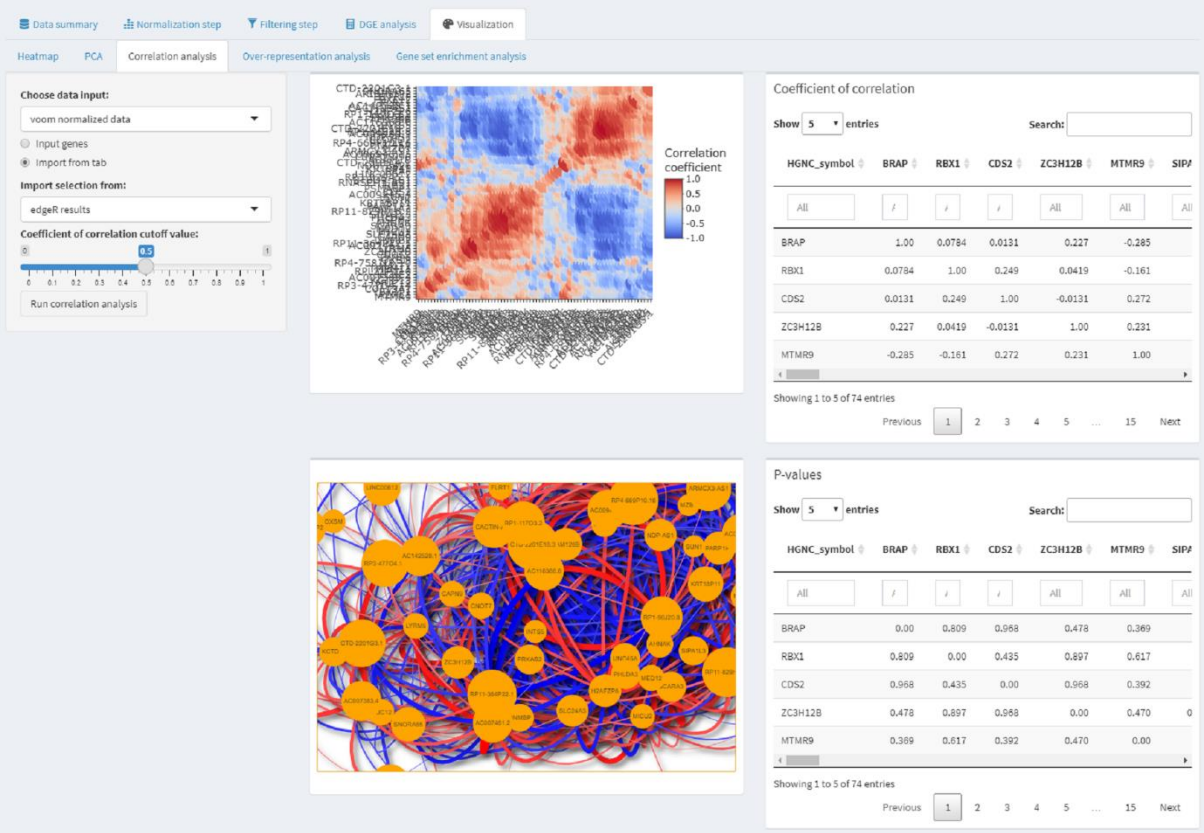


Figure 14. The "Correlation analysis" sub-tab. The tab has a sidebar panel where the input of data and gene selection is made, and the correlation dot plot and gene network plot alongside tables displaying the coefficient of correlation and p-values of the correlation analysis. The gene selection is done in the edgeR results table using adjusted p-values <= 0.0001 (74 genes).

4.1.10 Over-representation analysis

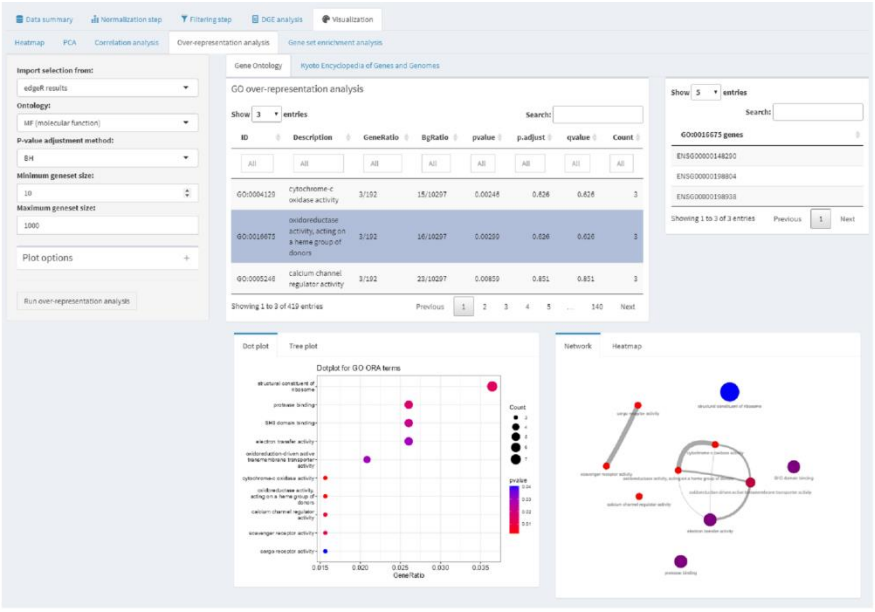
The “Over-representation analysis” sub-tab (**Figure 13**) is one of the two downstream functional analysis features that the application offers to the user. The ORA sub-tab is organized in further sub-tabs, one for each database that can be queried for information on over-represented terms. The user can select genes from a previous table (e.g., a DGEA results table), an ontology (MF: molecular function, BP: biological process, CC: cellular component), a p-value adjustment method, and minimum and maximum gene set sizes. Results are displayed

for the Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG). The results consist of a general table with statistics from the functional analysis such as the over-represented term, the gene ratio, the background ratio, the p-value and adjusted p-value, and count of genes present in the term. To explore the genes present in each term, another detail table is available which upon selecting a row from the first table, will display the features of the term. The ORA results are also visualized in the form of a dot plot, a tree plot, a terms network, and a heatmap, all showing only a select number of top over-represented terms (10 by default). There are also additional plot options regarding number of terms visualized and the color dependency (on p-value or adjusted p-value).

4.1.11 Gene set enrichment analysis

The “Gene set enrichment analysis” sub-tab (**Figure 14**) is the second downstream functional analysis feature available in the application. Similarly to the ORA tab, the GSEA tab is organized into two other sub-tabs for GO and KEGG enrichment analysis results. The user can select genes from a previous table, an ontology, a p-value adjustment method, and minimum and maximum gene set sizes. The results of the GSEA are displayed in two tables, a general table with information on GO/KEGG enriched terms and statistics about the enrichment analysis. Genes present in each term can be explored by selecting a row of the general table. A dot plot, a tree plot, a terms network, and a running score plot are used to visualize the top enriched terms (10 by default). There are additional plot options to change the number of terms visualized and the color dependency (on p-value or adjusted p-value).

A



B

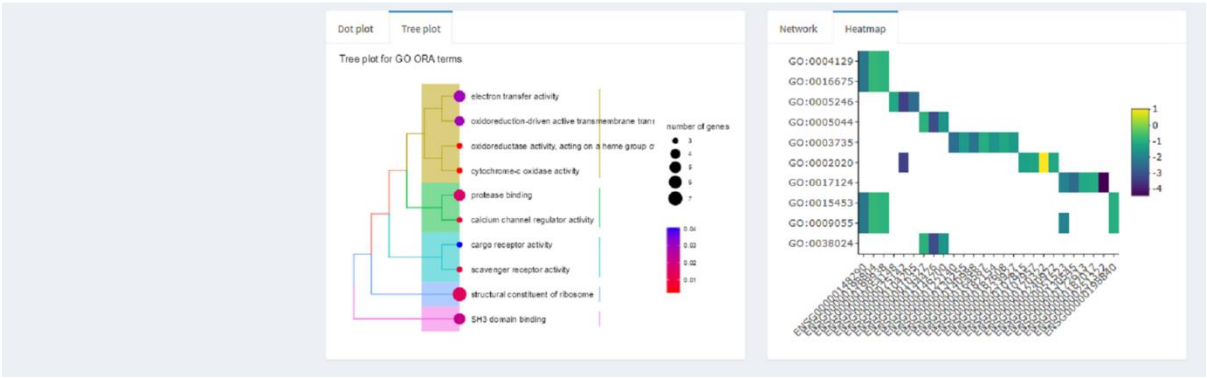
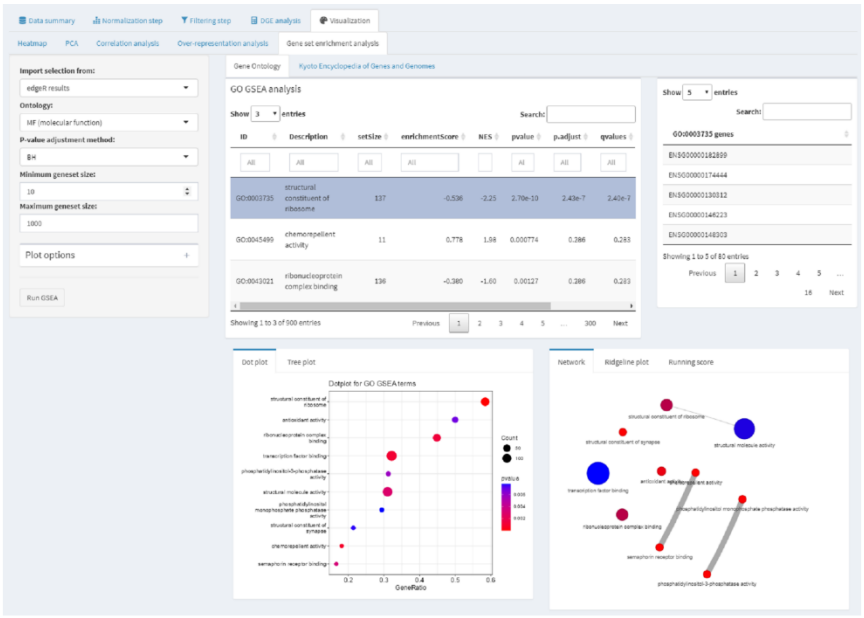


Figure 15. The "Over-representation analysis" sub-tab. **A.** Results of the ORA for a selection of genes from the edgeR results (adjusted p-value <= 0.05, 411 genes) using Gene Ontology database for querying data for the molecular function ontology. A table with results, a details table, dot plot and GO terms network is displayed (top 10 GO terms visualized). **B.** Additional plots for the ORA results in **A** are shown, including a tree plot and heatmap of features present in the top 10 GO terms.

A



B

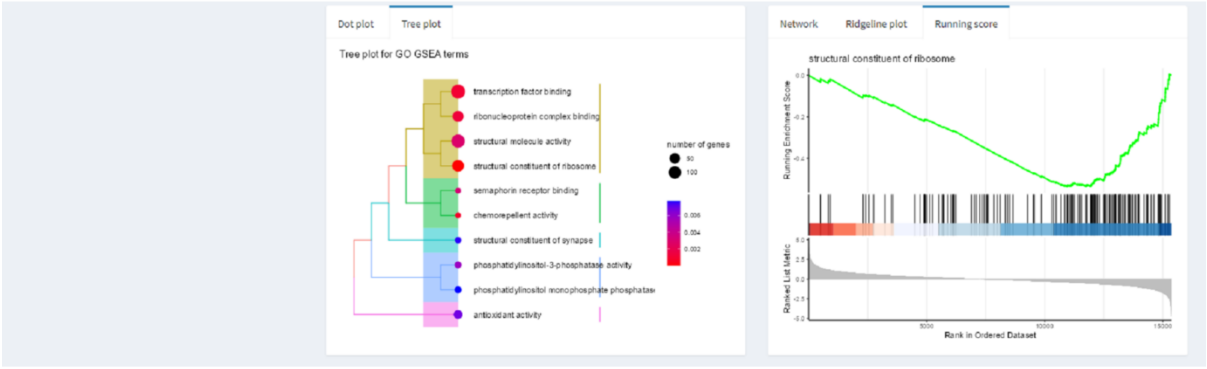


Figure 16. The "Gene set enrichment analysis" sub-tab. A. Results of the GSEA for a selection of genes from the edgeR results (adjusted p-value <= 0.05, 411 genes) using Gene Ontology database for querying data for the molecular function ontology. A table with results, a details table, dot plot and GO terms network is displayed (top 10 GO terms visualized). **B.** Additional plots for the GSEA results in **A** are shown, including a tree plot and a running score of features present in the selected GO term (structural component of ribosome).

4.2 Case study: MR overexpression study in breast cancer cell lines

4.2.1 RNA-seq file metadata

The RNA-seq dataset analyzed in this section was obtained by Dr. Seniye Targen. Dr. Targen performed an overexpression experiment using the MCF-7 and ZR-75-1 cell lines, starting with their transfection and overexpression of the mineralocorticoid receptor (OV) or empty vector (EV), hormone treatment (aldosterone, or aldosterone and spironolactone), RNA isolation, and RNA sequencing, which was purchased as a service from EMBL that provided FASTQ files (Targen, 2020). The experimental conditions, RNA integrity numbers (RIN), and group names, obtained from Seniye Targen, are given for each sample in **Table 6**. All samples except MCF_7_OVA1 had a RIN higher than 7, indicating a level of high RNA integrity (Puchta et al., 2020; Schroeder et al., 2006).

Table 6. RNA-seq metadata. RIN data provided by Dr. Seniye Targen.

Sample name	Group name	Condition	RIN
MCF_7_EV	MCF_7_EV	MCF-7 control	7.70
MCF_7_EVA	MCF_7_EVA	aldosterone	9.70
MCF_7_EVAS	MCF_7_EVAS	aldosterone + spironolactone	8.10
MCF_7_OV	MCF_7_OV	MR overexpression	7.70
MCF_7_OVA1	MCF_7_OVA	MR overexpression	6.20
MCF_7_OVA2		+ aldosterone	8.10
MCF_7_OVAS1	MCF_7_OVAS	MR overexpression	8.70
MCF_7_OVAS2		+ aldosterone + spironolactone	9.10
ZR_75_1_EV	ZR_75_1_EV	ZR-75-1 control	10.0
ZR_75_1_EVA	ZR_75_1_EVA	aldosterone	9.70
ZR_75_1_OV	ZR_75_1_OV	MR overexpression	10.0
ZR_75_1_OVA	ZR_75_1_OVA	MR overexpression + aldosterone	10.0

4.2.2 FASTQC reports

FASTQ files were analyzed using the FASTQC tool and found to have passing scores for mean quality of across all positions and reads (**Figure 17A**). The distribution of sequences quality scores showed that the majority of sequences in the FASTQ files were of high quality (Phred score >= 28, **Figure 17B**). There was a low percentage of adapter content in reads, and a low level of overrepresented sequences in the sequencing data (**Figure 17C-D**). Samples had between 18-30 unique sequences, and the distribution of sequence GC content percentage followed the normal distribution with a mean value near 50% (**Figure 17E-F**). A table of statistics that summarized the FASTQC step showed that the files failed 25% of FASTQC modules. These were the per base sequence content, sequence duplication, and k-mer content. The files were aligned using the STAR aligner (Dobin et al., 2013) and quantified with HTSeq-count (Anders et al., 2015), which mitigated some of these failures such as the sequence duplication and k-mer content.

Table 7. General statistics from FASTQC analysis, summarized with MultiQC tool (Ewels et al., 2016). Out of 12 modules the FASTQ files failed the per base sequence content, sequence duplication, and k-mer content modules.

Sample name	Duplicate Reads	GC content	Read Length	Modules Failed	Reads (millions)
MCF_7_EV	69.6%	51%	85 bp	25%	95.3
MCF_7_EVA	68.0%	50%	85 bp	25%	83.8
MCF_7_EVAS	68.1%	49%	85 bp	25%	74.0
MCF_7_OV	70.8%	51%	85 bp	25%	106.1
MCF_7_OVA1	70.7%	52%	85 bp	25%	61.0
MCF_7_OVA2	69.9%	51%	85 bp	33%	82.3
MCF_7_OVAS1	67.4%	49%	85 bp	25%	71.8
MCF_7_OVAS2	67.0%	50%	85 bp	25%	78.2
ZR_75_1_EV	63.8%	50%	85 bp	25%	77.8

ZR_75_1_EVA	67.7%	50%	85 bp	25%	100.6
ZR_75_1_OV	65.1%	51%	85 bp	25%	81.9
ZR_75_1_OVA	64.7%	49%	85 bp	25%	76.1

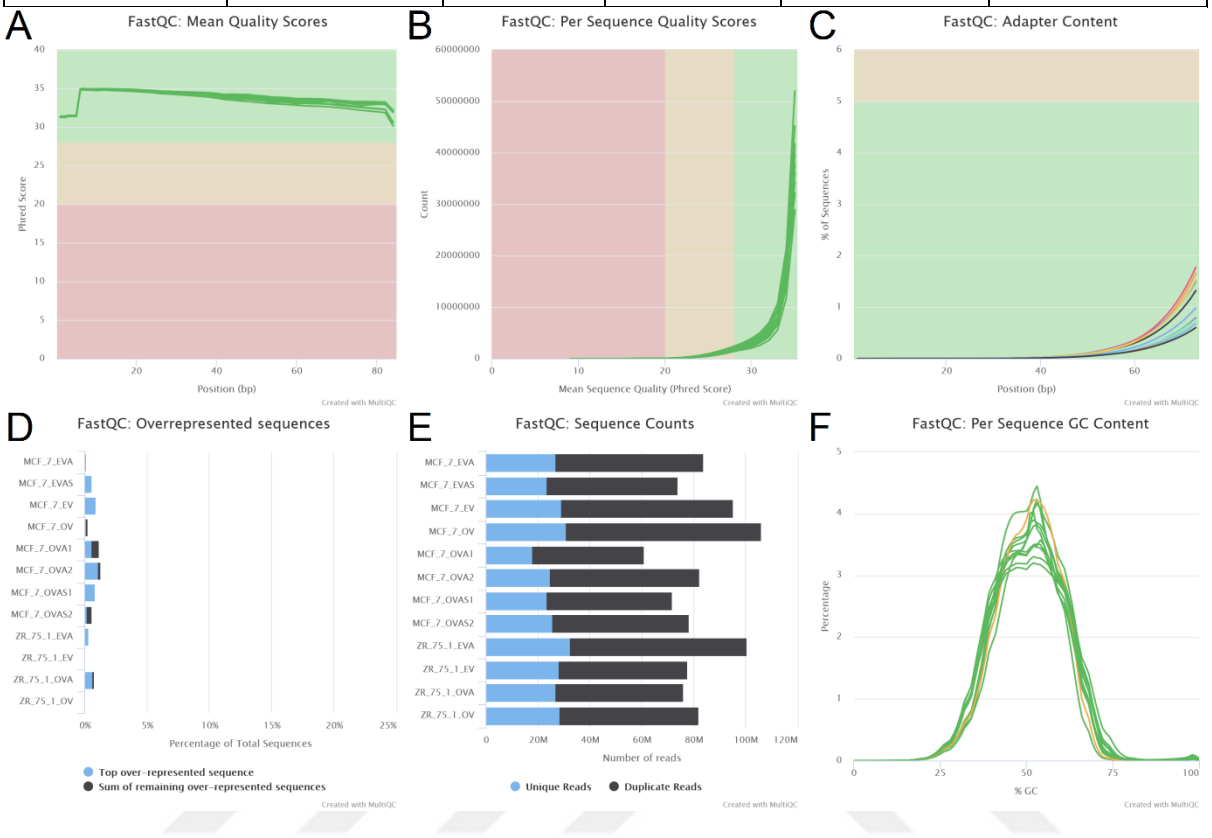


Figure 17. Quality control plots of the RNA-seq FASTQ files generated using FASTQC and MultiQC (green: passing score, yellow: warning, red: failing score). **A.** Mean quality scores plot per base position showing that all positions in the reads analyzed have high quality scores (Phred score > 30). **B.** Distribution of reads with regard to their mean quality score. The majority of reads have a high-quality score (>= 28). **C.** Adapter content per sample. The percentage of sequences with adapters increases with length of read, remaining under 2%. D. Over-represented sequences as a percentage of total sequences (blue: top over-represented sequence, black: sum of remaining over-represented sequences). Only MCF_7_OVA2 has a failing score because the top over-represented sequence has a frequency higher than 1%. **E.** Unique and duplicate read counts for all samples in dataset. There is between 63-70% duplication of sequences (**Table 7**). **F.** Distribution of GC content across all reads in all samples (yellow: MCF_7_OVAS2, green: other samples).

4.2.3 Exploratory data analysis results

The RNA-seq raw read counts were visualized in the "Data summary" tab of the application using several plots that were summarized in **Figure 18** and **Figure 19**. The RNA-seq samples had library sizes ranging from 2.3-4.2 million reads. Based on this difference in library sizes across samples, the data needed to be normalized to eliminate the effect the library size has on differential analysis results. After a preliminary normalization step using the DESeq method, the samples were plotted in a multidimensional scaling (MDS) plot and their normalized counts distribution were shown in **Figure 18B-C**. There seemed to be a clear distinction between MCF-7 and ZR-75-1 samples in the first dimension of the MDS plot, and some distinction could be made between conditions of the same cell line in the second dimension. The box plot showed that normalized counts (log scale because DESeq method is not log-transformed) were distributed evenly across the samples with their average count close to the dataset average count. The expression distribution (density) plot showed that the distribution of log-scale expression values was similar across samples with slight difference between samples of different cell lines (**Figure 18D**).

The second part of the exploratory analysis has been achieved by using functions of the {NOISeq} package. An annotation data object was generated using the {biomaRt} package to query information on the biotype of the genes in the dataset. This annotation data were then used to plot the biodetection plots in **Figure 19A-E**. In these plots, it has been shown that transcripts that are translated into proteins, processed pseudogenes, and long non-coding RNAs (lncRNAs) were the most abundant types of features in the dataset (on the left side of the dotted green line). Overall, the biotype percentages of the samples (solid red bar) were on a similar level with the genome (gray bar). The biodetected features that had at least 1 CPM were shown in stripped red bars. Across the four samples in our comparison, 25-30% of protein coding

features, 5-10% of lncRNAs, and 1-3% of processed pseudogenes were undetected (< 1 CPM). When comparing the RNA composition across MCF-7 samples with the MCF_7_EV control sample as reference, the distribution of M values, which were the ratios of read counts of the sample to read counts of the reference, indicated that a normalization step was necessary to eliminate the bias in sequence read distribution. Biotype-wise counts distribution plots for five MCF-7 samples (one MCF_7_OV, two MCF_7_OVA and two MCF_7_OVAS) were shown in **Figure 20**. The most common biotypes (protein coding, lncRNAs, processed pseudogenes, transcribed pseudogenes) had the highest average read counts as well as the widest range of counts.

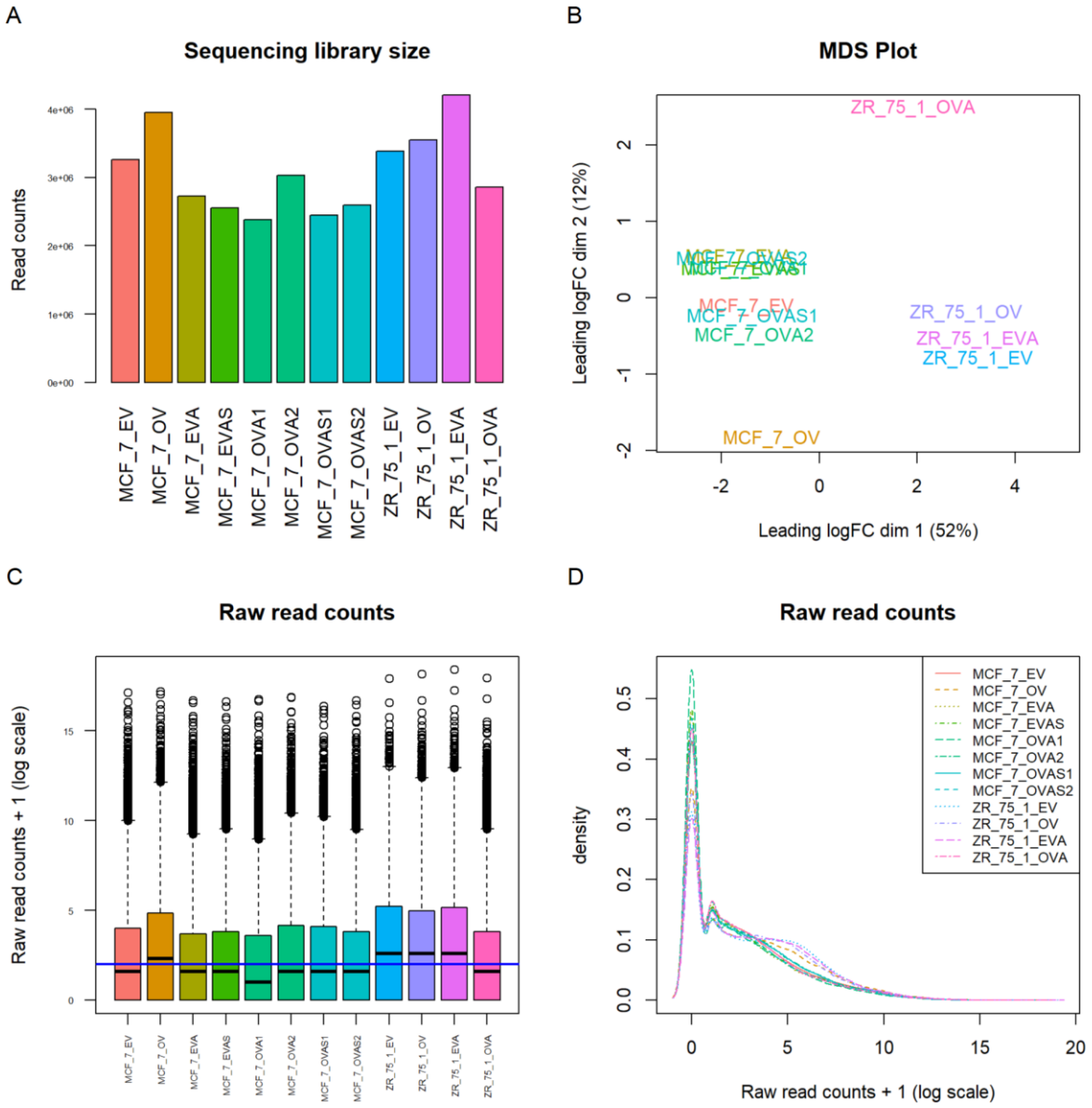


Figure 18. RNA-seq raw read count data summary plots. **A.** Sequencing library size barplot showing total read counts per sample. **B.** Multidimensional scaling plot showing samples arranged in 2-D space using normalized read count data. **C.** Boxplot of DESeq-normalized counts (log scale) for all samples. The blue line indicates the average expression value across all samples. **D.** Density plot showing the distribution of expression values (log scale) for all samples. The dotted red line indicates 1.5 raw read counts.

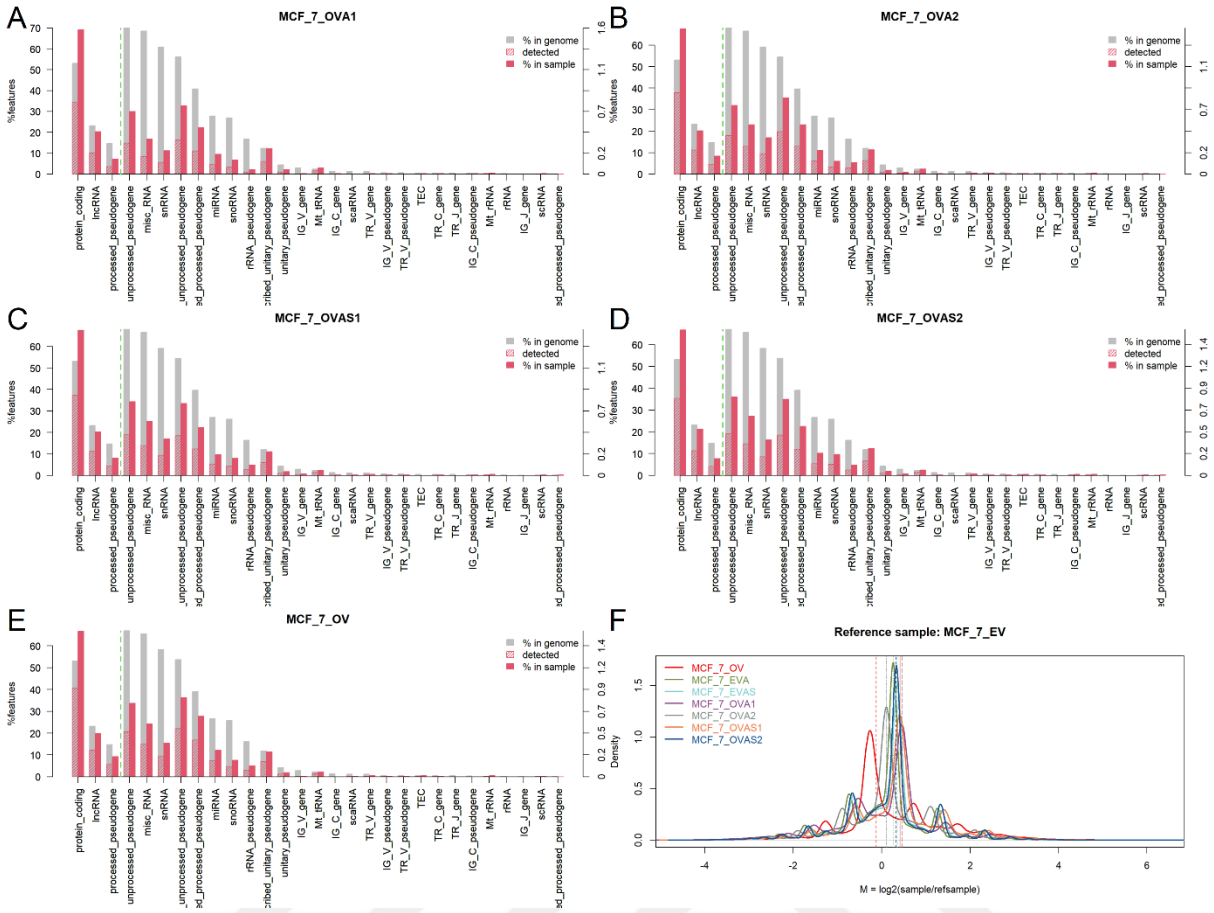


Figure 19. Data quality plots generated using the {NOISeq} R/Bioconductor package. **A-E.** Biodetection plots of five RNA-seq samples showing the percentage of features detected (*k* >= 1) in the sample compared to the frequency of the biotype in the genome. The dotted green line divides the most abundant biotypes (left-hand side) from the rest. **F.** RNA composition plot. The distribution of M values shows that there is a bias in the distribution of sequencing reads across features, therefore a normalization method is applied in further steps.

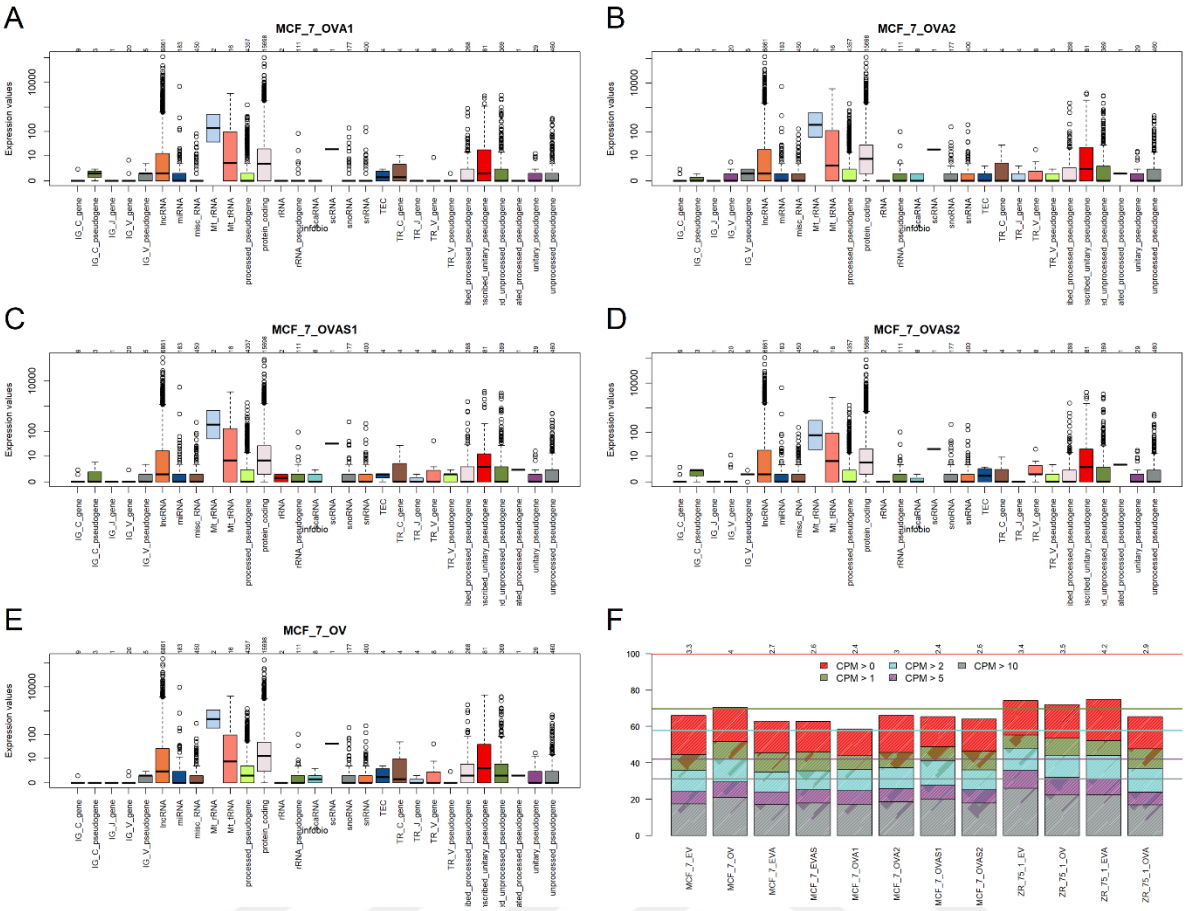


Figure 20. A-E. Counts distribution plots of the MCF-7 OV, OVA and OVAS samples (raw read counts). The boxplots show the counts distribution for each biotype available in the dataset. The number of features of each biotype is shown above each boxplot. **F.** Saturation plot for all samples. Horizontal colored lines indicate the percentage of features with the given CPM value in at least one of the samples, whereas the bars indicate the percentage of features within each sample having the given CPM value (0, 1, 2, 5, or 10 CPM).

4.2.4 Normalization and filtering results

The differences in sequence library sizes and RNA composition across samples indicated that the read counts needed to be normalized. The voom normalization method was chosen as it normalizes counts by calculating weights and minimizes the standard deviation of counts to the mean of the dataset. The sample count averages approach the geometric mean count of the

dataset, and the counts were log-transformed similarly to the log2CPM method. The distribution of the read counts for log2CPM and voom normalized counts were shown in **Figure 21** as boxplots and density plots.

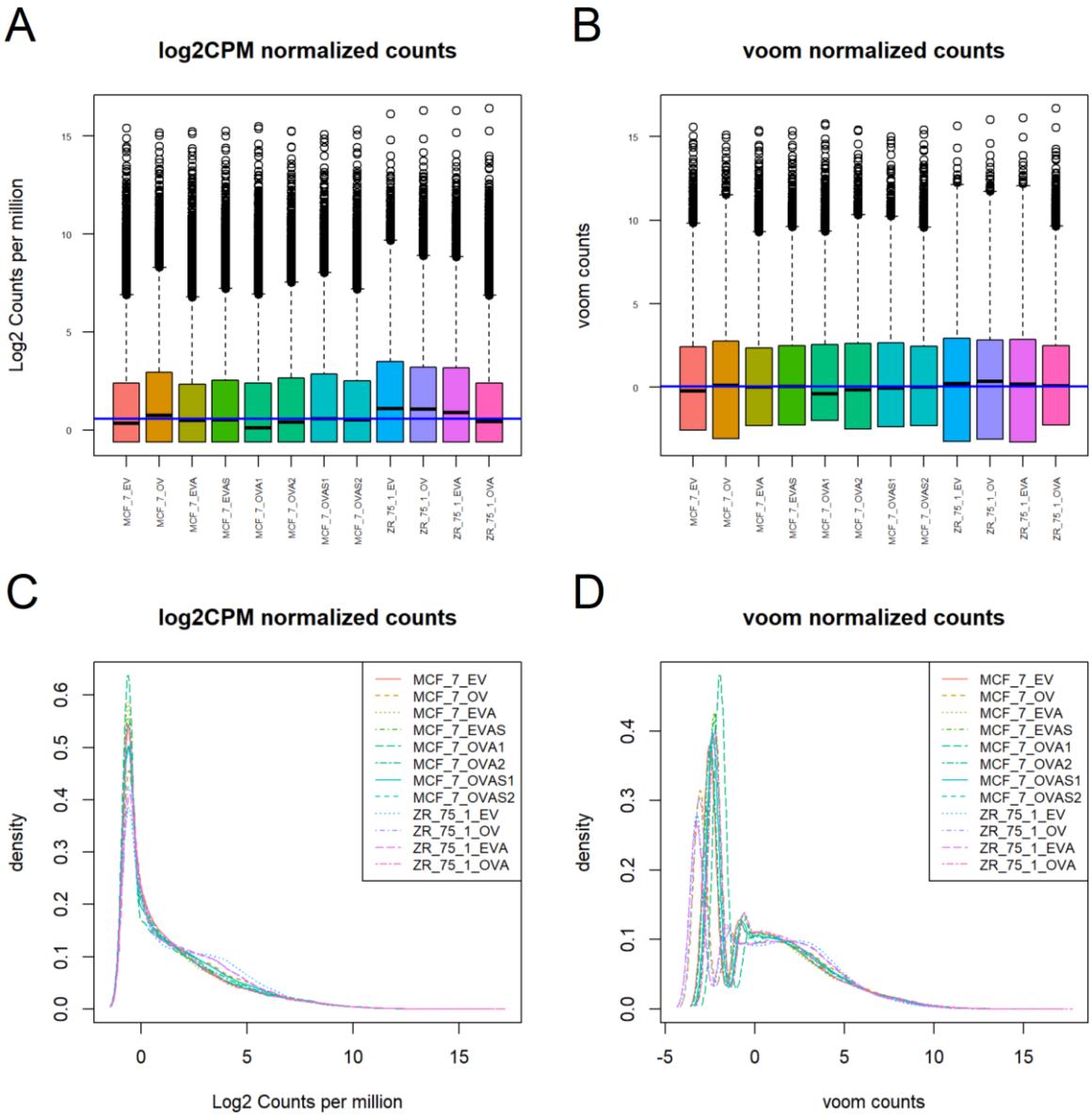


Figure 21. Distribution boxplots and density plots for normalized count data. **A, C.** log2CPM normalized counts distribution as a boxplot and density plot. **B, D.** voom normalized counts distribution as a boxplot and density plot.

The filtering step eliminated 7,622 genes as lowly expressed genes (minimum count of 1 and minimum total count of 15). A summary of the filtering step was given in **Table 8**. The remaining dataset consisted of 21,992 genes and the removal of lowly expressed genes did not affect the library sizes or counts distribution significantly (an average reduction of 1543.25 reads in library sizes ranging from 2.3-4.2 million reads).

Table 8. Summary statistics of the filtering step.

Summary statistic description	Value
Remaining genes	21,992 genes
Removed genes	7,622 genes
Removed counts mean	0.202 reads
Sample library size reduction average	1543.25 reads

The normalized, filtered counts were ranked by average mean expression across samples and the top 75% of all genes were used for principal component analysis. **Figure 22** showed the similarity between samples. In most PCA plots, the ZR-75-1 samples were clustered away from the MCF-7 samples due to inherent cell line differences in transcriptomes. The MCF-7 samples had more pronounced differences when plotting PC2-PC3 and PC2-PC3 (plots in the middle column in **Figure 22**).

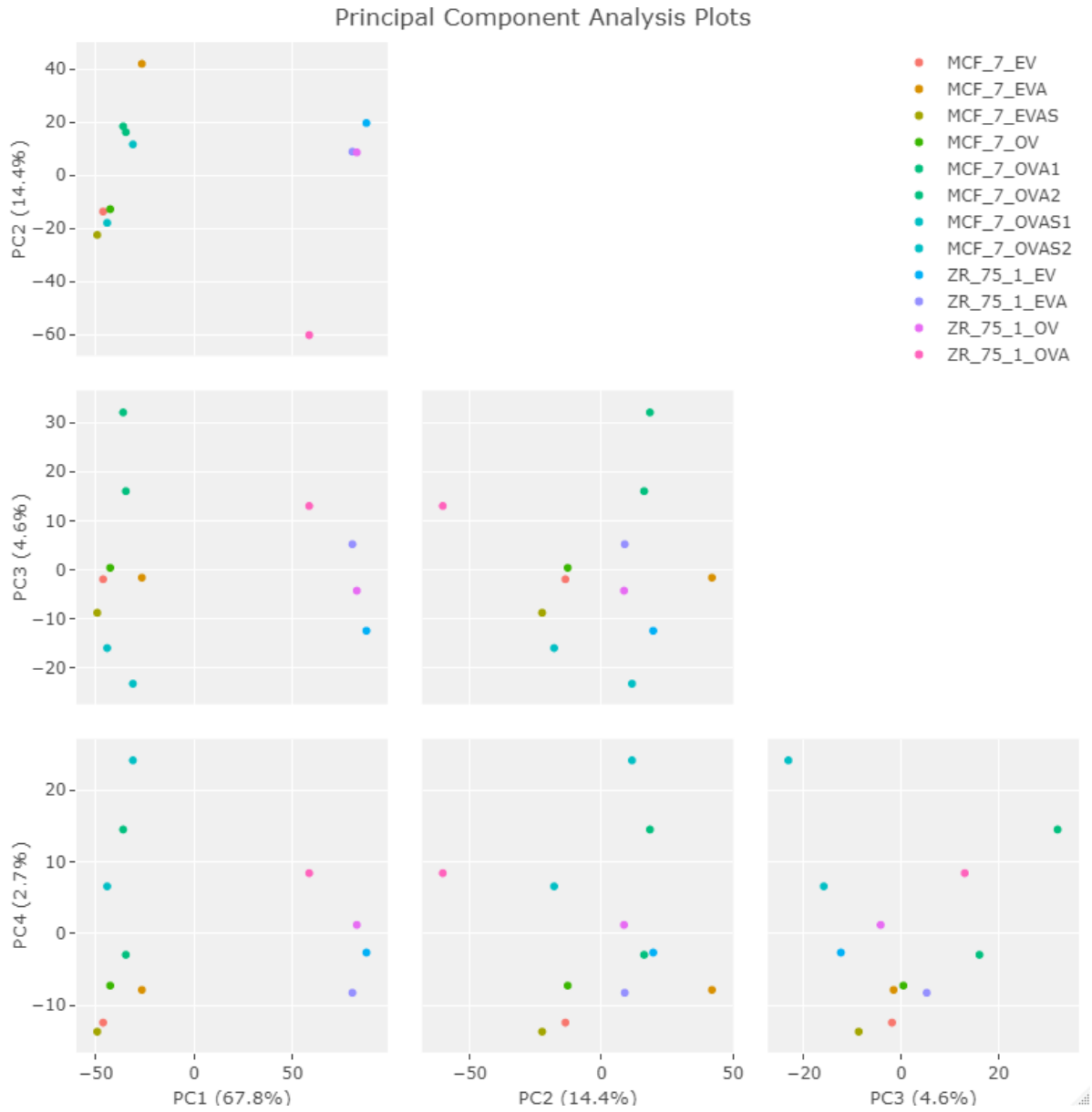


Figure 22. PCA plots for all combinations of first four PCs. The top 75% of most highly expressed genes were used to perform the principal component analysis.

4.2.5 Differential gene expression analysis (DGEA)

DESeq2, edgeR, limma, and NOISeq were used as methods for making two comparisons:

1. MCF-7 MR overexpression aldosterone treatment (OVA) *vs.* MCF-7 MR overexpression control (OV)
2. MCF-7 MR overexpression aldosterone + spironolactone treatment (OVAS) *vs.* MCF-7 MR overexpression aldosterone (OVA) treatment

MCF-7 MR overexpression aldosterone treatment vs. MCF-7 MR overexpression

The effect of the treatment of aldosterone in MR-overexpressing MCF-7 cells was examined using the comparative analyses between the MCF_7_OVA and MCF_7_OV samples. The results from the differential analysis were collected into two tables that list top upregulated and downregulated genes as a result of aldosterone treatment (**Table *9*** and **Table *10***).

The DGE analysis methods used to implement different strategies for finding DEGs, which led to different test results for the genes in the dataset. To confirm the discovery of DEGs, a comparison of log2 fold change values and adjusted p-values from all methods was done and the results were summarized in the UpSet plot shown in **Figure 24** and summarized in **Table 11**.

Generally, the limma and edgeR methods discovered more DEGs as evidenced by the UpSet plots for all three comparisons made (**Figure 24**, **Figure 29**) even though the log2 fold change and p-value thresholds were the same for all methods. Thus, it was important to incorporate the findings from the other two methods to validate the DEGs, and downstream analyses and results visualization was performed with genes that were discovered from at least three methods. The top 10 upregulated and top 10 downregulated genes named named in the volcano

plots in **Figure 23**. The top common genes discovered by all methods were listed in **Table 9** and **Table 10** (tables truncated when the list of genes was long, full lists were included in **Appendix 1.**) There were more downregulated genes than upregulated, but the log2FC values matched closely between different methods (**Figure 25**).

195 genes were discovered in agreement between the four methods, and an additional 189 genes were discovered from any combination of three methods. In total, a gene list consisting of 384 genes was used to perform over-representation analysis. The top over-represented molecular function terms involved functions related to DNA binding and transcription factor activity (**Table 12**); the biological processes that wered over-represented included cell death positive regulation and cell migration (**Table 13**); and the over-represented cellular component terms described the ribosomal subunits, focal adhesion, and components related to the cytoskeleton (**Table 14**). Overall, the treatment of MCF-7 cells that overexpressed the mineralocorticoid receptor caused the differential expression of genes that were involved in transcription regulation, cell death, and cell migration. Gene set enrichment analysis was performed using log2 fold change data for all genes and its results supported the findings of the over-representation analysis and further confirmed the enrichment of terms related with transcription activity, binding of chromatin, histones, actin, and involvement in phosphatase and kinase activities (**Table 15**, **Table 16**, and **Table 17**). The gene ratios of the ORA and GSEA terms, clustering of ORA terms and the expression distribution of core enrichment genes of GSEA terms were shown in **Figure 26** and **Figure 27**.

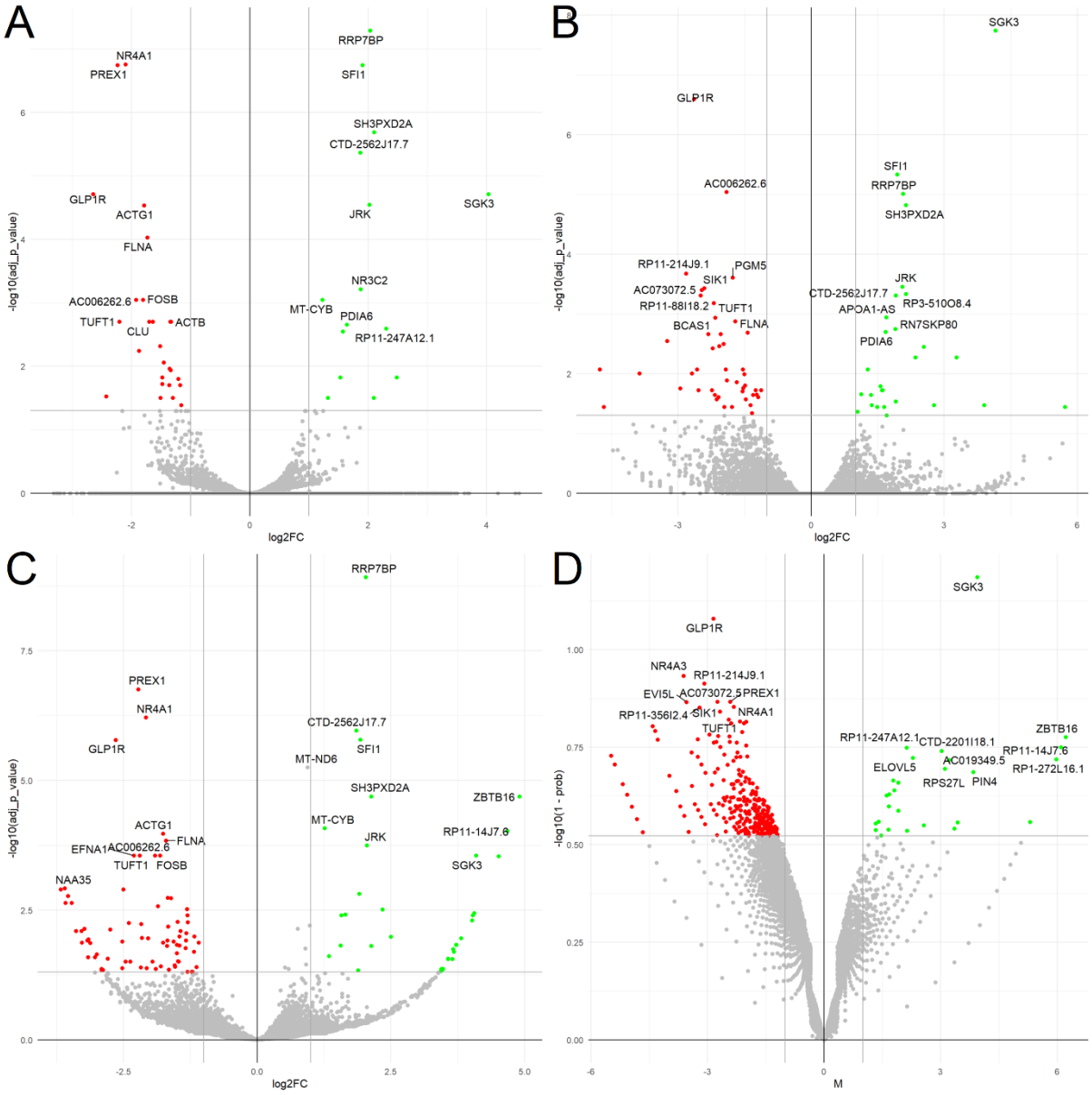


Figure 23. The volcano plots visualizing DGEA results of MCF_7_OVA vs MCF_7_OV. (**A.** DESeq2, **B.** edgeR, **C.** limma, **D.** NOISeq). The top 10 upregulated (green color) and downregulated genes (red color) are labeled (ranked by adjusted p-values). p-value threshold selected as 0.05 or prob value threshold as 0.70, and log2 fold change absolute value threshold selected as 1.

Table 9. Full list of upregulated genes discovered in common using all four methods in the comparison of MCF_7_OVA vs. MCF_7_OV groups. Average voom normalized counts for the analyzed groups and DGEA statistics are given.

HGNC symbol	average voom counts OV	average voom counts OVA	DESeq log2FC	DESeq p value	edgeR log2FC	edgeR p value	limma log2FC	limma p value	NOISeq M value	NOISeq prob value
RRP7BP	8.264	10.337	2.03	1.03E-11	2.07	2.23E-09	2.03	5.50E-14	1.78	0.784
SFI1	6.834	8.798	1.91	1.29E-10	1.94	6.32E-10	1.92	3.72E-10	1.68	0.765
SH3PXD2A	5.089	7.250	2.10	2.07E-09	2.13	4.10E-09	2.13	8.30E-09	1.91	0.781
CTD-2562J17.7	7.433	9.334	1.86	5.19E-09	1.90	2.89E-07	1.85	2.00E-10	1.61	0.764
SGK3	0.841	5.017	4.03	3.02E-08	4.15	8.20E-13	4.09	2.30E-07	3.94	0.935
JRK	4.748	6.835	2.02	5.15E-08	2.06	1.43E-07	2.05	1.14E-07	1.81	0.771
NR3C2	4.571	6.512	1.87	1.48E-06	1.91	8.05E-05	1.91	1.67E-06	1.67	0.748
PDIA6	5.224	6.918	1.64	9.34E-06	1.67	1.82E-06	1.65	5.95E-06	1.40	0.724
RP11-247A12.1	3.000	5.379	2.30	1.14E-05	2.35	7.26E-06	2.34	4.51E-06	2.12	0.821
RP11-69H7.3	5.645	7.264	1.57	1.32E-05	1.61	4.17E-05	1.57	6.48E-06	1.33	0.721
GNAO1	4.799	6.393	1.53	8.86E-05	1.56	3.09E-05	1.56	4.92E-05	1.33	0.710
ELOVL5	2.220	4.782	2.48	9.39E-05	2.54	4.35E-06	2.50	2.56E-05	2.28	0.810
RP3-510O8.4	2.817	4.989	2.09	2.47E-04	2.14	2.52E-07	2.13	5.16E-05	1.91	0.741
RN7SKP80	2.957	4.891	1.86	1.24E-03	1.90	1.54E-06	1.89	2.45E-04	1.66	0.711
APOA1-AS	3.344	5.066	1.66	2.15E-03	1.70	8.28E-07	1.69	4.38E-04	1.47	0.701
CTD-2201I18.1	0.104	3.354	3.07	3.42E-03	3.24	8.01E-04	3.13	7.25E-04	3.03	0.818
AC019349.5	-0.258	3.113	3.27	4.86E-03	3.48	1.09E-03	3.17	1.05E-03	3.22	0.808
RPS27L	-0.258	3.040	3.08	7.93E-03	3.28	7.13E-06	3.18	9.84E-04	3.11	0.798
ZBTB16	-3.066	2.471	4.55	1.22E-02	5.73	1.17E-04	4.90	8.34E-09	6.22	0.832
RP11-14J7.6	-3.066	2.293	4.49	1.52E-02	5.67	1.24E-03	4.68	4.76E-08	6.10	0.822
RP1-145M24.1	0.104	2.901	2.64	2.10E-02	2.77	9.99E-05	2.67	4.14E-03	2.57	0.718
PIN4	-1.481	2.397	3.46	2.21E-02	3.90	9.58E-05	3.65	1.11E-04	3.84	0.794
RP1-272L16.1	-3.066	1.922	4.19	2.96E-02	5.39	3.79E-03	4.51	2.64E-07	5.97	0.809

Table 10. List of top 30 downregulated genes discovered in common using all four methods in the comparison of MCF_7_OVA vs. MCF_7_OV groups. Average voom normalized counts for the analyzed groups and DGEA statistics are given. Full list is in **Appendix 1**.

HGNC symbol	average voom counts OV	average voom counts OVA	DESeq log2FC	DESeq p value	edgeR log2FC	edgeR p value	limma log2FC	limma p value	NOISeq M value	NOISeq prob value
NR4A1	8.152	6.109	-2.10	7.04E-11	-2.07	4.06E-06	-2.08	8.38E-11	-2.32	0.860
PREX1	8.095	5.885	-2.24	1.45E-10	-2.22	4.80E-06	-2.22	1.60E-11	-2.42	0.864
GLP1R	6.448	3.785	-2.65	3.13E-08	-2.64	2.30E-11	-2.65	4.51E-10	-2.85	0.917
ACTG1	7.796	6.065	-1.79	5.88E-08	-1.76	3.36E-04	-1.76	5.82E-08	-2.01	0.847
FLNA	7.807	6.122	-1.74	2.07E-07	-1.71	1.08E-06	-1.71	8.51E-08	-1.93	0.785
FOSB	7.174	5.409	-1.81	2.45E-06	-1.78	7.31E-04	-1.82	2.02E-07	-2.06	0.846
AC006262.6	6.694	4.823	-1.92	2.73E-06	-1.91	1.66E-09	-1.91	2.31E-07	-2.16	0.847
EEF2	8.500	7.227	-1.33	7.05E-06	-1.30	2.53E-03	-1.30	6.67E-06	-1.54	0.756
FOS	7.181	5.587	-1.64	7.22E-06	-1.62	4.14E-04	-1.61	2.29E-06	-1.84	0.780
CLU	6.916	5.272	-1.70	7.67E-06	-1.68	2.53E-05	-1.67	2.16E-06	-1.92	0.780
TUFT1	5.992	3.832	-2.21	7.82E-06	-2.20	4.52E-07	-2.20	2.44E-07	-2.46	0.849
ACTB	8.654	7.363	-1.34	8.06E-06	-1.32	3.20E-03	-1.32	4.25E-06	-1.53	0.756
SPTSSB	7.157	5.695	-1.52	2.34E-05	-1.49	2.92E-05	-1.49	9.40E-06	-1.73	0.774
TFF1	6.222	4.399	-1.87	2.90E-05	-1.86	4.92E-03	-1.85	3.61E-06	-2.10	0.837
LMNA	7.133	5.734	-1.46	4.60E-05	-1.43	1.95E-06	-1.43	1.78E-05	-1.68	0.759
JUNB	7.451	6.147	-1.36	5.96E-05	-1.33	1.51E-04	-1.33	3.27E-05	-1.56	0.752
ISG15	7.759	6.474	-1.34	6.55E-05	-1.30	3.35E-03	-1.33	2.06E-05	-1.58	0.757
MKI67	7.015	5.577	-1.48	9.05E-05	-1.46	1.76E-03	-1.45	2.45E-05	-1.67	0.757
MYH9	8.105	6.957	-1.21	1.03E-04	-1.18	6.34E-03	-1.19	7.12E-05	-1.43	0.744
JUN	6.721	5.295	-1.48	1.27E-04	-1.46	1.11E-02	-1.45	4.97E-05	-1.70	0.758
EIF4G1	7.284	5.962	-1.37	1.37E-04	-1.34	1.00E-02	-1.34	5.85E-05	-1.55	0.751
DSCAM-AS1	8.769	7.640	-1.18	1.41E-04	-1.14	5.38E-03	-1.17	2.41E-05	-1.42	0.746
AC073072.5	5.109	2.672	-2.43	2.18E-04	-2.47	1.99E-07	-2.50	1.33E-06	-2.74	0.864
UBC	6.487	5.026	-1.51	2.46E-04	-1.49	3.81E-03	-1.50	4.69E-05	-1.75	0.768
BASP1	8.242	6.950	-1.30	2.56E-04	-1.28	2.94E-04	-1.30	9.82E-06	-1.46	0.748
BHLHE40-AS1	7.760	6.660	-1.16	3.41E-04	-1.13	4.09E-05	-1.14	1.77E-04	-1.39	0.732
ADAR	6.180	4.709	-1.53	4.65E-04	-1.51	1.75E-05	-1.50	7.82E-05	-1.75	0.765
HSPB1	7.070	5.886	-1.24	4.85E-04	-1.22	2.61E-03	-1.21	2.70E-04	-1.45	0.741
EGR1	7.421	6.312	-1.17	4.89E-04	-1.14	3.77E-03	-1.15	3.13E-04	-1.39	0.731
MYC	7.002	5.807	-1.25	5.14E-04	-1.23	2.35E-03	-1.22	2.68E-04	-1.46	0.741

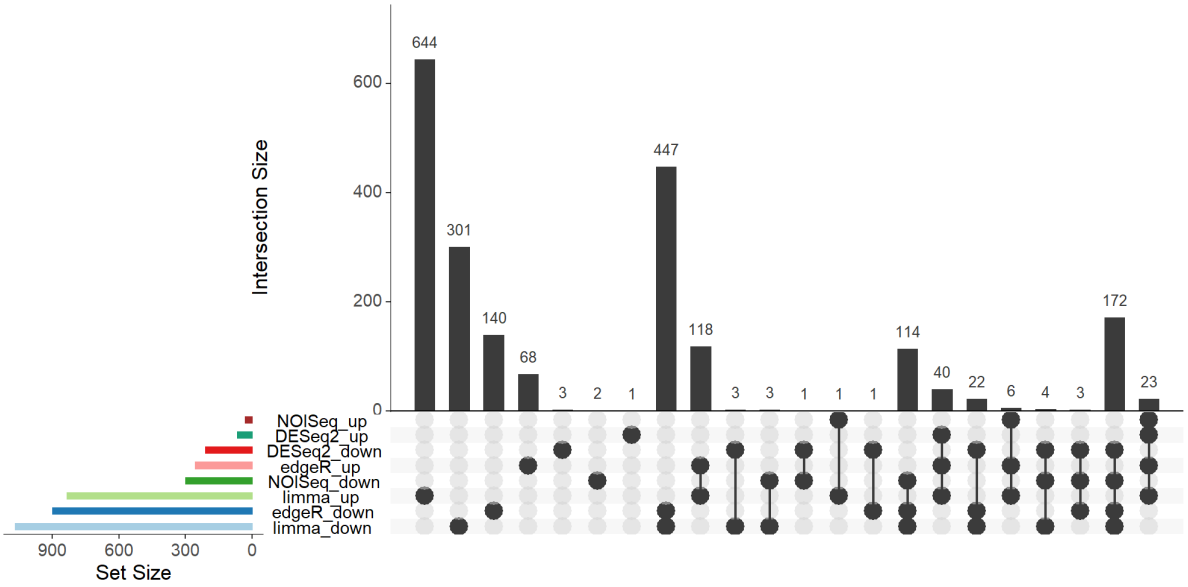


Figure 24. UpSet plot showing the intersections of DEGs discovered by different methods. Intersections are indicated by black dots corresponding to each gene set (2 for each method).

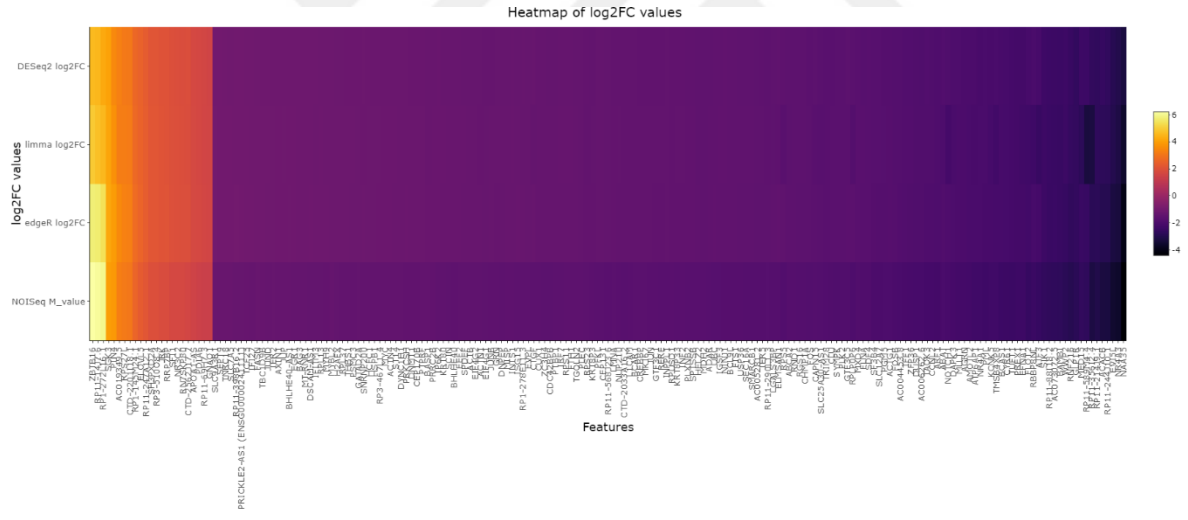


Figure 25. Heatmap of log2 fold change values for the 195 genes discovered by all methods. There is very little difference between the log2FC values calculated by different methods.

Table 11. Summarized results from the comparison of the four DGEA methods.

Intersection	**Upregulated genes**	**Downregulated genes**
DESeq2, edgeR, limma, NOISeq	23	172
DESeq2, edgeR, limma	40	22
DESeq2, edgeR, NOISeq	0	3
DESeq2, limma, NOISeq	0	4
edgeR, limma, NOISeq	6	114
DESeq2, edgeR	0	1
DESeq2, limma	0	3
DESeq2, NOISeq	0	1
edgeR, limma	118	447
edgeR, NOISeq	0	0
limma, NOISeq	1	3
DESeq2	1	3
edgeR	68	140
limma	644	301
NOISeq	0	2
Number of methods	**Upregulated genes**	**Downregulated genes**
all 4 methods	23	172
any 3 methods	46	143
any 2 methods	119	455
any 1 method	713	446
Number of methods	**Upregulated genes**	**Downregulated genes**
all 4 methods	23	172
at least 3 methods	69	315
at least 2 methods	188	770
at least 1 method	901	1,216

Table 12. Top 20 over-represented Gene Ontology molecular function terms.

ID	Description	GeneRatio	BgRatio	p.adjust	Count
GO:0140297	DNA-binding transcription factor binding	24/281	400/12866	0.001809	24
GO:0001228	DNA-binding transcription activator activity, RNA polymerase II-specific	21/281	324/12866	0.001809	21
GO:0001216	DNA-binding transcription activator activity	21/281	328/12866	0.001809	21
GO:0001221	transcription coregulator binding	10/281	97/12866	0.005956	10
GO:0031386	protein tag	4/281	11/12866	0.005956	4
GO:0050839	cell adhesion molecule binding	24/281	461/12866	0.005956	24
GO:0061629	RNA polymerase II-specific DNA-binding transcription factor binding	18/281	294/12866	0.005956	18
GO:0003735	structural constituent of ribosome	12/281	149/12866	0.006567	12
GO:0008013	beta-catenin binding	8/281	70/12866	0.007726	8
GO:0003713	transcription coactivator activity	15/281	238/12866	0.011609	15
GO:0098918	structural constituent of synapse	4/281	15/12866	0.011609	4
GO:0002039	p53 binding	7/281	59/12866	0.012028	7
GO:0031625	ubiquitin protein ligase binding	15/281	254/12866	0.018348	15
GO:0043495	protein-membrane adaptor activity	4/281	18/12866	0.019418	4
GO:0003712	transcription coregulator activity	21/281	443/12866	0.024512	21
GO:0031490	chromatin DNA binding	8/281	90/12866	0.024512	8
GO:0044389	ubiquitin-like protein ligase binding	15/281	270/12866	0.026366	15
GO:0005080	protein kinase C binding	5/281	40/12866	0.04763	5
GO:0008083	growth factor activity	7/281	88/12866	0.082714	7
GO:0005178	integrin binding	8/281	114/12866	0.084896	8

Table 13. Top 20 over-represented Gene Ontology cellular component terms.

ID	Description	GeneRatio	BgRatio	p.adjust	Count
GO:0022626	cytosolic ribosome	12/289	92/13086	0.000357	12
GO:0005938	cell cortex	19/289	255/13086	0.000856	19
GO:0005667	transcription regulator complex	23/289	391/13086	0.001942	23
GO:0030863	cortical cytoskeleton	10/289	87/13086	0.001942	10
GO:0005925	focal adhesion	22/289	368/13086	0.001942	22
GO:0015629	actin cytoskeleton	23/289	400/13086	0.001942	23
GO:0030055	cell-substrate junction	22/289	375/13086	0.001942	22
GO:0030864	cortical actin cytoskeleton	8/289	66/13086	0.005143	8
GO:0022627	cytosolic small ribosomal subunit	6/289	35/13086	0.005143	6
GO:0030018	Z disc	9/289	90/13086	0.007232	9
GO:0044391	ribosomal subunit	12/289	162/13086	0.010211	12
GO:0031674	I band	9/289	98/13086	0.011518	9
GO:0032432	actin filament bundle	7/289	63/13086	0.015632	7
GO:0042641	actomyosin	7/289	65/13086	0.017593	7
GO:0017053	transcription repressor complex	7/289	67/13086	0.019757	7

GO:0030016	myofibril	11/289	160/13086	0.023601	11
GO:0005840	ribosome	12/289	189/13086	0.026379	12
GO:0043292	contractile fiber	11/289	168/13086	0.031262	11
GO:0022625	cytosolic large ribosomal subunit	6/289	56/13086	0.031674	6
GO:0001725	stress fiber	6/289	57/13086	0.031674	6

Table 14. Top 20 over-represented Gene Ontology biological process terms.

ID	Description	GeneRatio	BgRatio	p.adjust	Count
GO:0010942	positive regulation of cell death	32/279	444/12549	1.58E-05	32
GO:0030099	myeloid cell differentiation	24/279	292/12549	6.58E-05	24
GO:0043065	positive regulation of apoptotic process	28/279	395/12549	7.47E-05	28
GO:0043068	positive regulation of programmed cell death	28/279	402/12549	8.11E-05	28
GO:0001525	angiogenesis	26/279	375/12549	0.000211	26
GO:0048514	blood vessel morphogenesis	28/279	435/12549	0.000275	28
GO:0001667	ameboidal-type cell migration	23/279	323/12549	0.000477	23
GO:0001568	blood vessel development	29/279	498/12549	0.000984	29
GO:0043542	endothelial cell migration	15/279	164/12549	0.001575	15
GO:0040017	positive regulation of locomotion	25/279	414/12549	0.001978	25
GO:0043534	blood vessel endothelial cell migration	11/279	92/12549	0.001978	11
GO:0009611	response to wounding	24/279	389/12549	0.001978	24
GO:0030335	positive regulation of cell migration	24/279	393/12549	0.002171	24
GO:0071363	cellular response to growth factor stimulus	27/279	481/12549	0.002532	27
GO:0050821	protein stabilization	14/279	157/12549	0.002676	14
GO:2000147	positive regulation of cell motility	24/279	406/12549	0.003041	24
GO:0051272	positive regulation of cellular component movement	24/279	417/12549	0.003948	24
GO:0002573	myeloid leukocyte differentiation	13/279	145/12549	0.003948	13
GO:0061061	muscle structure development	25/279	446/12549	0.003948	25
GO:1902895	positive regulation of miRNA transcription	7/279	39/12549	0.003948	7

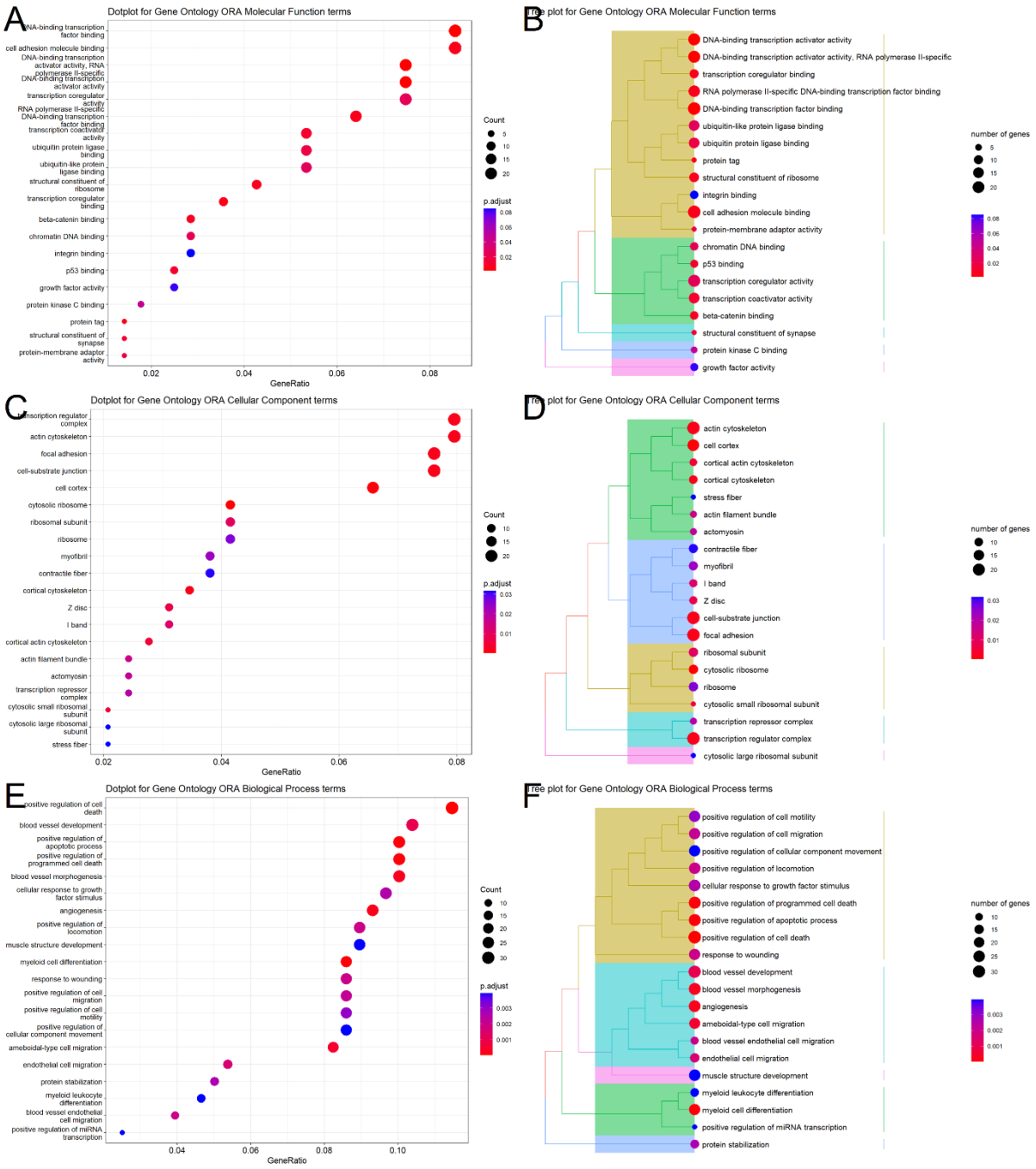


Figure 26. Dot plots and tree plots for the top 20 over-represented Gene Ontology MF, CC, and BP terms. **A, C, E.** Dot plot showing the gene ratios for each term. **B, D, F.** Tree plots showing the hierarchical clustering of terms indicating similarities and differences in the genes grouped under the terms. The dot size reflects the gene count for the term and the color reflects the adjusted p-value.

Table 15. Top 20 enriched Gene Ontology molecular function terms.

ID	Description	setSize	enrichmentScore	NES	p.adjust
GO:0003712	transcription coregulator activity	443	-0.367721411	-1.75510333	9.38E-06
GO:0003713	transcription coactivator activity	238	-0.427338934	-1.911377804	1.22E-05
GO:0003682	chromatin binding	482	-0.341548407	-1.645213435	2.41E-05
GO:0005096	GTPase activator activity	220	-0.374523777	-1.668809451	0.005451
GO:0035091	phosphatidylinositol binding	232	-0.359172199	-1.613699073	0.01338
GO:0042393	histone binding	218	-0.352960724	-1.573586406	0.01338
GO:0005543	phospholipid binding	375	-0.31885018	-1.501008928	0.01338
GO:0003779	actin binding	351	-0.318647905	-1.492898018	0.01338
GO:0030695	GTPase regulator activity	416	-0.308387012	-1.465912254	0.01338
GO:0060589	nucleoside-triphosphatase regulator activity	416	-0.308387012	-1.465912254	0.01338
GO:0045296	cadherin binding	303	-0.325209346	-1.499226978	0.01801
GO:0106310	protein serine kinase activity	319	-0.315449024	-1.463197893	0.021866
GO:0008083	growth factor activity	88	-0.447757615	-1.743414335	0.02199
GO:0003735	structural constituent of ribosome	149	-0.378119748	-1.605334402	0.02199
GO:0004712	protein serine/threonine/tyrosine kinase activity	383	-0.30612939	-1.44494018	0.023064
GO:0033549	MAP kinase phosphatase activity	11	-0.788178237	-1.902546181	0.02989
GO:0002039	p53 binding	59	-0.485605785	-1.752104451	0.02989
GO:0017017	MAP kinase tyrosine/serine/threonine phosphatase activity	10	-0.810553042	-1.896526189	0.029904
GO:0003714	transcription corepressor activity	171	-0.366495371	-1.598328068	0.031498
GO:0140030	modification-dependent protein binding	145	-0.375640032	-1.586583781	0.031554

Table 16. Top 20 enriched Gene Ontology cellular component terms.

ID	Description	setSize	enrichmentScore	NES	p.adjust
GO:0030496	midbody	190	-0.41587	-1.82663	0.000396
GO:0016607	nuclear speck	357	-0.35992	-1.69284	0.000396
GO:0032432	actin filament bundle	63	-0.55826	-2.05089	0.000613
GO:0005819	spindle	369	-0.34021	-1.60338	0.001093
GO:0010008	endosome membrane	427	-0.32745	-1.56722	0.001093
GO:0005925	focal adhesion	368	-0.33738	-1.591	0.001268
GO:0030055	cell-substrate junction	375	-0.3292	-1.55498	0.001526
GO:0015629	actin cytoskeleton	400	-0.32496	-1.54425	0.001682
GO:0005911	cell-cell junction	396	-0.32227	-1.52719	0.002452
GO:0042641	actomyosin	65	-0.52421	-1.93166	0.002726
GO:0005938	cell cortex	255	-0.35031	-1.59406	0.003145
GO:0001725	stress fiber	57	-0.52183	-1.87862	0.004151
GO:0097517	contractile actin filament bundle	57	-0.52183	-1.87862	0.004151
GO:0005874	microtubule	378	-0.31896	-1.50556	0.004166
GO:0031252	cell leading edge	352	-0.32267	-1.52134	0.004858

GO:0005770	late endosome	229	-0.34803	-1.5715	0.006853
GO:0005667	transcription regulator complex	391	-0.31066	-1.47086	0.006853
GO:0044391	ribosomal subunit	162	-0.38556	-1.66032	0.008723
GO:0071013	catalytic step 2 spliceosome	84	-0.45635	-1.77608	0.011868
GO:0005774	vacuolar membrane	376	-0.30118	-1.41893	0.015367

Table 17. Top 20 enriched Gene Ontology biological process terms.

ID	Description	setSize	enrichmentScore	NES	p.adjust
GO:0043408	regulation of MAPK cascade	458	-0.36473	-1.75475	2.10E-05
GO:0006325	chromatin organization	482	-0.3464	-1.68188	0.000115
GO:0120035	regulation of plasma membrane bounded cell projection organization	491	-0.3441	-1.67181	0.000115
GO:0006914	autophagy	471	-0.34365	-1.66474	0.000115
GO:0061919	process utilizing autophagic mechanism	471	-0.34365	-1.66474	0.000115
GO:0010942	positive regulation of cell death	444	-0.34695	-1.6641	0.000115
GO:0045936	negative regulation of phosphate metabolic process	347	-0.36617	-1.71397	0.000195
GO:0010563	negative regulation of phosphorus metabolic process	348	-0.36474	-1.70927	0.000195
GO:0031098	stress-activated protein kinase signaling cascade	197	-0.41934	-1.86111	0.00029
GO:1902532	negative regulation of intracellular signal transduction	411	-0.34766	-1.65823	0.00029
GO:0043068	positive regulation of programmed cell death	402	-0.34969	-1.66806	0.000349
GO:0043065	positive regulation of apoptotic process	395	-0.35076	-1.66845	0.000361
GO:0051403	stress-activated MAPK cascade	192	-0.42093	-1.85526	0.000362
GO:0048729	tissue morphogenesis	437	-0.33276	-1.59548	0.000413
GO:0016570	histone modification	430	-0.33894	-1.61903	0.000596
GO:0010506	regulation of autophagy	295	-0.3673	-1.69466	0.000886
GO:0002009	morphogenesis of an epithelium	373	-0.34754	-1.64152	0.000886
GO:0006397	mRNA processing	429	-0.33389	-1.5952	0.000886
GO:0090130	tissue migration	234	-0.38029	-1.71706	0.001366
GO:0043409	negative regulation of MAPK cascade	132	-0.45529	-1.90543	0.002177

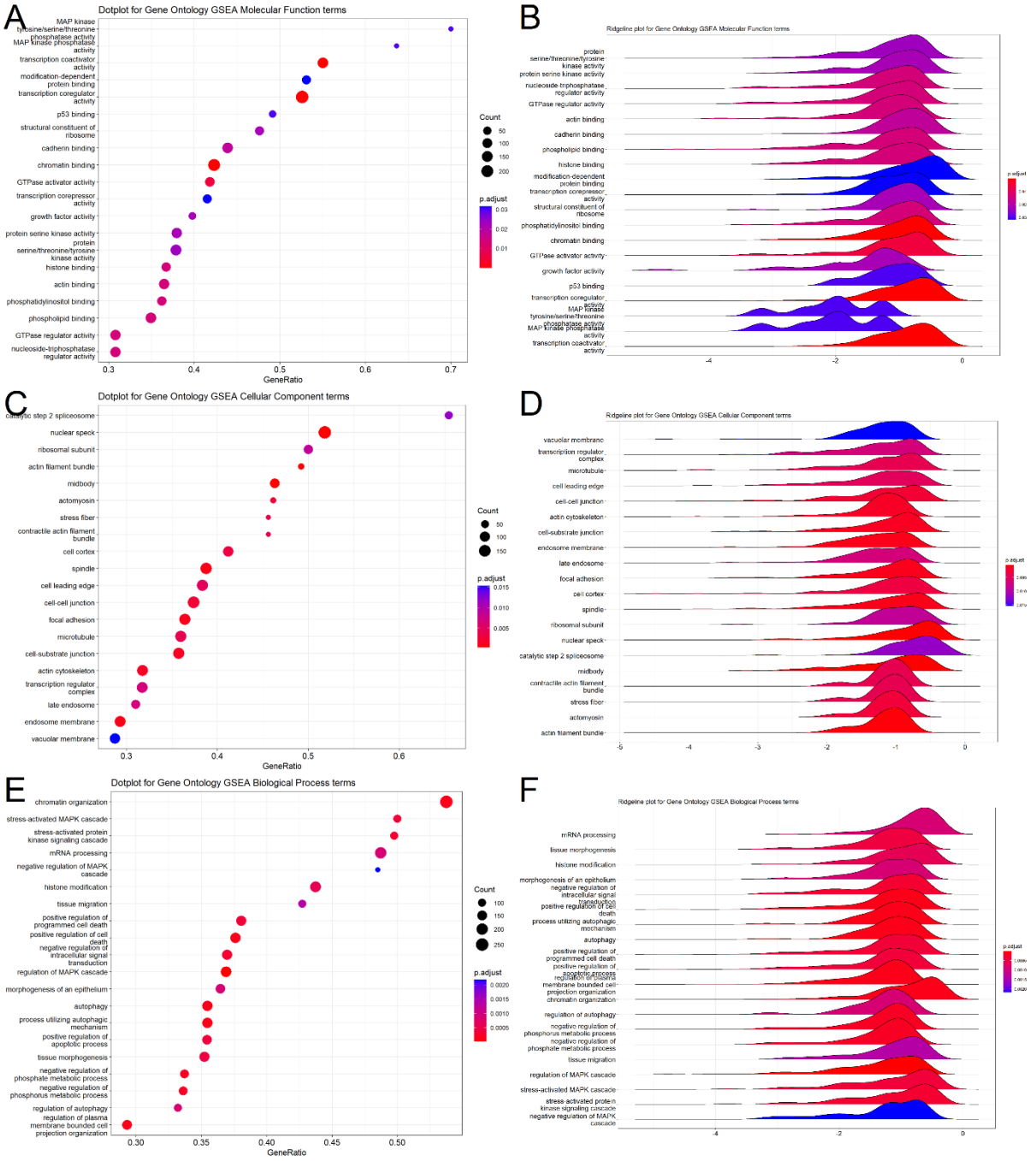


Figure 27. Dot plots and ridgeline plots for the top 20 enriched Gene Ontology MF, CC, and BP terms. **A, C, E.** Dot plot showing the gene ratios for each term. The dot size reflects the gene count for the term and the color reflects the adjusted p-value. **B, D, F.** Ridgeline plots showing the distribution of log2 fold change values for the core enrichment genes for each term. The color reflects the adjusted p-value.

MCF-7 MR overexpression aldosterone + spironolactone treatment vs. MCF-7 MR overexpression aldosterone treatment

Spironolactone is one of the antagonists of aldosterone, so a combinatorial treatment with aldosterone and spironolactone can provide insights into a potential reversal of effects that aldosterone treatment has on the breast cancer cell line. To test the effect of spironolactone treatment, MCF-7 cells were treated with a combination of aldosterone and spironolactone. A comparison between MCF_7_OVAS and MCF_7_OVA RNA-seq samples was conducted and the results of the differential analysis were summarized in the form of top 30 upregulated and downregulated genes in **Table 18** and **Table 19** (full lists were included in **Appendix 1**). The distribution of log2 fold change values and adjusted p-values was plotted in **Figure 28**, where the top 10 upregulated and top 10 downregulated genes were also labeled. The heatmap in **Figure 30** showed that the log2FC values very closely matched across all methods.

As in the previous section, the selection of genes for further analyses was done on the basis that two methods (edgeR and limma) to discover more DEGs for this dataset (**Figure 29**), therefore genes that were discovered by at least three methods were used to obtain results from functional analyses. Following this criterion, a gene set was selected that consisted of 143 genes that were discovered by all four methods and 319 genes that were discovered by any combination of three methods.

From the over-representation analysis using the molecular function (MF), cellular component (CC), and biological process (BP) the results showed that very few Gene Ontology terms were over-represented in the selected gene list. CC terms including terms that involved mitochondrial inner membrane, ribosomal subunits, and other complexes found in the cell were over-represented (top 20 terms for three ontologies are listed in **Table 21**, **Table 22**, and **Table 23**). There were some terms that involved similar biological processes and molecular functions as the terms from the results of the functional analysis using differentially expressed genes in

the OVAS vs. OVA comparison (previous section). These included DNA, chromatin, binding, transcription activity, and some cell death terms. However, in most cases the GO terms in this part of the ORA were not statistically significant and the dot plots and tree plots reflect this fact (**Figure 31**).

The similarities were also lacking in GSEA enriched terms. With exception to some GO terms that were related to ribosomal subunits, GO terms from MF, CC, and BP ontologies were different from one comparison to the other (**Table *24***, **Table *25***, and **Table *26***). To add to this, the GSEA terms had mostly non-significant enrichment values, meaning that the spironolactone treatment in combination with aldosterone has affected the expression of genes such that enriched terms had little similarity with OVA vs OV comparisons other than a few terms.

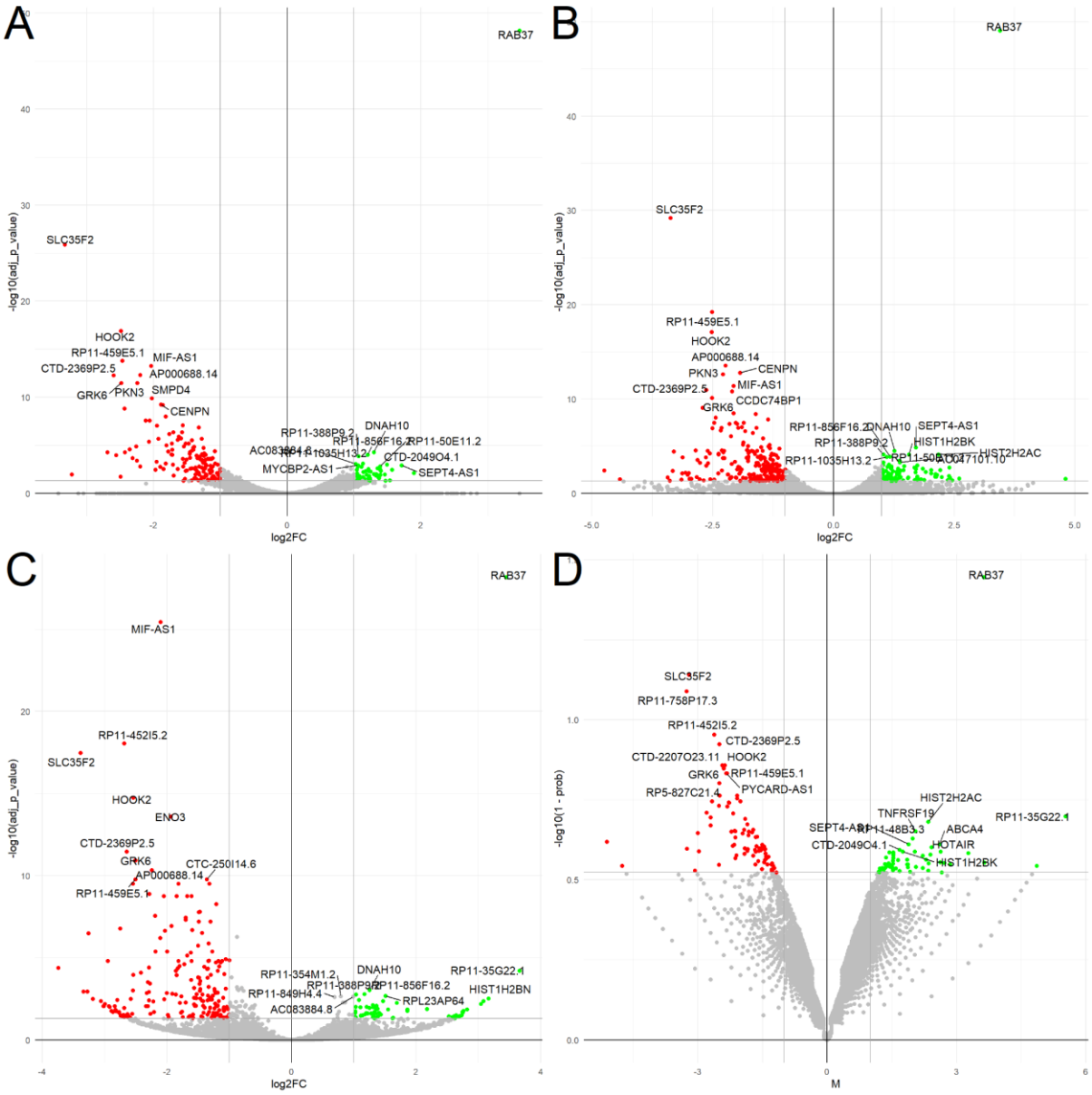


Figure 28. The volcano plots visualizing DGEA results of MCF_7_OVAS vs MCF_7_OVA. (**A.** DESeq2, **B.** edgeR, **C.** limma, **D.** NOISeq). The top 10 upregulated (green color) and downregulated genes (red color) are labeled (ranked by adjusted p-values). p-value threshold selected as 0.05 or prob value threshold as 0.70, and log2 fold change absolute value threshold selected as 1.

Table 18. List of top 30 upregulated genes discovered in common using all four methods in the comparison of MCF_7_OVAS vs. MCF_7_OVA groups. Average voom normalized counts for the analyzed groups and DGEA statistics are given. Full list is in **Appendix 1**.

HGNC symbol	average voom counts OVA	average voom counts OVAS	DESeq log2FC	DESeq p value	edgeR log2FC	edgeR p value	limma log2FC	limma p value	NOISeq M value	NOISeq prob value
RAB37	6.483	9.926	3.47	1.54E-52	3.44	3.86E-54	3.44	3.12E-33	3.63	0.964
DNAH10	6.494	7.751	1.29	5.34E-07	1.26	7.12E-08	1.25	3.70E-06	1.44	0.741
RP11-856F16.2	6.822	7.988	1.20	1.22E-06	1.17	4.13E-07	1.17	7.72E-06	1.37	0.731
RP11-388P9.2	7.959	8.992	1.07	1.86E-06	1.03	2.12E-07	1.03	8.53E-06	1.22	0.708
RP11-1035H13.2	6.172	7.279	1.13	1.52E-05	1.11	4.70E-07	1.10	8.63E-05	1.28	0.709
CTD-2049O4.1	4.539	5.987	1.49	2.10E-05	1.47	6.86E-06	1.46	2.76E-05	1.67	0.745
MYCBP2-AS1	6.817	7.855	1.06	2.46E-05	1.03	2.25E-06	1.03	8.28E-05	1.20	0.702
SEPT4-AS1	3.551	5.249	1.71	2.48E-05	1.70	2.38E-08	1.69	3.82E-05	1.89	0.755
RP11-50E11.2	4.693	6.075	1.39	3.53E-05	1.37	1.32E-06	1.36	6.42E-05	1.53	0.735
RP11-793H13.3	7.106	8.203	1.11	3.85E-05	1.08	3.09E-05	1.08	2.17E-05	1.23	0.706
TMEM186	6.454	7.477	1.06	3.89E-05	1.03	4.83E-06	1.02	1.79E-04	1.21	0.701
RP11-317G6.1	5.084	6.375	1.35	5.97E-05	1.32	9.53E-05	1.31	5.59E-05	1.53	0.740
FCMR	4.881	6.186	1.37	6.44E-05	1.35	7.27E-05	1.31	8.51E-05	1.54	0.737
RP11-110I1.11	5.502	6.730	1.22	8.06E-05	1.19	1.66E-05	1.20	7.98E-05	1.34	0.718
AC017101.10	4.396	5.804	1.39	9.23E-05	1.37	2.17E-06	1.38	1.11E-04	1.53	0.729
HIST1H2BK	3.654	5.207	1.57	9.59E-05	1.56	2.07E-07	1.55	1.26E-04	1.75	0.742
WDFY3-AS1	4.447	5.832	1.40	1.31E-04	1.38	2.76E-05	1.36	1.31E-04	1.54	0.730
NONOP2	5.107	6.253	1.17	1.47E-04	1.15	4.56E-06	1.14	3.92E-04	1.33	0.708
MBNL1-AS1	4.975	6.195	1.22	1.53E-04	1.20	7.96E-06	1.20	2.50E-04	1.36	0.711
TAB3-AS2	5.109	6.394	1.25	1.68E-04	1.23	3.90E-05	1.25	1.00E-04	1.37	0.715
HIST2H2AC	1.931	4.200	2.14	2.41E-04	2.17	2.25E-07	2.18	1.21E-04	2.35	0.792
TNFRSF19	2.703	4.644	1.90	2.51E-04	1.91	4.86E-05	1.86	1.27E-04	2.05	0.777
RP11-18D7.2	4.380	5.752	1.33	3.34E-04	1.31	4.79E-05	1.33	2.07E-04	1.45	0.716
CTD-3032J10.2	4.853	6.104	1.30	3.37E-04	1.28	3.24E-04	1.22	2.65E-04	1.42	0.719
RP11-815J21.2	4.264	5.566	1.34	3.72E-04	1.32	9.05E-05	1.32	3.10E-04	1.53	0.724
C16orf45	3.746	5.169	1.46	3.79E-04	1.45	2.60E-05	1.43	3.29E-04	1.65	0.726
SP2-AS1	5.000	6.215	1.19	4.25E-04	1.17	1.17E-04	1.18	3.03E-04	1.31	0.705

CTD-2310F14.1	3.984	5.439	1.40	4.35E-04	1.38	3.17E-05	1.40	2.35E-04	1.52	0.719
ANKRD61	4.206	5.607	1.33	6.19E-04	1.31	9.70E-05	1.35	2.55E-04	1.45	0.712
CTB-193M12.1	3.817	5.171	1.35	7.68E-04	1.34	1.65E-05	1.33	7.59E-04	1.51	0.711

Table 19. List of top 30 downregulated genes discovered in common using all four methods in the comparison of MCF_7_OVAS vs. MCF_7_OVA groups. Average voom normalized counts for the analyzed groups and DGEA statistics are given. Full list is in **Appendix 1**.

HGNC symbol	**average voom counts OVA**	**average voom counts OVAS**	**DESeq log2FC**	**DESeq p value**	**edgeR log2FC**	**edgeR p value**	**limma log2FC**	**limma p value**	**NOISeq M value**	**NOISeq prob value**
SLC35F2	7.884	4.496	-3.33	5.79E-30	-3.37	5.79E-34	-3.39	6.11E-22	-3.20	0.928
HOOK2	8.706	6.174	-2.49	7.97E-21	-2.52	1.38E-21	-2.54	4.05E-19	-2.37	0.862
RP11-459E5.1	7.219	4.709	-2.47	1.48E-17	-2.51	8.31E-24	-2.50	8.48E-14	-2.32	0.853
MIF-AS1	13.149	11.050	-2.04	6.14E-17	-2.07	1.51E-15	-2.10	3.35E-30	-1.90	0.796
AP000688.14	8.041	5.804	-2.20	6.70E-16	-2.24	6.53E-18	-2.24	1.93E-14	-2.08	0.828
CTD-2369P2.5	7.531	4.900	-2.60	8.40E-16	-2.63	4.42E-15	-2.65	1.10E-15	-2.50	0.881
GRK6	7.636	5.150	-2.48	5.78E-15	-2.52	3.83E-14	-2.51	4.36E-15	-2.40	0.858
PKN3	7.375	5.083	-2.24	6.67E-15	-2.29	7.72E-17	-2.28	8.69E-13	-2.08	0.824
SMPD4	7.981	5.918	-2.02	2.79E-13	-2.07	2.00E-12	-2.05	1.47E-12	-1.85	0.788
CENPN	7.459	5.528	-1.89	1.35E-12	-1.93	4.61E-17	-1.93	1.49E-10	-1.74	0.775
ENO3	10.478	8.537	-1.86	1.74E-12	-1.90	9.54E-11	-1.94	6.99E-18	-1.72	0.780
PYCARD-AS1	7.082	4.551	-2.43	4.40E-12	-2.47	2.98E-11	-2.55	1.98E-13	-2.34	0.853
GDF9	9.369	7.533	-1.82	3.27E-11	-1.87	5.94E-09	-1.82	2.03E-13	-1.62	0.768
CFAP57	7.042	4.862	-2.12	8.79E-11	-2.16	4.07E-11	-2.19	2.93E-11	-2.01	0.820
CCDC74BP1	6.613	4.502	-2.06	9.71E-11	-2.10	7.32E-15	-2.11	9.93E-10	-1.92	0.779
PDAP1	9.901	8.301	-1.56	3.12E-10	-1.59	1.20E-10	-1.61	1.28E-12	-1.44	0.746
RP11-196G11.6	7.029	4.999	-1.95	3.81E-10	-1.99	4.27E-11	-2.03	3.18E-10	-1.82	0.776
CTC-250I14.6	11.875	10.516	-1.32	6.04E-10	-1.36	1.12E-11	-1.36	7.95E-14	-1.18	0.700
NDUFB2-AS1	8.722	6.999	-1.71	2.06E-09	-1.75	1.23E-07	-1.70	5.61E-11	-1.51	0.754
H1FX	7.439	5.827	-1.57	2.09E-09	-1.61	2.47E-12	-1.60	3.25E-08	-1.41	0.729
AP006621.8	8.819	6.987	-1.76	3.25E-09	-1.79	4.85E-08	-1.85	1.34E-12	-1.66	0.768
PRR19	8.771	7.067	-1.62	6.49E-09	-1.66	8.77E-08	-1.70	3.92E-11	-1.47	0.751
RP3-416H24.1	9.586	7.908	-1.59	1.06E-08	-1.63	1.96E-07	-1.68	1.52E-12	-1.44	0.746
MROH8	6.512	4.524	-1.93	1.18E-08	-1.97	1.31E-10	-2.00	7.19E-09	-1.82	0.770

LA16c-366D1.3	7.383	5.721	-1.64	1.60E-08	-1.67	5.27E-10	-1.67	1.37E-08	-1.52	0.748
LAMA5	10.031	8.537	-1.43	1.80E-08	-1.47	7.11E-08	-1.49	1.69E-11	-1.28	0.719
CTD-3065J16.9	6.127	3.949	-2.07	2.68E-08	-2.12	1.22E-09	-2.18	7.25E-09	-1.96	0.774
P2RY11	8.703	7.188	-1.47	4.43E-08	-1.52	3.59E-07	-1.50	3.06E-09	-1.29	0.717
CTB-31O20.3	6.861	5.113	-1.73	4.45E-08	-1.78	5.40E-08	-1.73	6.18E-08	-1.54	0.744
LBHD1	8.813	7.212	-1.52	4.63E-08	-1.56	4.63E-07	-1.60	2.84E-10	-1.36	0.732

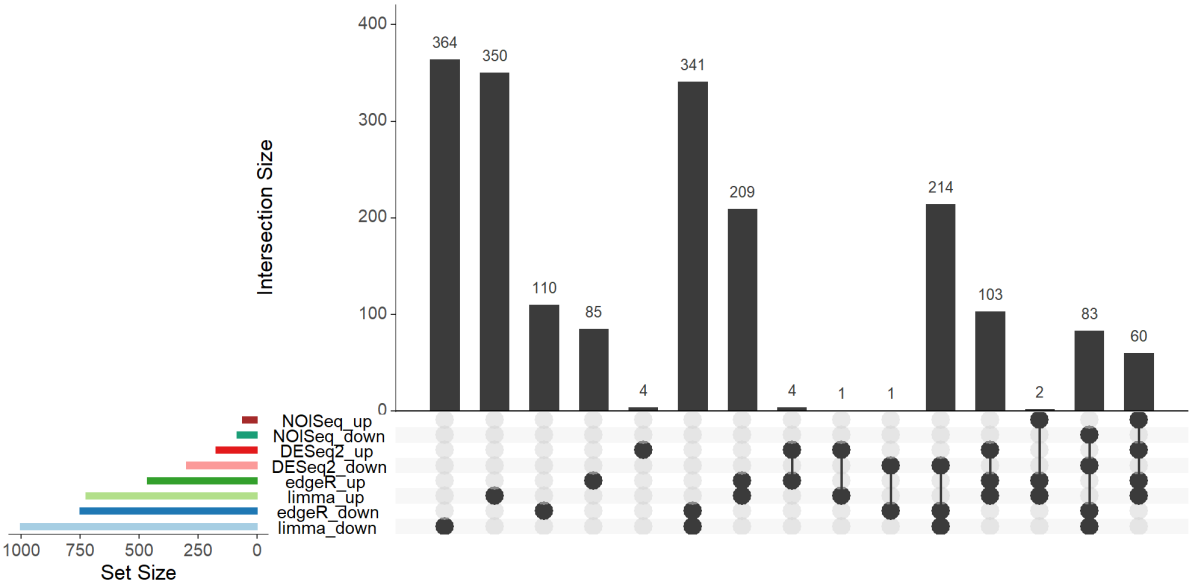


Figure 29. UpSet plot showing the intersections of DEGs discovered by different methods. Intersections are indicated by black dots corresponding to each gene set (2 for each method).

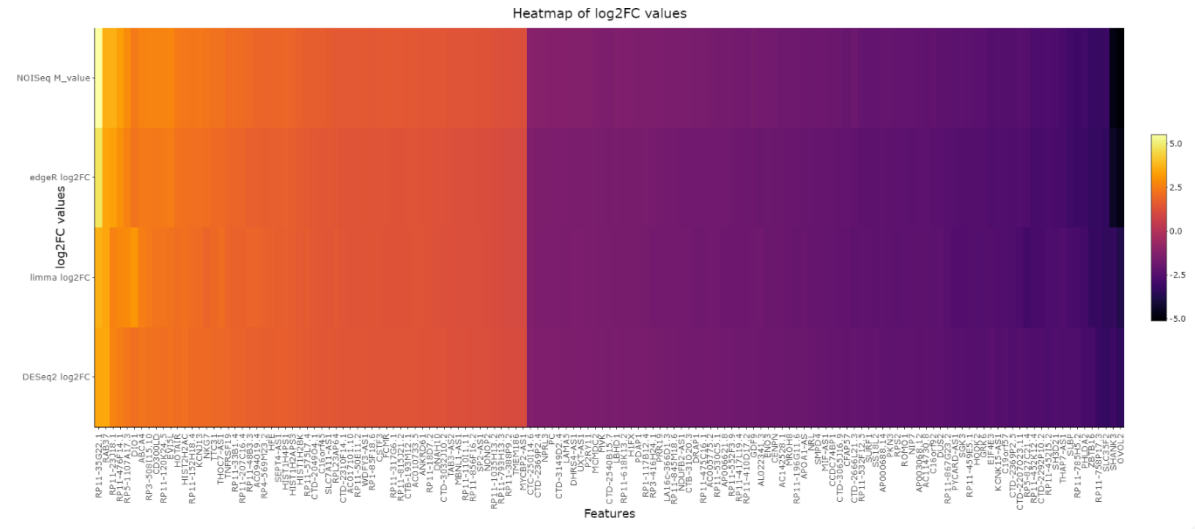


Figure 30. Heatmap of log2 fold change values for the 143 genes discovered by all methods.

Table 20. Summarized results from the comparison of the four DGEA methods.

Intersection	**Upregulated genes**	**Downregulated genes**
DESeq2, edgeR, limma, NOISeq	60	83
DESeq2, edgeR, limma	103	214
DESeq2, edgeR, NOISeq	0	0
DESeq2, limma, NOISeq	0	0
edgeR, limma, NOISeq	2	0
DESeq2, edgeR	4	1
DESeq2, limma	1	0
DESeq2, NOISeq	0	0
edgeR, limma	209	341
edgeR, NOISeq	0	0
limma, NOISeq	0	0
DESeq2	4	0
edgeR	85	110
limma	350	364
NOISeq	0	0
Number of methods	**Upregulated genes**	**Downregulated genes**
all 4 methods	60	83
any 3 methods	105	214
any 2 methods	214	342
any 1 method	439	474
Number of methods	**Upregulated genes**	**Downregulated genes**
all 4 methods	60	83
at least 3 methods	165	297
at least 2 methods	379	639
at least 1 method	818	1,113

Table 21. Top 20 over-represented Gene Ontology molecular function terms.

ID	Description	GeneRatio	BgRatio	p.adjust	Count
GO:0005246	calcium channel regulator activity	3/222	27/12866	0.864204	3
GO:0002020	protease binding	5/222	97/12866	0.864204	5
GO:0017124	SH3 domain binding	5/222	103/12866	0.864204	5
GO:0009931	calcium-dependent protein serine/threonine kinase activity	2/222	19/12866	0.864204	2
GO:0022829	wide pore channel activity	2/222	19/12866	0.864204	2
GO:0003735	structural constituent of ribosome	6/222	149/12866	0.864204	6
GO:0046982	protein heterodimerization activity	8/222	229/12866	0.864204	8
GO:0010857	calcium-dependent protein kinase activity	2/222	20/12866	0.864204	2
GO:0003697	single-stranded DNA binding	5/222	115/12866	0.864204	5
GO:0004722	protein serine/threonine phosphatase activity	4/222	82/12866	0.864204	4
GO:0051959	dynein light intermediate chain binding	2/222	23/12866	0.864204	2
GO:0090482	vitamin transmembrane transporter activity	2/222	25/12866	0.864204	2
GO:0046966	nuclear thyroid hormone receptor binding	2/222	28/12866	0.864204	2
GO:0003924	GTPase activity	8/222	271/12866	0.864204	8
GO:0016251	RNA polymerase II general transcription initiation factor activity	2/222	33/12866	0.864204	2
GO:0004721	phosphoprotein phosphatase activity	5/222	147/12866	0.864204	5
GO:0015453	oxidoreduction-driven active transmembrane transporter activity	3/222	68/12866	0.864204	3
GO:0009055	electron transfer activity	4/222	109/12866	0.864204	4
GO:0031492	nucleosomal DNA binding	2/222	35/12866	0.864204	2
GO:0032451	demethylase activity	2/222	35/12866	0.864204	2

Table 22. Top 20 over-represented Gene Ontology cellular component terms.

ID	Description	GeneRatio	BgRatio	p.adjust	Count
GO:0000786	nucleosome	6/224	67/13086	0.215677	6
GO:0032993	protein-DNA complex	9/224	151/13086	0.215677	9
GO:0044815	DNA packaging complex	7/224	129/13086	0.765995	7
GO:0015935	small ribosomal subunit	4/224	58/13086	0.765995	4
GO:0022627	cytosolic small ribosomal subunit	3/224	35/13086	0.765995	3
GO:0008287	protein serine/threonine phosphatase complex	3/224	40/13086	0.765995	3
GO:1903293	phosphatase complex	3/224	40/13086	0.765995	3
GO:0032426	stereocilium tip	2/224	17/13086	0.765995	2
GO:0098685	Schaffer collateral - CA1 synapse	3/224	49/13086	0.765995	3
GO:0005963	magnesium-dependent protein serine/threonine phosphatase complex	2/224	23/13086	0.765995	2
GO:0031307	integral component of mitochondrial outer membrane	2/224	23/13086	0.765995	2
GO:0044391	ribosomal subunit	6/224	162/13086	0.765995	6
GO:0031306	intrinsic component of mitochondrial outer membrane	2/224	24/13086	0.765995	2

GO:0000794	condensed nuclear chromosome	3/224	56/13086	0.765995	3
GO:0005689	U12-type spliceosomal complex	2/224	26/13086	0.765995	2
GO:0022626	cytosolic ribosome	4/224	92/13086	0.765995	4
GO:0005775	vacuolar lumen	5/224	133/13086	0.765995	5
GO:0042788	polysomal ribosome	2/224	29/13086	0.765995	2
GO:0090543	Flemming body	2/224	30/13086	0.765995	2
GO:0035861	site of double-strand break	3/224	64/13086	0.765995	3

Table 23. Top 20 over-represented Gene Ontology biological process terms.

ID	Description	GeneRatio	BgRatio	p.adjust	Count
GO:0031640	killing of cells of another organism	3/218	12/12549	0.893487	3
GO:0014067	negative regulation of phosphatidylinositol 3-kinase signaling	3/218	14/12549	0.893487	3
GO:0007190	activation of adenylate cyclase activity	3/218	17/12549	0.893487	3
GO:0042255	ribosome assembly	5/218	58/12549	0.893487	5
GO:0061844	antimicrobial humoral immune response mediated by antimicrobial peptide	3/218	20/12549	0.893487	3
GO:0051339	regulation of lyase activity	4/218	43/12549	0.893487	4
GO:0045762	positive regulation of adenylate cyclase activity	3/218	23/12549	0.893487	3
GO:0060674	placenta blood vessel development	3/218	25/12549	0.893487	3
GO:0006084	acetyl-CoA metabolic process	3/218	27/12549	0.893487	3
GO:0031281	positive regulation of cyclase activity	3/218	27/12549	0.893487	3
GO:0051349	positive regulation of lyase activity	3/218	27/12549	0.893487	3
GO:0006086	acetyl-CoA biosynthetic process from pyruvate	2/218	10/12549	0.893487	2
GO:0051081	nuclear membrane disassembly	2/218	10/12549	0.893487	2
GO:0071480	cellular response to gamma radiation	3/218	29/12549	0.893487	3
GO:1901797	negative regulation of signal transduction by p53 class mediator	3/218	29/12549	0.893487	3
GO:0030397	membrane disassembly	2/218	11/12549	0.893487	2
GO:0001906	cell killing	5/218	86/12549	0.893487	5
GO:2001020	regulation of response to DNA damage stimulus	9/218	226/12549	0.893487	9
GO:0045822	negative regulation of heart contraction	2/218	12/12549	0.893487	2
GO:1903523	negative regulation of blood circulation	2/218	12/12549	0.893487	2

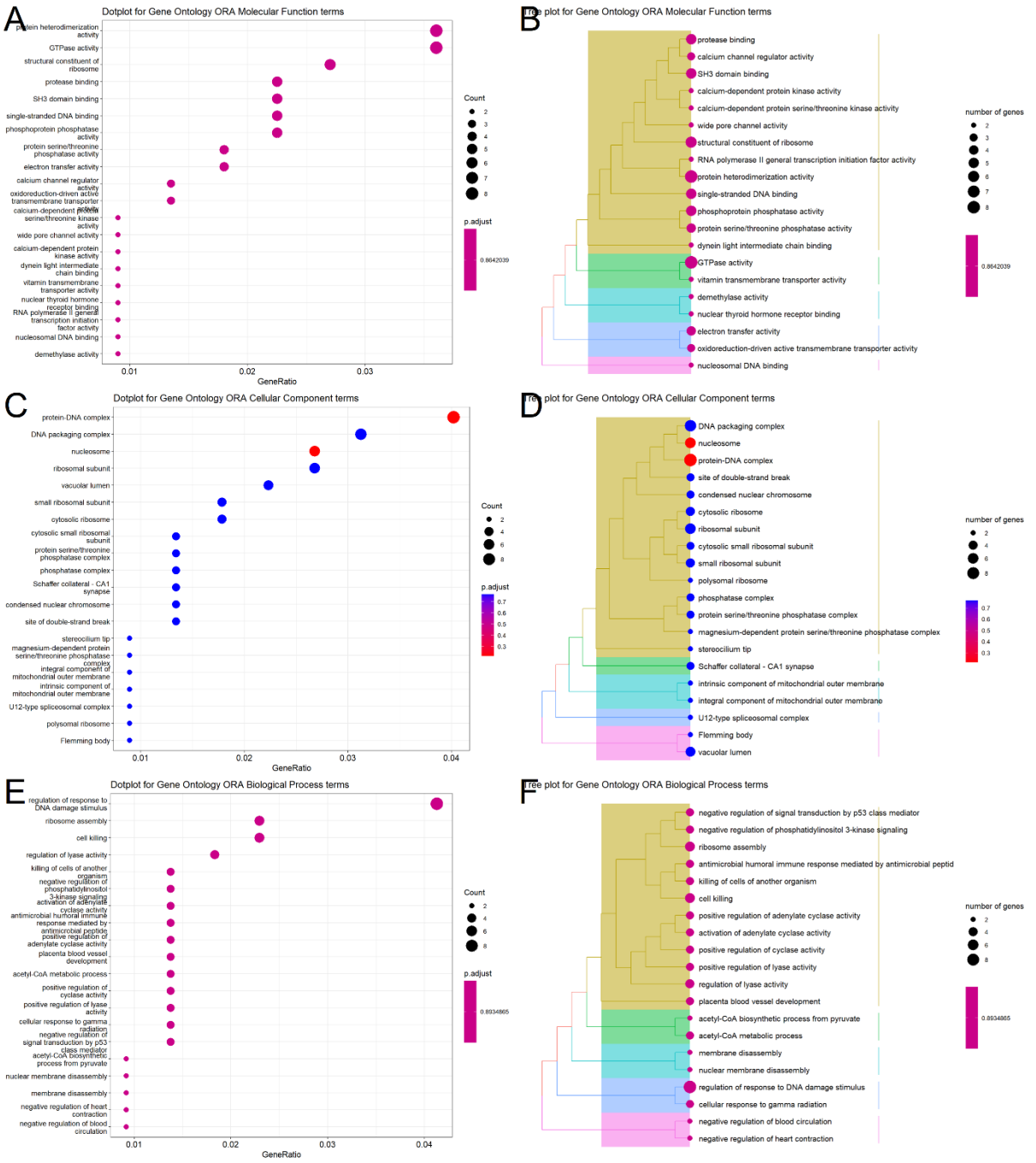


Figure 31. Dot plots and tree plots for the top 20 over-represented Gene Ontology MF, CC, and BP terms. **A, C, E.** Dot plot showing the gene ratios for each term. **B, D, F.** Tree plots showing the hierarchical clustering of terms indicating similarities and differences in the genes grouped under the terms. The dot size reflects the gene count for the term and the color reflects the adjusted p-value.

Table 24. Top 20 enriched Gene Ontology molecular function terms.

ID	Description	setSize	enrichmentScore	NES	p.adjust
GO:0003735	structural constituent of ribosome	149	-0.53181	-2.18356	3.94E-07
GO:0140333	glycerophospholipid flippase activity	13	0.752762	1.874746	0.537156
GO:0140327	flippase activity	16	0.690616	1.821846	0.689543
GO:0008569	minus-end-directed microtubule motor activity	13	0.688995	1.715936	0.984675
GO:0004984	olfactory receptor activity	21	0.603548	1.68431	0.984675
GO:0030215	semaphorin receptor binding	17	0.627441	1.672214	0.984675
GO:0005164	tumor necrosis factor receptor binding	20	0.596651	1.671059	0.984675
GO:0000049	tRNA binding	72	-0.43209	-1.58546	0.984675
GO:0070742	C2H2 zinc finger domain binding	12	-0.6947	-1.66158	1
GO:0016783	sulfurtransferase activity	10	-0.72203	-1.63877	1
GO:0070182	DNA polymerase binding	19	-0.592	-1.62887	1
GO:0098918	structural constituent of synapse	15	0.614612	1.595242	1
GO:0016229	steroid dehydrogenase activity	20	-0.56564	-1.57334	1
GO:0001530	lipopolysaccharide binding	17	0.588381	1.568116	1
GO:0003785	actin monomer binding	21	-0.5442	-1.54756	1
GO:0008191	metalloendopeptidase inhibitor activity	10	-0.68167	-1.54716	1
GO:0022884	macromolecule transmembrane transporter activity	25	-0.51595	-1.53163	1
GO:0017046	peptide hormone binding	26	-0.51244	-1.53082	1
GO:0033764	steroid dehydrogenase activity, acting on the CH-OH group of donors, NAD or NADP as acceptor	18	-0.55963	-1.52875	1
GO:0052744	phosphatidylinositol monophosphate phosphatase activity	18	0.560735	1.515567	1

Table 25. Top 20 enriched Gene Ontology cellular component terms.

ID	Description	setSize	enrichmentScore	NES	p.adjust
GO:0005743	mitochondrial inner membrane	442	-0.39792	-1.84717	6.99E-08
GO:0019866	organelle inner membrane	488	-0.38	-1.77655	1.11E-07
GO:0005840	ribosome	189	-0.48079	-2.03789	3.79E-07
GO:0044391	ribosomal subunit	162	-0.48768	-2.01057	3.17E-06
GO:0098798	mitochondrial protein-containing complex	261	-0.41661	-1.83301	1.71E-05
GO:0022626	cytosolic ribosome	92	-0.51564	-1.98968	0.000268
GO:0015934	large ribosomal subunit	107	-0.46476	-1.8219	0.003709
GO:0098800	inner mitochondrial membrane protein complex	142	-0.4203	-1.71816	0.008371
GO:0000313	organellar ribosome	75	-0.49589	-1.86196	0.011287
GO:0005761	mitochondrial ribosome	75	-0.49589	-1.86196	0.011287
GO:0015935	small ribosomal subunit	58	-0.51927	-1.85651	0.011409
GO:0022625	cytosolic large ribosomal subunit	56	-0.51391	-1.82965	0.019121
GO:0000502	proteasome complex	54	-0.51284	-1.8159	0.019121
GO:1905369	endopeptidase complex	70	-0.46893	-1.72838	0.027856
GO:0030687	preribosome, large subunit precursor	22	-0.63174	-1.8304	0.062135

GO:0005684	U2-type spliceosomal complex	86	-0.41451	-1.5825	0.09234
GO:0098799	outer mitochondrial membrane protein complex	22	-0.61	-1.76739	0.098823
GO:0022627	cytosolic small ribosomal subunit	35	-0.54455	-1.74787	0.098823
GO:0000786	nucleosome	67	0.449479	1.644761	0.110967
GO:0019867	outer membrane	202	-0.32946	-1.40804	0.132791

Table 26. Top 20 enriched Gene Ontology biological process terms.

ID	Description	setSize	enrichmentScore	NES	p.adjust
GO:0007005	mitochondrion organization	475	-0.32101	-1.49277	0.302464
GO:0002181	cytoplasmic translation	140	-0.41397	-1.69434	0.340101
GO:1903321	negative regulation of protein modification by small protein conjugation or removal	84	-0.45459	-1.7173	0.692283
GO:0043011	myeloid dendritic cell differentiation	15	-0.69795	-1.80198	0.791357
GO:0009415	response to water	10	-0.79053	-1.80052	0.791357
GO:0060004	reflex	11	0.744286	1.794743	0.791357
GO:0046794	transport of virus	16	-0.67967	-1.78981	0.791357
GO:0046500	S-adenosylmethionine metabolic process	14	-0.70131	-1.7713	0.791357
GO:0097028	dendritic cell differentiation	27	-0.58105	-1.74492	0.791357
GO:0048841	regulation of axon extension involved in axon guidance	26	0.574354	1.737372	0.791357
GO:1900102	negative regulation of endoplasmic reticulum unfolded protein response	14	-0.68583	-1.73222	0.791357
GO:0021872	forebrain generation of neurons	31	-0.55621	-1.73028	0.791357
GO:0060143	positive regulation of syncytium formation by plasma membrane fusion	14	-0.68447	-1.72879	0.791357
GO:0001569	branching involved in blood vessel morphogenesis	22	-0.60774	-1.72257	0.791357
GO:0045581	negative regulation of T cell differentiation	35	-0.53573	-1.7178	0.791357
GO:0000963	mitochondrial RNA processing	20	-0.61241	-1.71727	0.791357
GO:0042255	ribosome assembly	58	-0.4875	-1.71657	0.791357
GO:0036498	IRE1-mediated unfolded protein response	20	-0.60827	-1.70566	0.791357
GO:0031397	negative regulation of protein ubiquitination	76	-0.45362	-1.68431	0.791357
GO:0007007	inner mitochondrial membrane organization	40	-0.51339	-1.67493	0.791357

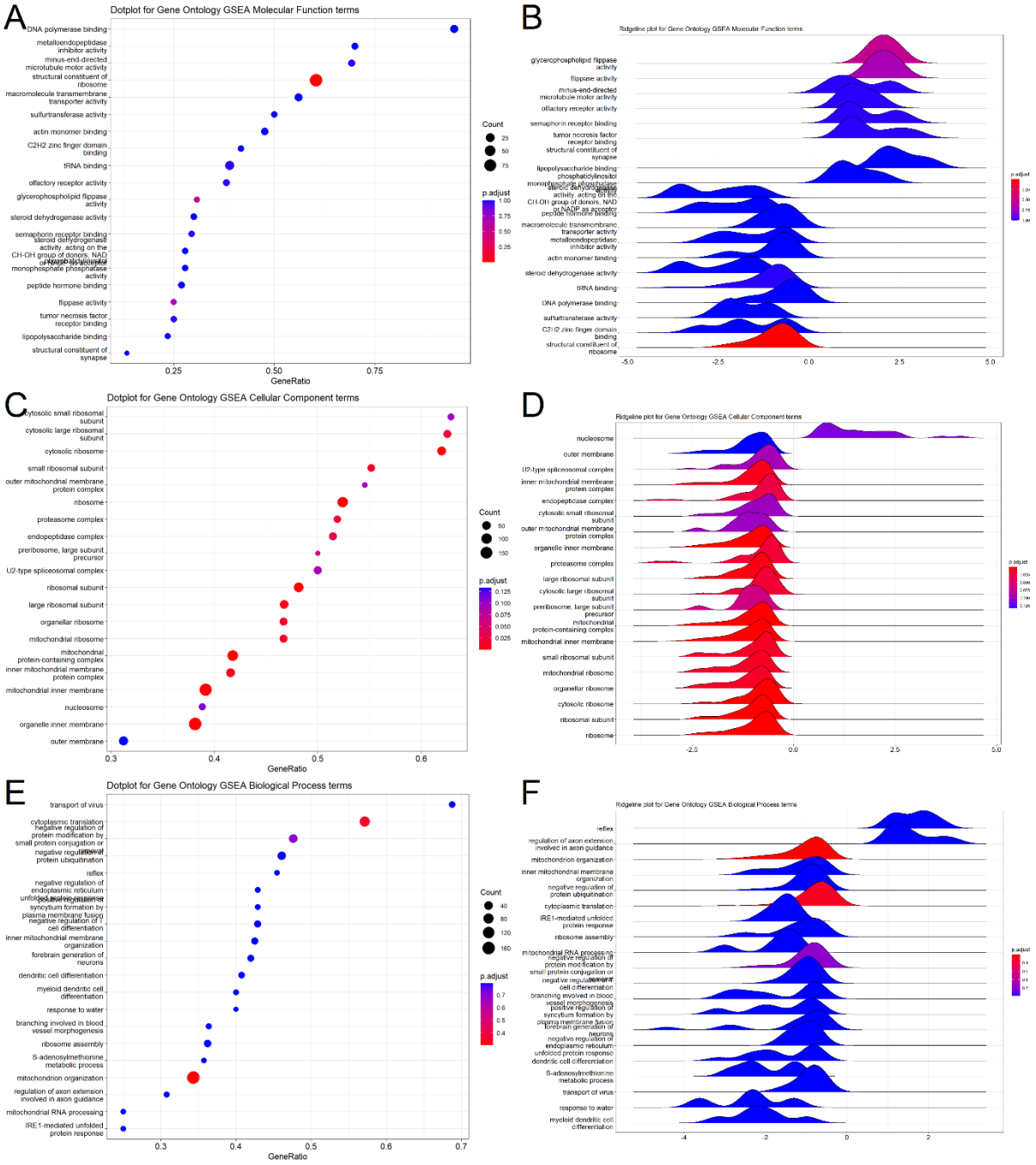


Figure 32. Dot plots and ridgeline plots for the top 20 enriched Gene Ontology MF, CC, and BP terms. **A, C, E.** Dot plot showing the gene ratios for each term. The dot size reflects the gene count for the term and the color reflects the adjusted p-value. **B, D, F.** Ridgeline plots showing the distribution of log2 fold change values for the core enrichment genes for each term. The color reflects the adjusted p-value.

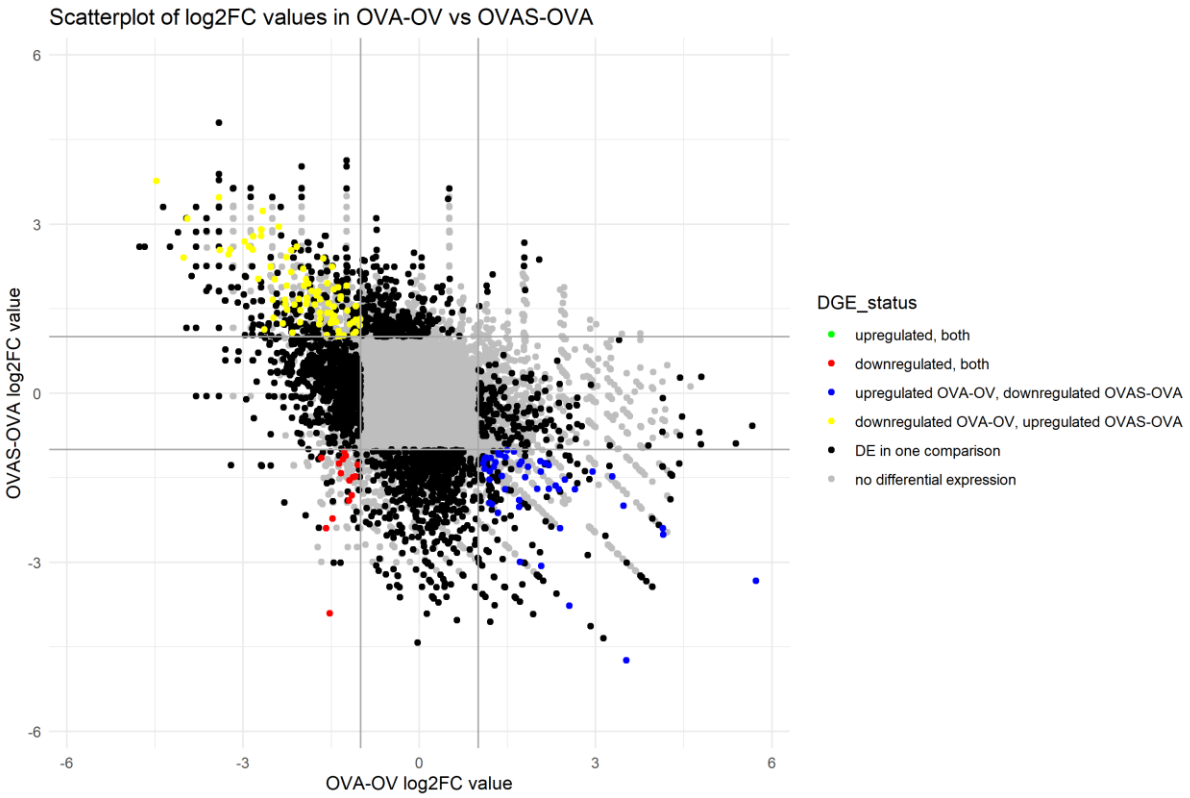


Figure 33. Scatterplot of log2FC values from the MCF-7 OVA *vs.* OV and MCF-7 OVAS *vs.* OVA comparisons.

To visualize the effect of spironolactone treatment in combination with aldosterone, a scatterplot of the log2 fold change values from the OVA *vs.* OV and OVAS *vs.* OVA was graphed. The log2 fold change of all genes in the comparisons were plotted in **Figure 33**. The dots were colored based on a scheme that was dependent on p-values: green, significantly upregulated genes in both comparisons; red, significantly downregulated genes in both comparisons; yellow, significantly downregulated in OVA vs. OV and upregulated in OVAS vs. OVA; blue, significantly upregulated in OVA vs. OV and downregulated in OVAS vs. OVA; black, significantly differentially expressed in one comparison, but not the other; and gray, not differentially expressed. The level of differential expression was set with a log2FC absolute value threshold of 1 and p-value threshold of 0.05. Interestingly, there were a

considerable number of genes that changed pattern of differential expression when the MCF-7 cells were treated with spironolactone in addition to aldosterone. 48 genes changed from upregulation to downregulation, and 83 genes changed from downregulation to upregulation upon combinatorial treatment with aldosterone and spironolactone. Another 1,581 genes changed from significant to non-significant differential expression or vice versa. A tentative list of five genes was chosen based on the level of change in differential expression and the statistical significance of that change. These genes were listed in **Table 27** which included ZBTB16, SGK3, RAB37, PKN3, and MIF-AS1. ZBTB16 and SGK3 expression underwent a significant change from upregulation under aldosterone treatment to downregulation under aldosterone and spironolactone treatment. These genes could be found in the lower right quadrant (blue color dots) in the scatterplot in **Figure *33***. The remaining three genes (RAB37, PKN3, MIF-AS1) were in the top genes that were significantly differentially expressed under aldosterone and spironolactone treatment, that were not differentially expressed when comparing OVA *vs.* OV groups (black dots in **Figure *33***).

These provided some explanation and validation for the differences in the functional analyses results between the two comparisons. It might also mean that ultimately, the MR and aldosterone antagonist spironolactone have profound effects in the MR signaling pathway and that it involves disrupting the interaction between aldosterone and MR. However, it has to be tested whether changing of the value of pseudocount added to the raw count values (e.g., 0.25 to 2) setting the transcript exclusion criteria based on the percentage of samples with 0 values and the mean counts affect this plot and list of differentially expressed genes. Future studies should focus on identification of the most optimal pseudocount value to be added for a series of low to high library sizes while examining the effects of setting different filtering thresholds and changing other variables (e.g., mapping methodologies) on very high and very low fold changes in one or both of the axes in **Figure 33**.

Table 27. Tentative list of genes of interest and the DGEA statistics in two comparisons.

HGNC symbol	**OVA vs. OV log2FC**	**OVA vs. OV p-value**	**OVAS vs. OVA log2FC**	**OVAS vs. OVA p-value**
ZBTB16	5.73	1.17E-04	-3.33	3.38E-05
SGK3	4.15	8.20E-13	-2.51	1.55E-10
RAB37	0.49	7.28E-02	3.44	3.86E-54
PKN3	0.42	1.82E-01	-2.29	7.72E-17
MIF-AS1	0.36	2.61E-01	-2.07	1.51E-15

Chapter 5. Conclusions, Discussion and Future Perspectives

Early efforts in transcriptomics started in the 1990s with technologies like cDNA microarrays and SAGE with Sanger sequencing. The development of next-generation sequencing methods and the first mapped transcriptomes of yeast, mouse and human cells started a movement that would make RNA-sequencing the most common technique for transcriptomic studies (Lister et al., 2008; Mortazavi et al., 2008; Nagalakshmi et al., 2008; Z. Wang et al., 2009). The generation of high-throughput RNA-seq datasets was followed by the development of tools that can extract information from these datasets, including tools that can do mapping, preprocessing, quantifying, normalizing and analyzing count data for differential expression (D. Li, 2019).

In this thesis I have presented the development of a shiny application that can be used for the analysis and visualization of RNA sequencing data. It allows: a) selection of an available dataset or the uploading of a user-generated dataset; b) exploratory analysis of the dataset; c) enables the normalization of the dataset according to multiple methods (CPM, log2CPM, voom, and DESeq2) (Law et al., 2014; Love et al., 2014; Robinson & Oshlack, 2010) and the filtering using multiple counts thresholds; d) differential gene expression analysis using four different methods (DESeq2, edgeR, limma, and NOISeq) (Love et al., 2014; Ritchie et al., 2015; Robinson et al., 2010; Tarazona et al., 2011) and the comparison of the results across these methods; e) further visualization and analyses such as PCA, correlation analysis and heatmaps; and f) integrates downstream functional analyses like the over-representation analysis and gene set enrichment analysis (Boyle et al., 2004; Subramanian et al., 2005).

The features presented here were developed in a way that would fulfill a few aims. The application needs to be easy to use therefore users are only required to provide a raw read count matrix with a row of metadata that follows the rules explained in this thesis. The app also needs to provide adequate methods for RNA-seq data analysis. These include methods for normalization and filtering of data, such that the process of analysis and visualization is streamlined and made easy for the user. In its core the goal of developing shiny apps with underlying statistical R packages is to provide a graphical user interface so users who are less familiar with R programming can still use the functions of several R packages to generate results in the form of tables and plots in a reproducible manner.

The operational flow of the application is also important. Data normalization and filtering is a part of the RNA-seq analysis pipeline that is contested with regard to order of the operation. Firstly, the application applies a filter that removes genes with zero raw read counts on all samples. This is done to increase the power to detect true positives in differential gene expression analysis (Bourgon et al., 2010). The DGEA consists of a large number of hypotheses, where many true null hypotheses produce small p-values, and to avoid false positives, a multiple test adjustment must be performed. This adjustment helps control the rate of false positives, but also reduces the power to detect true positives. The reduction in number of tests by removing zero genes can reduce the impact of the multiple test adjustment and improve detection power (Hackstadt & Hess, 2009; Tritchler et al., 2009). After normalization, another filtering threshold is used to further eliminate lowly expressed genes. This filtering step is also performed so that the stringency of the multiple test adjustment is reduced. This order of operation is followed by established methods like edgeR and NOISeq, specifically in the filterByExpr() in the {edgeR} package and the data quality functions such as biodetection and sensitivity in the {NOISeq} package (Chen et al., 2022; Robinson et al., 2010; Tarazona et al., 2015). Filtering is applied after normalization to ensure that the filtering threshold is

applied to all samples at the same level. However, it would be important not to lose genes whose expressions are absent in one condition but gets slightly yet significantly increased in others. These significantly modulated lowly expressed genes could still be biologically informative. Therefore, future capabilities of the RNAseq app presented here should include options to allow users to change their threshold selections and test the effects of these choices on differentially expressed gene lists and the value of fold changes and p-values obtained from the app.

This thesis also aims to contribute an original feature to the long list of applications that are designed for RNA-seq data analysis. When surveying the literature, one feature that was not found in any of the published online applications was the ability to compare the results of statistical R/Bioconductor packages that are designed for RNA-seq data analysis. To this end, the application in this thesis contains the most used and diverse methods for DGE analysis and a comparison section where results of different methods can be analyzed and intersected to produce gene sets that are discovered as DEGs in agreement between different statistical tests. edgeR and DESeq2 are commonly used in many published applications, limma provides another empirical Bayesian approach to discovering differentially expressed genes that is dependent on linear models, and NOISeq implements a noise distribution and another approach to testing for differential expression. These packages were also selected for the fact that they implement methods that can overcome the issue of having single replicates in a dataset. This was done so that the application has a broader selection of datasets that can be analyzed using predetermined dispersion statistics (edgeR, limma), or read count simulation (NOISeq).

The results of the application development part can be summarized as follows:

- An application for RNA-seq data analysis was developed and hosted locally for lab members to use and is planned to be hosted publicly to allow other users to use it.
- The application can perform data exploratory analysis using library size plots, multidimensional scaling of samples, other data quality plots like counts distribution, biotype detection, and RNA composition analysis.
- The application offers the user the ability to normalize and filter data before differential gene expression analysis.
- Differential gene expression analysis is offered using four statistical methods and the results can be compared to ensure the generation of gene sets that are discovered commonly across multiple tests.
- The application can perform PCA, correlation analysis and functional analysis based on gene selection from DGE analysis to find out more about the patterns of differential analysis in context of genes and pathways or ontologies.
- The application is overall easy to use, can generate reproducible results and is interactive, meaning that the user has many options available for customizing results and plots.

One of the main future perspectives of this application development is the goal of publicly hosting a version of this application that can be used by any researcher. While it is designed to work perfectly with a dataset generated in-house, efforts were made to make it as flexible as possible, meaning that any user that has a dataset that complies with the input file requirements, can theoretically and in practice upload the file and use every feature that is included in the application. Further improvements can be made to the application, to make it more flexible and more capable of analyzing a wider range of datasets. Borrowing from the already published applications, more normalization and DGEA methods can be added to the tool. Specifically, the batch normalization done using the Combat (REF) and Combat-seq (Zhang et al., 2020)

methods like in the DEBrowser tool must be added, as batch effects introduce variability to data which needs to be minimized or eliminated (Kucukural et al., 2019). Another feature that could be added is the ability to directly connect to Gene Expression Omnibus (GEO) and download raw counts from GEO datasets upon request from the user (feature available in the GREIN app) (Mahi et al., 2019). While the GREIN app, focused on RNA-seq data, already has a substantial number of these datasets analyzed and ready to be explored (>28,000 datasets and constantly adding user-requested datasets), the ability to add data from public datasets directly to our application would mean that a whole set of new features that are not available in GREIN (e.g., data normalization, and functional analyses such as over-representation analysis and gene set enrichment analysis) can be used to analyze these GEO datasets and visualize results as needed. Another limitation of the RNAseq app presented here could be the number of samples that can be uploaded without having memory or processing problems. The RNAseq app has not been tested with more than18 samples uploaded, and future studies are planned to test the ability of the app and the server to cope with large datasets. Moreover, the number of experimental groups at a time is two, which can be increased to more than two allowing the use of limma or other nonparametric tests to analyze more complicated experimental designs (Law et al., 2020). In addition, use of covariates can be added into the DGE analysis that can be a novel addition when compared with other apps (Choudhary et al., 2021; Gadepalli et al., 2019; Khomtchouk et al., 2016; Marini et al., 2020; Monier et al., 2019; Su et al., 2019).

In the present thesis the RNAseq app has been used to analyze a novel dataset from our lab that contained multiple groups exposed to different treatments in different cell lines. The MR signaling pathway consists of the mineralocorticoid receptor (MR), a transcription factor that may be dependent on binding of a ligand to regulate gene expression. Its two physiological ligands are aldosterone and cortisol (Griekspoor et al., 2007). Upon ligand binding, the MR

changes conformation, translocates into the nucleus, and starts a sequence of events that ultimately leads to the activation of the basal transcription apparatus and gene expression (Yang & Fuller, 2012). MR was found to have very low to no expression in breast cancer cell lines, however some studies regarding MR signaling in breast cancer have shown that the expression of MR coupled with activation of other receptors can have an effect on cancer growth (Konu & Targen, 2019). Co-activation of MR and retinoic acid receptor (RAR), or MR with glucocorticoid receptor (GR) and progesterone receptor (PR) is correlated with inhibition of tumor growth (Doan et al., 2020; Leo et al., 2004; Pavlides et al., 2009).

Spironolactone, which is an MR antagonist, was found to have a passive version of antagonism, which involves changes in the MR that prevent dimerization and binding of DNA, and promotes receptor degradation (Fuller et al., 2019). To understand the resulting molecular effects of this antagonistic relation between spironolactone and MR via activation with aldosterone, an experiment was conducted using MCF-7 breast cancer cells transfecting them to overexpress MR and treating with aldosterone or a combination of aldosterone and spironolactone (Targen, 2020). My analyses performed with these RNA-seq data from this experiment showed the data to be of good quality using the FASTQC and MultiQC tools, and from the exploratory analysis it was concluded that normalization for library size was necessary. Filtering of lowly expressed genes was done to lower the multiple test burden (Bourgon et al., 2010; Hackstadt & Hess, 2009). The dataset was analyzed in this thesis and the differential expression patterns of cells treated with aldosterone were compared to patters of cells treated with spironolactone. However, future studies should test what the effects of changing preprocessing and read alignments methods, and/or filtering thresholds would be on the differentially expressed genes and their log fold changes and p-values across the four methods used in this thesis.

Differential gene expression analysis of OVA *vs.* OV groups resulted in discovery of hundreds of genes using the four DGEA methods in the application. Two methods (edgeR and limma) consistently discovered more when DEGs compared to the other two methods (DESeq2 and NOISeq). To narrow down the significant DEGs, a minimum of three methods were needed to discover 384 DEGs and eliminate other method specific genes discovered by edgeR and limma. The requirement of three methods was also set in order to ensure the DEGs were found to be significantly differentially expressed by different approaches as described in the thesis introduction. Functional analysis of these 384 DEGs showed that the top over-represented Gene Ontology molecular function terms involve functions related to DNA binding and transcription factor activity; biological processes that are over-represented include cell death positive regulation and cell migration; and the over-represented cellular component terms describe the ribosomal subunits, focal adhesion, and components related to the cytoskeleton. Future studies might be done by separating upregulated and downregulated genes before functional analyses. Overall, the treatment of MCF-7 cells that overexpress the mineralocorticoid receptor causes differential expression of genes that are involved in transcription regulation, cell death, and cell migration.

Differential gene expression analysis of OVAS *vs.* OVA groups resulted in a list of 462 DEGs discovered by at least three DGEA methods. Over-representation analysis and gene set enrichment analysis showed that these genes significantly were enriched in pathways related to ribosomal subunits and cellular components related to the mitochondria. These analyses were inconclusive about the biological processes and molecular functions that were changed upon spironolactone treatment. This is likely a result of the sample size of the dataset, meaning that the DGEA methods were not able to extract a lot of information from the read counts that could lead to a better understanding of the spironolactone effect on the MR signaling pathway

in breast cancer cells. In the future, changing the p-value threshold to have more genes as the number of genes could be the limiting factor in functional analysis.

A simple comparison of log2 fold change values from the OVA *vs.* OV against OVAS *vs.* OVA showed that there was a notable reversal in differential gene expression patterns in genes as it was observed with a negative correlation shown in **Figure *33*** that 48 genes changed from upregulation to downregulation, whereas 83 genes changed from downregulation to upregulation upon combinatorial treatment with aldosterone and spironolactone when used together. Another type of change observed in **Figure 33** is the change from non-significant differential expression to significant upregulation/downregulation (black dots). Future studies should also examine the genes exhibiting very high log fold change values in one or both of the axes (between treatment comparisons) and how their values change when DGE method used changes.

Given the thresholds used in this thesis, out of this gene set, significantly reverted the effects of aldosterone with spironolactone and consisting of at least 131 genes, a list of five genes of interest was chosen and reported as a tentative list that can be studied further. Of these five genes, ZTBT16 was previously studied in the laboratory and its expression was found to be increased upon aldosterone treatment, and reduced after spironolactone was added (Targen, 2020). These genes also emerged using the RNA-seq tool in this thesis (**Table 27**). SGK3 is another gene that follows this pattern of aldosterone-induced upregulation and spironolactone-induced downregulation. Both these genes have been implicated in breast cancer growth and metastasis. ZBTB16 has been shown to inhibit breast cancer proliferation and metastasis through the upregulation of ZBTB28 and the antagonism of BCL6/ZBTB27 (He et al., 2020). SGK3, however, has been shown to promote breast cancer survival under estrogen induction (Y. Wang et al., 2011, p. 3). The findings of this thesis showed that Aldosterone induced significant transcriptomics changes which were reverted back by Spironolactone; and could

have a mixed response with respect to cell survival. This can be due to cell type or dose of the drugs used and warrants further research, e.g., comparisons with ZR75-1 cell line results followed by qRT-PCR confirmations.

The other three genes that were selected were in the top 10 genes that were significantly differentially expressed genes in the OVAS *vs.* OVA comparison but did not show any differential expression in OVA *vs.* OV. RAB37, a low molecular mass GTPase that is a member of the RAS oncogene family (NCBI, 2022), is one these genes; it is significantly upregulated in OVAS samples suggesting that this induction could be driven by spironolactone only. The importance of this gene in RAB37 has been studied and it was found through TCGA patient data that patients with breast invasive carcinoma might have homozygous deletion of RAB37 (Tsai et al., 2014). PKN3 (protein kinase 3) and MIF-AS1 (macrophage migration inhibitory factor anti-sense RNA 1) are both implicated in promoting tumor malignant growth and migration (Ding et al., 2019; Gemperle et al., 2019). Downregulation of these genes after spironolactone treatment can be biologically significant, and therefore validation experiments using breast cancer cells and migration, proliferation, and kinase assays can be helpful in checking the effect of spironolactone on breast cancer.

Chapter 6. References

Almende B. V. and Contributors, Benoit Thieurmel, & Titouan Robert. (2021). *visNetwork: Network Visualization using "vis.js" Library* (2.1.0). https://CRAN.R-project.org/package=visNetwork

Anders, S., & Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biology*, *11*(10), R106. https://doi.org/10.1186/gb-2010-11-10-r106

Anders, S., Pyl, P. T., & Huber, W. (2015). HTSeq—A Python framework to work with high-throughput sequencing data. *Bioinformatics*, *31*(2), 166–169. https://doi.org/10.1093/bioinformatics/btu638

Andrews, S. (2015). *FastQC: A Quality Control Tool for High Throughput Sequence Data [Online]*. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., & Sherlock, G. (2000). Gene Ontology: Tool for the unification of biology. *Nature Genetics*, *25*(1), 25–29. https://doi.org/10.1038/75556

Attali [aut, D., & cre. (2022). *shinybrowser: Find Out Information About a User's Web Browser in "Shiny"* (1.0.0). https://CRAN.R-project.org/package=shinybrowser

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Benjamini, Y., & Speed, T. P. (2012). Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Research*, *40*(10), e72. https://doi.org/10.1093/nar/gks001

Biggar, R. J., Andersen, E. W., Wohlfahrt, J., & Melbye, M. (2013). Spironolactone use and the risk of breast and gynecologic cancers. *Cancer Epidemiology*, *37*(6), 870–875. https://doi.org/10.1016/j.canep.2013.10.004

Bourgon, R., Gentleman, R., & Huber, W. (2010). Independent filtering increases detection power for high-throughput experiments. *Proceedings of the National Academy of Sciences*, *107*(21), 9546–9551. https://doi.org/10.1073/pnas.0914005107

Boyle, E. I., Weng, S., Gollub, J., Jin, H., Botstein, D., Cherry, J. M., & Sherlock, G. (2004). GO::TermFinder—Open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. *Bioinformatics*, *20*(18), 3710–3715. https://doi.org/10.1093/bioinformatics/bth456

Bullard, J. H., Purdom, E., Hansen, K. D., & Dudoit, S. (2010). Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics*, *11*(1), 94. https://doi.org/10.1186/1471-2105-11-94

Chang, W., Cheng, J., Allaire, J. J., Sievert, C., Schloerke, B., Xie, Y., Allen, J., McPherson, J., Dipert, A., Borges, B., RStudio, library), jQuery F. (jQuery library and jQuery U., inst/www/shared/jquery-AUTHORS.txt), jQuery contributors (jQuery library; authors listed in, inst/www/shared/jqueryui/AUTHORS.txt), jQuery U. contributors (jQuery U. library; authors listed in, library), M. O. (Bootstrap, library), J. T. (Bootstrap, library), B. contributors (Bootstrap, Twitter, library), I. (Bootstrap, … R), R. C. T. (tar implementation from. (2022). *shiny: Web Application Framework for R* (1.7.2). https://CRAN.R-project.org/package=shiny

Chen, Y., Lun, A. T. L., & Smyth, G. K. (2016). *From reads to genes to pathways: Differential expression analysis of RNA-Seq experiments using Rsubread and the edgeR quasi-likelihood pipeline* (5:1438). F1000Research. https://doi.org/10.12688/f1000research.8987.2

Chen, Y., McCarthy, D., Ritchie, M., Robinson, M., & Smyth, G. (2022). *EdgeR: differential analysis of sequence read count data User's Guide*. 119.

Choudhary, K. S., Caldwell, A. B., & Subramaniam, S. (2021). *DEGenR: An R Shiny app for differential gene expression and enrichment analysis*. Zenodo. https://doi.org/10.5281/zenodo.4815134

Comşa, Ş., Cîmpean, A. M., & Raica, M. (2015). The Story of MCF-7 Breast Cancer Cell Line: 40 years of Experience in Research. *Anticancer Research*, *35*(6), 3147–3154.

Ding, J., Wu, W., Yang, J., & Wu, M. (2019). Long non-coding RNA MIF-AS1 promotes breast cancer cell proliferation, migration and EMT process through regulating miR-1249-3p/HOXB8 axis. *Pathology - Research and Practice*, *215*(7), 152376. https://doi.org/10.1016/j.prp.2019.03.005

Doan, T. B., Cheung, V., Clyne, C. D., Hilton, H. N., Eriksson, N., Young, M. J., Funder, J. W., Muscat, G. E. O., Fuller, P. J., Clarke, C. L., & Graham, J. D. (2020). A tumour suppressive relationship between mineralocorticoid and retinoic acid receptors activates a transcriptional program consistent with a reverse Warburg effect in breast cancer. *Breast Cancer Research*, *22*(1), 122. https://doi.org/10.1186/s13058-020-01355-x

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., & Gingeras, T. R. (2013). STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics*, *29*(1), 15–21. https://doi.org/10.1093/bioinformatics/bts635

Durinck, S., Spellman, P. T., Birney, E., & Huber, W. (2009). Mapping Identifiers for the Integration of Genomic Datasets with the R/Bioconductor package biomaRt. *Nature Protocols*, *4*(8), 1184–1191. https://doi.org/10.1038/nprot.2009.97

Evans, C., Hardin, J., & Stoebel, D. M. (2018). Selecting between-sample RNA-Seq normalization methods from the perspective of their assumptions. *Briefings in Bioinformatics*, *19*(5), 776–792. https://doi.org/10.1093/bib/bbx008

Ewels, P., Magnusson, M., Lundin, S., & Käller, M. (2016). MultiQC: Summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, *32*(19), 3047–3048. https://doi.org/10.1093/bioinformatics/btw354

Fuller, P. J., Yang, J., & Young, M. J. (2019). Chapter Three—Mechanisms of Mineralocorticoid Receptor Signaling. In G. Litwack (Ed.), *Vitamins and Hormones* (Vol. 109, pp. 37–68). Academic Press. https://doi.org/10.1016/bs.vh.2018.09.004

Gadepalli, V. S., Ozer, H. G., Yilmaz, A. S., Pietrzak, M., & Webb, A. (2019). BISR-RNAseq: An efficient and scalable RNAseq analysis workflow with interactive report generation. *BMC Bioinformatics*, *20*(24), 670. https://doi.org/10.1186/s12859-019-3251-1

Galili, T., O'Callaghan, A., Sidi, J., & Sievert, C. (2018). heatmaply: An R package for creating interactive cluster heatmaps for online publishing. *Bioinformatics (Oxford, England)*, *34*(9), 1600–1602. https://doi.org/10.1093/bioinformatics/btx657

Gautier, L., Cope, L., Bolstad, B. M., & Irizarry, R. A. (2004). affy—Analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics*, *20*(3), 307–315. https://doi.org/10.1093/bioinformatics/btg405

Gemperle, J., Dibus, M., Koudelková, L., Rosel, D., & Brábek, J. (2019). The interaction of p130Cas with PKN3 promotes malignant growth. *Molecular Oncology*, *13*(2), 264–289. https://doi.org/10.1002/1878-0261.12401

Griekspoor, A., Zwart, W., Neefjes, J., & Michalides, R. (2007). Visualizing the action of steroid hormone receptors in living cells. *Nuclear Receptor Signaling*, *5*, e003. https://doi.org/10.1621/nrs.05003

Guo, K., & McGregor, B. (2022). *VennDetail: A package for visualization and extract details* [R]. https://github.com/guokai8/VennDetail (Original work published 2018)

Hackstadt, A. J., & Hess, A. M. (2009). Filtering for increased power for microarray data analysis. *BMC Bioinformatics*, *10*(1), 11. https://doi.org/10.1186/1471-2105-10-11

Hardcastle, T. J., & Kelly, K. A. (2010). baySeq: Empirical Bayesian methods for identifying differential expression in sequence count data. *BMC Bioinformatics*, *11*(1), 422. https://doi.org/10.1186/1471-2105-11-422

He, J., Wu, M., Xiong, L., Gong, Y., Yu, R., Peng, W., Li, L., Li, L., Tian, S., Wang, Y., Tao, Q., & Xiang, T. (2020). BTB/POZ zinc finger protein ZBTB16 inhibits breast cancer proliferation and metastasis through upregulating ZBTB28 and antagonizing BCL6/ZBTB27. *Clinical Epigenetics*, *12*(1), 82. https://doi.org/10.1186/s13148-020-00867-9

Holbeck, S., Chang, J., Best, A. M., Bookout, A. L., Mangelsdorf, D. J., & Martinez, E. D. (2010). Expression Profiling of Nuclear Receptors in the NCI60 Cancer Cell Panel Reveals Receptor-Drug and Receptor-Gene Interactions. *Molecular Endocrinology*, *24*(6), 1287–1296. https://doi.org/10.1210/me.2010-0040

Husi, H. (Ed.). (2019). *Computational Biology*. Codon Publications. http://www.ncbi.nlm.nih.gov/books/NBK550339/

Kanehisa, M., & Goto, S. (2000). KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, *28*(1), 27–30. https://doi.org/10.1093/nar/28.1.27

Kapp, A. V., Jeffrey, S. S., Langerød, A., Børresen-Dale, A.-L., Han, W., Noh, D.-Y., Bukholm, I. R., Nicolau, M., Brown, P. O., & Tibshirani, R. (2006). Discovery and

validation of breast cancer subtypes. *BMC Genomics*, *7*(1), 231. https://doi.org/10.1186/1471-2164-7-231

Kaya, T. (2018). *Co-expression pairs and modules (CoEX-PM): A shiny application and an example case study on chromogranins* [Thesis, Bilkent University]. http://repository.bilkent.edu.tr/handle/11693/48074

Khomtchouk, B. B., Hennessy, J. R., & Wahlestedt, C. (2016). MicroScope: ChIP-seq and RNA-seq software analysis suite for gene expression heatmaps. *BMC Bioinformatics*, *17*(1), 390. https://doi.org/10.1186/s12859-016-1260-x

Konu, O., & Targen, S. (2019). Investigating the Role of Mineralocorticoid Receptor Signaling in Cancer Biology in the Genomic Era. In *Aldosterone-Mineralocorticoid Receptor—Cell Biology to Translational Medicine*. IntechOpen. https://doi.org/10.5772/intechopen.87233

Kucukural, A., Yukselen, O., Ozata, D. M., Moore, M. J., & Garber, M. (2019). DEBrowser: Interactive differential expression analysis and visualization tool for count data. *BMC Genomics*, *20*(1), 6. https://doi.org/10.1186/s12864-018-5362-x

Lau, J. W., Lehnert, E., Sethi, A., Malhotra, R., Kaushik, G., Onder, Z., Groves-Kirkby, N., Mihajlovic, A., DiGiovanna, J., Srdic, M., Bajcic, D., Radenkovic, J., Mladenovic, V., Krstanovic, D., Arsenijevic, V., Klisic, D., Mitrovic, M., Bogicevic, I., Kural, D., … for The Seven Bridges CGC Team. (2017). The Cancer Genomics Cloud: Collaborative, Reproducible, and Democratized—A New Paradigm in Large-Scale Computational Research. *Cancer Research*, *77*(21), e3–e6. https://doi.org/10.1158/0008-5472.CAN-17-0387

Law, C. W., Chen, Y., Shi, W., & Smyth, G. K. (2014). voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biology*, *15*(2), R29. https://doi.org/10.1186/gb-2014-15-2-r29

Law, C. W., Zeglinski, K., Dong, X., Alhamdoosh, M., Smyth, G. K., & Ritchie, M. E. (2020). A guide to creating design matrices for gene expression experiments. *F1000Research*, *9*, 1444. https://doi.org/10.12688/f1000research.27893.1

Leng, N., Dawson, J. A., Thomson, J. A., Ruotti, V., Rissman, A. I., Smits, B. M. G., Haag, J. D., Gould, M. N., Stewart, R. M., & Kendziorski, C. (2013). EBSeq: An empirical Bayes hierarchical model for inference in RNA-seq experiments. *Bioinformatics (Oxford, England)*, *29*(8), 1035–1043. https://doi.org/10.1093/bioinformatics/btt087

Leo, J. C. L., Guo, C., Woon, C. T., Aw, S. E., & Lin, V. C. L. (2004). Glucocorticoid and Mineralocorticoid Cross-Talk with Progesterone Receptor to Induce Focal Adhesion and Growth Inhibition in Breast Cancer Cells. *Endocrinology*, *145*(3), 1314–1321. https://doi.org/10.1210/en.2003-0732

Li, D. (2019). Statistical Methods for RNA Sequencing Data Analysis. In H. Husi (Ed.), *Computational Biology*. Codon Publications. http://www.ncbi.nlm.nih.gov/books/NBK550334/

Li, J., & Tibshirani, R. (2013). Finding consistent patterns: A nonparametric approach for identifying differential expression in RNA-Seq data. *Statistical Methods in Medical Research*, *22*(5), 519–536. https://doi.org/10.1177/0962280211428386

Li, J., Witten, D. M., Johnstone, I. M., & Tibshirani, R. (2012). Normalization, testing, and false discovery rate estimation for RNA-sequencing data. *Biostatistics*, *13*(3), 523–538. https://doi.org/10.1093/biostatistics/kxr031

Lister, R., O'Malley, R. C., Tonti-Filippini, J., Gregory, B. D., Berry, C. C., Millar, A. H., & Ecker, J. R. (2008). Highly Integrated Single-Base Resolution Maps of the Epigenome in Arabidopsis. *Cell*, *133*(3), 523–536. https://doi.org/10.1016/j.cell.2008.03.029

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, *15*(12), 550. https://doi.org/10.1186/s13059-014-0550-8

Mahi, N. A., Najafabadi, M. F., Pilarczyk, M., Kouril, M., & Medvedovic, M. (2019). GREIN: An Interactive Web Platform for Re-analyzing GEO RNA-seq Data. *Scientific Reports*, *9*(1), 7580. https://doi.org/10.1038/s41598-019-43935-8

Marini, F., Linke, J., & Binder, H. (2020). ideal: An R/Bioconductor package for interactive differential expression analysis. *BMC Bioinformatics*, *21*(1), 565. https://doi.org/10.1186/s12859-020-03819-5

Messaoudi, S., & Jaisser, F. (2011). Aldosterone and the mineralocorticoid receptor. *European Heart Journal Supplements*, *13*(suppl_B), B4–B9. https://doi.org/10.1093/eurheartj/sur004

Monier, B., McDermaid, A., Wang, C., Zhao, J., Miller, A., Fennell, A., & Ma, Q. (2019). IRIS-EDA: An integrated RNA-Seq interpretation system for gene expression data analysis. *PLOS Computational Biology*, *15*(2), e1006792. https://doi.org/10.1371/journal.pcbi.1006792

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., & Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*, *5*(7), 621–628. https://doi.org/10.1038/nmeth.1226

Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M., & Snyder, M. (2008). The Transcriptional Landscape of the Yeast Genome Defined by RNA Sequencing. *Science*, *320*(5881), 1344–1349. https://doi.org/10.1126/science.1158441

NCBI. (2022). *RAB37 RAB37, member RAS oncogene family [Homo sapiens (human)]—Gene—NCBI*. https://www.ncbi.nlm.nih.gov/gene/326624

Neve, R. M., Chin, K., Fridlyand, J., Yeh, J., Baehner, F. L., Fevr, T., Clark, L., Bayani, N., Coppe, J.-P., Tong, F., Speed, T., Spellman, P. T., DeVries, S., Lapuk, A., Wang, N. J., Kuo, W.-L., Stilwell, J. L., Pinkel, D., Albertson, D. G., … Gray, J. W. (2006). A collection of breast cancer cell lines for the study of functionally distinct cancer subtypes. *Cancer Cell*, *10*(6), 515–527. https://doi.org/10.1016/j.ccr.2006.10.008

Nie, H., Li, J., Yang, X., Cao, Q., Feng, M., Xue, F., Wei, L., Qin, W., Gu, J., Xia, Q., & Zhang, Z. (2015). Mineralocorticoid receptor suppresses cancer progression and the Warburg effect by modulating the miR-338-3p-PKLR axis in hepatocellular carcinoma. *Hepatology (Baltimore, Md.)*, *62*(4), 1145–1159. https://doi.org/10.1002/hep.27940

Oshlack, A., & Wakefield, M. J. (2009). Transcript length bias in RNA-seq data confounds systems biology. *Biology Direct*, *4*(1), 14. https://doi.org/10.1186/1745-6150-4-14

Ozhan, A., Tombaz, M., & Konu, O. (2021). SmulTCan: A Shiny application for multivariable survival analysis of TCGA data with gene sets. *Computers in Biology and Medicine*, *137*, 104793. https://doi.org/10.1016/j.compbiomed.2021.104793

Pavlides, S., Whitaker-Menezes, D., Castello-Cros, R., Flomenberg, N., Witkiewicz, A. K., Frank, P. G., Casimiro, M. C., Wang, C., Fortina, P., Addya, S., Pestell, R. G., Martinez-Outschoorn, U. E., Sotgia, F., & Lisanti, M. P. (2009). The reverse Warburg effect: Aerobic glycolysis in cancer associated fibroblasts and the tumor stroma. *Cell Cycle*, *8*(23), 3984–4001. https://doi.org/10.4161/cc.8.23.10238

Perou, C. M., Sørlie, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S., Rees, C. A., Pollack, J. R., Ross, D. T., Johnsen, H., Akslen, L. A., Fluge, Ø., Pergamenschikov, A., Williams, C., Zhu, S. X., Lønning, P. E., Børresen-Dale, A.-L., Brown, P. O., & Botstein, D. (2000). Molecular portraits of human breast tumours. *Nature*, *406*(6797), 747–752. https://doi.org/10.1038/35021093

Perrier, V., Meyer, F., Granjon, D., updateMultiInput), I. F. (Methods for mutating vertical tabs &, function), W. D. (numericRangeInput, methods), S. M. (autoNumeric, & LICENSE.md), J. and C. libraries authors (All authors are listed in. (2022). *shinyWidgets: Custom Inputs Widgets for Shiny* (0.7.2). https://CRAN.R-project.org/package=shinyWidgets

Puchta, M., Boczkowska, M., & Groszyk, J. (2020). Low RIN Value for RNA-Seq Library Construction from Long-Term Stored Seeds: A Case Study of Barley Seeds. *Genes*, *11*(10), 1190. https://doi.org/10.3390/genes11101190

R Core Team. (2020). *R: A language and environment for statistical computing*. https://www.r-project.org/

Rapaport, F., Khanin, R., Liang, Y., Pirun, M., Krek, A., Zumbo, P., Mason, C. E., Socci, N. D., & Betel, D. (2013). Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. *Genome Biology*, *14*(9), 3158. https://doi.org/10.1186/gb-2013-14-9-r95

Reyes, A. L. P., Silva, T. C., Coetzee, S. G., Plummer, J. T., Davis, B. D., Chen, S., Hazelett, D. J., Lawrenson, K., Berman, B. P., Gayther, S. A., & Jones, M. R. (2019). GENAVi: A shiny web application for gene expression normalization, analysis and visualization. *BMC Genomics*, *20*(1), 745. https://doi.org/10.1186/s12864-019-6073-7

Risso, D., Schwartz, K., Sherlock, G., & Dudoit, S. (2011). GC-Content Normalization for RNA-Seq Data. *BMC Bioinformatics*, *12*(1), 480. https://doi.org/10.1186/1471-2105-12-480

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., & Smyth, G. K. (2015). Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, *43*(7), e47. https://doi.org/10.1093/nar/gkv007

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, *26*(1), 139–140. https://doi.org/10.1093/bioinformatics/btp616

Robinson, M. D., & Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology*, *11*(3), R25. https://doi.org/10.1186/gb-2010-11-3-r25

RStudio Team. (2022). *RStudio: Integrated Development Environment for R*. https://www.rstudio.com/

Schroeder, A., Mueller, O., Stocker, S., Salowsky, R., Leiber, M., Gassmann, M., Lightfoot, S., Menzel, W., Granzow, M., & Ragg, T. (2006). The RIN: An RNA integrity number for assigning integrity values to RNA measurements. *BMC Molecular Biology*, *7*(1), 3. https://doi.org/10.1186/1471-2199-7-3

Shehwana, H. (2017). *Development of novel tools for cancer diagnosis, prognosis and treatment using intra- or inter-species transcriptome metaanalysis* [Thesis, Bilkent University]. http://repository.bilkent.edu.tr/handle/11693/33632

Sica, D. A. (2015). Mineralocorticoid Receptor Antagonists for Treatment of Hypertension and Heart Failure. *Methodist DeBakey Cardiovascular Journal*, *11*(4), 235–239. https://doi.org/10.14797/mdcj-11-4-235

Sievert, C., Parmer, C., Hocking, T., Chamberlain, S., Ram, K., Corvellec, M., Despouy, P., Brüggemann, S., & Inc, P. T. (2021). *plotly: Create Interactive Web Graphics via "plotly.js"* (4.10.0). https://CRAN.R-project.org/package=plotly

Simpson, S. A., Tait, J. F., Wettstein, A., Neher, R., Von Euw, J., & Reichstein, T. (1953). [Isolation from the adrenals of a new crystalline hormone with especially high effectiveness on mineral metabolism]. *Experientia*, *9*(9), 333–335. https://doi.org/10.1007/BF02155834

Sørlie, T., Perou, C. M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., Hastie, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S., Thorsen, T., Quist, H., Matese, J. C., Brown, P. O., Botstein, D., Lønning, P. E., & Børresen-Dale, A.-L. (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proceedings of the National Academy of Sciences*, *98*(19), 10869–10874. https://doi.org/10.1073/pnas.191367098

Sørlie, T., Tibshirani, R., Parker, J., Hastie, T., Marron, J. S., Nobel, A., Deng, S., Johnsen, H., Pesich, R., Geisler, S., Demeter, J., Perou, C. M., Lønning, P. E., Brown, P. O., Børresen-Dale, A.-L., & Botstein, D. (2003). Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proceedings of the National Academy of Sciences*, *100*(14), 8418–8423. https://doi.org/10.1073/pnas.0932692100

Soule, H. D., Vazquez, J., Long, A., Albert, S., & Brennan, M. (1973). A Human Cell Line From a Pleural Effusion Derived From a Breast Carcinoma2. *JNCI: Journal of the National Cancer Institute*, *51*(5), 1409–1416. https://doi.org/10.1093/jnci/51.5.1409

Su, W., Sun, J., Shimizu, K., & Kadota, K. (2019). TCC-GUI: A Shiny-based application for differential expression analysis of RNA-Seq count data. *BMC Research Notes*, *12*(1), 133. https://doi.org/10.1186/s13104-019-4179-2

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., & Mesirov, J. P. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, *102*(43), 15545–15550. https://doi.org/10.1073/pnas.0506580102

Sun, J., Nishiyama, T., Shimizu, K., & Kadota, K. (2013). TCC: An R package for comparing tag count data with robust normalization strategies. *BMC Bioinformatics*, *14*(1), 219. https://doi.org/10.1186/1471-2105-14-219

Tarazona, S., Furió-Tarí, P., Turrà, D., Pietro, A. D., Nueda, M. J., Ferrer, A., & Conesa, A. (2015). Data quality aware analysis of differential expression in RNA-seq with NOISeq R/Bioc package. *Nucleic Acids Research*, *43*(21), e140. https://doi.org/10.1093/nar/gkv711

Tarazona, S., García-Alcalde, F., Dopazo, J., Ferrer, A., & Conesa, A. (2011). Differential expression in RNA-seq: A matter of depth. *Genome Research*, *21*(12), 2213–2223. https://doi.org/10.1101/gr.124321.111

Targen, S. (2020). *Development of a specialized zebrafish xenotransplantation database and establishment of ALU-based tumor DNA quantification method in zebrafish: Focus on models of overexpression and microenvironment* [Thesis, Bilkent University]. http://repository.bilkent.edu.tr/handle/11693/54266

Targen, S., Kaya, T., Avci, M. E., Gunes, D., Keskus, A. G., & Konu, O. (2020). ZenoFishDb v1.1: A Database for Xenotransplantation Studies in Zebrafish. *Zebrafish*, *17*(5), 305–318. https://doi.org/10.1089/zeb.2020.1869

The Gene Ontology Consortium. (2021). The Gene Ontology resource: Enriching a GOld mine. *Nucleic Acids Research*, *49*(D1), D325–D334. https://doi.org/10.1093/nar/gkaa1113

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., Pimentel, H., Salzberg, S. L., Rinn, J. L., & Pachter, L. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature Protocols*, *7*(3), 562–578. https://doi.org/10.1038/nprot.2012.016

Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., Salzberg, S. L., Wold, B. J., & Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching

during cell differentiation. *Nature Biotechnology*, *28*(5), 511–515. https://doi.org/10.1038/nbt.1621

Trapp, T., & Holsboer, F. (1995). Ligand-Induced Conformational Changes in the Mineralocorticoid Receptor Analyzed by Protease Mapping. *Biochemical and Biophysical Research Communications*, *215*(1), 286–291. https://doi.org/10.1006/bbrc.1995.2464

Tritchler, D., Parkhomenko, E., & Beyene, J. (2009). Filtering genes for cluster and network analysis. *BMC Bioinformatics*, *10*, 193. https://doi.org/10.1186/1471-2105-10-193

Tsai, C.-H., Cheng, H.-C., Wang, Y.-S., Lin, P., Jen, J., Kuo, I.-Y., Chang, Y.-H., Liao, P.-C., Chen, R.-H., Yuan, W.-C., Hsu, H.-S., Yang, M.-H., Hsu, M.-T., Wu, C.-Y., & Wang, Y.-C. (2014). Small GTPase Rab37 targets tissue inhibitor of metalloproteinase 1 for exocytosis and thus suppresses tumour metastasis. *Nature Communications*, *5*(1), 4804. https://doi.org/10.1038/ncomms5804

Vural-Ozdeniz, M., Akturk, A., Demirdizen, M., Leka, R., Acar, R., & Konu, O. (2021). CoVrimer: A tool for aligning SARS-CoV-2 primer sequences and selection of conserved/degenerate primers. *Genomics*, *113*(5), 3174–3184. https://doi.org/10.1016/j.ygeno.2021.07.020

Wang, Y., Zhou, D., Phung, S., Masri, S., Smith, D., & Chen, S. (2011). SGK3 is an estrogen-inducible kinase promoting estrogen-mediated survival of breast cancer cells. *Molecular Endocrinology (Baltimore, Md.)*, *25*(1), 72–82. https://doi.org/10.1210/me.2010-0294

Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: A revolutionary tool for transcriptomics. *Nature Reviews. Genetics*, *10*(1), 57–63. https://doi.org/10.1038/nrg2484

Wei, C., Bovonratwet, P., Gu, A., Moawad, G., Silverberg, J. I., & Friedman, A. J. (2020). Spironolactone use does not increase the risk of female breast cancer recurrence: A retrospective analysis. *Journal of the American Academy of Dermatology*, *83*(4), 1021–1027. https://doi.org/10.1016/j.jaad.2020.05.081

Wei, T., & Simko, V. (2021). *corrplot: Visualization of a Correlation Matrix* [R]. https://github.com/taiyun/corrplot (Original work published 2011)

Wu, T., Hu, E., Xu, S., Chen, M., Guo, P., Dai, Z., Feng, T., Zhou, L., Tang, W., Zhan, L., Fu, X., Liu, S., Bo, X., & Yu, G. (2021). clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *The Innovation*, *2*(3), 100141. https://doi.org/10.1016/j.xinn.2021.100141

Yang, J., & Fuller, P. J. (2012). Interactions of the mineralocorticoid receptor – Within and without. *Molecular and Cellular Endocrinology*, *350*(2), 196–205. https://doi.org/10.1016/j.mce.2011.07.001

Yu, G. (2022). *enrichplot: Visualization of Functional Enrichment Result*. https://yulab-smu.top/biomedical-knowledge-mining-book/

Zhang, Y., Parmigiani, G., & Johnson, W. E. (2020). ComBat-seq: Batch effect adjustment for RNA-seq count data. *NAR Genomics and Bioinformatics*, *2*(3), lqaa078. https://doi.org/10.1093/nargab/lqaa078

Appendix 1: Differentially expressed genes tables

MCF-7 OV Aldosterone vs. MCF-7 OV

Table 28. Full list of downregulated genes discovered in common using all four methods in the comparison of MCF_7_OVA vs. MCF_7_OV groups.

HGNC symbol	average voom counts OV	average voom counts OVA	DESeq log2FC	DESeq p value	edgeR log2FC	edgeR p value	limma log2FC	limma p value	NOISeq M value	NOISeq prob value
NR4A1	8.152	6.109	-2.10	7.04E-11	-2.07	4.06E-06	-2.08	8.38E-11	-2.32	0.860
PREX1	8.095	5.885	-2.24	1.45E-10	-2.22	4.80E-06	-2.22	1.60E-11	-2.42	0.864
GLP1R	6.448	3.785	-2.65	3.13E-08	-2.64	2.30E-11	-2.65	4.51E-10	-2.85	0.917
ACTG1	7.796	6.065	-1.79	5.88E-08	-1.76	3.36E-04	-1.76	5.82E-08	-2.01	0.847
FLNA	7.807	6.122	-1.74	2.07E-07	-1.71	1.08E-06	-1.71	8.51E-08	-1.93	0.785
FOSB	7.174	5.409	-1.81	2.45E-06	-1.78	7.31E-04	-1.82	2.02E-07	-2.06	0.846
AC006262.6	6.694	4.823	-1.92	2.73E-06	-1.91	1.66E-09	-1.91	2.31E-07	-2.16	0.847
EEF2	8.500	7.227	-1.33	7.05E-06	-1.30	2.53E-03	-1.30	6.67E-06	-1.54	0.756
FOS	7.181	5.587	-1.64	7.22E-06	-1.62	4.14E-04	-1.61	2.29E-06	-1.84	0.780
CLU	6.916	5.272	-1.70	7.67E-06	-1.68	2.53E-05	-1.67	2.16E-06	-1.92	0.780
TUFT1	5.992	3.832	-2.21	7.82E-06	-2.20	4.52E-07	-2.20	2.44E-07	-2.46	0.849
ACTB	8.654	7.363	-1.34	8.06E-06	-1.32	3.20E-03	-1.32	4.25E-06	-1.53	0.756
SPTSSB	7.157	5.695	-1.52	2.34E-05	-1.49	2.92E-05	-1.49	9.40E-06	-1.73	0.774
TFF1	6.222	4.399	-1.87	2.90E-05	-1.86	4.92E-03	-1.85	3.61E-06	-2.10	0.837
LMNA	7.133	5.734	-1.46	4.60E-05	-1.43	1.95E-06	-1.43	1.78E-05	-1.68	0.759
JUNB	7.451	6.147	-1.36	5.96E-05	-1.33	1.51E-04	-1.33	3.27E-05	-1.56	0.752
ISG15	7.759	6.474	-1.34	6.55E-05	-1.30	3.35E-03	-1.33	2.06E-05	-1.58	0.757
MKI67	7.015	5.577	-1.48	9.05E-05	-1.46	1.76E-03	-1.45	2.45E-05	-1.67	0.757
MYH9	8.105	6.957	-1.21	1.03E-04	-1.18	6.34E-03	-1.19	7.12E-05	-1.43	0.744
JUN	6.721	5.295	-1.48	1.27E-04	-1.46	1.11E-02	-1.45	4.97E-05	-1.70	0.758
EIF4G1	7.284	5.962	-1.37	1.37E-04	-1.34	1.00E-02	-1.34	5.85E-05	-1.55	0.751
DSCAM-AS1	8.769	7.640	-1.18	1.41E-04	-1.14	5.38E-03	-1.17	2.41E-05	-1.42	0.746
AC073072.5	5.109	2.672	-2.43	2.18E-04	-2.47	1.99E-07	-2.50	1.33E-06	-2.74	0.864
UBC	6.487	5.026	-1.51	2.46E-04	-1.49	3.81E-03	-1.50	4.69E-05	-1.75	0.768
BASP1	8.242	6.950	-1.30	2.56E-04	-1.28	2.94E-04	-1.30	9.82E-06	-1.46	0.748
BHLHE40-AS1	7.760	6.660	-1.16	3.41E-04	-1.13	4.09E-05	-1.14	1.77E-04	-1.39	0.732
ADAR	6.180	4.709	-1.53	4.65E-04	-1.51	1.75E-05	-1.50	7.82E-05	-1.75	0.765
HSPB1	7.070	5.886	-1.24	4.85E-04	-1.22	2.61E-03	-1.21	2.70E-04	-1.45	0.741
EGR1	7.421	6.312	-1.17	4.89E-04	-1.14	3.77E-03	-1.15	3.13E-04	-1.39	0.731

MYC	7.002	5.807	-1.25	5.14E-04	-1.23	2.35E-03	-1.22	2.68E-04	-1.46	0.741
CYR61	5.109	2.979	-2.16	5.33E-04	-2.17	5.78E-05	-2.17	1.12E-05	-2.44	0.833
RP11-398B16.2	9.026	7.979	-1.09	5.34E-04	-1.05	1.11E-03	-1.09	4.00E-05	-1.34	0.735
PLXNB2	6.117	4.662	-1.51	6.49E-04	-1.49	7.08E-04	-1.48	1.24E-04	-1.72	0.756
PGM5	5.582	3.830	-1.78	6.57E-04	-1.77	9.00E-08	-1.76	3.95E-05	-2.00	0.824
SIK1	4.774	2.416	-2.37	8.29E-04	-2.41	1.67E-07	-2.40	1.05E-05	-2.68	0.856
THBS1	6.902	5.736	-1.22	8.74E-04	-1.20	6.55E-05	-1.20	3.42E-04	-1.45	0.740
FASN	7.469	6.379	-1.14	8.75E-04	-1.11	1.06E-02	-1.11	4.71E-04	-1.32	0.725
ELMSAN1	6.276	4.779	-1.52	8.78E-04	-1.50	6.10E-04	-1.56	3.67E-05	-1.79	0.767
TRIM28	5.820	4.119	-1.69	9.38E-04	-1.68	2.73E-03	-1.69	5.19E-05	-1.87	0.765
BHLHE40	6.460	5.191	-1.33	1.04E-03	-1.30	2.18E-04	-1.31	2.70E-04	-1.56	0.743
AGRN	5.109	3.035	-2.04	1.18E-03	-2.04	2.19E-06	-2.04	2.80E-05	-2.25	0.827
CREBBP	6.151	4.685	-1.48	1.18E-03	-1.47	6.93E-04	-1.47	1.32E-04	-1.66	0.746
RP11-214J9.1	4.318	1.576	-2.72	1.29E-03	-2.82	6.70E-08	-2.75	1.60E-05	-3.08	0.878
ACTN4	6.562	5.366	-1.25	1.31E-03	-1.23	5.48E-03	-1.23	5.50E-04	-1.47	0.736
BCAS1	4.883	2.771	-2.13	1.32E-03	-2.16	8.98E-07	-2.15	2.70E-05	-2.42	0.828
IER5	5.739	4.230	-1.56	1.34E-03	-1.55	1.58E-03	-1.54	1.71E-04	-1.79	0.759
PRKCSH	6.555	5.359	-1.25	1.55E-03	-1.22	1.87E-03	-1.24	4.72E-04	-1.48	0.736
PRICKLE2-AS1 (ENSG00000241111)	7.164	6.118	-1.10	1.84E-03	-1.07	8.49E-04	-1.09	7.71E-04	-1.34	0.727
SLC9A3R1	7.195	6.177	-1.07	1.89E-03	-1.05	5.08E-03	-1.04	1.24E-03	-1.27	0.716
LGALS3BP	5.706	4.215	-1.53	2.24E-03	-1.52	1.90E-04	-1.55	1.59E-04	-1.79	0.758
IER2	6.864	5.749	-1.15	2.26E-03	-1.13	5.40E-04	-1.13	7.83E-04	-1.34	0.724
DNMT1	6.240	5.024	-1.27	2.35E-03	-1.25	3.91E-05	-1.24	6.91E-04	-1.49	0.733
PLEC	6.087	4.682	-1.40	2.78E-03	-1.39	4.75E-03	-1.40	3.00E-04	-1.58	0.741
BAG3	6.475	5.357	-1.18	2.82E-03	-1.15	3.47E-02	-1.15	1.45E-03	-1.39	0.721
JUP	6.514	5.406	-1.16	2.84E-03	-1.14	9.35E-03	-1.14	1.33E-03	-1.38	0.721
PLK2	5.261	3.641	-1.66	2.90E-03	-1.66	9.16E-04	-1.65	2.24E-04	-1.91	0.755
RPS5	5.948	4.353	-1.56	2.94E-03	-1.53	1.03E-02	-1.68	3.36E-05	-1.86	0.767
INF2	5.575	4.126	-1.49	3.10E-03	-1.48	7.20E-05	-1.47	4.05E-04	-1.71	0.741
SLC7A5	6.701	5.643	-1.11	3.52E-03	-1.08	1.05E-02	-1.08	1.75E-03	-1.30	0.714
FTH1	5.820	4.475	-1.38	3.69E-03	-1.36	1.14E-03	-1.35	6.29E-04	-1.58	0.737
ARF1	4.735	2.814	-1.95	3.93E-03	-1.98	3.59E-06	-1.96	1.28E-04	-2.23	0.817
RPL13	6.268	5.131	-1.19	4.06E-03	-1.17	1.70E-02	-1.18	1.34E-03	-1.42	0.729
RND1	5.314	3.739	-1.59	4.12E-03	-1.59	6.68E-03	-1.57	3.82E-04	-1.80	0.750
KRT8P3	5.610	4.196	-1.44	4.16E-03	-1.43	1.99E-02	-1.42	6.85E-04	-1.64	0.736
SPDEF	5.820	4.522	-1.34	4.49E-03	-1.32	3.04E-03	-1.31	9.03E-04	-1.54	0.732
DUSP1	4.709	2.807	-1.91	4.75E-03	-1.93	1.36E-05	-1.90	2.10E-04	-2.16	0.812
BCL9L	5.331	3.856	-1.52	4.94E-03	-1.51	1.04E-02	-1.52	5.38E-04	-1.77	0.750
RP11-290L1.3	5.340	3.770	-1.56	5.05E-03	-1.55	4.57E-05	-1.56	3.60E-04	-1.76	0.750
RP11-88I18.2	4.074	1.553	-2.43	6.15E-03	-2.49	3.11E-07	-2.46	1.32E-04	-2.73	0.833

MISP	5.603	4.309	-1.34	6.25E-03	-1.33	2.63E-03	-1.32	1.23E-03	-1.56	0.729
AMOTL2	4.527	2.404	-2.04	6.43E-03	-2.06	8.53E-04	-2.06	1.93E-04	-2.26	0.811
AXIN1	6.208	5.075	-1.17	6.45E-03	-1.14	6.81E-04	-1.14	1.79E-03	-1.36	0.716
DYNC1H1	5.789	4.573	-1.26	6.83E-03	-1.24	4.99E-03	-1.23	1.74E-03	-1.47	0.724
NR4A3	3.564	0.337	-3.04	6.87E-03	-3.30	4.20E-04	-3.16	3.08E-05	-3.62	0.883
TNRC18	6.285	5.253	-1.09	7.34E-03	-1.07	2.38E-03	-1.06	3.20E-03	-1.30	0.708
GTF3C1	5.215	3.767	-1.48	7.64E-03	-1.47	3.41E-04	-1.46	9.07E-04	-1.70	0.733
AC005281.2	5.058	3.551	-1.55	7.83E-03	-1.54	1.33E-05	-1.53	7.59E-04	-1.78	0.744
PTBP1	5.381	4.028	-1.39	8.07E-03	-1.38	2.95E-04	-1.37	1.22E-03	-1.60	0.730
TCF25	6.156	5.097	-1.11	8.31E-03	-1.09	2.05E-02	-1.10	2.93E-03	-1.35	0.715
FIS1	5.069	3.591	-1.52	8.44E-03	-1.52	3.28E-04	-1.51	8.52E-04	-1.77	0.744
KCNK5	4.369	2.106	-2.11	8.56E-03	-2.13	7.32E-05	-2.18	1.84E-04	-2.33	0.807
RP3-467L1.4	5.699	4.498	-1.25	8.70E-03	-1.23	5.35E-05	-1.22	2.02E-03	-1.46	0.722
CDC42BPB	5.331	3.993	-1.39	8.84E-03	-1.38	7.75E-04	-1.38	1.24E-03	-1.63	0.730
TBC1D9B	5.992	4.898	-1.14	9.22E-03	-1.12	2.07E-02	-1.11	3.35E-03	-1.35	0.712
BAMBI	3.900	1.300	-2.46	9.22E-03	-2.54	4.32E-05	-2.52	1.92E-04	-2.77	0.828
Sep-09	6.166	5.141	-1.08	9.29E-03	-1.06	1.22E-02	-1.06	3.91E-03	-1.30	0.707
SMARCB1	4.973	3.474	-1.54	9.70E-03	-1.54	2.88E-02	-1.52	1.20E-03	-1.78	0.742
RPLP0P6	5.027	3.318	-1.66	9.89E-03	-1.64	2.35E-02	-1.80	1.68E-04	-1.97	0.812
USP36	4.962	3.462	-1.53	1.02E-02	-1.53	3.41E-05	-1.51	1.02E-03	-1.77	0.741
IGSF3	4.984	3.521	-1.51	1.10E-02	-1.51	2.38E-04	-1.50	1.04E-03	-1.76	0.741
ATF3	4.335	1.931	-2.16	1.10E-02	-2.20	1.26E-02	-2.52	3.61E-05	-2.57	0.814
IFI27	5.987	4.919	-1.12	1.10E-02	-1.09	4.92E-02	-1.11	3.39E-03	-1.36	0.713
EVPL	5.314	4.017	-1.34	1.15E-02	-1.34	8.20E-04	-1.34	1.60E-03	-1.59	0.728
SDF4	4.683	2.924	-1.72	1.16E-02	-1.73	8.60E-04	-1.72	6.51E-04	-1.93	0.741
SLC25A30-AS1	4.951	3.167	-1.66	1.16E-02	-1.67	2.16E-03	-1.72	3.57E-04	-1.84	0.745
PFKL	5.129	3.763	-1.41	1.18E-02	-1.41	2.67E-03	-1.40	1.54E-03	-1.66	0.726
RPS11	5.215	3.884	-1.37	1.19E-02	-1.36	1.73E-02	-1.35	2.06E-03	-1.59	0.726
IP6K2	5.398	4.156	-1.29	1.20E-02	-1.28	3.14E-04	-1.28	2.09E-03	-1.53	0.724
RP11-356I2.4	3.780	0.518	-2.67	1.26E-02	-2.85	5.44E-04	-3.23	1.48E-05	-3.21	0.859
KRT80	5.296	4.022	-1.32	1.29E-02	-1.31	2.48E-04	-1.29	2.29E-03	-1.54	0.721
IFITM3	5.783	4.655	-1.17	1.29E-02	-1.15	6.27E-03	-1.18	2.40E-03	-1.42	0.720
TGOLN2	5.243	3.847	-1.39	1.33E-02	-1.38	1.60E-03	-1.38	1.50E-03	-1.58	0.725
CUX1	5.158	3.832	-1.37	1.33E-02	-1.36	1.11E-02	-1.37	1.94E-03	-1.62	0.726
BCAR1	5.119	3.631	-1.46	1.35E-02	-1.46	8.91E-03	-1.46	1.20E-03	-1.65	0.726
PRRC2B	5.406	4.124	-1.30	1.39E-02	-1.28	8.99E-04	-1.28	2.19E-03	-1.49	0.717
AC004453.8	4.352	2.533	-1.83	1.41E-02	-1.85	2.60E-04	-1.82	8.57E-04	-2.09	0.794
CEP170B	5.356	4.134	-1.27	1.42E-02	-1.26	9.94E-04	-1.24	2.91E-03	-1.49	0.716
HELZ2	4.940	3.382	-1.52	1.47E-02	-1.51	2.26E-04	-1.53	1.06E-03	-1.72	0.732
SYMPK	4.556	2.928	-1.66	1.49E-02	-1.67	2.64E-03	-1.65	1.18E-03	-1.92	0.736
PSMC3	5.422	4.244	-1.22	1.56E-02	-1.21	3.71E-02	-1.20	4.01E-03	-1.44	0.715
CCNL2	4.173	2.245	-1.92	1.58E-02	-1.96	1.12E-04	-1.92	8.58E-04	-2.20	0.797

MYBL2	5.732	4.518	-1.20	1.61E-02	-1.19	1.30E-02	-1.20	2.49E-03	-1.37	0.709
JUND	5.596	4.504	-1.14	1.67E-02	-1.13	3.92E-04	-1.12	4.75E-03	-1.37	0.705
MDH2	4.811	3.255	-1.53	1.74E-02	-1.53	5.94E-03	-1.53	1.49E-03	-1.73	0.735
CTGF	5.322	3.988	-1.33	1.74E-02	-1.32	1.29E-02	-1.41	9.95E-04	-1.62	0.730
EIF4G2	5.158	3.834	-1.34	1.74E-02	-1.33	5.80E-03	-1.32	2.68E-03	-1.54	0.718
MT-RNR1	5.493	4.358	-1.18	1.75E-02	-1.16	1.02E-02	-1.15	4.75E-03	-1.39	0.704
MED15	3.805	0.826	-2.41	1.78E-02	-2.48	2.24E-04	-2.80	1.06E-04	-2.67	0.822
GATAD2A	5.406	4.197	-1.23	1.81E-02	-1.22	7.96E-03	-1.21	3.62E-03	-1.42	0.712
EVI5L	3.201	0.094	-2.92	1.82E-02	-3.25	3.11E-06	-3.00	8.33E-05	-3.54	0.864
RBBP8NL	4.094	1.824	-2.08	1.87E-02	-2.17	5.09E-04	-2.37	1.30E-04	-2.48	0.803
L1CAM	4.774	3.342	-1.46	1.94E-02	-1.46	8.78E-03	-1.44	2.50E-03	-1.69	0.721
RPS23	4.995	3.516	-1.43	1.96E-02	-1.43	1.07E-02	-1.45	1.70E-03	-1.62	0.721
RPL37	5.296	4.136	-1.21	2.03E-02	-1.20	8.46E-03	-1.20	4.48E-03	-1.45	0.713
NUP210	4.735	3.318	-1.45	2.14E-02	-1.45	1.44E-03	-1.43	2.71E-03	-1.68	0.717
SEC16A	4.571	3.083	-1.52	2.22E-02	-1.54	3.62E-04	-1.52	2.19E-03	-1.78	0.730
SF3B4	4.229	2.511	-1.74	2.23E-02	-1.77	1.39E-03	-1.74	1.63E-03	-2.02	0.787
DAPK3	4.094	1.922	-1.96	2.27E-02	-2.00	1.30E-03	-2.07	6.83E-04	-2.18	0.791
INTS1	4.940	3.654	-1.33	2.27E-02	-1.32	3.39E-04	-1.31	3.67E-03	-1.56	0.712
ZFP36	4.053	2.193	-1.86	2.29E-02	-1.90	2.30E-05	-1.86	1.42E-03	-2.14	0.789
SCIN	4.951	3.674	-1.32	2.32E-02	-1.31	5.77E-05	-1.30	3.83E-03	-1.55	0.713
HCFC1	5.168	3.993	-1.22	2.37E-02	-1.22	3.73E-03	-1.21	4.72E-03	-1.47	0.711
INPPL1	4.655	3.228	-1.46	2.37E-02	-1.47	1.66E-02	-1.47	2.68E-03	-1.73	0.730
RP11-568J23.6	4.735	3.354	-1.42	2.41E-02	-1.43	3.20E-04	-1.42	2.61E-03	-1.68	0.717
RNF216	3.344	0.663	-2.53	2.48E-02	-2.69	1.68E-05	-2.60	4.74E-04	-2.95	0.835
EFNA1	5.909	3.544	-2.25	2.48E-02	-2.24	4.17E-05	-2.30	2.31E-07	-2.39	0.846
RP11-244J10.1	3.201	0.337	-2.68	2.61E-02	-2.95	6.45E-04	-2.80	2.96E-04	-3.25	0.830
RERE	4.571	3.130	-1.47	2.63E-02	-1.48	3.52E-04	-1.45	3.14E-03	-1.71	0.715
MXD4	4.283	2.529	-1.69	2.65E-02	-1.71	1.41E-03	-1.71	1.80E-03	-1.92	0.726
KRT8P33	4.542	3.085	-1.48	2.70E-02	-1.49	4.78E-02	-1.46	3.76E-03	-1.72	0.720
PTPRF	5.058	3.785	-1.27	2.89E-02	-1.26	1.39E-02	-1.26	4.64E-03	-1.46	0.707
SLC12A7	4.074	2.346	-1.74	2.90E-02	-1.79	1.16E-04	-1.75	2.08E-03	-2.03	0.782
CEP131	4.655	3.292	-1.40	2.91E-02	-1.41	5.40E-04	-1.41	3.38E-03	-1.66	0.712
ZC3H4	4.655	3.321	-1.37	3.13E-02	-1.38	6.05E-04	-1.37	4.04E-03	-1.63	0.712
MLLT1	3.780	1.679	-1.98	3.19E-02	-2.03	6.77E-04	-2.03	1.59E-03	-2.25	0.783
ST14	4.917	3.701	-1.25	3.21E-02	-1.24	1.48E-02	-1.23	6.64E-03	-1.47	0.704
FLOT2	3.593	1.310	-2.16	3.21E-02	-2.28	9.09E-04	-2.33	7.03E-04	-2.58	0.784
GTF3C5	4.229	2.630	-1.61	3.22E-02	-1.64	8.55E-03	-1.65	2.40E-03	-1.91	0.724
TAOK3	3.829	1.940	-1.89	3.22E-02	-1.96	1.73E-04	-1.91	1.83E-03	-2.21	0.783
NINJ1	4.402	2.937	-1.50	3.23E-02	-1.52	4.67E-04	-1.50	3.27E-03	-1.77	0.722
IRF7	4.628	3.293	-1.37	3.29E-02	-1.37	9.81E-05	-1.35	4.86E-03	-1.61	0.710
PML	3.621	1.481	-2.07	3.30E-02	-2.14	3.58E-04	-2.09	1.71E-03	-2.38	0.779
H2AFX	4.962	3.784	-1.21	3.50E-02	-1.20	1.82E-03	-1.19	7.43E-03	-1.42	0.701
WWP2	3.238	0.761	-2.39	3.56E-02	-2.57	1.27E-05	-2.44	1.02E-03	-2.84	0.827

TMSB4XP8	3.676	1.255	-2.08	3.70E-02	-2.17	1.80E-02	-2.28	1.00E-03	-2.33	0.779
GRN	4.642	3.305	-1.34	3.71E-02	-1.35	3.42E-02	-1.33	6.13E-03	-1.56	0.704
CTD-2033A16.2	4.386	2.973	-1.45	3.79E-02	-1.46	5.90E-04	-1.44	4.48E-03	-1.71	0.708
RP1-278E11.3	4.585	3.292	-1.33	4.01E-02	-1.33	1.90E-02	-1.34	5.84E-03	-1.59	0.707
RPUSD1	4.335	2.841	-1.48	4.03E-02	-1.49	9.49E-03	-1.47	4.83E-03	-1.70	0.706
ABCA2	4.210	2.562	-1.58	4.04E-02	-1.59	1.16E-03	-1.60	3.40E-03	-1.80	0.717
ACTN1	4.696	3.315	-1.34	4.08E-02	-1.33	6.38E-03	-1.35	4.88E-03	-1.53	0.704
ATP6AP1	3.564	1.372	-2.03	4.12E-02	-2.09	6.53E-05	-2.11	1.88E-03	-2.32	0.773
SNRNP200	4.917	3.749	-1.20	4.25E-02	-1.19	1.22E-02	-1.23	6.16E-03	-1.47	0.704
DNPEP	4.527	3.211	-1.34	4.26E-02	-1.34	1.27E-03	-1.32	6.84E-03	-1.57	0.703
CHMP1A	4.094	2.501	-1.58	4.28E-02	-1.61	5.72E-03	-1.57	4.69E-03	-1.83	0.715
RP11-589M4.4	3.309	-0.184	-2.42	4.32E-02	-2.55	2.04E-03	-3.29	1.83E-05	-2.73	0.798
NOTCH1	3.829	1.455	-1.93	4.49E-02	-1.96	1.17E-03	-2.20	8.72E-04	-2.13	0.779
NAA35	2.489	-1.436	-3.27	4.61E-02	-4.01	2.73E-04	-3.60	1.16E-06	-4.41	0.843
RGS16	4.074	2.458	-1.58	4.69E-02	-1.60	2.16E-02	-1.58	4.99E-03	-1.81	0.712
CAPN15	4.094	2.335	-1.62	4.71E-02	-1.64	6.00E-03	-1.68	3.22E-03	-1.83	0.715
ABAT	3.780	1.653	-1.86	4.76E-02	-1.90	7.02E-03	-2.00	1.95E-03	-2.08	0.771
ACACB	2.912	-0.032	-2.59	4.98E-02	-2.95	3.50E-05	-2.79	3.22E-04	-3.25	0.812

MCF-7 OV Aldosterone + Spironolactone vs. MCF-7 OV Aldosterone

Table 29. Full list of upregulated genes discovered in common using all four methods in the comparison of MCF_7_OVAS vs. MCF_7_OVA groups.

HGNC symbol	average voom counts OVA	average voom counts OVAS	DESeq log2FC	DESeq p value	edgeR log2FC	edgeR p value	limma log2FC	limma p value	NOISeq M value	NOISeq prob value
RAB37	6.483	9.926	3.47	1.54E-52	3.44	3.86E-54	3.44	3.12E-33	3.63	0.964
DNAH10	6.494	7.751	1.29	5.34E-07	1.26	7.12E-08	1.25	3.70E-06	1.44	0.741
RP11-856F16.2	6.822	7.988	1.20	1.22E-06	1.17	4.13E-07	1.17	7.72E-06	1.37	0.731
RP11-388P9.2	7.959	8.992	1.07	1.86E-06	1.03	2.12E-07	1.03	8.53E-06	1.22	0.708
RP11-1035H13.2	6.172	7.279	1.13	1.52E-05	1.11	4.70E-07	1.10	8.63E-05	1.28	0.709
CTD-2049O4.1	4.539	5.987	1.49	2.10E-05	1.47	6.86E-06	1.46	2.76E-05	1.67	0.745
MYCBP2-AS1	6.817	7.855	1.06	2.46E-05	1.03	2.25E-06	1.03	8.28E-05	1.20	0.702
SEPT4-AS1	3.551	5.249	1.71	2.48E-05	1.70	2.38E-08	1.69	3.82E-05	1.89	0.755
RP11-50E11.2	4.693	6.075	1.39	3.53E-05	1.37	1.32E-06	1.36	6.42E-05	1.53	0.735
RP11-793H13.3	7.106	8.203	1.11	3.85E-05	1.08	3.09E-05	1.08	2.17E-05	1.23	0.706
TMEM186	6.454	7.477	1.06	3.89E-05	1.03	4.83E-06	1.02	1.79E-04	1.21	0.701
RP11-317G6.1	5.084	6.375	1.35	5.97E-05	1.32	9.53E-05	1.31	5.59E-05	1.53	0.740
FCMR	4.881	6.186	1.37	6.44E-05	1.35	7.27E-05	1.31	8.51E-05	1.54	0.737
RP11-110I1.11	5.502	6.730	1.22	8.06E-05	1.19	1.66E-05	1.20	7.98E-05	1.34	0.718
AC017101.10	4.396	5.804	1.39	9.23E-05	1.37	2.17E-06	1.38	1.11E-04	1.53	0.729
HIST1H2BK	3.654	5.207	1.57	9.59E-05	1.56	2.07E-07	1.55	1.26E-04	1.75	0.742
WDFY3-AS1	4.447	5.832	1.40	1.31E-04	1.38	2.76E-05	1.36	1.31E-04	1.54	0.730
NONOP2	5.107	6.253	1.17	1.47E-04	1.15	4.56E-06	1.14	3.92E-04	1.33	0.708
MBNL1-AS1	4.975	6.195	1.22	1.53E-04	1.20	7.96E-06	1.20	2.50E-04	1.36	0.711
TAB3-AS2	5.109	6.394	1.25	1.68E-04	1.23	3.90E-05	1.25	1.00E-04	1.37	0.715
HIST2H2AC	1.931	4.200	2.14	2.41E-04	2.17	2.25E-07	2.18	1.21E-04	2.35	0.792
TNFRSF19	2.703	4.644	1.90	2.51E-04	1.91	4.86E-05	1.86	1.27E-04	2.05	0.777
RP11-18D7.2	4.380	5.752	1.33	3.34E-04	1.31	4.79E-05	1.33	2.07E-04	1.45	0.716
CTD-3032J10.2	4.853	6.104	1.30	3.37E-04	1.28	3.24E-04	1.22	2.65E-04	1.42	0.719
RP11-815J21.2	4.264	5.566	1.34	3.72E-04	1.32	9.05E-05	1.32	3.10E-04	1.53	0.724
C16orf45	3.746	5.169	1.46	3.79E-04	1.45	2.60E-05	1.43	3.29E-04	1.65	0.726
SP2-AS1	5.000	6.215	1.19	4.25E-04	1.17	1.17E-04	1.18	3.03E-04	1.31	0.705

CTD-2310F14.1	3.984	5.439	1.40	4.35E-04	1.38	3.17E-05	1.40	2.35E-04	1.52	0.719
ANKRD61	4.206	5.607	1.33	6.19E-04	1.31	9.70E-05	1.35	2.55E-04	1.45	0.712
CTB-193M12.1	3.817	5.171	1.35	7.68E-04	1.34	1.65E-05	1.33	7.59E-04	1.51	0.711
AC010733.5	3.864	5.240	1.35	7.90E-04	1.34	2.80E-05	1.34	6.24E-04	1.50	0.712
HIST1H2APS3	2.773	4.435	1.67	9.99E-04	1.67	1.16E-05	1.63	6.93E-04	1.85	0.727
RP1-85F18.6	3.744	5.188	1.37	1.08E-03	1.36	7.92E-05	1.39	5.08E-04	1.50	0.711
RP11-48B3.3	2.676	4.466	1.73	1.24E-03	1.73	5.90E-04	1.86	1.66E-04	1.98	0.765
CSTF3	3.567	4.958	1.37	1.38E-03	1.36	4.18E-05	1.35	1.05E-03	1.52	0.705
HFE	2.368	4.172	1.70	2.23E-03	1.70	1.74E-05	1.72	9.79E-04	1.87	0.717
RP4-569M23.2	2.335	4.203	1.69	2.55E-03	1.69	5.37E-05	1.76	8.11E-04	1.85	0.717
RP11-207C16.4	2.047	3.888	1.82	2.57E-03	1.85	1.66E-05	1.80	1.34E-03	2.04	0.742
RP11-575L7.4	2.781	4.506	1.54	3.00E-03	1.54	3.52E-04	1.62	7.41E-04	1.67	0.704
SLC7A11-AS1	3.224	4.698	1.47	3.38E-03	1.47	1.92E-03	1.40	1.63E-03	1.59	0.709
HIST1H4PS1	2.103	3.853	1.71	4.28E-03	1.73	5.47E-06	1.70	2.18E-03	1.92	0.702
AC094019.4	2.036	4.048	1.72	4.38E-03	1.74	3.05E-04	1.86	8.84E-04	1.87	0.711
TTC31	1.481	3.516	1.96	4.56E-03	2.02	7.54E-05	1.97	1.80E-03	2.22	0.734
ABCA4	0.415	3.084	2.27	6.47E-03	2.39	1.18E-05	2.53	5.16E-04	2.60	0.755
DIO1	-0.104	3.219	2.15	8.78E-03	2.24	2.14E-04	3.08	2.73E-05	2.39	0.737
HOTAIR	0.905	3.084	2.08	9.00E-03	2.21	2.67E-05	2.11	2.45E-03	2.42	0.750
RP11-33B1.4	1.481	3.354	1.82	1.06E-02	1.88	6.58E-04	1.79	4.81E-03	2.06	0.712
THOC7-AS1	0.977	3.187	1.96	1.11E-02	2.02	2.60E-05	2.08	2.46E-03	2.22	0.710
KCNJ13	0.905	3.015	2.03	1.23E-02	2.14	1.19E-04	2.04	3.61E-03	2.35	0.720
RP11-35G22.1	-2.228	1.992	3.48	1.30E-02	4.80	4.47E-04	3.66	1.83E-07	5.53	0.800
CD300LD	0.518	2.849	2.18	1.33E-02	2.37	6.43E-04	2.28	1.89E-03	2.63	0.742
NKG7	1.026	2.975	2.01	1.42E-02	2.11	3.87E-03	1.86	7.53E-03	2.30	0.726
RP3-508I15.10	0.149	2.646	2.21	1.88E-02	2.39	1.52E-04	2.38	1.43E-03	2.62	0.722
RP11-152H18.4	0.663	2.834	2.01	1.91E-02	2.13	2.09E-04	2.07	4.23E-03	2.34	0.704
RPL23AP64	4.703	6.279	1.46	2.75E-02	1.44	7.32E-04	1.51	1.09E-05	1.53	0.739
EVI5L	0.094	2.425	2.19	2.76E-02	2.46	7.80E-04	2.22	3.31E-03	2.71	0.720
RP11-476F14.1	-1.067	2.137	2.48	3.15E-02	2.95	1.72E-03	2.76	1.66E-04	3.27	0.739
RP5-1107A17.3	-0.824	2.238	2.32	3.39E-02	2.60	3.82E-04	2.82	1.31E-04	2.85	0.717
RP11-120K24.5	-0.032	2.419	2.12	3.45E-02	2.41	2.81E-04	2.28	2.59E-03	2.66	0.700
RP11-23J18.1	-1.436	1.497	2.65	4.86E-02	3.30	3.91E-03	2.60	5.68E-04	3.65	0.721

Table 30. Full list of downregulated genes discovered in common using all four methods in the comparison of MCF_7_OVAS vs. MCF_7_OVA groups.

HGNC symbol	average voom counts OVA	average voom counts OVAS	DESeq log2FC	DESeq p value	edgeR log2FC	edgeR p value	limma log2FC	limma p value	NOISeq M value	NOISeq prob value
SLC35F2	7.884	4.496	-3.33	5.79E-30	-3.37	5.79E-34	-3.39	6.11E-22	-3.20	0.928
HOOK2	8.706	6.174	-2.49	7.97E-21	-2.52	1.38E-21	-2.54	4.05E-19	-2.37	0.862
RP11-459E5.1	7.219	4.709	-2.47	1.48E-17	-2.51	8.31E-24	-2.50	8.48E-14	-2.32	0.853
MIF-AS1	13.149	11.050	-2.04	6.14E-17	-2.07	1.51E-15	-2.10	3.35E-30	-1.90	0.796
AP000688.14	8.041	5.804	-2.20	6.70E-16	-2.24	6.53E-18	-2.24	1.93E-14	-2.08	0.828
CTD-2369P2.5	7.531	4.900	-2.60	8.40E-16	-2.63	4.42E-15	-2.65	1.10E-15	-2.50	0.881
GRK6	7.636	5.150	-2.48	5.78E-15	-2.52	3.83E-14	-2.51	4.36E-15	-2.40	0.858
PKN3	7.375	5.083	-2.24	6.67E-15	-2.29	7.72E-17	-2.28	8.69E-13	-2.08	0.824
SMPD4	7.981	5.918	-2.02	2.79E-13	-2.07	2.00E-12	-2.05	1.47E-12	-1.85	0.788
CENPN	7.459	5.528	-1.89	1.35E-12	-1.93	4.61E-17	-1.93	1.49E-10	-1.74	0.775
ENO3	10.478	8.537	-1.86	1.74E-12	-1.90	9.54E-11	-1.94	6.99E-18	-1.72	0.780
PYCARD-AS1	7.082	4.551	-2.43	4.40E-12	-2.47	2.98E-11	-2.55	1.98E-13	-2.34	0.853
GDF9	9.369	7.533	-1.82	3.27E-11	-1.87	5.94E-09	-1.82	2.03E-13	-1.62	0.768
CFAP57	7.042	4.862	-2.12	8.79E-11	-2.16	4.07E-11	-2.19	2.93E-11	-2.01	0.820
CCDC74BP1	6.613	4.502	-2.06	9.71E-11	-2.10	7.32E-15	-2.11	9.93E-10	-1.92	0.779
PDAP1	9.901	8.301	-1.56	3.12E-10	-1.59	1.20E-10	-1.61	1.28E-12	-1.44	0.746
RP11-196G11.6	7.029	4.999	-1.95	3.81E-10	-1.99	4.27E-11	-2.03	3.18E-10	-1.82	0.776
CTC-250I14.6	11.875	10.516	-1.32	6.04E-10	-1.36	1.12E-11	-1.36	7.95E-14	-1.18	0.700
NDUFB2-AS1	8.722	6.999	-1.71	2.06E-09	-1.75	1.23E-07	-1.70	5.61E-11	-1.51	0.754
H1FX	7.439	5.827	-1.57	2.09E-09	-1.61	2.47E-12	-1.60	3.25E-08	-1.41	0.729
AP006621.8	8.819	6.987	-1.76	3.25E-09	-1.79	4.85E-08	-1.85	1.34E-12	-1.66	0.768
PRR19	8.771	7.067	-1.62	6.49E-09	-1.66	8.77E-08	-1.70	3.92E-11	-1.47	0.751
RP3-416H24.1	9.586	7.908	-1.59	1.06E-08	-1.63	1.96E-07	-1.68	1.52E-12	-1.44	0.746
MROH8	6.512	4.524	-1.93	1.18E-08	-1.97	1.31E-10	-2.00	7.19E-09	-1.82	0.770
LA16c-366D1.3	7.383	5.721	-1.64	1.60E-08	-1.67	5.27E-10	-1.67	1.37E-08	-1.52	0.748
LAMA5	10.031	8.537	-1.43	1.80E-08	-1.47	7.11E-08	-1.49	1.69E-11	-1.28	0.719
CTD-3065J16.9	6.127	3.949	-2.07	2.68E-08	-2.12	1.22E-09	-2.18	7.25E-09	-1.96	0.774
P2RY11	8.703	7.188	-1.47	4.43E-08	-1.52	3.59E-07	-1.50	3.06E-09	-1.29	0.717
CTB-31O20.3	6.861	5.113	-1.73	4.45E-08	-1.78	5.40E-08	-1.73	6.18E-08	-1.54	0.744
LBHD1	8.813	7.212	-1.52	4.63E-08	-1.56	4.63E-07	-1.60	2.84E-10	-1.36	0.732
RP11-417L19.4	6.166	4.302	-1.81	7.75E-08	-1.86	2.55E-11	-1.86	1.83E-07	-1.67	0.745

RP1-117B12.4	7.772	6.167	-1.55	1.05E-07	-1.58	6.23E-08	-1.62	9.12E-09	-1.43	0.739
AP003068.12	5.469	3.187	-2.26	1.06E-07	-2.32	1.30E-10	-2.29	1.03E-07	-2.18	0.804
NPRL3	10.200	8.730	-1.39	1.21E-07	-1.43	6.31E-07	-1.47	1.49E-11	-1.25	0.714
PC	9.711	8.228	-1.41	1.40E-07	-1.44	5.29E-07	-1.49	7.90E-11	-1.27	0.718
RP11-867G23.2	5.167	2.746	-2.36	2.31E-07	-2.43	6.33E-12	-2.42	2.39E-07	-2.28	0.819
DRAP1	6.690	4.861	-1.72	2.67E-07	-1.76	1.36E-07	-1.82	3.47E-08	-1.57	0.748
CTD-2369P2.4	9.633	8.181	-1.41	2.72E-07	-1.45	4.53E-06	-1.44	4.51E-10	-1.22	0.709
CTD-3149D2.4	7.082	5.615	-1.42	3.35E-07	-1.46	2.12E-09	-1.46	9.19E-07	-1.27	0.704
RP11-552F3.9	6.033	4.187	-1.78	4.32E-07	-1.83	3.15E-10	-1.85	3.69E-07	-1.66	0.742
DHRS4-AS1	6.896	5.406	-1.45	4.75E-07	-1.50	2.16E-08	-1.48	1.40E-06	-1.29	0.704
RP11-452I5.2	9.439	6.778	-2.69	5.39E-07	-2.71	4.62E-13	-2.69	1.25E-22	-2.62	0.889
SGK3	5.017	2.431	-2.43	5.86E-07	-2.51	1.55E-10	-2.54	3.46E-07	-2.32	0.814
UXT-AS1	7.204	5.703	-1.46	8.61E-07	-1.51	9.24E-07	-1.49	5.35E-07	-1.28	0.706
RP11-530C5.1	6.555	4.779	-1.73	1.03E-06	-1.76	7.59E-07	-1.80	8.00E-08	-1.64	0.752
RP11-475C16.1	6.386	4.552	-1.72	1.19E-06	-1.76	4.10E-07	-1.83	1.21E-07	-1.58	0.743
RP5-827C21.4	5.066	2.107	-2.56	1.35E-06	-2.64	5.05E-07	-2.95	3.58E-08	-2.50	0.842
DUS2	5.029	2.746	-2.31	3.02E-06	-2.38	1.69E-07	-2.31	1.06E-06	-2.27	0.819
RP11-618K13.2	6.598	4.970	-1.56	4.15E-06	-1.61	6.62E-06	-1.61	7.65E-07	-1.39	0.717
MCMDC2	7.156	5.676	-1.51	4.68E-06	-1.53	6.21E-06	-1.50	5.07E-07	-1.42	0.733
NIP7	4.742	2.415	-2.25	6.76E-06	-2.33	2.59E-10	-2.29	5.22E-06	-2.14	0.778
APOA1-AS	5.066	3.053	-1.96	8.85E-06	-2.02	2.82E-11	-2.00	6.28E-06	-1.85	0.743
KCNK15-AS1	4.518	1.984	-2.50	9.17E-06	-2.61	6.68E-08	-2.54	4.04E-06	-2.49	0.828
MVK	6.358	4.813	-1.54	1.23E-05	-1.57	4.57E-06	-1.57	2.64E-06	-1.45	0.726
RP11-410D17.2	5.245	3.392	-1.82	1.41E-05	-1.88	1.04E-08	-1.86	8.41E-06	-1.72	0.734
AC005775.2	5.354	3.588	-1.71	2.08E-05	-1.77	1.62E-08	-1.77	1.17E-05	-1.60	0.723
RPS2	4.651	2.313	-2.20	3.39E-05	-2.28	1.06E-07	-2.32	6.66E-06	-2.13	0.777
CTD-2540B15.7	5.825	4.278	-1.56	3.42E-05	-1.59	2.47E-06	-1.57	1.46E-05	-1.46	0.722
AC142528.1	5.177	3.294	-1.90	3.84E-05	-1.95	9.83E-06	-1.92	7.47E-06	-1.84	0.746
RP11-845M18.6	5.212	3.512	-1.64	7.25E-05	-1.69	7.14E-08	-1.70	3.44E-05	-1.52	0.707
AL022341.3	5.062	3.214	-1.85	7.61E-05	-1.93	4.43E-05	-1.79	3.91E-05	-1.67	0.717
SURF1	4.406	2.188	-2.17	7.76E-05	-2.27	6.31E-08	-2.22	2.92E-05	-2.13	0.767
NRL	4.741	2.720	-1.97	9.09E-05	-2.04	1.88E-06	-2.04	1.90E-05	-1.91	0.738
C19orf57	4.149	1.590	-2.48	1.01E-04	-2.60	2.92E-06	-2.58	1.85E-05	-2.51	0.815

AC114730.8	4.389	2.137	-2.23	1.32E-04	-2.31	2.84E-05	-2.29	2.32E-05	-2.22	0.777
CTD-2292P10.2	3.908	1.112	-2.61	1.71E-04	-2.76	2.45E-06	-2.80	2.16E-05	-2.67	0.820
C16orf92	4.110	1.711	-2.27	1.86E-04	-2.38	2.19E-07	-2.33	6.86E-05	-2.18	0.752
ROMO1	4.118	1.776	-2.20	2.42E-04	-2.31	5.60E-08	-2.31	6.70E-05	-2.15	0.754
RP11-758P17.3	5.484	2.273	-3.22	3.72E-04	-3.30	8.46E-08	-3.26	4.93E-10	-3.26	0.918
SH3D21	3.541	0.834	-2.61	5.39E-04	-2.81	1.12E-06	-2.69	1.01E-04	-2.71	0.798
RP11-452K12.4	3.483	0.555	-2.63	7.15E-04	-2.85	5.78E-08	-2.84	6.82E-05	-2.71	0.786
RP11-532F12.5	3.890	1.661	-2.14	7.20E-04	-2.26	6.86E-07	-2.21	2.10E-04	-2.10	0.735
CTD-2207O23.11	5.910	3.175	-2.49	8.98E-04	-2.56	1.53E-07	-2.75	2.22E-10	-2.43	0.861
THAP7-AS1	3.339	0.834	-2.65	1.03E-03	-2.83	2.24E-04	-2.56	2.86E-04	-2.80	0.805
SS18L2	3.731	1.438	-2.15	1.28E-03	-2.29	7.03E-07	-2.27	2.82E-04	-2.13	0.729
PHLDA2	3.055	-0.014	-2.81	1.97E-03	-3.12	5.23E-07	-2.92	8.35E-05	-3.00	0.774
EIF4E3	2.854	0.229	-2.39	8.20E-03	-2.64	6.70E-06	-2.51	8.12E-04	-2.50	0.717
ZBTB16	2.471	-1.175	-2.85	1.00E-02	-3.33	3.38E-05	-3.28	5.08E-06	-3.25	0.747
CTD-2666L21.3	5.153	2.799	-2.11	1.02E-02	-2.18	9.74E-05	-2.27	1.33E-06	-1.91	0.748
SLBP	2.471	-0.382	-2.62	1.20E-02	-3.02	1.15E-04	-2.77	2.38E-04	-2.97	0.742
OVOL2	2.012	-2.336	-3.42	1.50E-02	-4.74	2.93E-05	-3.74	1.10E-07	-5.11	0.760
RP11-785H5.2	2.290	-0.751	-2.65	1.87E-02	-3.13	7.20E-05	-2.83	1.36E-04	-3.07	0.704
SHANK3	1.626	-2.336	-3.13	3.42E-02	-4.42	5.17E-04	-3.34	5.04E-06	-4.76	0.714

Comparison of log2FC values of the MCF-7 OVA vs. OV and MC-7 OVAS vs. OVA comparisons

Table 31. Genes that were significantly upregulated in the MCF-7 OVA vs. OV comparison and significantly downregulated in the MCF-7 OVAS vs. OVA comparison.

HGNC symbol	OVA vs. OV log2FC	OVA vs. OV p-value	OVAS vs. OVA log2FC	OVAS vs. OVA p-value
GLP1R	-2.64	2.30E-11	1.13	2.19E-03
BCAS1	-2.16	8.98E-07	1.07	1.29E-02
CXCL12	-2.32	2.28E-06	1.23	4.32E-02
EVI5L	-3.25	3.11E-06	2.46	7.80E-04
MED15	-2.48	2.24E-04	1.33	4.92E-02
NAA35	-4.01	2.73E-04	2.40	4.39E-02
RERE	-1.48	3.52E-04	1.31	4.23E-04
TTC31	-1.93	3.74E-04	2.02	7.54E-05
RPL10AP6	-3.95	5.16E-04	3.09	4.28E-03
HOTAIR	-1.98	5.47E-04	2.21	2.67E-05
MLLT1	-2.03	6.77E-04	1.26	2.84E-02
FLOT2	-2.28	9.09E-04	1.66	1.51E-02
DSCAM	-2.25	9.61E-04	1.48	3.53E-02
MBNL1-AS1	-1.02	9.69E-04	1.20	7.96E-06
GNB5	-3.40	1.09E-03	2.54	4.47E-02
MICAL3	-2.05	1.20E-03	1.67	6.03E-03
AMOTL1	-1.91	1.33E-03	1.69	2.55E-03
RP3-486I3.4	-1.57	1.37E-03	1.03	3.26E-02
BBX	-2.26	1.38E-03	1.90	7.70E-03
TGOLN2	-1.38	1.60E-03	1.10	5.04E-03
NRDE2	-2.03	1.92E-03	1.47	3.88E-02
PCLO	-4.48	2.02E-03	3.76	2.60E-02
RP11-120K24.5	-2.26	2.30E-03	2.41	2.81E-04
SEZ6L2	-1.51	2.78E-03	1.25	8.86E-03
VRK3	-2.28	3.11E-03	1.57	4.85E-02
RP11-207C16.4	-1.46	3.37E-03	1.85	1.66E-05
RPL23AP42	-1.96	3.44E-03	1.91	2.61E-03
BCAT2	-2.50	3.47E-03	1.66	4.71E-02
PRDX3P1	-1.70	3.52E-03	1.22	2.51E-02
USE1	-2.90	4.04E-03	2.59	1.56E-02
LRRC27	-2.74	4.06E-03	2.03	3.36E-02
LTBP1	-1.55	4.69E-03	1.59	1.54E-03
KIAA0556	-1.93	4.82E-03	1.58	2.26E-02
ATG2A	-2.13	5.33E-03	1.57	4.76E-02
KIAA2012	-1.92	5.95E-03	1.63	1.43E-02
NCOA6	-1.70	8.29E-03	1.43	1.44E-02

FDFT1	-1.29	8.39E-03	1.02	2.81E-02
RP11-127O4.3	-2.53	8.61E-03	2.24	3.67E-02
RP11-815J21.2	-1.01	9.22E-03	1.32	9.05E-05
CDK19	-2.83	9.49E-03	2.78	1.70E-02
RPAP1	-2.84	1.01E-02	2.55	1.53E-02
ABCA4	-1.64	1.39E-02	2.39	1.18E-05
RP11-486F17.1	-1.50	1.41E-02	1.42	2.01E-02
ENTPD3	-1.18	1.42E-02	1.22	5.98E-03
UBA52P5	-1.75	1.45E-02	1.73	6.21E-03
RP5-1107A17.3	-2.09	1.49E-02	2.60	3.82E-04
RBM23	-1.37	1.49E-02	1.01	4.89E-02
PPP2R2A	-1.48	1.52E-02	1.27	3.02E-02
CT62	-1.42	1.55E-02	1.26	2.06E-02
DAZAP2P1	-2.46	1.66E-02	2.02	4.09E-02
FOXO3	-3.22	1.76E-02	2.54	4.47E-02
POT1-AS1	-1.34	1.83E-02	1.68	6.51E-04
GNAI3	-1.46	1.85E-02	1.27	2.83E-02
RP11-336A10.2	-1.08	1.93E-02	1.30	1.25E-03
SLF1	-2.18	1.94E-02	2.53	2.88E-03
CETP	-2.98	2.00E-02	2.69	1.60E-02
BRD8	-1.78	2.03E-02	1.57	3.77E-02
PTPN13	-2.69	2.08E-02	2.90	2.64E-03
ANGPTL1	-1.46	2.12E-02	1.54	6.61E-03
FANCC	-1.12	2.18E-02	1.05	1.45E-02
TNFRSF19	-1.24	2.22E-02	1.91	4.86E-05
SLC7A11-AS1	-1.22	2.27E-02	1.47	1.92E-03
NEDD9	-1.89	2.30E-02	1.94	1.22E-02
HS1BP3	-1.69	2.35E-02	1.32	4.83E-02
RP11-33B1.4	-1.38	2.40E-02	1.88	6.58E-04
WWC2	-2.69	2.42E-02	2.90	3.39E-03
TCF12	-2.69	2.55E-02	2.79	6.18E-03
ATP6V0A2	-1.83	2.73E-02	1.82	2.74E-02
RP11-245J9.5	-1.08	2.93E-02	1.55	2.57E-04
RP11-6E9.4	-1.34	3.06E-02	1.71	1.41E-03
PHACTR2P1	-1.45	3.21E-02	1.37	1.72E-02
DIO1	-1.49	3.59E-02	2.24	2.14E-04
NAB1	-3.41	3.83E-02	3.48	4.19E-02
RP11-458I7.1	-1.56	3.88E-02	1.95	4.10E-03
RP11-507K2.2	-1.56	3.89E-02	1.43	3.51E-02
RPL7AP10	-1.08	4.09E-02	1.08	1.98E-02
ZCCHC6	-1.13	4.13E-02	1.26	1.15E-02
RP11-476F14.1	-2.40	4.21E-02	2.95	1.72E-03
PRNP	-2.53	4.23E-02	2.25	4.14E-02
URB1	-2.19	4.26E-02	2.15	3.08E-02

SCAPER	-2.67	4.40E-02	3.23	7.12E-03
TFPI2	-1.69	4.59E-02	1.67	3.29E-02
Y_RNA (ENSG00000199349)	-1.72	4.94E-02	1.81	2.90E-02

Table 32. Genes that were significantly downregulated in the MCF-7 OVA vs. OV comparison and significantly upregulated in the MCF-7 OVAS vs. OVA comparison.

HGNC symbol	OVA vs. OV log2FC	OVA vs. OV p-value	OVAS vs. OVA log2FC	OVAS vs. OVA p-value
SGK3	4.15	8.20E-13	-2.51	1.55E-10
SH3PXD2A	2.13	4.10E-09	-1.24	4.38E-06
JRK	2.06	1.43E-07	-1.21	3.64E-05
APOA1-AS	1.70	8.28E-07	-2.02	2.82E-11
RPS27L	3.28	7.13E-06	-1.48	3.14E-03
RP11-69H7.3	1.61	4.17E-05	-1.03	5.96E-04
SYCP3	1.35	5.80E-05	-1.08	3.52E-05
GPATCH11	1.49	1.09E-04	-1.01	1.26E-03
ZBTB16	5.73	1.17E-04	-3.33	3.38E-05
SPATA9	1.70	1.67E-04	-1.26	5.85E-04
AC004510.3	1.20	7.73E-04	-1.39	2.12E-06
AC019349.5	3.48	1.09E-03	-2.00	4.02E-03
KLHL8	1.27	1.40E-03	-1.30	2.38E-04
OVOL2	3.53	1.51E-03	-4.74	2.93E-05
FANCF	2.20	2.51E-03	-1.70	4.32E-03
RP11-188C12.2	2.95	2.57E-03	-1.39	3.28E-02
SRRM2-AS1	1.34	2.69E-03	-1.05	4.69E-03
HINT1	1.11	3.04E-03	-1.34	2.10E-05
PRR4	2.08	4.99E-03	-3.06	5.60E-05
TAMM41	2.01	5.05E-03	-1.70	5.94E-03
C5orf22	2.65	5.17E-03	-1.71	1.16E-02
MCMDC2	1.19	5.33E-03	-1.53	6.21E-06
MIR3074	2.21	5.82E-03	-1.28	3.43E-02
SNRPFP1	1.11	7.72E-03	-1.18	4.68E-04
SMPD5	1.23	8.17E-03	-1.97	8.64E-06
RP11-363E6.3	1.85	9.01E-03	-1.30	2.01E-02
RP11-332H14.2	2.48	9.49E-03	-1.53	2.66E-02
RP11-392O18.2	1.34	1.49E-02	-2.12	2.84E-04
AC102953.6	1.71	1.65E-02	-2.99	6.63E-05
RC3H2	1.74	1.79E-02	-1.22	3.14E-02
ETFA	1.38	2.01E-02	-1.09	2.22E-02
UBE2J2	1.20	2.18E-02	-1.95	8.17E-05
RP1-212P9.3	2.55	2.31E-02	-3.77	1.96E-03

PRKAG1	1.47	2.48E-02	-1.13	4.11E-02
WDYHV1	2.37	2.48E-02	-1.71	3.76E-02
OSBPL9P4	1.29	2.50E-02	-1.23	8.69E-03
RPL22L1	1.70	2.57E-02	-1.90	7.21E-03
RP11-61K9.2	2.07	2.82E-02	-1.39	3.74E-02
CTD-2396E7.9	1.80	2.89E-02	-1.49	4.06E-02
HPRT1	1.40	3.23E-02	-1.47	1.81E-02
AKR1C6P	2.39	3.47E-02	-2.40	2.15E-02
MT1P3	2.39	3.47E-02	-2.40	2.15E-02
RN7SKP98	4.14	3.47E-02	-2.40	2.15E-02
C1orf216	1.13	3.65E-02	-1.14	2.03E-02
NUDT9	2.32	3.71E-02	-1.64	3.86E-02
PNRC1	1.46	3.88E-02	-1.71	1.05E-02
AGMAT	1.22	4.59E-02	-1.15	3.35E-02
LDHD	1.09	4.94E-02	-1.24	7.45E-03