A Comprehensive review of RNA-Seq Data Analysis in R using DESeq2

Maddie Medeiros

Github: <https://github.com/Obstghost/bioinfo_final_project>

Video: <https://youtu.be/nqlMEr6cdWw>

DESeq2: <https://bioconductor.org/packages/release/bioc/vignettes/DESeq2/inst/doc/DESeq2.html>

Airway (package used to access data): <https://bioconductor.org/packages/release/data/experiment/vignettes/airway/inst/doc/airway.html>

Paper: <https://pubmed.ncbi.nlm.nih.gov/24926665/>

Data: <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE52778>

Table of Contents

[Purpose 2](#_Toc1937003175)

[Background Information 3](#_Toc1368254842)

[Data 4](#_Toc638123217)

[DESeq2 4](#_Toc1372970721)

[Code Walkthrough 5](#_Toc717960454)

[Step 1: Get the Data 5](#_Toc1541198137)

[Step 2: Loading Packages 6](#_Toc1272682756)

[Step 3: Set Working Directory 7](#_Toc1500023797)

[Step 4: Read in data 7](#_Toc1961508360)

[Step 5: Make sure data frames match 7](#_Toc729623597)

[Step 7: Prefiltering 8](#_Toc494963140)

[Step 8: Setting factor level 8](#_Toc1149217663)

[Step 9: Run DESeq 9](#_Toc1619392934)

[Step 10: Visualize results 9](#_Toc1574110755)

[Work Cited 10](#_Toc1439560157)

# Purpose

The purpose of this tutorial is to conduct an analysis of RNA-seq counts data to identify differentially expressed genes across different conditions. We will be utilizing the DESeq2 package in R to conduct this analysis. DESeq2 employs statistical techniques to determine what genes are significantly upregulated or downregulated between the conditions being studied. This guide will cover what packages are needed, how to read in data, how to make a DESeq dataset object, filtering our data, running the DESeq function and finally, how to visualize our results by using a volcano plot. A volcano plot is a type of scatter plot that is commonly used in the field of bioinformatics to visualize large scale data such as RNA-seq data.

# Background Information

The data set we will be analyzing comes from a study addressing the underlying mechanisms of how glucocorticoids suppress inflammation in airway smooth muscle (ASM). [(Himes BE, 2014)](https://pubmed.ncbi.nlm.nih.gov/24926665/)Glucocorticoids are a class of corticosteroids, which Is a group of steroids produced in the adrenal cortex. They are named in their role in glucose metabolism but have potent anti-inflammatory and immunosuppressive properties. Glucocorticoids are the primary treatment strategy for treating asthma because they have anti-inflammatory effects in many lung tissues. Until this study, the mechanism in which glucocorticoids suppress inflammation in ASM was unknown. The researchers used RNA-seq to characterize transcriptomic changes in four primary human ASM cell lines. The lines were treated with dexamethasone which is a potent synthetic glucocorticoid. They found 316 differentially expressed genes this list of 316 genes underwent further analysis to see what sets of genes were statistically significant enriched in the samples. CRISPLD2 along with many other genes where differentially expressed and enriched. CRISPLD2 was further studied because it hasn’t been investigated.

CRISPLD2 which encodes a secreted protein which has been implicated in lung development and endotoxin regulation. Through western blotting and QRT-PCR, CRISPLD2 showed increase expression when cells were treated with dexamethasone. Overall, this study outlined the effects of glucocorticoid on ASM transcriptome and identified CRSPLD2 as a candidate gene that regulates anti-inflammatory effects of glucocorticoids in ASM [(Himes BE, 2014)](https://pubmed.ncbi.nlm.nih.gov/24926665/).

This tutorial will not be able to reach the same conclusion as the study but that does not mean that our analysis doesn’t work or isn’t useful. In this tutorial we are only investigating the differential expression in genes between untreated and treated groups. We are only conducting the first step of this study to find out what genes we should be interested in. We would have to conduct gene set enrichment analysis to gather more context of what genes are used in certain pathways. We need this because we could have genes that are differentially expressed but have nothing to do with anti-inflammatory responses.

# Data

The data are publicly available and can be accessed with the GEO accession number [GSE52778](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE52778). It is also available in a package called [‘airway’](https://bioconductor.org/packages/release/data/experiment/html/airway.html) within BiocManager. However, the data within the airway package requires data wrangling in order to get the counts data. Wrangling the data is outside the scope of this tutorial so instead I have made csv files (count\_data.csv, sample\_info.csv) that we will use in our analysis, which can be found on my [Github](https://github.com/Obstghost/bioinfo_final_project). I will include the code I used to wrangle the data in case anyone is curious, but it is not necessary for this tutorial.

# DESeq2

DESeq2 is one of the most widely used R packages for bioinformatics to analyze count data from sequencing experiments such as RNA-seq ([Pluto](https://pluto.bio/blog/deseq2-an-overview-of-popular-rna-seq-analysis-package)). In this tutorial, we will be analyzing differential gene expression on cell lines that have been either treated with dexamethasone or not (Himes BE, 2014).

But how does DESeq2 work? DESeq2 takes count data as input, which is usually organized in a matrix. In this matrix, genes are listed in rows and the samples are listed in columns. DESeq2 will normalize the counts to account for sequencing depth and RNA composition effects. DESeq2 normalizes the counts by calculating size factors to make each sample comparable to one another. This calculation is done by using the median-of-ratios method. This method finds the median of the ratio of counts for each gene relative to a geometric mean per gene across all samples. DESeq2 will then estimate the measure of dispersion, which shows the variability of counts across samples. DESeq2 will assume that counts for genes are over dispersed relative to a Poisson distribution. It is better modeled against a negative binomial distribution. The dispersion estimation is important because it can influence the statistical power to detect differential expression. Once these estimates are calculated DESeq2 will fit a negative binomial generalized linear model to the counts for each gene. The generalized linear model fitting process considers the dispersion and the size factors to adjust the counts. Now DESeq2 will perform hypothesis testing for each gene to determine if there is statistical evidence for differential expression across conditions. Likelihood ratio tests are performed to calculate the p-values for each gene (Love, 2014).Because DESeq2 is testing thousands of genes all at the same time it needs to apply multiple testing corrections to control the false discovery rate. This is accomplished through the Benjamin-Hochberg procedure which will adjust p-values to account for multiple hypothesis testing. Finally, the final output will include estimated log2 fold change and adjusted p-values (q-value) for each gene. Genes who have a p-value less than 0.05 are considered differentially expressed (Love, 2014).

In summary, DESeq2 employs a statistical framework to normalize count data, estimates variability and tests for differential expression. DESeq2’s strength comes from its ability to model count data accurately and provide reliable results even with low counts and high variability. These reasons are why DESeq2 is a popular bioinformatics tool used to analyze differential gene expression.

# Code Walkthrough

Now that we know how DESeq2 works and where our data comes from let’s start coding! All these steps will be outlined in the video but here is all the information in written form. This tutorial was run on macOS after any shortcut listed the windows short cut will be given in parenthesis.

## Step 1: Get the Data

1. Click on the [github](https://github.com/Obstghost/bioinfo_final_project) link provided on the cover page. Once there, download the csv files: [counts\_data.csv](https://github.com/Obstghost/bioinfo_final_project/blob/main/counts_data.csv) and [sample\_info.csv](https://github.com/Obstghost/bioinfo_final_project/blob/main/sample_info.csv)
2. Make sure you know where these files are being downloaded, I suggest creating a new directory (folder) in your home directory that will hold the data along with your script
   1. Open a new finder window (file explorer)
   2. Use the short cut: command + shift + H (windows+shift+H) to get to your home directory
   3. Right click and create a new folder and save it to whatever name you like in my case my directory is called ‘bioinfo\_final\_project’
   4. Once created move the data files from your download folder to the folder you just created

## Step 2: Loading Packages

1. If you already have these packages installed: tidyverse, DESeq2 and airway you can move on to the next step but if not, I will show you how to install these packages.
2. Open RStudio and create a new script file by going to the top right of your screen select File > New File > R script
3. Installing tidyverse
   1. A blue and pink text

      Description automatically generatedThis package can be installed by using the line of code above
   2. In R tidyverse Is an open source collection of packages used in data science for cleaning data.
4. Installing DESeq2 and airwayA computer screen shot of text

   Description automatically generated with medium confidence
   1. DESeq2 and airway can be installed with these lines of code
   2. If errors are occurring visit this [website](https://bioconductor.org/packages/release/bioc/html/DESeq2.html) and copy and paste the installation code
5. Once everything has been successfully installed, we can now load the libraries

A group of white and pink text

Description automatically generated

## Step 3: Set Working Directory

1. Set your working directory by using the line of code below
2. Remember to change the directory name (“bioinfo\_final\_project”) if need be

A blue and yellow text

Description automatically generated

## Step 4: Read in data

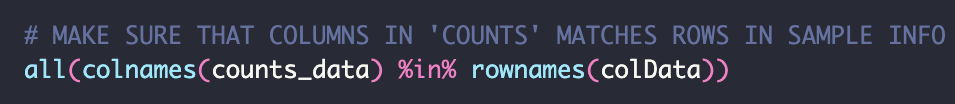
1. Read in counts and sample info data by using the read.csv() function

A close-up of text

Description automatically generated

## Step 5: Make sure data frames match

1. We need to make sure that our column names in the counts dataframe match with the row names in our colData dataframe we can test this by using this line of code



1. We also have to make sure that the samples are presented in the same order as well

A close up of text

Description automatically generated

1. Both of these lines of code should return TRUE. If the data doesn’t match or it is not in the right order DESeq will throw an error!

Step 6: Construct DESeq dataset

1. To create this object we will use the DESeqDataSetFromMatrix() function. And our arguments will be: countData, colData and design.

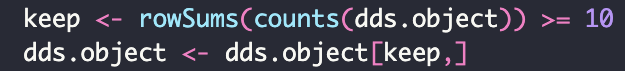
A blue and white text

Description automatically generated

1. countData argument is where we specify where our counts are
2. colData argument is where we specify sample info data frame
3. design argument is referring to our different conditions
   1. indicates how to model the samples or measure the effect of the condition.
   2. The value should be the column name in colData

## Step 7: Prefiltering

1. We are going to remove rows that have less then 10 reads across all samples by using this line of code



1. This is a highly recommended step because this is going to reduce the size of the object and will increase the speed of computation.

## Step 8: Setting factor level

1. We want to compare cell lines that have been treated with dexamethasone versus untreated. We need to tell DESeq2 which is our reference level. In our case it would be untreated.
2. We can accomplish this by using the relevel() function



## Step 9: Run DESeq

1. Now we are ready to run DESeq! We will do this by using the DESeq() function



1. We will now look at the results of our analysis by using the results() function



## Step 10: Visualize results

1. Congratulations you have officially ran DESeq in R! Now we can take this data and visualize it! In this tutorial we will make a volcano plot
2. Volcano Plot
   1. We will now use ggplot2 to make our plot. We have already installed tidyverse so all you need to do is load your package:
   2. I have also included two variables threshold and pval threshold these are just storing values that will be used in our plot to tell what genes are significant.
   3. A screenshot of a computer

      Description automatically generated
   4. Here is the code on how to make your graph:
   5. A computer screen with text and symbols

      Description automatically generated
   6. Here is the graph:
   7. A graph with red dots

      Description automatically generated
   8. From this graph we can see that there are many differentially expressed genes within our study. Each gene is represented by a dot. On the y-axis is the log of our p-value and on the x-axis is our log fold change. Genes who have a log fold change less than 0 means there was a decrease in expression. While those greater than 0 means these genes had an increase in their expression. Genes that are labeled red have a log fold change greater than 1.5 and have a p-value less than 0.05. From this graph we would want to focus more on the red genes because it indicates there is differential expression occurring.

# 

# Work Cited

Camp, D. (n.d.). Retrieved from Rdocumentation: <https://www.rdocumentation.org/packages/DESeq2/versions/1.12.3/topics/plotPCA>

“DESeq2: An Overview of a Popular RNA-Seq Analysis - Pluto Bioinformatics.” *DESeq2: An Overview of a Popular RNA-Seq Analysis - Pluto Bioinformatics*, pluto.bio/blog/deseq2-an-overview-of-popular-rna-seq-analysis-package. Accessed 12 Dec. 2023.

Himes BE, J. X.-S. (2014). RNA-Seq transcriptome profiling identifies CRISPLD2 as a glucocorticoid responsive gene that modulates cytokine function in airway smooth muscle cells. *PLOS ONE*.

Love, M. H. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*.

Michael I. Love, S. A. (2023, 10 24). *Analyzing RNA-seq data with DESeq2*. Retrieved from bioconductor: https://bioconductor.org/packages/release/bioc/vignettes/DESeq2/inst/doc/DESeq2.html#the-deseq2-model