Michelle Gregoire

Bio539

Dr. Rachel Schwartz

2021 April 22

*Final Project Essay*

Introduction

Asthma affects 1 in 13 people in the United States and is the leading chronic disease in children. Each day about 10 people die from asthma (Asthma Facts and Figures). With increasing levels of pollution, these numbers are expected to increase. Current treatments of asthma involve the use of drugs that including β2-agonists and glucocorticosteroids such as dexamethasone. These treatments primarily target receptors in the smooth muscle tissues of the airway. The response of the cells to these treatments were not well characterized, so an RNA sequencing (RNA seq) experiment of human airway smooth muscle cells in response to treatment asthma medications was performed by Himes BE, Jiang X, Wagner P, Hu R et al. The dataset of this RNA seq experiment is available for public access from the Gene Expression Omnibus (GEO) from NCBI (Accession: PRJNA229998 ID: 229998). GEO is a database repository of high throughput gene expression data and hybridization arrays, chips, microarrays, and RNA seq data. Raw data from the RNA seq experiment was obtained from SRA files and analyzed through a pipeline to check for sequence quality before checking differential expression and principal components analysis. Results from the study indicate that dexamethasone does indeed alter the transcriptome of human airway smooth muscle cells.

Methods

Finding and downloading a dataset:

The dataset “Human Airway Smooth Muscle Transcriptome Changes in Response to Asthma Medications” was obtained from <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE52778>. This file contains 16 Sequence Read Archives (SRA files) containing the raw sequencing reads. These files were obtained by clicking on the "SRA Run Selector" button towards the bottom of the page, clicking on the SRR file name, the "Data access" tab, and then copying the link address from NCBI. These files were downloaded onto the HPC terminal command line using the command "wget" followed by the links.

Preparing the samples for a quality check:

To check the quality of the sequences, the SRA files were converted into fastq files using the SRA Toolkit from NCBI. Fastq generates two files for each SRR. One will be \_1.fastq left reads and the other \_2.fastq right rights. These files are the paired ends.

Checking the quality of the samples:

Sequence quality is rated with quality scores (Q) which are derived from the formula Q = -10log10(e). "e" is the estimated probability of the base call being wrong. A Q of 30 indicates that there is a 1 in 1000 probability of an incorrect base call and that the Inferred Base Call Accuracy is 99.9%. Likewise a Q of 20 indicates there is a 1 in 100 probability of an incorrect base call and has an accuracy of 99%. Generally a Q of 20 or above indicates good quality sequence data. The quality of the sequences was analyzed with a program called Fastqc. To view the html output files from Fastqc, they were download form the HPC locally with a program called CyberDuck. The fastqc report gives information on: basic statistics (such as the type of sequencing platform and %GC content), per base sequence quality (in the form of Q scores to represent how accurate each base was --again a Q of 30 would mean a 1 in 1000 chance that the base was wrongly incorporated), per sequence quality scores (the average Q score for the whole sequence), per base sequence content (this should ideally show parallel lines), perbase N content, per base GC content (can indicate sample contamination), sequence length distribution, sequence duplication levels, overrepresented sequences, and adaptor content. Since the quality of the reads were good, trimming was not necessary.

Aligning to a Reference Genome:

The reads were then aligned to a reference genome using the program STAR. The human reference genome GRCh38 (hg38) was used because it is a well cited and relatively recent reference genome. It was released in December 2013 and has the latest assembly of the human genome with greatly expanded alternate (ALT) contigs. These alternate haplotypes include highly variable HLA loci and represent common complex variations of the human genome. To align to this reference, an index of the reference genome was first created on STAR with the hg38 reference genome FASTA file and its annotation which provides information about the gene structure and splicing sites (in Gene Transfer Format (GTF)). Both of these files were download to the command line from the UCSC genome browser. The output file from STAR was specified to save the results as a BAM sorted by genome coordinates (binary SAM file)

Preparing the BAM files for differential gene expression analysis:

To analyze the files for differential gene expression, they were changed from coordinate sorted BAM files to a merged .gtf file with the program Stringtie. The output from Stringtie were then converted into a matrix of counts as a .txt file using a slight modification to the python script supplied by Stringtie in order to specify the specific files. The .txt file was then converted into a .csv file to be read in R.

Testing differential gene expression:

The .csv file was then download locally using CyberDuck and uploaded to R-Studio. The following packages were then installed on R: "DESeq2", ggplot2, and tidyverse. Differential gene expression for treatment of the human airway smooth muscle cells with dexamethasone (dex) was plotted for one of the genes (as an example for how to visualize differential expression for a specific gene) and it was also plotted for the top six significant changes. Principal component analysis for treatment of human airway smooth muscle cells with dexamethasone (dex) was also evaluated. Differential expression was evaluated using the raw data from the csv file, but for other downstream analysis of RNA seq, the raw count data needs to be transformed because for these analyses it is not clear how to best compute a distance metric. Therefore, variance stabilizing transformation which removes the dependence of the variance on the mean, was performed on the data prior to plotting principal component analysis.

Results

Quality:

For the fastqc results, since there are 16 reads, figures will only be shown for the first read SRR1039508.1, but results will be reported for all the reads in general. For per base sequence quality (Q score) the samples were generally in the mid 30s, meaning that there is a 1 in 1000 chance of a wrong base call (Figure 1A). This was also true for the per sequence quality scores (Figure 1B). This means these are good quality reads. The per base sequence content was a little variable at beginning but this is okay because RNA seq data has a little variation from the random hexamers used to generate the cDNA and since there is the same pattern of bias across files it is okay (Figure 1C). The per base GC for the reads also reflected the theoretical distribution meaning the samples are not likely to be contaminated (Figure 1D). The per base n content was also low across the files (Figure 1E). The sequence length was 63 bp (Figure 1F). Again, some sequence duplication was seen at beginning but that is okay because RNA seq data reflects the transcriptome where some transcripts may be duplicated in high numbers (Figure 1G). Likewise, there were some overrepresented sequences but these are alright. The reads also had low adaptor contents (close to 0) (Figure 1H).



Figure 1. Fastqc analysis output graphs for SRR1039508.1 (Sorry my cursor was caught in graph A during the screenshot!).

Differential Gene Expression:

The summary of the results of the differential gene expression output can be seen in table 1. In this table the base mean, log2 fold change, and p values can be seen across the samples in a range from the minimum to the maximum. This summary is good to get a general idea of the spread of the data and indicates that there are some significant changes in transcript expression across the samples.

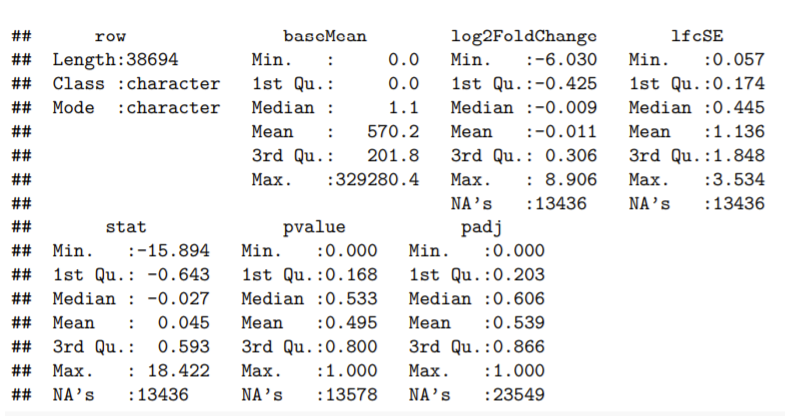


Table 1. Summary of the results of the deseq2 analysis on the counts.

A graph of one of the genes (ENSG00000103196/ CRISPLD2 -promotes matrix assembly) was made to get a general idea for how to visualize the information for each gene (Figure2). This graph demonstrated a clear and significant (p<0.5) increase in expression of ENSG00000103196/ CRISPLD2 for the cells treated with dexamethasone compared to the untreated cells.

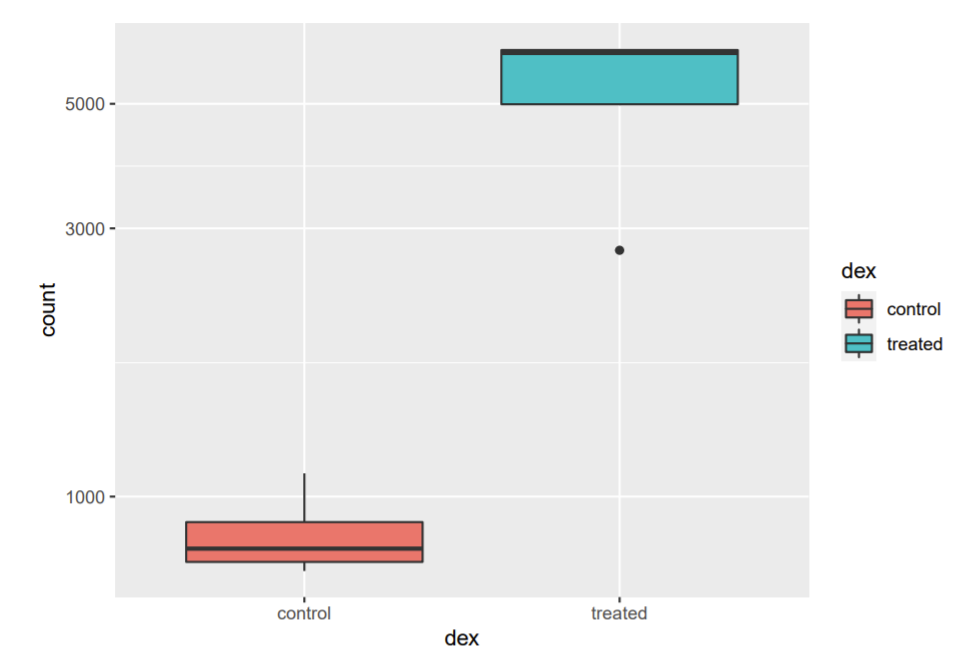


Figure 2. Expression of ENSG00000103196 (CRISPLD2) in human airway smooth muscle cells treated with dexamethasone.

The top six significantly differentially expressed genes were then sorted from the data by the adjusted p value and can be seen in table 2. The output of this sort can be seen visually in Figure 3. In general treatment with dexamethasone caused a significant increase in expression of genes ENSG00000152583 (SPARCL1), ENSG00000179094 (PER1), ENSG00000189221 (MAOA), ENSG00000120129 (DUSP1), and ENSG00000148175 (STOM), but showed a significant down regulation for ENSG00000116584 (ARHGEF2).

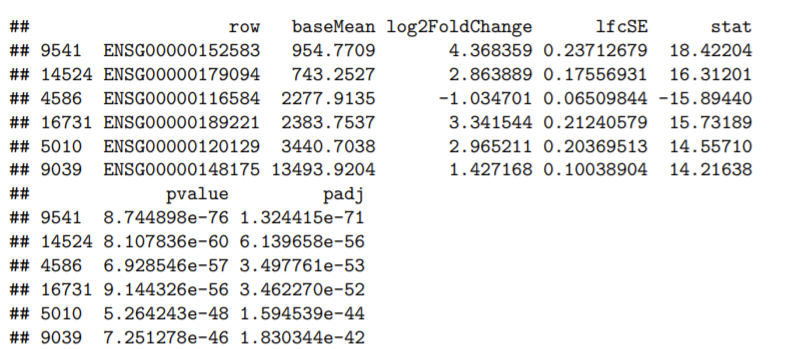


Table 2. Summary of top six significantly differentially expressed genes.

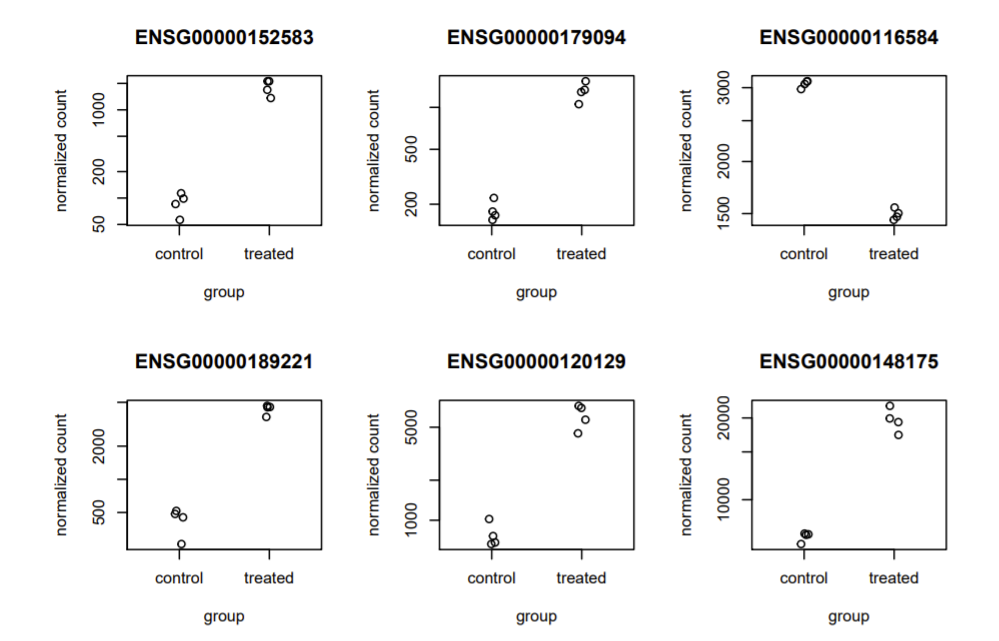


Figure 3. Graphs of top six significantly differentially expressed genes. Up regulation of ENSG00000152583 (SPARCL1), ENSG00000179094 (PER1), ENSG00000189221 (MAOA), ENSG00000120129 (DUSP1), and ENSG00000148175 (STOM). Down regulation for ENSG00000116584 (ARHGEF2).

Principal component analysis:

Principal component analysis demonstrated that for PC1 (the most variation) that there was 32% variance and for PC2 (the second most variation) that there was 24% variance. This means that the control and the treated were significantly different from one another and that even within the groups there were some significant differences.

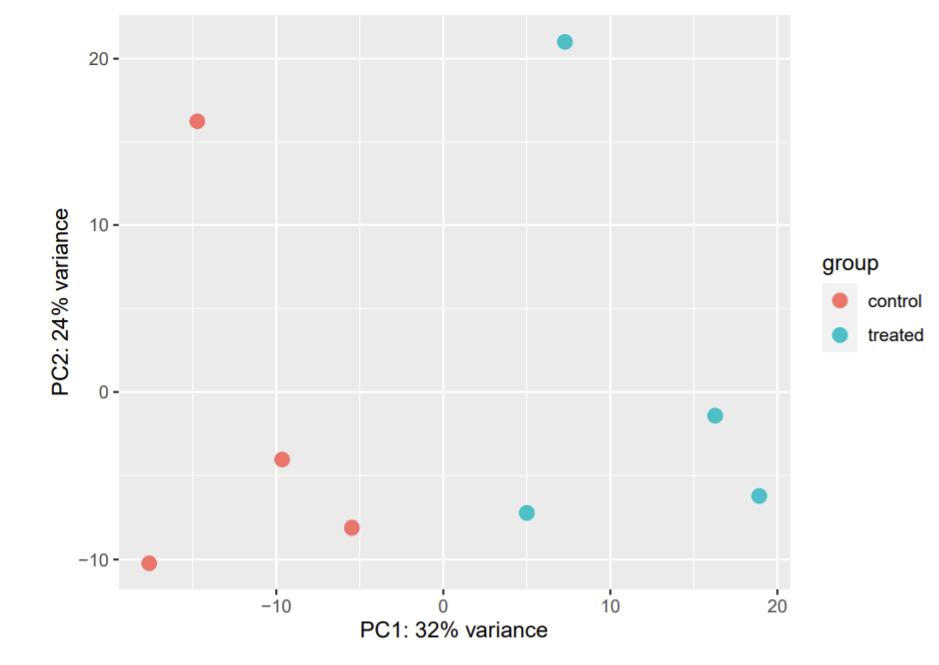


Figure 4. Principal component analysis.

Discussion

Asthma is a condition that causes airways narrow and swell. Asthma is generally treated with anti-inflammatory drugs such as dexamethasone. These drugs interact with receptors on the surface of airway smooth muscle cells. The results from this study investigated if treatment with asthma therapeutics changed the transcriptome of airway smooth muscle cells.

The raw data RNA sequencing reads demonstrated a good quality after they were checked with Fastqc. Differential expression analysis showed that there were differences between control cells versus those treated with dexamethasone, particularly that there was a significant up regulation of ENSG00000152583 (SPARCL1 – interacts with extracellular matrix to create intermediate states of cell adhesion), ENSG00000179094 (PER1 -involved in circadian rhythm and apoptosis induction), ENSG00000189221 (MAOA - encode mitochondrial enzymes that catalyze the oxidative deamination of amines), ENSG00000120129 (DUSP1 - (aka MKP-1) dephosphorylates MAPKs), and ENSG00000148175 (STOM - regulates ion channel activity and transmembrane ion transport) and a significant down regulation for ENSG00000116584 (ARHGEF2 - activates Rho-GTPases by promoting the exchange of GDP for GTP). This indicates that dexamethasone is indeed causing transcriptome level changes in gene expression. More work would need to be done to investigate if these changes are positive or negative.

The principal components analysis demonstrated that there was a difference between the controls and the treatments. It also showed that there was some variation between the samples themselves in the control and treatment groups. This could be due to individual differences between the genomes of the donors of those samples.

Asthma is one of the most common chronic childhood disorders and about 10 people die of asthma each day in the United States. With many individuals in the United States affected by asthma (1 in 13) who likely take treatments such as dexamethasone, the importance of investigating how these treatments may affect gene expression is important. This report only looked at differential expression for the top six most significant gene changes for treatment with dexamethasone and at the principal component analysis between controls and treatment with anti-asthma drugs on human airway smooth muscle cells. This concludes the analysis in this report, however there are many more analyses that can be performed such as heat map analysis. This project only introduces the basics of RNA seq workflow analysis using Shell/Linux on an HPC to process the raw data and Deseq2 on R Studio to analyze the results.

References

Asthma Facts and Figures. Asthma and Allergy Foundation of America. Available from: <https://www.aafa.org/asthma-facts/#:~:text=According%20to%20the%20Centers%20for,in%2013%20people%20have%20asthma.&text=More%20than%2025%20million%20Americans,age%2C%20sex%20and%20racial%20groups>.

Count-Based Differential Expression Analysis of RNA-seq Data. Available from: https://4va.github.io/biodatasci/r-rnaseq-airway.html#pca

Dobin, A., & Gingeras, T. R. (2015). Mapping RNA-seq Reads with STAR. *Current protocols in bioinformatics*, *51*, 11.14.1–11.14.19. <https://doi.org/10.1002/0471250953.bi1114s51>

Fastqc. Available from : <https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>

GEO NCBI. Human Airway Smooth Muscle Transcriptome Changes in Response to Asthma Medications. Available from : <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE52778>

GRCh38. Available from: ftp://ftp.ebi.ac.uk/pub/databases/gencode/Gencode\_human/release\_37/GRCh38.p13.genome.fa.gz

GRCh38 Annotation. Available from: <ftp://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_37/gencode.v37.chr_patch_hapl_scaff.basic.annotation.gtf.gz>

Himes BE, Jiang X, Wagner P, Hu R et al. RNA-Seq transcriptome profiling identifies CRISPLD2 as a glucocorticoid responsive gene that modulates cytokine function in airway smooth muscle cells. PLoS One 2014;9(6):e99625. PMID: 24926665

Roberts, E.M. 2020. "Pipeline to perform differential expression analysis with transcriptomes". <https://github.com/erinroberts/apoptosis_data_pipeline>. The University of Rhode Island.

SRA Toolkit. Available from : <https://ftp-trace.ncbi.nlm.nih.gov/sra/sdk/2.11.0/sratoolkit.2.11.0-centos_linux64.tar.gz>

STAR. Available from : <https://github.com/alexdobin/STAR>

Stringtie. Available from : https://ccb.jhu.edu/software/stringtie/index.shtml?t=manual

Stringtie Python prepDE Script : Available from : https://ccb.jhu.edu/software/stringtie/dl/prepDE.py