RNA sequence data and Differential Expression analysis in asthma treatment

**Introduction**

Asthma is a chronic and common condition, and treatment with dexamethasone is a common maneuver. This study was done by using dexamethasone on four people with different genes and using a controlled trial to see if the dexamethasone treatment had an effect. This study demonstrates the use of R for bioanalysis to determine differential genes expression, identify significant genes and enrichment analysis. This research will focus on learning how to load packages, import data, perform exploratory analysis with built in functions as well as functions from packages installed, performing differential expression analysis of RNA-seq data with the DESeq2 package, and visualizing the results using ggplot2.

**Method**

This research will work with the airway dataset. This data set comes from an RNA-Seq experiment, a high throughput sequencing method, on four human airway smooth muscle cell lines treated and untreated with dexamethasone. The data set has four control groups up to 38,695 rows. I have adopted 1) T test and Linear Regression, 2) Differential Expression Analysis and 3) enrichment analysis. In this case we choose variables cell and dex because we care about the cell line and which samples are treated with dexamethasone versus which samples are untreated controls. The basis for this type of analysis is common when analyzing high-throughput data. It has the following steps: Extract the expression values for a single gene. Run compare the mean expression between two groups using a statistical test. Repeat former steps for every gene. Then, from dplyr package to filter out results based on padj < 0.01. Gene set enrichment analysis (GSEA), is a method to identify classes of genes that are over-represented in a large set of genes.

**Discussion**

First, I converted the data into a matrix and performed Ttest to see if the effect of medication in the treatment was significant. As shown in Figure 1, a single gene is extracted to establish a linear regression. Because it is a simple linear relationship, the linear regression here is also equivalent to the Ttest result. The t-test is a common choice for performing a differential analysis. Next we will perform a simple differential test comparing treated and control groups in our gene expression data. The "dex" column in `metadata` gives group values ​​for treated and control samples . I created a new data.frame called `genedata` with two columns: 1) log-transformed expression values ​​of "ENSG00000002549" and 2) group values ​​from the "dex" variable. Call the columns "ex" and "group", respectively .

Hypothesis H0: Dexamethasone is not related to gene expression, H1: Dexamethasone is related to gene expression. According to Figure 1 and Figure 2, the pvalue of this gene is as high as 0.9981. Although the variable is significant, the Pvalue in Ttest is much greater than 0.05, so accept the null hypothesis that dex has nothing to do with gene expression. After performing the T test on all genes, select the mean difference and Pvalue value for comparison. The results are shown in Figure 3. Looking at all the Ttest results, most of the Pvalue ​​are far greater than 0.05. Next we have to do Differential Expression Analysis using DESeq2.

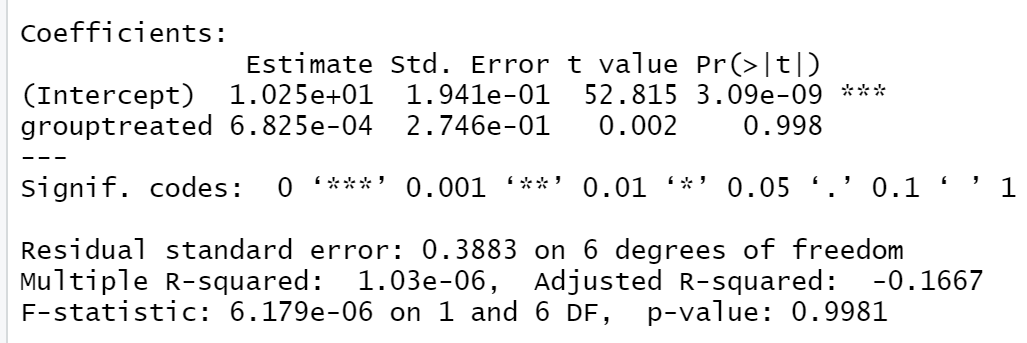


Figure1. Liner regression of ENSG00000002549

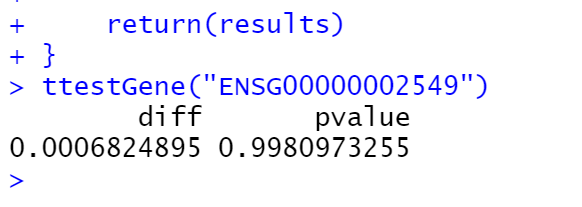


Figure2. Ttest of ENSG00000002549

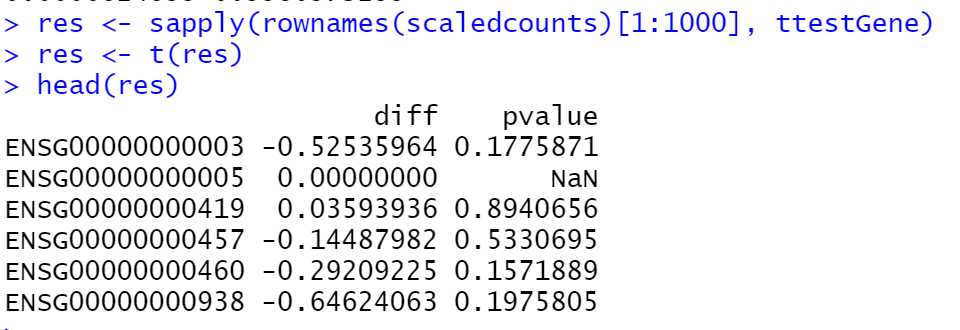


Figure3. Ttest result of whole genes

Now, I will use DESeq2 package for differential expression analysis of the airway data set to find differentially expressed genes between untreated and treated samples. I first loaded DESeq2 and set up the data to be compatible with DESeq by using the function DESeqDataSet().

The function DESeqDataSet includes an argument called design which asks for a formula that expresses how the counts for each gene depends on the variables in colData. In this case I choosed variables cell and dex because care about the cell line and which samples are treated with dexamethasone versus which samples are untreated controls.

I set control group reference level for comparison in our differential expression analysis. I run the differential expression analysis steps through the function DESeq(). I generated a results table with log2 fold changes, p values and adjusted p values for each gene. The log2 fold change and the Wald test p value is based on the last variable in the design formula, in this case variable dex. Afterward, I added a column showing whether the significance is TRUE or FALSE based on cutoff padj < 0.01. Then, I used the filter() function from dplyr to filter out results based on padj < 0.01, and write this to a csv file using write\_csv() function from readr. Therefore results will show which genes are differentially expressed (Figure4.).

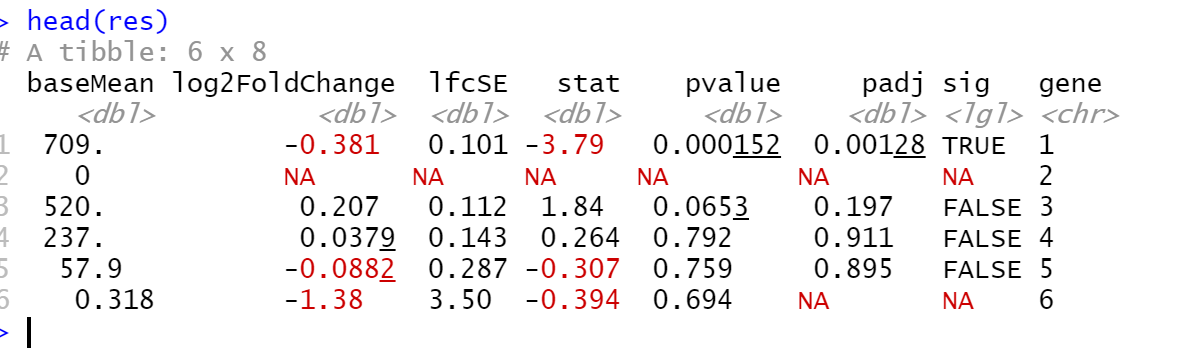


Figure4. DESeq2 result for differential expression

In DESeq2, the function ggplot2 generates a Volcano Plot (Figure5) commonly used to visualize the differential expression results. The plot shows the log2 fold changes attributable to a given variable over the mean of normalized counts for all the samples in the DESeqDataSet. Points represent genes and will be colored red if the adjusted p value is less than 0.1. Points which fall out of the figure5 are plotted as open triangles pointing either up or down. Volcano map can conveniently and intuitively display the distribution of gene differential expression between two samples. Usually the abscissa is expressed by log2 (fold change), the genes with the greater difference are distributed at both ends, the ordinate is represented by -log10 (pvalue), and the negative logarithm of the P value of the T test significance. Generally, the larger the difference multiple is, the more significant the gene T test is, so the values in the upper left corner and the upper right corner are often concerned.

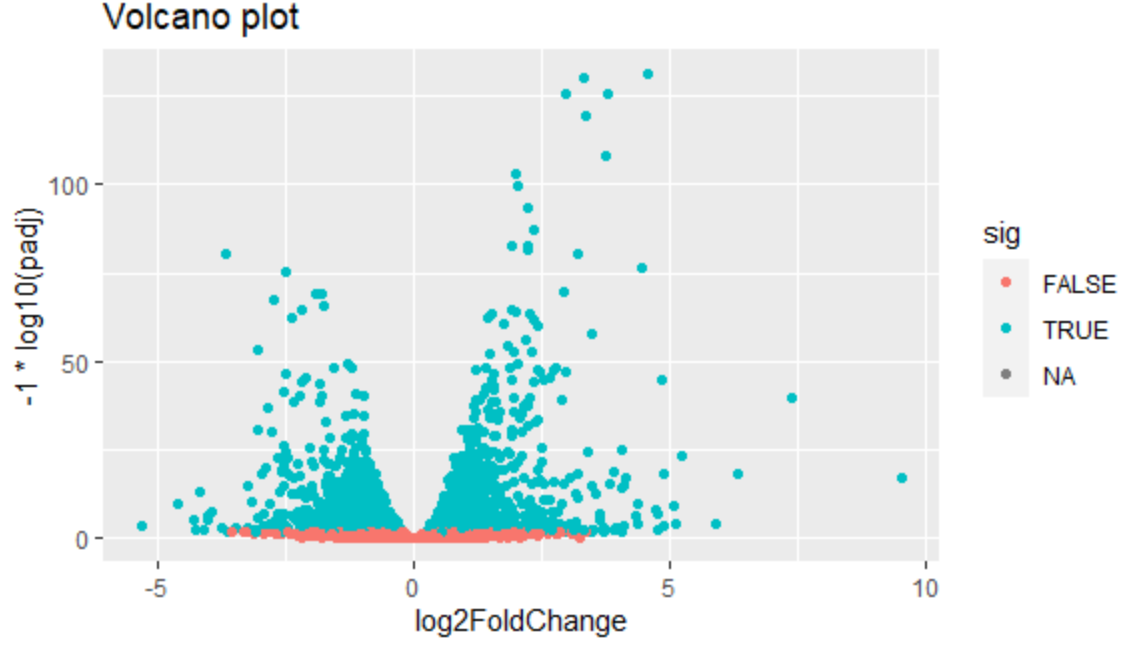


Figure5. Volcano plot of differential genes expression

**Gene set enrichment analysis**

Gene set enrichment analysis (GSEA) is a method to identify classes of genes that are over-represented in a large set of genes. This is performed by comparing the input gene set with annotated gene sets from online functional databases such as Gene Ontology (GO) and KEGG. This is a common step in bioinformatics as it aids with the biological interpretation of results.

Keeping only the set of genes that showed statistically-significant change in expression between conditions. Our threshold is be a false discovery rate (FDR) of 0.1 (i.e. no more than 10% chance that the observed change in expression is due to chance). Not all the genes in the results from DESeq2 were assigned p-values so I filtered out the genes without p-values followed by storing the significant genes separately (Figure6).

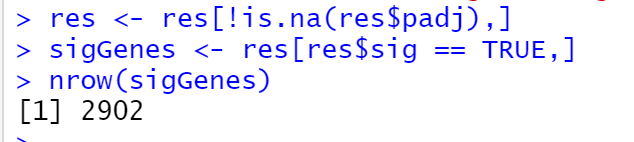


Figure6. Significant Genes result

I found the list of all available databases from package and set up list with databases of interest (Figure7). I used the Gene Ontology (GO) databases: GO Biological Process, GO Molecular Function, and GO Cellular Component.

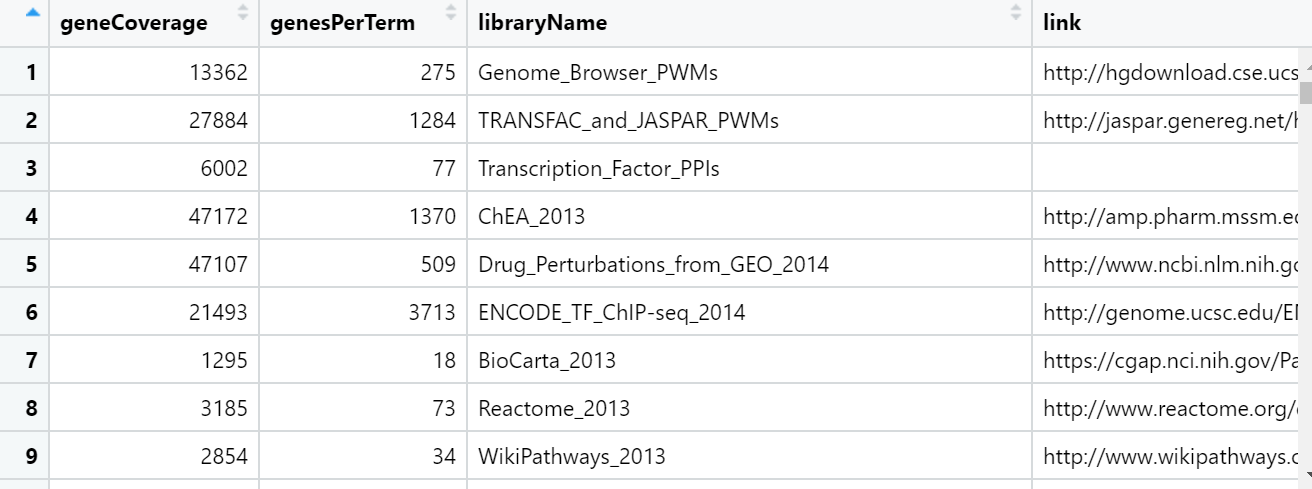
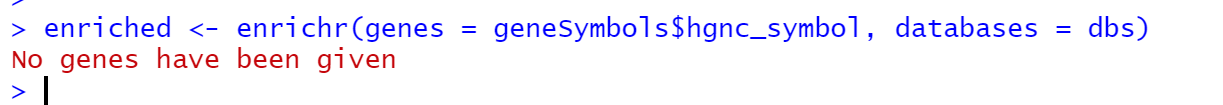


Figure7. Enrichment available databases

 The genes in this dataset use Ensembl indentifiers, while enrichR expects gene symbols. We'll use the biomaRt package to map our Ensembl IDs to gene symbols.



No genes can be compared.

**Conclusion**

Through the T test of a single gene, the difference in mean is very small relative to the variance, hence the large p-value. With control experiments and filtering non-significant cells, it is found that the significant cells are low. There is no significant relationship between genes and dexamethasone. Gene set enrichment analysis (GSEA) is a method to identify classes of genes that are over-represented in a large set of genes. After database comparison, there are no similar genes.

Overall, this analysis has no data showing that Dex can relate to the RNA sequence of asthma patients.