IL4R highly expressed in allergic asthma?

INTRODUCTION

Asthma is a common respiratory disease and is considered the third leading cause of hospitalization. It is the chronic inflammatory disease of the airways which results in hyper-responsiveness and air-flow limitation. Asthma includes respiratory symptoms as wheeze, cough and shortness of breath. At molecular level, asthma includes activation and differentiation of B lymphocytes toward the production of IgE. IL4R role is important in B cell development, B cell proliferation and isotype secretion and the ability of B cells to regulate the immune system. I will analyze B cell expression profile RNAseq data (**GEO Accession GSE52742**) from the research paper **Genome-wide expression profiling of B Lymphocytes reveals IL4R increase in allergic asthma**. For this study they collected peripheral blood $CD19^+$ using immunomagnetic procedures. The generated libraries were sequenced on the Illumina HiSeq 2000 platform.

My goal will be to asses expression differences in B $CD19^+$ Lymphocytes from house dust mite patients and healthy controls and to know B-cell-specific targets with therapeutic potentials. There are 6 single end reads from 6 samples: three control and three house dust mites allergic. I will download the data by using **sratoolkit**. The next step will be aligning the reads to the human reference genome index by using **Kallisto** with the k-mer size less than the length of the fastq sequences. The output of Kallisto will be further analyzed by using **DESeq2** in R to know which genes expressed differentially across the disease and normal conditions. I will use R packages to plot the abundance.tsv output files for each sample to see if there are differences between the libraries. In order to know the pattern of clustering of the genes, I will perform principal component analysis (PCA). Finally, I will prepare a heatmap with differentially expressed and statistically significant genes.

Materials:

The six, single end reads are first aligned to human reference genome index of size 21kmer by using kallisto. The ouput of kallisto are then analysed by using a Bioconductor package DESeq2. DESeq2 performs differential expression analysis based on negative binomial distribution (Love, Huber, & Anders, 2014).The null hypothesis is tested that logarithmic fold change (LFC) between control and house dust mite for a gene's expression is exactly zero.

Since we know that a list of transcript id based on p-value indicating highly significant might not be the interesting candidates, the produced list is again sorted based on LFC to get the interesting candidates. A more quantitative analysis is performed using DESeq2 shrinkage. For all the data analysis part R studio is used (Ross & Robert, 1996). To run DSeq2 I need to create two different matrixes. First matrix will include all abundance.tsv files where each column in the dataframe will correspond to one of the RNA-seq runs and each row will correspond to a target ID. This matrix will be called **asthma_d**. Another matrix will include SRA run ID and run description. This matrix will be named annotation.

DATA ANALYIS

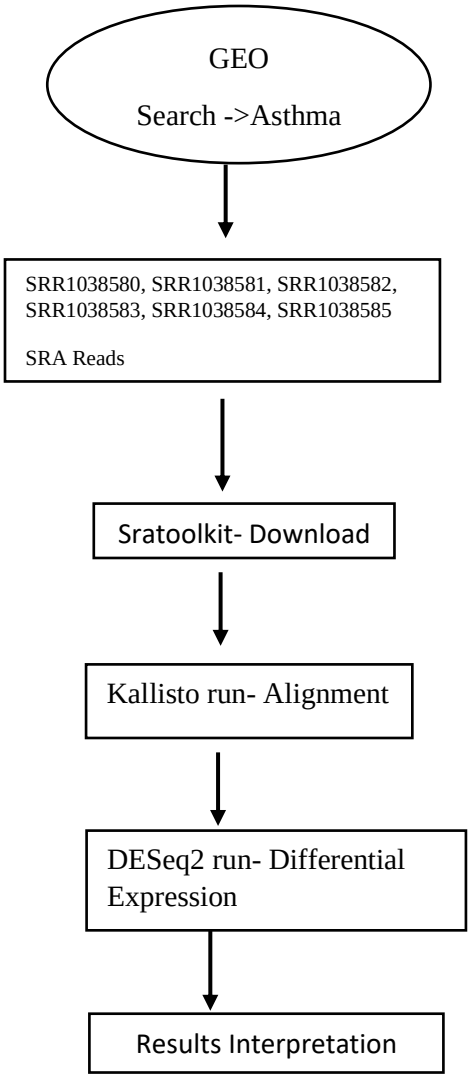


Fig 1: Overview of analysis

Step 1:

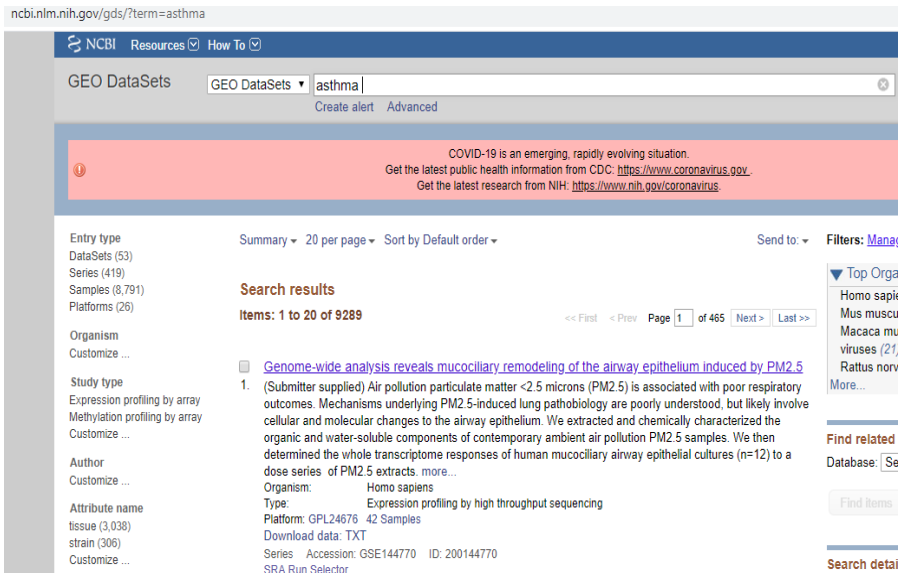


Fig 2**: GEO Search Database**

Step 2:

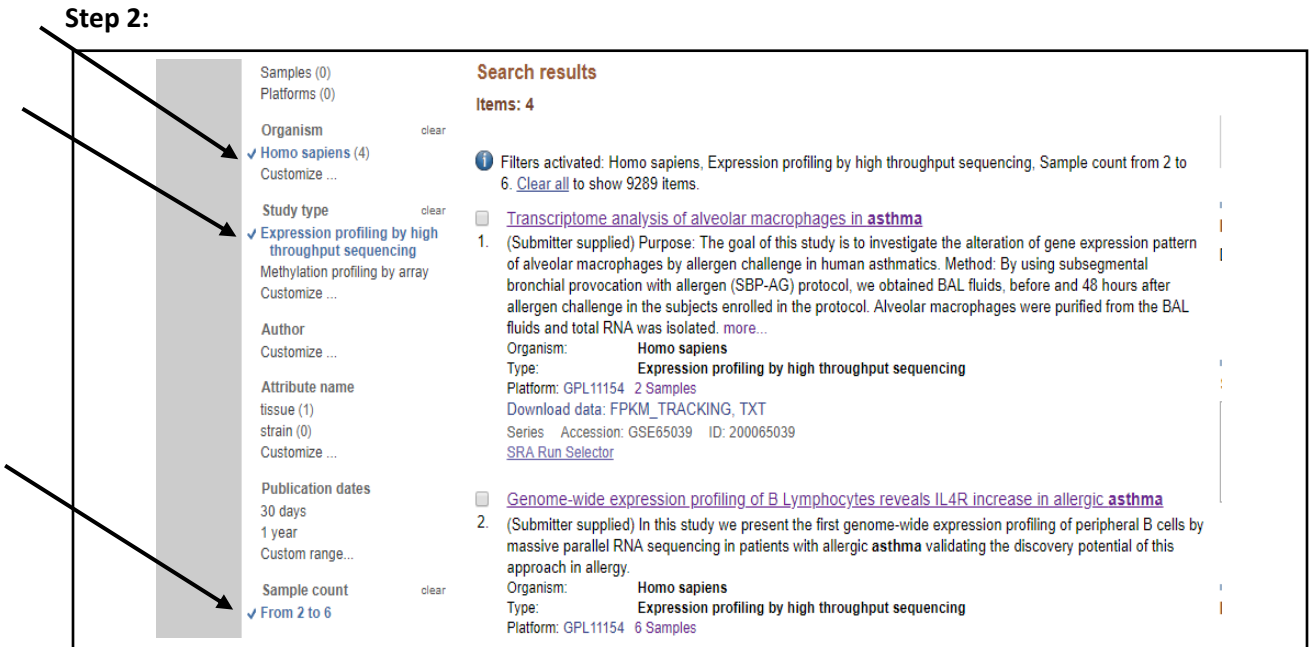


Fig 3: **Customize the search with Organism to Homo sapiens, Study Type to Expression profiling by high throughput sequencing and Sample count to 2 to 6.**

Step 3

	Run	BioSample	Bases	Bytes	Experiment	GEO_Accession	phenotype	Sample Name
1	SRR1038580	SAMN02422536	2.13 G	1.04 Gb	SRX383986	GSM1275173	Control	GSM1275173
2	SRR1038581	SAMN02422534	1.96 G	1,010.23 Mb	SRX383987	GSM1275174	Control	GSM1275174
3	SRR1038582	SAMN02422531	2.65 G	1.30 Gb	SRX383988	GSM1275175	Control	GSM1275175
4	SRR1038583	SAMN02422533	2.17 G	1.07 Gb	SRX383989	GSM1275176	House dust mite allergic	GSM1275176
5	SRR1038584	SAMN02422535	2.23 G	1.12 Gb	SRX383990	GSM1275177	House dust mite allergic	GSM1275177

Legend

Experiment	Library Name	Platform	Strategy	Source	Selection	Layout	Action
SRX383986		Illumina	RNA-Seq	TRANSCRIPTOMIC	cDNA	SINGLE	BLAST

Fig 4**: RNA-Seq Layout**

```
#!/bin/bash
#SBATCH -J SRAdownloadsamples      # Name of the job
#SBATCH -N 1                   # Number of nodes
#SBATCH -n 1                   # Number of cores (processors) per node
#SBATCH -t 24:00:00            # Runtime in HH:MM:SS
#SBATCH --mem=4G      # Memory requested in MB (see also --mem-per-cpu)
#SBATCH -o SRAdownload_%j.out  # File to write STDOUT, %j=jobid
#SBATCH -e SRAdownload_%j.err  # File to wrote STDERR, %j=jobid
#SBATCH --mail-type=ALL            # Send email when job starts, ends, fails, etc
#SBATCH --mail-user=bsharmachapagain@uh.edu

#Load the "sratoolkit" module
module load sratoolkit

# Make sure in the right directory to download
cd /project/meisel/bsharmac/Final_Project/

# Use fastq-dump, from the sratoolkit, to download SRA files


#fastq-dump --gzip SRR1038580
fastq-dump --gzip SRR1038581
fastq-dump --gzip SRR1038582
fastq-dump --gzip SRR1038583
fastq-dump --gzip SRR1038584
fastq-dump --gzip SRR1038585
```

Fig 5**: Six fastq data download with sratoolkit**

```
#!/bin/bash
#SBATCH -J ALIGNMENT1              # Name of the job
#SBATCH -N 1                   # Number of nodes
#SBATCH -n 1                   # Number of cores (processors) per node
#SBATCH -t 24:00:00            # Runtime in HH:MM:SS
#SBATCH --mem=4G               # Memory requested in MB (see also --mem-per-cpu)
#SBATCH -o GSE37704_%j.out     # File to write STDOUT, %j=jobid
#SBATCH -e GSE37704_%j.err     # File to wrote STDERR, %j=jobid
#SBATCH --mail-type=ALL            # Send email when job starts, ends, fails, etc
#SBATCH --mail-user=bsharmachapagain@uh.edu
#SBATCH --mem=4G
#Load the "kallisto" module
module load kallisto

# Move to your directory
cd /project/meisel/bsharmac/Final_Project



# Use kallisto to quantify gene expression


kallisto quant -i Homo_sapiens.GRCh38.cdna.all.release-94_k21.idx -o SRR1038580 --single -l  200 -s 20  SRR1038580.fastq.gz
kallisto quant -i Homo_sapiens.GRCh38.cdna.all.release-94_k21.idx -o SRR1038581 --single -l  200 -s 20  SRR1038581.fastq.gz
kallisto quant -i Homo_sapiens.GRCh38.cdna.all.release-94_k21.idx -o SRR1038582 --single -l  200 -s 20  SRR1038582.fastq.gz
kallisto quant -i Homo_sapiens.GRCh38.cdna.all.release-94_k21.idx -o SRR1038583 --single -l  200 -s 20  SRR1038583.fastq.gz
kallisto quant -i Homo_sapiens.GRCh38.cdna.all.release-94_k21.idx -o SRR1038584 --single -l  200 -s 20  SRR1038584.fastq.gz
kallisto quant -i Homo_sapiens.GRCh38.cdna.all.release-94_k21.idx -o SRR1038585 --single -l  200 -s 20  SRR1038585.fastq.gz
```

Fig 6: **Aligning individual single reads to Homo sapiens reference genome index.** Here, the index size is 21.

```
library(readr)
SRR1038580 <- read.delim("C:/Users/Rahul/Desktop/class_Data/final_project/SRR1038580/abundance.tsv")
SRR1038581 <- read.delim("C:/Users/Rahul/Desktop/class_Data/final_project/SRR1038581/abundance.tsv")
SRR1038582 <- read.delim("C:/Users/Rahul/Desktop/class_Data/final_project/SRR1038582/abundance.tsv")
SRR1038583 <- read.delim("C:/Users/Rahul/Desktop/class_Data/final_project/SRR1038583/abundance.tsv")
SRR1038584 <- read.delim("C:/Users/Rahul/Desktop/class_Data/final_project/SRR1038584/abundance.tsv")
SRR1038585 <- read.delim("C:/Users/Rahul/Desktop/class_Data/final_project/SRR1038585/abundance.tsv")

####check if gene id matched across all abundance.tsv
all(rownames(SRR1038580)%in%rownames(SRR1038581))
all(rownames(SRR1038580)%in%rownames(SRR1038582))
all(rownames(SRR1038580)%in%rownames(SRR1038583))
all(rownames(SRR1038580)%in%rownames(SRR1038584))
all(rownames(SRR1038580)%in%rownames(SRR1038585))
```

Fig 7: All the six abundance.tsv files loaded in R and gene id is checked across all abundance.tsv

8. a. ##creating dataframe1 **asthma_d**###

```
asthma_d <- data.frame(Control_1 = round(SRR1038580$est_counts),Control_2 =
round(SRR1038581$est_counts),Control_3 = round(SRR1038582$est_counts), HDMA_1=
round(SRR1038583$est_counts),HDMA_2 = round(SRR1038584$est_counts),HDMA_3 =
round(SRR1038585$est_counts))
```

Fig 8.a**: Dataframe with estimated-counts value for each sample**

```
rownames(asthma_d) <- SRR1038580$target_id
```

8.b

```
annotation <- data.frame( Condition=c("Control","Control","Control","HDMA","HDMA","HDMA"))

rownames(annotation) <- colnames(asthma_d)
```

Fig 8.b: **Dataframe with sample condition**

9.a. Creating object for DESeq2

```
asthma_d_seq <- DESeqDataSetFromMatrix(countData = asthma_d, colData = annotation,design =
~Condition)

#####runing deseq

asthma_d_d_seq <- DESeq(asthma_d_seq)

##########################
```

9.b. Summary result

```
asthma_d_d_result <- results(asthma_d_d_seq)

##########summary of the result

summary(asthma_d_d_result)
```

Fig 9: DESeq2

10.a Shrunken Fold Change

```
resultshrink <- lfcShrink(asthma_d_d_seq,coef="Condition_HDMA_vs_Control",type="apeglm")
```

10.b. Correlation between Shrunk and Unshrunk

```
plot(asthma_d_d_result$log2FoldChange,resultshrink$log2FoldChange,xlab ="Unshrunk
LFC",ylab="Shrunk LFC",main="Unshrunk vs Shrunk")
```

10.c. Shrunk fold change vs Shrunk baseMean

```
plot(resultshrink$baseMean,resultshrink$log2FoldChange,xlab="basemean",ylab="Shrunk logfold
change",main="Shrunk logfold vs baseMean Expression",pch=1)
```

10. d. Unshrunk fold change vs unshrunk baseMean

```
plot(asthma_d_d_result$baseMean,asthma_d_d_result$log2FoldChange,xlab="basemean",ylab="un
shrunk logfold change",main="Unshrunk logfold vs baseMean Expression",pch=2)
```

Fig 10: Fold Change visualization

```
res <- res[ ! is.na(res$pvalue), ]

asthma_d_d_result_sig <- asthma_d_d_result_sig[which(asthma_d_d_result_sig$pvalue < 0.01),]

log_sig <- asthma_d_d_result_sig[ order(asthma_d_d_result_sig$log2FoldChange),]

head(log_sig)
```

Fig 11: Top differentially expressed genes

```
hist(asthma_d_d_result_sig$pvalue[res$baseMean > 1], breaks=20, col="blue",
border="white",xlab="pvalue",main="Histogram of pvalue")
```

Fig 12: Histogram of pvalue

```
plotMA(asthma_d_d_result[asthma_d_d_result$baseMean >1 ,],colNonSig = "gray", colSig =
"red3",cex=0.9)
```

Fig 13: MA plot

```
14. a. plotCounts(asthma_d_d_seq,gene=which.min(asthma_d_d_result$padj),intgroup = "Condition")
14. b. plotCounts(asthma_d_d_seq,gene=which.max(asthma_d_d_result$pvalue),intgroup = "Condition")
```

Fig 14. a: Gene with minimum p-value. Fig 14.b: Gene with maximum p-value

```
vsd <- vst(asthma_d_d_seq,blind=FALSE)
plotPCA(vsd,intgroup=c("Condition"))
```

Fig 15: PCA plot

16. a. Heatmap

```
select <- order(rowMeans(counts(asthma_d_d_seq,normalized=TRUE)), decreasing=TRUE)[1:20]

library("pheatmap")

df <- as.data.frame(colData(asthma_d_d_seq)[("Condition")])

pheatmap(assay(vsd)[select,],cluster_rows = FALSE,show_rownames =FALSE,cluster_cols =
FALSE,annotation_col = df
```

16. b. Heatmap Clustering

```
sampleDists <- dist(t(assay(vsd)))

sampleDistMatrix <- as.matrix(sampleDists)

rownames(sampleDistMatrix) <- paste(vsd$Condition,sep ="-")

colnames(sampleDistMatrix) <- NULL

colors <- colorRampPalette(rev(brewer.pal(7,"Greens")))(255)

pheatmap(sampleDistMatrix, clustering_distance_rows = sampleDists, clustering_distance_cols =
sampleDists, col=colors)
```

Fig 16: Heatmap

RESULTS

LFC > 0 (up)	117, 0.09%
LFC < 0 (down)	85, 0.065%
Outliers	411, 0.32%
Low counts	86168, 66%

Table 1: Counts 0f genes LFC <0 , LFC>0 , Number of Outliers and Low counts for adjusted p-value < 0.1. Here, LFC >0 means expression is significantly higher in HDMA and LFC < 0 means expressions is significantly higher in Control. Hence, 117 genes are highly expressed in HDMA and 85 genes are highly expressed in Control. There are 411 of outliers and 86168 of the low counts.

LFC > 0 (up)	62, 0.048%
LFC < 0 (down)	82, 0.063%
Outliers	411, 0.32%
Low counts	88630, 68%

Table2: Counts f genes with LFC <0 , LFC>0 , Number of Outliers and Low counts for adjusted p-value < 0.1. Here, LFC >0 means expression is significantly higher in HDMA and LFC < 0 means expressions is significantly Control. Hence, 62 genes are highly expressed in HDMA and 82 genes are Control. There are 411 of outliers and 88630 of the low counts

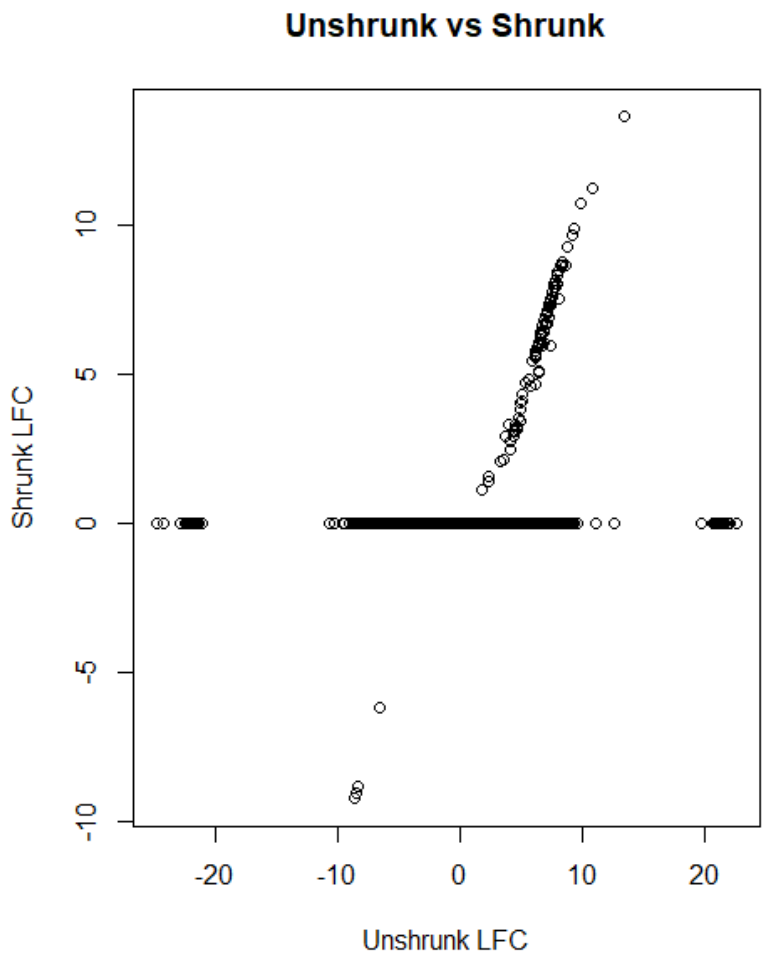
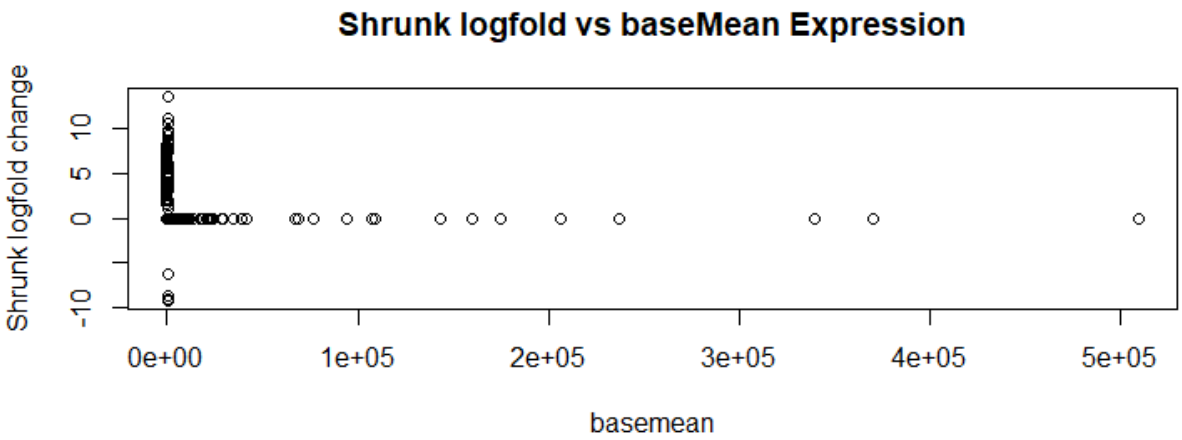


Fig 17: Correlation plot of shrunken and unshrunken LFC. It reduces variance in LFC estimates.



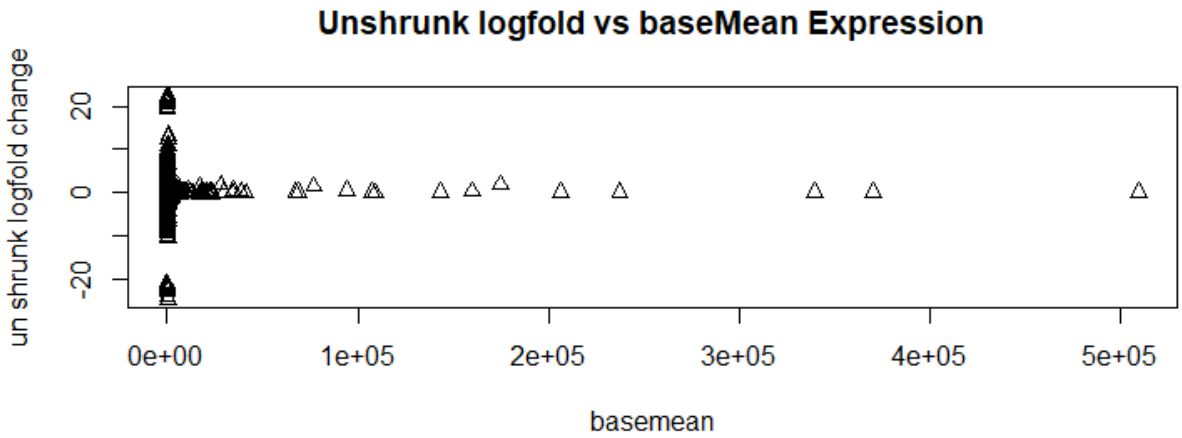


Fig 18: Top plot represents correlation between shrunk logfold change versus baseMean expression. Bottom plot represents correlation between unshrunk logfold change versus basemean expression.

	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj
	<numeric>	<numeric>	<numeric>	<numeric>	<numeric>	<numeric>
ENST00000472628.1	114.174257660048	-6.76081388890982	0.949137340897861	-7.12311442990353	1.05515145397767e-12	4.58336688578821e-08
ENST00000358075.10	73.7289029622492	-9.53458208297152	1.44879987874436	-6.58102076267078	4.67229325900175e-11	7.10537418758615e-07
ENST00000500450.6	77.1172970754547	9.89892077148912	1.50583176940133	6.57372289032299	4.90725230506894e-11	7.10537418758615e-07
ENST00000454264.6	125.430597415236	-10.3011039175684	1.64523664812757	-6.26116852508237	3.82103309867362e-10	3.3195607148037e-06
ENST00000371160.5	65.769025694586	-9.36951133653161	1.49025515569869	-6.28718599006544	3.23271999587833e-10	3.3195607148037e-06
ENST00000381461.6	41.0314471918625	8.98854707679602	1.45804411940417	6.1647977294879	7.05732932004278e-10	5.10927118340031e-06

Fig 16: Top differentially expressed genes.

Transcript ID	Function	Gene
ENST00000383251.6	Major histocompatibility complex, class II, DQ (HLA-DQA1-208) Located at ChromosomeCHR_HSCHR6_MHC_QBL_CTG:32,570,792-32,577,0071 forward strand	ENSG00000206305.12
ENST00000454264.6	MDM4 regulator of p53 (MDM4-207) Located at Chromosome 1: 204,516,379-204,558,119 forward strand.	ENSG00000198625.13
ENST00000358075.10	Magnesium transporter 1 (MAGT1-201) Located at: Chromosome X: 77,826,364-77,895,593 reverse strand.	ENSG00000102158.19
ENST00000371160.5	Stromal antigen 2 (STAG2); Located at Chromosome X: 123,961,314-124,102,656 forward strand.	ENSG00000101972.18
ENST00000337147.11	Ezrin (EZR-201) Located at Chromosome 6: 158,765,741-158,818,227 reverse strand.	ENSG00000092820.18
ENST00000396401.7	Transcription elongation factor A1 (TCEA1-201) Located at: Chromosome 8: 53,966,552-54,022,456 reverse strand.	ENSG00000187735

Table 2: Top five differentially transcripts with its corresponding gene ID.

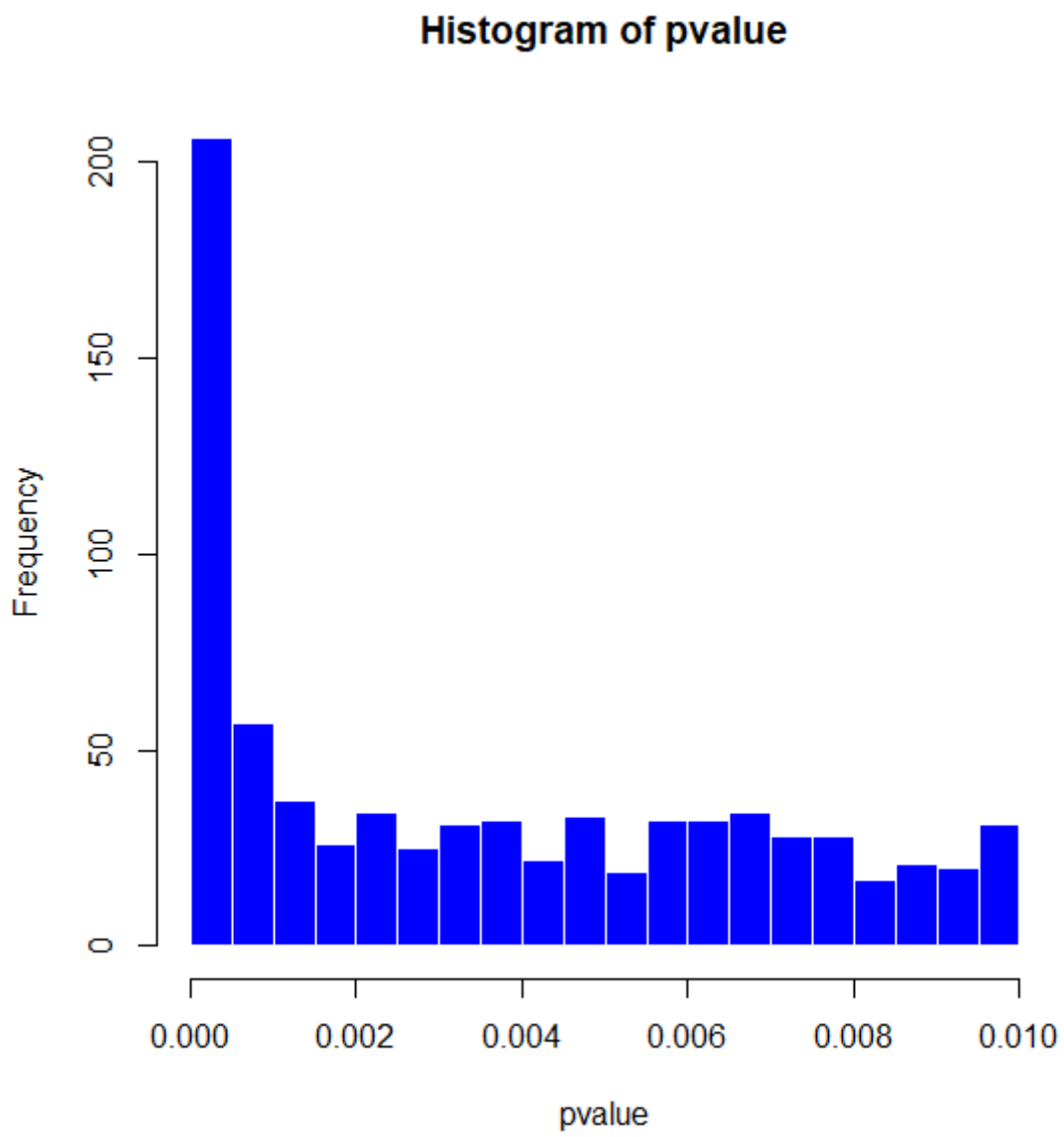
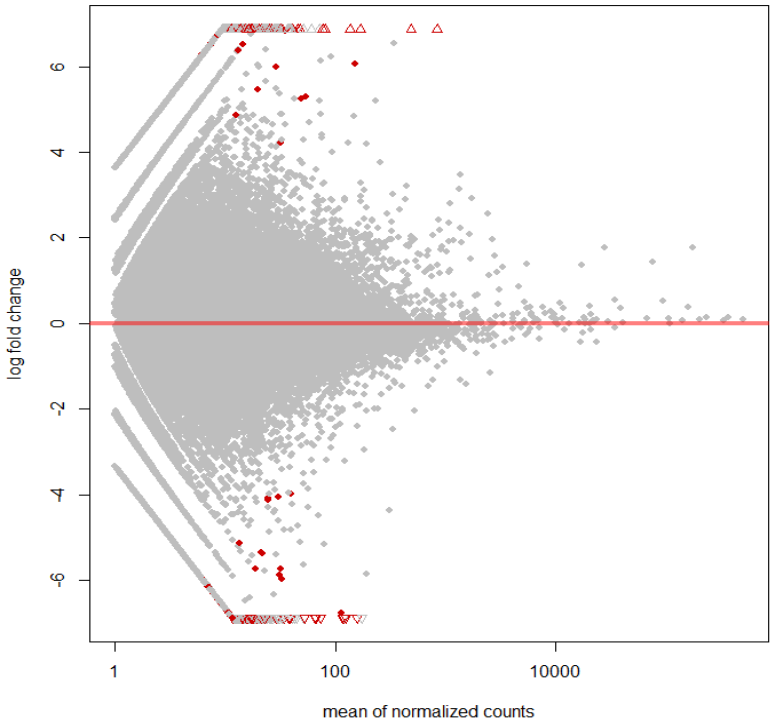
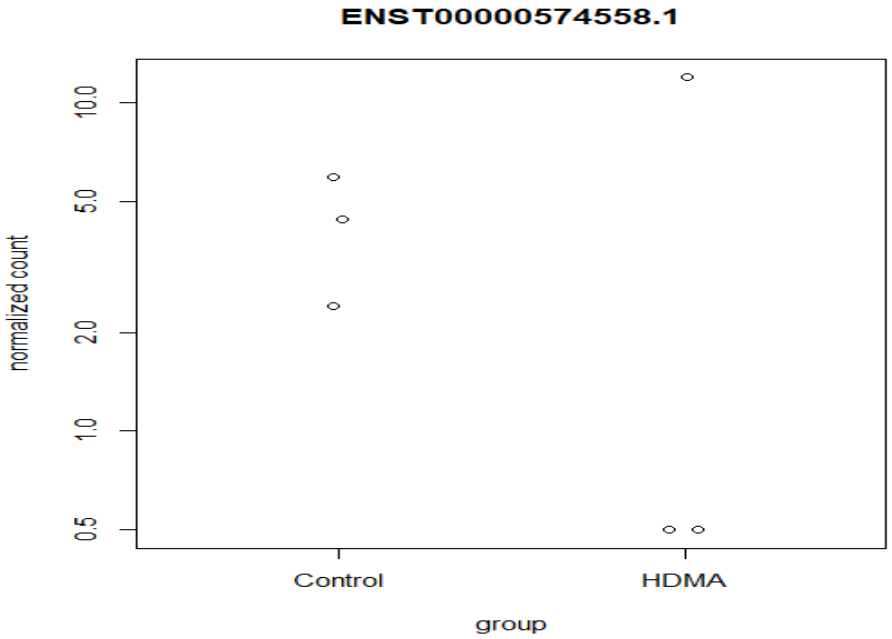


Fig 19: Histogram of pvalue. Here, the genes with small counts are excluded. The counts with pvalue less than 0.01 are about 49.



<u>Fig 20</u>: <u>MA plot where points below 0 are for down regulated and points above 0 are for up regulated</u>. The significant points are shown in red color and nonsignificant is shown in grey. The points which fall out of the window are plotted as open triangles pointing either up or down. Here, adjusted p value is less than 0.1.

21. a.



21.b.

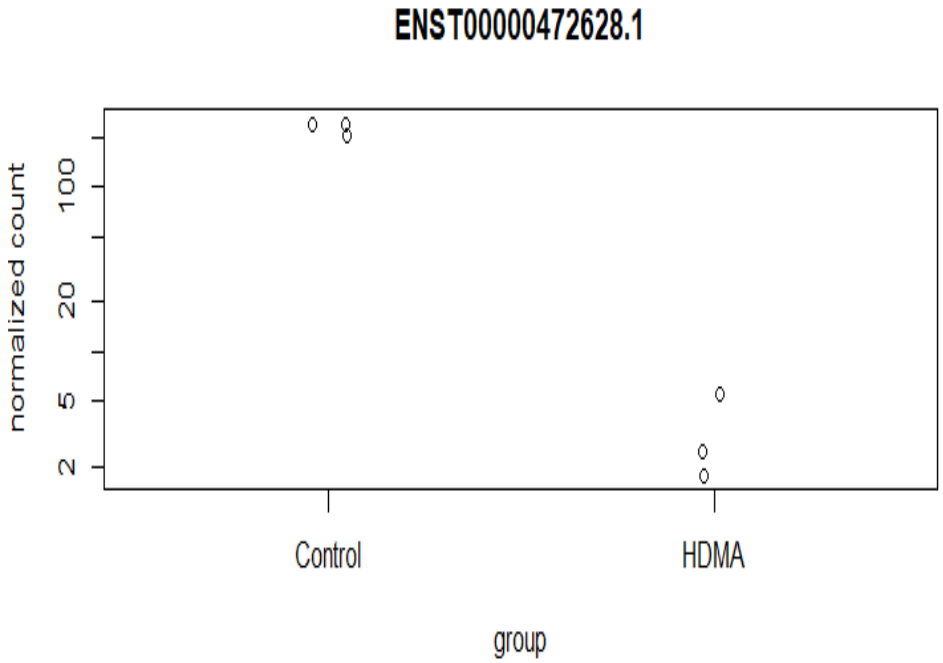


Fig 21: Transcript with maximum and minimum p-value. In Fig.18.a there are large variations within the group in HDMA**.** While in Fig. 18.b there are large variations between the groups. While there are less variations within the groups.

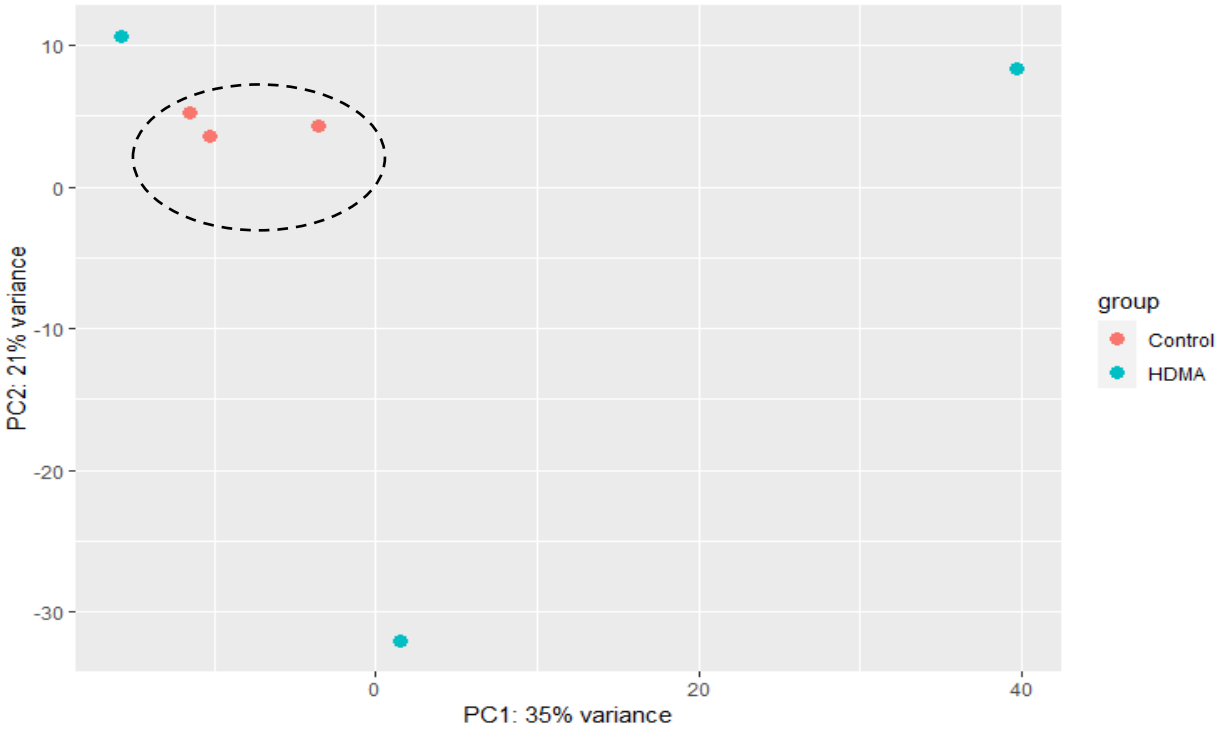


Fig 22: PCA Plot. Principal component analysis (PCA) represents the variation in gene expression across samples control and house dust mite allergic (HDMA). The samples are projected in 2D plane so that they separate out in two directions. Here, controls are clustered together indicating similar gene expression while HDMA are far apart indicating no similarity within them. Here, 35% variance is explained by PC1 and 21% by PC2.The two groups are control in red and HDMA in light blue.

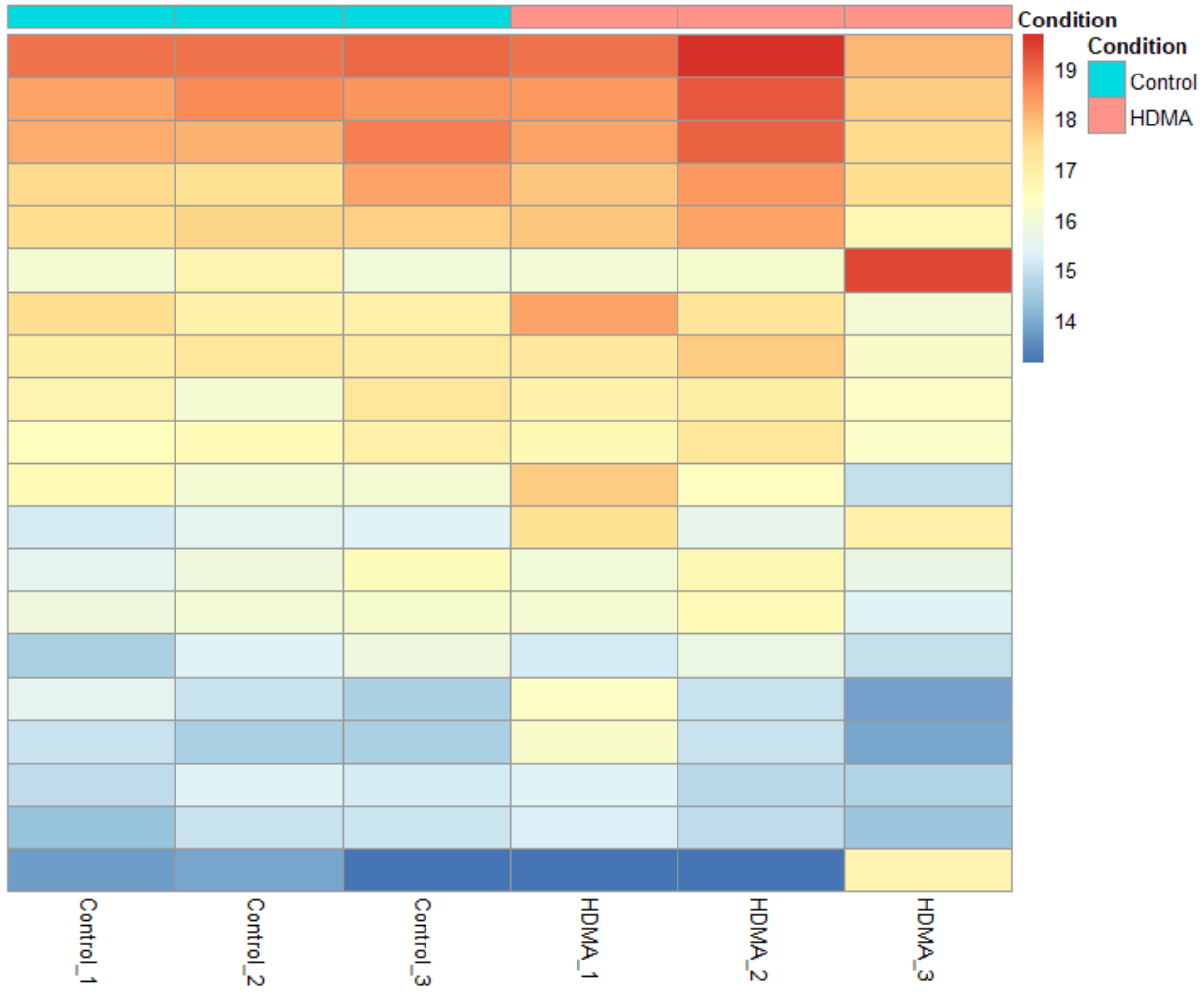


Fig 23: Heatmap. Samples control and house dust mite allergic (HDMA) represented by each columns and rows represent measurements from different genes. Here, red indicates higher expression of genes and blue and yellow represents low expression of genes.

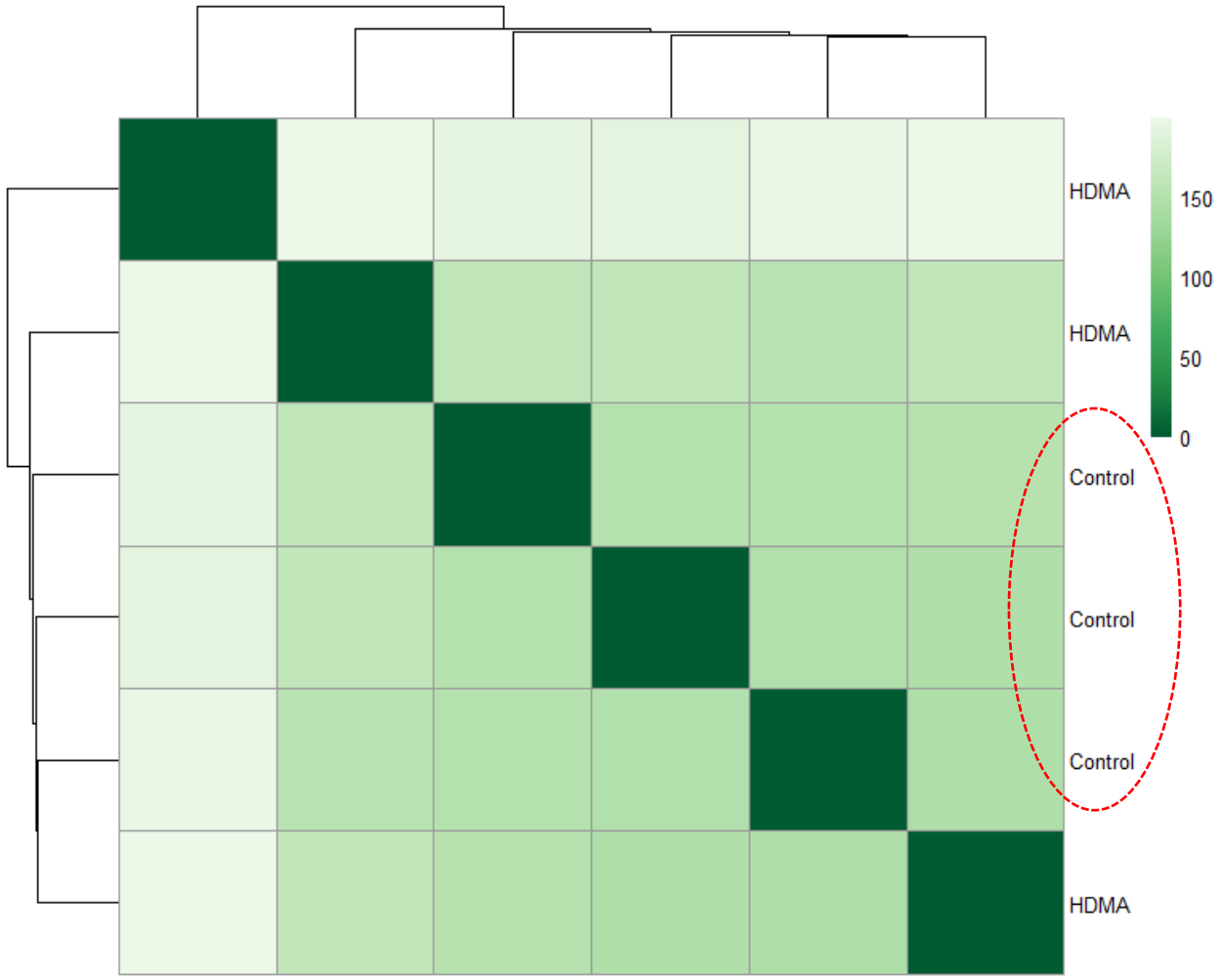


Fig 24: Heatmap Clustering. Here, control samples are clustered together while the HDMA samples they are not clustered together. This plot also supports that the HDMA samples are not similar.

Discussion:

It is important to understand the pathway involved in activation and differentiation of B lymphocytes towards the production of IgE (Barnes & Woolcock, 1998). In this process, top five differentially expressed genes are presented in Table 2. These genes are previously studied and are reported as a molecule involve in asthma pathway. PCA plot in Fig 19 indicates that there are large variations among the house dust mite allergic patient (117 genes) compared to the control group as shown. This could be due to the genetic variations or due to the experimental error. The expression value as indicated by logfold change is higher in-house dust mite allergic patient compared to the control. This indicates a clear difference between house dust mite allergic patient group and normal group. Out of 129606 genes there are 2254 genes with pvalue less than 0.05. To further understand the functions underlying these genes, pathway enrichment analysis can be performed using the **Gene Set Enrichment Analysis** method applied to the Molecular Signature Database (MSigDB). Also, this analysis includes no technical replicates. So, to make such kind of analysis more significant, the sample size can be increased considering at least two technical replicates.

Conclusion:

Within top 10 differentially expressed gens, I am not able to identify IL4R as a highly expressed, but here, I am reporting two key genes Human leukocytes Antigens (HLA) and Ezrin which are also associated with numerous immune response and may be a potential biomarkers of asthma control. HLA is located on chromosome 6p21 are among the most polymorphic genes which is associated with other aspects of immune response (Mahdi et al., 2018).Ezrin downregulation is associated with IL-13-induced epithelial damage(Jia et al., 2019). Another key finding is Transcription Elongation Factor A1 (TCEA1-201). As mention in the paper **Transcription factors in asthma: are transcription factors a new target for asthma therapy?** TCEA1 could be a novel way of treating asthma by connecting link transcription factor inhibitors/activators to a cell type specific delivery system. Also, MAGT1-201 could be one of the marker to consider for treatment of asthma Mg2+ therapies ((Trapani, Shomer, & Rajcan-Separovic, 2015). Hence, DESeq analysis of RNA-sed data set revealed Ensembl transcripts to be differentially expressed between control (nonallergic) and house dust mite allergic (allergic) groups. Thus, further detailed analysis involve mapping differentially expressed transcript id to IL pathway and understanding function.

References

Barnes, P. J., & Woolcock, A. J. (1998). Difficult asthma. *European Respiratory Journal*, *12*(5), 1209–1218. https://doi.org/10.1183/09031936.98.12051209

Jia, M., Yan, X., Jiang, X., Wu, Y., Xu, J., Meng, Y., … Yao, X. (2019). Ezrin, a membrane cytoskeleton cross-linker protein, as a marker of epithelial damage in asthma. *American Journal of Respiratory and Critical Care Medicine*, *199*(4), 496–507. https://doi.org/10.1164/rccm.201802-0373OC

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), 1–21. https://doi.org/10.1186/s13059-014-0550-8

Mahdi, B. M., Al-Hadithi, A. T. R., Raouf, H., Zalzala, H. H., Abid, L. A., & Nehad, Z. (2018). Effect of HLA on development of asthma. *Annals of Medicine and Surgery*, *36*(September), 118–121. https://doi.org/10.1016/j.amsu.2018.10.003

Ross, I., & Robert, G. (1996). R: a language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, Vol. 5, pp. 299–314.

Trapani, V., Shomer, N., & Rajcan-Separovic, E. (2015). The role of MAGT1 in genetic syndromes. *Magnesium Research*, *28*(2), 46. https://doi.org/10.1684/mrh.2015.0381