

Developmental regulation of human cortex transcription project

Vy K Phung

Load library

```
library(tidyverse)
library(limma)
library(preprocessCore)
library(RColorBrewer)
library(org.Hs.eg.db)
library(AnnotationDbi)
library(edge)
library(sva)
library(DESeq2)
library(broom)
library(readxl)
```

Data preprocessing

```
count <- read.delim("D:/word/bioinformatics/personal project/PRJNA245228/tidy_data/count.csv")
pheno <- read.csv("D:/word/bioinformatics/personal project/PRJNA245228/sample_data/pheno_sample.csv")
head(count)
```

##	ENTREZID	ENSEMBL	SYMBOL	SRR1554534	SRR1554535	SRR1554568	SRR1554561
## 1	1	ENSG00000121410	A1BG	444	378	328	650
## 2	10	ENSG00000156006	NAT2	11	8	4	0
## 3	100	ENSG00000196839	ADA	299	658	161	229
## 4	1000	ENSG00000170558	CDH2	7384	10623	43926	11016
## 5	10000	ENSG00000117020	AKT3	6837	7391	90930	13677
## 6	10000	ENSG00000275199	AKT3	6837	7391	90930	13677
##	SRR1554567	SRR1554536	SRR1554541	SRR1554539	SRR1554538	SRR1554537	
## 1	146	114	592	295	275	518	
## 2	8	0	8	8	12	4	
## 3	225	291	382	265	354	160	
## 4	52746	3270	44244	10639	47354	52346	
## 5	36246	937	74768	18216	48565	79685	
## 6	36246	937	74768	18216	48565	79685	

```
head(pheno)
```

##	Run	age_group	age	sex
## 1	SRR1554534	adult	40.4200	male

```
## 2 SRR1554535      adult 41.5800   male
## 3 SRR1554568      fetus -0.4986   male
## 4 SRR1554561      adult 43.8800   male
## 5 SRR1554567      fetus -0.4027   male
## 6 SRR1554536      adult 44.1700 female
```

It is clear that there are some duplications of gene symbol and Entrezid, for example, in line 5,6 of count table. Therefore I will remove duplicated genes and use gene symbol as row name.

```
dup <- duplicated(count$SYMBOL)
table(dup)
```

```
## dup
## FALSE  TRUE
## 23741  4955
```

```
count_symbol <- count[!dup, -1:-2]
na <- is.na(count_symbol$SYMBOL)
count_symbol <- count_symbol[!na,]
row.names(count_symbol) <- count_symbol$SYMBOL
count_symbol <- count_symbol[, -1]
head(count_symbol)
```

```
##      SRR1554534 SRR1554535 SRR1554568 SRR1554561 SRR1554567 SRR1554536
## A1BG          444         378         328         650         146         114
## NAT2           11          8          4          0          8          0
## ADA           299         658         161         229         225         291
## CDH2          7384        10623        43926        11016        52746        3270
## AKT3          6837        7391        90930        13677        36246         937
## GAGE12F         0          0          0          0          0          0
##      SRR1554541 SRR1554539 SRR1554538 SRR1554537
## A1BG           592         295         275         518
## NAT2            8          8         12          4
## ADA           382         265         354         160
## CDH2          44244        10639        47354        52346
## AKT3          74768        18216        48565        79685
## GAGE12F         0          0          0          0
```

Data exploration

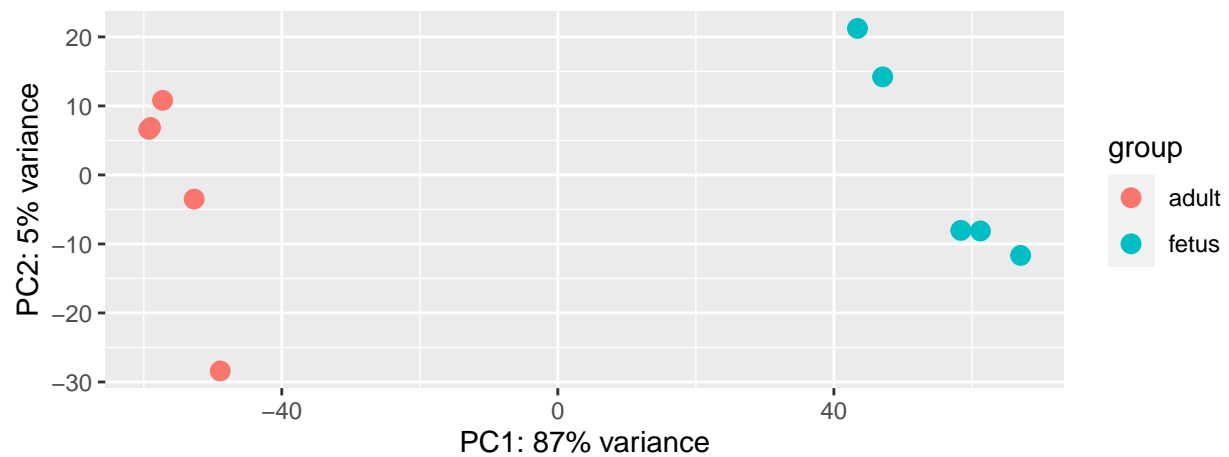
Although the target of this exploratory analysis is figuring out if there is a correlation between differential gene expression in fetus vs adult, I still plot PCA for the sex group in this data exploration to see if there is an association between sex variable (female vs male) with gene expression in fetus vs adult.

At first, I will explore unfiltered data which still has some genes' row names having 0 reads in count_symbol table by using DESeq2.

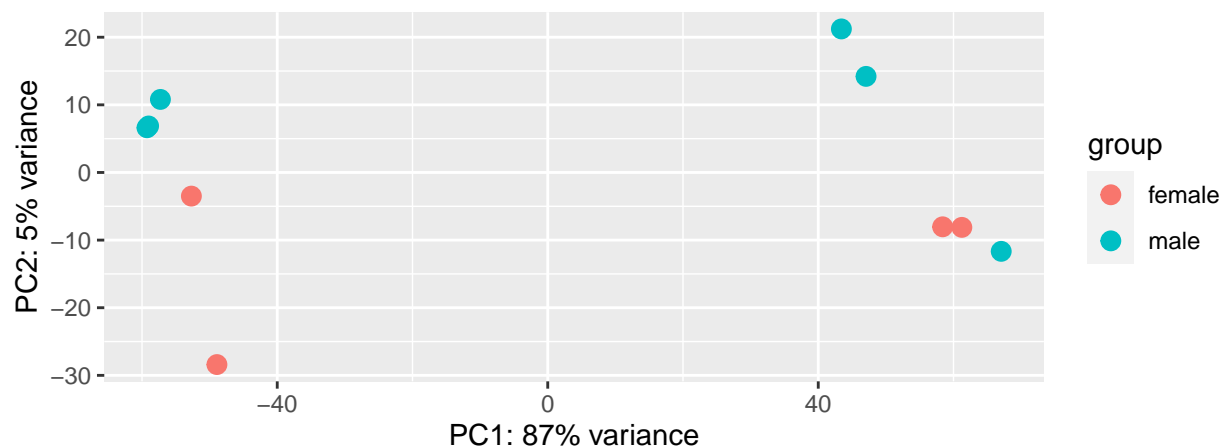
Unfiltered data

```
library(DESeq2)
edata <- DESeqDataSetFromMatrix(countData = count_symbol, colData = pheno, design = ~ age_group)

edata_tr <- rlog(edata, blind = FALSE)
plotPCA(edata_tr, intgroup = c("age_group"))
```



```
plotPCA(edata_tr, intgroup = c("sex"))
```



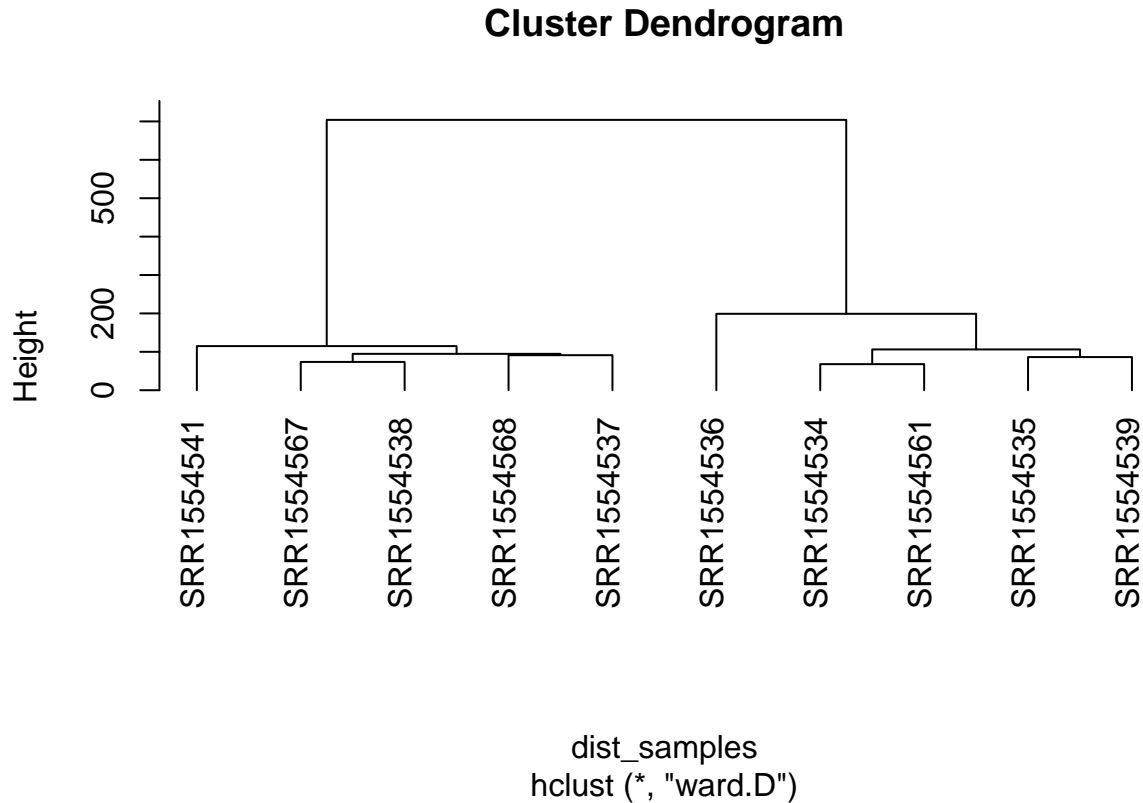
According to plotPCA having intgroup is “age_group”, it is clear that there can be an association between differential gene expression and age_group(fetus vs adult). However, in plotPCA having intgroup is “sex”, we can see that there could be no association between gender and gene expression.

I will also show the table of data transform of the above count_symbol table and the cluster of it in Dendrogram so that we could easily visualize the correlation between those samples.

```
edata_tr <- assay(edata_tr)
head(edata_tr)
```

##	SRR1554534	SRR1554535	SRR1554568	SRR1554561	SRR1554567	SRR1554536
## A1BG	9.190681	8.591025	8.087121	9.266152	6.808106	8.754521
## NAT2	2.898727	2.564131	2.169268	1.840736	2.306952	1.945286
## ADA	8.671241	9.225800	7.208636	7.967730	7.255809	9.862274
## CDH2	13.536261	13.571832	14.972885	13.633542	14.836363	13.764138
## AKT3	13.453204	13.144329	15.892417	13.906920	14.381850	12.281291
## GAGE12F	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
##	SRR1554541	SRR1554539	SRR1554538	SRR1554537		
## A1BG	8.388780	8.418109	7.553544	8.518947		
## NAT2	2.289385	2.615617	2.494186	2.140643		
## ADA	7.824756	8.248514	7.806678	7.090937		
## CDH2	14.563325	13.699760	14.722689	15.072426		
## AKT3	15.224047	14.367595	14.763178	15.606567		
## GAGE12F	0.000000	0.000000	0.000000	0.000000		

```
dist_samples <- dist(t(edata_tr))
gene_fit <- hclust(dist_samples, method="ward.D")
plot(gene_fit, hang=-1)
```

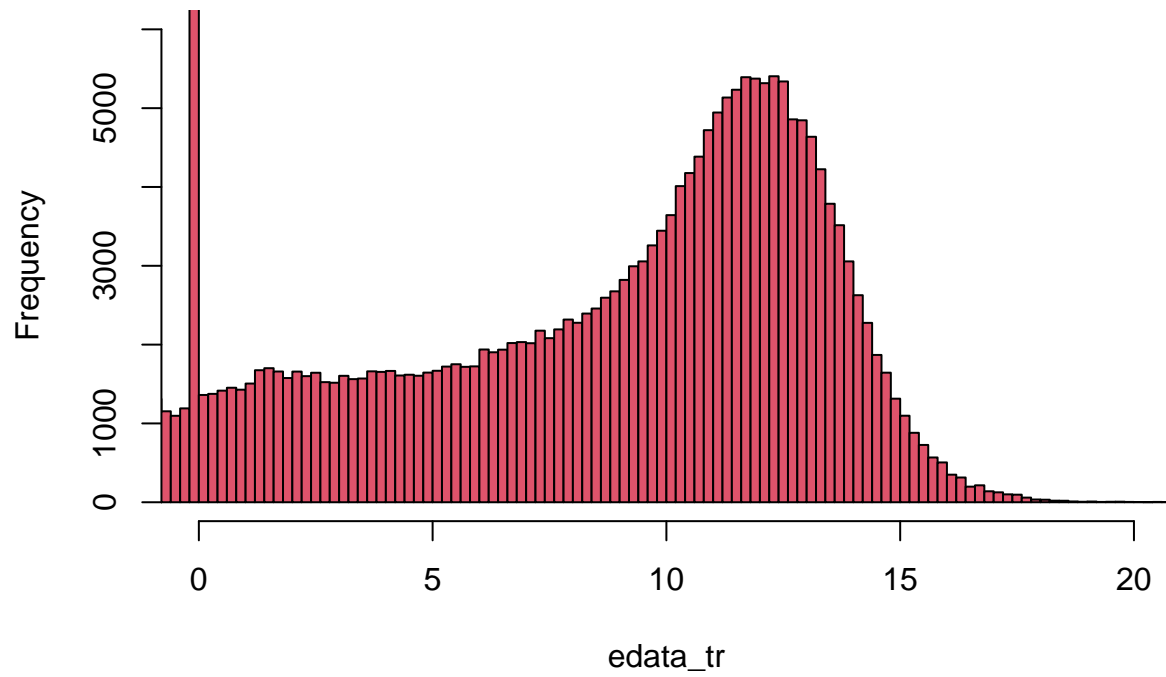


According to cluster, we can also see that there are 2 main branches. All the small branches SRR15545(41,68,37,67,38) on the main left branch are in the fetus group, while the others on the right main branch are in the adult group. This result reinforces the hypothesis that there can be a correlation between differential gene expression in fetus vs adult.

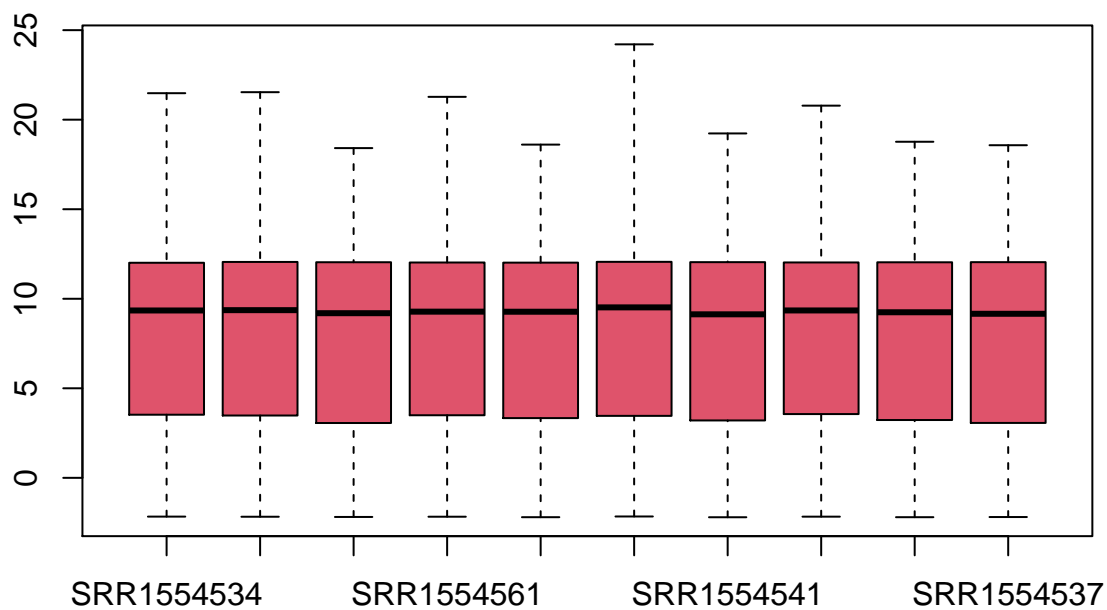
Additionally, I also visualize the frequency of number of reads in the count table when data has been transformed in histogram and boxplot. We can see that mostly the reads are below 20 and in the range between 5 to 15.

```
hist(edata_tr, breaks=100, col=2, xlim=c(0,20), ylim=c(0,6000))
```

Histogram of edata_tr



```
boxplot(edata_tr,col=2,range=0)
```

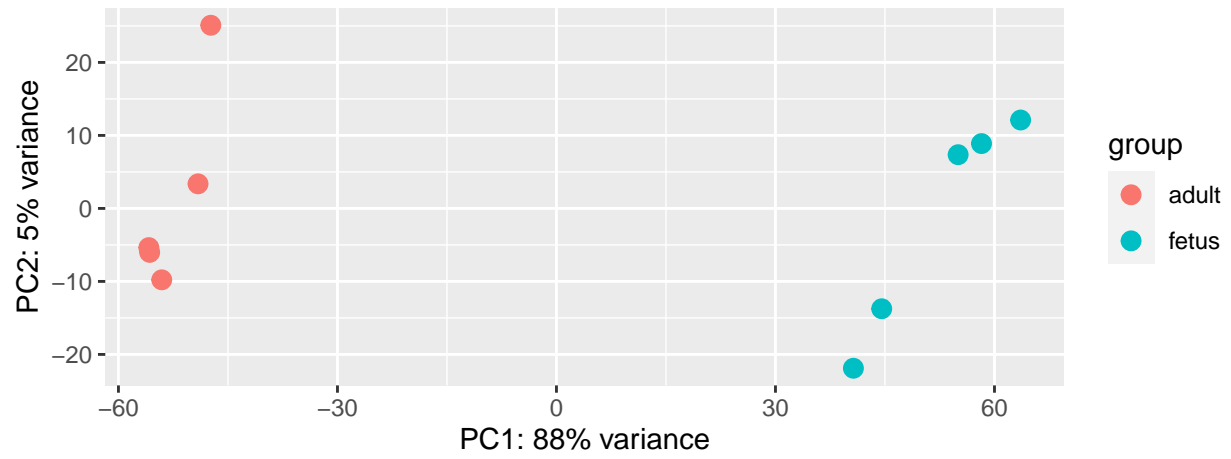


Filtering data

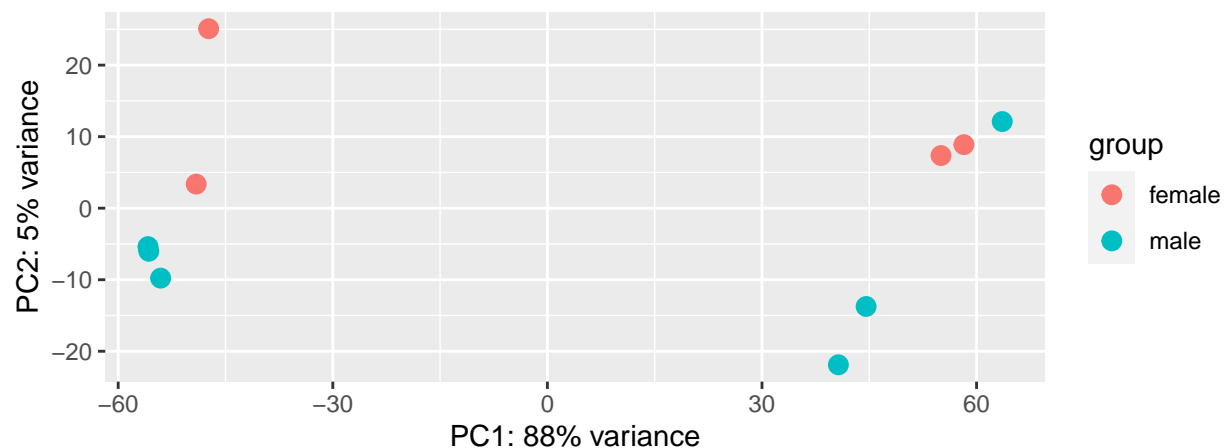
Removing lowly expressed genes and using DESeq to explore data as the same way as aforementioned unfiltered data in order to see if there are any differences between filtered and unfiltered.

```
count_filter = count_symbol[rowMeans(count_symbol) > 100,]
edata <- DESeqDataSetFromMatrix(countData = count_filter, colData = pheno, design = ~ age_group)

edata_tr <- rlog(edata, blind = FALSE)
plotPCA(edata_tr, intgroup = c("age_group"))
```



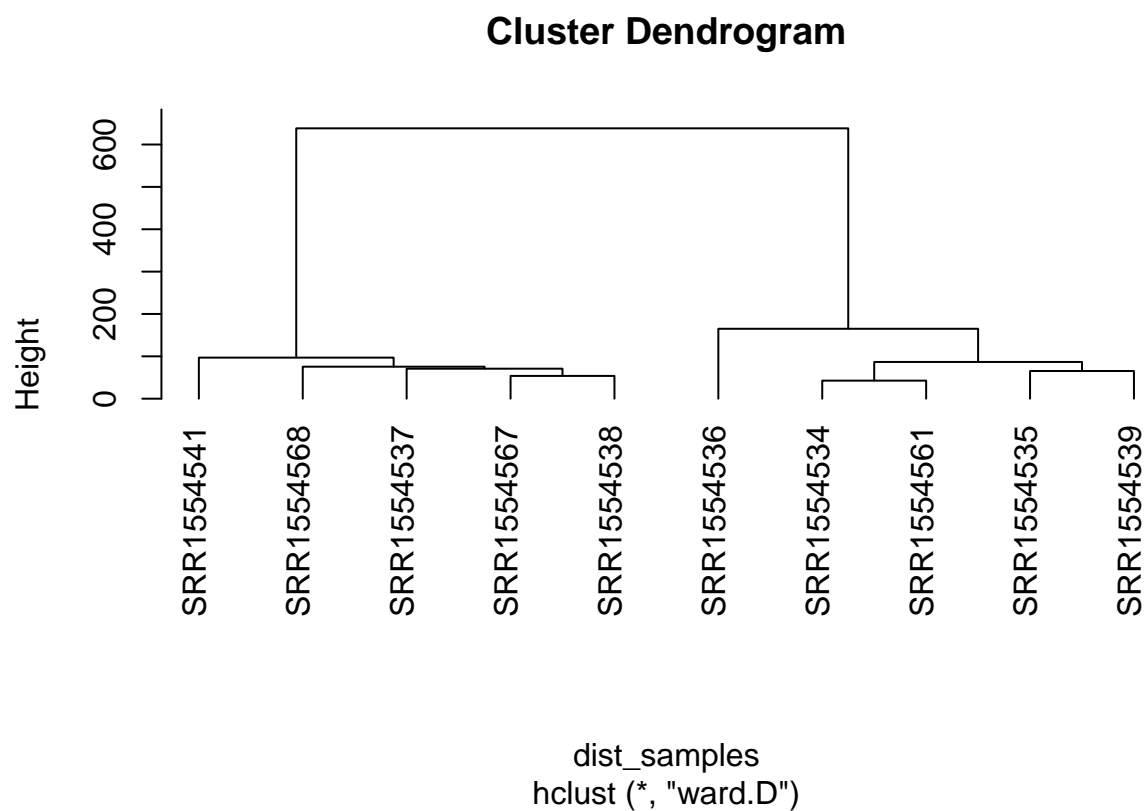
```
plotPCA(edata_tr, intgroup = c("sex"))
```

```
edata_tr <- assay(edata_tr)
head(edata_tr)
```

```
##          SRR1554534 SRR1554535 SRR1554568 SRR1554561 SRR1554567 SRR1554536
## A1BG          9.143633  8.583227  8.105891  9.206773  6.995598  8.762864
## ADA           8.646385  9.145974  7.303994  8.001912  7.351627  9.770115
## CDH2          13.562468 13.585906 14.957846 13.652418 14.829826 13.812075
## AKT3          13.479431 13.159273 15.876052 13.925121 14.376621 12.332018
## ZBTB11-AS1    7.746129  7.539217  7.489181  7.323988  7.706548  7.731734
## MED6          10.965760 10.953407 11.062971 10.825176 11.161363 10.689913
##          SRR1554541 SRR1554539 SRR1554538 SRR1554537
## A1BG          8.373665  8.429514  7.632873  8.489938
## ADA           7.843839  8.253862  7.831596  7.197164
## CDH2          14.545283 13.717079 14.708753 15.051336
## AKT3          15.204848 14.383520 14.749481 15.584859
## ZBTB11-AS1    7.684944  7.654166  7.625887  7.705363
## MED6          11.229573 11.095095 11.270915 11.087431
```

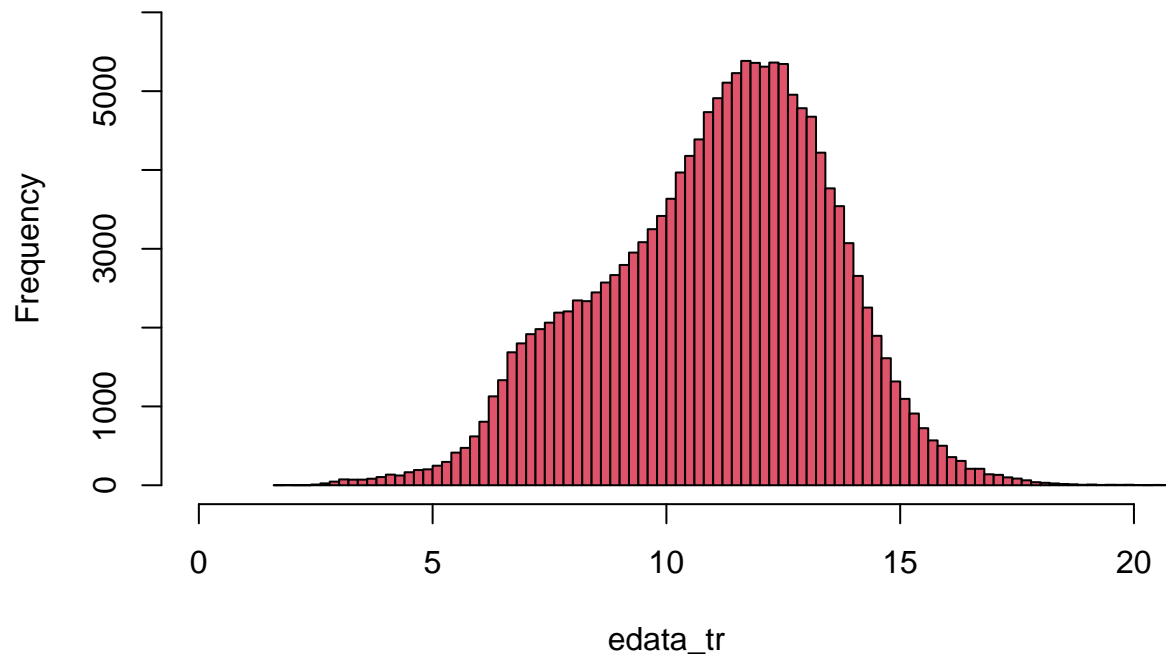
```
dist_samples <- dist(t(edata_tr))
gene_fit <- hclust(dist_samples, method="ward.D")
plot(gene_fit, hang=-1)
```



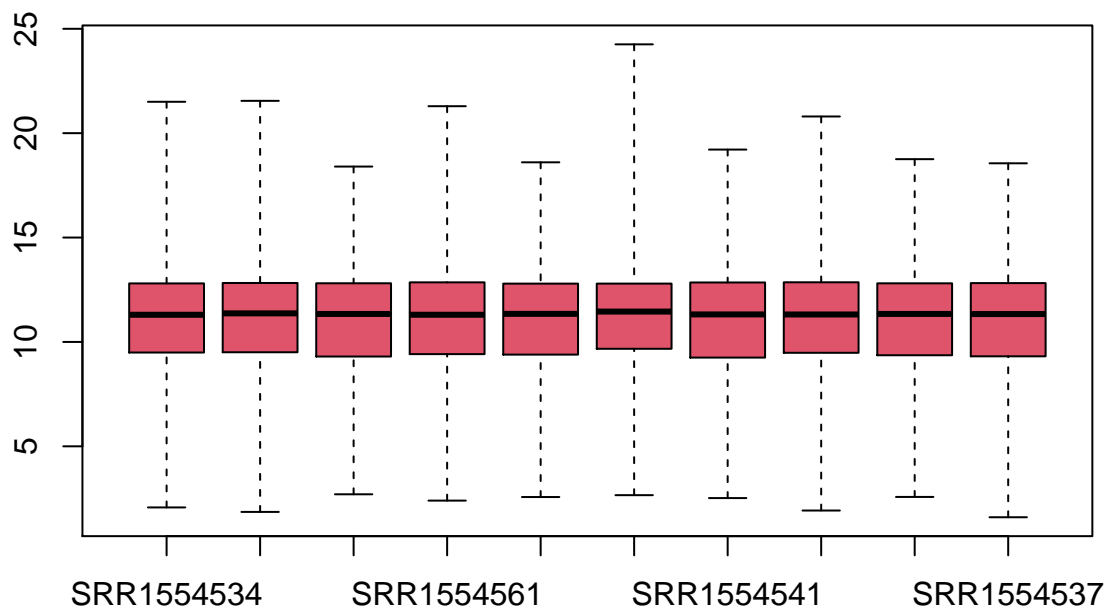
We can see that there is almost no difference in the results of unfiltered and filtered data except in cluster dendrogram of filtered data, the relation between SRR1554568 and SRR1554537 are not close-related as same as that of unfiltered data.

```
hist(edata_tr,breaks=100,col=2,xlim=c(0,20),ylim=c(0,6000))
```

Histogram of edata_tr



```
boxplot(edata_tr,col=2,range=0)
```



Sex-adjusted

Because I want to explore if the sex variable (male vs female) might effect the association between age_group variable and gene expression, I get the female data in 10 samples above and analyse the factor age_group in this gender and do the same with the male gender.

Female There are 4 samples of 10 above samples which sex is female

```
pheno_female = pheno[pheno$sex == "female",]
pheno_female
```

```
##      Run age_group    age  sex
## 6  SRR1554536   adult 44.1700 female
## 8  SRR1554539   adult 36.5000 female
## 9  SRR1554538   fetus -0.4027 female
## 10 SRR1554537   fetus -0.3836 female
```

```
female_run <- (colnames(count_symbol) %in% pheno_female$Run)
table(female_run)
```

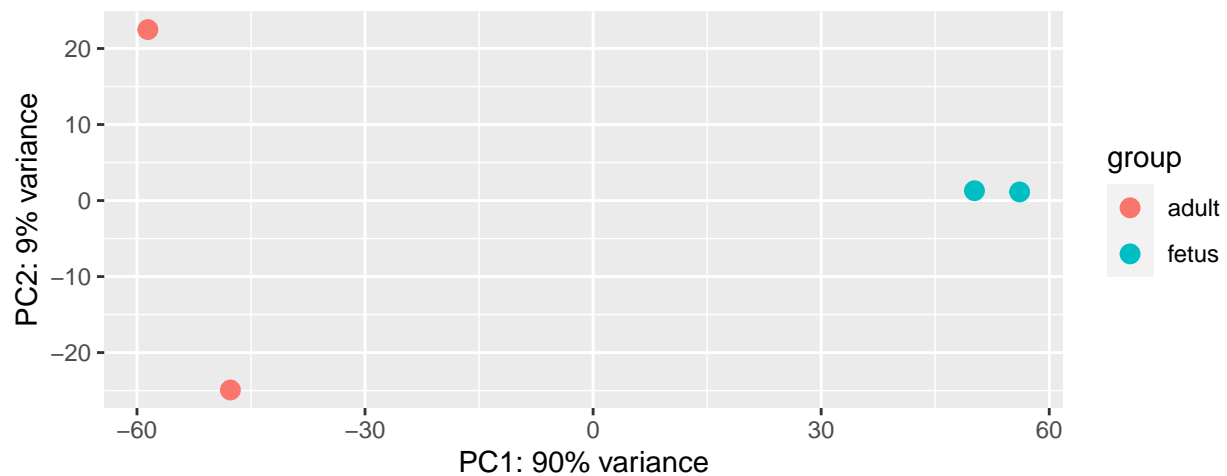
```
## female_run
## FALSE  TRUE
##      6      4
```

```
count_female = count_symbol[,female_run]
head(count_female)
```

```
##          SRR1554536 SRR1554539 SRR1554538 SRR1554537
## A1BG             114          295          275          518
## NAT2              0           8           12           4
## ADA              291          265          354          160
## CDH2             3270         10639         47354         52346
## AKT3              937         18216         48565         79685
## GAGE12F           0           0           0           0
```

```
edata_fe <- DESeqDataSetFromMatrix(count_female, pheno_female, ~age_group)
```

```
edata_fe_tr <- rlog(edata_fe, blind = FALSE)
plotPCA(edata_fe_tr, intgroup = c("age_group"))
```



```
pheno_male = pheno[pheno$sex == "male",]
pheno_male
```

Male

```
##          Run age_group    age sex
## 1 SRR1554534    adult 40.4200 male
## 2 SRR1554535    adult 41.5800 male
## 3 SRR1554568    fetus -0.4986 male
## 4 SRR1554561    adult 43.8800 male
## 5 SRR1554567    fetus -0.4027 male
## 7 SRR1554541    fetus -0.3836 male
```

```
male_run <- (colnames(count_symbol) %in% pheno_male$Run)
table(male_run)
```

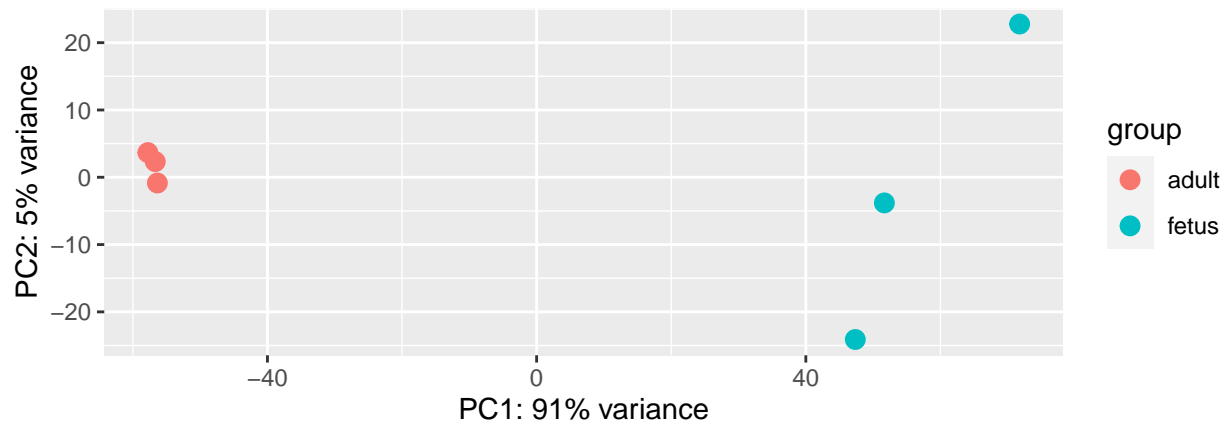
```
## male_run
## FALSE  TRUE
##      4      6
```

```
count_male = count_symbol[,male_run]
head(count_male)
```

```
##          SRR1554534 SRR1554535 SRR1554568 SRR1554561 SRR1554567 SRR1554541
## A1BG             444          378          328          650          146          592
## NAT2              11           8           4           0           8           8
## ADA              299          658          161          229          225          382
## CDH2             7384         10623         43926         11016         52746         44244
## AKT3             6837          7391          90930         13677         36246         74768
## GAGE12F           0           0           0           0           0           0
```

```
edata_ma <- DESeqDataSetFromMatrix(count_male, pheno_male, ~age_group)
```

```
edata_ma_tr <- rlog(edata_ma, blind = FALSE)
plotPCA(edata_ma_tr, intgroup = c("age_group"))
```



We can see that in female and male, there is a difference between fetus and adult, which is as same as the above result of age_group variable in unfiltered and filtering data section without adjustment factor, sex variable. Thus, the sex variable might not effect much to the result of age_group.

Statistical analysis

The target of this statistical analysis is examining the correlation between age_group variable (fetus or adult) and gene expression so that in the end we can see the difference between up-regulated genes and down-regulated genes in fetus vs adult.

In this statistical analysis, I will use limma package and DESeq package to see if there are any the different results between two methods.

I will use DESeq package to analyze edata (output of DESeq) and Limma package to analyze edata_tr (edata has been transformed)

Unadjusted data

At first, I will analyse statistically variable age_group without adjustment factor (sex variable).

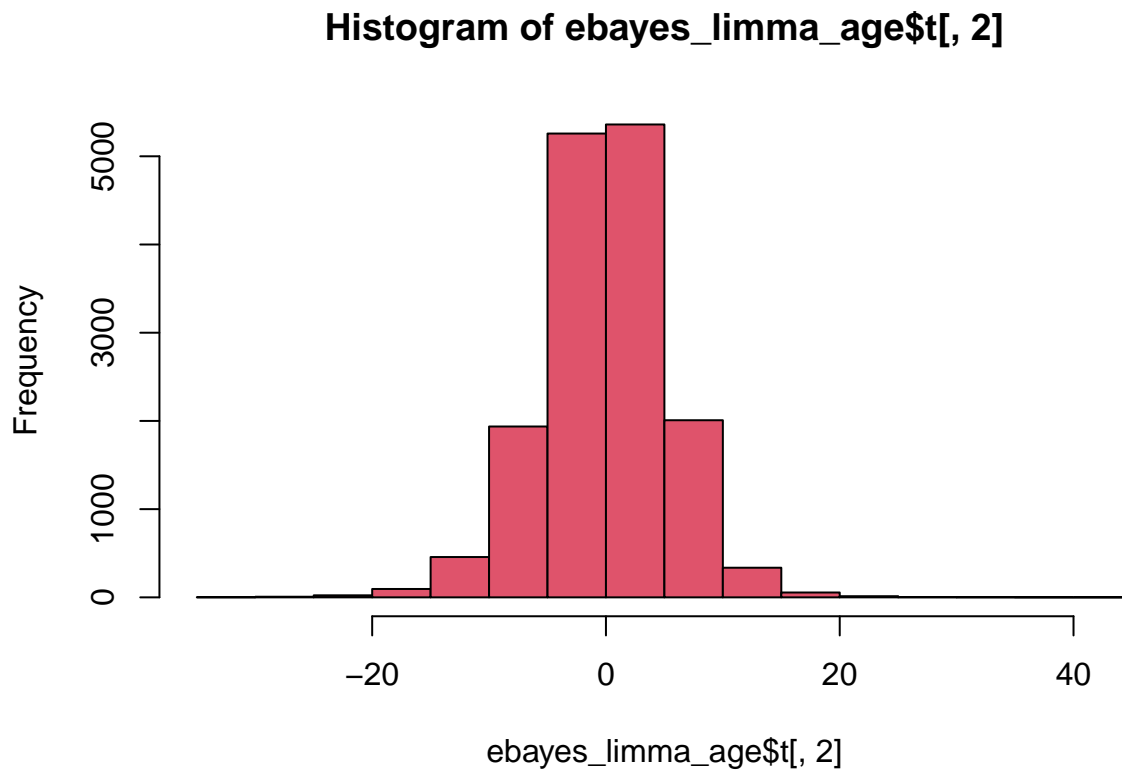
```
# age_group
mod_age = model.matrix(~ pheno$age_group)
fit_limma_age = lmFit(edata_tr,mod_age)
```

```
ebayes_limma_age = eBayes(fit_limma_age)
re = topTable(ebayes_limma_age, number=dim(count_symbol)[1])
head(re)
```

Fit regression with limma package

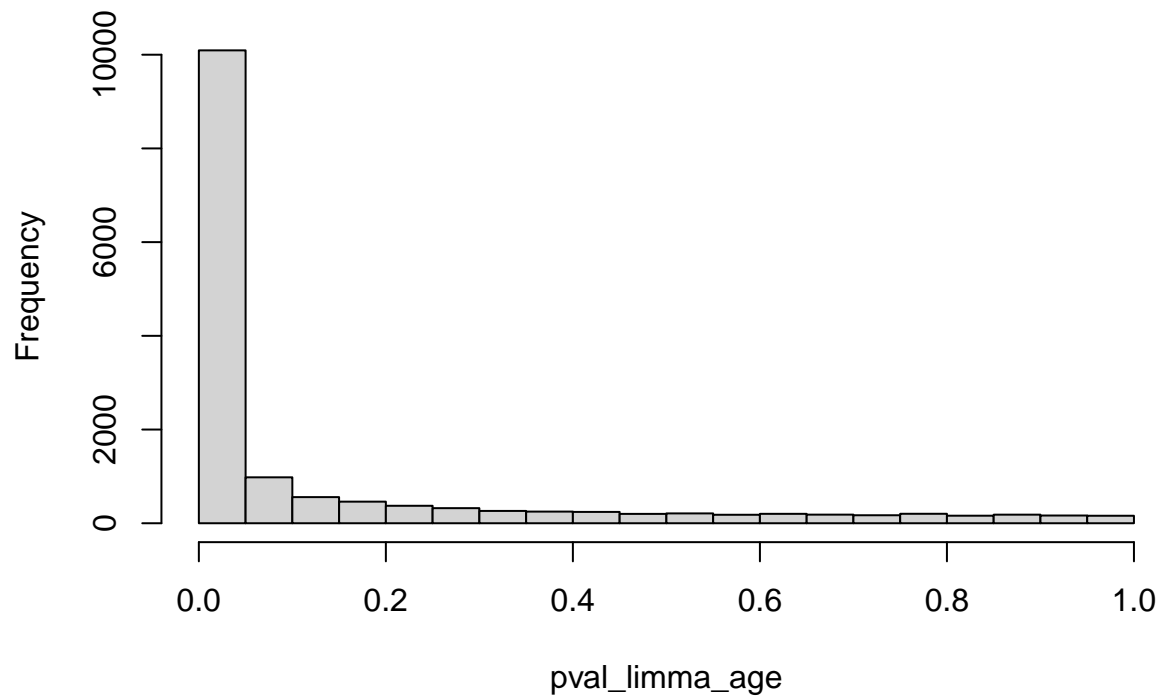
##		logFC	AveExpr	t	P.Value	adj.P.Val	B
##	ST8SIA2	6.264017	11.767199	40.30384	2.684402e-13	2.770718e-09	20.15611
##	SOX11	7.196578	13.549118	39.27425	3.561563e-13	2.770718e-09	19.94839
##	TRIM54	-6.448248	6.397455	-34.32829	1.547426e-12	8.025467e-09	18.81527
##	SLA	5.755159	12.718546	33.37950	2.100303e-12	8.169655e-09	18.56884
##	FBN3	4.918986	11.616488	30.36669	5.882270e-12	1.382686e-08	17.71288
##	SNCG	-6.707473	10.307404	-30.34373	5.930876e-12	1.382686e-08	17.70589

```
# Statistics
hist(ebayes_limma_age$t[,2], col=2)
```



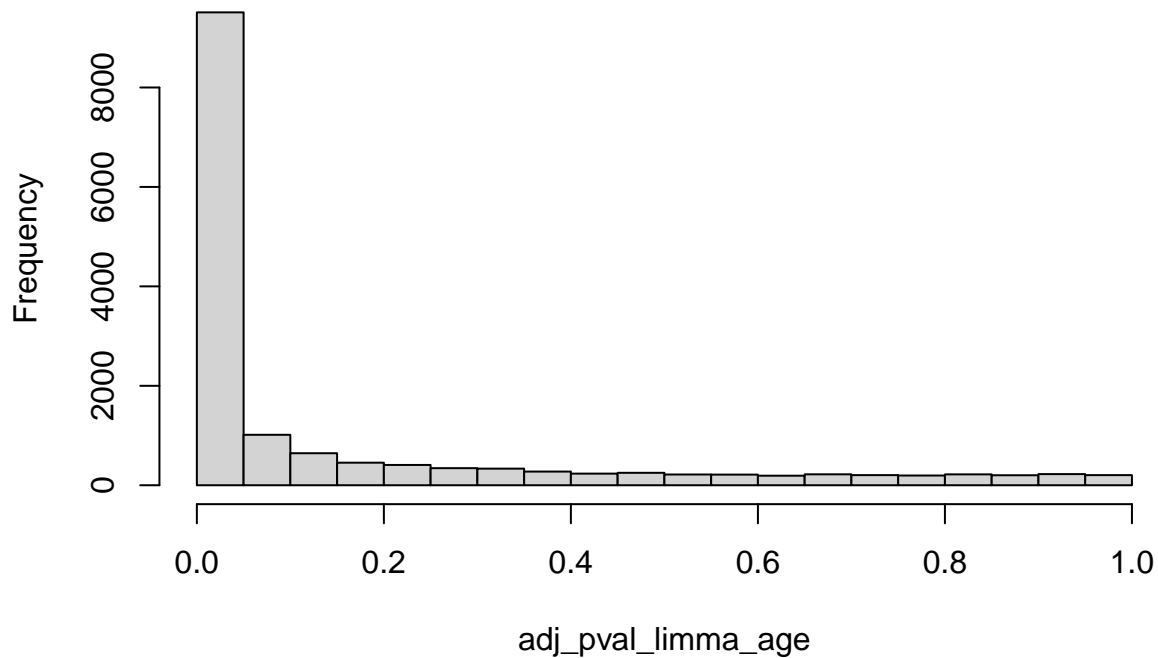
```
# P-values
pval_limma_age = topTable(ebayes_limma_age, number=dim(edata_tr)[1])$P.Value
hist(pval_limma_age)
```


Histogram of pval_limma_age



```
# Adjusted p-values  
adj_pval_limma_age = topTable(ebayes_limma_age, number=dim(edata_tr)[1])$adj.P.Val  
hist(adj_pval_limma_age)
```

Histogram of adj_pval_limma_age



According to histogram of adjusted p-value, it suggests that there might be an association between age_group variable and gene expression, which means there is a difference between fetus and adult.

A number of genes have adjusted p-value less than 0.05

```
sum(re$adj.P.Val < 0.05)
```

```
## [1] 9511
```

```
dds <- DESeq(edata)
```

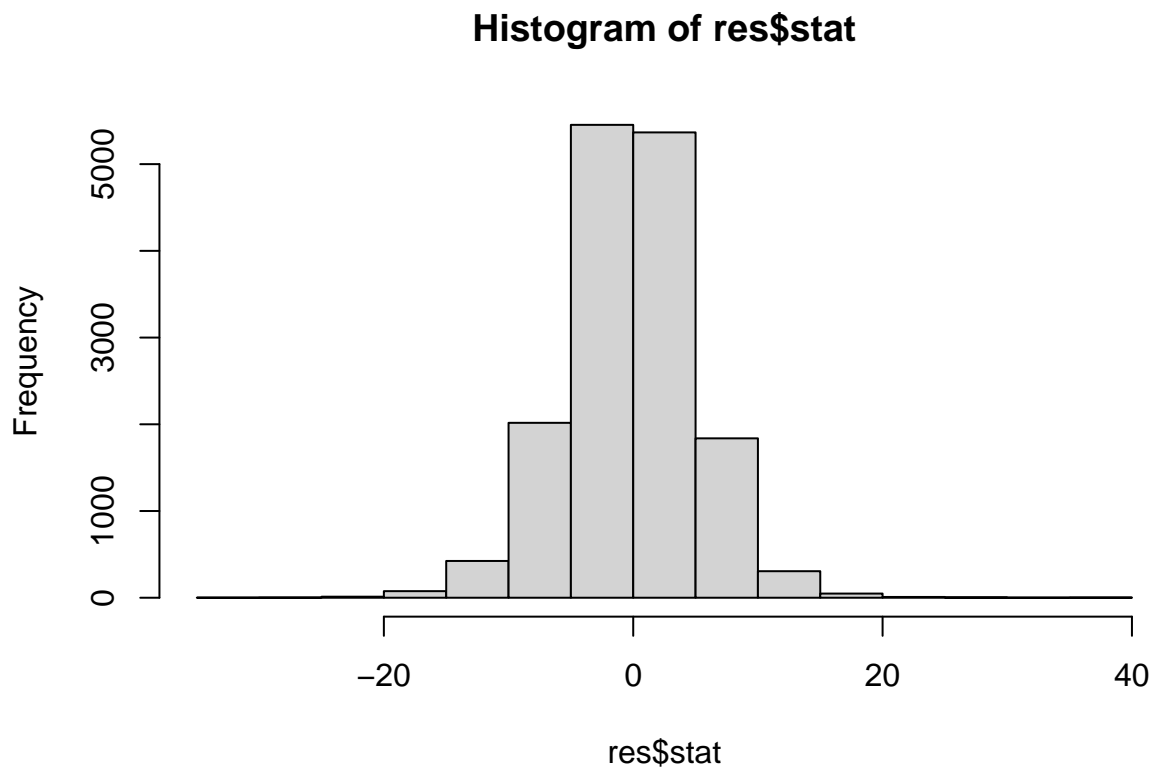
```
res <- results(dds)
res = as.data.frame(res)
head(res)
```

Fit regression with DESeq

	baseMean	log2FoldChange	lfcSE	stat	pvalue
## A1BG	380.8469	-1.10293610	0.4076920	-2.7053167	6.823930e-03
## ADA	372.3584	-1.86297881	0.4452617	-4.1840088	2.864130e-05
## CDH2	21681.2984	1.36767686	0.1515741	9.0231574	1.827447e-19

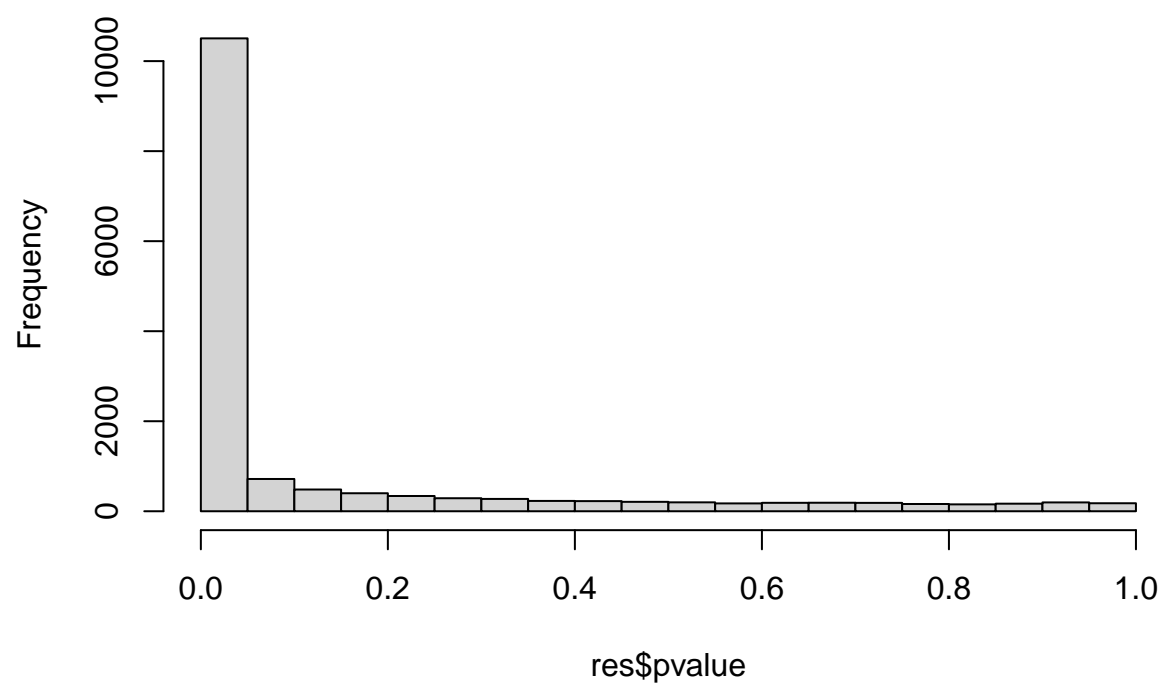
```
## AKT3      27928.4725      1.91781744 0.4513954  4.2486416 2.150707e-05
## ZBTB11-AS1 198.2623      0.06499846 0.2042175  0.3182805 7.502722e-01
## MED6      2117.0608      0.29961394 0.1355707  2.2100194 2.710382e-02
##
##          padj
## A1BG      1.190664e-02
## ADA       7.382847e-05
## CDH2      2.284087e-18
## AKT3      5.653192e-05
## ZBTB11-AS1 7.941369e-01
## MED6      4.215816e-02
```

```
# Statistic
hist(res$stat)
```

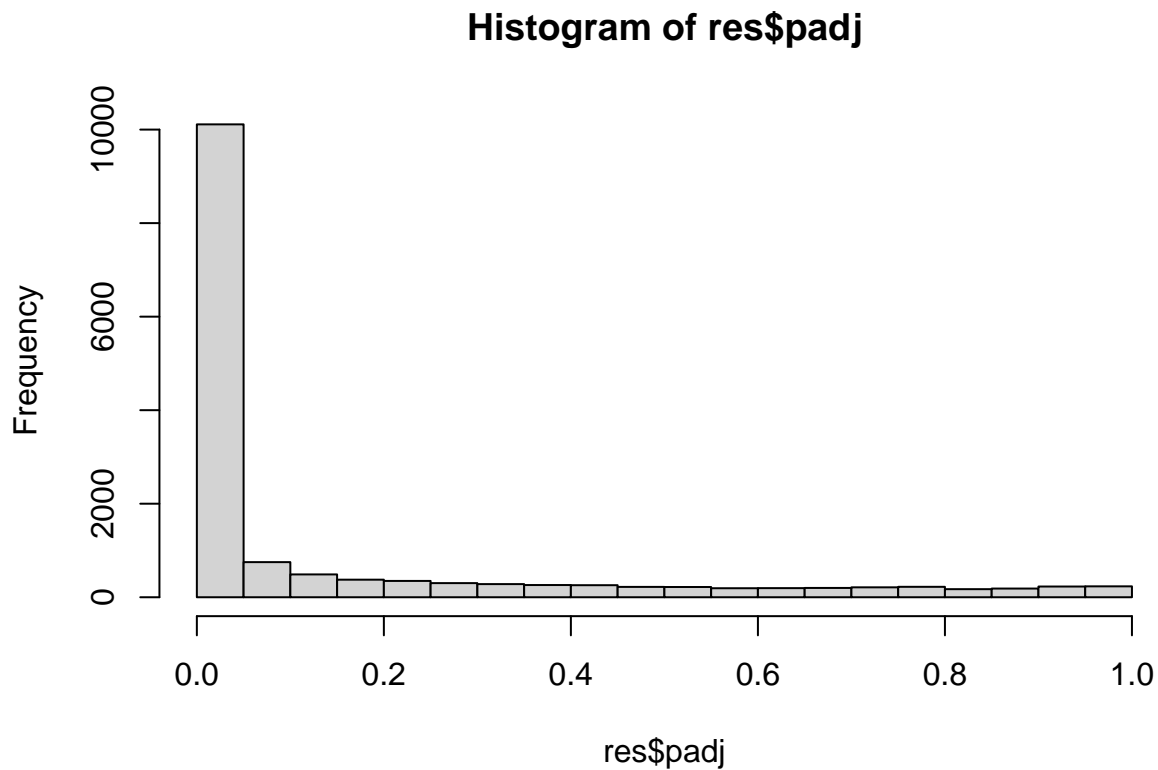


```
# P-values
hist(res$pvalue)
```

Histogram of res\$pvalue



```
#Adjusted p-values  
hist(res$padj)
```



We can see that there is also an association between age_group and gene expression that is as same as limma package.

A number of genes have adjusted p-value less than 0.05

```
table(res$padj < 0.05)
```

```
##
## FALSE  TRUE
##  5349 10112
```

Adjusted data

Because I suspect that sex variable can adjust the association between age_group factor with gene expression, we will analyse statistically data with adjustment factor.

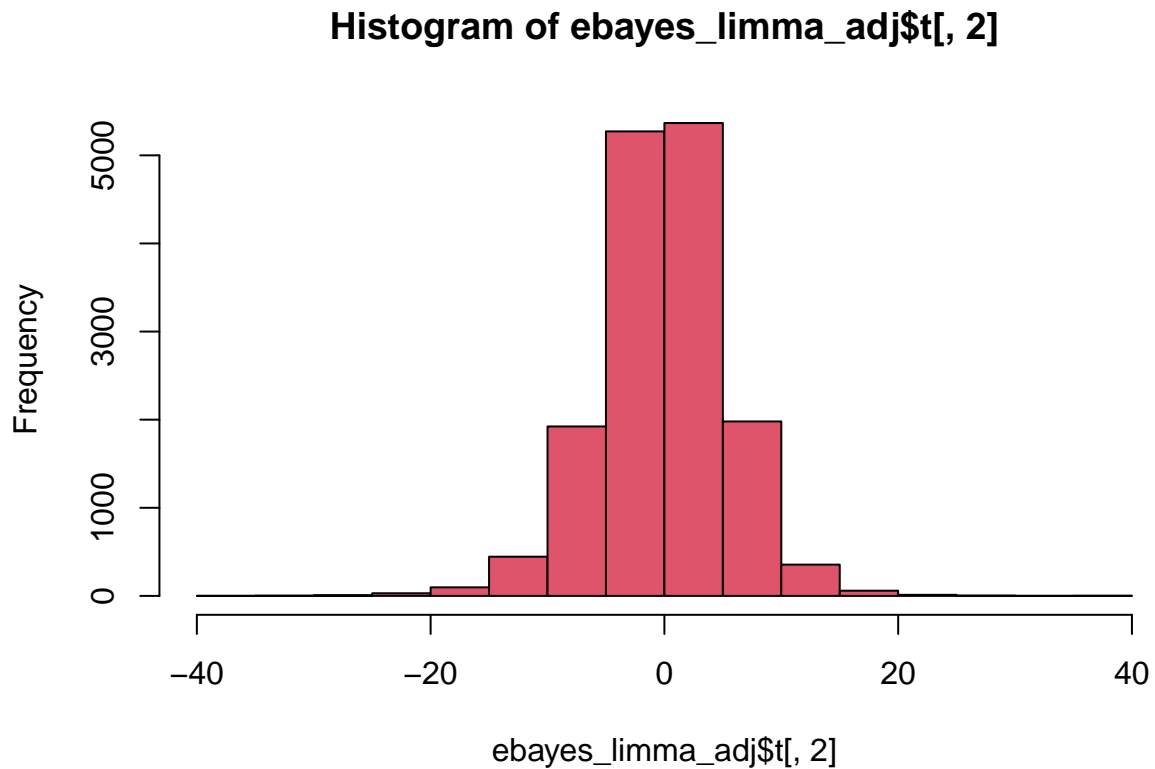
```
mod_adj = model.matrix(~ pheno$age_group+pheno$sex)
fit_limma_adj = lmFit(edata_tr,mod_adj)
ebayes_limma_adj <- eBayes(fit_limma_adj)
names(ebayes_limma_adj)
```

Fit regression with limma package with adjustment factor

```
## [1] "coefficients"      "rank"              "assign"            "qr"
## [5] "df.residual"       "sigma"             "cov.coefficients"  "stdev.unscaled"
## [9] "pivot"            "Amean"             "method"            "design"
## [13] "df.prior"          "s2.prior"          "var.prior"         "proportion"
## [17] "s2.post"           "t"                 "df.total"          "p.value"
## [21] "lods"              "F"                 "F.p.value"
```

```
# Statistics
```

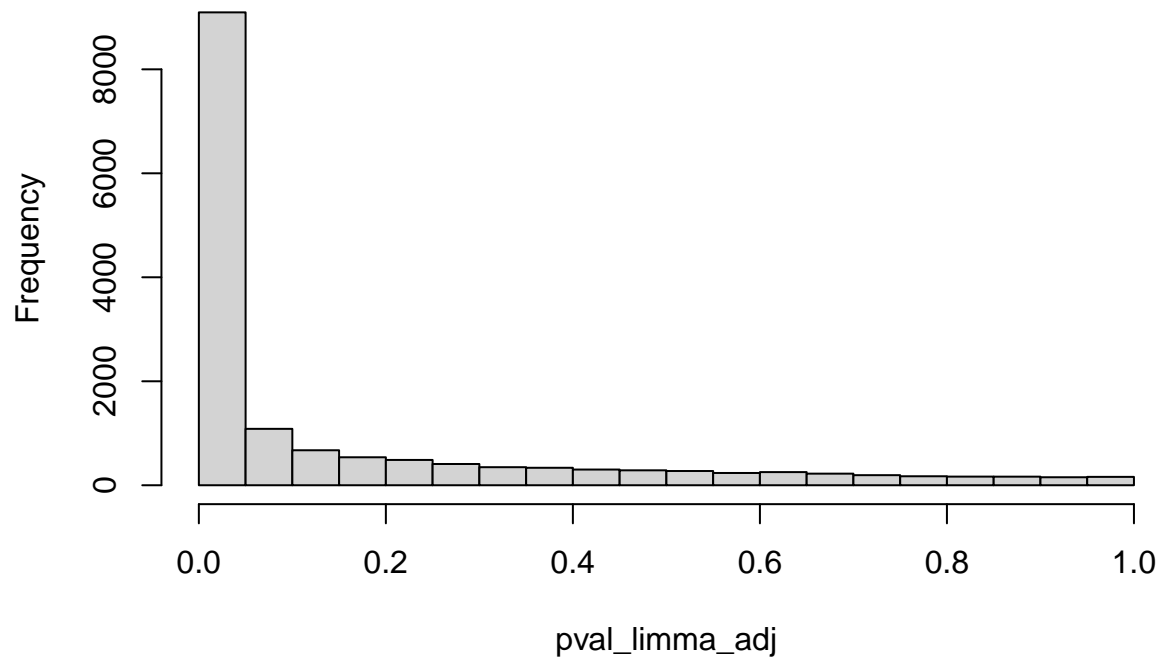
```
hist(ebayes_limma_adj$t[,2], col=2)
```



```
# P-values
```

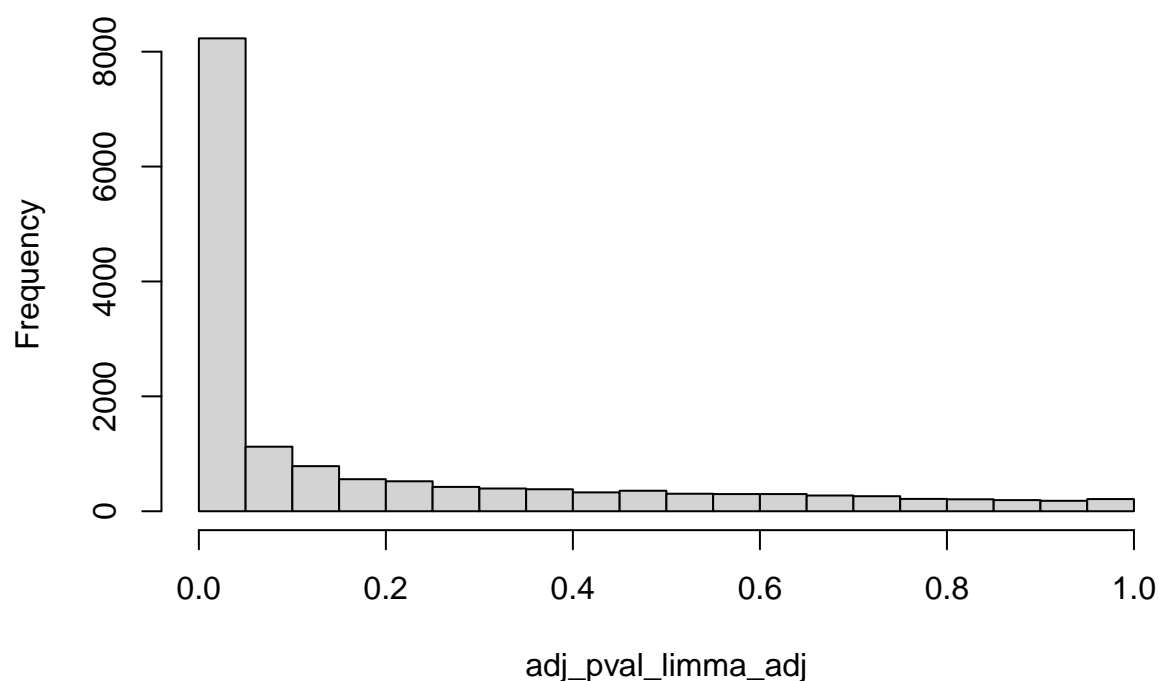
```
pval_limma_adj = topTable(ebayes_limma_adj, number=dim(edata_tr)[1])$P.Value
hist(pval_limma_adj)
```

Histogram of pval_limma_adj



```
# Adjusted p-values  
adj_pval_limma_adj = topTable(ebayes_limma_adj, number=dim(edata_tr)[1])$adj.P.Val  
hist(adj_pval_limma_adj)
```

Histogram of adj_pval_limma_adj



```
re_adj = topTable(ebayes_limma_adj, number=dim(edata_tr)[1])
head(re_adj)
```

```
##      pheno.age_groupfetus pheno.sexmale AveExpr      F      P.Value
## ST8SIA2      6.264017    -0.09717847 11.767199 758.4099 1.177104e-11
## SOX11       7.196578    -0.14784492 13.549118 739.4981 1.334594e-11
## SNCG       -6.707473     0.44733611 10.307404 650.4990 2.524224e-11
## NKX6-2     -7.302594     0.82472619  6.955729 588.5204 4.150116e-11
## TRIM54     -6.448248     0.06076597  6.397455 537.9631 6.480537e-11
## SLA        5.755159    -0.06607231 12.718546 509.9261 8.450284e-11
##      adj.P.Val
## ST8SIA2 1.038247e-07
## SOX11   1.038247e-07
## SNCG    1.309147e-07
## NKX6-2  1.614291e-07
## TRIM54  2.016614e-07
## SLA     2.191300e-07
```

```
sum(re_adj$adj.P.Val < 0.05)
```

```
## [1] 8232
```



```
de_adj = DESeqDataSetFromMatrix(countData = count_filter, colData = pheno, ~ age_group + sex)
glm_all_adj = DESeq(de_adj)
```

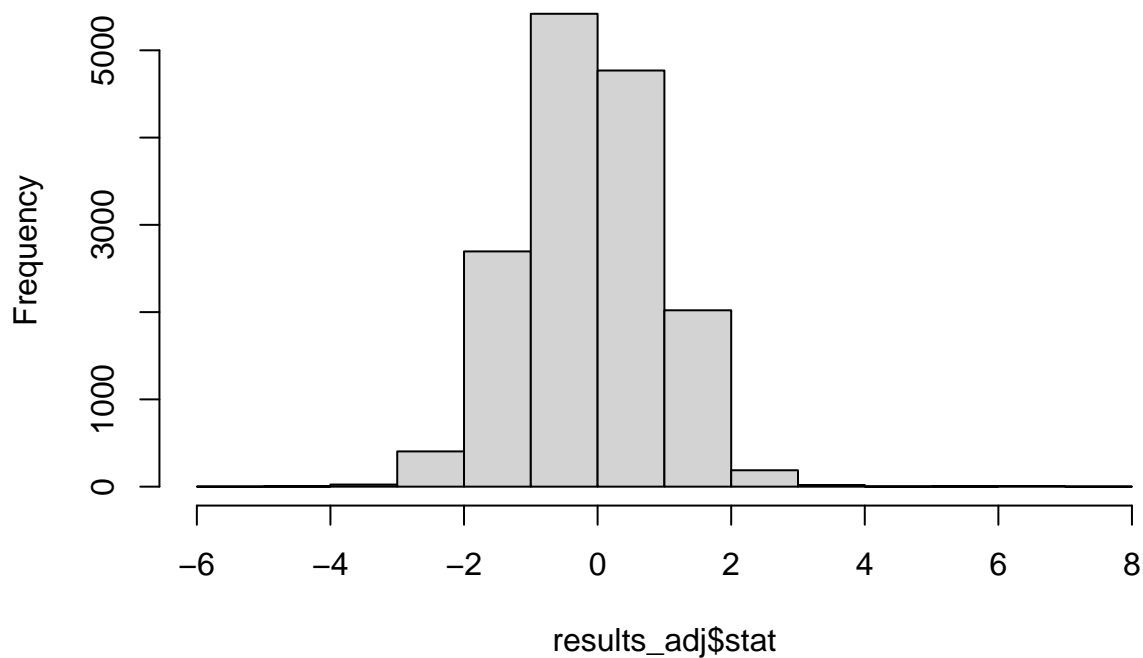
```
results_adj = results(glm_all_adj)
results_adj = as.data.frame(results_adj)
head(results_adj)
```

Fit regression with DESeq with adjustment factor

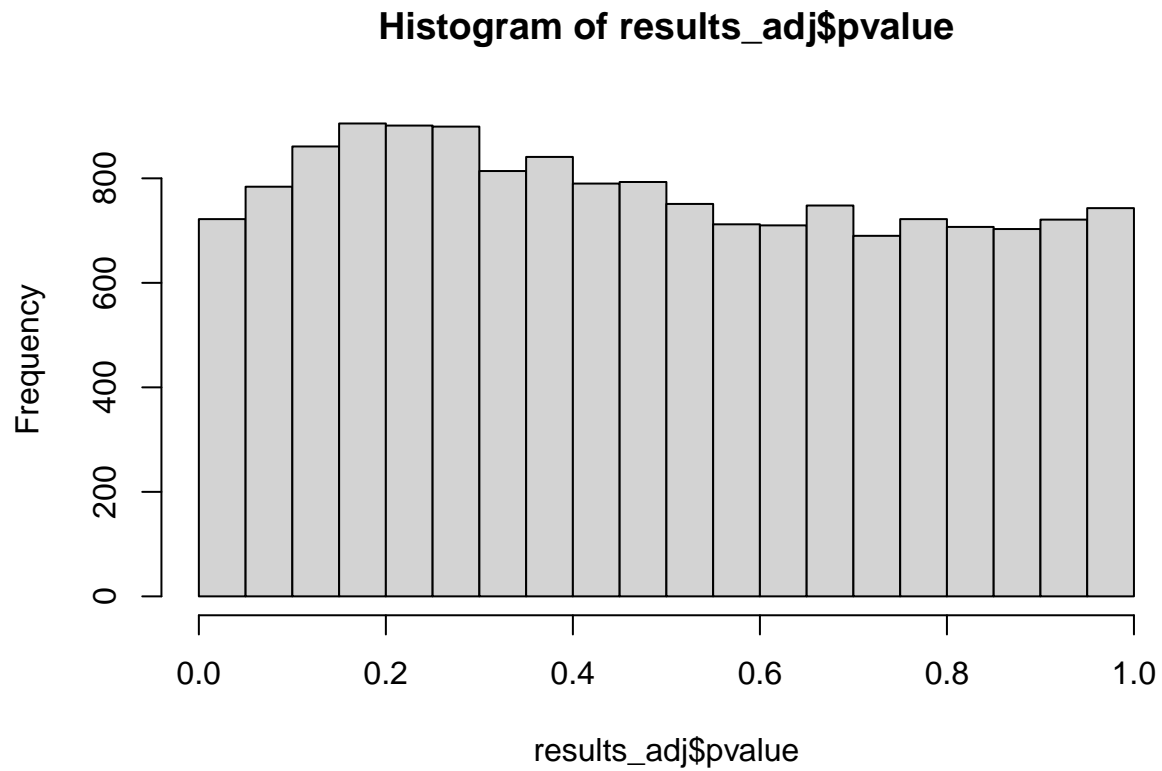
##	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj
## A1BG	380.8469	0.12906526	0.4352513	0.2965304	0.7668250	0.9776339
## ADA	372.3584	-0.38152598	0.4613657	-0.8269492	0.4082658	0.9230627
## CDH2	21681.2984	-0.15749391	0.1544875	-1.0194605	0.3079844	0.9171161
## AKT3	27928.4725	-0.07403085	0.4782344	-0.1548003	0.8769787	0.9913890
## ZBTB11-AS1	198.2623	-0.12674096	0.2223552	-0.5699933	0.5686822	0.9462994
## MED6	2117.0608	-0.01682903	0.1510819	-0.1113901	0.9113070	0.9942778

```
# Statistic
hist(results_adj$stat)
```

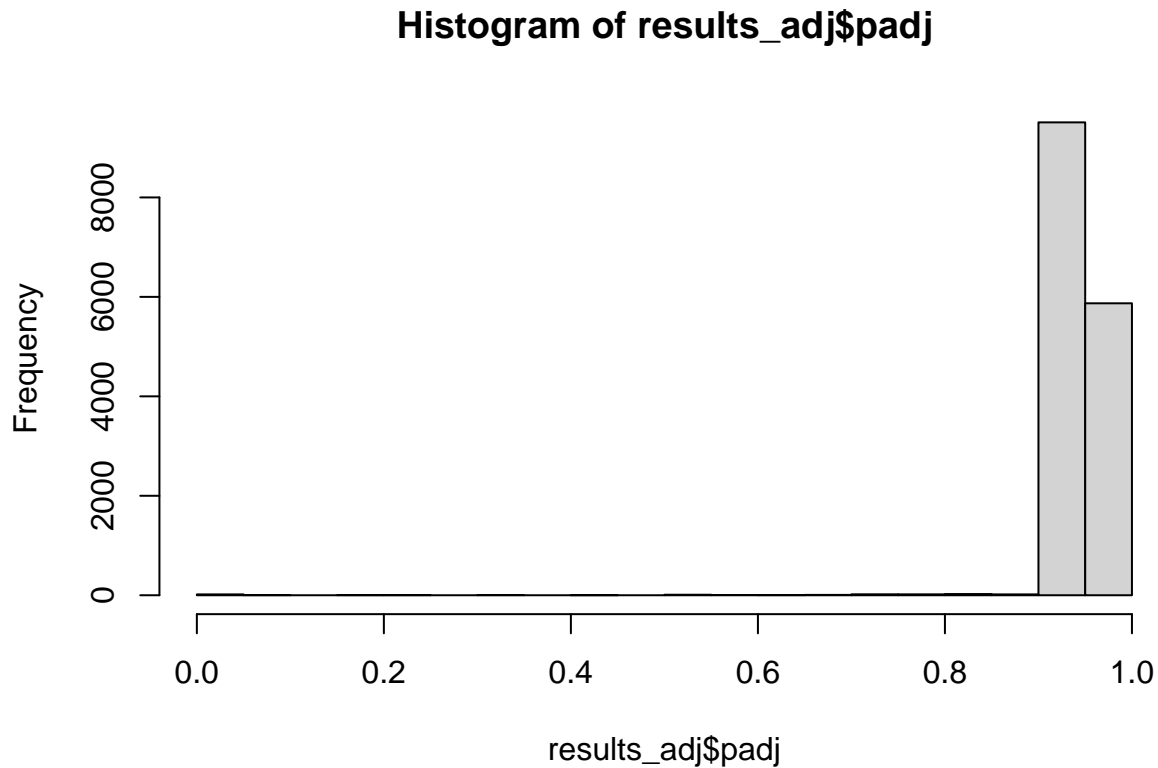
Histogram of results_adj\$stat



```
# P-values  
hist(results_adj$pvalue)
```



```
# Adjusted p-values  
hist(results_adj$padj)
```



In the adjustment section, we can see that in limma package, the results which adjusted p-value is less than 0.05 are 8232 accounting for about 52.9% in the total of 15559 genes. However, in DESeq package, the results less than 0.05 are little, and according to the the histogram of results_adj pvalue in DESeq, it is likely that there is no association between adjusted data and gene expression.

Get up and down regulated genes

Because of clear evidence of correlation when analyzing unadjusted data, which only have age_group variable, we can see that there is a correlation between the age_group factor and gene expression. Therefore, I will count the up-regulated genes, which are highly expressed and up-regulation in human especially for fetus, and down-regulated genes, which are down-regulation in human especially when getting older and as a result highly appear in the adult than the fetus

I will get up-regulated and down-regulated genes from unadjusted data of both limma package and DESeq package.

Limma package The number of up-regulated genes

```
sum(re$adj.P.Val < 0.05 & re$logFC > 1)
```

```
## [1] 2560
```

```
up_limma <- re %>% filter (logFC > 1 & adj.P.Val < 0.05) %>% arrange(adj.P.Val)
head(up_limma)
```

	logFC	AveExpr	t	P.Value	adj.P.Val	B
## ST8SIA2	6.264017	11.76720	40.30384	2.684402e-13	2.770718e-09	20.15611
## SOX11	7.196578	13.54912	39.27425	3.561563e-13	2.770718e-09	19.94839
## SLA	5.755159	12.71855	33.37950	2.100303e-12	8.169655e-09	18.56884
## FBN3	4.918986	11.61649	30.36669	5.882270e-12	1.382686e-08	17.71288
## VASH2	5.011997	11.06801	29.96521	6.798356e-12	1.382686e-08	17.58961
## DCX	5.904494	14.73240	29.23579	8.886727e-12	1.382686e-08	17.35962

The number of down-regulated genes

```
sum(re$adj.P.Val < 0.05 & re$logFC < -1)
```

```
## [1] 3080
```

```
down_limma <- re %>% filter (logFC < -1 & adj.P.Val < 0.05) %>% arrange(adj.P.Val)
head(down_limma)
```

	logFC	AveExpr	t	P.Value	adj.P.Val	B
## TRIM54	-6.448248	6.397455	-34.32829	1.547426e-12	8.025467e-09	18.81527
## SNCG	-6.707473	10.307404	-30.34373	5.930876e-12	1.382686e-08	17.70589
## OPALIN	-8.692659	8.768713	-29.51529	8.013662e-12	1.382686e-08	17.44869
## SOHLH1	-7.503042	7.626357	-29.37968	8.424990e-12	1.382686e-08	17.40562
## KRT17	-6.499428	5.955451	-27.47641	1.743578e-11	2.466212e-08	16.77083
## UAP1L1	-4.556272	8.839835	-26.51408	2.566615e-11	3.135364e-08	16.42680

DESeq package The number of up-regulated genes

```
sum(res$padj < 0.05 & res$log2FoldChange > 1, na.rm=TRUE)
```

```
## [1] 3178
```

```
up_de <- res %>% filter (log2FoldChange > 1 & padj < 0.05) %>% arrange(padj)
head(up_de)
```

	baseMean	log2FoldChange	lfcSE	stat	pvalue
## ST8SIA2	21470.986	7.442980	0.2003001	37.15915	3.119957e-302
## SOX11	105896.449	8.543693	0.2330527	36.65991	3.180749e-294
## SLA	33778.704	6.781894	0.2163388	31.34849	1.020083e-215
## FBN3	11490.654	5.816179	0.2007956	28.96568	1.781244e-184
## MEX3B	12548.477	4.157591	0.1503316	27.65614	2.354317e-168
## VASH2	8164.861	5.920934	0.2269659	26.08732	5.077630e-150

	padj
## ST8SIA2	4.823766e-298
## SOX11	2.458878e-290
## SLA	5.257168e-212
## FBN3	5.507963e-181
## MEX3B	6.066682e-165
## VASH2	9.813154e-147

The number of down-regulated genes

```
sum(res$padj < 0.05 & res$log2FoldChange < -1, na.rm=TRUE)
```

```
## [1] 3853
```

```
down_de <- res %>% filter (log2FoldChange < -1 & padj < 0.05) %>% arrange(padj)
head(down_de)
```

```
##          baseMean log2FoldChange      lfcSE      stat      pvalue
## BCL2L2  21968.220      -2.676407 0.08831507 -30.30521 9.788276e-202
## SNCG    9473.167      -8.005651 0.29376717 -27.25169 1.587048e-163
## CLMN    3227.245      -3.760479 0.14437310 -26.04695 1.456668e-149
## UAP1L1  1538.034      -5.591225 0.23073417 -24.23232 1.015679e-129
## ITPKA   5882.483      -7.047788 0.29600941 -23.80934 2.673017e-125
## OPALIN  7380.398     -10.922203 0.45988078 -23.75008 1.096772e-124
##                padj
## BCL2L2  3.783413e-198
## SNCG    3.505336e-160
## CLMN    2.502394e-146
## UAP1L1  1.427583e-126
## ITPKA   3.443960e-122
## OPALIN  1.304400e-121
```

```
up <- rownames(up_de) %in% rownames(up_limma)
table(up)
```

```
## up
## FALSE  TRUE
##   638   2540
```

```
down <- rownames(down_de) %in% rownames(down_limma)
table(down)
```

```
## down
## FALSE  TRUE
##   827   3026
```

```
up = up_de[up,]
head(up)
```

```
##          baseMean log2FoldChange      lfcSE      stat      pvalue
## ST8SIA2  21470.986       7.442980 0.2003001  37.15915 3.119957e-302
## SOX11    105896.449       8.543693 0.2330527  36.65991 3.180749e-294
## SLA      33778.704       6.781894 0.2163388  31.34849 1.020083e-215
## FBN3     11490.654       5.816179 0.2007956  28.96568 1.781244e-184
## MEX3B    12548.477       4.157591 0.1503316  27.65614 2.354317e-168
## VASH2     8164.861       5.920934 0.2269659  26.08732 5.077630e-150
##                padj
## ST8SIA2  4.823766e-298
## SOX11    2.458878e-290
## SLA      5.257168e-212
## FBN3     5.507963e-181
## MEX3B    6.066682e-165
## VASH2    9.813154e-147
```

```
down = down_de[down,]
head(down)
```

```
##          baseMean log2FoldChange      lfcSE      stat      pvalue
## BCL2L2  21968.220      -2.676407  0.08831507 -30.30521 9.788276e-202
## SNCG    9473.167      -8.005651  0.29376717 -27.25169 1.587048e-163
## CLMN    3227.245      -3.760479  0.14437310 -26.04695 1.456668e-149
## UAP1L1  1538.034      -5.591225  0.23073417 -24.23232 1.015679e-129
## ITPKA   5882.483      -7.047788  0.29600941 -23.80934 2.673017e-125
## OPALIN  7380.398     -10.922203  0.45988078 -23.75008 1.096772e-124
##                padj
## BCL2L2  3.783413e-198
## SNCG    3.505336e-160
## CLMN    2.502394e-146
## UAP1L1  1.427583e-126
## ITPKA   3.443960e-122
## OPALIN  1.304400e-121
```

As we can see, there are 2540 common up-regulated genes and 3026 common down-regulated genes between limma package and DESeq package.

In addition to analyzing the correlation in R, I also want to predict and classify some characteristics of those 10 samples such as gender, or age by using Python to analyze their genes from count table. Below is the preparation for that process.

Preparing data for prediction and classification

```
up_down_reg = rbind(up,down)
dim(up_down_reg)
```

```
## [1] 5566      6
```

```
up_down_tr <- (rownames(edata_tr) %in% rownames(up_down_reg))
table(up_down_tr)
```

```
## up_down_tr
## FALSE  TRUE
##  9993  5566
```

```
up_down = edata_tr[up_down_tr,]
head(up_down)
```

```
##          SRR1554534 SRR1554535 SRR1554568 SRR1554561 SRR1554567 SRR1554536
## ADA             8.646385   9.145974   7.303994   8.001912   7.351627   9.770115
## CDH2            13.562468  13.585906  14.957846  13.652418  14.829826  13.812075
## AKT3            13.479431  13.159273  15.876052  13.925121  14.376621  12.332018
## ACOT8           11.946038  11.757942  10.652184  12.014095  10.470291  11.250633
## ZBTB33          11.724310  12.137270  13.056753  11.858819  12.975194  11.759541
## ZSCAN30         11.187060  11.583213  12.612542  11.040663  12.383769  11.699403
```

```
##          SRR1554541 SRR1554539 SRR1554538 SRR1554537
## ADA          7.843839   8.253862   7.831596   7.197164
## CDH2         14.545283  13.717079  14.708753  15.051336
## AKT3         15.204848  14.383520  14.749481  15.584859
## ACOT8        10.522167  11.461451  10.601639  10.700495
## ZBTB33       12.988798  12.410141  13.263124  13.212136
## ZSCAN30      12.182915  11.462477  12.586344  12.277068
```

```
df <- merge(up_down_reg, up_down, by = 0)
row.names(df) <- df$Row.names
df = df[, -1]
head(df)
```

```
##          baseMean log2FoldChange    lfcSE      stat      pvalue      padj
## A2M      13185.3832      -1.670680  0.4204894  -3.973179  7.091974e-05  1.717291e-04
## A2ML1    484.6328      -2.748295  0.5692828  -4.827644  1.381576e-06  4.322248e-06
## A4GALT    412.3115      -2.790862  0.5719644  -4.879432  1.063917e-06  3.379052e-06
## AARD      124.8856      -2.105662  0.5372000  -3.919699  8.865947e-05  2.115377e-04
## AARS1    33645.6493      -1.564934  0.2709131  -5.776517  7.626273e-09  3.150984e-08
## AATK      21134.6178      -3.684225  0.3588961 -10.265435  1.008658e-24  1.959154e-23
##          SRR1554534 SRR1554535 SRR1554568 SRR1554561 SRR1554567 SRR1554536
## A2M      13.657416   13.795969   12.584194   13.003245   12.597035   15.235020
## A2ML1     8.848175    9.062477    6.909735    8.644313    7.457088   10.622862
## A4GALT     8.868851    8.344972    6.895125    8.568670    6.710507   10.362960
## AARD       7.347104    7.314591    5.402734    6.636227    6.395016    7.306073
## AARS1     15.757594   15.345289   14.133470   15.908997   14.037013   14.566318
## AATK       15.412314   15.065392   11.532492   15.465764   11.828298   13.675639
##          SRR1554541 SRR1554539 SRR1554538 SRR1554537
## A2M      12.938130   13.881265   12.930901   12.535831
## A2ML1     6.885452    7.708383    7.500564    7.351208
## A4GALT     7.260562    7.533763    7.292369    6.723675
## AARD       6.484457    7.470666    5.264764    5.737391
## AARS1     14.312565   15.532491   14.157486   14.182287
## AATK      12.056377   14.628722   11.641881   12.151114
```

I will also examine the Python's results of those above 10 samples in wider context by using new 4 samples for the purpose of testing. Below is the preparation of 4 samples in R, which then is processed in Python.

Preparing count_symbol for test sample

```
library(readxl)
```

```
test_ann_count <- read_excel("D:/word/bioinformatics/personal project/PRJNA245228/tidy data/test_ann_count.xlsx")
dup_test <- duplicated(test_ann_count$SYMBOL)
table(dup_test)
```

```
## dup_test
## FALSE  TRUE
## 23741 45153
```

```

test_count = test_ann_count[!dup_test,c(-1,-2,-4)]
na_test <- is.na(test_count$SYMBOL)
test_count = test_count[!na_test,]
test_count = as.data.frame(test_count)
rownames(test_count) <- test_count[,1]
test_count <- test_count[,-1]
head(test_count)

```

```

##          SRR1554555 SRR1554554 SRR1554556 SRR1554566
## A1BG           462         717         702         624
## NAT2            17           5           6           8
## ADA            643         463        1224         367
## CDH2           4899        24385        13301        51721
## AKT3           1498        41020        10917        73863
## GAGE12F          0           0           0           0

```

```

test_sample <- read.csv("D:/word/bioinformatics/personal project/PRJNA245228/sample_data/test_sample.csv")
head(test_sample)

```

```

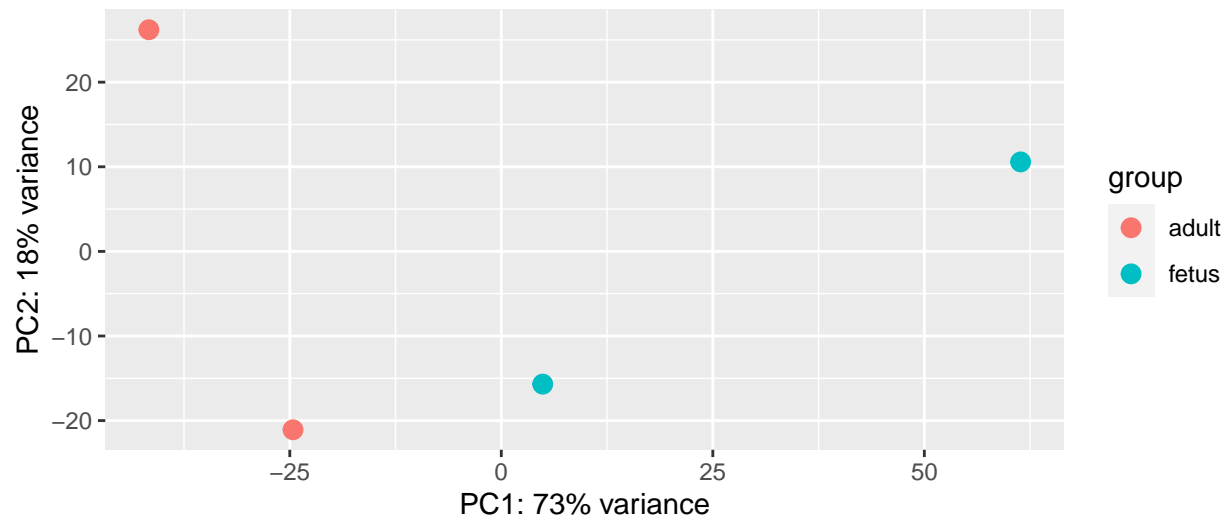
##          Run age_group  age  sex
## 1 SRR1554555    adult 66.72 female
## 2 SRR1554554    fetus  0.30 female
## 3 SRR1554556    adult 36.98  male
## 4 SRR1554566    fetus -0.50  male

```

```

test <- DESeqDataSetFromMatrix(countData = test_count, colData = test_sample, design = ~ age_group)
test_tr <- rlog(test, blind = FALSE)
plotPCA(test_tr, intgroup = c("age_group"))

```

```
test_tr = assay(test_tr)
head(test_tr)
```

```
##          SRR1554555 SRR1554554 SRR1554556 SRR1554566
## A1BG          9.842594  9.116122  9.255977  8.899507
## NAT2          4.064995  3.053218  3.133106  3.140652
## ADA          10.194456  8.701244  9.820673  8.412066
## CDH2          13.666168 14.107108 13.665275 14.802656
## AKT3          12.536838 14.617142 13.432602 15.160773
## GAGE12F       0.000000  0.000000  0.000000  0.000000
```