Analysis of Dataset 3 - Expression profiling by array (GSE87435)

Data Exploration and Preprocessing

```r
head(summary(exprs(abatch1)))[,1:5]
```

```
##  GSM2331292_000149_HGU133A.CEL.gz GSM2331294_000151_HGU133A.CEL.gz
##  Min.   :   54.0                  Min.   :   45.3                 
##  1st Qu.:  125.3                  1st Qu.:  124.0                 
##  Median :  192.5                  Median :  204.5                 
##  Mean   :  358.6                  Mean   :  438.9                 
##  3rd Qu.:  335.0                  3rd Qu.:  394.5                 
##  Max.   :20480.8                  Max.   :22916.5                 
##  GSM2331296_000152_HGU133A.CEL.gz GSM2331298_000153_HGU133A.CEL.gz
##  Min.   :   51.0                  Min.   :   48.5                 
##  1st Qu.:  109.3                  1st Qu.:  135.4                 
##  Median :  171.3                  Median :  233.3                 
##  Mean   :  362.0                  Mean   :  506.6                 
##  3rd Qu.:  315.4                  3rd Qu.:  457.0                 
##  Max.   :21190.8                  Max.   :46139.0                 
##  GSM2331300_000154_HGU133A.CEL.gz
##  Min.   :   44.5                 
##  1st Qu.:  117.5                 
##  Median :  196.0                 
##  Mean   :  429.6                 
##  3rd Qu.:  377.0                 
##  Max.   :28452.0                 
```

```r
dim(exprs(abatch1))
```

```
## [1] 506944     18
```

```r
#arrayQualityMetrics(
 # exprs(abatch1),
#  outdir = "QC_rawdata",
 # force = TRUE
#)
```

```r
#arrayQualityMetrics(
#  exprs(abatch1),
 # outdir = "QC_rawdata_log",
  #force = TRUE,
  #do.logtransform = TRUE
#)
```

```r
HumanRMA <- affy::rma(abatch1)
```

```
## Warning: replacing previous import 'AnnotationDbi::tail' by 'utils::tail' when
## loading 'hgu133acdf'
## Warning: replacing previous import 'AnnotationDbi::head' by 'utils::head' when
## loading 'hgu133acdf'
```

```
##
```

```
## Background correcting
## Normalizing
## Calculating Expression
```

```r
dim(exprs(abatch1))
```

```
## [1] 506944     18
```

```r
dim(HumanRMA)
```

```
## Features  Samples
##    22283       18
```

```r
length(exprs(HumanRMA))
```

```
## [1] 401094
```

```r
my_id <- "GSE87435"
gse <- getGEO(my_id)
```

```
## Found 2 file(s)
```

```
## GSE87435-GPL96_series_matrix.txt.gz
```

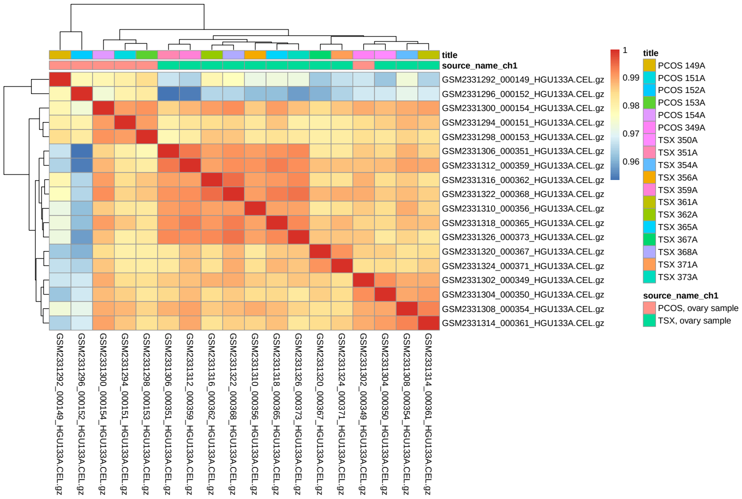
```
## GSE87435-GPL97_series_matrix.txt.gz
```

```r
gse <- gse[[1]] #For first batch
gse2 <- gse[[2]] #For second batch
```

```r
pdata= pData(gse)
fdata = fData(gse)
sampleInfo= pdata[,c("source_name_ch1","title")]
```

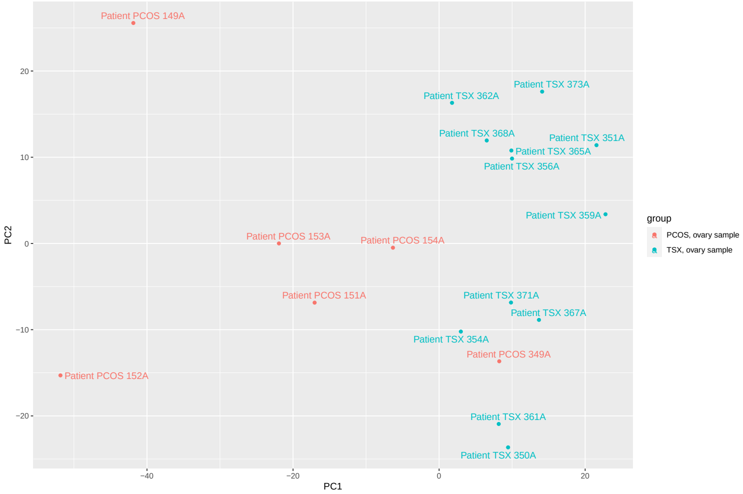
```r
library(pheatmap)
corMatrix <- cor(exprs(HumanRMA),use="c")
```

```r
rownames(sampleInfo) <- colnames(corMatrix)
pheatmap(corMatrix,
         annotation_col=sampleInfo)
```



```r
sampleInfo <- select(pdata, source_name_ch1 ,title)
sampleInfo<- rename(sampleInfo,group =source_name_ch1 , patient= title)
```

```r
pca <- prcomp(t(exprs(HumanRMA)))

## Join the PCs to the sample information
cbind(sampleInfo, pca$x) %>%
ggplot(aes(x = PC1, y=PC2, col=group,label=paste("Patient", patient))) + geom_point() + geom_text_repel
```



Differential Expression Analysis by LIMMA

```r
group = factor(pdata$source_name_ch1)
design = model.matrix(~ 0 + group)
colnames(design) = c("PCOS","Normal")
fit <- lmFit(exprs(HumanRMA), design)
head(fit$coefficients)
```

```
##               PCOS   Normal
## 1007_s_at 9.264967 9.038700
## 1053_at   6.631785 6.802279
## 117_at    6.693852 6.631233
## 121_at    9.326375 9.114591
## 1255_g_at 4.619216 4.523439
## 1294_at   8.849684 8.783237
```

```
matrix_data = makeContrasts(Normal-PCOS ,levels = design)
fit1 = contrasts.fit(fit,matrix_data)
fit2 = eBayes(fit1)
```

```
Limma_out = topTable(fit2,coef = 1,adjust.method = "BH",number = "Inf")
decideTests(fit2)
```

```
## TestResults matrix
##             Contrasts
##              Normal - PCOS
##   1007_s_at             0
##   1053_at               0
##   117_at                0
##   121_at                0
##   1255_g_at             0
## 22278 more rows ...
```

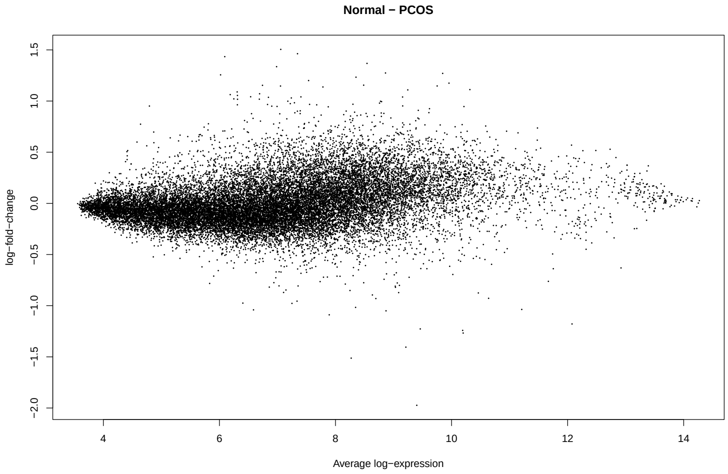
```
table(decideTests(fit2))
```

```
##
##    -1     0     1
##  2773 16945  2565
```

```
head(Limma_out)
```

```
##                   logFC  AveExpr         t      P.Value   adj.P.Val        B
## 64474_g_at  -0.5265489 7.827172 -7.880734 1.770671e-07 0.003483982 7.292687
## 217991_x_at -0.7757138 8.485872 -7.585632 3.127032e-07 0.003483982 6.781208
## 201990_s_at  0.6585609 7.742738  6.881356 1.274015e-06 0.003727136 5.504621
## 204433_s_at -0.5080181 6.093444 -6.820683 1.442408e-06 0.003727136 5.390974
## 160020_at   -0.5113104 8.670438 -6.800017 1.504875e-06 0.003727136 5.352133
## 221354_s_at -0.4182821 6.287904 -6.764327 1.619411e-06 0.003727136 5.284896
```

```
limma::plotMA(fit2)
```



```
sig_prob_names = rownames(head(Limma_out,1))
row_name_ematrix = rownames(exprs(HumanRMA))
row_selector = row_name_ematrix %in% sig_prob_names
e = exprs(HumanRMA)[row_selector,]
e = exprs(HumanRMA)[rownames(HumanRMA)%in% rownames(head(Limma_out,1)),]
```

```
rowMeans(exprs(HumanRMA)[rownames(exprs(HumanRMA)) %in%rownames(head(Limma_out,5)),1:6])
```

```
##   160020_at 201990_s_at 204433_s_at 217991_x_at   64474_g_at
##    9.011312    7.303697    6.432123    9.003014    8.178205
```

```
rowMeans(exprs(HumanRMA)[rownames(exprs(HumanRMA)) %in%rownames(head(Limma_out,5)),7:18])
```

```
##   160020_at 201990_s_at 204433_s_at 217991_x_at   64474_g_at
##    8.500001    7.962258    5.924104    8.227300    7.651656
```

```
probe_name = fdata$ID
sorted_indx = sort(probe_name,index.return = T)$ix
annot_ma = fdata[sorted_indx,]
dim(annot_ma)
```

```
## [1] 22283    16
```

```
dim(Limma_out)
```

```
## [1] 22283     6
```

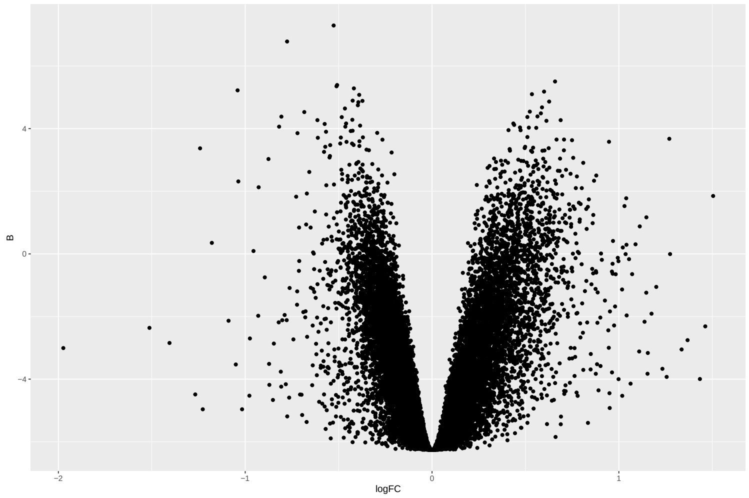
```
Limma_out_sorted = Limma_out[sort(rownames(Limma_out),index.return = T)$ix,]

Limma_out_sorted$gene = annot_ma$`Gene Symbol`

Limma_out_annot = Limma_out_sorted[sort(Limma_out_sorted$adj.P.Val,index.return = T)$ix,]

head(Limma_out_annot)
```

```
##                   logFC  AveExpr         t      P.Value   adj.P.Val        B
## 217991_x_at -0.7757138 8.485872 -7.585632 3.127032e-07 0.003483982 6.781208
## 64474_g_at  -0.5265489 7.827172 -7.880734 1.770671e-07 0.003483982 7.292687
## 160020_at   -0.5113104 8.670438 -6.800017 1.504875e-06 0.003727136 5.352133
## 200702_s_at  0.5643422 7.197586  6.298666 4.282748e-06 0.003727136 4.389659
## 201200_at    0.5234048 9.607172  6.376655 3.631814e-06 0.003727136 4.541891
## 201556_s_at -0.3724101 6.789247 -6.554093 2.503270e-06 0.003727136 4.884825
##                          gene
## 217991_x_at             SSBP3
## 64474_g_at  DGCR8 /// MIR1306
## 160020_at               MMP14
## 200702_s_at             DDX24
## 201200_at               CREG1
## 201556_s_at             VAMP2
```

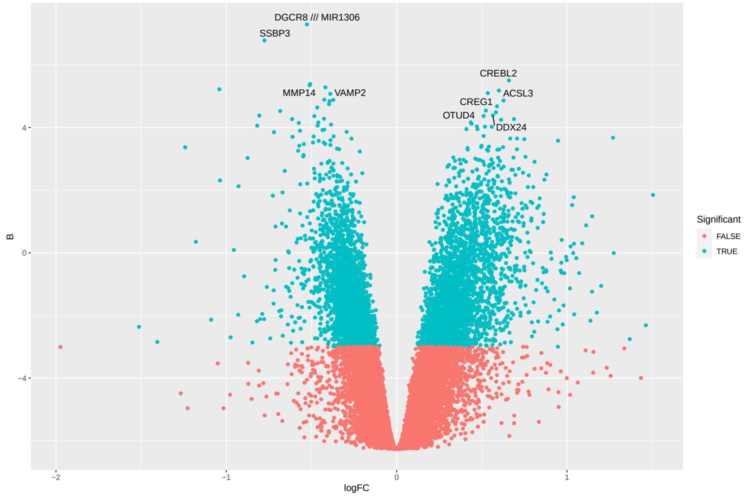
```
selected_columns = Limma_out_annot %>% select(adj.P.Val,logFC,gene)
head(selected_columns)
```

```
##               adj.P.Val      logFC              gene
## 217991_x_at 0.003483982 -0.7757138             SSBP3
## 64474_g_at  0.003483982 -0.5265489 DGCR8 /// MIR1306
## 160020_at   0.003727136 -0.5113104             MMP14
## 200702_s_at 0.003727136  0.5643422             DDX24
## 201200_at   0.003727136  0.5234048             CREG1
## 201556_s_at 0.003727136 -0.3724101             VAMP2
```

```
ggplot(Limma_out,aes(x = logFC, y=B)) + geom_point()
```



```
p_cutoff <- 0.05
fc_cutoff <- 1.5
topN <- 10

Limma_out_annot %>%
  mutate(Significant = adj.P.Val < p_cutoff, abs(logFC) > fc_cutoff ) %>%
  mutate(Rank = 1:n(), Label = ifelse(Rank < topN,gene,"")) %>%
  ggplot(aes(x = logFC, y = B, col=Significant,label=Label)) + geom_point() + geom_text_repel(col="blac
```



GENE SET ANALYSIS

```
limma_out_filtered = Limma_out_annot[Limma_out_annot$adj.P.Val < 0.05,]
```

```
print(paste("Amount of significant genes:", as.numeric(nrow(limma_out_filtered))))
```

```
## [1] "Amount of significant genes: 5338"
```

```
#for all genes.
EntrezIDs = mapIds(org.Hs.eg.db,gsub("///.*","",limma_out_filtered$gene),"ENTREZID",keytype = "SYMBOL")
```

```
## 'select()' returned 1:1 mapping between keys and columns
```

```
length(EntrezIDs)
```

```
## [1] 5338
```

Unique genes

```
ids_unique = unlist(EntrezIDs[!(duplicated(EntrezIDs) | is.na(EntrezIDs))])
length(ids_unique)
```

```
## [1] 3650
```

Overrepresentation analysis

```
goanaOut = goana(de = ids_unique, species = "Hs",trend =T)
head(goanaOut)
```

```
##                                                                     Term Ont
## GO:0008150                                           biological_process  BP
## GO:0000003                                                 reproduction  BP
## GO:0001553                                                luteinization  BP
## GO:0001867                        complement activation, lectin pathway  BP
## GO:0001868          regulation of complement activation, lectin pathway  BP
## GO:0001869 negative regulation of complement activation, lectin pathway  BP
##                N   DE         P.DE
## GO:0008150 18614 3444 1.546001e-47
## GO:0000003  1506  330 1.315023e-06
## GO:0001553    12    3 3.466790e-01
## GO:0001867    11    1 8.770201e-01
## GO:0001868     2    0 1.000000e+00
## GO:0001869     2    0 1.000000e+00
```

```
goanaOut = goanaOut[order(goanaOut$P.DE, decreasing = TRUE),]
goanaOut$FDR.DE = p.adjust(goanaOut$P.DE,method = "BH")
topGO = topGO(goanaOut,ontology = "MF",number = 100)
head(topGO)
```

```
##                                       Term Ont     N   DE          P.DE
## GO:0005515                 protein binding  MF 13998 2976 2.009422e-112
## GO:0003674              molecular_function  MF 18369 3490  7.290934e-91
## GO:0005488                         binding  MF 16581 3273  1.673113e-86
## GO:0097159 organic cyclic compound binding  MF  6217 1397  3.151150e-36
## GO:1901363   heterocyclic compound binding  MF  6147 1382  8.170726e-36
## GO:0019899                  enzyme binding  MF  2068  556  1.349356e-30
##                   FDR.DE
## GO:0005515 4.608410e-108
## GO:0003674  8.360514e-87
## GO:0005488  1.279039e-82
## GO:0097159  2.007458e-33
## GO:1901363  5.064525e-33
## GO:0019899  6.189224e-28
```

```
goanaOut_MF <- goanaOut[goanaOut$Ont == "MF",]
print(paste("Amount of significant GO MF terms:",
as.character(sum(goanaOut_MF$FDR.DE < 0.05))))
```

```
## [1] "Amount of significant GO MF terms: 186"
```

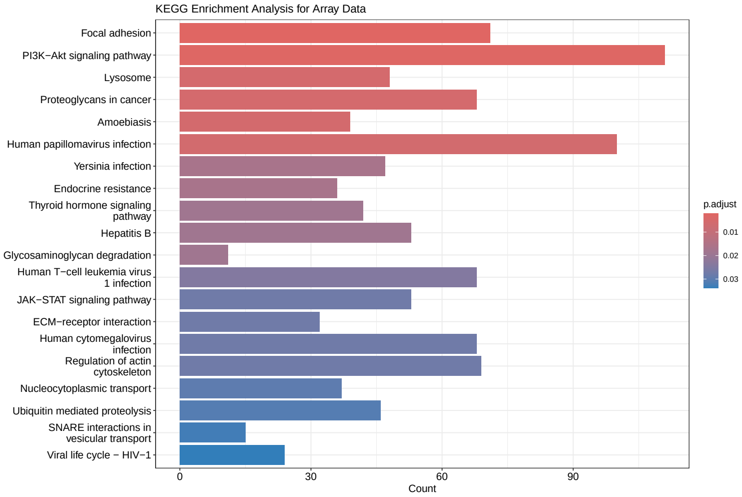
```
sign_genes = goanaOut_MF[which(goanaOut_MF$FDR.DE < 0.05),]
```

```
# Load the packages

# Perform KEGG enrichment analysis with a p-value cutoff of 0.05
kegg_enrichment <- enrichKEGG(gene = ids_unique, organism = "hsa", keyType = "kegg", pvalueCutoff = 0.05
```

```
## Reading KEGG annotation online: "https://rest.kegg.jp/link/hsa/pathway"...
```

```
## Reading KEGG annotation online: "https://rest.kegg.jp/list/pathway/hsa"...
```

```
par(mar = c(5, 8, 4, 2) + 0.1)  # Adjust the margin to accommodate longer labels
barplot(kegg_enrichment, showCategory = 20, title = "KEGG Enrichment Analysis for Array Data", xlab = "
```


KEGG Enrichment Analysis for Array Data
Focal adhesion
PI3K−Akt signaling pathway
Lysosome
Proteoglycans in cancer
Amoebiasis
Human papillomavirus infection
Yersinia infection
Endocrine resistance
Thyroid hormone signaling pathway
Hepatitis B
Glycosaminoglycan degradation
Human T−cell leukemia virus 1 infection
JAK−STAT signaling pathway
ECM−receptor interaction
Human cytomegalovirus infection
Regulation of actin cytoskeleton
Nucleocytoplasmic transport
Ubiquitin mediated proteolysis
SNARE interactions in vesicular transport
Viral life cycle − HIV−1
Count
p.adjust