Analysis of Dataset 1 - Expression profiling by Array (GSE124226)

2023-12-22

Data Exploration and Preprocessing

```
# Set the CRAN mirror
options(repos = c(CRAN = "https://cran.rstudio.com"))

# Download the package of interest
BiocManager::install('ArrayExpress')
BiocManager::install("affy")
BiocManager::install("limma")
BiocManager::install("siggenes")
BiocManager::install("arrayQualityMetrics")
BiocManager::install("GEOquery")
BiocManager::install("hgu133plus2.db")

#Load the package
library('ArrayExpress')
library("affy")
library("limma")
library("siggenes")
library("arrayQualityMetrics")
library("GEOquery")
library("hgu133plus2.db")
```

The celfiles were downloaded from the NCBI GEO website. Sample info was added manually as the metadata was not present.

```
celpath = "C:/Users/maxim/OneDrive - UGent/1 Master Bio-informatics/Semester 1/Applied High Throughput A

sample_info <- data.frame(
  SampleNumber = c("1", "2", "3", "4", "5", "6", "7", "8"),
  SampleName = c("GSM3526020_3004_Control", "GSM3526021_3006_PCOS", "GSM3526022_3010_PCOS", "GSM3526023_
                 "GSM3526024_3027_Control", "GSM3526025_3034_Control", "GSM3526026_3036_Control", "GSM35
  Condition = c("Control", "PCOS", "PCOS", "PCOS", "Control", "Control", "Control", "PCOS") )

# Load in the Affybatch object
PCOSExp <- ReadAffy(celfile.path=celpath, phenoData=sample_info)
```

```
## Warning: Mismatched phenoData and celfile names!
##
## Please note that the row.names of your phenoData object should be identical to what you get from list
## Otherwise you are responsible for ensuring that the ordering of your phenoData object conforms to the
## If not, errors may result from using the phenoData for subsetting or creating linear models, etc.
```

```r
# Have a look to the data you just loaded
head(exprs(PCOSExp))
```

```
##       1     2     3     4     5     6     7     8
## 1   223   136   117   172    94   164   172   126
## 2 14476 14691 15068 12624 14743 16990 20960 14976
## 3   300   215   223   197   154   224   223   188
## 4 14718 14719 15144 12676 14729 16629 20299 14924
## 5   117    80    75    75   116   135    79    77
## 6   163   132   115   121    88   130    71   118
```

```r
pData(PCOSExp)
```

```
##   SampleNumber              SampleName Condition
## 1            1 GSM3526020_3004_Control   Control
## 2            2    GSM3526021_3006_PCOS      PCOS
## 3            3    GSM3526022_3010_PCOS      PCOS
## 4            4    GSM3526023_3019_PCOS      PCOS
## 5            5 GSM3526024_3027_Control   Control
## 6            6 GSM3526025_3034_Control   Control
## 7            7 GSM3526026_3036_Control   Control
## 8            8    GSM3526027_3042_PCOS      PCOS
```

Quality control

```r
arrayQualityMetrics(PCOSExp, intgroup = colnames(pData(PCOSExp))[3], outdir = "QC_rawdata", force = TRU
```

```
## The report will be written into directory 'QC_rawdata'.

## Warning: replacing previous import 'AnnotationDbi::head' by 'utils::head' when
## loading 'hgu133plus2cdf'

## Warning: replacing previous import 'AnnotationDbi::tail' by 'utils::tail' when
## loading 'hgu133plus2cdf'

##

## Warning in svgStyleAttributes(style, svgdev): Removing non-SVG style attribute
## name(s): subscripts, group.number, group.value

## Warning in svgStyleAttributes(style, svgdev): Removing non-SVG style attribute
## name(s): subscripts, group.number, group.value

## Warning in svgStyleAttributes(style, svgdev): Removing non-SVG style attribute
## name(s): subscripts, group.number, group.value

## Warning in svgStyleAttributes(style, svgdev): Removing non-SVG style attribute
## name(s): subscripts, group.number, group.value
```

```
## Warning in svgStyleAttributes(style, svgdev): Removing non-SVG style attribute
## name(s): subscripts, group.number, group.value

## Warning in svgStyleAttributes(style, svgdev): Removing non-SVG style attribute
## name(s): subscripts, group.number, group.value

## Warning in svgStyleAttributes(style, svgdev): Removing non-SVG style attribute
## name(s): subscripts, group.number, group.value

## Warning in svgStyleAttributes(style, svgdev): Removing non-SVG style attribute
## name(s): subscripts, group.number, group.value

## (loaded the KernSmooth namespace)

## Warning in KernSmooth::bkde2D(x, gridsize = nbin, bandwidth = bandwidth):
## Binning grid too coarse for current (small) bandwidth: consider increasing
## 'gridsize'

## Warning in KernSmooth::bkde2D(x, gridsize = nbin, bandwidth = bandwidth):
## Binning grid too coarse for current (small) bandwidth: consider increasing
## 'gridsize'

## Warning in KernSmooth::bkde2D(x, gridsize = nbin, bandwidth = bandwidth):
## Binning grid too coarse for current (small) bandwidth: consider increasing
## 'gridsize'

## Warning in KernSmooth::bkde2D(x, gridsize = nbin, bandwidth = bandwidth):
## Binning grid too coarse for current (small) bandwidth: consider increasing
## 'gridsize'

## Warning in KernSmooth::bkde2D(x, gridsize = nbin, bandwidth = bandwidth):
## Binning grid too coarse for current (small) bandwidth: consider increasing
## 'gridsize'

## Warning in KernSmooth::bkde2D(x, gridsize = nbin, bandwidth = bandwidth):
## Binning grid too coarse for current (small) bandwidth: consider increasing
## 'gridsize'

## Warning in KernSmooth::bkde2D(x, gridsize = nbin, bandwidth = bandwidth):
## Binning grid too coarse for current (small) bandwidth: consider increasing
## 'gridsize'

## Warning in KernSmooth::bkde2D(x, gridsize = nbin, bandwidth = bandwidth):
## Binning grid too coarse for current (small) bandwidth: consider increasing
## 'gridsize'

## Warning in KernSmooth::bkde2D(x, gridsize = nbin, bandwidth = bandwidth):
## Binning grid too coarse for current (small) bandwidth: consider increasing
## 'gridsize'

## Warning in KernSmooth::bkde2D(x, gridsize = nbin, bandwidth = bandwidth):
## Binning grid too coarse for current (small) bandwidth: consider increasing
```

```
## 'gridsize'

## Warning in KernSmooth::bkde2D(x, gridsize = nbin, bandwidth = bandwidth):
## Binning grid too coarse for current (small) bandwidth: consider increasing
## 'gridsize'

## Warning in KernSmooth::bkde2D(x, gridsize = nbin, bandwidth = bandwidth):
## Binning grid too coarse for current (small) bandwidth: consider increasing
## 'gridsize'

## Warning in KernSmooth::bkde2D(x, gridsize = nbin, bandwidth = bandwidth):
## Binning grid too coarse for current (small) bandwidth: consider increasing
## 'gridsize'

## Warning in KernSmooth::bkde2D(x, gridsize = nbin, bandwidth = bandwidth):
## Binning grid too coarse for current (small) bandwidth: consider increasing
## 'gridsize'

## Warning in KernSmooth::bkde2D(x, gridsize = nbin, bandwidth = bandwidth):
## Binning grid too coarse for current (small) bandwidth: consider increasing
## 'gridsize'

## Warning in KernSmooth::bkde2D(x, gridsize = nbin, bandwidth = bandwidth):
## Binning grid too coarse for current (small) bandwidth: consider increasing
## 'gridsize'
```

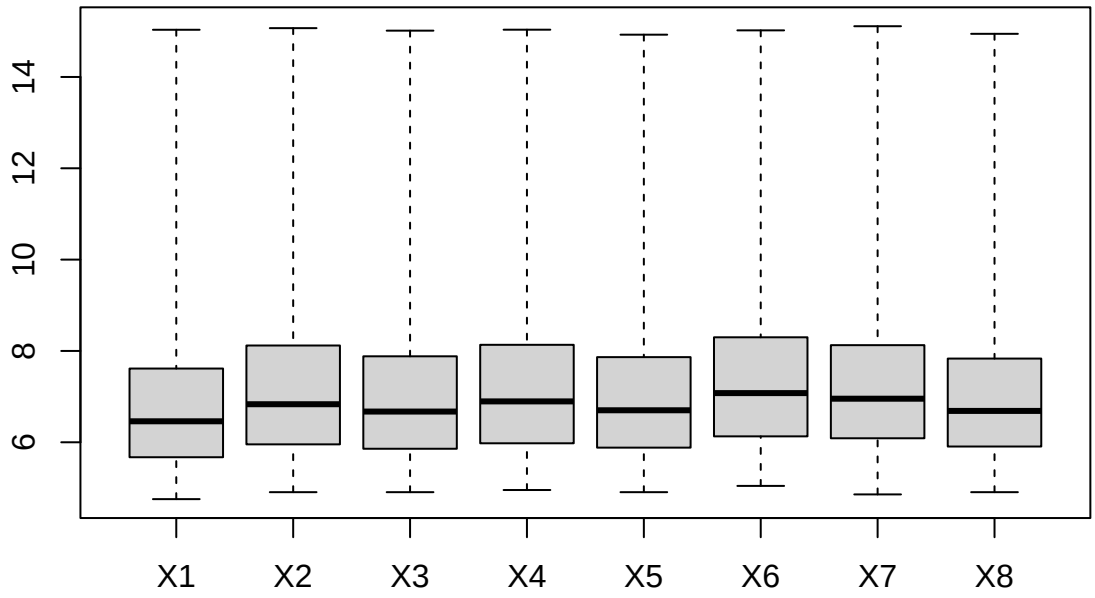
```
arrayQualityMetrics(PCOSExp, intgroup = colnames(pData(PCOSExp))[3], outdir = "QC_rawdata_log", force =
```

```
## The report will be written into directory 'QC_rawdata_log'.

## Warning in svgStyleAttributes(style, svgdev): Removing non-SVG style attribute
## name(s): subscripts, group.number, group.value

## Warning in svgStyleAttributes(style, svgdev): Removing non-SVG style attribute
## name(s): subscripts, group.number, group.value

## Warning in svgStyleAttributes(style, svgdev): Removing non-SVG style attribute
## name(s): subscripts, group.number, group.value

## Warning in svgStyleAttributes(style, svgdev): Removing non-SVG style attribute
## name(s): subscripts, group.number, group.value

## Warning in svgStyleAttributes(style, svgdev): Removing non-SVG style attribute
## name(s): subscripts, group.number, group.value

## Warning in svgStyleAttributes(style, svgdev): Removing non-SVG style attribute
## name(s): subscripts, group.number, group.value

## Warning in svgStyleAttributes(style, svgdev): Removing non-SVG style attribute
## name(s): subscripts, group.number, group.value

## Warning in svgStyleAttributes(style, svgdev): Removing non-SVG style attribute
## name(s): subscripts, group.number, group.value
```

Preprocessing

A boxplot is used to check the distribution of the expression data.

```
boxplot(PCOSExp)
```



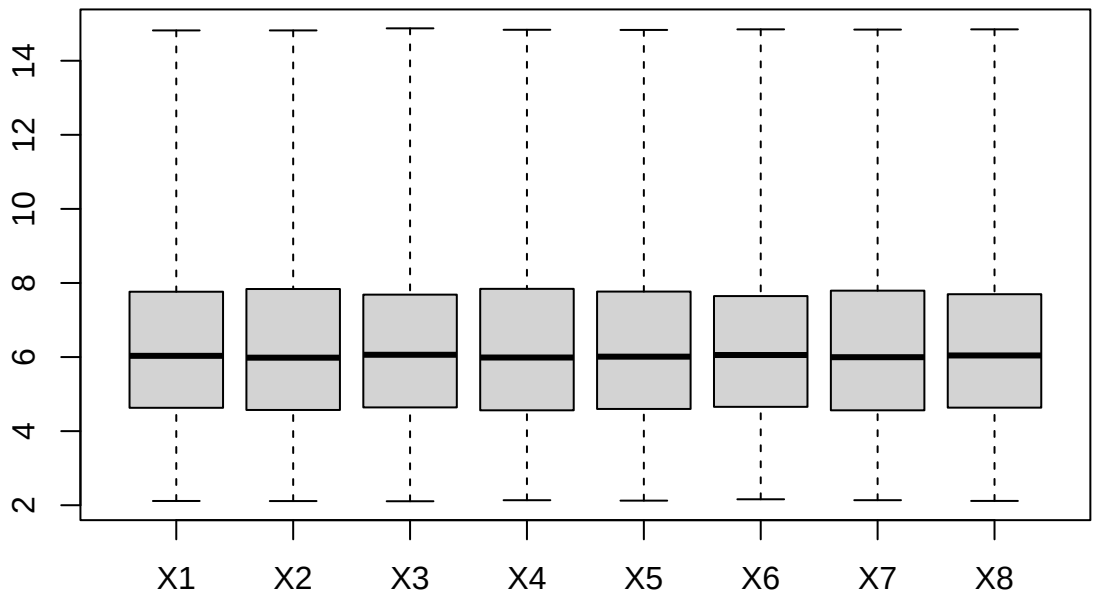
Although the boxplots are similar, they can be further optimized. Using the rma function of the affy package, background correction and quantile normalisation is performed.

```
# Background correction, quantile normalisation
PCOSRMA <- affy::rma(PCOSExp)
```

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

Again, a boxplot is used to check the distribution of the expression data that has been background corrected and quantile normalised.

```
boxplot(PCOSRMA)
```



The boxplots are now ideal, showing an equal mean, equal quantiles and equal variances.

Quality control on processed data

RMA produces logtransformed data => do.logtransform = FALSE

```
arrayQualityMetrics(PCOSRMA, intgroup = colnames(pData(PCOSExp))[3], outdir = "QC_preprocessed", force =
```

```
## The report will be written into directory 'QC_preprocessed'.

## Warning in svgStyleAttributes(style, svgdev): Removing non-SVG style attribute
## name(s): subscripts, group.number, group.value

## Warning in svgStyleAttributes(style, svgdev): Removing non-SVG style attribute
## name(s): subscripts, group.number, group.value

## Warning in svgStyleAttributes(style, svgdev): Removing non-SVG style attribute
## name(s): subscripts, group.number, group.value

## Warning in svgStyleAttributes(style, svgdev): Removing non-SVG style attribute
## name(s): subscripts, group.number, group.value

## Warning in svgStyleAttributes(style, svgdev): Removing non-SVG style attribute
## name(s): subscripts, group.number, group.value
```

```
## Warning in svgStyleAttributes(style, svgdev): Removing non-SVG style attribute
## name(s): subscripts, group.number, group.value

## Warning in svgStyleAttributes(style, svgdev): Removing non-SVG style attribute
## name(s): subscripts, group.number, group.value

## Warning in svgStyleAttributes(style, svgdev): Removing non-SVG style attribute
## name(s): subscripts, group.number, group.value
```

Differential Expression Analysis

A design matrix is made. There is no intercept.

```
design <- model.matrix(~ 0 + factor(c(1,2,2,2,1,1,1,2)))
colnames(design) <- c("control", "PCOS")
design
```

```
##   control PCOS
## 1       1    0
## 2       0    1
## 3       0    1
## 4       0    1
## 5       1    0
## 6       1    0
## 7       1    0
## 8       0    1
## attr(,"assign")
## [1] 1 1
## attr(,"contrasts")
## attr(,"contrasts")$`factor(c(1, 2, 2, 2, 1, 1, 1, 2))`
## [1] "contr.treatment"
```

A linear model is fit to the expression data.

```
fit <- lmFit(PCOSRMA, design)
contrast.matrix <- makeContrasts(PCOS-control, levels=design)
fit2 <- contrasts.fit(fit, contrast.matrix)
fit2 <- eBayes(fit2)
```

Top Genes are called.

```
LIMMAout <- topTable(fit2,
                     coef = 1, # which contrast do you want to test, 1 is without intercept, 2 is with
                     adjust = 'BH', # Benjamini-Hochberg procedure
                     number=Inf)
head(LIMMAout)
```

```
##                  logFC  AveExpr         t      P.Value adj.P.Val         B
## 219073_s_at -0.9404248 8.169289 -6.786883 7.062747e-05 0.9999858 -2.905081
## 213802_at   -0.8614519 7.877965 -6.616723 8.610008e-05 0.9999858 -2.925224
```

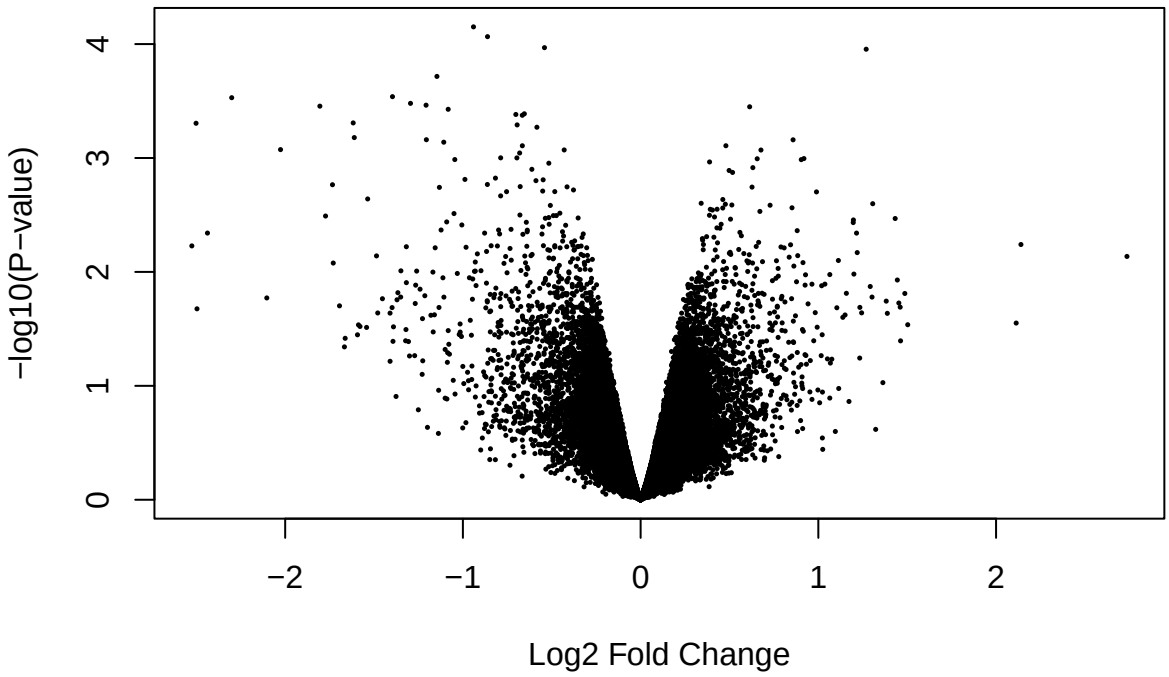
```
## 203200_s_at -0.5410565 9.625066 -6.429321 1.075287e-04 0.9999858 -2.948733
## 213975_s_at  1.2691885 5.194109  6.402909 1.109890e-04 0.9999858 -2.952163
## 204320_at   -1.1461782 7.829749 -5.955744 1.923307e-04 0.9999858 -3.015029
## 37892_at    -1.3963016 8.254226 -5.636194 2.895403e-04 0.9999858 -3.066081
```

Volcano plot

```
jpeg("Volcanoplotdataset1.jpg")
limma::volcanoplot(fit2)
dev.off()
```

```
## pdf
##   2
```

```
limma::volcanoplot(fit2)
```

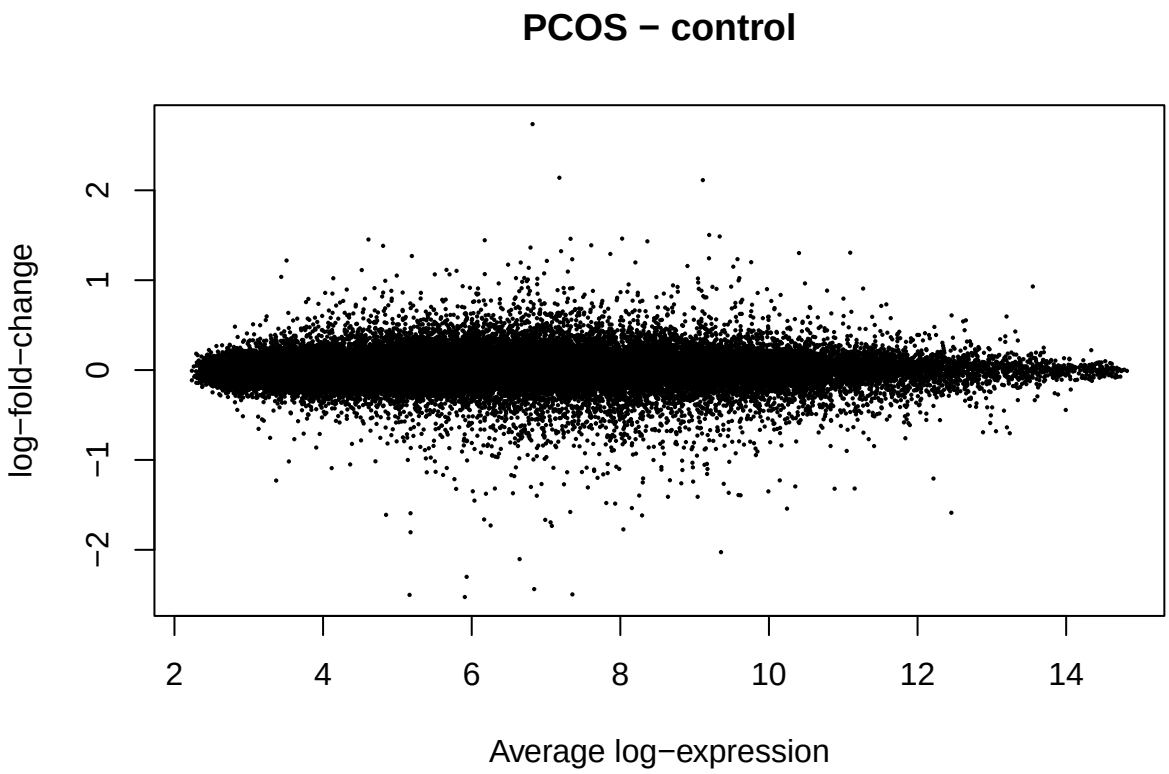


MA plot

```
jpeg("MAplotdataset1.jpg")
limma::plotMA(fit2)
dev.off()
```

```
## pdf
##   2
```

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



Gene Set Analysis

Gene annotation

As the package WebGestaltR is used, which requires EntrezIDs or Ensembl IDs to perform an Over Representation Analysis, annotation of the genes is performed first.

```
# Get gene annotations based on probe IDs
probe_ids <- rownames(LIMMAout)
gene_annotations <- select(hgu133plus2.db, keys = probe_ids, columns = c("SYMBOL", "GENENAME", "ENTREZI
```

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

```
duplicate_probes <- probe_ids[duplicated(gene_annotations$PROBEID)]
cat("Duplicate probe IDs found:", length(duplicate_probes), "\n")
```

```
## Duplicate probe IDs found: 2480
```

```r
# Remove duplicate rows (keeping the first occurrence)
gene_annotations_unique <- gene_annotations[!duplicated(gene_annotations$PROBEID), ]

# Sort gene_annotations_unique by probe name alphabetically
gene_annotations_unique_sorted <- gene_annotations_unique[order(gene_annotations_unique$PROBEID), ]

# Sort LIMMA output alphabetically on probe name
LIMMAout_sorted <- LIMMAout[sort(rownames(LIMMAout), index.return=T)$ix, ]

# Add gene names to LIMMA output
LIMMAout_sorted$Gene <- gene_annotations_unique_sorted$GENENAME
LIMMAout_sorted$Symbol <- gene_annotations_unique_sorted$SYMBOL
LIMMAout_sorted$EntrezID <- gene_annotations_unique_sorted$ENTREZID
LIMMAout_sorted$Gene <- as.character(LIMMAout_sorted$Gene)
LIMMAout_sorted$Symbol <- as.character(LIMMAout_sorted$Symbol)
LIMMAout_sorted$EntrezID <- as.character(LIMMAout_sorted$EntrezID)

LIMMAout_annot <- LIMMAout_sorted[sort(LIMMAout_sorted$P.Value, index.return = T)$ix,]
head(LIMMAout_annot)
```

```
##                  logFC  AveExpr         t      P.Value adj.P.Val         B
## 219073_s_at -0.9404248 8.169289 -6.786883 7.062747e-05 0.9999858 -2.905081
## 213802_at   -0.8614519 7.877965 -6.616723 8.610008e-05 0.9999858 -2.925224
## 203200_s_at -0.5410565 9.625066 -6.429321 1.075287e-04 0.9999858 -2.948733
## 213975_s_at  1.2691885 5.194109  6.402909 1.109890e-04 0.9999858 -2.952163
## 204320_at   -1.1461782 7.829749 -5.955744 1.923307e-04 0.9999858 -3.015029
## 37892_at    -1.3963016 8.254226 -5.636194 2.895403e-04 0.9999858 -3.066081
##                                                                        Gene
## 219073_s_at                                    oxysterol binding protein like 10
## 213802_at                                                   serine protease 12
## 203200_s_at 5-methyltetrahydrofolate-homocysteine methyltransferase reductase
## 213975_s_at                                                         lysozyme
## 204320_at                                     collagen type XI alpha 1 chain
## 37892_at                                      collagen type XI alpha 1 chain
##              Symbol EntrezID
## 219073_s_at OSBPL10   114884
## 213802_at    PRSS12     8492
## 203200_s_at    MTRR     4552
## 213975_s_at     LYZ     4069
## 204320_at   COL11A1     1301
## 37892_at    COL11A1     1301
```

Over Representation Analysis

The top 300 genes are selected based on p-value and logFC, these genes are most likely to be differently expressed.

```r
logFC_interest <- LIMMAout_annot[abs(LIMMAout_annot$logFC) > 0.7, ]
logFC_interest_sorted <- logFC_interest[sort(logFC_interest$P.Value, index.return = T)$ix, ]
top_genes <- head(logFC_interest_sorted, 300)
head(top_genes)
```

```
##                  logFC  AveExpr         t      P.Value adj.P.Val         B
## 219073_s_at -0.9404248 8.169289 -6.786883 7.062747e-05 0.9999858 -2.905081
## 213802_at   -0.8614519 7.877965 -6.616723 8.610008e-05 0.9999858 -2.925224
## 213975_s_at  1.2691885 5.194109  6.402909 1.109890e-04 0.9999858 -2.952163
## 204320_at   -1.1461782 7.829749 -5.955744 1.923307e-04 0.9999858 -3.015029
## 37892_at    -1.3963016 8.254226 -5.636194 2.895403e-04 0.9999858 -3.066081
## 241404_at   -2.3008717 5.931977 -5.620142 2.956602e-04 0.9999858 -3.068793
##                                          Gene  Symbol EntrezID
## 219073_s_at oxysterol binding protein like 10 OSBPL10   114884
## 213802_at                  serine protease 12  PRSS12     8492
## 213975_s_at                           lysozyme     LYZ     4069
## 204320_at      collagen type XI alpha 1 chain COL11A1     1301
## 37892_at       collagen type XI alpha 1 chain COL11A1     1301
## 241404_at                                 <NA>    <NA>     <NA>
```

From these genes, we select the upregulated and downregulated genes. We store the entrez ids of these genes for further use in the Over Representation Analysis. The reference gene set is equal to all genes in the array.

```
# Upregulated genes
upreg_genes <- top_genes$EntrezID[top_genes$logFC > 0]
upreg_genes <- upreg_genes[!is.na(upreg_genes)]
cat("Number of downregulated genes:", length(upreg_genes), "\n")
```

```
## Number of downregulated genes: 115
```

```
# Downregulated genes
downreg_genes <- top_genes$EntrezID[top_genes$logFC < 0]
downreg_genes <- downreg_genes[!is.na(downreg_genes)]
cat("Number of downregulated genes:", length(downreg_genes), "\n")
```

```
## Number of downregulated genes: 176
```

```
# Reference genes
ref_genes <- LIMMAout_annot$EntrezID[!is.na(LIMMAout_annot$EntrezID)]
cat("Number of reference genes:", length(ref_genes), "\n")
```

```
## Number of reference genes: 44662
```

KEGG

```
# Perform KEGG pathway analysis on the upregulated genes
kegg_upreg <- limma::kegga(de = upreg_genes, universe = ref_genes, species = "Hs")

# Sort results and calculate FDR
kegg_upreg <- kegg_upreg[sort(kegg_upreg$P.DE, index.return=T)$ix, ]
kegg_upreg$P.DE.adj <- p.adjust(kegg_upreg$P.DE, n=nrow(kegg_upreg), method="BH")
head(kegg_upreg)
```

```
##                                                    Pathway   N DE
```

```
## hsa04061 Viral protein interaction with cytokine and cytokine receptor  96  6
## hsa04062                                   Chemokine signaling pathway 187  6
## hsa05323                                         Rheumatoid arthritis  88  4
## hsa04657                                     IL-17 signaling pathway  93  4
## hsa04668                                       TNF signaling pathway 114  4
## hsa04060                        Cytokine-cytokine receptor interaction 291  6
##                  P.DE    P.DE.adj
## hsa04061 3.907414e-06 0.001391039
## hsa04062 1.681870e-04 0.029937282
## hsa05323 5.931269e-04 0.065011427
## hsa04657 7.304655e-04 0.065011427
## hsa04668 1.557558e-03 0.101755325
## hsa04060 1.714977e-03 0.101755325
```

```
# Perform KEGG pathway analysis on the downregulated genes
kegg_downreg <- limma::kegga(de = downreg_genes, universe = ref_genes, species = "Hs")

# Print or further process the results
# Sort results and calculate FDR
kegg_downreg <- kegg_downreg[sort(kegg_downreg$P.DE, index.return=T)$ix, ]
kegg_downreg$P.DE.adj <- p.adjust(kegg_downreg$P.DE, n=nrow(kegg_downreg), method="BH")
head(kegg_downreg)
```

```
##                              Pathway   N DE         P.DE  P.DE.adj
## hsa05200          Pathways in cancer 526 11 0.0004028282 0.1434068
## hsa05224               Breast cancer 145  5 0.0020154816 0.2680263
## hsa04390     Hippo signaling pathway 154  5 0.0026180917 0.2680263
## hsa05410 Hypertrophic cardiomyopathy  97  4 0.0030115318 0.2680263
## hsa04310       Wnt signaling pathway 169  5 0.0038979973 0.2775374
## hsa04360               Axon guidance 181  5 0.0052042787 0.3087872
```

Gene Ontology: Molecular Function (MF)

```
# Perform Gene Ontology: Molecular Function Analysis on the upregulated genes
goana_upreg <- limma::goana(de = upreg_genes, universe = ref_genes, species = "Hs")

# Print or further process the results
goana_upreg <- goana_upreg[goana_upreg$Ont == "MF", ]
goana_upreg <- goana_upreg[sort(goana_upreg$P.DE, index.return=T)$ix, ]
goana_upreg$P.DE.adj <- p.adjust(goana_upreg$P.DE, n=nrow(goana_upreg), method="BH")
head(goana_upreg)
```

```
##                                         Term Ont   N DE         P.DE
## GO:0008009                 chemokine activity  MF  46  5 1.668135e-06
## GO:0042379         chemokine receptor binding  MF  63  5 8.069423e-06
## GO:0004962       endothelin receptor activity  MF   2  2 1.874149e-05
## GO:0045236    CXCR chemokine receptor binding  MF  18  3 6.212164e-05
## GO:0001664 G protein-coupled receptor binding  MF 273  7 1.912088e-04
## GO:0005518                   collagen binding  MF  68  4 2.213311e-04
##               P.DE.adj
## GO:0008009 0.008402397
```

```
## GO:0042379 0.020322841
## GO:0004962 0.031466953
## GO:0045236 0.078226670
## GO:0001664 0.185807459
## GO:0005518 0.185807459
```

```
# Perform Gene Ontology: Molecular Function Analysis using WebGestaltR on the downregulated genes
goana_downreg <- limma::goana(de = downreg_genes, universe = ref_genes, species = "Hs")

# Print or further process the results
goana_downreg <- goana_downreg[goana_downreg$Ont == "MF", ]
goana_downreg <- goana_downreg[sort(goana_downreg$P.DE, index.return=T)$ix, ]
goana_downreg$P.DE.adj <- p.adjust(goana_downreg$P.DE, n=nrow(goana_downreg), method="BH")
head(goana_downreg)
```

```
##                                                Term Ont    N DE         P.DE
## GO:0005102                  signaling receptor binding  MF 1434 29 8.986291e-09
## GO:0048018                    receptor ligand activity  MF  477 16 3.892548e-08
## GO:0030546       signaling receptor activator activity  MF  482 16 4.494615e-08
## GO:0030545       signaling receptor regulator activity  MF  515 16 1.112389e-07
## GO:0008083                      growth factor activity  MF  159  8 6.283248e-06
## GO:0005201 extracellular matrix structural constituent  MF  163  8 7.543403e-06
##                P.DE.adj
## GO:0005102 4.526395e-05
## GO:0048018 7.546459e-05
## GO:0030546 7.546459e-05
## GO:0030545 1.400776e-04
## GO:0008083 6.329744e-03
## GO:0005201 6.332687e-03
```

Writing out Data for comparison

Results of the analysis of Dataset 1 are saved for comparison between dataset analysis. The results of limma analysis are saved in txt file for further use.

```
write.table(LIMMAout_annot, sep= "\t", file="Dataset1_limma_results.txt")
```

The results of gene set analysis are saved in a txt file for further use.

```
write.table(kegg_upreg, sep= "\t", file="Dataset1_PathwayAnalysis_upreg_results.txt")
write.table(kegg_downreg, sep= "\t", file="Dataset1_PathwayAnalysis_downreg_results.txt")
write.table(goana_upreg, sep= "\t", file="Dataset1_MolecularFunction_upreg_results.txt")
write.table(goana_downreg, sep= "\t", file="Dataset1_MolecularFunction_downreg_results.txt")
```