Analysis of Dataset 4 - Methylation profiling by genome tiling array(GSE80468)

Data Exploration and Preprocessing

```
infinium_data = infinium_data[rowSums(is.na(exprs(infinium_data)))==0,]
#Take subset
infdata =infinium_data[,c(grep("healthy",pdata[,34]),grep("PCOS",pdata[,34]))]
annot_data_inf = pdata[c(grep("healthy",pdata[,34]),grep("PCOS",pdata[,34])),]
sampleNames(infdata) = paste(pdata[,1],sep="_")
```

```
infdata.pf = pfilter(infdata)
```

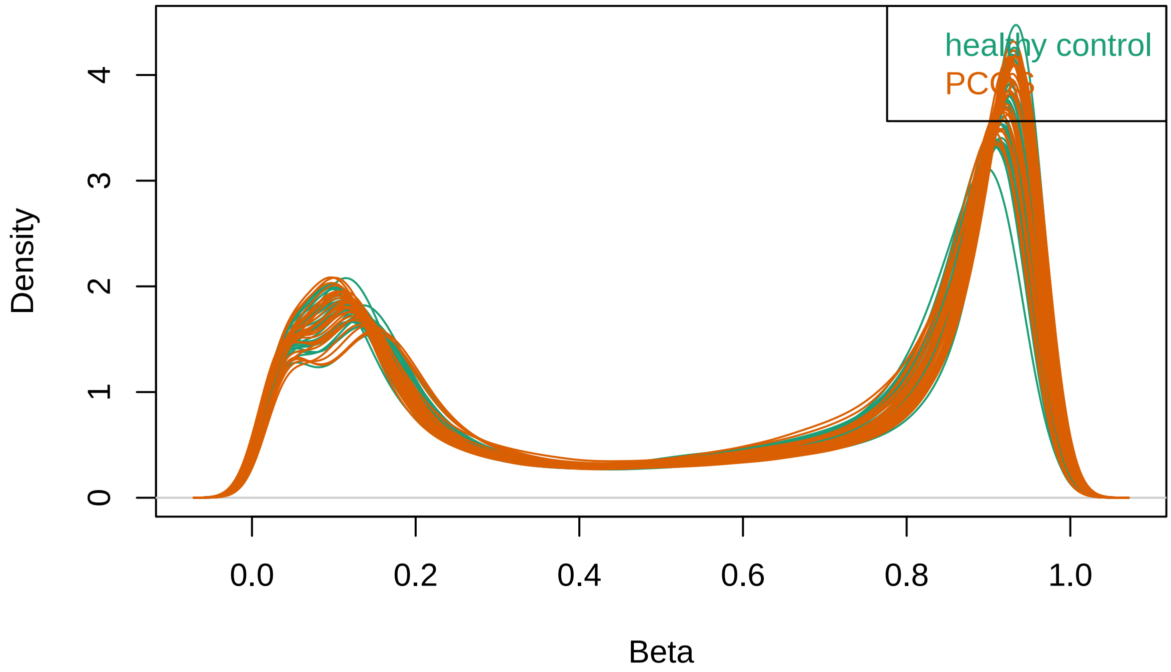
```
## 0 samples having 1 % of sites with a detection p-value greater than 0.05 were removed
## Samples removed:
## 260 sites were removed as beadcount <3 in 5 % of samples
## 0 sites having 1 % of samples with a detection p-value greater than 0.05 were removed
```

```
meth_mean = colMeans(betas(infdata))
meth_mean_healthy = meth_mean[0:30]
meth_mean_pcos =meth_mean[31:60]
```

```
t_test_res = t.test(meth_mean_healthy,meth_mean_pcos,var.equal = F)
t_test_res
```

```
##
##  Welch Two Sample t-test
##
## data:  meth_mean_healthy and meth_mean_pcos
## t = -1.677, df = 53.772, p-value = 0.09934
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.0084189891  0.0007500976
## sample estimates:
## mean of x mean of y
## 0.5638146 0.5676490
```

```
phenoData <- pData(geo_data)
densityPlot(betas(infdata), sampGroups = pdata$`diagnosis:ch1`)
```



Normalization

```
infdata.db = nanes(infdata.pf)
infdata.norm = dasen(infdata.db) # used for color balance normalization of Illumina methylation data. D
```

```
infdataM = as(infdata,"MethyLumiM")
infdataP = as(infdata.pf,"MethyLumiM")
infdataD = as(infdata.db,"MethyLumiM")
infdataN = as(infdata.norm,"MethyLumiM")
```

```
des = factor(pdata$`diagnosis:ch1`)
design = model.matrix(~0 + des)
colnames(design) = c("Healthy","PCOS")
cont.matrix = makeContrasts(Healthy - PCOS, levels = design)
```

```
fit = lmFit(infdataN,design)
fit2 = contrasts.fit(fit,cont.matrix)
fit2 = eBayes(fit2)
dim(fit2)
```

```
## [1] 484175      1
```

```r
limma_inf = topTable(fit2,adjust.method = "BH",number = nrow(exprs(infdataM)))
head(limma_inf)
```

```
##              Probe_ID DESIGN COLOR_CHANNEL      logFC   AveExpr         t
## cg23647968 cg23647968      I           Grn -0.3007335 2.0588751 -7.005270
## cg04737885 cg04737885     II          Both -0.7155935 1.4505993 -6.532893
## cg03197935 cg03197935     II          Both -0.5485005 2.0684252 -6.496843
## cg09456760 cg09456760      I           Grn -0.2145767 1.1294714 -6.261053
## cg03349251 cg03349251     II          Both -0.4278613 0.5681522 -5.666390
## cg08092966 cg08092966     II          Both  0.3068525 1.6792284  5.655763
##                 P.Value   adj.P.Val        B
## cg23647968 2.106791e-09 0.001020056 9.493648
## cg04737885 1.374686e-08 0.002558333 8.035692
## cg03197935 1.585170e-08 0.002558333 7.924349
## cg09456760 4.012525e-08 0.004856911 7.196555
## cg03349251 4.046752e-07 0.030915780 5.372000
## cg08092966 4.215249e-07 0.030915780 5.339638
```

```r
sum(limma_inf$adj.P.Val < 0.1)
```

```
## [1] 146
```

```r
dim(limma_inf)
```

```
## [1] 484175      9
```

```r
exprs(infdataN)[rownames(infdataN) %in% rownames(head(limma_inf)),][,1:4]
```

```
##            14286A-N1  14286A-N2 14286A-N3 14286A-N4
## cg03197935 1.3499246 1.51938408 1.6312616 1.6840806
## cg03349251 0.4643780 0.05064704 0.2467608 0.1482951
## cg04737885 0.5209632 1.15658581 0.7642958 1.2491249
## cg08092966 1.9121186 1.79115679 1.9128692 1.5802166
## cg09456760 0.8690290 0.96461514 1.0131581 1.0367054
## cg23647968 1.7416374 1.83427246 1.8736134 1.8681575
```

```r
dim(exprs(infdataN)[rownames(infdataN) %in% rownames(head(limma_inf)),])
```

```
## [1]  6 60
```

```r
head(betas(infdataN)[rownames(infdataN)%in%rownames(head(limma_inf))])
```

```
## [1] 0.7182295 0.5797829 0.5893077 0.7900753 0.6461972 0.7698051
```

```r
length(betas(infdataN)[rownames(infdataN)%in%rownames(head(limma_inf))])
```

```
## [1] 360
```

```r
data("IlluminaHumanMethylation450kanno.ilmn12.hg19")
annot_MA_inf = getAnnotation(IlluminaHumanMethylation450kanno.ilmn12.hg19)
annot_MA_inf= annot_MA_inf[sort(rownames(annot_MA_inf),index.return = T)$ix,]
dim(annot_MA_inf)
```

```
## [1] 485512     33
```

```r
# Find common Probe IDs and merge the data frames
merged_data <- merge(limma_inf, annot_MA_inf, by.x = "Probe_ID", by.y = "Name", all.x = TRUE)

# Create a new column 'Genes' in limma_inf, initialized with NAs
limma_inf$Genes <- NA

# Assign corresponding gene names to limma_inf for matched Probe IDs
limma_inf$Genes <- merged_data$UCSC_RefGene_Name

# Check the number of matching Probe IDs
cat("Number of matching Probe IDs:", sum(!is.na(limma_inf$Genes)), "\n")
```

```
## Number of matching Probe IDs: 484110
```

```r
cat("Number of Probe IDs in the data:", length(limma_inf$Genes), "\n")
```

```
## Number of Probe IDs in the data: 484175
```

```r
limma_out_sorted_inf = limma_inf[sort(limma_inf$Probe_ID,index.return = T)$ix,]
annot_MA_inf = annot_MA_inf[rownames(annot_MA_inf)%in%limma_inf$Probe_ID,]
limma_out_sorted_inf$CHR <- NA
limma_out_sorted_inf$POS <- NA

limma_out_sorted_inf$CHR = merged_data$chr
limma_out_sorted_inf$POS = merged_data$pos
dim(limma_out_sorted_inf)
```

```
## [1] 484175     12
```

```r
dim(annot_MA_inf)
```

```
## [1] 484110     33
```

```r
head(limma_out_sorted_inf)
```

```
##              Probe_ID DESIGN COLOR_CHANNEL       logFC    AveExpr          t
## cg00000029 cg00000029     II          Both  0.04079244 -0.5681203  0.6308328
## cg00000108 cg00000108     II          Both  0.04314586  3.3165936  0.6280095
## cg00000109 cg00000109     II          Both  0.13684003  2.0440933  1.8071540
## cg00000165 cg00000165     II          Both -0.02100752 -2.2938981 -0.4328383
## cg00000236 cg00000236     II          Both -0.05756591  0.8611430 -0.7591227
## cg00000289 cg00000289     II          Both  0.09989620  0.4772885  1.6745241
##               P.Value adj.P.Val         B                       Genes   CHR
## cg00000029 0.53046880 0.9418767 -5.407856 SCAND1;SCAND1;SCAND1;SCAND1 chr16
## cg00000108 0.53230450 0.9421796 -5.409399                              chr3
## cg00000109 0.07559121 0.7871182 -4.194278                   TFEB;TFEB  chr3
## cg00000165 0.66663448 0.9620385 -5.499474                       SPNS2  chr1
## cg00000236 0.45065611 0.9278180 -5.330556                       SPON1  chr8
## cg00000289 0.09906598 0.8127183 -4.385519               EIF2C2;EIF2C2 chr14
##                  POS
## cg00000029  53468112
## cg00000108  37459206
## cg00000109 171916037
## cg00000165  91194674
## cg00000236  42263294
## cg00000289  69341139
```

```r
limma_out_sorted_inf$Genes = gsub(";.*","",limma_out_sorted_inf$Genes)
```

```r
selected_columns_inf = limma_out_sorted_inf %>% select(adj.P.Val,logFC,Genes)
```

```r
cat("Dimensions of limma_out_sorted:", dim(limma_out_sorted_inf), "\n")
```

```
## Dimensions of limma_out_sorted: 484175 12
```

```r
cat("Number of unique Probe_IDs in limma_out_sorted:", length(unique(limma_out_sorted_inf$Probe_ID)), "
```

```
## Number of unique Probe_IDs in limma_out_sorted: 484175
```

```r
# Extract row names from betas(infdata)
beta_row_names <- rownames(betas(infdata))
# Identify row names not present in limma_out_sorted$Probe_ID
beta_row_names <- beta_row_names[( beta_row_names%in% limma_out_sorted_inf$Probe_ID)]
limma_out_sorted_inf = limma_out_sorted_inf[beta_row_names,]
```

```r
limma_out_sorted_inf$PCOS = rowMeans(betas(infdata)[rownames(infdata)%in%limma_out_sorted_inf$Probe_ID,
```

```r
limma_out_sorted_inf$PCOS_meth = rowMeans(betas(infdata)[rownames(infdata)%in%limma_out_sorted_inf$Probe
```

Control data

```r
limma_out_sorted_inf$Control_meth = rowMeans(betas(infdata)[rownames(infdata)%in%limma_out_sorted_inf$P
```

```r
limma_out_sorted_inf$abs_diff_meth = abs(limma_out_sorted_inf$PCOS_meth-limma_out_sorted_inf$Control_me
limma_out_annot_inf = limma_out_sorted_inf[sort(limma_out_sorted_inf$P.Value,index.return = T)$ix,]
```

```r
sign_re = limma_out_annot_inf[which(limma_out_annot_inf$adj.P.Val < 0.1),]

topgenes_prom = unique(sign_re)

sign_genes = sign_re |> select(Genes,CHR,POS,abs_diff_meth,P.Value,logFC,adj.P.Val)
head(sign_genes)
```

```
##               Genes   CHR      POS abs_diff_meth      P.Value      logFC
## cg23647968     RBL2 chr19 15051936    0.04298198 2.106791e-09 -0.3007335
## cg04737885  C3orf35 chr16  4014095    0.05324310 1.374686e-08 -0.7155935
## cg03197935   FNDC3B chr11 77885410    0.02820257 1.585170e-08 -0.5485005
## cg09456760          chr22 51206645    0.04128453 4.012525e-08 -0.2145767
## cg03349251    VDAC3  chr6 10832472    0.04300896 4.046752e-07 -0.4278613
## cg08092966    ACTN1  chr8 19009591    0.00862358 4.215249e-07  0.3068525
##               adj.P.Val
## cg23647968 0.001020056
## cg04737885 0.002558333
## cg03197935 0.002558333
## cg09456760 0.004856911
## cg03349251 0.030915780
## cg08092966 0.030915780
```

```
print(paste("Amount of significant genes:",
as.character(sum(nrow(sign_genes)))))
```

```
## [1] "Amount of significant genes: 146"
```

Gene Set Analysis (GSA)

```
library("org.Hs.eg.db")
limma_out_filtered_inf = limma_out_annot_inf[limma_out_annot_inf$adj.P.Val < 0.1,]

#For all genes

limma_out_annot_inf$Entrez_id = mapIds(org.Hs.eg.db,gsub("///.*","",limma_out_annot_inf$Genes),"ENTREZI
```

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

```
EntrezIDs_inf = limma_out_annot_inf$Entrez_id
```

```
ids_unique_inf = unlist(EntrezIDs_inf[!(duplicated(EntrezIDs_inf) | is.na(EntrezIDs_inf))])
```

```
goanaOut_inf = goana(de = ids_unique_inf, species = "Hs",trend =T)
head(goanaOut_inf)
```

```
##                                                                    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 15356 7.286299e-103
## GO:0000003  1506  1287  5.193012e-08
## GO:0001553    12    12  7.233538e-02
## GO:0001867    11    11  9.003950e-02
## GO:0001868     2     2  6.455697e-01
## GO:0001869     2     2  6.455697e-01
```

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

```
##                           Term Ont     N    DE          P.DE        FDR.DE
## GO:0003674 molecular_function  MF 18369 15514 4.916715e-310 1.127599e-305
## GO:0005488            binding  MF 16581 14115 7.321766e-234 8.395869e-230
## GO:0005515    protein binding  MF 13998 12109 3.961998e-216 3.028815e-212
## GO:0043167        ion binding  MF  6024  5435 6.855674e-129 2.246115e-125
## GO:0043169     cation binding  MF  4346  3907  6.436998e-80  4.613316e-77
## GO:0046872  metal ion binding  MF  4260  3832  5.606445e-79  3.896309e-76
```

```
goanaOut_inf <- goanaOut_inf[goanaOut_inf$Ont == "MF",]
print(paste("Amount of significant GO MF terms:",
as.character(sum(topGO_inf$FDR.DE < 0.1))))
```

```
## [1] "Amount of significant GO MF terms: 100"
```

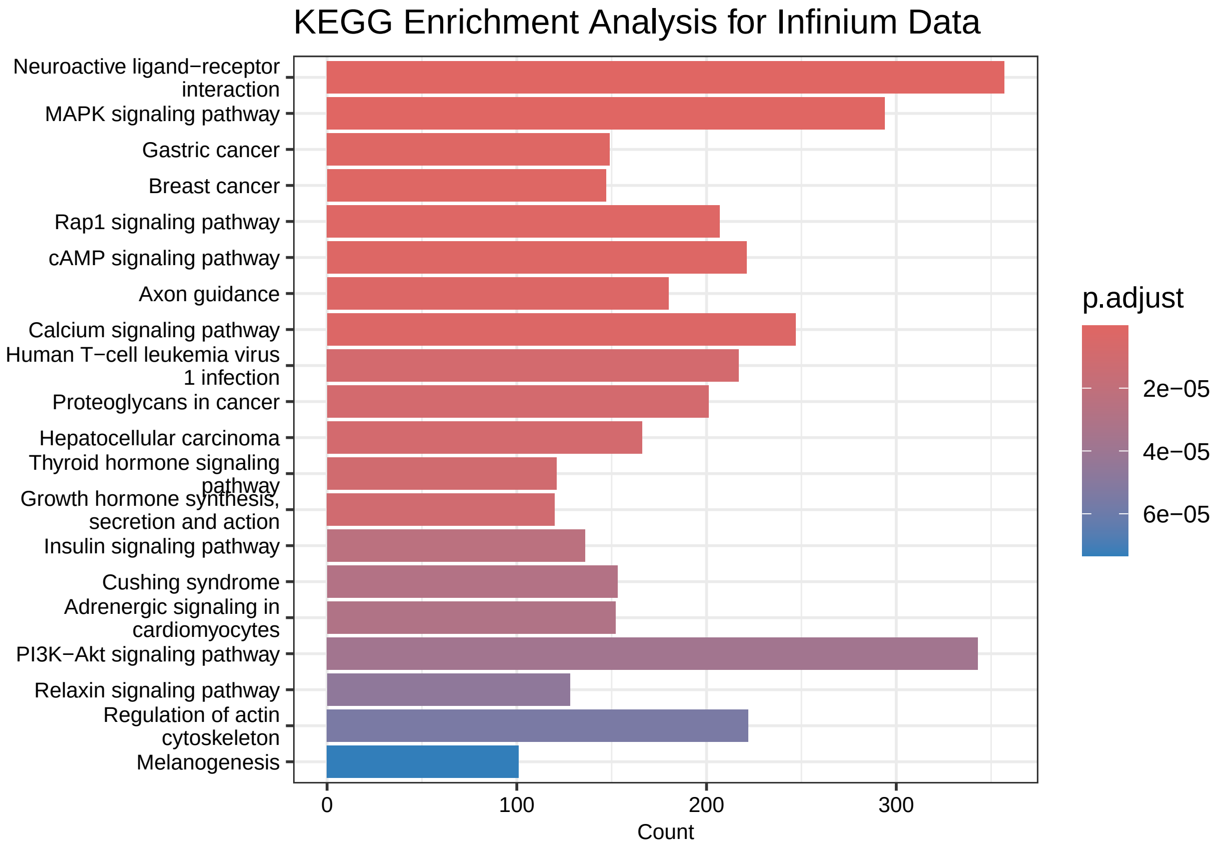
```
#if (!requireNamespace("BiocManager", quietly = TRUE))
 #   install.packages("BiocManager")
#BiocManager::install("clusterProfiler")


# Perform KEGG enrichment analysis with a p-value cutoff of 0.1
kegg_enrichment_inf <- enrichKEGG(gene = ids_unique_inf, organism = "hsa", keyType = "kegg", pvalueCuto
```

```
## 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_inf, showCategory = 20, title = "KEGG Enrichment Analysis for Infinium Data", xl
```


KEGG Enrichment Analysis for Infinium Data
Neuroactive ligand−receptor interaction
MAPK signaling pathway
Gastric cancer
Breast cancer
Rap1 signaling pathway
cAMP signaling pathway
Axon guidance
Calcium signaling pathway
Human T−cell leukemia virus 1 infection
Proteoglycans in cancer
Hepatocellular carcinoma
Thyroid hormone signaling pathway
Growth hormone synthesis, secretion and action
Insulin signaling pathway
Cushing syndrome
Adrenergic signaling in cardiomyocytes
PI3K−Akt signaling pathway
Relaxin signaling pathway
Regulation of actin cytoskeleton
Melanogenesis
0
100
200
300
Count
p.adjust
2e−05
4e−05
6e−05