Downstream analysis

Magali Eisik

Contents


```r
#Load require libraries
library(dplyr)      # Data manipulation
library(tidyr)      # Data cleaning / tidying
library(ggplot2)    # Data visualization
library(kableExtra) # Table formatting
library(knitr)      # Reporting tools
library(IlluminaHumanMethylation450kanno.ilmn12.hg19) # CpG annotation
library(clusterProfiler) # Pathway enrichment
library(org.Hs.eg.db)    # Gene IDs
library(enrichplot)      # Enrichment plots
library(readr)      # Read data
library(patchwork)  # Combine plots
library(kableExtra) # Table formatting
```

```r
#Read csv file with top20cpgs
top_cpgs <- read.csv("Top20_CpGs_ElasticNet.csv",
                     stringsAsFactors = FALSE)
#See first 3 rows
head(top_cpgs,3)
```

```
##          CpG     Weight Methylation_Change
## 1 cg00860090 -1.2912952     Hypomethylated
## 2 cg00653615 -0.9615959     Hypomethylated
## 3 cg02901139 -0.8243922     Hypomethylated
```

0.1 Separate Cpgs by methylation status

```
hyper_cpgs <- top_cpgs$CpG[top_cpgs$Methylation_Change == "Hypermethylated"]
hypo_cpgs <- top_cpgs$CpG[top_cpgs$Methylation_Change == "Hypomethylated"]

df_cpgs_met <- data.frame(
  CpG = c(hyper_cpgs, hypo_cpgs),
  Status = c(
    rep("Hypermethylated", length(hyper_cpgs)),
    rep("Hypomethylated", length(hypo_cpgs))
  )
)

df_cpgs_met %>%
  knitr::kable(
    format = "latex",
    escape = FALSE,
    booktabs = TRUE,
    caption = "CpG Sites by Methylation Status"
  ) %>%
  kableExtra::kable_styling(
    latex_options = c("striped", "hold_position")
  )
```

Table 1: CpG Sites by Methylation Status

CpG	Status
cg01832549	Hypermethylated
cg01901101	Hypermethylated
cg00116092	Hypermethylated
cg01557798	Hypermethylated
cg00019759	Hypermethylated
cg02177231	Hypermethylated
cg01352586	Hypermethylated
cg02431597	Hypermethylated
cg00860090	Hypomethylated
cg00653615	Hypomethylated
cg02901139	Hypomethylated
cg01561629	Hypomethylated
cg01974375	Hypomethylated
cg02294302	Hypomethylated
cg01850505	Hypomethylated
cg00251125	Hypomethylated
cg01785514	Hypomethylated
cg00891995	Hypomethylated
cg01459453	Hypomethylated
cg00039326	Hypomethylated

0.2 CpG Annotation

```
# The Illumina 450k annotation is loaded and converted
#into a data frame to map each CpG site to its associated genomic information
ann450k <- getAnnotation(IlluminaHumanMethylation450kanno.ilmn12.hg19)
ann450k_df <- as.data.frame(ann450k)
```

0.3 Mapping CpG Sites to Their Associated Genes

```
# Merge top CpGs with gene names
annotated_cpgs <- top_cpgs %>%
  left_join(ann450k_df[, c("Name", "UCSC_RefGene_Name")],
            by = c("CpG" = "Name"))
annotated_cpgs
```

```
##           CpG     Weight Methylation_Change                  UCSC_RefGene_Name
## 1  cg00860090 -1.2912952     Hypomethylated                              ITPKB
## 2  cg00653615 -0.9615959     Hypomethylated                  CDC42SE1;CDC42SE1
## 3  cg02901139 -0.8243922     Hypomethylated                          NENF;NENF
## 4  cg01561629 -0.8229217     Hypomethylated       OSBPL9;OSBPL9;OSBPL9;OSBPL9
## 5  cg01832549  0.7664121    Hypermethylated                             CAPZB
## 6  cg01901101  0.6819186    Hypermethylated
## 7  cg00116092  0.6810944    Hypermethylated
## 8  cg01974375 -0.6790483     Hypomethylated                             PI4KB
## 9  cg01557798  0.6621631    Hypermethylated                      OBSCN;OBSCN
## 10 cg02294302 -0.6176032     Hypomethylated                      FOXD2;FOXD2
## 11 cg00019759  0.5615314    Hypermethylated
## 12 cg01850505 -0.5614891     Hypomethylated                              RGL1
## 13 cg00251125 -0.5540297     Hypomethylated                      OBSCN;OBSCN
## 14 cg01785514 -0.5422039     Hypomethylated TATDN3;TATDN3;TATDN3;TATDN3;TATDN3
## 15 cg02177231  0.5182653    Hypermethylated                            TBX15
## 16 cg01352586  0.5165745    Hypermethylated
## 17 cg00891995 -0.5147423     Hypomethylated                           SPRR2C
## 18 cg02431597  0.5020730    Hypermethylated                          CACNA1E
## 19 cg01459453 -0.4969226     Hypomethylated                             SELP
## 20 cg00039326 -0.4861703     Hypomethylated                          TRAPPC3
```

```
#UCSC_RefGene_Name contains the gene(s) associated with each CpG site
#according to the UCSC RefGene annotation


# Split multiple genes per CpG
df_annotation <- annotated_cpgs %>%
  tidyr::separate_rows(UCSC_RefGene_Name, sep = ";") %>%
  dplyr::rename(Gene = UCSC_RefGene_Name) %>%
  dplyr::filter(Gene != "") %>%
  dplyr::select(Methylation_Change, Gene)
```

0.4 Gene Conversion for KEGG Analysis

```
#separates genes by methylation status and converts
#their symbols into ENTREZ IDs for downstream pathway enrichment analysis.

# Hypermethylated
hyper_genes <- unique(df_annotation$Gene[df_annotation$Methylation_Change == "Hypermethylated"])
hyper_ids <- bitr(hyper_genes, fromType = "SYMBOL", toType = "ENTREZID", OrgDb = org.Hs.eg.db)  #conver
# OrgDb = org.Hs.eg.db--> homo sapiens data base
# Hypomethylated
hypo_genes <- unique(df_annotation$Gene[df_annotation$Methylation_Change == "Hypomethylated"])
hypo_ids <- bitr(hypo_genes, fromType = "SYMBOL",
                 toType = "ENTREZID", OrgDb = org.Hs.eg.db)
```

0.5 KEGG Enrichment Analysis by Methylation Status

KEGG pathway enrichment analysis was performed on the ENTREZ IDs of genes associated with hyper- and hypomethylated CpGs, identifying significantly overrepresented biological pathways with adjusted p-values, gene counts, and lists of contributing genes for each pathway

```
# KEGG for hypermethylated
hyper_kegg <- enrichKEGG(gene = hyper_ids$ENTREZID, organism = 'hsa', pvalueCutoff = 0.05)
```

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

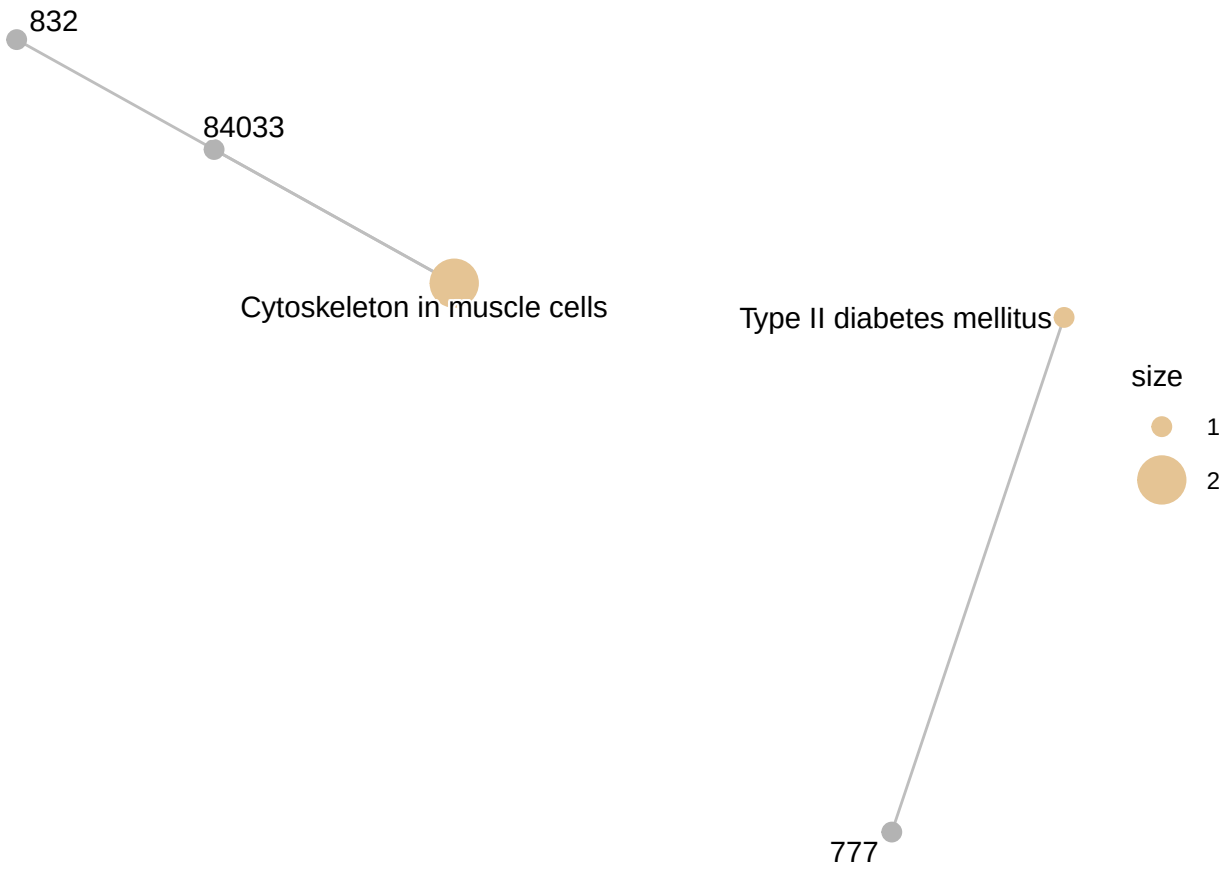
```
## Reading KEGG annotation online: "https://rest.kegg.jp/list/pathway/hsa"...
```

```
# KEGG for hypomethylated
hypo_kegg <- enrichKEGG(gene = hypo_ids$ENTREZID, organism = 'hsa', pvalueCutoff = 0.05)

#organism = 'hsa'--> homo sapiens
#pvalueCutoff = 0.05 →  only significantly enriched pathways whose p-value< 0.05 are considered.
```

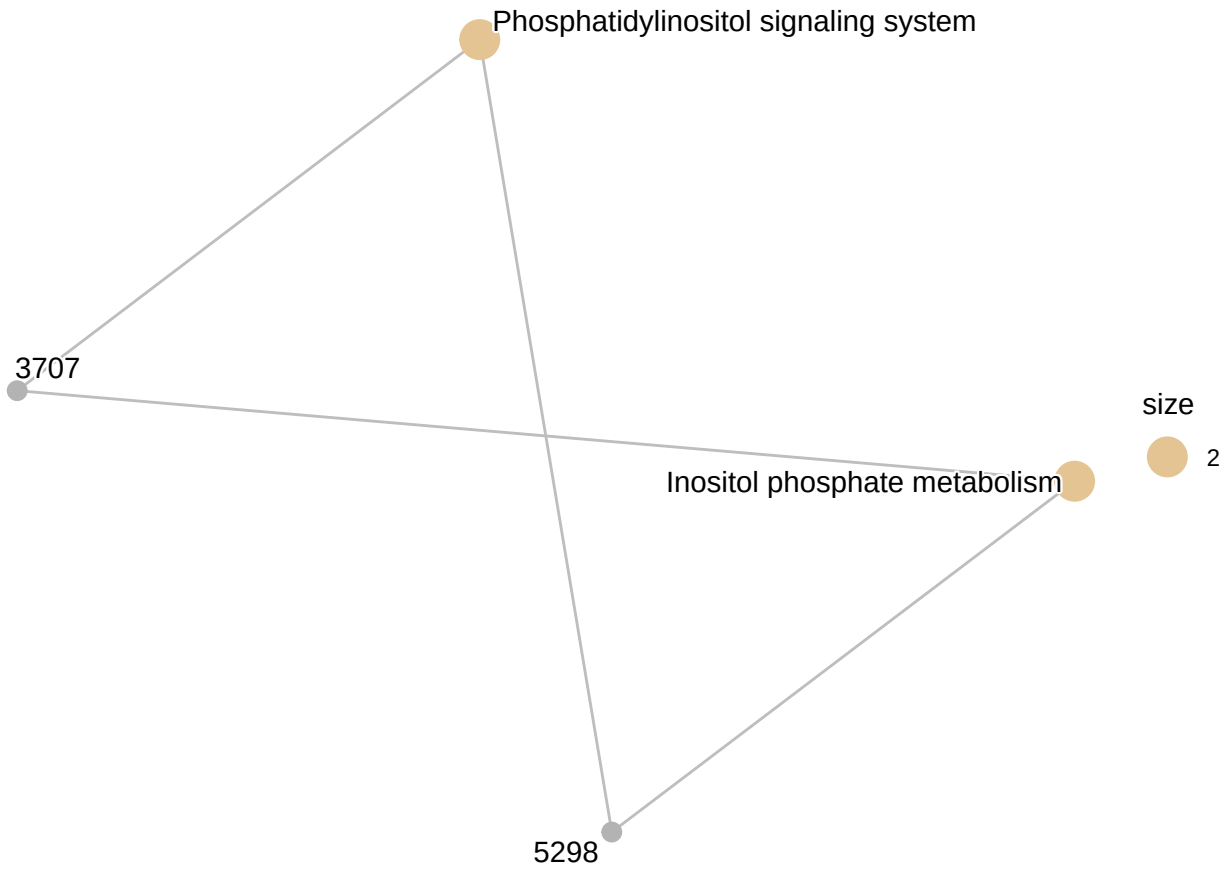
0.6 Data visualization

#Cnetplots (Category-Net Plot)

```
cnetplot(hyper_kegg, showCategory = min(5, nrow(hyper_kegg)))
```




```
cnetplot(hypo_kegg, showCategory = min(5, nrow(hypo_kegg)))
```



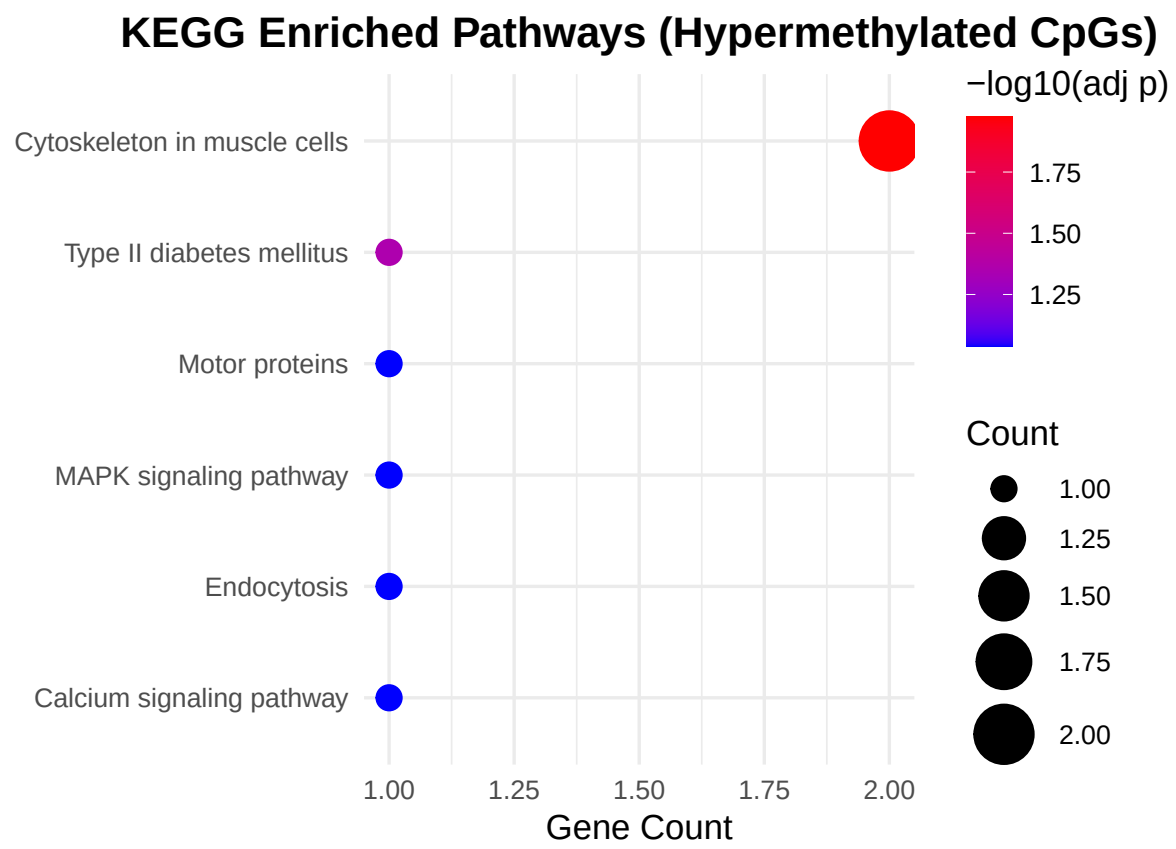
The cnetplots reveal the relationship between genes and pathways, which are significantly enriched for hypermethylated and hypomethylated gene sets. According to the first cnetplot (hypermethylated genes), the most represented pathways include the cytoskeleton in muscle cells and type II diabetes mellitus. In the second plot, hypomethylated genes are concentrated in signaling and metabolic pathways, such as the phosphatidylinositol signaling system and inositol phosphate metabolism

0.7 Dotplots

```
prepare_top_kegg <- function(enrich_obj, top_n = 10) {
  enrich_obj@result %>%
    mutate(
      ID = as.character(ID),
      Description = as.character(Description),
      Count = as.numeric(Count),
      p.adjust = as.numeric(p.adjust)
    ) %>%
    arrange(p.adjust) %>%
    slice(1:top_n)
}


# Prepare top pathways

top_hyper <- prepare_top_kegg(hyper_kegg, top_n = 10)
```

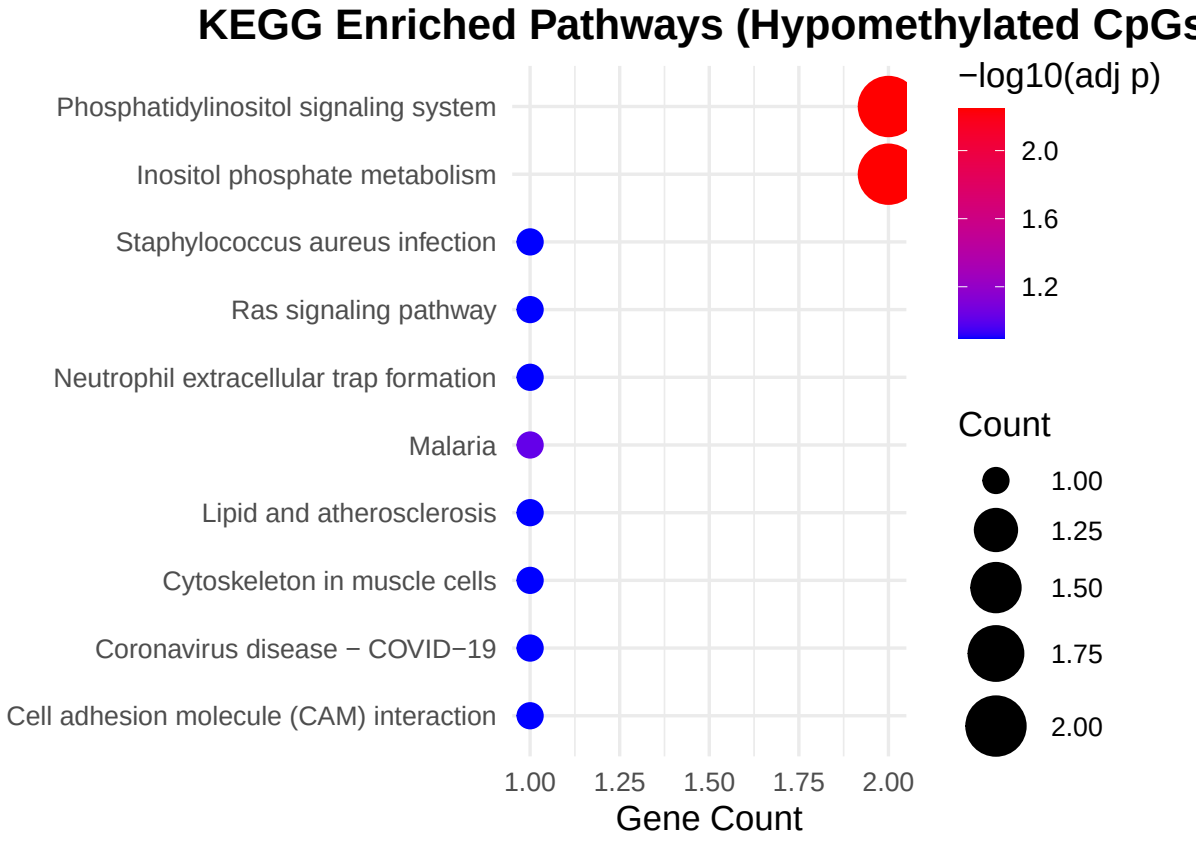
```r
top_hypo  <- prepare_top_kegg(hypo_kegg, top_n = 10)


# Create dotplot for hypermethylated CpGs

p_hyper <- ggplot(top_hyper, aes(x = reorder(Description, Count), y = Count)) +
  geom_point(aes(size = Count, color = -log10(p.adjust))) +
  coord_flip() +
  scale_color_gradient(low = "blue", high = "red") +
  scale_size_continuous(range = c(4, 10)) +  # Larger points
  labs(
    title = "KEGG Enriched Pathways (Hypermethylated CpGs)",
    x = "",
    y = "Gene Count",
    color = "-log10(adj p)"
  ) +
  theme_minimal(base_size = 13) +
  theme(
    plot.title = element_text(face = "bold", hjust = 0.5),
    axis.text.y = element_text(size = 10)
  )


# Create dotplot for hypomethylated CpGs

p_hypo <- ggplot(top_hypo, aes(x = reorder(Description, Count), y = Count)) +
  geom_point(aes(size = Count, color = -log10(p.adjust))) +
  coord_flip() +
  scale_color_gradient(low = "blue", high = "red") +
  scale_size_continuous(range = c(4, 10)) +
  labs(
    title = "KEGG Enriched Pathways (Hypomethylated CpGs)",
    x = "",
    y = "Gene Count",
    color = "-log10(adj p)"
  ) +
  theme_minimal(base_size = 13) +
  theme(
    plot.title = element_text(face = "bold", hjust = 0.5),
    axis.text.y = element_text(size = 10)
  )


ggsave("KEGG_Hypermethylated.png", plot = p_hyper, width = 10, height = 8, dpi = 300)
ggsave("KEGG_Hypomethylated.png", plot = p_hypo, width = 10, height = 8, dpi = 300)

ggplot(top_hyper, aes(x = reorder(Description, Count), y = Count)) +
  geom_point(aes(size = Count, color = -log10(p.adjust))) +
  coord_flip() +
  scale_color_gradient(low = "blue", high = "red") +
  scale_size_continuous(range = c(4, 10)) +  # Larger points
  labs(
    title = "KEGG Enriched Pathways (Hypermethylated CpGs)",
```

```
    x = "",
    y = "Gene Count",
    color = "-log10(adj p)"
  ) +
  theme_minimal(base_size = 13) +
  theme(
    plot.title = element_text(face = "bold", hjust = 0.5),
    axis.text.y = element_text(size = 10)
  )
```




```
ggplot(top_hypo, aes(x = reorder(Description, Count), y = Count)) +
  geom_point(aes(size = Count, color = -log10(p.adjust))) +
  coord_flip() +
  scale_color_gradient(low = "blue", high = "red") +
  scale_size_continuous(range = c(4, 10)) +
  labs(
    title = "KEGG Enriched Pathways (Hypomethylated CpGs)",
    x = "",
    y = "Gene Count",
    color = "-log10(adj p)"
  ) +
  theme_minimal(base_size = 13) +
  theme(
    plot.title = element_text(face = "bold", hjust = 0.5),
    axis.text.y = element_text(size = 10)
```

```
)
```



The dot plots illustrate pathway-specific methylation changes. Hypermethylated CpGs are mainly enriched in pathways such as Cytoskeleton in muscle cells and Type II diabetes mellitus , whereas hypomethylated CpGs are enriched in Inositol phosphate metabolism and Phosphatidylinositol signaling system. Dot size corresponds to the number of genes in each pathway, color represents statistical significance (FDR), and gene ratio indicates the proportion of input genes associated with each pathway. These findings indicate that methylation changes selectively impact biological pathways, potentially repressing some while activating others.

0.8 KEGG Pathway Enrichment and CpG Annotation by Methylation Status

```
# Table with hyperlinks to KEGG pathways
# Hyper-methylated
hyper_table <- hyper_kegg@result
hyper_table$Methylation_Status <- "Hypermethylated"

# Hypo-methylated
hypo_table <- hypo_kegg@result
hypo_table$Methylation_Status <- "Hypomethylated"

# Combine tables
kegg_results <- rbind(
  hyper_table[, c("Description", "ID", "pvalue", "p.adjust", "Methylation_Status")],
```

```
  hypo_table[, c("Description", "ID", "pvalue", "p.adjust", "Methylation_Status")]
)

# Rename columns
colnames(kegg_results) <- c("Pathway_Name", "KEGG_ID", "PValue", "Adjusted_PValue", "Methylation_Status

kegg_results <- kegg_results %>%
  mutate(KEGG_Link = paste0("https://www.kegg.jp/dbget-bin/www_bget?", KEGG_ID))

kegg_results <- kegg_results %>%
  mutate(Pathway_Name_LaTeX = paste0("\\href{", KEGG_Link, "}{", Pathway_Name, "}"))

kegg_results %>%
  dplyr::select(Pathway_Name_LaTeX, PValue, Adjusted_PValue, Methylation_Status) %>%
  knitr::kable(format = "latex", escape = FALSE, booktabs = TRUE,
               caption = "KEGG Pathways Enriched by Methylation Status",
               col.names = c("Pathway", "P-value", "Adjusted P-value", "Methylation Status"),
               align = "c") %>%
  kable_styling(latex_options = c("striped", "hold_position"), full_width = FALSE)
```

Table 2: KEGG Pathways Enriched by Methylation Status

	Pathway	P-value	Adjusted P-value	Methylation Status
hsa04820	Cytoskeleton in muscle cells	0.0017635	0.0105810	Hypermethylated
hsa04930	Type II diabetes mellitus	0.0147518	0.0442554	Hypermethylated
hsa04814	Motor proteins	0.0608604	0.0916739	Hypermethylated
hsa04144	Endocytosis	0.0773995	0.0916739	Hypermethylated
hsa04020	Calcium signaling pathway	0.0779972	0.0916739	Hypermethylated
hsa04010	MAPK signaling pathway	0.0916739	0.0916739	Hypermethylated
hsa00562	Inositol phosphate metabolism	0.0006533	0.0056633	Hypomethylated
hsa04070	Phosphatidylinositol signaling system	0.0010297	0.0056633	Hypomethylated
hsa05144	Malaria	0.0260131	0.0953815	Hypomethylated
hsa05150	Staphylococcus aureus infection	0.0524897	0.1265981	Hypomethylated
hsa04514	Cell adhesion molecule (CAM) interaction	0.0813386	0.1265981	Hypomethylated
hsa04613	Neutrophil extracellular trap formation	0.0988884	0.1265981	Hypomethylated
hsa05417	Lipid and atherosclerosis	0.1085217	0.1265981	Hypomethylated
hsa048201	Cytoskeleton in muscle cells	0.1166452	0.1265981	Hypomethylated
hsa04014	Ras signaling pathway	0.1190231	0.1265981	Hypomethylated
hsa05171	Coronavirus disease - COVID-19	0.1190231	0.1265981	Hypomethylated
hsa040201	Calcium signaling pathway	0.1265981	0.1265981	Hypomethylated

0.9 Save results of KEGG analysis in CSV files

```
# Hypermethylated
hyper_table <- hyper_kegg@result
```

```r
hyper_table$Methylation_Status <- "Hypermethylated"

# Hypomethylated
hypo_table <- hypo_kegg@result
hypo_table$Methylation_Status <- "Hypomethylated"

# Combine tables
kegg_results <- rbind(
  hyper_table[, c("Description", "ID", "pvalue", "p.adjust", "Methylation_Status")],
  hypo_table[, c("Description", "ID", "pvalue", "p.adjust", "Methylation_Status")]
)

# Rename columns
colnames(kegg_results) <- c(
  "Pathway_Name", "KEGG_ID", "PValue", "Adjusted_PValue", "Methylation_Status")

# Create KEGG link column
kegg_results$KEGG_Link <- paste0("https://www.kegg.jp/dbget-bin/www_bget?", kegg_results$KEGG_ID)
kegg_results <- kegg_results[order(kegg_results$Adjusted_PValue), ]

# Save to CSV
write.csv(kegg_results, "KEGG_Pathways_Methylation.csv", row.names = FALSE)


#Gene names

final_cpg_genes <- annotated_cpgs %>%
  tidyr::separate_rows(UCSC_RefGene_Name, sep = ";") %>%
  dplyr::rename(Gene = UCSC_RefGene_Name) %>%
  dplyr::filter(Gene != "") %>%
  dplyr::select(CpG, Gene, Methylation_Change)


final_cpg_genes <- distinct(final_cpg_genes)

write.csv(final_cpg_genes,
          "Top20_CpGs_Annotated.csv",
          row.names = FALSE)
#Save csv
```

0.10 Annotated CpGs with Associated Genes and Methylation Status

```r
#Display table with genes
cpgs_display <- final_cpg_genes %>% head(20)

cpgs_display %>%
  head(20) %>%
  knitr::kable(
    format = "latex",
    booktabs = TRUE,
    caption = "CpG Sites with Associated Genes and Methylation Status",
    col.names = c("CpG Site", "Associated Gene", "Methylation Status"),
```

```
    align = "c",
    escape = TRUE
  ) %>%
  kableExtra::kable_styling(
    latex_options = c("striped", "hold_position"),
    font_size = 11,
    full_width = FALSE
  )
```

Table 3: CpG Sites with Associated Genes and Methylation Status

CpG Site	Associated Gene	Methylation Status
cg00860090	ITPKB	Hypomethylated
cg00653615	CDC42SE1	Hypomethylated
cg02901139	NENF	Hypomethylated
cg01561629	OSBPL9	Hypomethylated
cg01832549	CAPZB	Hypermethylated
cg01974375	PI4KB	Hypomethylated
cg01557798	OBSCN	Hypermethylated
cg02294302	FOXD2	Hypomethylated
cg01850505	RGL1	Hypomethylated
cg00251125	OBSCN	Hypomethylated
cg01785514	TATDN3	Hypomethylated
cg02177231	TBX15	Hypermethylated
cg00891995	SPRR2C	Hypomethylated
cg02431597	CACNA1E	Hypermethylated
cg01459453	SELP	Hypomethylated
cg00039326	TRAPPC3	Hypomethylated

```
cpgs_display %>%
  knitr::kable(format = "latex", escape = FALSE, booktabs = TRUE,
               caption = "Annotated Top 20 CpGs with Associated Genes and Methylation Status",
               col.names = c("CpG Site", "Associated Gene", "Methylation Status"),
               align = "c") %>%
  kable_styling(latex_options = c("striped", "hold_position"), font_size = 10)
```

Table 4: Annotated Top 20 CpGs with Associated Genes and Methylation Status

CpG Site	Associated Gene	Methylation Status
cg00860090	ITPKB	Hypomethylated
cg00653615	CDC42SE1	Hypomethylated
cg02901139	NENF	Hypomethylated
cg01561629	OSBPL9	Hypomethylated
cg01832549	CAPZB	Hypermethylated
cg01974375	PI4KB	Hypomethylated
cg01557798	OBSCN	Hypermethylated
cg02294302	FOXD2	Hypomethylated
cg01850505	RGL1	Hypomethylated
cg00251125	OBSCN	Hypomethylated
cg01785514	TATDN3	Hypomethylated
cg02177231	TBX15	Hypermethylated
cg00891995	SPRR2C	Hypomethylated
cg02431597	CACNA1E	Hypermethylated
cg01459453	SELP	Hypomethylated
cg00039326	TRAPPC3	Hypomethylated