

Methods/Results/Discussion

[Illumina Infinium 850K data analysis](#)

[Method](#)

[Result](#)

[Interactome analysis of the related Differentially Methylated Genes](#)

[Method](#)

[Result](#)

[Enrichment analysis of the Interactome](#)

[Method](#)

[Result](#)

Methods/Results/Discussion

Illumina Infinium 850K data analysis

Method

Data generated in Illumina Infinium 850K arrays, were taken to the *minfi*, *missmethyl* and *limma* packages in RStudio (v 1.2.1) combined with R (v3.6.3) and other bioconductor packages. The filtration applied to remove unbound probes, remove X and Y chromosomes, multi-hit probes, and also stringent p-value ($p > 0.01$) detected probes, resulted in ~ 750K CpG sites. The differential methylation analysis between the patient and the control samples were calculated by using the Bonferroni-Hochberg (BH) corrected p -values (< 0.05). The CpG sites were annotated using hg38.13 database from UCSC.

```
if (!requireNamespace("BiocManager", quietly = TRUE))
  install.packages("BiocManager")
BiocManager::install("missMethyl")
BiocManager::install("minfiData")
library(missMethyl)
library(limma)
library(minfi)
library(minfiData)

baseDir <- "/Users/jyoda68/Documents/ML/COVID-
19/NewData_201025/Analysis_201026/"
targets <- read.metharray.sheet(baseDir)
rgSet <- read.metharray.exp(targets = targets)
mSet <- preprocessRaw(rgSet)
mSetSw <- SWAN(mSet, verbose=TRUE)
detP <- detectionP(rgSet)
keep <- rowSums(detP < 0.01) == ncol(rgSet)
mSetSw <- mSetSw[keep, ]
mset_reduced <- mSetSw[sample(1:nrow(mSetSw), ), ]
meth <- getMeth(mset_reduced)
unmeth <- getUnmeth(mset_reduced)
Mval <- log2((meth + 100) / (unmeth + 100))
beta <- getBeta(mset_reduced)
dim(Mval)
group <- factor(targets$Sample_Group, levels=c("CON", "COVID"))
design <- model.matrix(~group)
```

```

design
fit.reduced <- lmFit(Mval,design)
fit.reduced <- eBayes(fit.reduced)
summary(decideTests(fit.reduced))
top<-topTable(fit.reduced,coef=2)
top
colnames(beta) <- targets$Sample_Name
colnames(Mval) <- targets$Sample_Name
write.table(beta, "BetaValueData_COVID-19.txt", sep = "\t", quote = F)
write.table(Mval, "MethylationValueData_COVID-19.txt", sep = "\t", quote = F)

```

Result

The analysis resulted into 2 CpG sites and annotation revealed that these two are present in two different chromosomes, Chr 17 and Chr 22. The calculated β values for each participants were showed in the violin plot (**Figure 1**).

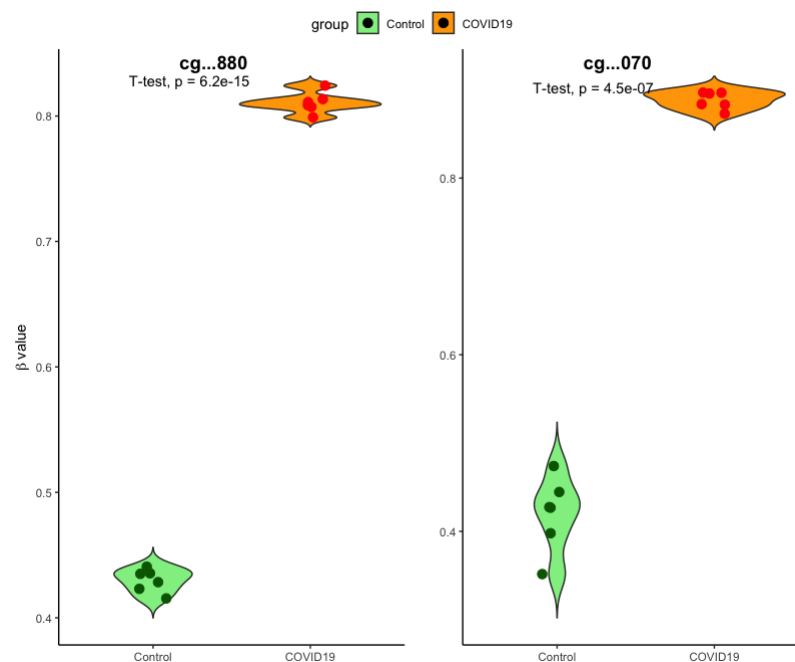


Figure 1: The violin plot illustrates the β values of the significantly methylated CpGs in each individual. The orange color violin represents the patients and the red dots are each individual while light green color violins are from the control population and dark green circles represent each individual. The Y and X axes show the β values and sample group, respectively.

```

# Violin plot of significant CpGs
dat880 <- read.table("cg880.txt", stringsAsFactors = F, header = T)
dat070 <- read.table("cg070.txt", stringsAsFactors = F, header = T)
con <- grepl("Control", dat880$group)
p <- ggplot(dat880, aes(x=group, y=beta, fill=group)) +
  geom_violin(trim=FALSE)+
  geom_jitter(position = position_jitter(0.1), color = ifelse(con, "darkgreen",
"red"), size = 3)+
  theme_classic()+
  scale_fill_manual(values = c("lightgreen", "orange"))+
  labs(y=beta~"value", x=NULL)+
  stat_compare_means(method = "t.test")
p1 <-
ggplot(dat070, aes(x=group, y=beta, fill=group)) +
  geom_violin(trim=FALSE)+

```

```

geom_jitter(position = position_jitter(0.1), color = ifelse(con, "darkgreen",
"red"), size = 3) +
theme_classic() +
scale_fill_manual(values = c("lightgreen", "orange")) +
labs(y=NULL, x=NULL) +
stat_compare_means(method = "t.test")

install.packages("ggpubr")
library(ggpubr)

ggarrange(p, p1,
          labels = c("cg...880", "cg...070"),
          ncol = 2, nrow = 1, align = "v", hjust = -2.5, vjust = 2,
          common.legend = TRUE)

```

The annotation result showed the first CpG site is located in the CpG island (shore region) in the database and the second CpG site is located at the transcription start site (TSS1500).

Interactome analysis of the related Differentially Methylated Genes

Method

The significant differentially methylated genes (DMGs) were taken into account to search in the BioGRID (v4.1) database which has almost 2 million interactions among ~75000 genes from different species. We extracted only human [species ID: 9606] interactions from the all downloaded dataset and compared them with the two DMGs from our analysis.

```

setwd("~/Desktop/Link to BackUp_All/SignatureData")
FNs <- read.table("PrIDfrom2CpGs.txt", stringsAsFactors = F, header = T)
biogrid <- read.table("BIOGRID_HUMAN_Interactors.txt", stringsAsFactors = F,
header = T)
colnames(biogrid)
biogrid <- read.table("BIOGRID_HUMAN_Interactors.txt", stringsAsFactors = F,
header = F)
colnames(biogrid)
head(biogrid)
colnames(biogrid) <- c("PrID", "InteractorB")
mergeData <- merge(FNs, biogrid, by = "PrID")
write.table(mergeData, "MatchedData.txt", sep = "\t", row.names = F, quote = F)

```

Next, we performed StringDB analysis on the first interacting neighbors of those 2 DMGs and used strong stringent filtration criteria (minimum required interaction score = 0.900) and generated the interactome comprises of 60 proteins. We used Cytoscape (v3.8) and StringDB API package to generate the interactome.

Result

We obtained a total of 76 interacting human proteins using these 2 DMGs using the BioGRID data. We extracted the interactors from the dataset and used in the Cytoscape (v3.8) to visualize the interacting network. We then performed the StringDB analysis on these interactome (see method) and obtained strong interaction result (**Figure 2**). In the analysis, we observed that Fibronectin (FN1) and Tran-Golgi Network Protein 2 (TGOLN2) are common interactors between the two

DMGs. Four genes, ITGA8, ITGA9, RASGRP2 and TRIM71 are showed to not annotated by StringDB dataset.

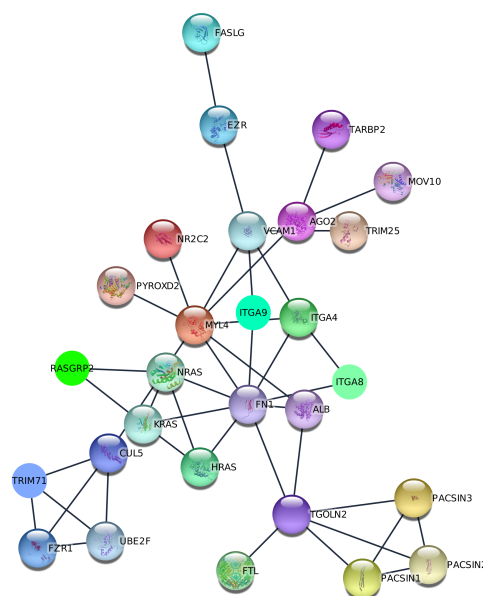


Figure 2: The figure illustrates the interactome of the two identified significant DMGs. The light glass colors applied from the default StringDB database and the edges show the physical interaction among the interactors. Four genes which have solid colored circles (green: RASGRP2, blue: TRIM71, sea green: ITGA9 and light sea green: ITGA8) are not annotated by StringDB network.

Enrichment analysis of the Interactome

Method

In the Cytoscape (v3.8), we used StringDB enrichment analysis API package to calculate the over-represented functional categories from Gene Ontology and pathway enrichment using KEGG and reactome databases. The default filter was changed and we applied stringent filtration on the calculation of the statistics, two-sided hyper-geometric test with Benjamini-Hochberg correction (p -value < 0.05), genes per pathway is 3 or more with 1000 permutation.

Result

The above interactome was placed as input file to perform the enrichment analysis of functional categories and pathways. We used string pathway analysis using gene ontology biological processes, reactome pathways and KEGG pathways. A total of 142, 99 and 78 enriched terms were found in GO Biological processes, Reactome pathways and KEGG pathways, respectively (**Figure 3** ; Supplementary File S1).

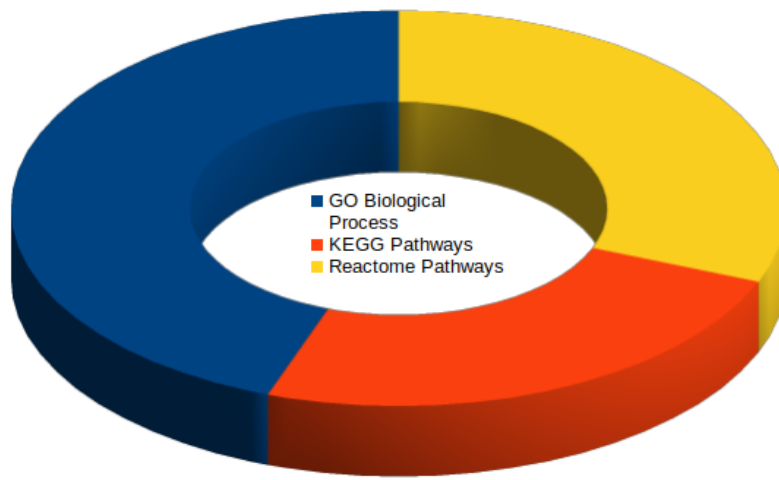


Figure 3: The donut chart portrays the number of enriched terms associated with the ontology and pathway analyses. The color scale were described in the figure legend.

Among these pathways, four were related to both DMGs, two from KEGG and two from Reactome databases (**Table-1**).

Category	Description	FDR	Gene ratio
KEGG	VEGF signaling pathway	0.0021	0.05
KEGG	Mitophagy - animal	0.0021	0.047
Reactome	EGFR Transactivation by Gastrin	2.05e-05	0.375
Reactome	Clathrin-mediated endocytosis	0.0006	0.29

Table -1: Four significantly overrepresented pathways derived from the enrichment analysis of StringDB contains both differentially methylated genes.

We calculated the gene ratio and the false discovery rates to locate the strongest pathway among these four and observed that Clathrin-mediated endocytosis from the reactome database has more genes from our list including those two DMGs (**Figure 4**).

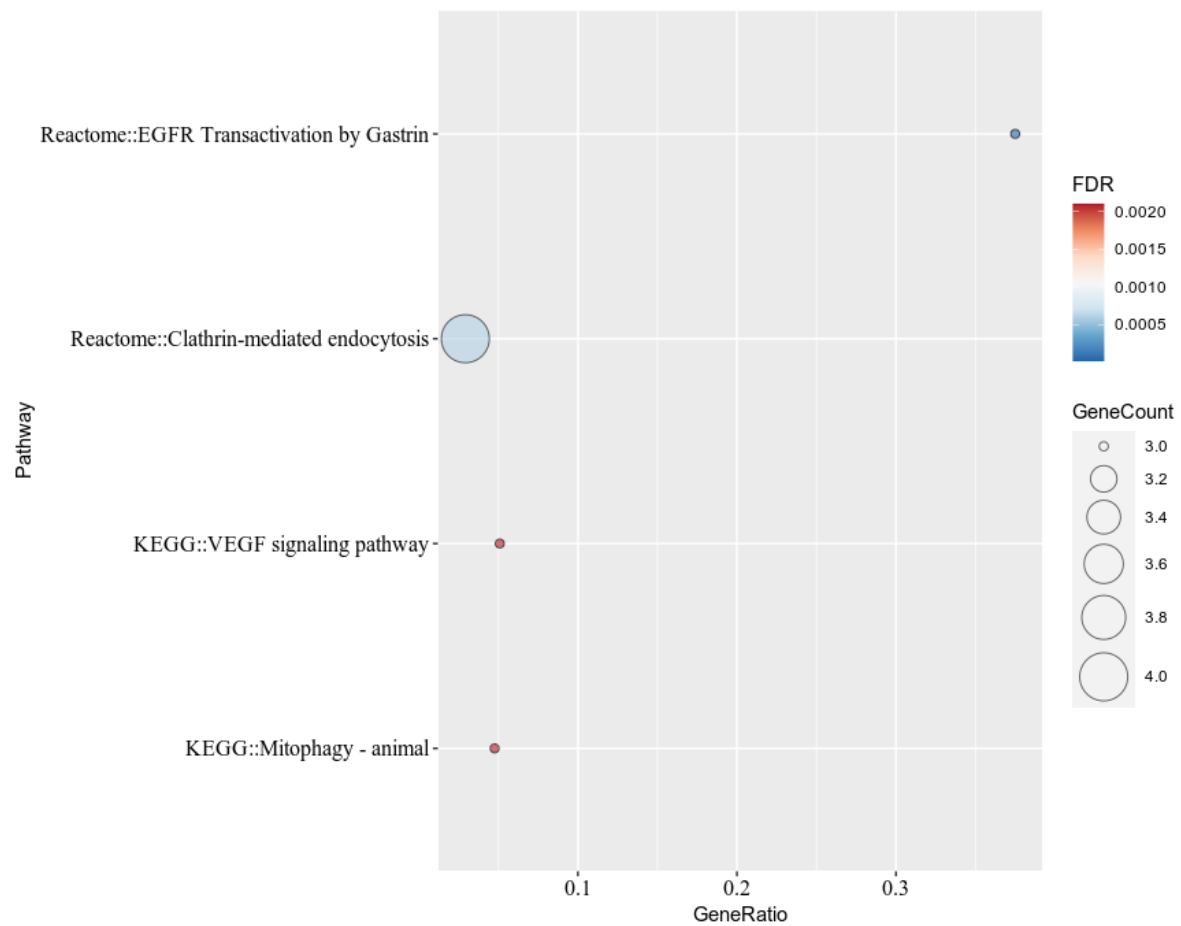


Figure 4: Top 4 pathways those contain both DMGs from our data. Gene count calculates the number of genes from our data, Gene ratio was calculated as the ratio of genes present in our data over the number of genes for the respective pathway in the databases. FDR represents the BH-corrected p -values. Circle size determines the gene count and color depicts the FDR value.