#Analisis Differential Gene Expression untuk Memprediksi Keparahan COVID-19 Pasien Mild dan Severe pada Dataset GSE164805 Menggunakan R

#1. Pendahuluan

#Pandemi COVID-19 merupakan wabah yang muncul pada tahun 2019 dan telah menyebabkan hampir lebih dari ratusan juta orang terinfeksi (Al-Jindeel et al., 2021). Penyakit ini disebabkan oleh virus *Severe Acute Respiratory Syndrome Coronavirus* 2 (SARS-CoV-2) yang menginfeksi saluran pernapasan (Hao et al., 2022). Keparahan COVID-19 dapat berkembang dalam bentuk yang lebih berat dan berakibat komplikasi serius apabila tidak ditangani dengan tepat. Perubahan COVID-19 dari gejala ringan menjadi gejala berat sangat sulit untuk diprediksi secara dini. Oleh karena itu sangat penting untuk menemukan metode deteksi dini pada penderita COVID-19 untuk mencegah peningkatan keparahan.

#Peningkatan teknologi pada analisis biomolekuler dan bioinformatika saat ini memberikan wawasan baru untuk pencegahan keparahan penyakit secara dini. Salah satu pendekatan yang digunakan saat ini untuk deteksi dini penyakit atau infeksi yaitu transkriptomik. Analisis transkriptomik banyak digunakan karena membantu menganalisis perubahan ekspresi gen yang dapat digunakan sebagai biomarker awal gejala infeksi ringan sebelum berkembang menjadi gejala berat. Adanya analisis ini dapat membantu identifikasi pasien yang berisiko tinggi. Oleh karena itu, pada analisis ini bertujuan untuk identifikasi *Differentially Expressed Genes* (DEGs) antara pasien mild dan severe menggunakan dataset GSE164805 dengan bantuan program R dan website ShinyGo.

#2. Metode

#2.1. Sumber Dataset

#Dataset yang digunakan yaitu GSE164805 diambil dari database Gene Expression Omnibus (GEO) (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE164805). Dataset ini dipilih karena berisi data mengenai profil ekspresi gen dari sampel *Peripheral Blood Mononuclear Cells* (PBMC) yang diambil dari pasien COVID-19 ringan dan berat. Sampel dataset ini diambil dari darah 10 pasien yang terbagi menjadi 2 kelompok yaitu 5 sampel pasien COVID-19 ringan (mild), 5 sampel pasien COVID-19 berat (severe) (Zhang et al., 2021).

#2.2. Persiapan Perangkat Lunak dan Pengolahan Data

#a. Terlebih dahulu mengunduh dan menginstal perangkat lunak R for Window (versi 4.5.2) di (https://cran.r-project.org/) dan Rstudio Dekstop (versi 2026.01.1+403) di (: https://posit.co/download/rstudio-desktop/).

#b. Setelah itu, menginstal manager paket Bioconductor yaitu BiocManager. Dengan menggunakan *script* tertera:

```
if (!require("BiocManager", quietly = TRUE))
install.packages("BiocManager")
BiocManager::install(version = "3.22")
```

```
BiocManager::install("hgu133a.db")
```
#c. BiocManager yang digunakan dalam analisis ini yaitu GEOquery dan limma. Dengan menggunakan *script* tertera:
```
BiocManager::install(c("GEOquery", "limma"), ask = FALSE, update = FALSE)
```
#d. Selanjutnya menginstal dplyr untuk memotong dan merapikan tabel data agar mudah diolah. Dengan menggunakan *script* tertera:
```
install.packages("dplyr")
```
#e. Setelah itu menginstal paket CRAN untuk visualisasi meliputi ggplot (Volcano Plot), pheatmap (Heatmap), gplots & RcolorBrewer (Diagram Venn dan palet warna), dan UMAP. Dengan menggunakan *script* tertera:
```
install.packages(c("ggplot2", "pheatmap", "gplots", "RColorBrewer"))
```
#f. Tahap terakhir cek instalasi. Dengan menggunakan *script* tertera:
```
library(GEOquery)
library(limma)
library(dplyr)
library(ggplot2)
library(pheatmap)
library(gplots)
library(RColorBrewer)
library(umap)
library(hgu133a.db)
library(AnnotationDbi)
```

#2.3. Pengambilan Data dari GEO

#Proses analisis DEG dengan dataset GSE164805 menggunakan *script* tertera:
```
library(GEOquery)
gset <- getGEO("GSE164805", GSEMatrix = TRUE, AnnotGPL = TRUE)[[1]]
expr_matrix <- exprs(gset)
sample_info <- pData(gset)
feature_info <- fData(gset)
```

#2.4. Pre-Processing Data Ekspresi

#Proses pre-processing dilakukan untuk memastikan bahwa data ekspresi gen sesuai sebelum dianalisis DEG. Proses pre-processing menggunakan *script* tertera:

#Mengambil matriks ekspresi menggunakan *script* tertera:
```
ex <- exprs(gset)
```
#Menghitung distribusi nilai kuantil menggunakan *script* tertera:
```
qx <- as.numeric(quantile(ex,
c(0, 0.25, 0.5, 0.75, 0.99, 1),
na.rm = TRUE))
```
#Menentukan apakah perlu transform log2 menggunakan *script* tertera:

```
LogTransform <- (qx[5] > 100) ||
(qx[6] - qx[1] > 50 && qx[2] > 0)
```
#Melakukan transformasi log2 jika diperlukan menggunakan *script* tertera:
```
if (LogTransform) {
ex[ex <= 0] <- NA
ex <- log2(ex)
}
```

#2.5. Definisi Kelompok Sampel

#Proses definisi kelompok sampel dilakukan untu mengekompokkan sampel yang terdapat pada dataset GSE164805 berdasarkan kondisi biologis.

#Melihat metadata sampel menggunakan *script* tertera:
```
sample_info <- pData(gset)
group_info <- sample_info$source_name_ch1
```
#Mendefinisi kelompok sampel menggunakan *script* tertera:
```
groups <- ifelse(
grepl("mild", group_info, ignore.case = TRUE), "mild",
ifelse(grepl("severe", group_info, ignore.case = TRUE), "severe", "HC")
)
groups <- factor(groups)
table(groups)
```
#Menghilangkan kelompok sampel kontrol (HC) menggunakan *script* tertera:
```
keep_samples <- groups != "HC"
gset <- gset[, keep_samples]
groups <- droplevels(groups[keep_samples])
gset$group <- groups
table(gset$group)
```

#2.6. Design Matrix (Kerangka Statistik)

#Membuat design matrix tanpa intercept menggunakan *script* tertera:
```
design <- model.matrix(~0 + gset$group)
```
#Memberi nama kolom sesuai kelompok menggunakan *script* tertera:
```
colnames(design) <- levels(gset$group)
```
#Menentukan perbandingan biologi yang akan dianalisis menggunakan *script* tertera:
```
library(limma)
contrast.matrix <- makeContrasts(
severe_vs_mild = severe - mild,
levels = design
)
contrast.matrix
```

#2.7. Analisis Differentially Expressed Gene (DEG) dengan Limma

#Analisis DEG dilakukaan menggunakan paket linear model linear dan metode Empirical Bayes. menggunakan *script* tertera:

```
fit <- lmFit(exprs(gset), design)
fit2 <- contrasts.fit(fit, contrast.matrix)
fit2 <- eBayes(fit2)
deg <- topTable(fit2,
coef="severe_vs_mild",
number=Inf,
adjust.method="BH")
head(deg)
```

Untuk menentukan DEGs signifikan digunakan kriteria: Adjusted p-value (FDR) < 0.05 |log2 Fold Change| ≥ 1 menggunakan *script* tertera:

```
deg_filtered <- deg[
  abs(deg$logFC) >= 1 &
  deg$adj.P.Val < 0.05,
]
nrow(deg_filtered)
```

#catatan: gen dengan nilai logFC positif menunjukkan peningkatan ekspresi gen pada kelompok DHF, sedangkan nilai logFC negatif menunjukkan peningkatan ekspresi gen pada kelompok DF

#2.8. Anotasi Nama Gen

#Mengecek anotasi bawaan GEO menggunakan *script* tertera:

```
head(fData(gset))
colnames(fData(gset))
```

#Mengambil informasi anotasi gen bawaan dari GEO menggunakan *script* tertera:

```
feature_info <- fData(gset)
```

#Menyalin hasil DEG menggunakan *script* tertera:

```
topTableResults <- deg
```

#Menambahkan ID gen (rownames) sebagai kolom menggunakan *script* tertera:

```
topTableResults$ID <- rownames(topTableResults)
```

#Memberi ID sebegai simbol gen (apabila tidak ada gen simbol) menggunakan *script* tertera:

```
topTableResults$SYMBOL <- topTableResults$ID
```

#Menambahkan deskripsi gen dari GEO menggunakan *script* tertera:

```
topTableResults$GENENAME <- feature_info[
match(topTableResults$ID, feature_info$ID),
"GENE DESCRIPTION"
]
topTableResults$SYMBOL[topTableResults$SYMBOL == topTableResults$ID] <- NA
```

#Memeriksa hasil anotasi menggunakan *script* tertera:

```
head(topTableResults[, c("ID","SYMBOL","GENENAME")])
```

#2.12. Visualisasi Volcano Plot

#Menyiapkan data untuk volcano plot menggunakan *script* tertera:

```
volcano_data <- data.frame(
logFC = topTableResults$logFC,
adj.P.Val = topTableResults$adj.P.Val,
Gene = topTableResults$SYMBOL
)
```

#Mengklasifikasikan status gen menggunakan *script* tertera:

```
volcano_data$status <- "NO"
volcano_data$status[
volcano_data$logFC > 1 &
volcano_data$adj.P.Val < 0.05
] <- "UP"
volcano_data$status[
volcano_data$logFC < -1 &
volcano_data$adj.P.Val < 0.05
] <- "DOWN"
```

#Membuat Volcano Plot menggunakan *script* tertera:

```
ggplot(volcano_data,
      aes(x = logFC,
          y = -log10(adj.P.Val),
          color = status)) +
  geom_point(alpha = 0.6) +
  scale_color_manual(values = c("DOWN" = "blue",
                                "NO" = "black",
                                "UP" = "red")) +
  geom_vline(xintercept = c(-1, 1), linetype = "dashed") +
  geom_hline(yintercept = -log10(0.05), linetype = "dashed") +
  theme_minimal() +
  ggtitle("Volcano Plot DEGs severe vs mild (GSE164805)")
```

#2.13. Visualisasi Heatmap

#Melakukan pengecekan awal jumlah gen menggunakan *script* tertera:

```
dim(mat_heatmap)
```

#Mengambil top 50 DEG terlebih dahulu (apabila hasil tidak ada gen tersisa) menggunakan *script* tertera:

```
top50 <- deg_filtered[order(deg_filtered$adj.P.Val), ][1:50, ]
```

#Mengambil ekspresi dari gset (menggunakan ID gen sebagai rownames) menggunakan *script* tertera:

```
mat_heatmap <- exprs(gset)[rownames(top50), ]
```

#Menghapus NA (wajib sebelum heatmap) menggunakan *script* tertera:

```
mat_heatmap <- mat_heatmap[complete.cases(mat_heatmap), ]
```

#Membuang gen tanpa variasi menggunakan *script* tertera:

```
mat_heatmap <- mat_heatmap[
apply(mat_heatmap, 1, sd, na.rm = TRUE) > 0,
]
```
#Melakukan pengencekan jumlah gen kembali menggunakan *script* tertera:
```
dim(mat_heatmap)
```
#Membuat annotation sampel menggunakan *script* tertera:
```
annotation_col <- data.frame(
  Group = gset$group
)
rownames(annotation_col) <- colnames(mat_heatmap)
dim(annotation_col)
```
#catatan: kode dim(annotation_col) harus 10 1
#Menjalankan heatmap menggunakan *script* tertera:
```
pheatmap(
  mat_heatmap,
  scale = "row",
  annotation_col = annotation_col,
  show_colnames = FALSE,
  show_rownames = TRUE,
  fontsize_row = 7,
  clustering_distance_rows = "euclidean",
  clustering_distance_cols = "euclidean",
  clustering_method = "complete",
  main = "Top 50 DEGs severe vs mild (GSE164805)"
)
```

#2.14. Menyimpan Hasil

#Menyimpan seluruh worskpace (backup) menggunakan *script* tertera:
```
save.image("backup.RData")
```
#Menghapus objek besar yang sudah tidak dipakai menggunakan *script* tertera:
```
rm(fit, fit2)
gc()
```
#Menyimpan hasil analisis ke file csv menggunakan menggunakan *script* tertera:
```
write.csv(topTableResults,
"Hasil_GSE164805_DEG.csv",
row.names = FALSE)
```
#Analisis selesai menggunakan *script* tertera:
```
message("Analisis selesai. File hasil telah disimpan.")
```

#2.15. Analisis Enrichment Gene Ontology (GO) dan KEGG Pathway

#a. Menginstal dan load package menggunakan *script* tertera:
```
if (!require("BiocManager", quietly = TRUE))
install.packages("BiocManager")
```

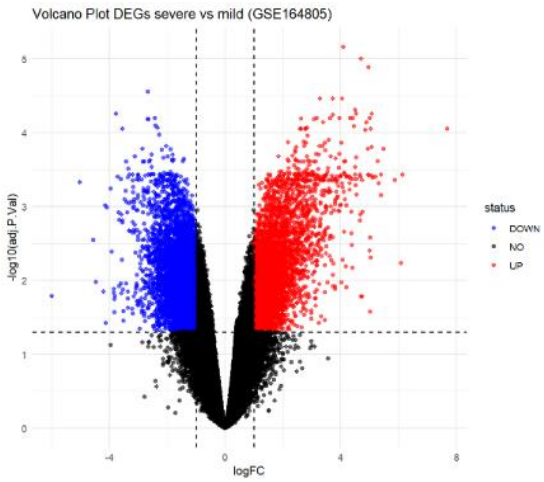
```
BiocManager::install(c(
"clusterProfiler",
"org.Hs.eg.db",
"AnnotationDbi"
), ask = FALSE, update = FALSE)
install.packages("ggplot2")
```
#b. Meload library menggunakan *script* tertera:
```
library(clusterProfiler)
library(org.Hs.eg.db)
library(AnnotationDbi)
library(ggplot2)
```
#c. Memastikan platform dataset menggunakan *script* tertera:
```
annotation(gset)
```
#d. Mengambil data DEG signifikan menggunakan *script* tertera:
```
deg_sig <- topTableResults[
  abs(topTableResults$logFC) >= 1 &
  topTableResults$adj.P.Val < 0.05,
]
nrow(deg_sig)
```
#e. Melakukan pembersihan nama gen menggunakan *script* tertera:
```
symbol_clean <- deg_sig$GENENAME
symbol_clean <- gsub("\\s*\\[.*\\]", "", symbol_clean)
symbol_clean <- trimws(symbol_clean)
symbol_clean <- symbol_clean[symbol_clean != ""]
symbol_clean <- symbol_clean[!grepl(
"novel|query|LOC|ENSG|uncharacterized",
symbol_clean,
ignore.case = TRUE
)]
symbol_clean <- unique(symbol_clean)
length(symbol_clean)
head(symbol_clean)
```
#f. Melakukan convert GENENAME menjadi ENTREZID menggunakan *script* tertera:
```
gene_df <- bitr(symbol_clean,
            fromType = "GENENAME",
            toType   = c("SYMBOL","ENTREZID"),
            OrgDb    = org.Hs.eg.db)
head(gene_df)
nrow(gene_df)
```
#g. Melakukan GO Enrichment (Biological process) menggunakan *script* tertera:
```
ego <- enrichGO(
  gene         = gene_df$ENTREZID,
  OrgDb        = org.Hs.eg.db,
  ont          = "BP",
```

```
  pAdjustMethod = "BH",
  pvalueCutoff  = 0.05,
  readable      = TRUE
)
head(ego)
```
#h. Membuat Grafik Gene Ontology (GO) menggunakan *script* tertera:
```
dotplot(ego, showCategory = 15) +
  ggtitle("GO Biological Process")
barplot(ego, showCategory = 15)
```
#i. Membuat Grafik KEGG Pathway menggunakan *script* tertera:
```
ekegg <- enrichKEGG(
  gene         = gene_df$ENTREZID,
  organism     = "hsa",
  pvalueCutoff = 0.05
)
head(ekegg)
dotplot(ekegg, showCategory = 15) +
ggtitle("KEGG Pathway Enrichment")
barplot(ekegg, showCategory = 15)
ggtitle("KEGG Pathway Enrichment")
```

#3. Hasil dan Interpretasi

#3.1. Identifikasi Differentially Expressed Genes (DEGs) – Volcano Plot

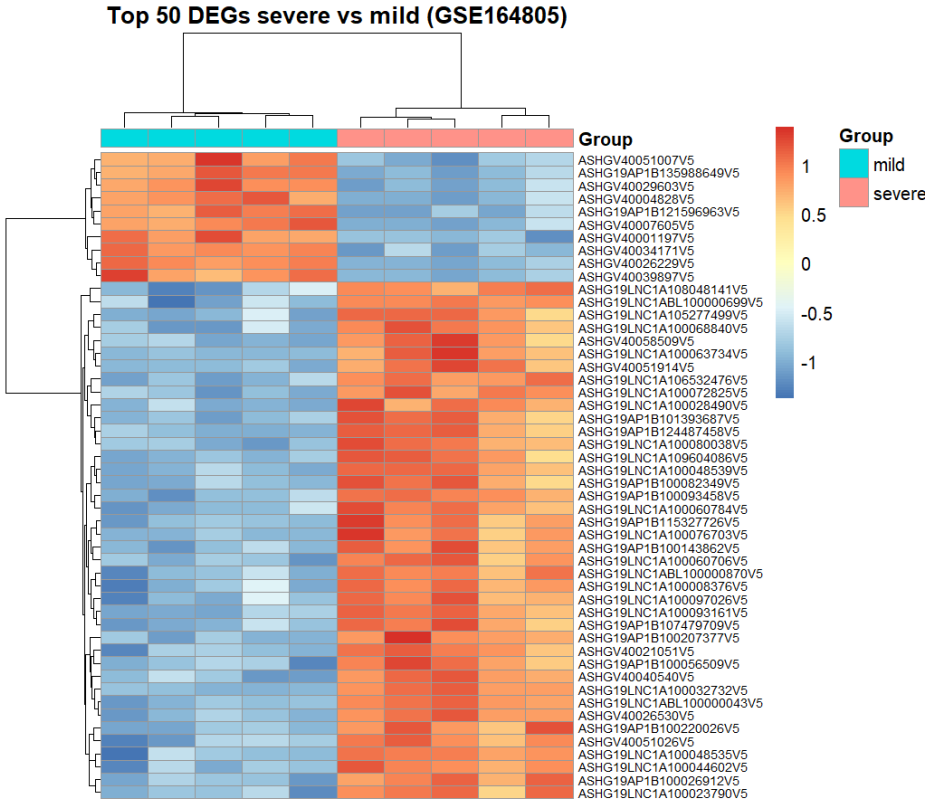
#Gambar 1 menunjukkan volcano plot dari dataset GSE 164805. Titik berwarna merah menunjukkan peningkatan ekspresi gen (*upregulation*) yang signifikan dan titik berwarna biru menunjukkan gen yang mengalami penurunan ekspresi (*downregulation*) yang signifikan. Berdasarkan kriteria seleksi |log2 Fold Change| ≥ 1 dan adjusted p-value (FDR) < 0.05, diperoleh total 9.878 gen signifikan. Dimana sebanyak 5.642 gen mengalami peningkatan ekspresi (*upregulation*) dan sebanyak 4.236 gen mengalami penurunan ekspresi (*downregulation*). Nilai logFC maksimum mencapai >7, menunjukkan adanya perubahan ekspresi gen yang kuat pada penderita COVID-19 *severe*. Secara biologis, jumlah gen upregulation yang meningkat pada penderita COVID-19 *severe* mengindikasikan adanya aktivasi inflamasi dan imun.



#Gambar 1. *Volcano plot* dari DEGs antara sampel penderita COVID-19 mild dan severe menggunakan dataset GSE164805

#3.2. Identifikasi Differentially Expressed Genes (DEGs) Teratas – Heatmap

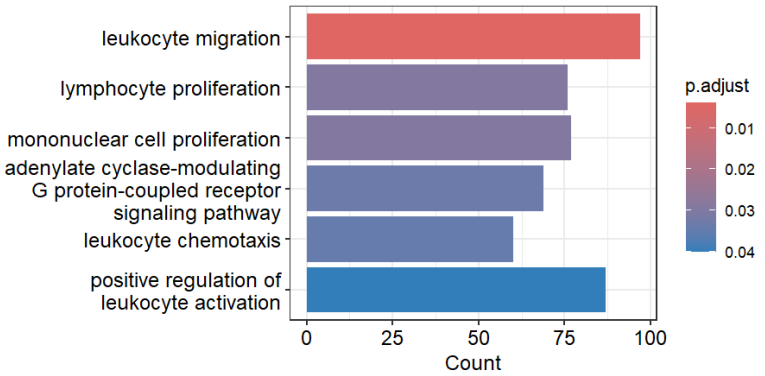
#Gambar 2 menunjukkan heatmap dari 50 DEGs dengan nilai adjusted p-valuenya paling kecil. Pada heatmap memperlihatkan variasi ekspresi gen yang teratur antar kelompok dan adanya pola klaster yang memisahkan kelompok *mild* dan *severe*. Pada kelompok *severe* telihat ekspresi gennya membentuk klaster sendiri. Warna merah padda heatmap menunjukkan ekspresi gen yang tinggi, sedangkan warna biru menunjukkan ekspresi gen yang rendah.



#Gambar 2. *Heatmap* dari DEGs antara sampel penderita COVID-19 mild dan severe menggunakan dataset GSE164805

#3.3. Analisis Enrichment Gene Ontology (GO)

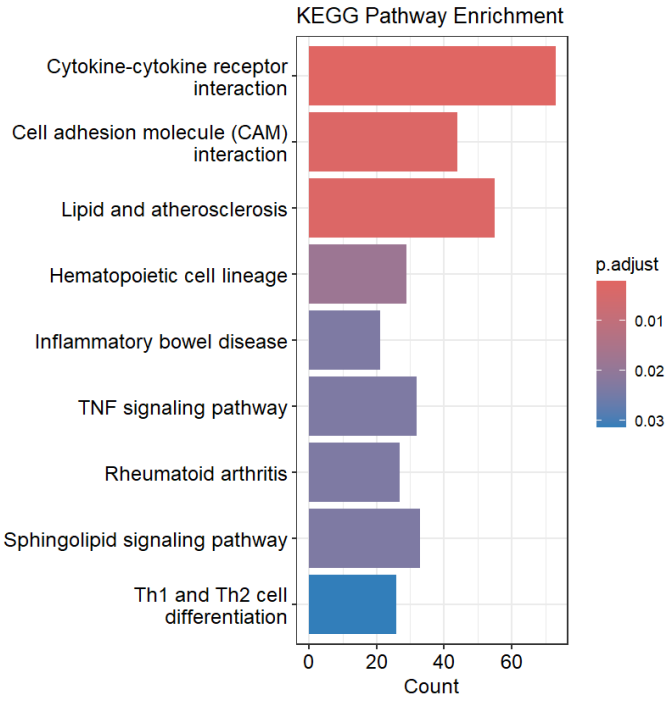
#Gambar 3 menunjukkan hasil *enrichment Gene Ontology* (GO) kategori Biological Process (BP). Pada gambar barplot GO warna merah menunjukkan semakin signifikan, sedangkan warna biru signifikan relatif rendah. Hasil analisis memperlihatkan beberapa proses biologi enrichment utama yang signifikan seperti *leucocyte migration* dengan jumlah gen tertinggi dan paling signifikan, *Lymphocyte proliferation* yang menunjukkan adanya aktivasi dan peningkatan limfosit sebagai respon imun bawaan, *Mononuclear cell proliferation* yang menunjukkan adanya aktivasi sel imun pada penderita COVID-19 severe (monosit dan limfosit), Adenylate cyclase–modulating *G protein–coupled receptor signaling pathway* yang merupakan jalur transmisi sinyal antar sel dan regulasi respon imun, *Leukocyte chemotaxis* yang menunjukkan adanya sinyal kemotaktik untuk mengarahkan sel imun ke lokasi infeksi, dan *Positive regulation of leukocyte activation* yang menunjukkan adnaya peningkatan aktivasi sel imun pada penderita COVID-19 *severe*.



#Gambar 3. Analisis *Gen Ontology* (GO) *enrichment* dari sampel penderita COVID-19 mild dan severe menggunakan dataset GSE164805

#3.4. Analisis Enrichment KEGG Pathway

#Gambar 4 menunjukkan hasil *enrichment KEGG pathway*. Terlihat beberapa jalur yang di *enrichment* seperti *JAK-STAT signaling pathway, Viral protein interaction with cytokine and cytokine receptor, Toll-like receptor signaling pathway, NF-kappa B signaling pathway,* dan *Cytokine-cytokine receptor interaction*. Jalur-jalur ini berperan penting dalam proses aktivasi dan regulasi sinyal imun adaptif, serta produksi mediator inflamasi. Jalur-jalur tersebut apabila diaktivasi secara berlebihan dapat menyebabkan kerusakan jaringan.



#Gambar 4. Analisis KEGG pathway enrichment dari DEGs sampel penderita COVID-19 mild dan severe menggunakan dataset GSE164805

#4. Kesimpulan

#Hasil analisis DEGs pada dataset GSE164805 didapatkan total DEGs sebanyak 9.878 gen, dengan 5.642 gen *upregulation* dan 4.236 gen *downregulation*. Hasil visualisasi volcano plot menunjukkan bahwa dominasi gen mengalami upregulation pada kelompok COVID-19 severe. Heatmap 50 DEGs teratas menunjukkan adanya perbedaan ekspresi gen antara penderita COVID-19 *mild* dan *severe*. Hasil analisis GO dan KEGG menunjukkan bahwa progresi *mild* menjadi *severe* berhubungan dengan aktivasi respon inflamasi, sistem imum bawaan, dan jalur sitokin, NF-kB, dan JAK-STAT. Ekspresi gen yang ditemukan dapat menjadi dasar pengembangan biomarker untuk deteksi keparahan COVID-19.

#5. Daftar Pustaka

#Al-Jindeel, T. J., Al-Karawi, A. S., Kready, H. O., & Mohammed, M. (2022). Prevalence of COVID-19 virus infection in asymptomatic volunteers in Baghdad city / Iraq during 2021. *International Journal of Health Sciences*, 6(52), 552-5530. Doi: 10.53730/ijhs.v6nS2.6396.

#Hao, Y. J., Yu, L. W., Mei, Y. W., Lan, Z., Jian, Y. S., Ji, M. C., & De, P. W. (2022). The origins of COVID-19 pandemic: a brief overview. *Transboundary and Emerging Disease*, 69(6), 3181-3197. Doi:10.1111/tbed.14732

#Zhang, Q., Yuting, M., Kaihang, W., Zujun, Z., Wnbiao, C., Jifang, S., Yunqing, Q., Hongyang, D., & Lanjuan, L. (2021). Inflammation and antiviral immune response associated with severe progression of COVID-19. *Frontier in Immunology*, 12, 1-12. Doi: 10.3389/fimmu.2021.631226