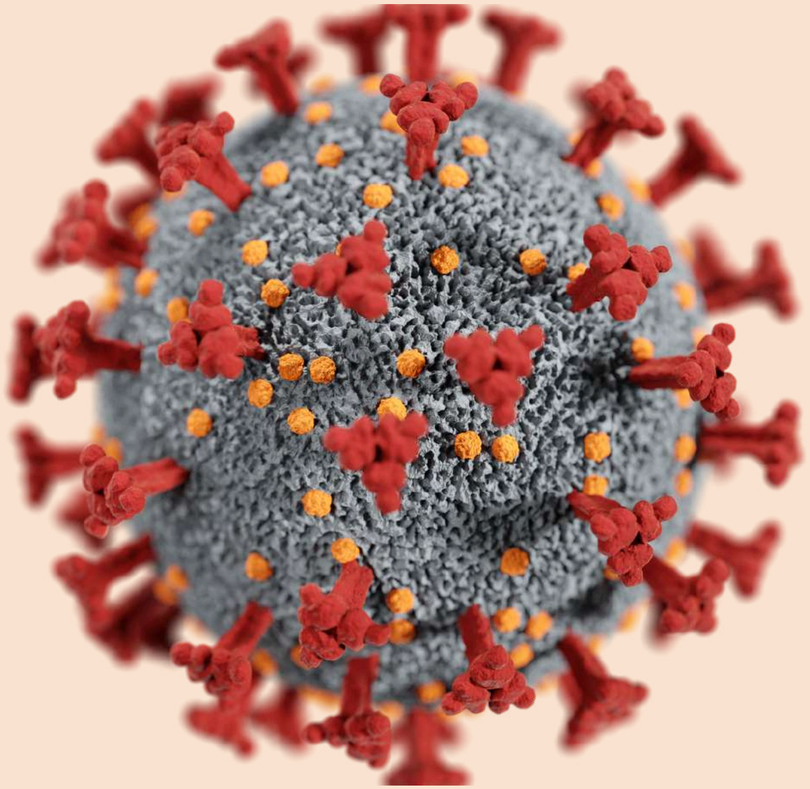
EXPLORING GENE EXPRESSION PATTERNS AS PREDICTIVE MARKERS FOR COVID-19 TRANSMISSION: A COMPUTATIONAL ANALYSIS

By Group 6

Introduction

- COVID-19, short for Coronavirus Disease 2019, is caused by the novel coronavirus SARS-CoV-2.
- Emerged in December 2019 in Wuhan, China, COVID-19 swiftly evolved into a global pandemic.
- It has presented unparalleled challenges to public health, economies, and societies worldwide.



Problem Statement & Solution

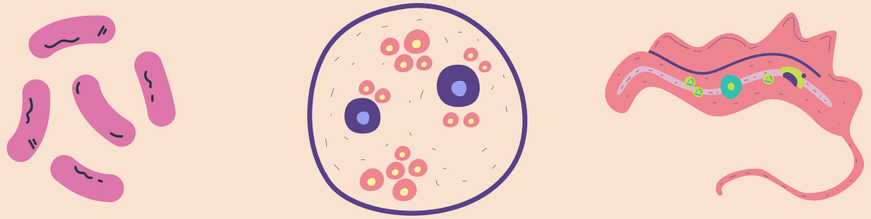
- SARS-CoV-2, with one of the largest RNA virus genomes, presents unique challenges for treatment and vaccine development.
- Using RNA sequencing (RNA-Seq) data, this study explores variations in human gene expression in response to SARS-CoV-2 infection.
- The research assesses if the expression levels of certain human genes vary across different severities of COVID-19.
- Additionally, the study aims to uncover the molecular basis of differing patient responses to SARS-CoV-2 infection.

DATASETS

- We identified three datasets related to SARS-CoV-2, namely **GSE166424**, **GSE178824**, and comprising a total of 38 samples sourced from GEO datasets.
- In total, these datasets encompass 35,413 genes across all samples.
- To narrow down our search in the GEO database, we utilized keywords such as "SARS-CoV-2," "Human," and "RNA-Seq."

WHY THESE DATASETS?

We chose these two datasets because they best matched our study's requirements, offering suitable characteristics and a sufficient number of samples for statistical significance, unlike other datasets we considered.



Workflow

Data Collection

- Access the Gene Expression Omnibus (GEO) database and retrieve the raw expression data (e.g., microarray or RNA-seq data) for the two datasets: GSE166424, GSE178824.

Differential Expression Analysis

We will use GEO2R to differentially analyse each dataset with the given accession numbers. We will focus on the overexpressed genes coming out from the following comparisons between experimental groups:

- "Severe" vs "Asymptomatic”
- "Severe" vs "Mild"
- "Severe" vs "Moderate"

Workflow

Gene Enrichment Analysis

- Once the differentially expressed genes (DEGs) have been identified, conduct gene enrichment analysis to uncover biological pathways or functions associated with these genes using DAVID.

Potential Biomarkers Identification

- Focus on the overexpressed genes identified from the differential expression analysis for each comparison between experimental groups, as specified in the study.
- Evaluate the biological significance and potential as biomarkers for COVID-19 severity based on the identified gene expression patterns and enrichment analysis results.

Data Collection

Access the Gene Expression Omnibus (GEO) database and retrieve the raw expression data (e.g., microarray or RNA-seq data) for the two datasets: GSE166424 & GSE178824.

Series GSE166424	Query DataSets for GSE166424
Status	Public on Apr 24, 2021
Title	Asymptomatic COVID-19: disease tolerance with efficient anti-viral immunity against SARS-CoV-2
Organism	Homo sapiens
Experiment type	Expression profiling by high throughput sequencing
Summary	The immune responses and mechanisms limiting symptom progression in asymptomatic cases of SARS-CoV-2 infection remain unclear. We comprehensively characterized transcriptomic profiles, cytokine responses, neutralization capacity of antibodies and cellular immune phenotypes of asymptomatic patients with acute SARS-CoV-2 infection to identify potential protective mechanisms. Compared to symptomatic patients, asymptomatic patients had higher counts of mature neutrophils and lower proportion of CD169+ expressing monocytes in the peripheral blood. Systemic levels of pro-inflammatory cytokines were also lower in asymptomatic patients, accompanied by milder pro-inflammatory gene signatures. Mechanistically, a more robust systemic Th2 cell signature with a higher level of virus-specific Th17 cells and a weaker yet sufficient neutralizing antibody profile against SARS-CoV-2 was observed in asymptomatic patients. In addition, asymptomatic COVID-19 patients had higher systemic levels of growth factors that are associated with cellular repair. Together, asymptomatic patients mount less pro-inflammatory and more protective immune responses against SARS-CoV-2 indicative of disease tolerance. Insights from this study highlight key immune pathways that could serve as therapeutic targets to prevent disease progression in COVID-19.
Overall design	RNA was extracted from whole blood collected from 30 asymptomatic patients from the Singapore cohort. Timepoints selected for extraction were during active infection (PCR-positive). Bridging samples were used for data comparison to previous dataset (GSE155454; transcriptomic signatures of WT SARS-CoV-2 infection).

Series GSE178824	Query DataSets for GSE178824
Status	Public on Jun 25, 2021
Title	Severe COVID-19 is Characterized by an Impaired Type I Interferon Response and Elevated Levels of Arginase Producing Granulocytic Myeloid Derived Suppressor Cells
Organism	Homo sapiens
Experiment type	Expression profiling by high throughput sequencing
Summary	COVID-19 ranges from asymptomatic in 35% of cases to severe in 20% of patients. Differences in the type and degree of inflammation appear to determine the severity of the disease. Recent reports show an increase in circulating monocytic-myeloid-derived suppressor cells (M-MDSC) in severe COVID 19 that deplete arginine but are not associated with respiratory complications. Our data shows that differences in the type, function and transcriptome of granulocytic-MDSC (G-MDSC) may in part explain the severity COVID-19, in particular the association with pulmonary complications. Large infiltrates by Arginase 1+ G-MDSC (Arg+G-MDSC), expressing NOX-1 and NOX-2 (important for production of reactive oxygen species) were found in the lungs of patients who died from COVID-19 complications. Increased circulating Arg+G-MDSC depleted arginine, which impaired T cell receptor and endothelial cell function. Transcriptomic signatures of G-MDSC from patients with different stages of COVID-19, revealed that asymptomatic patients had increased expression of pathways and genes associated with type I interferon (IFN), while patients with severe COVID-19 had increased expression of genes associated with arginase production, and granulocyte degranulation and function. These results suggest that asymptomatic patients develop a protective type I IFN response, while patients with severe COVID-19 have an increased inflammatory response that depletes arginine, impairs T cell and endothelial cell function, and causes extensive pulmonary damage. Therefore, inhibition of arginase-1 and/or replenishment of arginine may be important in preventing/treating severe COVID-19.
Overall design	We did RNA-seq in Covid-19 patients (4 severe, 4 asymptomatic, 4 convalescent) and controls (n=4). RNA was extracted from purified granulocytic-myeloid derived suppressor cells (G-MDSC), qualified by Agilent 2100 and quanitfied by Qubit. Librarires were prepared using the Illumina's TruSeq Stranded mRNA kit and sequenced at 2 x 75bp (50M reads/sample).

Differential Gene Expression

We will use GEO2R to differentially analyse each dataset with the given accession numbers. We will focus on the overexpressed genes coming out from the following comparisons between experimental groups:

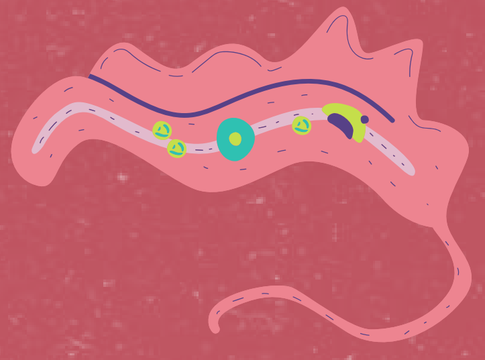
- "Severe" vs "Asymptomatic”
- "Severe" vs "Mild"
- "Severe" vs "Moderate"
- “Moderate to Severe” vs “Mild”

Differential Gene Expression

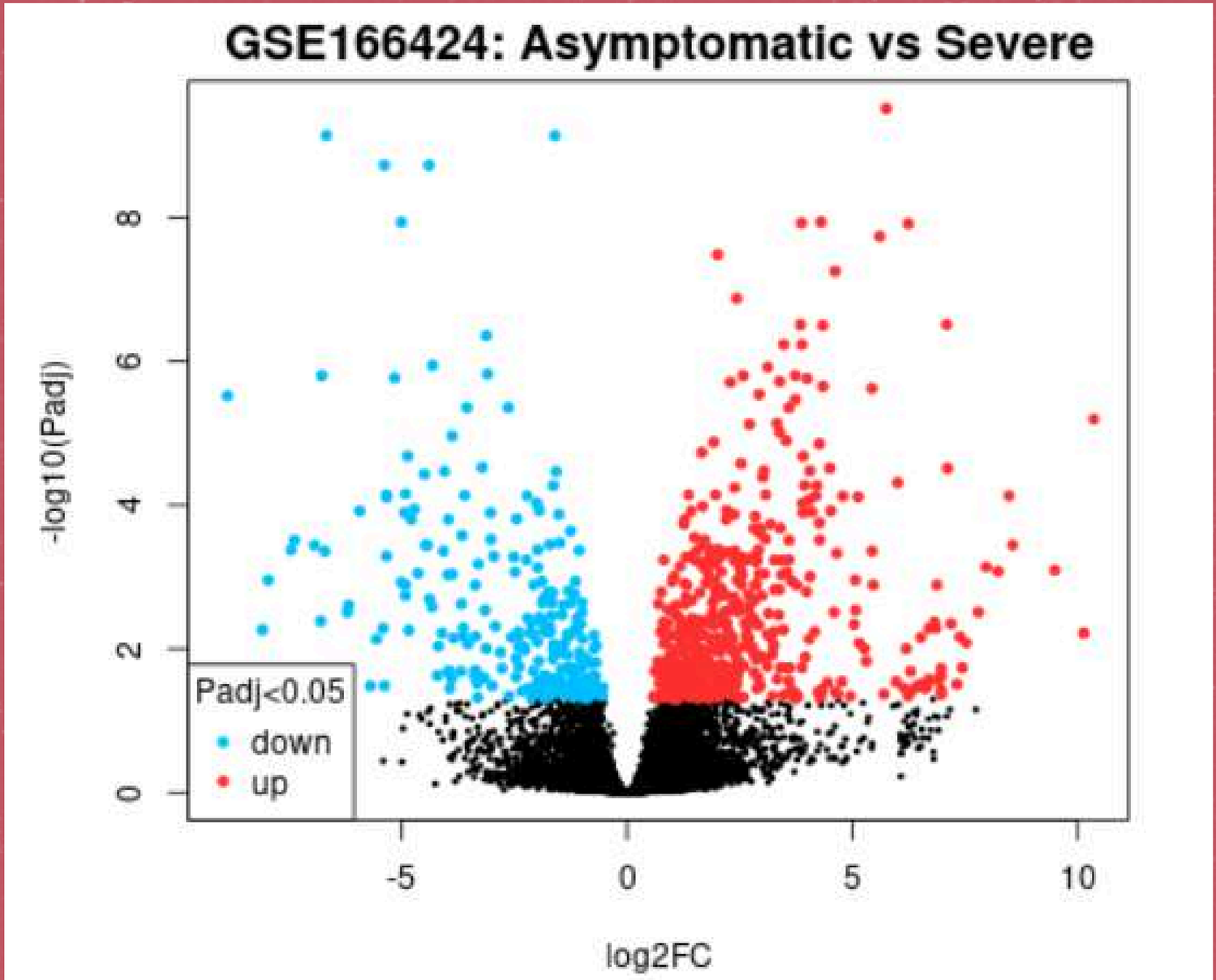
We decided to focus on the differentially expressed genes for “Asymptomatic” and “Severe” conditions in the next steps of the analysis. Both these conditions represent two key points in the fight against COVID-19:

- Asymptomatic individuals serve as significant vectors for viral spread due to challenges in identification and timely treatment, facilitating unchecked virus circulation.
- Severe patients face substantial risks, often requiring hospitalization and experiencing long-term consequences, even with successful therapy.

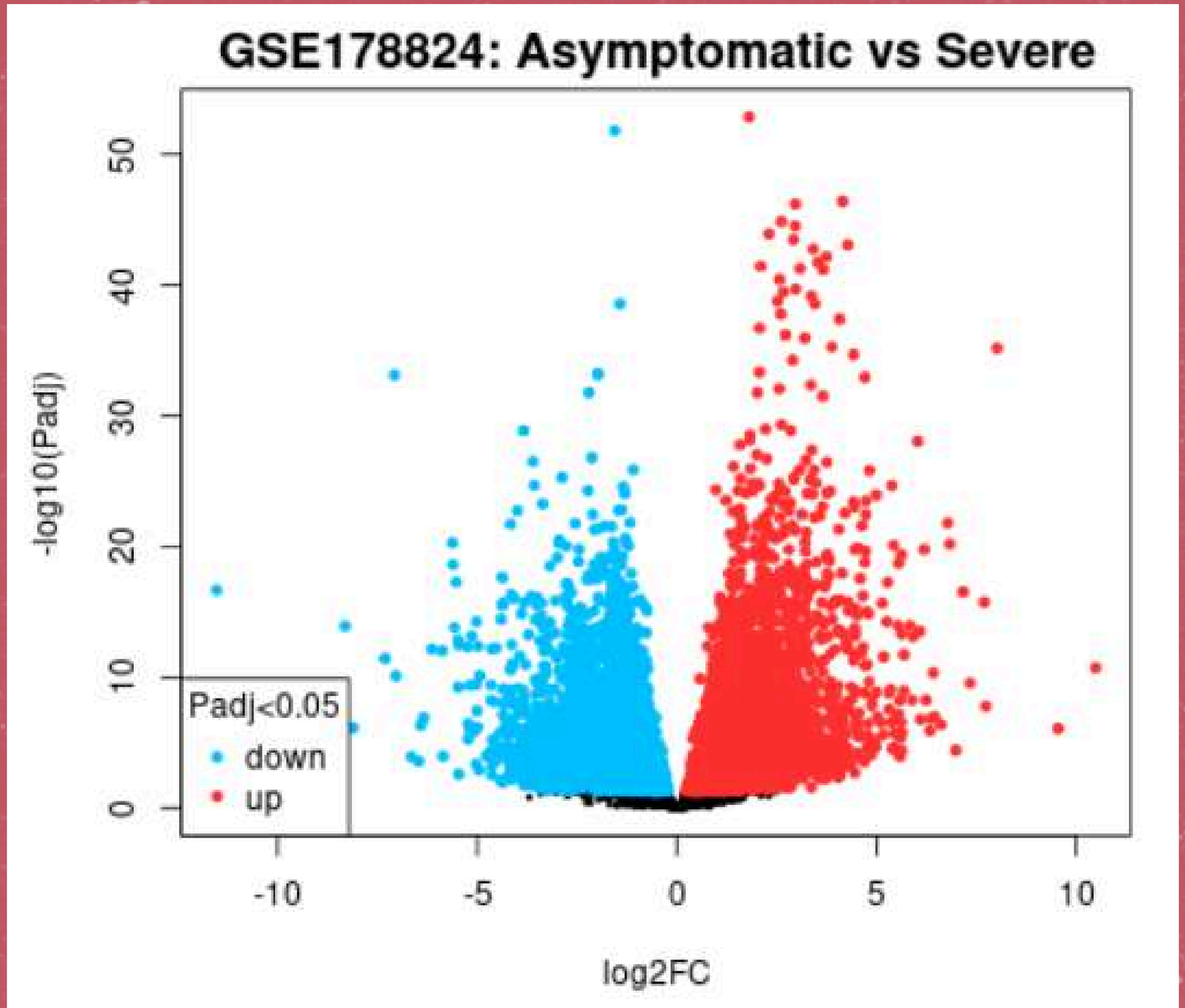
In both cases, early diagnosis is decisive in terms of both control of the spread and treatment of the disease



Differential Gene Expression



Differential Gene Expression



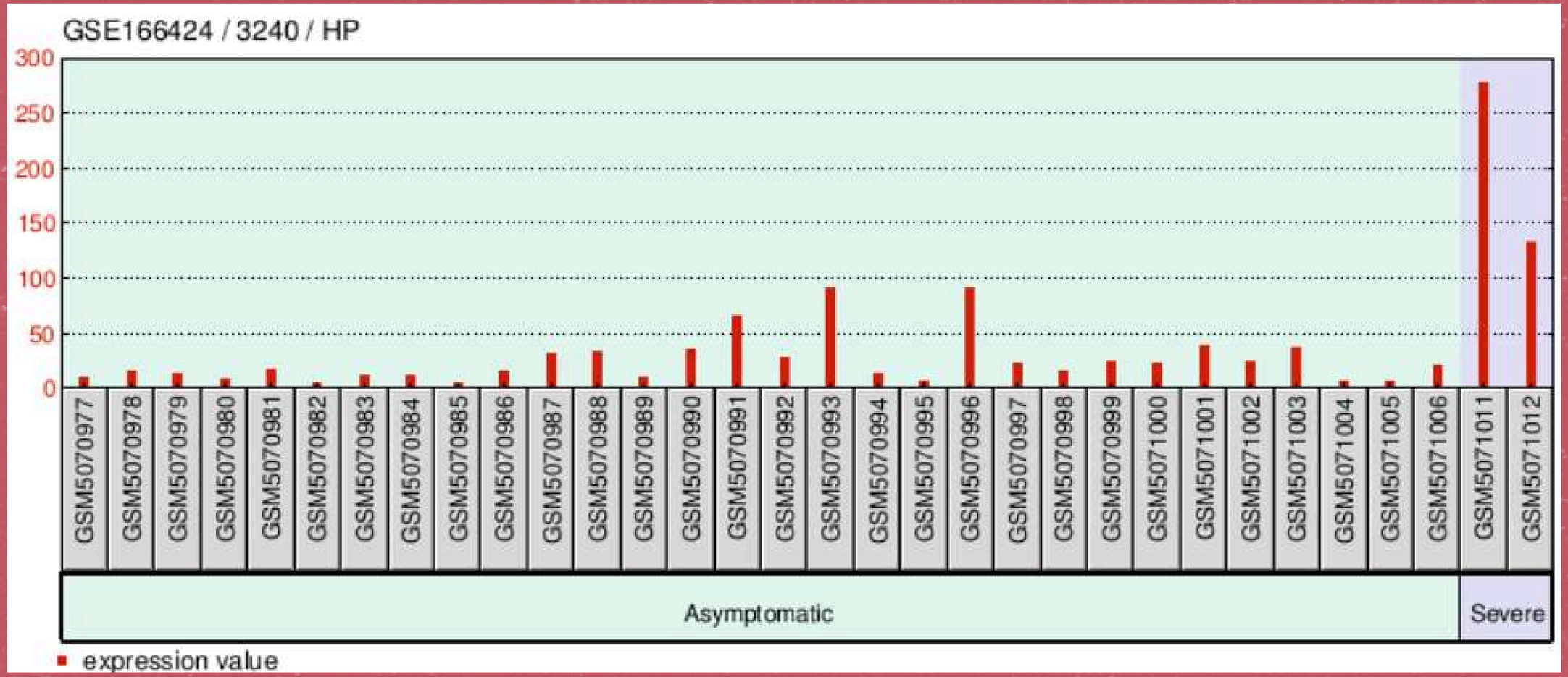
Differential Gene Expression

We identified three particular genes which resulted in being up-regulated in Severe condition and not that differentially expressed in the Asymptomatic one:

- ENSG00000257017:HP
- ENSG00000183019: MCEMP1
- ENSG00000204936:CD177

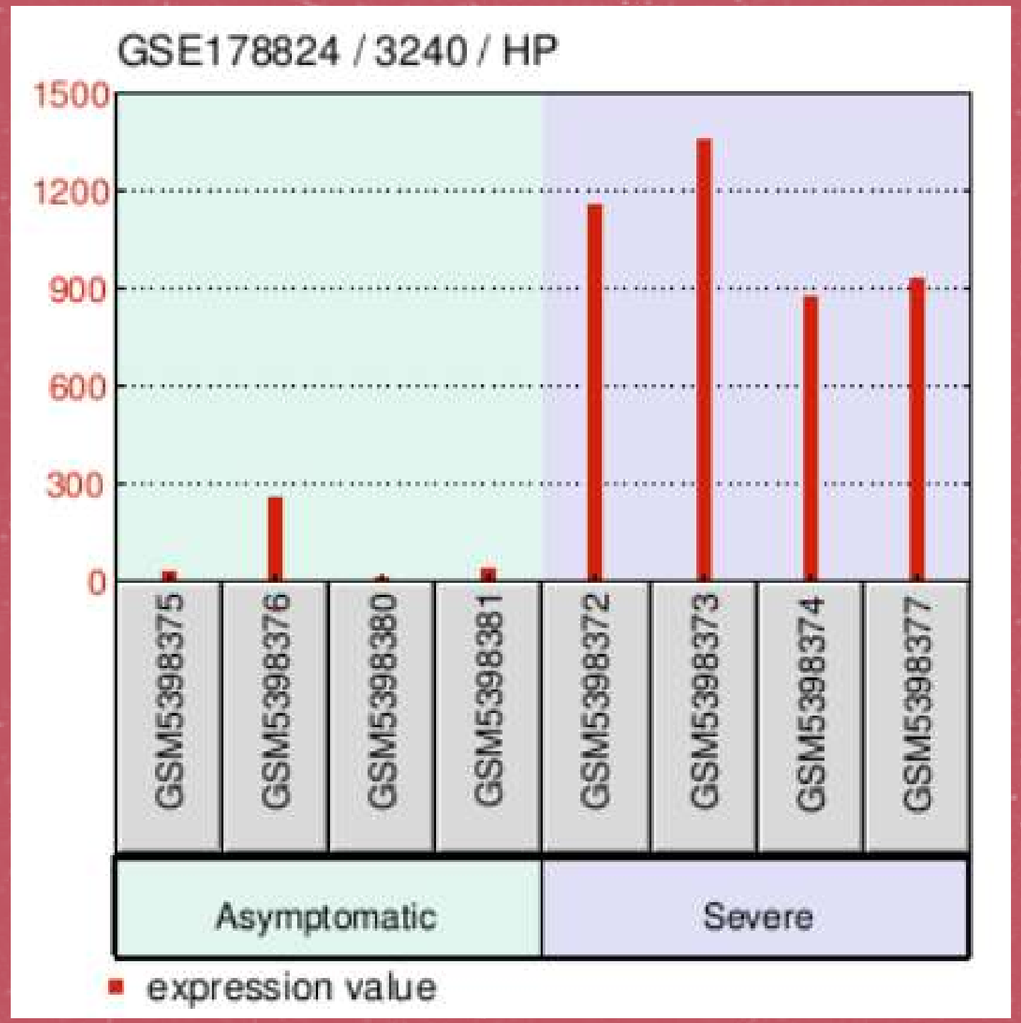
Differential Gene Expression

Gene Expression Value for HP in GSE166424



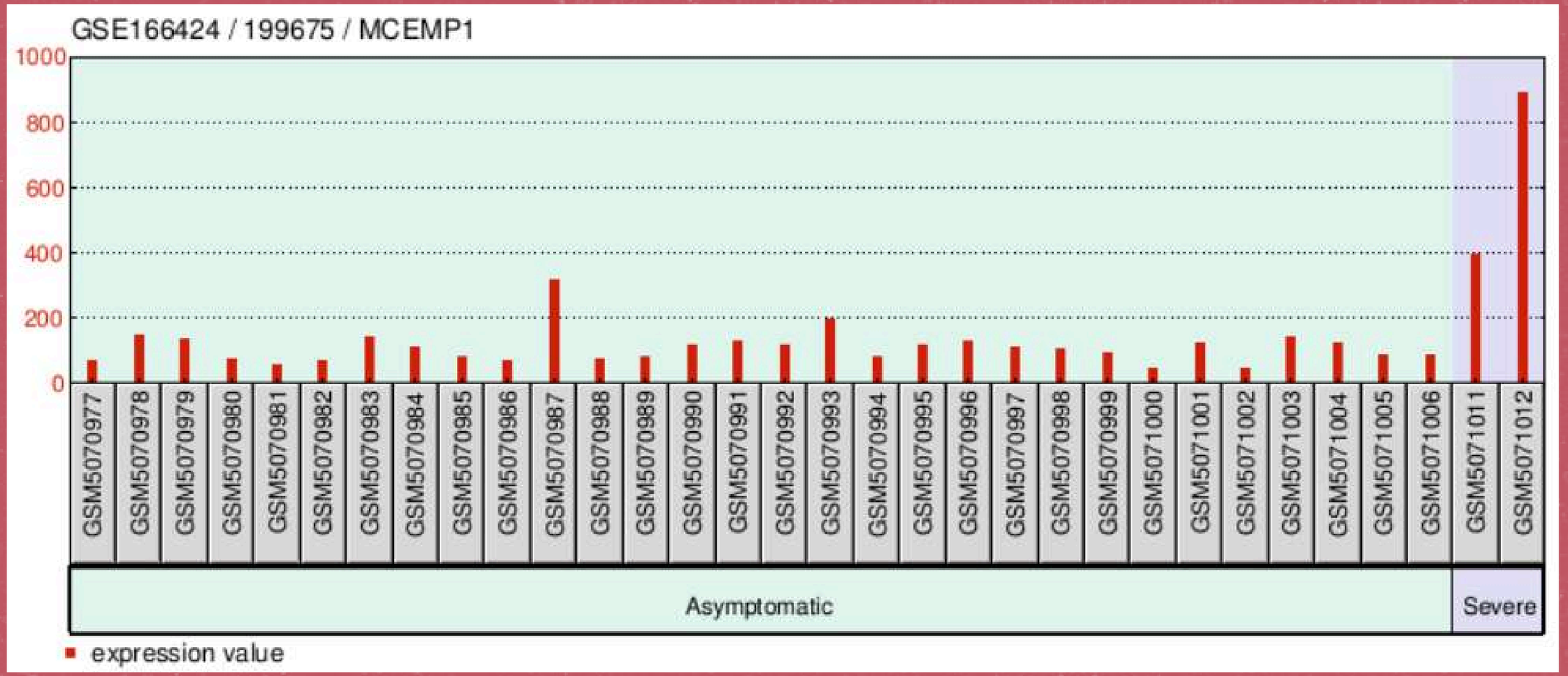
Differential Gene Expression

Gene Expression Value for HP in GSE178824



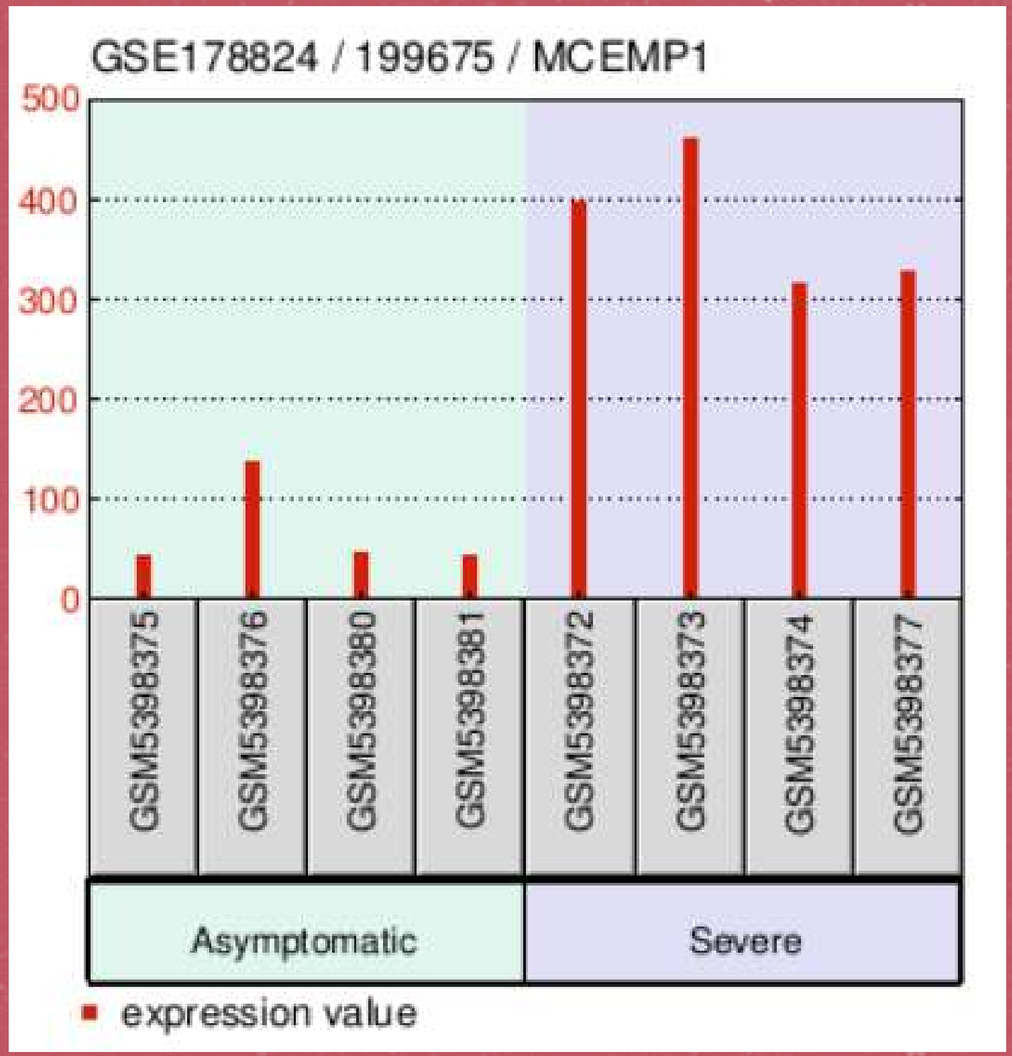
Differential Gene Expression

Gene Expression Value for MCEMP1 in GSE166424



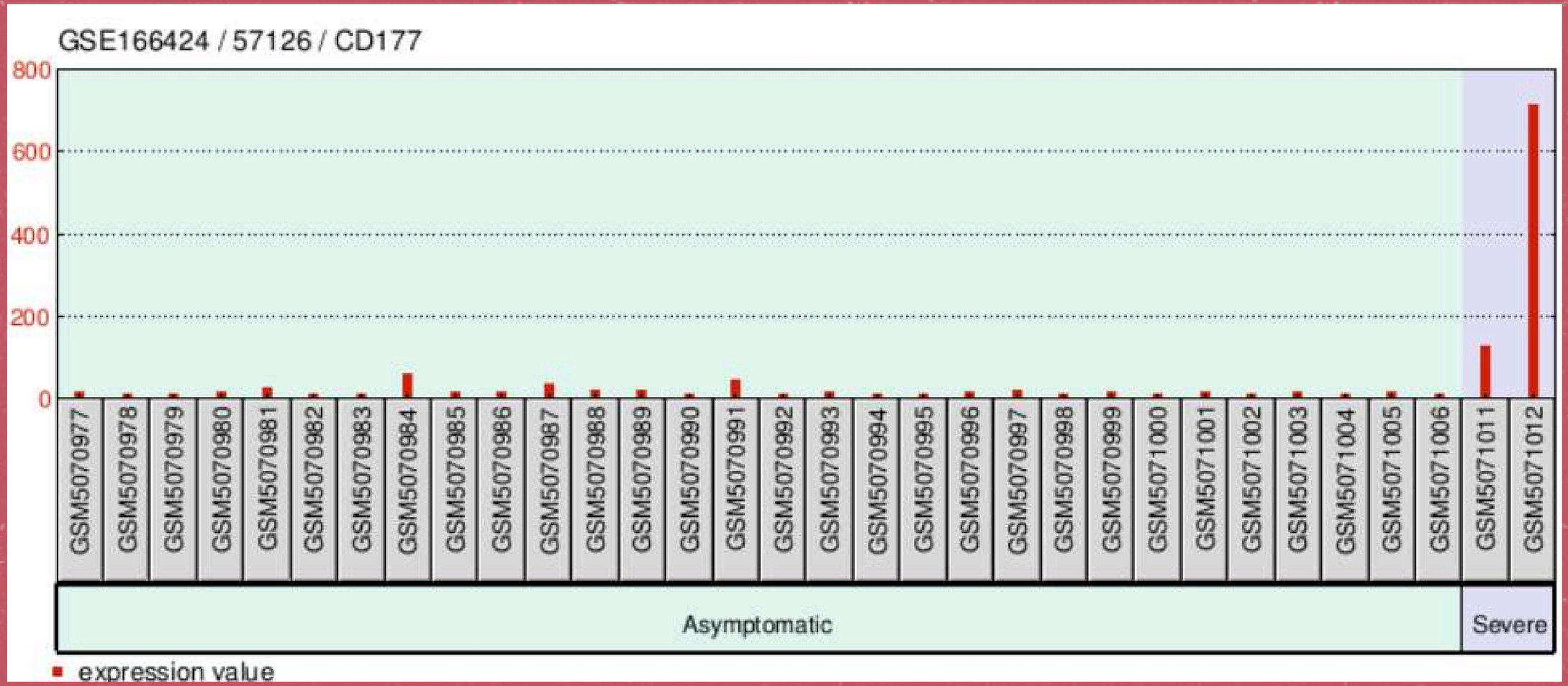
Differential Gene Expression

Gene Expression Value for MCEMP1 in GSE178824



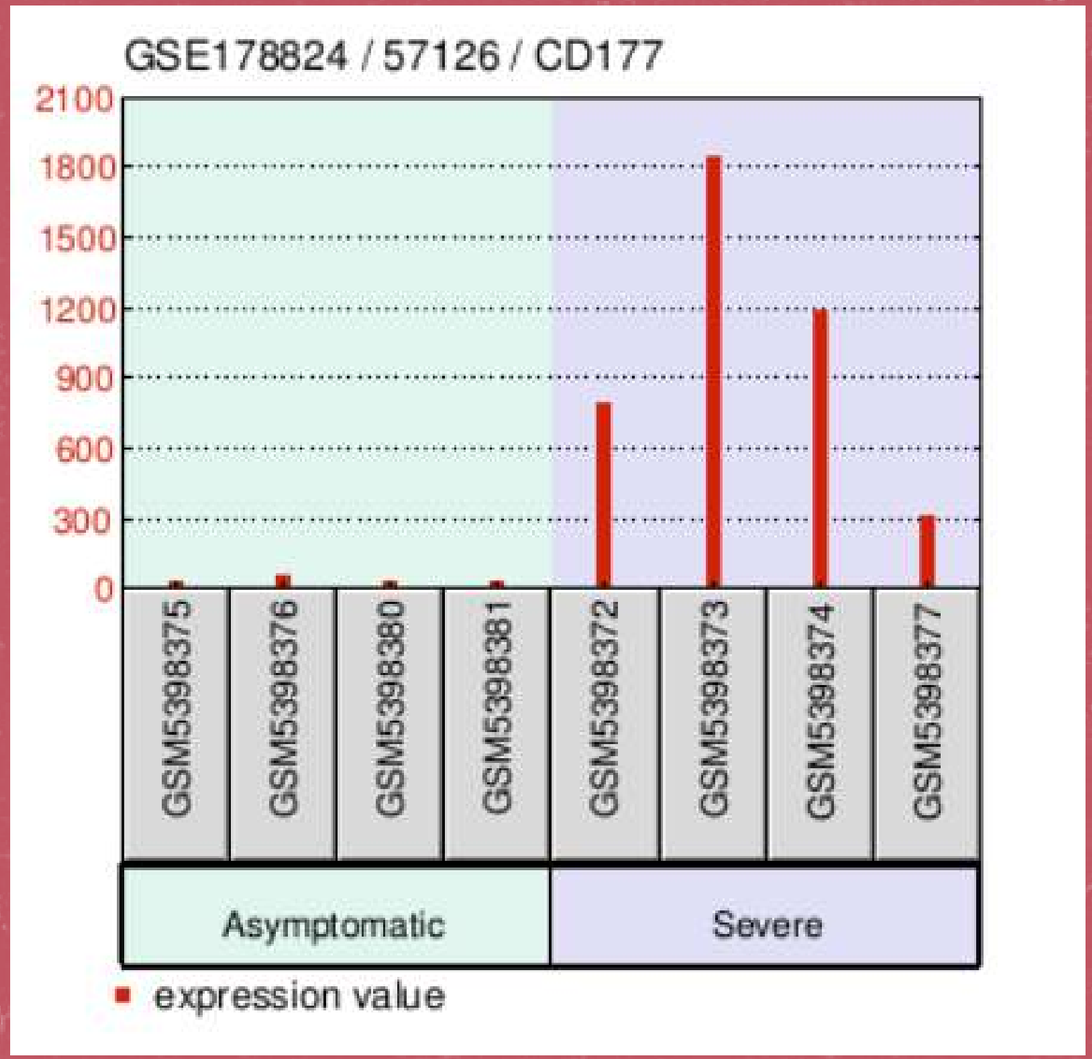
Differential Gene Expression

Gene Expression Value for CD177 in GSE166424



Differential Gene Expression

Gene Expression Value for CD177 in GSE178824



Gene Set Enrichment Analysis

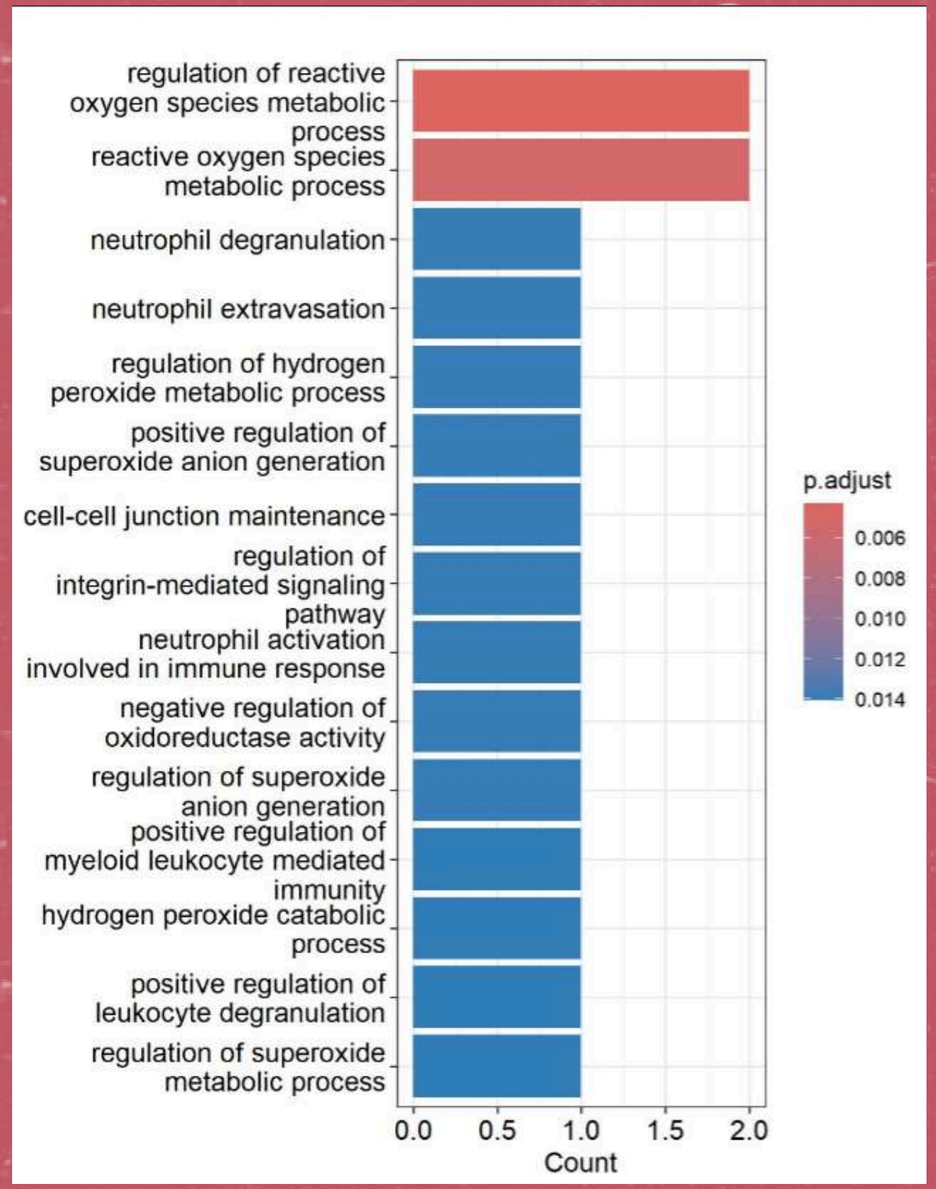
DOSE R package was used with the clusterProfiler R package to perform enrichment analysis in order to discover disease associations of high-throughput biological data.

Following analysis have been done using the above mentioned libraries:

- **Barplot of Over Representation Analysis results for “Asymptomatic vs Severe” condition.**

*Code for the same is provided in the file

Gene Set Enrichment Analysis



Barplot of Over Representation Analysis results for "Asymptomatic vs Severe" condition.

Y-axis lists immune-related processes, while the X-axis shows GeneRatio, representing pathway enrichment based on differentially expressed genes.

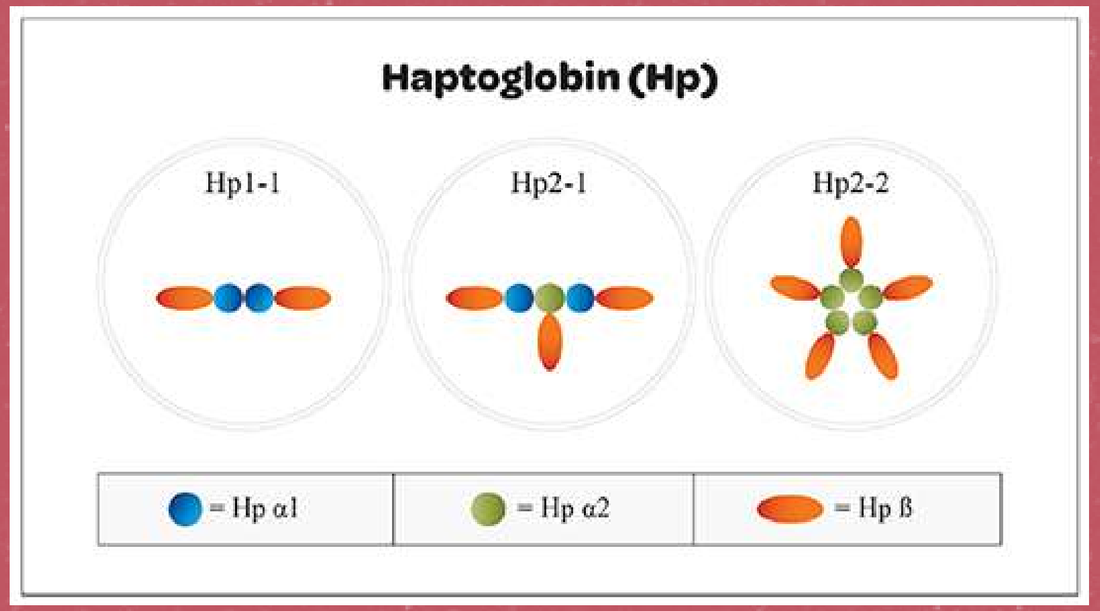
Bar plot illustrates process frequency at various enrichment levels, offering a visual summary of enrichment distribution.

Analysis reveals significant enrichment of immune response, neutrophil activation, and oxidative stress pathways in SARS-CoV-2 context, indicating their pivotal roles in host response to infection.

Potential Biomarkers Identification

We identified three up-regulated genes in the severe condition:

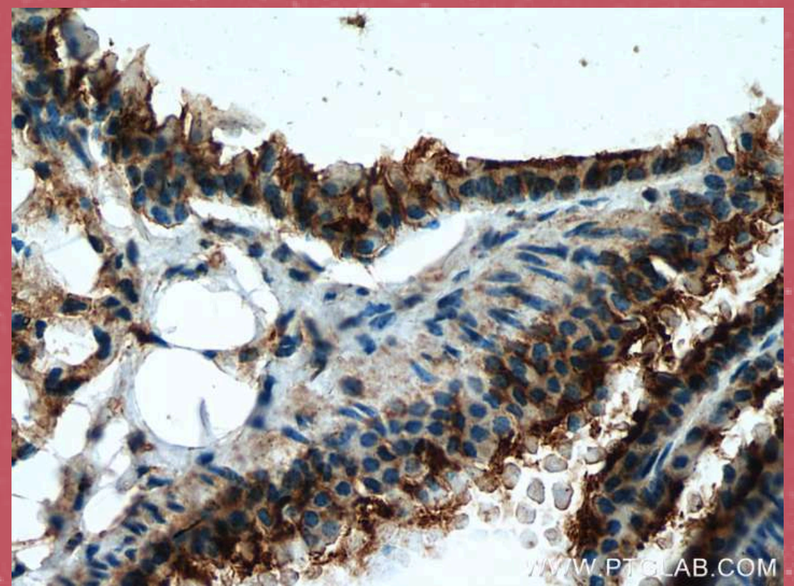
- Haptoglobin (Hp) is a plasma protein with antioxidant, immunomodulatory properties and has an important biological function in the host defense responses to infection and inflammation

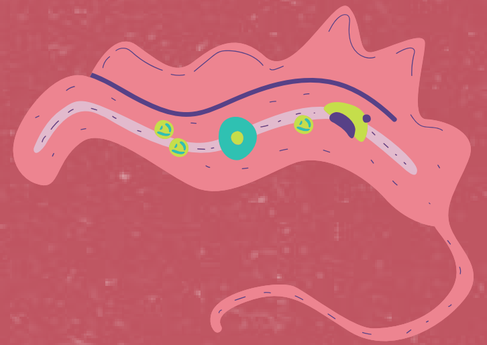


Potential Biomarkers Identification

We identified three up-regulated genes in the severe condition:

- Mast cell-expressed membrane protein1 (MCEMP1) is a type II transmembrane protein that is highly expressed in human lung tissue1. MCEMP1 is primarily expressed in myeloid lineage immune cells, such as lung-resident mast cells and alveolar macrophages. MCEMP1 is one of the top inducible genes in many inflammatory diseases, such as asthma,idiopathic pulmonary fibrosis,cancer,sepsis and stroke.

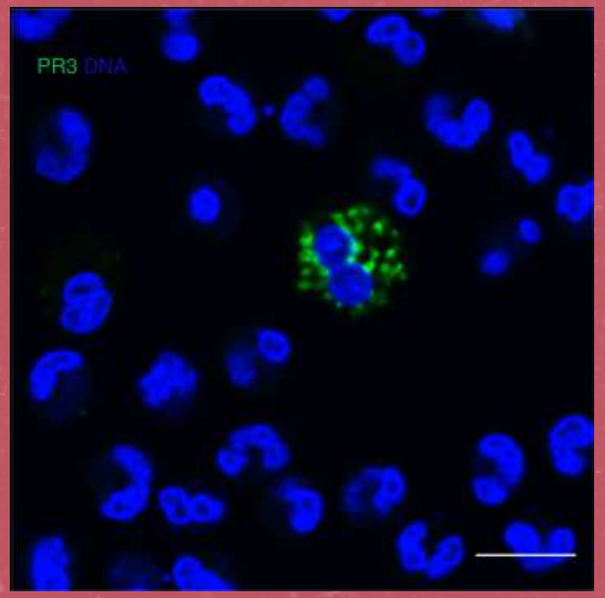




Potential Biomarkers Identification

We identified three up-regulated genes in the severe condition:

- CD177: This gene encodes a glycosyl-phosphatidylinositol (GPI)-linked cell surface glycoprotein that plays a role in neutrophil activation. The protein can bind platelet endothelial cell adhesion molecule-1 and function in neutrophil transmigration. Mutations in this gene are associated with myeloproliferative diseases. Over-expression of this gene has been found in patients with polycythemia rubra vera



Conclusion

Our research proposes hypotheses linking host genetic diversity to symptom profiles and laboratory characteristics, accounting for the varying pathophysiology and severity of COVID-19 among patients with underlying chronic diseases and risk factors.

RNA-Seq data allowed us to detect and highlight all those genes involved in COVID-19 progression whose expression level may vary changing from a specific grade of disease severity to another, helping us to go deeper into the genetic causes of COVID-19.

In particular, we focused on severe and asymptomatic conditions and performed an accurate literature search on some of the most up- and down-regulated genes in the correspondent DEGs list, which we found particularly interesting. We identified three genes being up-regulated in the severe condition and not-differentially expressed in the asymptomatic condition: HP,MCEMP1,CD177.

Conclusion

Our analysis produced a list of differentially expressed genes (DEGs) for each condition, upon which we also performed a Gene Set Enrichment Analysis (GSEA), to highlight any possible correlation between

For future study, a more detailed investigation of the identified genes may provide a context to introduce new human gene biomarkers for different conditions of COVID-19 disease for rapid identification and treatment.

Understanding the biology and functional significance of the three obtained genes (HP,MCEMP1,CD177) provides insights into the molecular processes associated with disease severity in COVID-19. Integrating these potential biomarkers into clinical practice may hold several potential implications

Thank you!

Our Team

Dhawal Garg
2022160
(Interpretation of Results)

Kartavya Chauhan
2022236
(Differential Expression Analysis)

Swayam Ghanghas
2022526
(GSEA)

Adtiya Dixit
2022030
(Differential Expression Analysis)

Nishchay Yadav
2022332
(GSEA)

Ayush Vikram Singh
2022606
(PPT and Data Collection)

Resources page

https://www.sciencedirect.com/science/article/pii/S2001037023002192

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7205601/

Tools used

https://www.ncbi.nlm.nih.gov/geo/geo2r/

Links to Dataset

https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE166424

https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE178824