PROTEOMIC ANALYSIS OF SARS-COV-2 IMPACT ON HOST PROTEIN EXPRESSION AND STABILITY

By
SUBBULAKSHMI NATARAJAN

Table of Contents

Introduction

SARS-CoV-2, the virus that causes COVID-19, has accelerated global research efforts to better understand viral mechanisms within host cells. Protein-protein interactions and their impact on cellular processes such as replication, immunological response, and protein synthesis are at the heart of this research. Mass spectrometry-based proteomics has emerged as a critical method for identifying host and viral proteins implicated in these interactions, providing insight into the virus's ability to hijack cellular machinery during its lifecycle. Advanced bioinformatics tools such as MSFragger and gProfiler allow for extensive analysis, revealing essential functional roles and pathways influenced by the infection.

Despite advancements, there are still information gaps in understanding how SARS-CoV-2 preferentially changes host protein synthesis over time. This study intends to fill these gaps by analyzing time-course proteome and translatome data to identify key proteins altered by the virus during infection. By pinpointing active and stable host proteins and pathways involved in infection, this research seeks to uncover potential therapeutic targets within these viral-host interactions.

This study looks at how SARS-CoV-2 affects host protein expression, with an emphasis on replication and immunity proteins. We hope to pinpoint the key protein

Methods

mzML File Search and Protein Identification

1. **Database Search Execution Using MSFragger**

MSFragger, which was installed on a Virtual Machine (VM) hosted by Monash, was used to search raw mzML files containing mass spectrometry data. Directories for SARS-CoV-2 bait and control samples were given to each group; these were set up in MSFragger's parameter file (`fragger.params`).

An output file (`combined_protein.tsv`) including comprehensive details on detected proteins, peptide counts, and intensity values was produced by running the MSFragger script (`runFraggerPipe.sh`). The essential information for analysing the protein interactions between SARS and CoV-2 was provided by this file.

Output Result Analysis

1. Data Inspection and Initial Filtering

For additional investigation, the result file `combined_protein.tsv` was downloaded to local systems. To verify the presence of proteins, data examination in Excel, R, or Python concentrated on columns that represented distinct intensity values, protein IDs, and peptide counts. By eliminating proteins found in the EGFP control samples, proteins were filtered to isolate those only found in SARS-CoV-2 bait samples. Finding SARS-CoV-2 protein interactions required this stage.

2. Biological Function and Pathway Analysis

To ascertain the biological functions and pathways connected to the discovered proteins, additional investigation was carried out. Using the gProfiler and GO databases, Gene Ontology (GO) enrichment analysis revealed important biological processes, molecular roles, and cellular constituents linked to the proteins. A confidence-based depiction of SARS-CoV-2 protein interactions was subsequently produced by visualising networks of protein-protein interactions using the IntAct Database, which gave information on biological activities.

Time-Course Analysis and Translatome Correlation

1. Temporal Expression Analysis

To monitor changes in expression, time-course data of protein expression was analysed for many infection timepoints (2, 6, 10, and 24 hours). Line plots were used to visualise expression profiles, and volcano plots, which plotted -log10 P values on the y-axis and log2 fold changes on the x-axis, were used to highlight important proteins that were up- and down-regulated. This revealed proteins whose expression varied significantly over time, offering information about the cellular course of SARS-CoV-2 infection.

2. Proteome and Translatome Comparison

To compare with the proteome data, more translatome data was analysed in Week 11 to evaluate alterations in freshly synthesised proteins. Consistency in protein production responses to infection was shown by visualising expression correlation analyses between proteome and translatome data. This method gave a more complete view of the viral influence on host protein synthesis and assisted in differentiating the translational effects of SARS-CoV-2 infection.

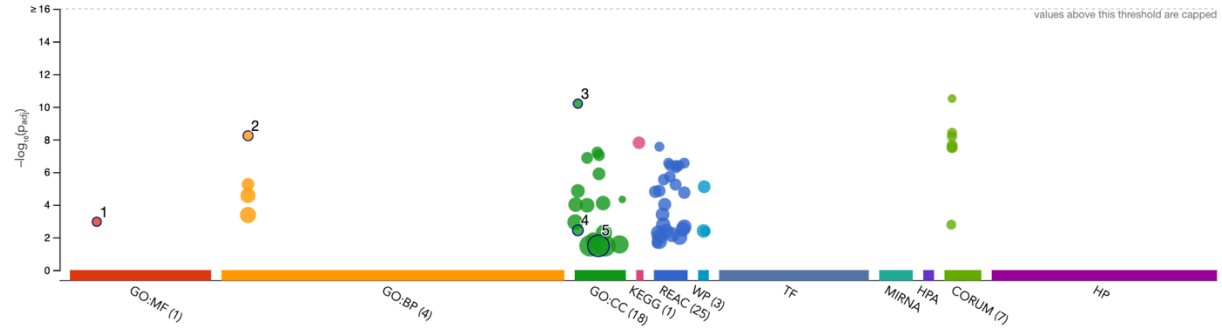
Software and Tools Utilised

1. **MSFragger:** Used for initial mzML database searching and protein identification.
2. **Virtual Machine (VM):** Hosted MSFragger and provided computational resources.
3. **Excel and Python:** Used for data inspection, filtering, and visualisation, converting TSV data into data frames for flexible handling.
4. **gProfiler and Gene Ontology Databases**: Facilitated GO enrichment analysis, revealing protein functions and pathway insights.
5. **IntAct Database:** Enabled verification and visualisation of protein-protein interactions.
6. **Data Visualization Tools:** created in Seaborn and Matplotlib illustrated data trends and protein expression variations.

From examining mzML data for initial protein identification to analysing biological pathways and expression patterns over several time points, this methodical approach offered a comprehensive methodology that led to a thorough understanding of host response and SARS-CoV-2 protein relationships.

Results

Identification of proteins that are confident hits for protein-protein interactions with your assigned SARS-CoV-2 protein



ID	Source	Term ID	Term Name	p_{adj} (query_1)
1	GO:MF	GO:0009374	biotin binding	1.091×10^{-3}
2	GO:BP	GO:0006269	DNA replication, synthesis of RNA primer	5.898×10^{-9}
3	GO:CC	GO:0005658	alpha DNA polymerase:primase complex	6.345×10^{-11}
4	GO:CC	GO:0005652	nuclear lamina	3.647×10^{-3}
5	GO:CC	GO:0043233	organelle lumen	3.221×10^{-2}

version	e111_eg58_p18_f463989d
date	29/10/2024, 12:33:14
organism	hsapiens

g:Profiler

Figure 1: *GO Enrichment Analysis of SARS-CoV-2 Host Protein Interactions*

This visualization from g represents the results of a Gene Ontology (GO) enrichment analysis. Here's a breakdown of each component:

- **X-Axis (GO Terms):** The categories on the x-axis indicate various GO terms related to biological processes, molecular functions, and cellular components.
- **Y-Axis (-log10(p-value)):** How closely each gene set is linked to the observed data, with higher rankings suggesting more substantial relationships. Points above the threshold are regarded statistically significant and highlight the most important biological processes.
- **Coloured dots:** distinct GO terms enriched in the dataset. The colours are often assigned to distinct functional categories, making it easier to distinguish across sorts of phrases or pathways.

Key Enrichment

- **DNA Replication & RNA Primer Synthesis:** These are processes required for cell division and DNA synthesis, implying that proteins in the dataset are involved in replication mechanisms.

- **DNA Polymerase Primase Complex:** This complex is required for the commencement of DNA synthesis and connects to the replication machinery, which can be exploited by viral replication.
- **Organelle Lumen:** This word refers to sub-cellular compartmentalization, highlighting the probable sites within cells where these proteins are localized or functional

Identification of viral proteins in the time course dataset and visualisation of their expression profile

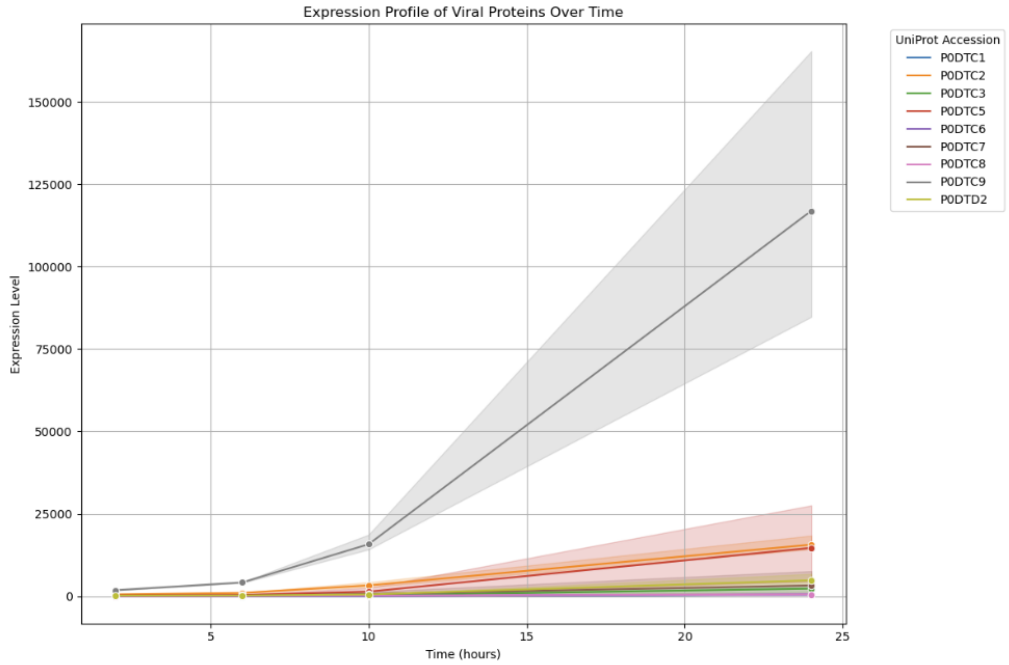


Figure 2: *Expression Profile of Viral Proteins Over Time*

Using data at 2, 6, 10, and 24 hours, this investigation looked at the expression profiles of SARS-CoV-2 viral proteins over the course of a 24-hour infection period, emphasising proteins with notable expression alterations.

❖ **Key Finding**

- **Elevated Expression Over Time**

Several viral proteins, most notably the one indicated by the black line, had a noticeable rise by the 24-hour mark, indicating a crucial function in the later phases of infection. As the infection worsens, this pattern suggests increased viral activity.

- **Variability in Expression**

Not every protein showed significant increases. Some stayed low, indicating jobs that don't need to be expressed as much. These patterns are easily distinguished in the figure, which makes it easier to identify the most pertinent proteins.

- **Confidence Intervals**

The robustness of our results is increased by the shaded areas surrounding each line, which show variability. Broader intervals identify proteins with higher measurement variability.

Visualisation of differentially expressed proteins at each time point of the time course dataset

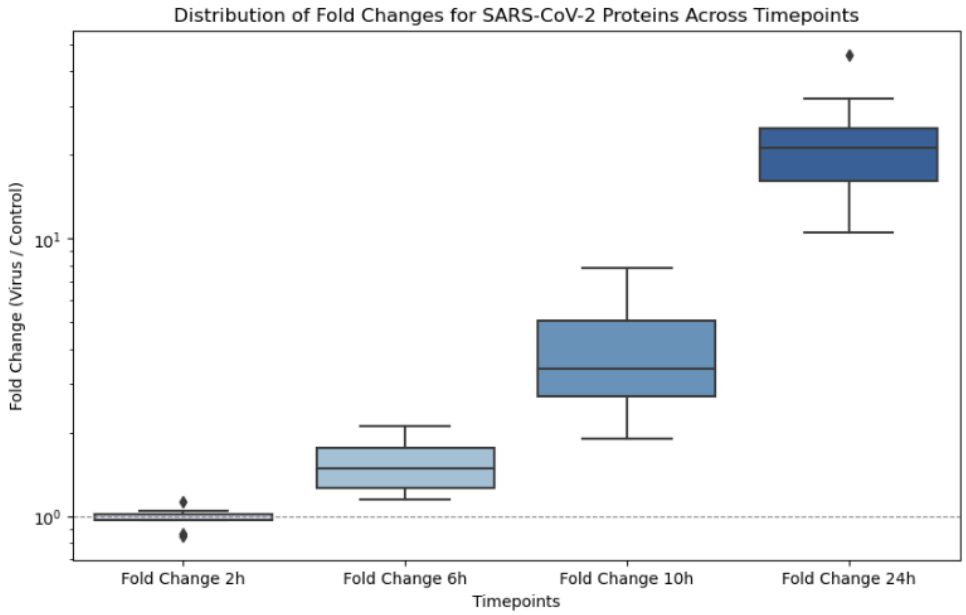


Figure 3: *Distribution of Fold Changes for SARS-CoV-2 Proteins Across Timepoints*

At four time points—two, six, ten, and twenty-four hours—the box plot shows the distribution of fold changes in SARS-CoV-2 protein expression in comparison to control samples.

Key Observations

- **Increasing Fold Change Over Time**

As the infection progresses, SARS-CoV-2 protein expression grows, as evidenced by the median fold change rising steadily at each time point. This is especially evident at the 24-hour mark.

- **Variability Across Timepoints**

Over time, particularly at 10 and 24 hours, the interquartile range, which represents the distribution of fold changes, gets wider. This implies that different expression responses are highlighted by the fact that some proteins exhibit significant fold changes while others have modest increases in expression.

- **Outliers**

A small number of proteins with abnormally low or high fold changes in comparison to others at that time point are observed at 2 and 24 hours. These could point to proteins that play special functions either early or late in the infection process.

Comparison of the proteome and translatome datasets

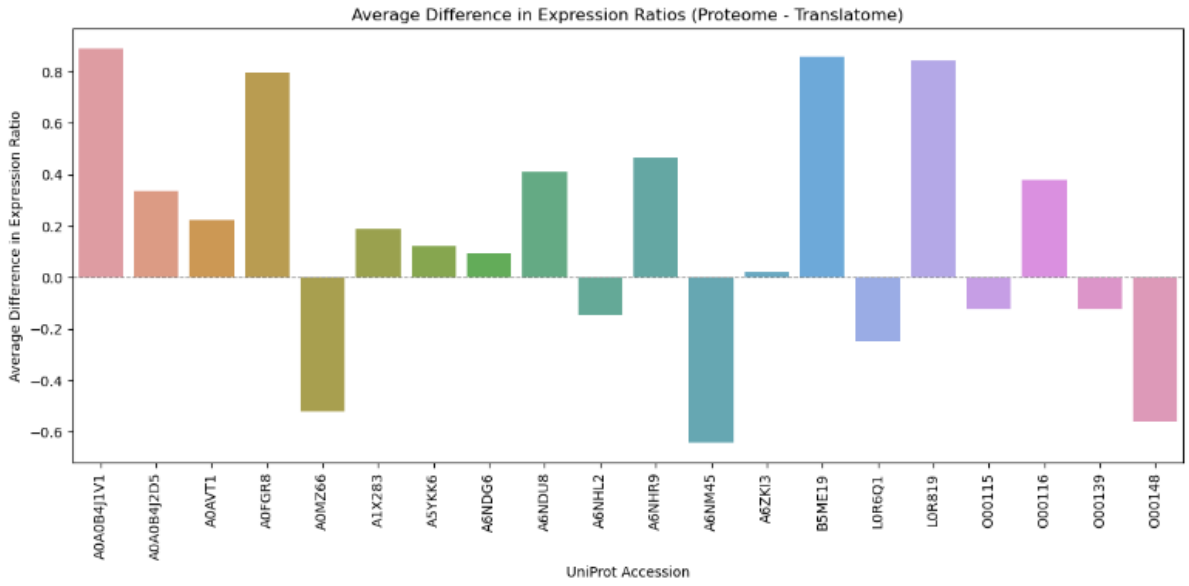


Figure 4: *Average Difference* in Expression Ratios (Proteome – Translatome) for SARS-COV-2

The average expression differences for SARS-CoV-2 proteins between the proteomes and translatome at four time points—two, six, ten, and twenty-four hours—are displayed in a bar plot.

Key Findings

- **Positive Differences**

During infection, proteins with larger proteome ratios (bars above 0) have less active synthesis or are more stable, indicating that they stay in the protein pool without undergoing substantial new synthesis.

- **Negative Differences**

Active synthesis is indicated by proteins with greater translatome ratios (bars below 0), most likely in response to SARS-CoV-2 infection. Over time, these proteins are actively generated, suggesting that they play a part in viral response mechanisms.

- **Implications**

This differentiation aids in distinguishing between proteins that are actively generated and those that are stable, offering information on how SARS-CoV-2 affects host protein synthesis and identifying possible areas for therapeutic investigation.

Identification of biological functions/pathways/GO enrichments, where relevant, for any of the above points

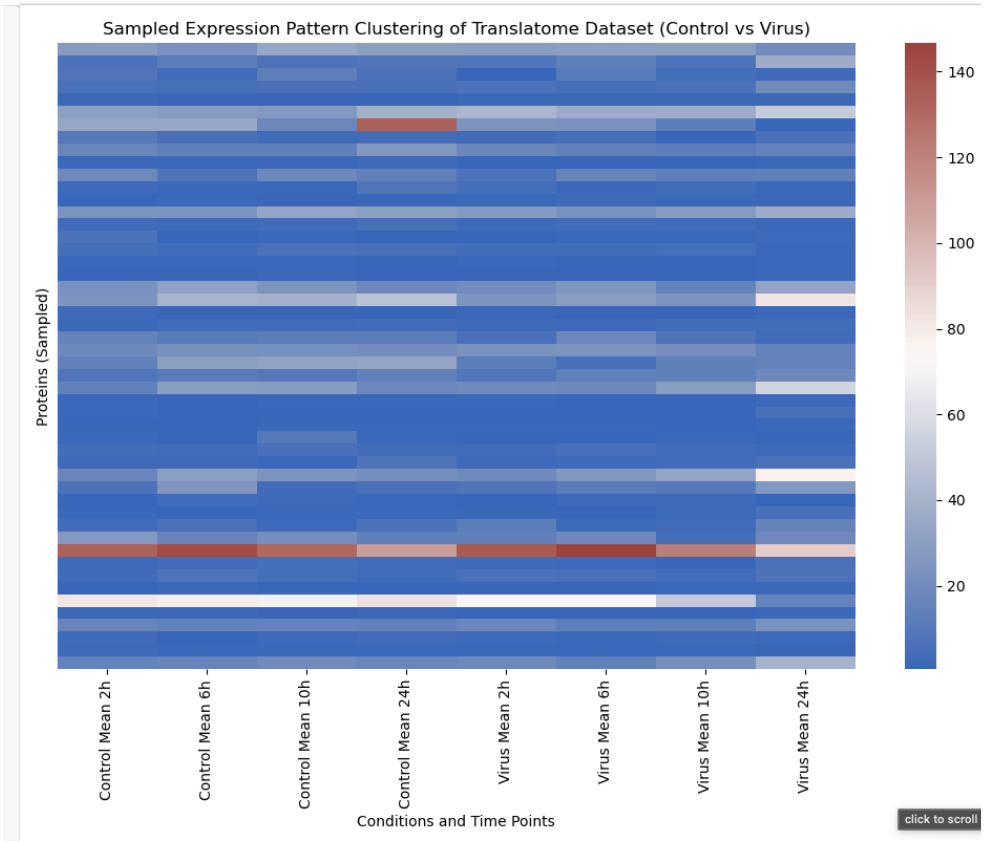


Figure 5: *Sampled Expression Pattern Clustering of Translatome Dataset (Control vs. Virus)*

The translatome dataset's heatmap compares the expression levels of specific proteins in control and virus-infected samples over time (2, 6, 10, and 24 hours).

- **Clustering of High Expression**

Clusters of higher expression (shown in red) are mainly seen in viral samples, particularly at later time points (10 and 24 hours), according to the heatmap. This implies that when the infection worsens, some proteins are increased.

- **Contrast between Control and Virus**

In general, the virus-infected samples exhibit more vibrant colours than the controls, suggesting that their bodies are producing more proteins in reaction to the infection.

- **Temporal Changes**

The gradual rise in expression from early to late time points demonstrates the dynamics of protein synthesis across time, mirroring the viral influence on host protein synthesis.

Discussion

By analysing time-course expression changes, protein interactions between SARS-CoV-2, and comparisons between proteome and translatome datasets, this work shed light on how the virus affects host cellular functions.

• Enrichment analysis using Gene Ontology (GO) and visualization.

This GO enrichment study reveals that SARS-CoV-2 targets host proteins involved in DNA replication and RNA primer production, which are critical for cell division and viral replication. The enrichment of the DNA Polymerase Primase Complex shows that SARS-CoV-2 uses these replication processes to increase its infection. Furthermore, GO keywords associated to organelle lumen point to specific subcellular regions where these interactions are most likely to occur, facilitating viral proliferation and immune evasion.

• Key Protein Interaction

Certain human proteins that interact with SARS-CoV-2 were found by the combined protein dataset. Numerous of them have similar functions in cellular replication and immunological signalling pathways, which the virus most likely uses to replicate. GO enrichment analysis revealed pathways linked to protein synthesis and immunological response, offering hints about how SARS-CoV-2 affects host processes.

- Modulation of Host Protein Expression

Time-course research revealed that active synthesis is indicated by the rise in expression of some host proteins during infection, particularly in the translatome. These proteins either indicate host responses that the virus manipulates or are probably engaged in antiviral defences. On the other hand, other proteins are inhibited, which would indicate that the virus is interfering with host functions.

- Viral Protein Expression Over Time

SARS-CoV-Two proteins displayed different patterns of temporal expression, some of which increased significantly by 24 hours. This rise points to crucial functions in later phases of infection, maybe in promoting viral reproduction or eluding immunological reactions. The modest expression of other proteins suggests that they played support roles during the infection.

- Comparison of the Proteome and the Translatome

According to the comparison, proteins with greater translatome ratios indicate active synthesis, whereas those with higher proteome ratios are more stable. This difference emphasizes how the virus selectively alters host protein synthesis, encouraging the synthesis of proteins that are necessary for its life cycle.

Conclusion

In summary, our study shed light on SARS-CoV-2's interactions with host proteins, notably those involved in replication and immunological function. Significant changes in the expression of viral and host proteins over the course of the infection demonstrate how the virus exploits and alters host cellular processes to maintain replication and dodge immune responses. The comparison of proteome and translatome data demonstrated SARS-CoV-2's selective effect on host protein stability and production.

However, there are several drawbacks, such as the use of in vitro infection models that may not fully reproduce in vivo circumstances. Future research should include more refined time points and cell types to corroborate these findings and investigate therapeutic targets that can disrupt crucial virus-host protein interactions, potentially improving antiviral defences.

Reference

Ghosh, A., Anantharaman, V., Melnick, D., & Aravind, L. (2020). Proteomics of SARS-CoV-2 infection. *Nature Communications, 11*(1), 4855. https://doi.org/10.1038/s41467-020-18533-3

Kong, A. T., Leprevost, F. V., Avtonomov, D. M., Mellacheruvu, D., & Nesvizhskii, A. I. (2017). MSFragger: Ultrafast and comprehensive peptide identification in mass spectrometry–based proteomics. *Nature Methods, 14*(5), 513-520. https://doi.org/10.1038/nmeth.4256

Strupat, K., Scheibner, O., & Bromirski, M. (n.d.). High-resolution, accurate-mass Orbitrap mass spectrometry – Definitions, opportunities, and advantages. Thermo Fisher Scientific (Bremen) GmbH.

De Las Rivas, J., & Fontanillo, C. (2010). Protein–Protein Interactions Essentials: Key Concepts to Building and Analyzing Interactome Networks. *PLoS Computational Biology, 6*(6), e1000807. https://doi.org/10.1371/journal.pcbi.1000807