Notes

PEA Visualizations – Abishek Nirupan Bhuvanaratnam Olink Proteomics

A chart of a virus

Description automatically generated with medium confidence

Figure 1: Initial Neutralization Responses to SARS-CoV-2 by COVID-19 Status

I wanted to replicate 3B labelled “Predictors of neutralization and its association with disease severity and age”, mostly due to my background in studying immune responses and working with clinical data. This box plot gives a clear comparison of the initial immune responses between COVID-19 negative and positive patients on Day 0 of observation. I believe that box plots are good for depicting the distribution and even the variability of the data across different categories! Apart from showing the median and IQR (showing the middle 50% of the data), which help look at the central tendency and spread, this box plot also shows outliers and extremes in the data.

I believe that the motivation behind creating this visualization for the researchers could be from the need to understand if there were any significant differences in how the patient’s immune system initially reacts to the virus, categorized by their COVID-19 status.

The COVID-19 positive group appeared to have a higher median neutralization level than the COVID-19 negative group. The box plot image you’ve provided showcases the SARS-CoV-2 spike pseudo virus neutralization levels categorized by COVID-19 status. Here’s an interpretation of the visualization: The COVID-19 positive group seemed to have a larger IQR, indicating more variability in neutralization levels among this group. Finally, looking at the scatter of points, which is also known as the jitter, the spread of points for the COVID-19 positive group is more towards the higher end of neutralization levels compared to the COVID-19 negative group.

- \*\*Two Categories\*\*: The data is split into two groups: "COVID-19 Negative" (shown in red) and "COVID-19 Positive" (shown in blue). Each group’s neutralization levels are depicted as a separate box plot.

- \*\*Central Tendency\*\*: For both groups, the median neutralization level (represented by the horizontal line within each box) is visible. The COVID-19 positive group appears to have a higher median neutralization level than the COVID-19 negative group.

- \*\*Spread and Variability\*\*: The boxes represent the interquartile range (IQR), showing the middle 50% of the data. The COVID-19 positive group seems to have a larger IQR, indicating more variability in neutralization levels among this group.

- \*\*Outliers\*\*: There are individual points plotted outside the upper and lower "whiskers" of the box plots, which typically represent 1.5 times the IQR from the upper and lower quartiles. These are considered outliers. The COVID-19 negative group appears to have several outliers below the lower whisker, whereas the COVID-19 positive group has outliers on both ends, suggesting a few individuals with exceptionally high or low neutralization levels.

- \*\*Overall Distribution\*\*: The scatter of points (jitter), representing individual neutralization levels, shows how the data is distributed across the range for each group. The spread of points for the COVID-19 positive group is more towards the higher end of neutralization levels compared to the COVID-19 negative group.

Based on this plot, one might infer that while there is a range of neutralization responses within each group, on average, COVID-19 positive individuals may have a higher neutralization level against the pseudovirus, indicating a more robust initial immune response detectable at Day 0. However, the significant variability, especially within the COVID-19 positive group, suggests a diverse immune response to the infection.

It's important to follow up on these findings with statistical testing to confirm whether the observed differences are significant. Additionally, the outliers warrant further investigation to understand the underlying reasons for their distinct neutralization levels, which could lead to insights into different immune response mechanisms or potential errors in data collection.

Outliers   
In a box plot, extremes and outliers provide important insights into the distribution and characteristics of the data. Here's what they can tell you about your data regarding SARS-CoV-2 spike pseudovirus neutralization levels:

1. \*\*Variability and Range\*\*:

- Outliers indicate that there are cases where the neutralization levels significantly differ from the majority. This can suggest high variability in how individuals respond to the virus at the initial stage.

- Extremes beyond the typical range can point to unusual immune responses, which might be particularly strong or particularly weak. Understanding these extremes could be crucial for identifying patterns or factors that contribute to severe or mild forms of the disease.

2. \*\*Population Diversity\*\*:

- The presence of outliers reflects the diversity in the biological responses among the population. This diversity could be due to genetic factors, pre-existing health conditions, differences in exposure levels, or previous immunity (such as from vaccines or other coronaviruses).

3. \*\*Potential Errors or Anomalies\*\*:

- Sometimes, outliers in a dataset could also indicate potential errors in data collection, processing, or anomalies in sample handling. It's important to review these outliers closely to ensure they represent true values and not mistakes.

4. \*\*Insights into Disease Mechanisms\*\*:

- Outliers might reveal critical insights into the disease mechanisms, particularly if these values can be correlated with other clinical data. For instance, unusually high or low neutralization levels might correlate with the severity of symptoms or the speed of disease progression.

5. \*\*Guidance for Further Investigation\*\*:

- Extremes and outliers can guide further investigations. For instance, examining the medical histories and other immunological markers of these outlier individuals might help uncover why their responses are different, leading to new hypotheses about immune responses or therapeutic targets.

By analyzing these outliers and extremes, researchers can gain deeper insights into the immune responses to SARS-CoV-2, potentially leading to more tailored treatment approaches or preventive measures.

Code

Certainly! Let's break down the rationale and the code you used for creating a boxplot of SARS-CoV-2 spike pseudovirus neutralization levels for COVID-19 negative and positive patients at Day 0. This visualization aims to illustrate how neutralization levels differ between these two groups at the initial time point of the study.

### Step-by-Step Explanation

#### Step 1: Loading the Data

```r

data3 <- read.csv("/Users/abishekbhuvanaratnam/Downloads/MGH\_Olink\_COVID\_Apr\_27\_2021/Percent Neutralization.csv")

```

- \*\*Rationale\*\*: The first step is to load the dataset containing neutralization percentages into R. This dataset is crucial for analyzing how patients' immune responses (measured as neutralization levels) differ based on their COVID-19 status at Day 0.

#### Step 2: Merging Data

```r

merge\_data <- merge(data3, data, by = "subject\_id")

```

- \*\*Rationale\*\*: This line merges the newly loaded `data3` with another dataset `data` based on `subject\_id`. The merge is essential to combine relevant clinical information or additional measurements that are stored separately but related by the `subject\_id`. This enriches the dataset by ensuring all pertinent information regarding each subject is in one place.

#### Step 3: Filtering Data for Day 0

```r

day0\_data <- filter(merge\_data, Day == 0)

```

- \*\*Rationale\*\*: Here, you filter the merged data to include only entries from Day 0. This focus on Day 0 is critical for assessing the initial neutralization response before any treatment or progression of the disease, providing a baseline measurement.

#### Step 4: Recoding COVID Status

```r

day0\_data$COVID <- factor(day0\_data$COVID, levels = c(0,1), labels = c("COVID-19-", "COVID-19+"))

```

- \*\*Rationale\*\*: This line converts the `COVID` variable from numeric (0 and 1) to a factor with more descriptive labels. This transformation enhances the readability of plots and analysis by replacing codes with meaningful descriptions: "COVID-19-" for negative and "COVID-19+" for positive.

#### Step 5: Creating the Boxplot

```r

ggplot(day0\_data, aes(x = COVID, y = Percent.Neutralization, fill = COVID)) +

geom\_boxplot() +

geom\_jitter(width = 0.2, aes(color = COVID), size = 1.5, alpha = 0.6) +

labs(x = "COVID-19 Status", y = "Neutralization Level (%)") +

scale\_fill\_manual(values = c("lightblue", "lightgreen"), labels = c("COVID-19 Negative", "COVID-19 Positive")) +

theme\_minimal() +

ggtitle("SARS-COV-2 Spike Pseudovirus Neutralization Levels")

```

- \*\*Rationale\*\*: This block uses `ggplot2` to create the visualization.

- `aes`: Defines aesthetics. `x = COVID` and `y = Percent.Neutralization` map COVID status to the x-axis and neutralization levels to the y-axis, respectively. `fill = COVID` assigns colors based on COVID status.

- `geom\_boxplot()`: Generates box plots, which summarize the distribution of neutralization levels for each group, displaying medians, quartiles, and potential outliers.

- `geom\_jitter()`: Adds a jitter plot over the boxplot to display individual data points, reducing overlap and aiding in visualizing the distribution density. The jitter's attributes (`width`, `color`, `size`, `alpha`) are set to enhance visibility and aesthetics.

- `labs`: Sets labels for axes and the plot title, enhancing clarity and providing context.

- `scale\_fill\_manual()`: Customizes the colors used in the plot, making it visually distinct between groups.

- `theme\_minimal()`: Applies a minimalistic theme to the plot to focus attention on the data.

- `ggtitle`: Sets the title of the plot, summarizing its content.

### Summary

This visualization aims to directly compare the initial immune response between COVID-19 negative and positive patients by visualizing the distribution of their neutralization levels. It highlights differences in how patients' immune systems are initially handling the virus, which could be vital for understanding the severity and progression of the disease. This plot can help identify if there is a significant difference in the initial immune response between these two groups, which could have implications for treatment and management strategies.

A graph showing a number of blue and red boxes

Description automatically generated

Figure 2: Comparative Box Plot of Percent Neutralization by Patient Severity at Different Days

This additional box plot gives a visual comparison of the percent neutralization between the patient severity groups, Non-Severe, and Severe, over the three days. I chose to use this box plot again because, it gives a great display of the central tendency, variability and the skewness of the data in a glance. In my opinion this box plot can be useful in the medical data context, where we can understand and see if there is in fact any difference between the variability between severe and non-severe patients.

As we can see in this plot

### Short Description of the Box Plot:

This box plot provides a visual comparison of the percent neutralization between two patient severity groups, "Non-Severe" and "Severe," over three different days of sample collection post-diagnosis (Day 0, Day 3, and Day 7). Each pair of box plots represents the interquartile range (IQR), median, and outliers for percent neutralization for each patient group at a given time point.

The choice to use a box plot for this visualization is driven by the plot's ability to succinctly display the central tendency, variability, and skewness of the data at a glance. Furthermore, it allows for an easy comparison between the different categories (in this case, "Severe" versus "Non-Severe" patients) across multiple time points. This is particularly useful for medical data, where understanding the distribution of a variable like percent neutralization can have diagnostic or prognostic implications.

### Extremes and Outliers:

The extremes (the "whiskers" of the box plot) extend to the highest and lowest values within 1.5 times the IQR from the upper and lower quartiles, respectively. Outliers are data points that lie beyond the extremes and are represented as individual points.

In this plot, the presence of outliers, especially in the "Severe" group, suggests that there is variability in percent neutralization that is not captured by the central 50% of the data. It could indicate that there are subpopulations within the "Severe" group that respond differently over time. The fact that outliers exist only for the "Severe" group could also imply that severe patients have a wider range of responses to the disease or treatment, indicating a less consistent or predictable course of illness.

### Explanation of the Result:

The result, depicted by the box plot, shows that there is a distribution of percent neutralization across both "Non-Severe" and "Severe" groups at all three time points. The median percent neutralization does not appear to vary significantly between the two severity groups; however, there is a visible difference in the range and outliers, especially in the "Severe" category. This could suggest that while the median response to treatment or disease progression is similar between severity groups, the variability in response is greater in severe cases. The plot also reveals that the distribution of responses changes over time, as evidenced by the differing box sizes and positions at each day of collection. This temporal aspect is key for understanding the disease progression and could be valuable for clinical decision-making.

Code

Sure, let's break down the code you've provided step by step, explaining each part along with the purpose of the functions and libraries used in the R script.

### Libraries

```r

library(tidyverse)

library(ggplot2)

```

These lines load the `tidyverse` and `ggplot2` libraries. `tidyverse` is a collection of R packages designed for data science that makes it easier to handle, process, and visualize data. `ggplot2` is specifically focused on data visualization and is also part of the `tidyverse` collection. Loading both ensures all necessary functions for data manipulation and plotting are available.

### Assumption Comment

```r

# Assuming you have already loaded your data into a dataframe called 'df'

# You might need to read the data if it's not already loaded:

# df <- read.csv("path\_to\_your\_dataset.csv")

```

This comment suggests that the dataset should be loaded into a dataframe called `df`. If it hasn't been loaded, it provides a comment on how to load it using `read.csv()`.

### Data Preprocessing

```r

processed\_data <- df %>%

# Filter for days of interest

filter(Day %in% c(0, 3, 7)) %>%

# Categorize the severity based on acuity levels

mutate(

Severity = case\_when(

Acuity\_0 %in% 1:2 | Acuity\_3 %in% 1:2 | Acuity\_7 %in% 1:2 ~ "Severe",

TRUE ~ "Non-Severe"

)

) %>%

# Filter for COVID positive patients

filter(COVID == 1)

```

This block of code uses the `dplyr` package from `tidyverse` for data manipulation:

1. `filter(Day %in% c(0, 3, 7))`: Filters rows where the `Day` column's values are 0, 3, or 7. This is used to focus the analysis on these specific days.

2. `mutate(...)` with `case\_when(...)`: Adds a new column `Severity`. This column categorizes patients as "Severe" if any of the `Acuity` measurements at days 0, 3, or 7 are in the range 1 to 2. If not, they are categorized as "Non-Severe".

3. Another `filter(COVID == 1)`: Further filters the data to include only COVID-positive patients.

### Plotting

```r

ggplot(processed\_data, aes(x = factor(Day), y = Percent.Neutralization, fill = Severity)) +

geom\_boxplot() +

scale\_fill\_brewer(palette = "Set1", name = "Patient Severity") +

labs(

x = "Day of Collection",

y = "Percent Neutralization",

title = "Neutralization Levels by Patient Severity Over Time"

) +

theme\_minimal() +

theme(legend.position = "right")

```

This part of the code uses `ggplot2` to create a box plot:

1. `ggplot(...)`: Initializes the plot with the processed data specifying aesthetics - the x-axis as the `Day` (converted to factor), y-axis as `Percent.Neutralization`, and box colors (`fill`) by `Severity`.

2. `geom\_boxplot()`: Adds box plots to visualize the distributions of percent neutralization by day and severity.

3. `scale\_fill\_brewer(...)`: Uses a color palette from Color Brewer for the fill scale, which provides aesthetically pleasing color schemes.

4. `labs(...)`: Labels the axes and provides a title for the plot.

5. `theme\_minimal()`: Uses a minimal theme for a cleaner look.

6. `theme(legend.position = "right")`: Positions the legend on the right side of the plot.

Overall, this script filters and processes the data to focus on specific conditions and then uses a box plot to visually compare the percent neutralization over time across different severity levels in COVID-positive patients.

A graph of different colored lines

Description automatically generated

Figure 3: Point-range plots of neutralization levels in non-severe and severe COVID-19+ patients over time. Color-coding by neutralization level at D3, grouped into 0%– 25%, 25%–50%, 50%–75%, and 75%–100%

Description:

The visualization presents a comparative analysis of COVID-19 neutralization percentages in patients, distinguishing between non-severe and severe cases over three different time points: Day 0, Day 3, and Day 7 post-infection. Each patient's trajectory is color-coded to represent the categorized neutralization level recorded on Day 3, ranging from 0%–25% (orange), 25%–50% (green), 50%–75% (blue), to 75%–100% (purple). The use of a line plot with points allows for the tracking of individual patient responses over time within the context of their illness severity. The facetted design segregates the data by severity status to facilitate a clear comparison.

Reasoning Behind Visualization:

This visualization was chosen to effectively demonstrate individual patient responses to COVID-19 over a specific timeline while emphasizing the variance in immune response between non-severe and severe cases. The color-coding by Day 3 neutralization levels provides immediate visual insight into the proportion of patients achieving various neutralization thresholds at an early point in their infection, which could be indicative of their subsequent immune response trajectory. Moreover, the visualization enables a quick assessment of patterns and outliers, such as whether severe cases are less likely to achieve high neutralization levels or if there's a distinct progression pattern between the severity categories. This detailed and categorized depiction can be instrumental in understanding the correlation between early immune response and disease outcome, informing therapeutic strategies and patient management.

The plot you've provided is a point-range plot that tracks the neutralization levels of COVID-19 patients over time, comparing those with non-severe and severe symptoms. It is a rich source of information, and here's a detailed breakdown of its elements and the insights one might infer:

### Plot Elements:

- \*\*X-Axis (Day of Collection):\*\* Represents three time points - Days 0, 3, and 7 post-infection, when the neutralization data were collected.

- \*\*Y-Axis (Percent Neutralization):\*\* Indicates the percentage of neutralization, presumably reflecting the ability of the patients' blood serum to neutralize the virus, which is a proxy for immune response.

- \*\*Colors:\*\* Each line is color-coded based on the Day 3 neutralization level, indicating the initial response to the infection.

- \*\*Facets:\*\* The plot is divided into two panels labeled "Non-Severe" and "Severe", showing the data for two distinct groups of patients based on the severity of their COVID-19 symptoms.

### Insights:

- \*\*Initial Response Variability:\*\* There's a wide range of initial neutralization responses on Day 3, as evidenced by the presence of all color categories in both severe and non-severe groups. This suggests variability in how different individuals' immune systems initially respond to COVID-19.

- \*\*Progression Over Time:\*\* Many patients, especially in the non-severe category, show an increasing trend in neutralization percentage from Day 0 to Day 7, suggesting an improving immune response over the first week of infection.

- \*\*Comparison by Severity:\*\* The severe group seems to have more patients with lower neutralization levels (orange and green lines) than the non-severe group, particularly noticeable on Day 3. This might imply a correlation between lower early neutralization levels and the severity of the disease.

- \*\*Day 7 Observations:\*\* By Day 7, there's a notable increase in higher neutralization percentages (blue and purple lines) in both groups, indicating that many patients' immune responses are ramping up by this point in their disease course.

- \*\*Individual Variations:\*\* The spread and overlap of lines demonstrate that individual patient responses vary widely, even within the same severity category, highlighting the complexity of the immune response to COVID-19.

- \*\*Potential Outliers:\*\* Some patients, particularly in the severe category, do not show the expected increase in neutralization levels over time, remaining in the lower categories or showing minimal change. These cases could be outliers or suggest a less effective immune response.

### Conclusion:

The plot is insightful in showing how neutralization levels change over the course of the first week post-infection and how these levels might relate to the severity of the disease. Such information can be valuable for clinicians and researchers as it might help predict the course of the disease based on early immune response or evaluate the effectiveness of treatments that aim to boost the body's ability to fight the virus.

Code

Certainly! Let's break down the R code you've shared step by step. This code involves using `tidyverse` and `ggplot2` libraries to manipulate a dataset and create a visualization.

### Libraries Used

1. `tidyverse`: A collection of R packages for data science, which includes dplyr for data manipulation, ggplot2 for plotting, and others.

2. `ggplot2`: A system for declaratively creating graphics, based on The Grammar of Graphics.

3. `scales`: Provides functions for adequately scaling the axes of plots.

### Code Explanation

#### Data Preparation

```r

merge\_data <- merge\_data %>%

mutate(

D3\_Neut\_Category = case\_when(

Day == 3 & Percent.Neutralization <= 0.25 ~ "0%–25%",

Day == 3 & Percent.Neutralization > 0.25 & Percent.Neutralization <= 0.50 ~ "25%–50%",

Day == 3 & Percent.Neutralization > 0.50 & Percent.Neutralization <= 0.75 ~ "50%–75%",

Day == 3 & Percent.Neutralization > 0.75 ~ "75%–100%",

TRUE ~ NA\_character\_ # Assign NA for other days to be filled forward

)

) %>%

group\_by(subject\_id) %>%

fill(D3\_Neut\_Category, .direction = "downup")

```

- \*\*Mutate with `case\_when`\*\*: The `mutate()` function is used to create a new column `D3\_Neut\_Category` in the `merge\_data` dataframe. `case\_when()` is used to assign categories based on the value of `Percent.Neutralization` on Day 3.

- \*\*Filling NAs\*\*: The `fill()` function fills in missing values (`NA`) in the `D3\_Neut\_Category` column. It uses data from the same subject (`group\_by(subject\_id)`) and fills missing data in both directions (`downup`), ensuring that all days for a subject have a category based on Day 3 values.

#### Data Filtering and Summarizing

```r

df\_covid\_positive <- merge\_data %>%

filter(COVID == 1, Day %in% c(0, 3, 7)) %>%

group\_by(subject\_id, Day, Severity, D3\_Neut\_Category) %>%

summarise(Percent.Neutralization = mean(Percent.Neutralization, na.rm = TRUE), .groups = "drop")

```

- \*\*Filtering\*\*: Keeps only data for subjects who tested positive for COVID (`COVID == 1`) and on Days 0, 3, and 7.

- \*\*Grouping and Summarising\*\*: Calculates the mean `Percent.Neutralization` for each `subject\_id`, `Day`, `Severity`, and `D3\_Neut\_Category`, removing NA values in the process.

#### Visualization with ggplot2

```r

custom\_colors <- c("0%–25%" = "orange", "25%–50%" = "green", "50%–75%" = "blue", "75%–100%" = "purple")

ggplot(df\_covid\_positive, aes(x = as.factor(Day), y = Percent.Neutralization,

group = subject\_id, color = D3\_Neut\_Category)) +

geom\_line() +

geom\_point() +

scale\_color\_manual(values = custom\_colors) +

facet\_wrap(~Severity) +

scale\_y\_continuous(labels = percent\_format()) +

labs(x = "Day of Collection", y = "Percent Neutralization (%)",

title = "Neutralization Over Time by Severity",

color = "D3 Neutralization Level") +

theme\_minimal() +

theme(legend.position = "bottom")

```

- \*\*Plotting\*\*: Uses `ggplot()` for plotting with line (`geom\_line()`) and point (`geom\_point()`) geometries.

- \*\*Coloring\*\*: Maps the `D3\_Neut\_Category` to colors defined in `custom\_colors`.

- \*\*Faceting\*\*: Creates separate plots for each `Severity` level using `facet\_wrap()`.

- \*\*Axes and Labels\*\*: Customizes y-axis labels to show percentages and adds labels and a title.

- \*\*Styling\*\*: Applies a minimal theme and positions the legend at the bottom.

This code efficiently categorizes the neutralization data based on Day 3 values, fills gaps for continuity in visualization, and then visualizes the trends of neutralization across different days and severities in a clear, color-coded manner.

A chart of a covid-19 status

Description automatically generated with medium confidence

The box plot you provided displays normalized protein expression (NPX) values for a set of proteins related to the viral response and interferon (IFN) pathway, differentiated by COVID status. Here are some insights and interpretations of the results shown:

1. \*\*Protein Expression Variability\*\*: The spread of the box plots (interquartile range) and the presence of outliers indicate variability in protein expression levels within the groups. Some proteins like IFNG, CXCL10, and CXCL11 show a broad range of expression levels across individuals, which could suggest a high variability in immune response to COVID.

2. \*\*Differential Expression\*\*: For proteins like IFNG and DDX58, the medians of the NPX values appear different between the 'COVID -' and 'COVID +' groups, suggesting that these proteins may be differentially expressed in individuals with COVID. This could imply a role for these proteins in the response to the virus or their potential as biomarkers for COVID infection.

3. \*\*Central Tendency\*\*: The median lines within the box plots provide a sense of the central tendency of the data. Proteins with medians that differ substantially between 'COVID -' and 'COVID +' could be key factors in the body's response to COVID.

4. \*\*Outliers\*\*: The presence of outliers (individual points that lie far from the rest of the data) might indicate exceptional responses in some individuals or could point to measurement errors or data entry issues. These outliers warrant further investigation to understand their causes.

5. \*\*Overall Trends\*\*: In several cases, such as with IFNG and DDX58, the 'COVID +' group appears to have higher median NPX values than the 'COVID -' group. This may suggest an upregulation of these proteins in response to the COVID infection.

6. \*\*Consistency Across Proteins\*\*: Some proteins do not show a marked difference between COVID positive and negative statuses, such as CCL7 and CCL24. This could mean that the expression of these proteins is not significantly affected by COVID status or that the sample size isn't large enough to detect a difference.

7. \*\*Facetted Design\*\*: The use of faceting to create individual plots for each protein allows for an easier comparison across proteins and between COVID statuses. This layout helps in identifying which specific proteins have the most pronounced differences in expression related to COVID.

In summary, the box plots suggest that certain proteins are differentially expressed in individuals with COVID compared to those without. These differences can provide insights into the biological processes and pathways activated in response to COVID infection and might be useful in understanding disease mechanisms or developing diagnostic tests. However, further statistical analysis would be needed to confirm these visual observations and understand their biological significance.

A graph showing different types of protein

Description automatically generated

This volcano plot is a visualization of protein expression data that compares the mean normalized protein expression (NPX) levels between patients with and without COVID-19. Each point represents a protein, with those that are significantly differentially expressed highlighted in blue and labeled accordingly. This approach is not only analytical but also serves a dual purpose for a molecular biology and immunology student with clinical experience like myself.

The motivation behind creating such a plot lies in the intersection of my scientific curiosity and the desire to convey complex data through visually engaging means. Having worked in various hospital settings, I understand the importance of translating molecular findings into clinical insights. This plot does just that by identifying potential biomarkers for COVID-19, which could be critical for diagnostics or therapeutic targets.

Moreover, as someone who appreciates the elegance of data visualization, the volcano plot offers an aesthetically pleasing approach to present significant biological changes in a clear and impactful way. It makes the data accessible to a broader audience, including those who might not specialize in molecular details, thereby fostering an interdisciplinary dialogue. This plot is a testament to the power of visual storytelling in science, where every dot can be a story of its own, capable of advancing our understanding of diseases like COVID-19.

From the volcano plot provided, several insights can be drawn:

1. \*\*Identification of Significant Proteins\*\*: The proteins labeled on the plot are those that have shown significant changes in expression levels between COVID-19 positive and negative patients. These proteins could potentially serve as biomarkers for COVID-19 infection or indicate the immune system's response to the virus.

2. \*\*Magnitude of Expression Changes\*\*: The horizontal axis (NPX Difference Day 0) indicates the magnitude of expression change. Points to the right or left indicate an increase or decrease in expression in COVID-19 positive patients, respectively. The further away from the center, the larger the magnitude of change.

3. \*\*Statistical Significance\*\*: The vertical axis (-Log10(p-values)) reflects the statistical significance of the expression changes. Higher points indicate more statistically significant changes. Proteins well above the dashed line, which represents a common significance threshold (p-value of 0.05), are considered highly significant.

4. \*\*Potential Targets for Further Study\*\*: The proteins that are both highly significant (high on the y-axis) and show a large magnitude of change (far from zero on the x-axis) could be of particular interest for further study. These may be involved in critical pathways affected by COVID-19.

5. \*\*Insight into Pathophysiology of COVID-19\*\*: By analyzing these proteins, researchers could gain a better understanding of the pathophysiological processes of COVID-19. For example, cytokines and chemokines that are significantly upregulated might suggest a heightened inflammatory response in COVID-19 positive patients.

6. \*\*Basis for Hypothesis Generation\*\*: This visual representation helps in formulating hypotheses regarding the molecular interactions and pathways that are active during the infection and could direct future experimental studies.

7. \*\*Visual Assessment of Data Quality and Distribution\*\*: The spread of points across the plot allows for an assessment of the overall data quality and variability. For example, a tightly clustered group of points might suggest low variability among the subjects' responses, while widely dispersed points could indicate the opposite.

8. \*\*Informing Clinical Decisions\*\*: For a clinical researcher or practitioner, understanding which proteins are markedly changed in expression can inform decisions about potential therapeutic targets or prognostic indicators in COVID-19 patients.

In summary, the volcano plot provides a snapshot of the differentially expressed proteins with potential biological relevance to COVID-19. It is a useful tool for identifying candidates for further research and possible clinical application.

```

- This snippet groups the data by assay type, calculates the mean NPX (Normalized Protein Expression) for COVID-positive and COVID-negative samples, computes the negative logarithm of the average p-value for the COVID-positive cases, and identifies whether each assay meets a significance threshold. A new column for log2 fold change is created by subtracting the mean NPX of COVID-negative from COVID-positive.

5. \*\*Plotting the data\*\*:

```r

ggplot(mean\_npx, aes(x = Log2FoldChange, y = Adjusted\_pval)) +

...

```

- Using `ggplot`, this code creates a volcano plot, which is a type of scatter plot commonly used to show changes in gene expression. Points are colored based on significance, and labels are added for target assays using `ggrepel`.

6. \*\*Create a filtered result table\*\*:

```r

result\_table1 <- mean\_npx %>%

filter(Assay %in% target\_assays) %>%

select(target\_assays = Assay, Significance\_level1 = Significant)

print(result\_table1)

```

- Filters `mean\_npx` to include only the target assays, selects relevant columns, and prints the table.

7. \*\*Export the results to a CSV file\*\*:

```r

write.csv(result\_table1, file = "result\_table1.csv", row.names = FALSE)

```

- The filtered result table is saved to a CSV file, which can be used for further analysis or reporting.

Overall, this script is used for analyzing and visualizing expression data from various biological assays, highlighting significant changes and exporting the results for specified target assays.

A graph of red and blue dots

Description automatically generated