**Identifying Gene Expression Biomarkers in Long-COVID Patients Through Transcriptomic Analysis (GSE270045)**

**Data**: The study uses whole-blood RNA-Seq data from GEO (GSE270045), including samples from 19 Long-COVID patients and 17 healthy controls (36 samples in total).

**Data Download**: Supplementary count files were downloaded via GEOquery, using the largest available file as the count matrix.

**Quality Control (QC)**: Genes with fewer than 10 counts in at least 25% of the samples were excluded to ensure reliable data.

**Analysis Model**: DESeq2 was employed to analyze differences between Long-COVID and healthy controls, with batch effects included when necessary. Surrogate effects were removed only for visualization purposes.

**Visualization**: Variance Stabilizing Transformation (VST) values were used for PCA and heatmaps. Differential expression was analyzed using raw counts with apeglm shrinkage.

**Differentially Expressed Genes (DEGs)**: Significant DEGs were defined as having an adjusted p-value < 0.05 and |log2FC| > 1.

**Volcano Plot**: A volcano plot was generated with stricter thresholds: adjusted p-value < 0.001 and |log2FC| > 2.

**Protein-Protein Interaction (PPI)**: STRING (human v11.5) was used to build a PPI network and identify key hub genes.

**Pathway Enrichment**: DEGs were mapped to ENTREZ IDs and analyzed for enrichment in KEGG, Reactome, and Gene Ontology (GO) terms.

**Outputs**: CSV tables and PNG figures were created for submission.

**Key Findings**: The analysis identified candidate biomarker genes and immune/neuronal pathways related to Long-COVID, creating a reproducible data structure for future studies.

**Introduction:**

Long-COVID is linked to ongoing immune and neurological disruptions, and identifying gene expression biomarkers could offer valuable insight into its mechanisms. RNA-Seq data from public databases can help identify differentially expressed genes (DEGs) and pathways that may serve as biomarkers for Long-COVID.

**Methods:**

**2.1 Data Acquisition**:  
R packages used: GEOquery, DESeq2, limma, ggplot2, STRINGdb, clusterProfiler, and others.  
Data was downloaded using GEOquery::getGEOSuppFiles and processed in the working directory.

**2.2 Metadata Construction**:  
Sample IDs were extracted from the count matrix, and group labels (HealthyControl vs LongCOVID) were assigned either from GEO metadata or manually.

**2.3 Filtering, Normalization, and PCA**:  
Genes with fewer than 10 counts in at least 25% of the samples were filtered out. VST values were used to perform PCA before and after batch effect removal to visualize any group separations.

**2.4 Differential Expression Analysis**:  
DESeq2 was used for differential expression analysis, and the data was shrunk using apeglm for improved interpretation of fold changes. Significant DEGs were defined by adjusted p-value < 0.05 and |log2FC| > 1.

**2.5 PPI Network and Hub Gene Identification**:  
The STRING database was used to build a PPI network, with hub genes identified based on their network degree. Cytoscape-compatible files were generated for further exploration.

**2.6 Pathway and Gene Ontology Enrichment**:  
Enrichment analyses were conducted using KEGG, Reactome, and GO terms for Molecular Function and Cellular Component, highlighting biological processes related to Long-COVID.

**2.7 Outputs**:  
Final results were exported as CSV tables and visualized in PNG format, ensuring reproducibility.

**Results:**

**3.1 Data Overview and QC**:After filtering, more than 5,000 genes remained for analysis. PCA showed initial separation by group and batch, with improved clustering after batch/SV removal.

**3.2 Differential Expression**:DESeq2, combined with apeglm shrinkage, revealed several DEGs between Long-COVID patients and healthy controls. A volcano plot highlighted the most significant genes (adjusted p-value < 0.001, |log2FC| > 2).

**3.3 PPI Network and Hub Genes**:STRING analysis identified key hub genes with high degrees in the PPI network. These hub genes were visualized with their log2FC values.

**3.4 Pathway and Enrichment**:KEGG, Reactome, and GO analyses revealed immune and neuronal pathways enriched in Long-COVID patients. The results pointed to key biological processes that are disrupted in Long-COVID.

**Discussion:**The combination of strong filtering, batch effect handling, and conservative DEG calling provides a robust method for identifying biomarkers in Long-COVID. The PPI network further highlights central genes that could play a role in the disease, while the pathway enrichment analysis emphasizes immune and neurological processes.

Hub genes identified in the PPI network are promising candidates for further biomarker studies, as they are both differentially expressed and central to the network's function.

**Limitations:**

* The GEO annotations may not always fully capture all batch and group information, requiring fallback strategies.
* Whole-blood RNA-Seq may dilute signals from tissue-specific genes.
* Accurate gene ID mapping is critical for successful enrichment analysis; mismatches can reduce the recovery of relevant terms

**Figure Analysis:**

1. **PCA (VST) before batch/SV removal:** Separation by group, shaped by batch.

![A graph with red and blue dots

AI-generated content may be incorrect.]()

**Interpretation:**

The PCA plot (before batch removal) provides a visual representation of how gene expression varies between Long-COVID and Healthy Control samples, based on the first two principal components (PCs).

1. **X-axis (PC1)**: This component explains 18.5% of the total variance in the data, showing the largest differences in gene expression.
2. **Y-axis (PC2)**: This component captures 14.8% of the total variance, adding another layer of variation between the samples.
3. **Batch**: Points are color-coded based on batch (black for Batch 1), suggesting that the samples from the same batch are clustering together, which indicates a potential batch effect influencing the results.
4. **Group**: The samples are also color-coded by group (red for Healthy Control and blue for Long-COVID). This color distinction highlights some separation between the two groups based on gene expression.

**Key Findings:**

1. **Group Separation**: There is some noticeable separation between Long-COVID and Healthy Control samples based on the first two principal components, though the separation is not perfect. A significant amount of the separation seems to be influenced by batch effects.
2. **Batch Effect**: The clustering of samples by batch (black points) suggests that batch effects are influencing the grouping, which could be confounding the biological differences between the two groups.
3. **Variation Within Groups**: Both the Long-COVID and Healthy Control groups show some variability within their respective clusters, indicating that gene expression varies within each group, which is common due to individual biological differences.

**Limitations:**

1. **Batch Effects**: The PCA plot shows that batch effects might be driving the separation of the groups, making it harder to distinguish between true biological differences between Long-COVID and Healthy Control samples.
2. **Limited Variance Representation**: The plot only shows the first two principal components, which together capture about 33.3% of the total variance. Other components may also contain important information that’s not visualized here.
3. **No Post-Batch Correction**: Since this PCA plot was generated before correcting for batch effects, it doesn’t fully represent how the groups might appear after these effects are accounted for.

This interpretation highlights the importance of addressing batch effects and considering additional components for a more complete understanding of the data.

1. **PCA after batch/SV removal**:

A graph of a group

AI-generated content may be incorrect.

**Interpretation:**

The PCA plot (after batch/SV removal) now shows the variance in gene expression between Long-COVID and Healthy Control groups, with batch effects and surrogate variables (SVs) removed. This adjustment helps provide a clearer understanding of the true biological differences between the groups.

1. **X-axis (PC1)**: The first principal component (PC1) now explains 22.3% of the total variance, showing the most significant differences between the groups.
2. **Y-axis (PC2)**: The second principal component (PC2) explains 6.5% of the variance, adding further differentiation between the samples.
3. **Group**: The samples are color-coded by group (red for Healthy Control, blue for Long-COVID). After batch/SV removal, the groups now show clearer separation along PC1, indicating more distinct differences between the two conditions.
4. **Batch/SV Removal**: The removal of batch effects and surrogate variables is evident in the improved separation between the groups, with less clustering based on batch and more on the group.

**Key Findings:**

1. **Clear Group Separation**: After addressing batch effects, there’s a much clearer separation between the Long-COVID and Healthy Control groups along PC1, suggesting that batch effects were previously masking true biological differences.
2. **Reduced Batch Influence**: Unlike the previous PCA plot, where batch effects were evident, this plot shows that the groups now primarily separate by condition (Long-COVID vs Healthy Control), with minimal batch-related clustering. This confirms the successful removal of batch and surrogate effects.
3. **Group Homogeneity**: Both groups (Long-COVID and Healthy Control) form tighter, more uniform clusters, showing reduced variability within each group after the correction.

**Limitations:**

1. **Variance Representation**: Even though batch effects have been reduced, the first two principal components only account for 28.8% of the total variance. Other components could contain additional important information that’s not visible in this plot.
2. **Post-Correction Heterogeneity**: Despite the clearer separation, some overlap between the two groups still exists, indicating that other sources of variance remain unaccounted for.
3. **No Full Biological Insight**: PCA is an unsupervised method and doesn’t explain the biological significance behind the observed variance. Further analyses are needed to validate and interpret these findings in terms of biological relevance.
4. **Volcano plot**: DEGs with adjusted p-value < 0.001 and |log2FC| > 2.

A graph of a graph

AI-generated content may be incorrect.

**Interpretation:**

The volcano plot provides a clear visual representation of the differential gene expression analysis results, helping to identify genes that show significant changes between Long-COVID patients and Healthy Controls. The plot uses the **log2 fold change** (x-axis) and the **adjusted p-value** (y-axis) to highlight which genes are most significantly altered.

* **X-axis (log2 Fold Change)**: This axis shows the log2 fold change between the two groups. Genes with a positive log2FC (to the right) are upregulated in Long-COVID, while genes with a negative log2FC (to the left) are downregulated.
* **Y-axis (-log10 adjusted p-value)**: This represents the negative logarithm of the adjusted p-value, where higher values indicate stronger statistical significance for gene expression differences.
* **Red Dots**: These represent the significant differentially expressed genes (DEGs) that meet the thresholds of **adjusted p-value < 0.001** and **|log2FC| > 2**. These genes have both large fold changes and high statistical significance.
* **Grey Dots**: These genes do not meet the significance thresholds and show either smaller changes or less statistical confidence, meaning they are not significantly different between Long-COVID and Healthy Controls.
* **Dotted Lines**: The horizontal and vertical dotted lines represent the significance thresholds for the adjusted p-value (p < 0.001) and log2FC (|log2FC| > 2). These lines help to clearly separate significant (red dots) from non-significant genes (grey dots).

**Key Findings:**

1. **Differential Expression**: A number of significant genes (red dots) show both large fold changes and low adjusted p-values, indicating clear, robust differences in gene expression between the Long-COVID and Healthy Control groups.
2. **Gene Distribution**: The majority of genes fall in the middle of the plot (grey dots), indicating that only a smaller subset of genes show significant changes. This suggests that while some genes are highly differentially expressed, most remain relatively unchanged.
3. **Thresholds**: The dashed lines provide clear cutoffs for significance, separating genes that show strong evidence of differential expression from those that do not.

**Limitations:**

1. **Threshold Choice**: The thresholds used for log2FC (> 2) and adjusted p-value (< 0.001) are somewhat arbitrary. Using different cutoff values may yield a different set of significant genes and impact the interpretation.
2. **No Multivariate Considerations**: The volcano plot only takes into account the fold change and adjusted p-value, which may overlook other factors or interactions that could influence gene expression. More complex models might be needed to account for other variables.
3. **Extreme Outliers**: Extreme outliers can affect the scale of both axes and potentially distort the appearance of the plot. This could lead to a misleading representation of the overall distribution of gene expression changes.
4. **Heatmap**: Top 10 up-regulated and down-regulated genes.

A screenshot of a graph

AI-generated content may be incorrect.

**Interpretation:**

The heatmap visualizes the gene expression of the top 10 upregulated and downregulated genes, scaled by row using Variance Stabilizing Transformation (VST), comparing Long-COVID patients and Healthy Controls.

* **X-axis (Samples)**: This axis represents individual samples, with colors indicating the sample's group (Healthy Control or Long-COVID) and batch (Batch 1).
* **Y-axis (Genes)**: The genes listed on this axis are the top 10 upregulated and downregulated genes based on differential expression.
* **Color Scale**: The colors represent gene expression levels, with **blue** indicating low expression, **yellow** representing moderate expression, and **red** signifying high expression.

**Key Findings:**

1. **Distinct Clustering**: The samples are grouped based on their gene expression patterns, with the Long-COVID and Healthy Control samples showing potential clustering according to gene expression. This suggests that gene expression patterns can differentiate the two groups.
2. **Gene Expression Differences**: Specific genes show clear differences in expression between the groups. **Upregulated genes** in Long-COVID (e.g., **JPH3**, **FEZF2**) are shown in **red**, indicating high expression in this group. Conversely, **downregulated genes** (e.g., **PRX**, **BICRA**) are shown in **blue**, indicating lower expression in Long-COVID compared to Healthy Controls.
3. **Batch Effects**: The batch information (Batch 1) is indicated at the top of the heatmap, and you can see that samples from the same batch are clustered together, suggesting some batch-related effects on the gene expression patterns.
4. **Group Differences**: The Long-COVID group exhibits distinct expression patterns compared to the Healthy Control group, as reflected in the color patterns for genes like **JPH3** and **FEZF2**. This indicates that gene expression in Long-COVID may be significantly different from that in Healthy Controls.

**Limitations:**

1. **Batch Effects**: Although batch information is shown, there may still be underlying batch effects influencing the clustering of samples. The presence of batch effects could mask true biological differences between the groups.
2. **Interpretation of Gene Expression**: While the genes are row-scaled for comparison, this scaling doesn’t capture the absolute magnitude of gene expression changes. Larger or smaller expression differences within each gene might not be fully represented.
3. **Clustering Artifacts**: The clustering of genes might be influenced by variability in gene expression within each group, which could obscure more subtle differences in gene expression between Long-COVID and Healthy Controls.
4. **STRING PPI hub genes**: Top 10 hub genes with their log2FC values.

A graph of red and blue squares

AI-generated content may be incorrect.

**Interpretation:**

This bar plot visualizes the **hub genes** identified from the Protein-Protein Interaction (PPI) network, with the bars representing the **log2 Fold Change (log2FC)** of these genes, which indicates the magnitude and direction of their expression changes between Long-COVID and Healthy Control groups.

* **X-axis (log2 Fold Change)**: This axis shows the log2 fold change of each hub gene, reflecting whether the gene is upregulated or downregulated in Long-COVID. Positive values indicate upregulation in Long-COVID, while negative values indicate downregulation.
* **Y-axis (Hub Genes)**: This axis lists the top hub genes, ranked by their **degree** (the number of interactions or connections they have in the PPI network). Higher degrees suggest these genes are central in the network and may have a key role in Long-COVID pathogenesis.
* **Bar Color**: The **red bars** represent genes that are **upregulated** in Long-COVID, while the **blue bars** represent genes that are **downregulated** in Long-COVID.

**Key Findings:**

1. **Upregulated Genes**: Genes such as **RPL17-C18orf32**, **RPS27**, and **RPL39** are upregulated in Long-COVID (indicated by red bars). These genes are highly connected in the PPI network, suggesting they may be central to the disease's biological mechanisms and could play significant roles in the pathogenesis of Long-COVID.
2. **Downregulated Genes**: Genes like **HBQ1** and **RPL29**, shown with blue bars, are downregulated in Long-COVID. These genes may be involved in immune or metabolic processes that are altered during the disease and could serve as biomarkers or therapeutic targets.
3. **Hub Genes**: The genes with the highest connectivity (degree) in the PPI network are labeled as **hub genes**. These genes are of particular interest because their central role in the network may indicate they have a pivotal function in the molecular pathways involved in Long-COVID.

**Limitations:**

1. **Biological Relevance**: Although these hub genes are highly connected in the PPI network, further experimental validation is needed to confirm their actual biological relevance to Long-COVID. High network connectivity doesn’t automatically translate to biological significance.
2. **Expression vs. Network Connectivity**: The plot focuses on gene expression (log2FC) and the connectivity of genes in the PPI network, but it does not account for other important factors such as **post-translational modifications** or environmental influences that could also affect gene function and contribute to Long-COVID.
3. **Potential False Positives**: Some hub genes may appear to be significant due to their high degree of connectivity in the PPI network, but they may not be biologically relevant. Further exploration and validation are needed to rule out false positives and better understand the actual role of these genes in Long-COVID.
4. **KEGG top 10 pathways**: Enriched pathways visualized in barplot and dotplot.

A graph of a number of patients with covid-19

AI-generated content may be incorrect.

A graph with red and blue dots

AI-generated content may be incorrect.

**DOTPLOT INTERPRETATION:**

The **dotplot** visualizes the top **KEGG pathways** enriched among the differentially expressed genes (DEGs), offering insight into their statistical significance and gene representation in the context of Long-COVID.

* **X-axis (Gene Ratio)**: Represents the GeneRatio, which is the proportion of DEGs involved in each pathway relative to the total number of genes tested. A higher GeneRatio means that a larger proportion of DEGs are associated with that pathway.
* **Y-axis (Pathways)**: Lists the top 10 KEGG pathways enriched in the data. Examples include pathways like **Salmonella infection**, **Endocytosis**, and **Coronavirus disease - COVID-19**.
* **Dot Size**: The size of the dots indicates the **count of genes** involved in each pathway. Larger dots correspond to pathways with more genes associated with them.
* **Color**: The color gradient represents the **adjusted p-value**. Red indicates more **statistically significant** pathways (lower p-value), and blue represents less significant pathways (higher p-value). Pathways such as **Salmonella infection** and **Endocytosis** are more significant, with p.adjust values around **0.004**.

**Key Findings:**

1. **Significant Pathways**: Pathways like **Salmonella infection**, **Endocytosis**, and **Shigellosis** are significantly enriched (with low p-values), suggesting they may have an important role in Long-COVID. These pathways may be central to immune or cellular processes disrupted in the disease.
2. **Pathways Related to COVID-19**: The pathway **Coronavirus disease - COVID-19** is enriched, emphasizing the relevance of COVID-19-related processes in Long-COVID, likely reflecting immune and viral response pathways still active in the disease.
3. **Gene Proportions**: Pathways such as **Lipid and atherosclerosis** and **Autophagy - animal** have higher **GeneRatios**, meaning a greater proportion of DEGs are involved in these pathways, indicating these processes could be more involved in the biological changes associated with Long-COVID.
4. **Gene Count**: Larger dots indicate pathways with a higher count of genes involved, such as **Salmonella infection** and **Endocytosis**, which may suggest broader biological relevance and a more extensive impact on cellular processes in Long-COVID.

**Limitations:**

1. **Limited Pathway Diversity**: The plot focuses on the **top 10 pathways** based on statistical significance, so potentially relevant pathways that don’t meet the threshold for inclusion may be overlooked.
2. **Biological Relevance**: While the pathways are statistically significant, their **biological relevance** to Long-COVID requires further validation through experimental studies. Some pathways, like **Apoptosis** and **Mitophagy**, might reflect underlying cellular processes, but more research is needed to determine their precise roles in the disease.
3. **Gene Ratio Interpretation**: Some pathways, like **Glycosylphosphatidylinositol (GPI)-anchor biosynthesis**, have a low **GeneRatio**, meaning only a small proportion of DEGs are involved. Despite being statistically significant, these pathways may have limited biological relevance due to the small number of DEGs involved.

**INTERPRETATION OF THE BARPLOT:**

The **barplot** also visualizes the **top KEGG pathways** enriched among DEGs, but this time it shows the **count of genes** involved in each pathway along with their statistical significance.

* **X-axis (Count)**: This axis represents the number of genes associated with each pathway. Larger bars indicate more genes are involved in that pathway.
* **Y-axis (Pathways)**: This axis lists the top 10 KEGG pathways, including pathways like **Legionellosis**, **Salmonella infection**, and **Autophagy**.
* **Color Gradient**: The **color** of the bars reflects the adjusted p-value, with red representing more **statistically significant** pathways (lower p-value) and blue indicating less significant pathways (higher p-value). The most significant pathways, like **Legionellosis** and **Lipid and atherosclerosis**, are shown in red.

**Key Findings:**

1. **Highly Significant Pathways**: Pathways like **Legionellosis**, **Salmonella infection**, and **Shigellosis** show strong enrichment with low p-values, suggesting a significant role in the differential expression of genes in Long-COVID.
2. **Gene Involvement**: Some pathways, like **Legionellosis** and **Endocytosis**, involve a large number of genes (over 40), suggesting these pathways have a broader impact on Long-COVID biology.
3. **Less Significant Pathways**: Some pathways, such as **Mitophagy - animal** and **Apoptosis**, are less enriched, with fewer genes and higher p-values. These pathways might still be involved in Long-COVID but have a smaller role compared to more significant pathways.

**Limitations:**

1. **Limited Pathway Exploration**: The plot focuses only on the **top 10 pathways** based on statistical significance, which may exclude other potentially relevant pathways that don’t meet the cutoff.
2. **Gene Ratio**: Unlike the dotplot, the barplot does not show the **GeneRatio** (the proportion of DEGs in each pathway), so it may overlook pathways with fewer DEGs but high relevance. Smaller counts do not necessarily mean smaller biological relevance.
3. **Interpretation of Significance**: While pathways with low p-values are statistically significant, their **biological relevance** to Long-COVID still requires confirmation. Further experimental studies are necessary to validate these pathways and their roles in the disease.

**Reactome top 10 pathways**: Barplot and dotplot showing the most enriched pathways.

**I tried my best, even though my semester final exam is in three days. But despite all my efforts, I couldn't manage to generate the GO and Reactome plots**

![A graph with a dot in the middle

AI-generated content may be incorrect.]()![A graph of a bar

AI-generated content may be incorrect.]()

1. **GO Molecular Function terms**: Top enriched terms visualized in barplot and dotplot.

A graph with a rectangle

AI-generated content may be incorrect.A graph with numbers and a dot

AI-generated content may be incorrect.

**GO Molecular Function: Barplot Interpretation**