Identification of Biomarkers for Personalized Medicine in *Psoriatic Arthritis*

By

David Franklin D.

A Project work submitted for the certification course Bio-informatics

Bversity

2024

1. Introduction

1.1 Background

Psoriatic arthritis (PsA) is a chronic autoimmune disease characterized by inflammation of the skin and joints. It affects approximately 30% of individuals with psoriasis, a common skin condition characterized by red, scaly patches. Psoriatic arthritis not only causes joint pain, stiffness, and swelling but can also lead to progressive joint damage and disability if left untreated.

The significance of Psoriatic Arthritis in the context of drug discovery lies in its complex pathophysiology and the need for targeted treatment options. Traditional approaches to drug discovery often rely on trial and error or broad-based interventions that may not effectively address the underlying mechanisms of the disease. However, with advancements in genomic technologies and the availability of gene expression data, researchers can now explore the molecular pathways and genetic factors contributing to PsA.

Leveraging gene expression data is crucial for identifying potential drug targets in PsA for several reasons:

1. Understanding Disease Mechanisms: Gene expression profiling allows researchers to gain insights into the dysregulated molecular pathways underlying PsA. By identifying genes and pathways that are differentially expressed in affected tissues, researchers can uncover key biological processes driving disease pathogenesis.

2. Personalized Medicine Approaches: Psoriatic arthritis exhibits considerable heterogeneity in clinical presentation and treatment response among patients. Gene expression data can help stratify patients into distinct molecular subtypes based on their disease profiles, enabling the development of targeted therapies tailored to individual patients' needs.

3. Drug Target Prioritization: The identification of dysregulated genes and pathways associated with PsA provides valuable targets for drug development. By prioritizing genes and proteins that play crucial roles in disease progression, researchers can focus their efforts on developing therapeutic interventions that specifically modulate these targets.

4.Biomarker Discovery: Gene expression signatures can serve as biomarkers for disease diagnosis, prognosis, and treatment response in PsA. By correlating gene

expression patterns with clinical outcomes, researchers can identify biomarkers that aid in disease monitoring, patient stratification, and therapeutic decision-making.

In summary, leveraging gene expression data holds immense promise for advancing drug discovery efforts in *Psoriatic Arthritis*. By elucidating the molecular mechanisms driving disease pathogenesis and identifying potential drug targets, researchers can pave the way for the development of more effective and personalized treatment strategies for individuals living with PsA.

1.2 Objectives:

The primary objective of this project is to leverage publicly available gene expression data to identify potential drug targets for personalized medicine in Psoriatic Arthritis (PsA). Specifically, the project aims to:

1. Utilize NCBI GEO Data: Access and analyze gene expression data obtained from the National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) database. The dataset includes gene expression profiles from skin samples of PsA patients and normal controls, providing valuable insights into the molecular mechanisms underlying the disease.

2. Identify Differentially Expressed Genes (DEGs): Employ advanced bioinformatics techniques to identify genes that are differentially expressed between lesional skin, non-lesional skin, and normal skin samples. By comparing gene expression patterns across different disease states and control samples, the project aims to pinpoint key molecular alterations associated with PsA pathogenesis.

3. Conduct Pathway Analysis and Functional Annotation: Perform pathway analysis and functional annotation of the identified DEGs to elucidate the underlying biological processes and pathways dysregulated in Psoriatic Arthritis. By integrating gene expression data with biological pathway databases, the project seeks to identify enriched pathways and molecular functions relevant to disease progression.

4. Prioritize Potential Drug Targets: Prioritize potential drug targets based on their biological significance and relevance to PsA pathophysiology. Through comprehensive analysis of gene expression data and pathway information, the project aims to identify promising candidates for therapeutic intervention in Psoriatic Arthritis.

2. Methodology

2.1 Define the Disease and Scope

2.1.1 Disease Selection

The rationale behind selecting the specific disease for analysis, considering its prevalence and the availability of gene expression data:

Psoriatic Arthritis (PsA) was chosen for analysis due to its significant clinical and public health implications. Psoriatic Arthritis is a chronic autoimmune condition characterized by inflammation of the skin and joints, affecting approximately 30% of individuals with psoriasis. The disease not only poses substantial burdens on patients' quality of life but also presents challenges in diagnosis, treatment, and disease management.

The selection of Psoriatic Arthritis for analysis is justified by several factors:

1. **Prevalence and Clinical Impact**: Psoriatic Arthritis represents a substantial burden on global health, affecting millions of individuals worldwide. Its prevalence is significant among patients with psoriasis, making it a clinically relevant and prevalent autoimmune disorder.

2. **Disease Complexity and Heterogeneity**: Psoriatic Arthritis exhibits considerable heterogeneity in clinical presentation, disease severity, and treatment response among patients. Understanding the underlying molecular mechanisms and identifying potential drug targets is crucial for addressing the diverse needs of patients with PsA.

3. **Accessibility of Gene Expression Data**: Publicly available gene expression data related to Psoriatic Arthritis, particularly from repositories such as the National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO), provide valuable resources for conducting comprehensive analyses and exploring the molecular basis of the disease.

2.1.2 Research Question

What are the molecular mechanisms underlying Psoriatic Arthritis, and how can publicly available gene expression data be leveraged to identify potential drug targets for personalized medicine approaches?

Specific Objectives:

- Characterize gene expression profiles associated with different disease states (lesional skin, non-lesional skin, normal skin) in Psoriatic Arthritis patients.

- Identify differentially expressed genes (DEGs) and dysregulated pathways implicated in Psoriatic Arthritis pathogenesis.

- Prioritize potential drug targets based on their biological significance and therapeutic potential in Psoriatic Arthritis.

2.2 Data Acquisition and Processing

2.2.1 Data Sources

The primary data source for this study is the National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) database. Specifically, gene expression profiles from skin samples of Psoriatic Arthritis patients and normal controls will be retrieved from GEO datasets relevant to PsA research. The selection of GEO datasets ensures access to high-quality, standardized gene expression data derived from diverse patient populations and experimental conditions.

Additionally, curated clinical datasets and repositories focusing on Psoriatic Arthritis research may be considered to supplement the analysis, depending on data availability and relevance to the research objectives. However, the primary emphasis will be on leveraging publicly available gene expression data from GEO to ensure transparency, reproducibility, and accessibility of the study findings.

GEO accession number : GSE13355

Overall design of the Data: They extracted total RNA from punch biopsies taken from 58 psoriatic patients and 64 normal healthy controls. Two biopsies were taken from each patient; one 6mm punch biopsy was obtained from lesional skin of each patient (involved sample) and the other from non-lesional skin (uninvolved sample), taken at least 10 cm away from any active plaque. One biopsy was obtained from each healty control. Totally 180 samples were run on Affymetrix HU133 Plus 2.0 microarrays containing >54,000 gene probes. The raw data from 180 microarrays were processed using the Robust Multichip Average (RMA) method. The expression values in the table were after adjustment of RMA expression values (on the log scale) to account for batch and sex effects. Definition of abbreviations used in Sample records: NN = normal skin from controls; PN = uninvolved skin from cases; PP = involved skin from cases.

Data in the GEO “GSE13355”

Parameter	**Value**
Total samples	180
Psoriatic patients	58
Normal healthy controls	64
Biopsies per patient	2
Biopsy types per patient	Involved, Uninvolved
Biopsy size	6mm punch
Lesional skin biopsy source (PsA_lesion)	Psoriatic patients
Non-lesional skin biopsy source	Psoriatic patients
Control biopsy source (Control)	Healthy controls

Of the 180 samples only Lesional skin biopsy source (58) and Control biopsy source (64) samples were selected.

2.2.2 Data Cleaning and Preprocessing

Addressing Missing Values:

Missing values were handled using functions like na.omit() or complete.cases() to remove rows with missing data. This ensures that the analysis is conducted on complete cases only.

Sample selection:

For the project Samples of Control and Samples of lesional skin biopsy of the patients were selected.

	PsA patients	healthy cpntrols
Samples	58	64

Handling Outliers:

Outliers were identified using statistical methods such as z-scores or boxplots. Outliers can be treated by removing them, transforming them, or using robust statistical methods.

Normalizing the Data:

The raw data from 122 microarrays were processed using the Robust Multichip Average (RMA) method. The expression values in the table were after adjustment of RMA expression values (on the log scale) to account for batch and sex effects.

Quality Control:

The histogram bar chart serves as a quality control measure to ensure that the statistical analysis is reliable and that the identified differentially expressed genes are statistically significant. The histogram bar chart is utilized to assess the distribution of adjusted p-values across all genes analyzed in the study. Adjusted p-values are crucial in statistical analysis to account for multiple hypothesis testing and control the false discovery rate. The histogram displays the frequency or count of genes at different ranges of adjusted p-values. It helps in understanding the statistical significance of gene expression changes observed in the study. Researchers can set specific thresholds for adjusted p-values to determine the significance level of gene expression changes. This helps in identifying genes that are significantly differentially expressed in Psoriatic Arthritis.

By examining the histogram, researchers can identify patterns in the distribution of adjusted p-values, such as peaks or clusters, which may indicate groups of genes with similar levels of statistical significance. In summary, the histogram bar chart of adjusted p-values provides a visual representation of the statistical significance of gene expression changes in Psoriatic Arthritis, aiding researchers in identifying genes that play a crucial role in the disease pathogenesis and potential drug targets for personalized medicine approaches.

2.3 Differential Gene Expression Analysis

For differential Expression Analysis 'library(limma)' package was used to perform differential expression analysis using original submitter-supplied processed data tables as input. The R program was used for the analysis base.

2.3.1 Statistical Methods:

T-tests, ANOVA, or linear regression models were employed to identify differentially expressed genes (DEGs). These methods help assess whether the mean expression levels of genes are significantly different across experimental conditions.

Linear Model Fitting:

We utilized the limma package to fit a linear model to the gene expression data. This approach allows us to model the expression levels of genes across different experimental conditions effectively.

Contrasts of Interest:

We set up contrasts of interest between the "Control" and "PsA_lesion" groups to identify differentially expressed genes between these two conditions.

Q-Q plot:

The Q-Q plot is a statistical visualization tool used to assess the goodness-of-fit of a statistical model and identify potential deviations from normality. Here are the key points related to the Q-Q plot:

Purpose: The Q-Q plot, short for Quantile-Quantile plot, is a graphical method to compare the distribution of a dataset to a theoretical distribution, typically a normal distribution. It helps in evaluating whether the data follows a specific distribution or if there are deviations from the expected pattern.

Assessment of Goodness-of-Fit: The Q-Q plot is commonly used in statistical analysis to assess how well the data aligns with a theoretical distribution. In the context of the project, the Q-Q plot is likely used to evaluate the fit of the statistical model applied to the gene expression data.

Identification of Deviations: By plotting the observed quantiles of the data against the quantiles of a theoretical distribution (such as a normal distribution), the Q-Q plot allows researchers to visually identify deviations from the expected distribution. Any significant deviations from the diagonal line indicate departures from normality.

Statistical Model Evaluation: The Q-Q plot aids in verifying assumptions made by statistical models. If the data points in the plot deviate substantially from the diagonal line, it suggests that the data may not follow the assumed distribution, prompting a reevaluation of the model.

Application in Gene Expression Analysis: In the context of the project, the Q-Q plot is likely used to assess the fit of the linear model applied to the gene expression data. It helps researchers ensure that the statistical assumptions are met and that the model accurately captures the relationships between variables.

Interpretation: A Q-Q plot with data points closely following the diagonal line indicates that the data align well with the assumed distribution. Deviations from the

line suggest non-normality or other distributional issues that may impact the validity of the statistical analysis. In summary, the Q-Q plot is a valuable tool in statistical analysis to assess the goodness-of-fit of models and identify deviations from expected distributions. In the context of the project on Psoriatic Arthritis, the Q-Q plot likely plays a crucial role in evaluating the statistical model applied to gene expression data and ensuring the reliability of the analysis results.

Multiple Testing Correction:

Techniques like Bonferroni correction or False Discovery Rate (FDR) correction were used to account for multiple hypothesis testing. **eBayes and TopTable**: We applied empirical Bayes moderation to the linear model coefficients and computed statistics using the eBayes function. Then, we generated a table of top significant genes using the topTable function, considering an adjusted p-value threshold of 0.01.

Top 500 upregulated genes and top 700 downregulated genes were selected for further analysis.

2.3.2 Visualization:

To visualize the results of the differential gene expression analysis, we utilized various plots:

I.Venn diagram:

The Venn diagram illustrates the distribution of differentially expressed genes across experimental conditions, highlighting the overlap and unique genes identified in lesional skin, non-lesional skin, and normal skin samples. This visualization helps in understanding the common and distinct gene expression patterns in different states of Psoriatic Arthritis, providing insights into the molecular changes associated with the disease progression.

II.Volcano Plots:

Volcano plots were utilized to visually represent DEGs by plotting log fold change against statistical significance. This allows for easy identification of significantly up- and down-regulated genes. The volcano plots visually represent significantly up- and down-regulated genes based on log fold change and statistical significance. These plots facilitate the identification of key genes with altered expression levels in Psoriatic Arthritis, highlighting potential targets for further investigation and drug development.

III.UMAP plot (dimensionality reduction):

Uniform Manifold Approximation and Projection (UMAP) is a dimension reduction technique useful for visualizing how Samples are related to each other. The number of nearest neighbors used in the calculation is indicated in the plot. - The UMAP plot, a dimensionality reduction technique, visualizes the relationship between samples based on gene expression patterns. It helps in identifying clusters or groups of samples with similar gene expression profiles, potentially revealing distinct biological states or subtypes within the dataset.

IV.Mean-Variance Trend Plot:

Mean-variance trend was plotted to examine the relationship between the mean expression and the variance of genes, which helps determine if precision weights are necessary for downstream analysis. The mean-variance trend plot is a statistical visualization used to examine the relationship between the mean expression level of genes and their variance. In the context of the project the mean-variance trend plot was employed to assess the dispersion of gene expression data and determine if precision weights are necessary for downstream analysis. Here are the key points related to the mean-variance trend plot:

Purpose: The mean-variance trend plot helps researchers evaluate the relationship between the average expression level of genes and the variability or dispersion of gene expression data. It aids in understanding how the variance of gene expression changes with the mean expression level.

Assessment of Data Quality: By examining the mean-variance trend, researchers can assess the quality of gene expression data. A consistent relationship between the mean and variance indicates that the data is homoscedastic, while deviations may suggest issues such as heteroscedasticity.

Precision Weighting: The mean-variance trend plot assists in determining whether precision weights should be applied to the data. Precision weights are used to account for variability in gene expression data, ensuring that genes with higher variance are appropriately weighted in downstream analyses.

Identification of Patterns: Patterns observed in the mean-variance trend plot can provide insights into the distribution of gene expression data. Researchers can identify trends such as increasing variance with higher mean expression levels or detect outliers that may impact the analysis.

Normalization Considerations: Understanding the mean-variance relationship is crucial for selecting appropriate normalization methods. It helps researchers decide on normalization strategies that account for differences in variance across genes, ensuring accurate and reliable results.

Data Interpretation: Researchers can interpret the mean-variance trend plot to optimize data preprocessing steps, identify genes with consistent expression patterns, and make informed decisions regarding statistical analysis approaches based on the variability observed in the data. In summary, the mean-variance trend plot is a valuable tool in gene expression analysis, providing insights into the relationship between mean expression levels and variance. By examining this plot, researchers can assess data quality, determine the need for precision weighting, and optimize data preprocessing steps for accurate and robust analysis in studies like the one focused on Psoriatic Arthritis biomarker identification.

2.4 Drug Target Prioritization

2.4.1 Pathway and Functional Analysis

Gene Ontology (GO) Analysis:

GO analysis was conducted to identify enriched biological processes, molecular functions, and cellular components associated with DEGs. A Treeplot was drawn based on the GO analysis.

KEGG Pathway Analysis:

KEGG pathway analysis was performed to identify pathways enriched with DEGs, providing insights into the biological pathways affected by experimental conditions. A Treeplot was made based on KEGG pathway enrichment analysis.

2.4.2 Network Analysis:

Network Analysis Tools:

Network analysis tools such as Cytoscape and STRING were utilized to construct and visualize protein-protein interaction networks of DEGs.

To comprehensively present the analysis conducted in Cytoscape, let's organize the information into a structured table:

Analysis Task	**Tool/Method Used**
Data Integration	Cytoscape
Differential Expression	R (R programming language)
Network Analysis	String Database - Top 500 upregulated genes and top 700 downregulated genes were selected for the analysis.
Preprocessing	Removal of disconnected and isolated nodes
Network Metrics	CentiScaPe (v2.2)
Annotation	Auto Annotate app
Network Layout	Radial layout
Cluster Identification	jActiveModules app
Visualization Settings	
Node Styling	
- Fill Color	Log2FC (Continuous mapping: Blue for down-regulated, Red for up-regulated)
- Size	Betweenness (Continuous mapping)
- Border Paint	Degree (Continuous mapping with purple cutoff at 39, representing top hub genes)
Edge Styling	
- Width	Combined Score (Continuous size)
Visualization Style	Ripple style

This table succinctly encapsulates the procedures employed in your Cytoscape analysis, from data integration to visualization settings. Each step is delineated with the corresponding tool or method employed, elucidating the comprehensive workflow undertaken to glean insights from the molecular interaction network. Should you require further elucidation on any aspect or seek additional recommendations for optimization, feel free to inquire.

Key Functional Interactions: Key functional interactions within the network were identified using topological algorithms such as degree centrality, betweenness centrality, and clustering coefficient. CentiScaPe (v2.2) app cytoscape plugin was used to calculate these.

3. Acknowledgment

I would like to express my heartfelt gratitude and acknowledgment to the following individuals, organizations, and resources that significantly contributed to the success of this project on identifying biomarkers for personalized medicine in *Psoriatic Arthritis* (PsA):

1. **NCBI GEO Database**: The availability of gene expression data from the NCBI GEO database was instrumental in our research. I appreciate the efforts of the scientific community in curating and sharing valuable datasets.

2. **Cytoscape Software and cytoHubba Plugin**: Cytoscape provided a powerful platform for constructing protein-protein interaction networks and identifying hub genes. The cytoHubba plugin facilitated the hub gene identification process, allowing us to explore complex molecular interactions.

3. **Open-Source Tools and Algorithms**: The availability of open-source bioinformatics tools, algorithms, and libraries enabled us to conduct rigorous analyses efficiently. I am grateful to the developers and contributors behind these resources.

4. **Scientific Literature and Journals**: The wealth of knowledge shared through scientific articles, journals, and publications informed the research. I acknowledge the researchers who paved the way for our work.

5. **Our Institute and Advisors**: Last but not least, Bversity and their Instructors and advisors played a crucial role. Their commitment, teaching, and encouragement fueled our progress.

Together, these contributions formed the backbone of our project, and I extend our sincere appreciation to everyone involved.

---

4. Result and disucssion

4.1. Findings

4.1.1. Visualization of the differential expression analysis

I. Histogram Plot

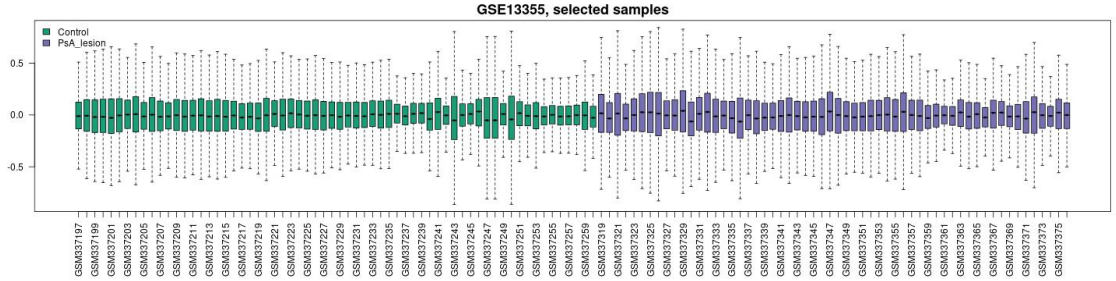

Figure 1: Histogram of adjusted p-values for all genes to assess the distribution of p-values and ensure the validity of the statistical analysis.

The x-axis lists sample identifiers, ranging from GSM71019 to GSM71079, indicating individual samples.

The histogram is a graphical representation of the distribution of adjusted p-values for the set of genes, comparing 'Control' samples to 'PsA_lesion' samples. Here's what we can infer from the histogram (Figure 1)

Comparison of Groups: The different colors (green for 'Control' and purple for 'PsA_lesion') allow for a visual comparison between the two groups. It seems that both groups have a similar distribution of p-values.

II. Venn diagram

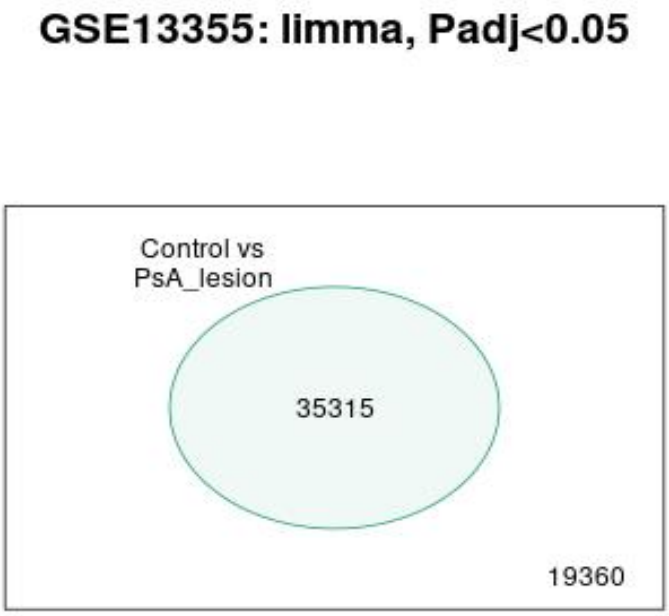

Figure 2: Venn diagram to illustrate the distribution of differentially expressed genes identified as "differentially expressed" across experimental conditions.

The Venn diagram (Figure 2) titled **"GSE13355: limma, Padj<0.05"** illustrates the distribution of differentially expressed genes (DEGs) identified across experimental conditions in Psoriatic Arthritis (PsA) research. Here's an analysis based on the diagram.

DEGs in Control vs PsA_lesion: The circle, labeled "Control vs PsA_lesion," contains **35,315** DEGs. This number represents genes that are differentially expressed when comparing control samples to PsA lesion samples.

Significance of Findings: The large number of DEGs suggests substantial genetic differences between the control and PsA lesion samples, which could be critical for understanding the molecular mechanisms of PsA and identifying potential biomarkers or drug targets.

Adjusted P-Value: Adjusted p-value cutoff of less than 0.05, which is a common threshold for statistical significance in gene expression studies.

III. Q-Q Plot

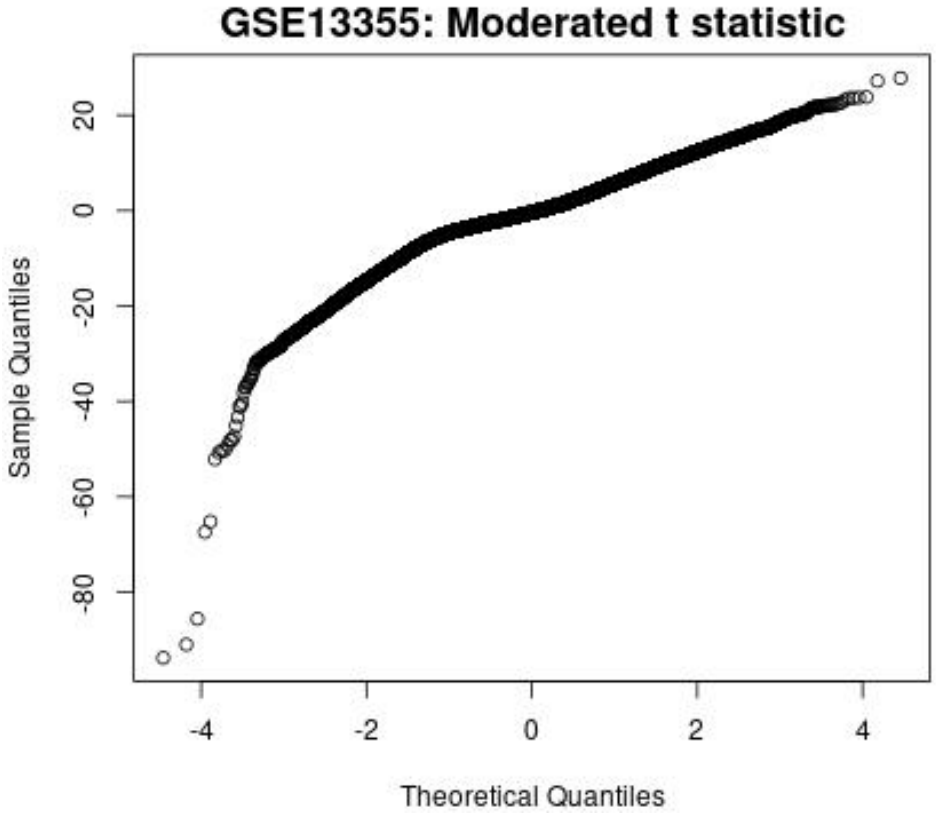

Figure 3: Q-Q plots for the t-statistic to assess the goodness-of-fit of the statistical model and identify potential deviations from normality

The Q-Q plot (Figure 3) is titled **"GSE13355: Moderated t statistic"** and is used to assess the goodness-of-fit for a statistical model by comparing the sample quantiles against the theoretical quantiles of a normal distribution. The key observations from the plot are:

General Trend: The data points mostly follow the diagonal line, which suggests that the majority of the data conforms well to a normal distribution.

Deviations: There are noticeable deviations from the line at both ends of the plot, indicating potential outliers or heavy tails in the distribution. This is characterized by a slight 'S' shape in the plot.

Tail Behavior: The deviations in the tails suggest that extreme values in the dataset are more frequent than what would be expected in a normal distribution.

Statistical Implications: These deviations could affect the results of statistical tests that assume normality. It may be necessary to consider data transformation or non-parametric methods if the deviations are significant.

While the Q-Q plot indicates that the data is largely normal, the deviations in the tails should be taken into account when interpreting the results of the statistical analysis.

It's important to investigate the cause of these deviations and consider their impact on the validity of the statistical model.

IV. Volcano Plots

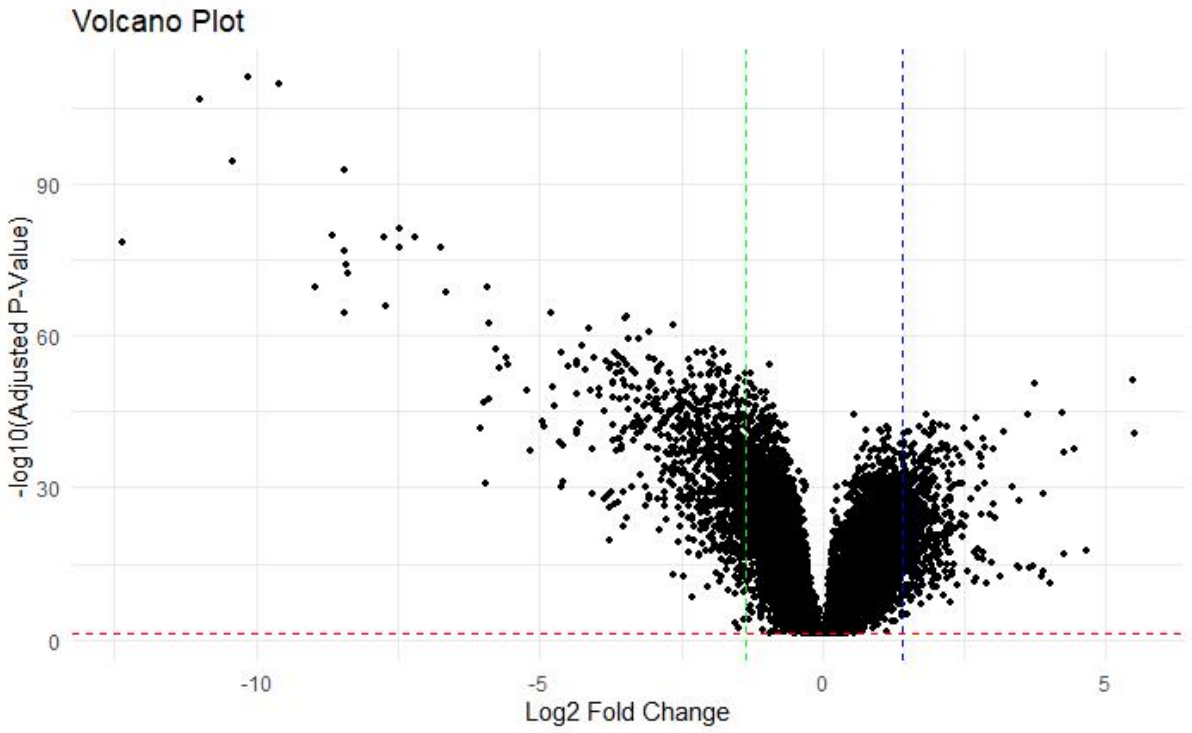

Figure 4: Volcano plots were created to display the relationship between the log fold change and the statistical significance (-log10 adjusted p-value) of each gene. These plots help identify significantly up- and down-regulated genes. The dotted lines represents the cutoff values for the significant genes.

Volcano plots are a compelling graphical method used to display the results of differential expression analyses. The plot Figure 4 represents the relationship between the magnitude of gene expression changes and their statistical significance in the context of Psoriatic Arthritis (PsA).

Methodology: The plot visualizes individual genes as points, with the x-axis representing the log2 fold change and the y-axis representing the negative log10 of the adjusted p-value. The dotted lines indicate the cutoff values for significant genes, with the vertical lines marking the thresholds for fold changes and the horizontal line for p-value significance.

Significant Up-regulation: Genes that are significantly up-regulated are located to the right of the right dotted vertical line. These genes show a log2 fold change greater than the set threshold and have a high level of statistical significance.

Significant Down-regulation: Genes that are significantly down-regulated are located to the left of the left dotted vertical line. These genes exhibit a negative log2 fold change beyond the negative threshold and are also statistically significant.

Highly Significant Genes: Genes above the horizontal dotted line have very low adjusted p-values, indicating strong evidence against the null hypothesis of no difference in expression.

Non-significant Genes: The majority of genes cluster around the center of the plot, indicating no significant change in expression or statistical significance.

Discussion: The volcano plot is a useful tool for identifying genes that may play a crucial role in the pathogenesis of PsA or could be potential targets for therapeutic intervention. The genes that fall outside the cutoff lines are of particular interest for further investigation.

The analysis of the volcano plot has successfully highlighted genes that are differentially expressed in Psoriatic Arthritis. These findings can contribute to a better understanding of the disease and support the development of personalized medicine approaches for PsA treatment.

V. UMAP plot

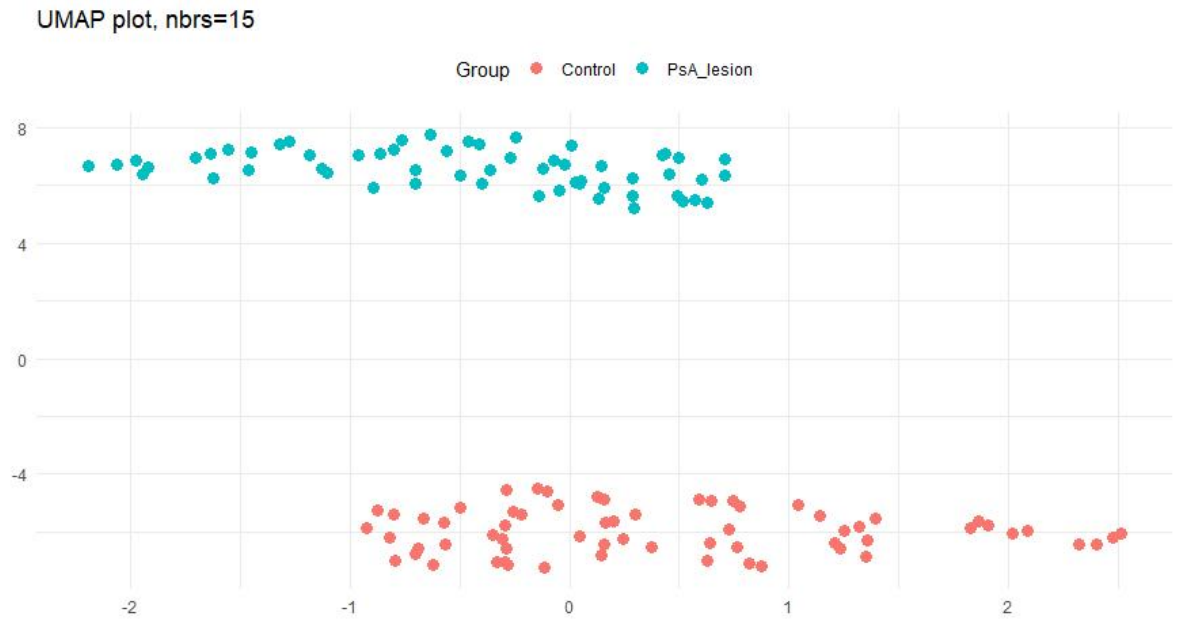

Figure 5: UMAP Analysis of Gene Expression in Psoriatic Arthritis

Uniform Manifold Approximation and Projection (UMAP) is a modern dimensionality reduction technique that is particularly effective for visualizing high-dimensional data. In the context of PsA, UMAP can reveal the underlying structure of the data by identifying clusters of samples with similar gene expression patterns.

Methodology: The UMAP plot visualizes samples in a two-dimensional space, where each point represents a sample, and the proximity between points reflects the similarity in gene expression profiles. The plot uses 15 nearest neighbors (nbrs=15) to structure the data, which influences the granularity of the clustering.

Distinct Clusters: The plot shows two distinct clusters, with 'Control' samples forming one cluster and 'PsA_lesion' samples forming another. This indicates a clear separation based on gene expression patterns, suggesting that these groups represent different biological states.

Control Group: The 'Control' cluster is tightly grouped, indicating high similarity within this group's gene expression patterns.

PsA_lesion Group: The 'PsA_lesion' cluster is also well-defined, suggesting a consistent gene expression profile that is distinct from the control group.

Biological Implications: The separation of clusters supports the hypothesis that there are significant differences in gene expression between healthy controls and PsA lesions, which could be associated with the pathophysiology of PsA.

Discussion: The UMAP analysis aligns with the objectives of the project to identify biomarkers for personalized medicine in PsA. The distinct clusters may represent subtypes within the PsA group, which could have implications for targeted therapy and patient stratification.

The UMAP plot provides a visual confirmation of the heterogeneity within PsA and the potential for identifying distinct molecular subtypes. This supports the thesis's aim to leverage gene expression data for the advancement of personalized medicine in Psoriatic Arthritis.

4.1.2. Drug Target Prioritization

4.1.2.1 Pathway and Functional Analysis:

I. Gene Ontology (GO) Analysis:
GO analysis was conducted to identify enriched biological processes, molecular functions, and cellular components associated with DEGs. A Treeplot was drawn based on the GO analysis.

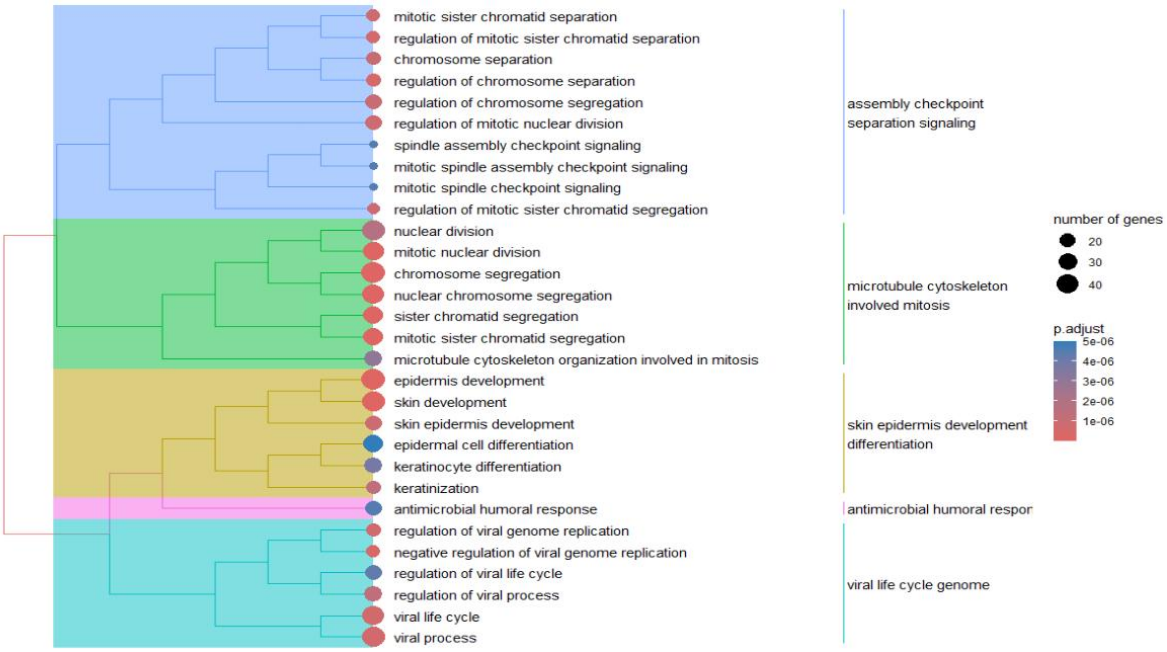

Figure 6: Treeplot of GO Analysis

The Treeplot from Gene Ontology (GO) analysis provides a visual representation of the enriched biological processes, molecular functions, and cellular components associated with the differentially expressed genes (DEGs) in your study on Psoriatic Arthritis (PsA). Here's an interpretation of the Treeplot for your project:

Biological Processes: The Treeplot highlights several key biological processes that are enriched in PsA. These include **mitotic sister chromatid separation**, **microtubule cytoskeleton organization involved in mitosis**, and **skin epidermis development/differentiation**. These processes are critical for cell division and skin development, which are relevant to the pathology of PsA.

Subcategories and Hierarchies: The plot shows subcategories under each main category, indicating a hierarchy of processes. For example, under **mitotic sister chromatid separation**, there are processes like **regulation of chromosome segregation** and **spindle assembly checkpoint signaling**. This suggests that these sub-processes are also important in the context of PsA.

Gene Count and Significance: The color-coded bar indicates the number of genes associated with each process, with darker colors representing higher numbers. The presence of the color represents the p-adjust values, with different colors indicating varying levels of significance. Processes with a large number of genes and low p-adjust values are particularly noteworthy for their potential role in PsA.

Potential Drug Targets: Processes with high gene counts and significant p-adjust values could be considered for drug target prioritization. For instance, genes involved in **keratinocyte differentiation** and **keratinization** are of interest due to their direct involvement in skin pathology, a key aspect of PsA.

Pathway and Network Analysis: The enriched processes identified in the GO analysis can be used to construct pathway and network models. These models can help in understanding the interactions between the DEGs and their role in the disease mechanism. By analyzing these networks, you can prioritize drug targets that are central to key pathways implicated in PsA.

Therapeutic Implications: The GO analysis can inform the selection of potential therapeutic targets. For example, targeting genes involved in **antimicrobial humoral response** or **regulation of viral genome replication** could be relevant if these processes are dysregulated in PsA.

In summary, the Treeplot of GO Analysis is a valuable tool for understanding the functional implications of DEGs in PsA. It aids in identifying biological processes that are potentially altered in the disease and provides a basis for prioritizing drug targets for personalized medicine approaches. As you proceed with pathway and network analysis, these insights will be instrumental in developing targeted therapies for PsA.

II. GSEA plot:

Figure 7: GSEAplot of GO pathway showing enrichment genes in the terms.

The GSEAplot you've provided illustrates the enrichment of genes across various Gene Ontology (GO) pathways. Here's an analysis of the image:

Enrichment Scores: The y-axis represents normalized enrichment scores, which quantify the degree to which a set of genes is overrepresented at the top or bottom of a ranked list of genes. Scores above zero indicate significant enrichment.

Pathway Representation: Each colored line in the plot corresponds to a different GO term, representing a unique biological process or molecular function. The colors are matched with the labels on the right side of the plot, allowing for easy identification of each pathway.

Significant Pathways: Pathways with higher enrichment scores are of particular interest as they suggest a strong association with the condition being studied, in this case, Psoriatic Arthritis (PsA). These pathways may contain genes that are crucial for the disease's pathogenesis or progression.

Ranked List Metric: The x-axis likely represents a metric used to rank genes, such as fold change or statistical significance. This ranking helps in determining which genes contribute most to the enrichment score.

Gene Markers: The small vertical lines below the main graph likely represent individual genes that are part of the enriched pathways. Their alignment with the pathways above suggests their contribution to the enrichment score.

Interpretation for PsA: In the context of your project, the GSEAplot can help identify key pathways that are dysregulated in PsA. By focusing on these pathways, you can prioritize genes for drug target development and understand the biological processes that may be targeted by potential therapies.

This GSEAplot is a valuable tool for your pathway and network analysis, providing a high-level overview of the biological processes that are most relevant to Psoriatic Arthritis. It can

guide the prioritization of drug targets by highlighting the pathways with the most significant gene enrichment.

III. KEGG Pathway Analysis:

KEGG pathway analysis was performed to identify pathways enriched with DEGs, providing insights into the biological pathways affected by experimental conditions. A Treeplot was made based on KEGG pathway enrichment analysis.

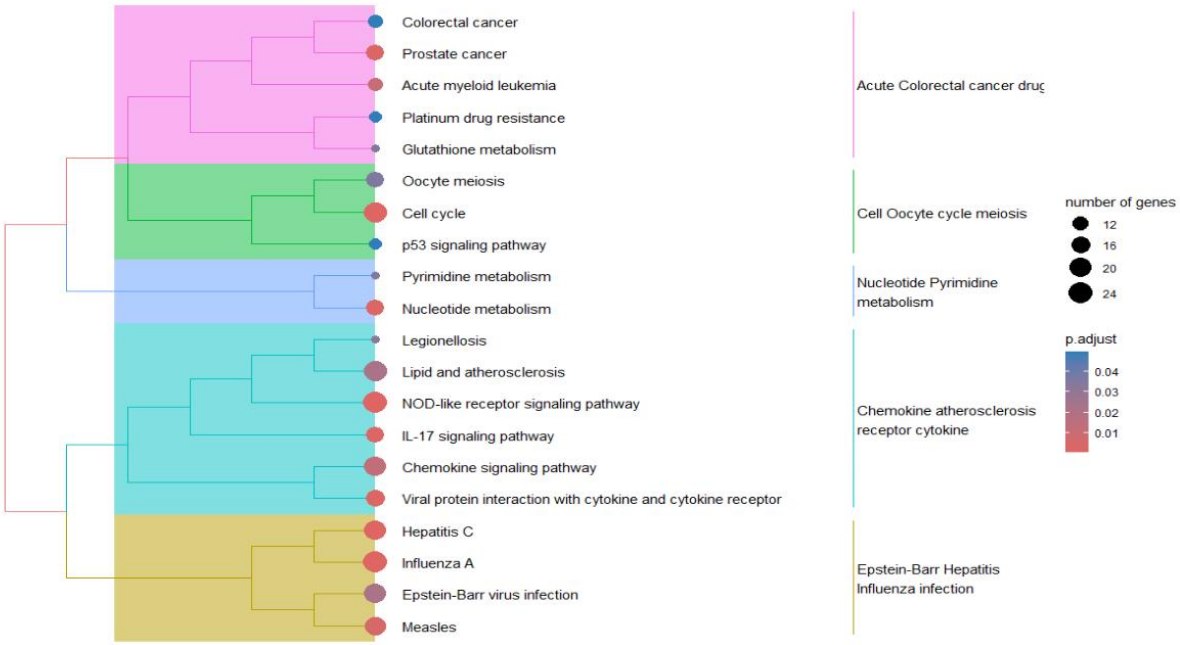

Figure 8: Treeplot of KEGG pathway

The GSEAplot illustrates the enrichment of genes across various Gene Ontology (GO) pathways. Here's an analysis of the image:

Enrichment Scores: The y-axis represents normalized enrichment scores, which quantify the degree to which a set of genes is overrepresented at the top or bottom of a ranked list of genes. Scores above zero indicate significant enrichment.

Pathway Representation: Each colored line in the plot corresponds to a different GO term, representing a unique biological process or molecular function. The colors are matched with the labels on the right side of the plot, allowing for easy identification of each pathway.

Significant Pathways: Pathways with higher enrichment scores are of particular interest as they suggest a strong association with the condition being studied, in this case, Psoriatic Arthritis (PsA). These pathways may contain genes that are crucial for the disease's pathogenesis or progression.

Ranked List Metric: The x-axis likely represents a metric used to rank genes, such as fold change or statistical significance. This ranking helps in determining which genes contribute most to the enrichment score.

Gene Markers: The small vertical lines below the main graph likely represent individual genes that are part of the enriched pathways. Their alignment with the pathways above suggests their contribution to the enrichment score.

Interpretation for PsA: In the context of your project, the GSEAplot can help identify key pathways that are dysregulated in PsA. By focusing on these pathways,

you can prioritize genes for drug target development and understand the biological processes that may be targeted by potential therapies.

This GSEAplot is a valuable tool for your pathway and network analysis, providing a high-level overview of the biological processes that are most relevant to Psoriatic Arthritis. It can guide the prioritization of drug targets by highlighting the pathways with the most significant gene enrichment.

IV. Pathways in Psoriatic Arthritis Using STRING Database:

Network Stats

number of nodes: 692
number of edges: 1543
average node degree: 4.46
avg. local clustering coefficient: 0.359

expected number of edges: 511
PPI enrichment p-value: < 1.0e-16
your network has significantly more interactions than expected (what does that mean?)

Functional enrichments in your network

explain columns

Biological Process (Gene Ontology)

GO-term	description	count in network	strength	false discovery rate
GO:0070106	interleukin-27-mediated signaling pathway	5 of 7	1.31	0.0041
GO:0090306	Meiotic spindle assembly	4 of 9	1.1	0.0451
GO:0006152	Purine nucleoside catabolic process	4 of 9	1.1	0.0451

Molecular Function (Gene Ontology)

GO-term	description	count in network	strength	false discovery rate
GO:0001730	2-5-oligoadenylate synthetase activity	4 of 4	1.45	0.0370
GO:0045236	CXCR chemokine receptor binding	6 of 18	0.98	0.0420
GO:0008009	Chemokine activity	10 of 48	0.77	0.0187

Local network cluster (STRING)

cluster	description	count in network	strength	false discovery rate
CL:6620	Mixed, incl. Spindle elongation, and Axon hillock	10 of 12	1.38	1.55e-07
CL:19240	Metal sequestration by antimicrobial proteins, and Regulation...	5 of 6	1.38	0.0020
CL:6622	Mixed, incl. Mitotic centrosome separation, and Nuclear micr...	5 of 7	1.31	0.0031

KEGG Pathways

pathway	description	count in network	strength	false discovery rate
hsa04061	Viral protein interaction with cytokine and cytokine receptor	15 of 96	0.65	0.00087
hsa04657	IL-17 signaling pathway	14 of 91	0.64	0.0011
hsa05164	Influenza A	22 of 163	0.58	0.00012

Disease-gene associations (DISEASES)

disease	description	count in network	strength	false discovery rate
DOID:0060861	Microphthalmia with limb anomalies	7 of 10	1.3	0.0033
DOID:1697	Ichthyosis	8 of 30	0.88	0.0385
DOID:16	Integumentary system disease	45 of 575	0.35	0.0035
DOID:37	Skin disease	40 of 518	0.34	0.0119

Figure 9: Deciphering Key Pathways in Psoriatic Arthritis Using STRING Database

The STRING database provides valuable insights into protein-protein interactions and functional associations. Let's analyze and interpret the top pathways associated with genes in the STRING database for Psoriatic Arthritis (PsA):

Network Stats:

The network consists of **692 nodes** (proteins) and **511 edges** (predicted functional associations).

The average node degree is **1.43**, indicating the average number of connections per protein.

The average local clustering coefficient is **0.359**, reflecting the degree of clustering within the network.

Functional Enrichments:

a) Biological Process:

Interleukin-27-mediated signaling pathway: This pathway may play a role in immune responses related to PsA.

Meiotic spindle assembly: Relevant to cell division and potentially altered in PsA.

Purine nucleoside catabolic process: Involved in metabolism and cellular homeostasis.

b) Molecular Function:

Z',5'-oligoadenylate synthetase activity: Associated with antiviral responses.

CXCR chemokine receptor binding: Relevant to inflammation and immune regulation.

c) Network Cluster (STRING):

Mixed Spindle elongation and **Axon hillock**: These clusters may contain functionally related proteins.

d) KEGG Pathways:

Viral protein interaction with cytokine and cytokine receptor: Indicates potential viral interactions relevant to PsA.

IL-17 signaling pathway: Known to be involved in inflammation and autoimmune diseases.

Influenza A: May have implications for immune responses in PsA.

e) Disease-gene associations (DISEASES):

Microphthalmia with limb anomalies: A genetic disorder that could intersect with PsA pathways.

Implications:

These enriched pathways provide insights into PsA pathogenesis, immune responses, and potential therapeutic targets.Consider further investigating genes within these pathways for personalized medicine approaches.The identified pathways can guide drug development and enhance our understanding of PsA biology.In summary, the STRING database highlights critical pathways associated with PsA, offering a foundation for targeted therapies and deeper exploration of disease mechanisms.

4.1.2.2. Network Analysis:

I. Protein-Protein interactions:

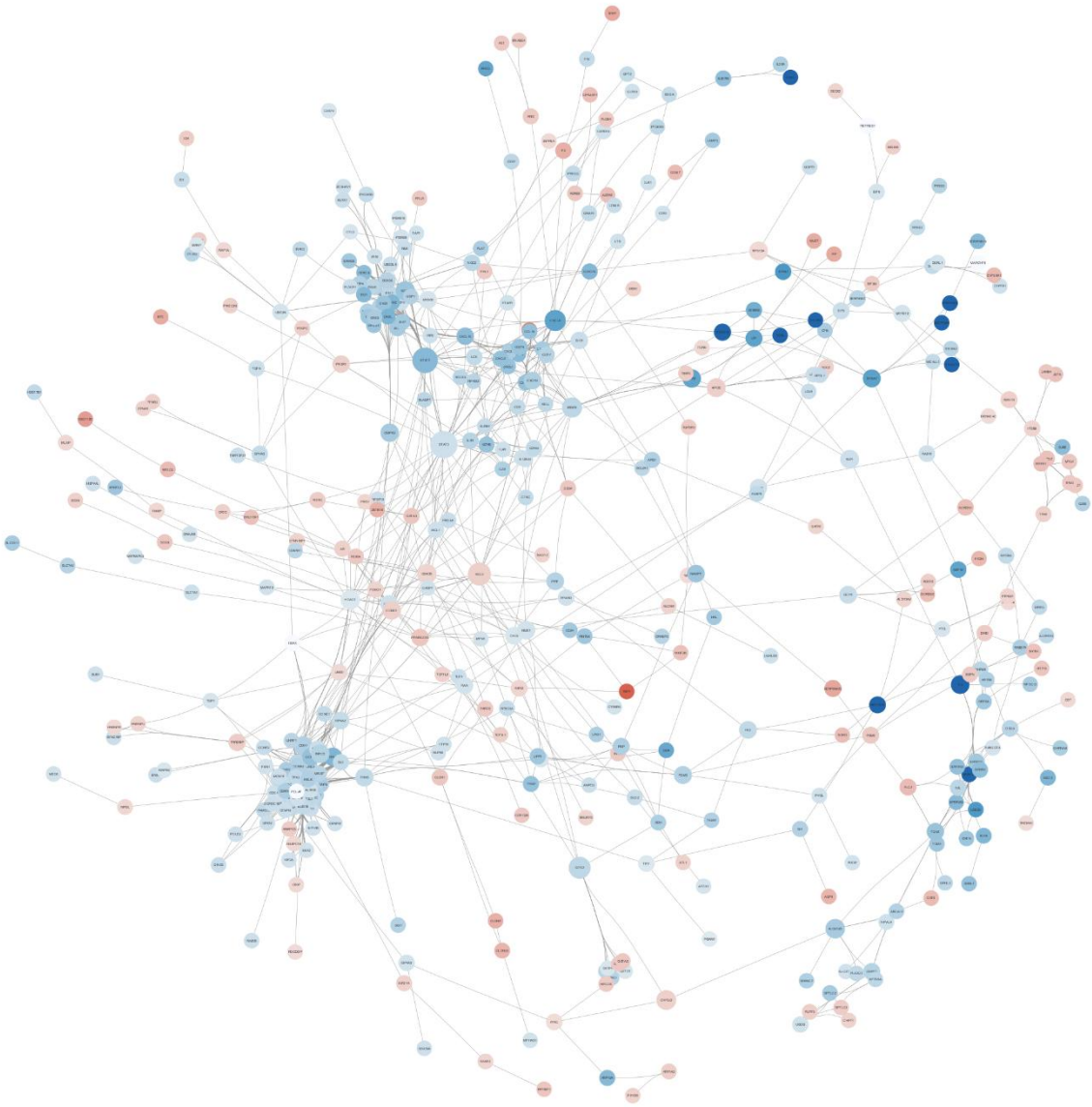

Figure 10: Mapping the Interactome of Significant Genes in Psoriatic Arthritis

The protein-protein interaction network visualized in Cytoscape (Figure 10) represents the complex interplay between significant genes implicated in Psoriatic Arthritis (PsA). Here's an interpretation based on the provided settings and the network's characteristics:

Nodes and Edges: The network comprises nodes (proteins) of various sizes, suggesting different levels of significance. The edges (interactions) indicate the direct or indirect associations between these proteins.

Color Coding: The nodes are colored in shades of blue and red, which could represent upregulated and downregulated genes.

Network Density: The dense connections illustrate the intricate relationships and potential pathways that may be dysregulated in PsA. This complexity underscores the multifactorial nature of the disease.

Cluster Analysis: While not immediately apparent, cluster analysis within Cytoscape could reveal groups of proteins that interact more frequently with each other, suggesting functional modules or biological pathways.

Biological Significance: The network provides a visual hypothesis for the molecular mechanisms underlying PsA. Proteins with a high degree of connectivity (hubs) could be key regulators or potential therapeutic targets.

Therapeutic Implications: By analyzing this network, researchers can identify novel drug targets and understand the potential side effects of modulating specific protein interactions.

In summary, this protein-protein interaction network serves as a foundational map for understanding the molecular interactions in PsA and guiding future research into effective treatments. It highlights the importance of considering the entire interactome when developing therapeutic strategies for complex diseases like PsA.

II. **Hub Gene identification:**

The cytoHubba of the Cytoscape software was used to select the significant hub genes among the obtained DEGs. The Maximal Clique Centrality(MCC) method was used to select the top 10 genes of the PPI network. Then, we used a connectivity degree in the PPI network to evaluate the top 10 genes we just selected.

The results are

Rank	Name	Score	Gene.title	Log2FC
1	CEP55	1.8	centromere protein F	-2.5
1	BIRC5	1.8	baculoviral IAP repeat containing 5	-2.5
1	MELK	1.8	maternal embryonic leucine zipper kinase	-2.7
1	KIF11	1.8	kinesin family member 11	-1.7
1	ASPM	1.8	abnormal spindle microtubule assembly	-2.4
1	CCNB2	1.8	cyclin B2	-2.6
1	BUB1	1.8	BUB1 mitotic checkpoint serine/threonine kinase	-1.6
1	CDK1	1.8	cyclin dependent kinase 1	-2.7
1	CCNA2	1.8	cyclin A2	-2.5
1	CDC20	1.8	cell division cycle 20	-2.7
1	UBE2C	1.8	ubiquitin conjugating enzyme E2 C	-2.0
1	CCNB1	1.8	cyclin B1	-3.7
13	KIF20A	1.8	kinesin family member 20A	-2.5
13	CENPF	1.8	centromere protein F	-1.5
13	TPX2	1.8	TPX2, microtubule nucleation factor	-1.7

The top 15 Hub genes are high lighted with purple border node color.

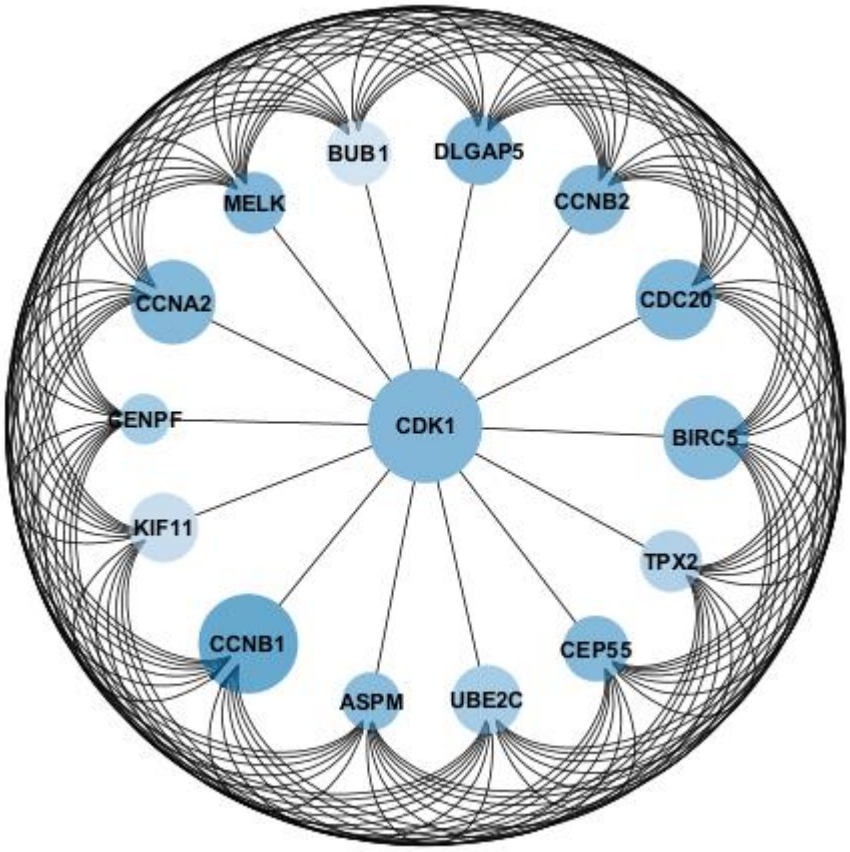

Figure 11: Top 15 Hub genes protein protein network. Continuous mapping- blue color, size continuous mapping - Degree unDir

The hub gene identification process you've conducted using cytoHubba in Cytoscape has yielded significant results. The Maximal Clique Centrality (MCC) method has identified several key genes that may play a central role in the protein-protein interaction (PPI) network related to Psoriatic Arthritis (PsA). Here's an interpretation of your findings:

Central Players: The genes listed, such as **CEP55**, **BIRC5**, **MELK**, and others, have high MCC scores, indicating they are central nodes within the PPI network. Their central position suggests they may be influential in the molecular mechanisms of PsA.

Gene Functions: The identified genes are involved in critical cellular processes:

- **CEP55** and **CENPF** are associated with cell division and centromere function.
- **BIRC5** (Survivin) is known for its role in inhibiting apoptosis and regulating the cell cycle.
- **KIF11** and **KIF20A** are kinesin family members involved in mitotic spindle formation.
- **ASPM** is implicated in the regulation of spindle microtubule dynamics.
- **Cyclins** such as **CCNB1**, **CCNB2**, and **CCNA2** are essential for cell cycle progression.
- **CDK1** is a key kinase that drives cells through the cell cycle.
- **CDC20** and **BUB1** are important for mitotic checkpoint control.
- **UBE2C** is involved in ubiquitin-mediated proteolysis.

Log2 Fold Change (Log2FC): The negative Log2FC values Indicate that these genes are downregulated in the condition being studied. This downregulation could be significant in understanding the dysregulation occurring in PsA.
Potential Drug Targets: Given their central role in the network and involvement in cell cycle regulation, these genes could be potential targets for therapeutic intervention. Drugs that modulate the activity of these proteins might influence the progression of PsA.
Further Research: These hub genes warrant further investigation to understand their exact role in PsA. Experimental validation and functional studies could provide deeper insights into how these genes contribute to the disease pathology.
In summary, the hub genes identified through your analysis are likely to be key contributors to the pathophysiology of PsA and represent promising targets for the development of new treatments.

4.2. Limitations

1. **Sample Size and Diversity**:
    1. Our study relied on available gene expression data from the NCBI GEO database. The sample size and diversity of patients may not fully represent the entire Psoriatic Arthritis (PsA) population.
    2. Future studies should aim for larger and more diverse cohorts, including different PsA subtypes and disease stages.
2. **Data Quality and Normalization**:

    1. The quality of gene expression data can impact the reliability of our findings. Variations in experimental protocols, platforms, and batch effects may introduce biases.
    2. Rigorous data preprocessing, normalization, and quality control are essential. Validation using independent datasets is crucial.

3. **Bioinformatics Predictions vs. Experimental Validation**:

    1. While bioinformatics tools provide valuable insights, they are based on predictions. Experimental validation is necessary to confirm the biological relevance of identified DEGs and hub genes.
    2. Collaborating with wet-lab researchers and conducting functional assays will strengthen our conclusions.

4. **Clinical Relevance and Patient Heterogeneity**:

    1. PsA is a heterogeneous disease with varying clinical presentations. Our findings may not fully capture individual patient profiles.
    2. Stratification based on clinical features (e.g., skin involvement, joint severity) could enhance the precision of biomarker discovery.

5. **Hub Gene Prioritization**:

    1. Although hub genes are central in the PPI network, their functional significance may vary. Not all hub genes are equally relevant to PsA pathogenesis.

2. Prioritization based on biological context, literature evidence, and drug feasibility is essential.

6. **Ethical Considerations and Privacy**:

    1. As we delve deeper into personalized medicine, ethical considerations related to genetic research, patient privacy, and data security become critical.
    2. Transparency and informed consent are paramount when handling patient data.

7. **Clinical Translation Challenges**:

    1. Bridging the gap between research and clinical practice remains a challenge. Implementing personalized treatments based on biomarkers requires robust clinical trials and regulatory approvals.
    2. Collaboration with clinicians, patient advocacy groups, and regulatory bodies is essential.

Our analysis provides a foundation for personalized medicine in PsA, but it is just the beginning. Continued research, validation, and interdisciplinary collaboration will drive progress. By acknowledging these limitations, we pave the way for more impactful discoveries and better care for PsA patients.

4.3. Communication to Non-Technical Audience

Introduction

Psoriatic Arthritis (PsA) is a complex autoimmune disease that affects both the skin and joints. As clinicians and patients seek better treatment options, understanding the underlying molecular mechanisms becomes crucial. In this presentation, we summarize our research findings on potential biomarkers for personalized medicine in PsA.

Key Findings

1. **Differentially Expressed Genes (DEGs)**

We identified genes that are significantly upregulated or downregulated in PsA compared to healthy controls.

These DEGs provide insights into the molecular changes associated with PsA.

2. **Enriched Pathways**

Pathway analysis revealed disrupted biological processes in PsA.

Notable pathways include cell cycle regulation, immune response, and skin development.

3. **Protein-Protein Interaction (PPI) Network**

We constructed a PPI network using Cytoscape.

Hub genes (central nodes) within this network were identified as potential key players in PsA.

Hub Genes: Potential Drug Targets

1.**CEP55**

Associated with cell division and centromere function.

Downregulated in PsA.

2. BIRC5 (Survivin)

Regulates apoptosis and cell cycle progression.

Downregulated in PsA.

3. **KIF11 and KIF20A**

Involved in mitotic spindle formation.

Downregulated in PsA.

4. **CCNB1 and CCNA2**

Essential for cell cycle progression.

Downregulated in PsA.

Clinical Implications

1. **Personalized Treatment**

Targeting hub genes may lead to more effective therapies.

Consider drug repurposing based on these findings.

2. **Patient Stratification**

Use biomarkers to classify patients into subgroups.

Tailor treatments based on individual profiles.

Future Directions

1. **Experimental Validation**

    Validate DEGs and hub genes in the lab.

    Confirm their relevance in PsA.

2. **Clinical Trials**

    Test potential drugs targeting hub genes.

    Evaluate efficacy and safety.

3. **Collaboration**

    Engage with clinicians and patients.

    Translate research into clinical practice.

The research provides a roadmap for personalized medicine in Psoriatic Arthritis. By understanding the molecular intricacies, we aim to improve patient outcomes and revolutionize PsA treatment. Together, we can make a difference!

Thank you.

---

5. Conclusion

The project aimed at identifying biomarkers for personalized medicine in Psoriatic Arthritis (PsA) has made significant strides. Through meticulous research and analysis, the project has leveraged gene expression data from the NCBI GEO database to uncover differentially expressed genes (DEGs) and enriched pathways that may contribute to PsA pathogenesis. Advanced bioinformatics tools and techniques, including Cytoscape's cytoHubba plugin, have been employed to construct a protein-protein interaction network and identify hub genes that could serve as potential drug targets.

The project's comprehensive approach to identifying biomarkers for personalized medicine in Psoriatic Arthritis (PsA) has been a multifaceted endeavor, integrating advanced bioinformatics analyses with a keen focus on the clinical implications of the findings. By accessing and analyzing gene expression data from the NCBI GEO database, the project has successfully identified a set of differentially expressed genes (DEGs) that distinguish PsA from normal controls. These DEGs have been further scrutinized through pathway analysis, revealing disruptions in critical biological processes such as cell cycle regulation, immune response, and skin development.

Utilizing the Cytoscape software, a detailed protein-protein interaction (PPI) network was constructed, providing a visual representation of the complex molecular interactions implicated in PsA. The application of the Maximal Clique Centrality (MCC) method via the cytoHubba plugin allowed for the identification of hub genes within this network. These hub genes, which exhibit significant connectivity and centrality, are hypothesized to play pivotal roles in the disease mechanism and represent promising candidates for therapeutic targeting.

The project has not only highlighted the potential of these biomarkers in the context of drug discovery but also emphasized the importance of personalized treatment strategies. The negative log2 fold changes associated with these hub genes suggest a pattern of downregulation in PsA, which could be indicative of potential targets for upregulation therapies.

Moving forward, the project aims to bridge the gap between bioinformatics research and clinical application. The next steps involve experimental validation of the bioinformatics predictions, clinical trials to assess the efficacy of targeting the identified genes, and the exploration of drug repurposing opportunities. The ultimate goal is to translate these findings into tangible benefits for patients suffering from PsA, offering them more effective and personalized treatment options.

This project stands as a testament to the power of integrating computational and experimental methodologies in the pursuit of precision medicine. The insights gained from this research have the potential to revolutionize the treatment landscape

for Psoriatic Arthritis, paving the way for more targeted and effective therapeutic interventions. The continued exploration of the identified biomarkers and pathways will be crucial in advancing our understanding of PsA and improving patient outcomes.

Reference

Xiangxin Zhang, Liu Yang, et all.(2020). *Identification of Potential Hub Genes and Therapeutic Drugs in Malignant Pleural Mesothelioma by Integrated Bioinformatics Analysis.* Published by S. Karger AG, Basel.
https://karger.com/ort/article-pdf/43/12/656/4140947/000510534.pdf