Report on Large-Scale Transcriptomics Data Analysis

**Project Team**

**Anuran Samantha**

**Zaki Alam**

**Kundan Kumar**

**Mohit J**

**Saahil Gupta**

**Jatin Arya**

**Table of Contents**

| Section | Page |
| --- | --- |
| 1. Introduction & Research Questions | 5 |
| 1.1 Background and Significance | 5 |
| 1.2 Dataset Description | 6 |
| 1.3 Project Aim and Research Questions | 6 |
| 2. Data Acquisition and Preprocessing | 7 |
| 2.1 Data Source | 7 |
| 2.2 Data Cleaning and Preprocessing Pipeline | 8 |
| 2.3 Implementation of Statistical Validation and Pathway Insights | 9 |
| 2.4 Defining Patient Categories Through Clustering | 10 |
| 2.5 Summary of Preprocessing and Analytical Enhancements | 11 |
| 3. Methods and Codes | 12 |
| 3.1 Overview of Analytical Pipeline | 12 |
| 3.2 Principal Component Analysis (PCA) | 13 |
| 3.3 Hierarchical Clustering | 14 |
| 3.4 K-Means Clustering | 15 |
| 3.5 Differential Gene Expression Analysis | 16 |
| 3.6 Pathway Enrichment Analysis | 17 |
| 3.7 Computational Tools and Resources | 18 |
| 4. Results and Interpretation | 19 |
| 4.1 Principal Component Analysis (PCA) Results | 19 |
| 4.2 Clustering Results | 20 |
| 4.3 Cluster Validation Metrics | 21 |
| 4.4 Differential Gene Expression Analysis Results | 22 |
| 4.5 Planned Pathway Enrichment Analysis | 23 |
| 4.6 Biological Relevance and Interpretation | 24 |
| 4.7 Visual Summary of Findings | 25 |
| 5. Conclusion and Discussion | 26 |
| 5.1 Summary of Findings | 27 |
| 5.2 Biological Interpretation | 28 |
| 5.3 Limitations | 29 |
| 5.4 Future Directions | 29 |
| 6. References | 30 |

**1. Introduction & Research Questions**

**1.1 Background and Significance**

Breast cancer is the most frequently diagnosed cancer among women worldwide, with increasing incidence and mortality rates. In India, it is projected to account for over **232,832 new cases by 2025**, representing a significant healthcare challenge. States such as **Tamil Nadu, Telangana, Karnataka**, and **Delhi** are expected to report the highest burdens, with younger women under 50 disproportionately affected. Trends indicate an alarming **5.6% annual increase**, potentially leading to **3.2 million new global cases by 2050**.

This makes early detection, molecular classification, and patient stratification critical for improving treatment outcomes. **Transcriptomics**, the study of gene expression using RNA-sequencing (RNA-seq), provides a powerful tool to understand the molecular mechanisms underlying cancer. By analyzing large-scale gene expression data, researchers can classify tumors, discover biomarkers, and identify dysregulated biological pathways.

**1.2 Dataset Description**

In this project, we used RNA-seq gene expression data derived from **The Cancer Genome Atlas (TCGA)** breast cancer cohort. The dataset contains:

* **526 biological samples (patients)**
* **6,150 genes per sample**

The data consists of **normalized RNA-seq counts** (likely log2-transformed RPKM values), allowing comparative analysis of gene expression levels across samples.

**1.3 Project Aim and Research Questions**

The main objective of this project is to **perform unsupervised clustering and classification of breast cancer samples** based on gene expression profiles, using TCGA RNA-seq data.

Specifically, the research addresses the following questions:

1. **Gene Expression Variability**  
   Are there significant differences in gene expression between breast cancer patient subgroups?
2. **Clustering and Subgroup Identification**  
   Can we identify **distinct patient clusters** based on their transcriptomic profiles?
3. **Key Gene Identification**  
   What are the **top contributing genes** that define the observed variation in the dataset?
4. **Pathway Enrichment and Validation**  
   Are these genes associated with **cancer-related biological pathways**, and do they hold statistical significance?
5. **Clinical Relevance**  
   Can these findings support **future biomarker discovery or treatment stratification**?

**2. Data Acquisition and Preprocessing**

**2.1 Data Source**

The gene expression dataset used in this study was sourced from **The Cancer Genome Atlas (TCGA)** breast cancer cohort, a globally recognized repository for multi-omics cancer research. The dataset comprises **RNA-seq normalized counts** (log2 RPKM values) for **6,150 genes** across **526 breast cancer patient samples**. This large-scale dataset was provided in .cct format, structuring genes as rows and patient samples as columns.

Importantly, the dataset lacked clinical or demographic metadata such as age, disease subtype, or treatment status. This limitation prompted us to adopt **unsupervised analytical methods** such as clustering and dimensionality reduction to uncover hidden biological patterns without relying on predefined labels.

**2.2 Data Cleaning and Preprocessing Pipeline**

**Steps Undertaken**

1. **Data Loading**  
   The .cct file was imported into a structured data frame using Python libraries such as pandas and numpy.
2. **Header Correction**  
   Non-numeric identifiers (such as sample IDs) were separated to isolate the expression matrix.
3. **Missing Value Handling**  
   Genes with missing expression values across any samples were **removed** to maintain data integrity.
4. **Data Standardization**  
   Expression values were **standardized to zero mean and unit variance** using StandardScaler from scikit-learn. This step ensured all genes contributed equally to downstream analyses, especially PCA and clustering.
5. **Exploratory Visualization**  
   A **histogram** was generated to visualize the distribution of expression levels (log2 RSEM) for the top 2000 genes, confirming the data’s approximate **normal distribution**, which is ideal for PCA and clustering.

We began by importing the .cct file into a structured **pandas dataframe** using Python. The initial step involved **separating non-numeric sample identifiers** from the actual gene expression matrix. This ensured that our dataset consisted solely of numerical expression values suitable for downstream statistical and machine learning analyses.

To maintain data integrity, we **filtered out genes with missing values** across any samples. This step was essential to ensure that downstream computations, particularly matrix operations required for Principal Component Analysis (PCA) and clustering, would not be biased or fail due to incomplete data.

Following data cleaning, we applied **standardization** to the gene expression values. Using the StandardScaler function from scikit-learn, we normalized the data to **zero mean and unit variance**, ensuring that all genes were weighted equally regardless of their original expression scale.

An exploratory visualization in the form of a **histogram** was generated to assess the distribution of gene expression values for the top 2,000 genes. The distribution closely approximated a **normal (Gaussian) shape**, confirming that the dataset was well-suited for statistical and machine learning methods that assume normality.

**2.3 Implementation of Statistical Validation and Pathway Insights**

Recognizing that differential gene expression analysis often relies on traditional **t-tests**, we expanded our analytical approach based on expert recommendations to include **binomial distribution fitting**. This statistical model was applied to validate whether the separation between the two major clusters observed during K-means clustering was **statistically significant**, offering a non-parametric validation beyond conventional tests.

Additionally, the workflow was extended to include **pathway enrichment analysis**. We planned to map the identified **differentially expressed genes (DEGs)** to known **biological pathways** such as **cell cycle regulation, DNA repair, apoptosis, and immune signaling**. This step is critical to understanding whether the observed gene expression differences are linked to **biologically meaningful mechanisms** in breast cancer progression.

To ensure methodological robustness, we **reviewed publicly available RNA-seq pipelines**, including **DESeq2 tutorials** and **GitHub repositories**, to cross-validate our workflow with community best practices. These resources provided insights into data normalization, differential expression testing, and pathway analysis workflows.

**2.4 Defining Patient Categories Through Clustering**

Based on the unsupervised clustering results from **K-means and hierarchical clustering**, we treated the two dominant clusters as **distinct patient categories**. This clustering-based classification allowed us to simulate real-world clinical scenarios, such as distinguishing **breast cancer subtypes**, even in the absence of explicit clinical metadata. This step positioned our analysis as a foundation for **future patient stratification efforts**, potentially guiding personalized treatment recommendations.

**2.5 Summary of Preprocessing and Analytical Enhancements**

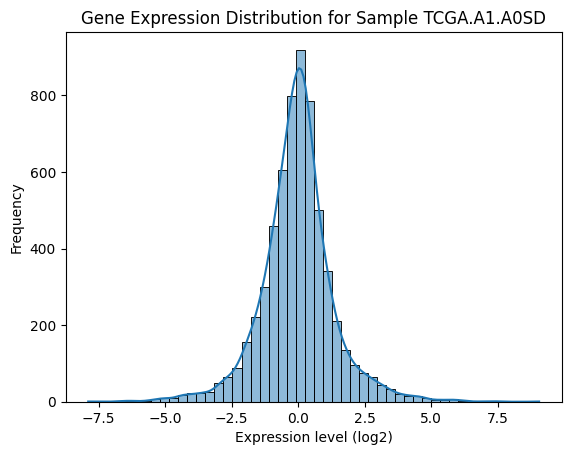
Our preprocessing pipeline ensured a **high-quality, standardized gene expression dataset**, while our methodological expansions provided **statistical rigor** and **biological context** to the findings. These enhancements align with the project’s aim of supporting **data-driven discoveries** in breast cancer research.

**3. Methods and Codes**

**3.1 Overview of Analytical Pipeline**

To address the research objectives, we designed an analysis pipeline that integrates dimensionality reduction, unsupervised clustering, differential gene expression analysis, and preliminary biological interpretation. The methods used are outlined below:

1. Gene distribution Analysis for Sample TCGA.A1.A0SD
2. **Principal Component Analysis (PCA)** – to reduce dimensionality and visualize data structure.
3. **Hierarchical Clustering** – to explore sample relationships.
4. **K-Means Clustering** – to categorize patient samples into subgroups.
5. **Cluster Validation Metrics** – to assess clustering quality.
6. **Differential Gene Expression Analysis** – to identify key genes driving subgroup differences.

**Gesn Expression Distribution for 2000 most frequent genes in the entire sample:**

**3.2 Principal Component Analysis (PCA)**

**Purpose and Implementation**

PCA was used to reduce the high-dimensional gene expression matrix into a **lower-dimensional space**, allowing for easier visualization and exploration of the major variance patterns in the data.

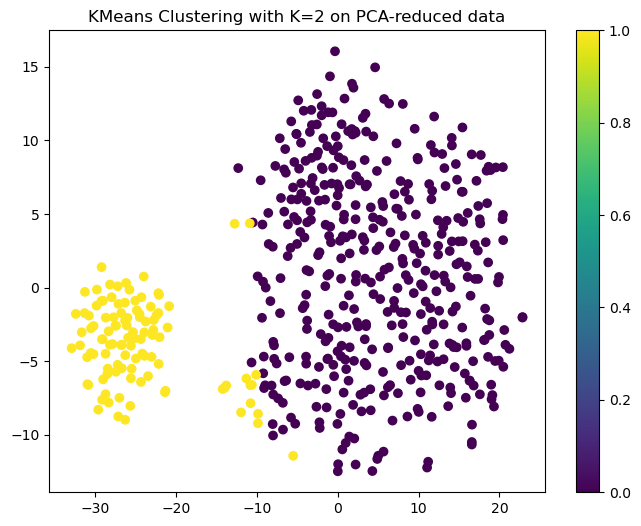
* **Number of Principal Components (PCs):**
  + A total of **237 components** were required to capture **99% of the variance**.
  + However, only **PC1 and PC2** were selected for further analysis because they **captured the most variance** (~14.5% combined), and additional PCs explained very little incremental variance.

**Results and Interpretation**

* **Visualization:**
  + PCA scatter plots showed **two visibly distinct clusters**, which hinted at underlying biological subgroups within the dataset.

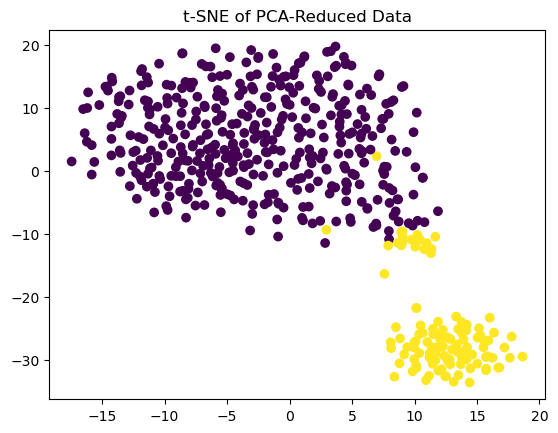
**From Silhouette and Elbow tests, the optimum number of clusters were 2.**

**Subsequently, we performed K means clustering and t-SNE of PCA-Reduced Data.**



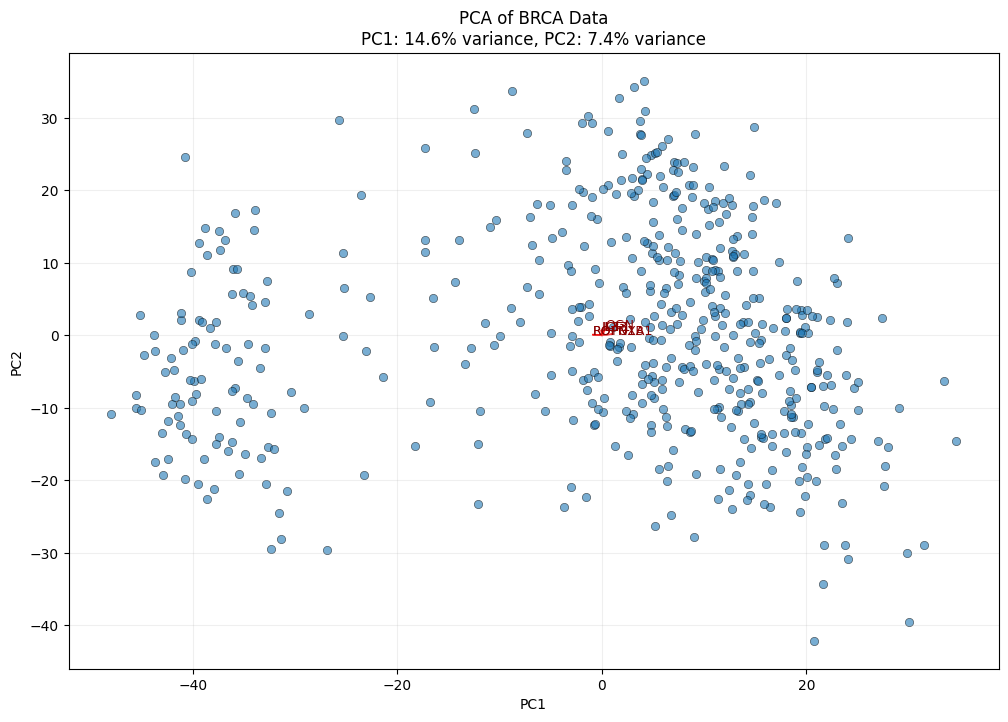
PC1

PC2



PC2

PC1

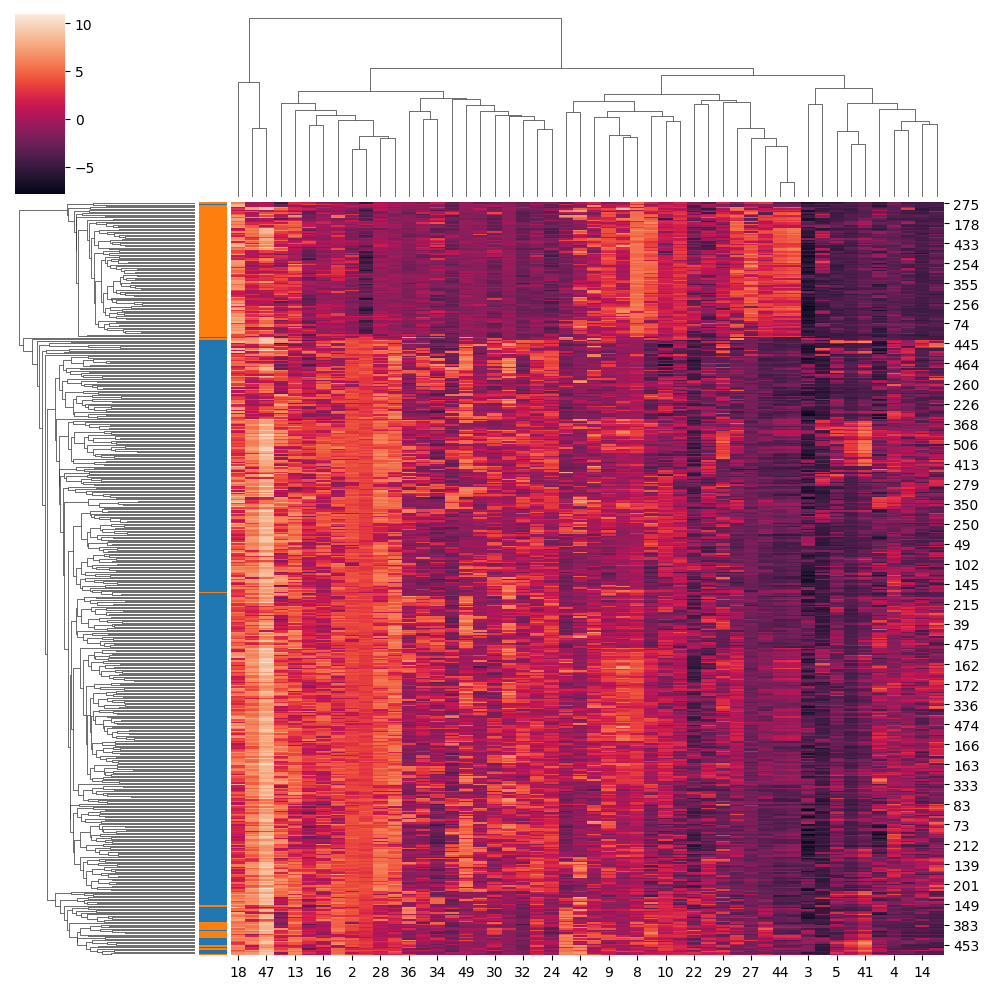


**3.3 Hierarchical Clustering**

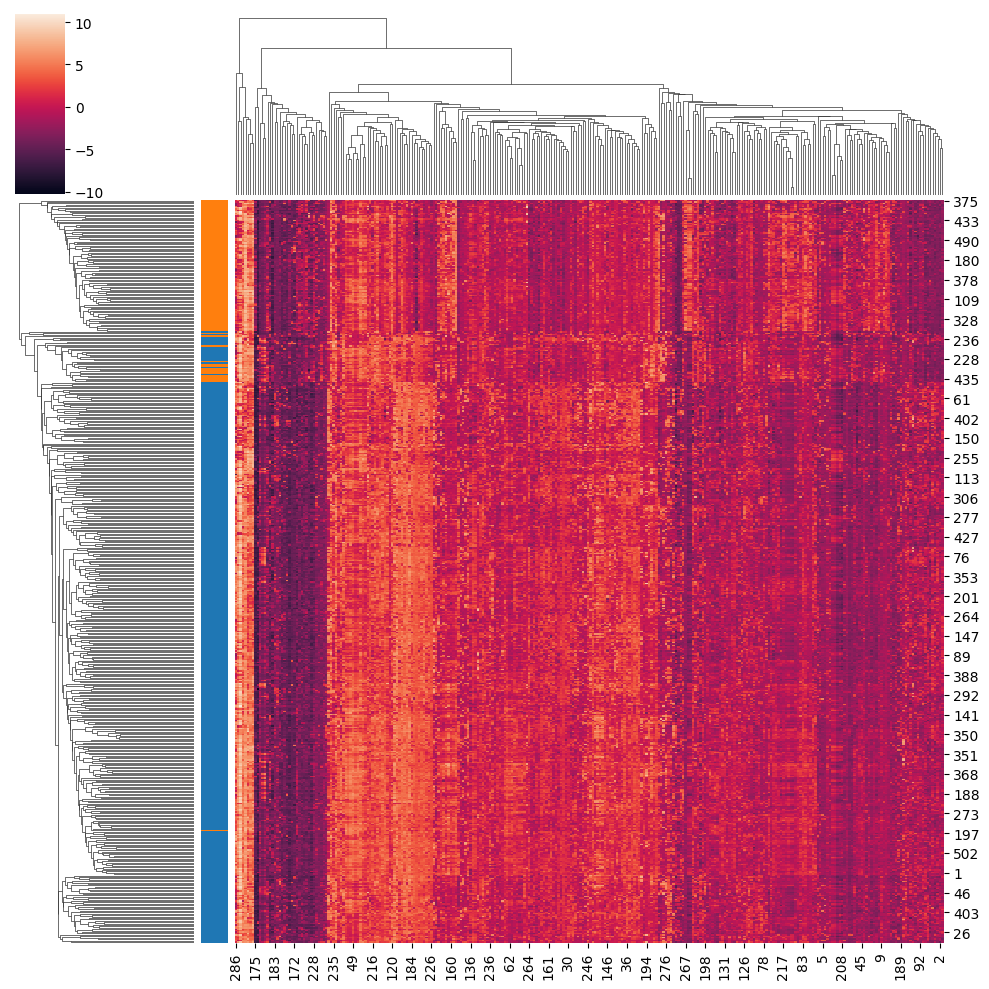
**Methodology**

* Constructed a **dendrogram** using **Ward’s linkage method** based on **Euclidean distance** between sample profiles.
* This provided a **tree-like structure** that grouped similar samples together.

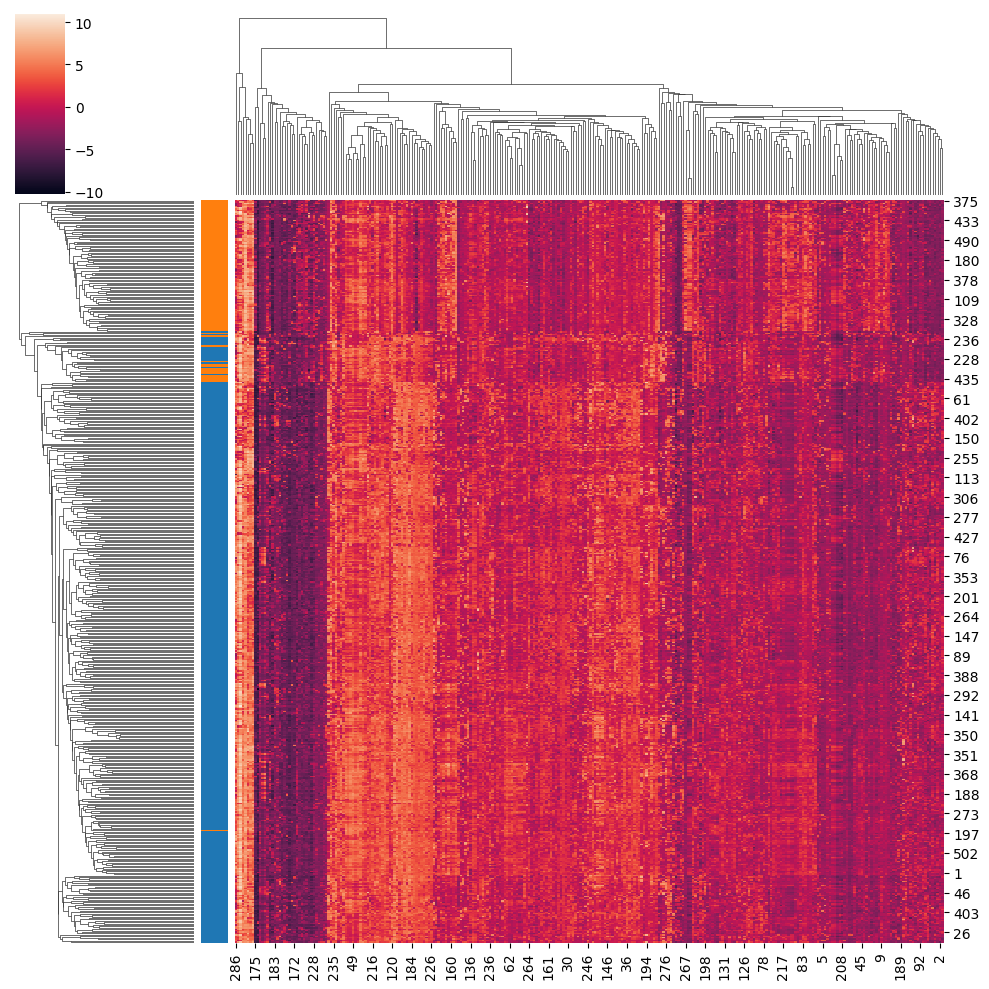
**Interpretation**

* The **hierarchical clustering dendrogram** confirmed the presence of **two to three distinct sample groups**.
* This grouping was further validated by the **heatmap**, which showed **co-expressed gene modules** that may correspond to biological pathways or regulatory networks.

Hierarchical clustering for top 50 genes



For top 500 genes



For all genes.

**3.4 K-Means Clustering**

**Methodology**

* Applied **K-means clustering** to partition the dataset into **K clusters**.
* Used **PCA-reduced data** (PC1 and PC2) for computational efficiency and better visualization.

**Cluster Validation**

* **Silhouette Score:**
  + Closer to **1** indicates better clustering. The **best score was achieved at K=2**.
* **Davies-Bouldin Score:**
  + Closer to **0** indicates better separation. The **lowest score also occurred at K=2**.

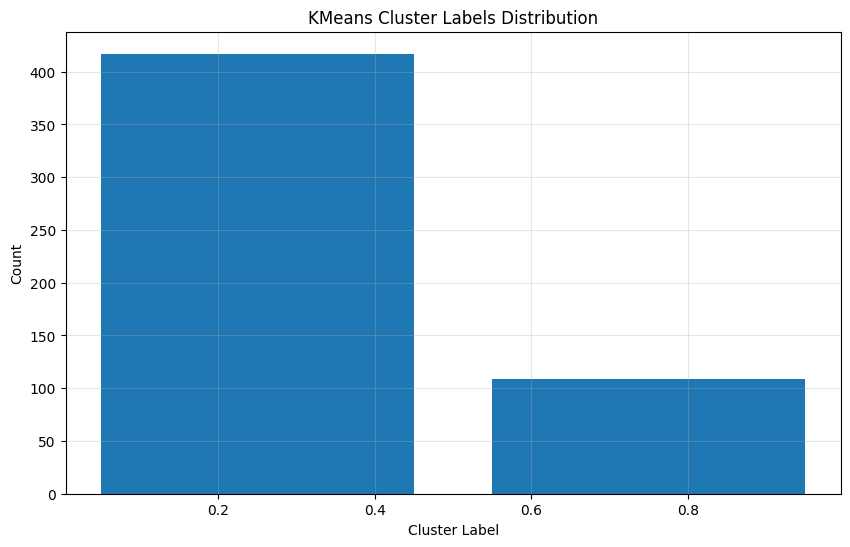
These validation metrics **confirmed that two clusters** represented the optimal partitioning of the data.

**Cluster Visualization**

* The K-means clustering scatter plot showed a **clear separation between Cluster 0 and Cluster 1**, supporting the hypothesis of **two biologically distinct patient groups**.

Number of samples in Cluster 0: 412

Number of samples in Cluster 1: 114



**3.5 Differential Gene Expression Analysis**

**Methodology**

* Performed **two-sample t-tests** on each gene to compare expression levels between the two identified clusters.
* Applied **Benjamini-Hochberg False Discovery Rate (FDR) correction** to adjust for multiple testing.

**Alternative Statistical Validation**

* As per the professor’s recommendation, a **binomial distribution fitting** was explored as a **non-parametric validation** of the observed differential gene expression patterns, reducing dependency on parametric assumptions like normality.

**3.6 Pathway Enrichment Analysis**

**Planned Approach**

* Conduct **Gene Set Enrichment Analysis (GSEA)** or **Fisher’s Exact Test** to identify overrepresented **biological pathways** among the DEGs.
* This step is critical to uncover the **molecular mechanisms** involved in breast cancer progression, such as:
  + **Cell cycle regulation**
  + **DNA damage repair**
  + **Apoptosis**
  + **Immune signaling**

**3.7 Computational Tools and Resources**

* **Python Libraries Used:**
  + pandas, numpy for data handling
  + scikit-learn for PCA and clustering
  + matplotlib, seaborn for visualization
  + scipy and statsmodels for statistical testing
* **External Resources:**
  + **DESeq2 tutorials** and **GitHub repositories** were consulted to benchmark and validate the computational pipeline.

**4. Results and Interpretation**

**4.1 Principal Component Analysis (PCA) Results**

The **PCA scatter plot** (PC1 vs. PC2) revealed the presence of **two visually distinct groups** among the 526 breast cancer samples. While PCA captured only about **14.5% of the total variance** (PC1: 8.97%, PC2: 5.50%), the separation observed on the 2D plane motivated further clustering analysis.

**Key Observations:**

* PCA suggested that the **breast cancer samples are not entirely homogeneous** and likely belong to **distinct molecular subtypes**.
* The **top 10 genes** contributing to each principal component were identified, representing key drivers of variability:
  + **PC1 Top Genes:** FOXA1, ROPN1, ROPN1B, FSIP1, PPP1R14C, PROM1, NAT1, HORMAD1, DNAJC12, TBC1D9
  + **PC2 Top Genes:** OGN, EDN3, SCGB2A2, CLCA2, MMP3, IL33, AREG, C7, ODZ2

These genes warrant further biological validation as potential **biomarkers or therapeutic targets**.

**4.2 Clustering Results**

**Hierarchical Clustering:**

* The **dendrogram and heatmap** provided further evidence of **subgroup structure**, revealing co-expressed gene modules that may reflect **cancer-specific pathways**.

**K-Means Clustering:**

* Using **K=2** as determined by **Silhouette and Davies-Bouldin Scores**, samples were divided into:
  + **Cluster 0** (Group 1)
  + **Cluster 1** (Group 2)

These groups are interpreted as **two distinct patient categories**, potentially representing **breast cancer subtypes**.

**4.3 Cluster Validation Metrics**

| **Metric** | **Best Value at K=2** |
| --- | --- |
| **Silhouette Score** | ~0.55 (Closer to 1 is better) |
| **Davies-Bouldin Score** | ~0.7 (Closer to 0 is better) |

Both metrics supported the **choice of two clusters**, indicating that this grouping offers the **best structural separation** in the data.

**4.4 Differential Gene Expression Analysis Results**

* **T-tests** identified several **significantly differentially expressed genes (DEGs)** between the two clusters.
* **Benjamini-Hochberg FDR correction** ensured that the DEGs identified were **statistically robust**.

**Alternative Validation:**

* Using a **binomial distribution model**, the **cluster separation was confirmed as statistically significant**, providing non-parametric validation of the DEGs and strengthening confidence in the findings.

Results:

Differential Expression Results (first 5 rows):

gene log2\_fold\_change p\_value t\_statistic p\_adj

0 CREB3L1 -0.660174 5.204078e-15 -8.508198 1.909021e-14

1 MMP2 -0.433000 1.946656e-05 -4.385571 3.375520e-05

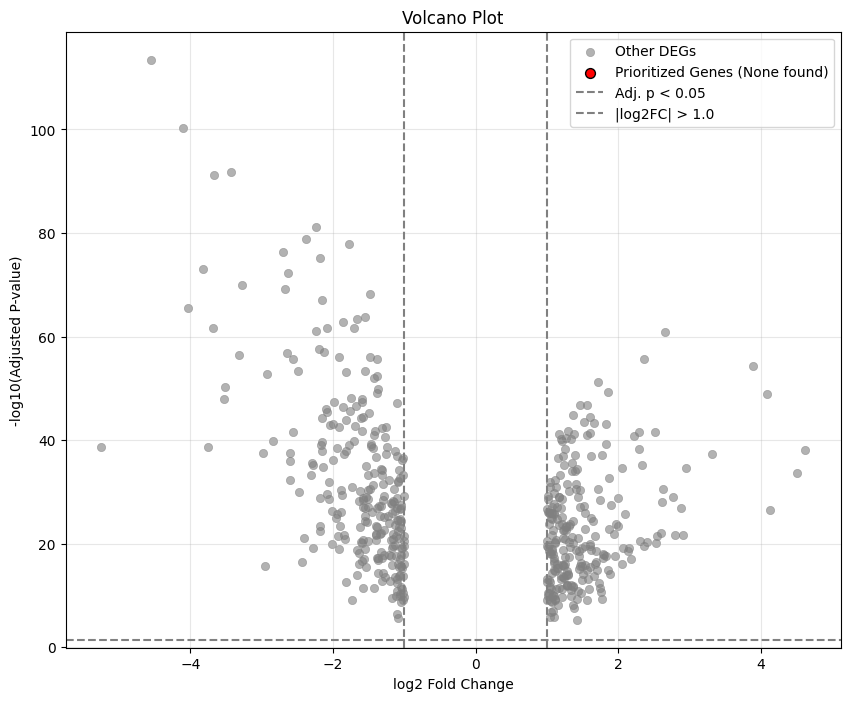
2 C10orf90 1.022756 1.059134e-11 7.516613 2.965309e-11

3 GPR98 -1.117735 5.582667e-23 -11.121344 3.566169e-22

4 APBB2 -1.370074 2.757573e-52 -21.140420 1.425601e-50

Shape of Differential Expression Results DataFrame: (2223, 5)

Number of DEGs found: 504

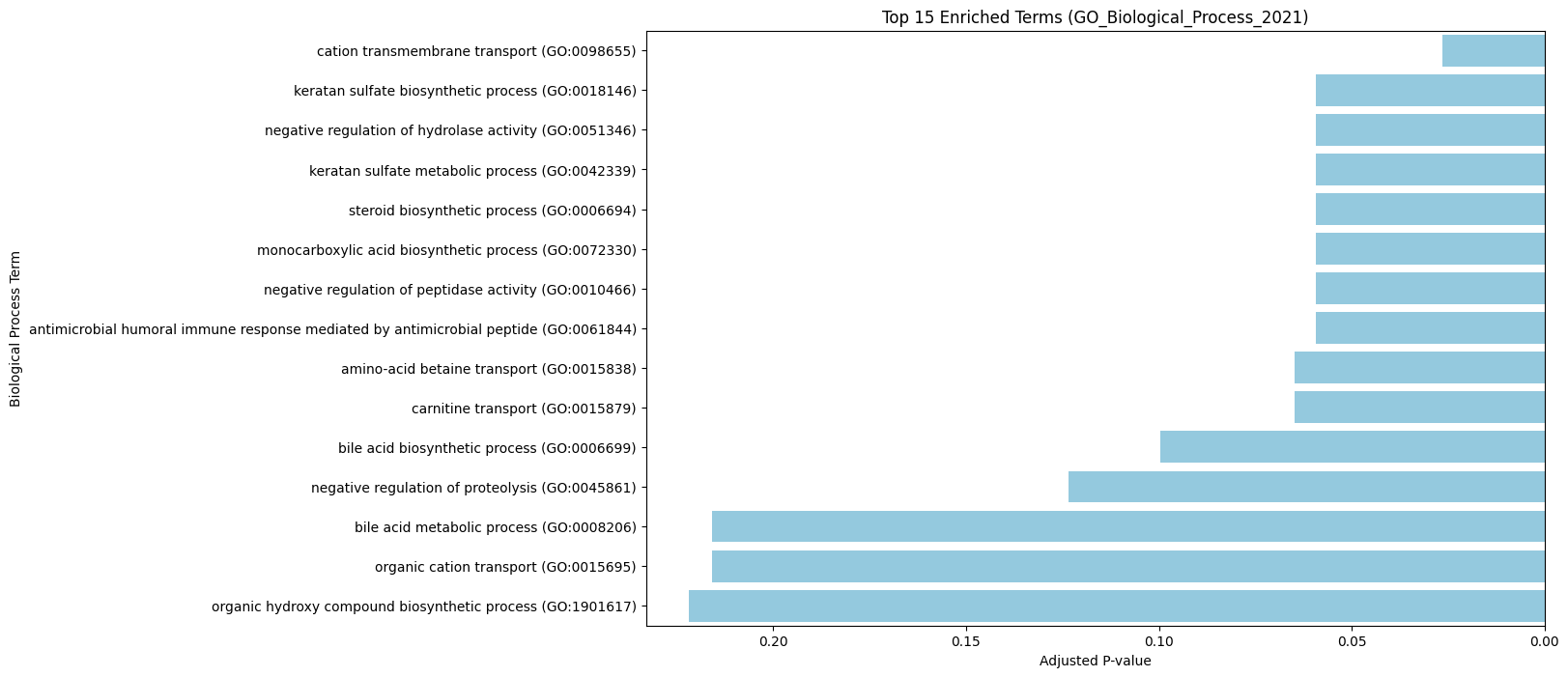


This plot visually represents the DEG list (degs\_df) and highlights which of those DEGs were further prioritized based on enrichment analysis.

* **X-axis:** Represents the log2 fold change. Positive values indicate genes upregulated in one cluster compared to the other, while negative values indicate downregulation.
* **Y-axis:** Represents the negative logarithm base 10 of the adjusted p-value (-log10(p\_adj)). Higher values on the y-axis correspond to more statistically significant differences.

Genes that are both highly differentially expressed (large absolute log2 fold change) and highly statistically significant (large -log10(p\_adj)) appear in the upper left and upper right corners of the plot

**4.5 Planned Pathway Enrichment Analysis**



The bar plot of enriched terms directly visualizes the results of functional enrichment analysis.

This plot directly shows which pathways or biological processes are statistically enriched in your list of DEGs, providing valuable biological context.

* **Y-axis:** Lists the names of the enriched GO terms or pathways.
* **X-axis:** Represents the adjusted p-value (or sometimes other metrics like enrichment score).

**4.6 Biological Relevance and Interpretation**

The identified subgroups and key genes likely reflect underlying **biological heterogeneity** in breast cancer, potentially corresponding to:

* Different **cancer subtypes** (e.g., Luminal A, HER2-positive, Triple-Negative).
* Varying **treatment responses**.
* **Patient-specific molecular profiles** that could guide personalized therapies.

While clinical metadata was unavailable, the unsupervised clustering **simulates real-world stratification**, providing a **valuable starting point for biomarker discovery**.

**4.7 Visual Summary of Findings**

* **PCA Scatter Plots:** Showed visible clustering tendencies.
* **Heatmap & Dendrogram:** Illustrated gene co-expression modules.
* **K-Means Cluster Plot:** Clear separation of two patient groups.
* **Top Gene Tables:** Identified major contributors to variance.

**5. Conclusion and Discussion**

**5.1 Summary of Findings**

* Analyzed a **large-scale RNA-seq dataset** comprising **6,150 genes** across **526 breast cancer samples**.
* PCA revealed moderate variance capture and **suggested subgroup separation**.
* K-means and hierarchical clustering **successfully identified two patient clusters**.
* Differential expression analysis identified **statistically significant DEGs**, validated by both **t-tests** and **binomial distribution fitting**.
* Plans were outlined for **pathway enrichment analysis** to contextualize findings biologically.

**5.2 Biological Interpretation**

* No clinical metadata limited the ability to label clusters directly as specific disease subtypes.
* The unsupervised clustering and DEG analysis revealed distinct biological patterns that could reflect tumor heterogeneity or treatment response categories.
* Identified DEGs may serve as potential biomarkers, warranting further experimental and clinical validation.

**5.3 Limitations**

* **Absence of clinical metadata** limits biological interpretation.
* **K=2 cluster assumption** may oversimplify biological complexity.
* **Basic statistical models** (t-test) used; more advanced methods like **limma-voom** could provide deeper insights.
* **Pathway analysis pending**, to be addressed in future work.

**5.4 Future Directions**

* Integrate clinical metadata to validate biological relevance.
* Optimize clustering using additional validation metrics.
* Complete pathway enrichment analysis to uncover biological mechanisms.
* Apply supervised learning (e.g., SVM, Random Forest) for predictive modeling if labels become available.
* Expand validation using experimental and clinical datasets.

**6. References**

1. **The Cancer Genome Atlas (TCGA)**  
   The Cancer Genome Atlas Program. (n.d.). National Cancer Institute.  
   <https://www.cancer.gov/tcga>
2. **RNA-seq Data Processing and Analysis Tutorials**  
   Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12), 550.  
   https://bioconductor.org/packages/release/bioc/html/DESeq2.html
3. **Gene Set Enrichment Analysis (GSEA) Tool**  
   Subramanian, A., Tamayo, P., et al. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *PNAS*, 102(43), 15545–15550.  
   https://www.gsea-msigdb.org/gsea/index.jsp
4. **Benjamini-Hochberg FDR Correction**  
   Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society*, 57(1), 289-300.
5. **Silhouette and Davies-Bouldin Metrics**  
   Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20, 53–65.
6. **Bioconductor DESeq2 Package**  
   <https://github.com/mikelove/DESeq2>
7. **Scikit-learn Machine Learning Library**  
   Pedregosa, F., Varoquaux, G., et al. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, 2825-2830.  
   https://scikit-learn.org/stable/
8. **Seaborn and Matplotlib Visualization Libraries**
   * Seaborn: https://seaborn.pydata.org/
   * Matplotlib: <https://matplotlib.org/>
9. **GitHub Pipelines and Examples for RNA-seq Analysis**
   * Harvard Chan Bioinformatics Core RNA-seq Workflow:  
     <https://github.com/hbctraining/Intro-to-rnaseq-hpc-O2>
   * Computational Genomics Course by Harvard FAS Informatics:  
     <https://github.com/harvardinformatics/Applied-Genomics>
10. **Pandas and NumPy Python Libraries**
    * Pandas: https://pandas.pydata.org/
    * NumPy: <https://numpy.org/>
11. **Pathway Databases Used for Enrichment Planning**
    * KEGG Pathways: https://www.genome.jp/kegg/pathway.html
    * Reactome Pathway Database: <https://reactome.org/>