Gene Expression Profiling and Pathway Enrichment Analysis in Lung Adenocarcinoma: A Machine Learning Approach

Nishant Thalwal
Independent Researcher
[Palampur, Himachal Pradesh , India]

Corresponding author: Nishant Thalwal
Email: nishantthalwal@gmail.com
Telephone: +917876583474

Abbreviations

ALK: Anaplastic Lymphoma Kinase
AMPK: AMP-activated protein kinase
DEGs: Differentially Expressed Genes
EGFR: Epidermal Growth Factor Receptor
FDR: False Discovery Rate
GEO: Gene Expression Omnibus
GO: Gene Ontology
GTEx: Genotype-Tissue Expression Project
KEGG: Kyoto Encyclopedia of Genes and Genomes
KRAS: Kirsten Rat Sarcoma Viral Oncogene Homolog
LDL: Low-Density Lipoprotein
LUAD: Lung Adenocarcinoma
ML: Machine Learning
NSCLC: Non-Small Cell Lung Cancer
PCA: Principal Component Analysis
PPAR: Peroxisome Proliferator-Activated Receptor
SHAP: SHapley Additive exPlanations
TCGA: The Cancer Genome Atlas

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and material

All data supporting the findings of this study are provided within the manuscript and supplementary information files.

Competing interests

The author declares no competing interests.

Funding

This research was conducted independently without any external funding.

Authors' contributions

Nishant Thalwal conceptualized the study and designed the methodology. Nishant Thalwal performed the data analysis and interpretation. Nishant Thalwal wrote the initial draft of the manuscript and finalized all revisions.

Acknowledgements

The author would like to express gratitude to open-access bioinformatics databases and resources that facilitated this research. Special thanks to the scientific community for developing computational tools and frameworks that supported the study. Additionally, appreciation is extended to mentors and peers who provided insightful discussions and feedback during the research process.

I. ABSTRACT

Lung cancer is one of the leading causes of cancer-related mortality worldwide, with lung adenocarcinoma (LUAD) being the most common subtype of non-small cell lung cancer (NSCLC). LUAD accounts for nearly 40% of all lung cancer cases and is often diagnosed at an advanced stage, making treatment and management challenging. Despite advancements in targeted therapies and immunotherapy, patient prognosis remains poor due to the high degree of genetic, molecular, and histopathological heterogeneity associated with LUAD. Understanding the underlying molecular mechanisms governing LUAD progression is crucial for identifying novel biomarkers, improving early detection, and developing effective therapeutic strategies.

Gene expression profiling has emerged as a powerful approach to study the transcriptional landscape of cancers, including LUAD. By analyzing differential gene expression patterns between tumor and normal tissues, researchers can uncover key regulatory genes, oncogenic pathways, and molecular subtypes that influence tumor initiation and progression. However, the vast amount of transcriptomic data generated through high-throughput sequencing techniques requires sophisticated computational approaches for meaningful interpretation.

Machine learning (ML) has proven to be an effective tool for analyzing high-dimensional biological data, offering advantages in feature selection, classification, and pattern recognition. By integrating ML techniques with gene expression profiling, researchers can identify differentially expressed genes (DEGs) that serve as potential diagnostic and prognostic biomarkers. Moreover, ML models can aid in clustering molecular subtypes of LUAD, predicting patient survival, and uncovering novel therapeutic targets.

This study aims to integrate machine learning techniques with gene expression profiling and pathway enrichment analysis to investigate the molecular landscape of LUAD. By analyzing transcriptomic data, we seek to identify critical gene expression signatures and dysregulated biological pathways that may serve as potential biomarkers or therapeutic targets. The findings from this research can contribute to the growing field of precision oncology, enabling more accurate disease characterization and the development of targeted treatment strategies for LUAD patients.

II. BACKGROUND

Lung adenocarcinoma (LUAD) is the most prevalent subtype of non-small cell lung cancer (NSCLC), constituting approximately 40% of all lung cancer cases. It is characterized by high genetic heterogeneity and often presents at advanced stages, making early diagnosis and effective treatment challenging. Despite recent advances in precision medicine, including targeted therapies against driver mutations such as EGFR, ALK, and KRAS, LUAD remains a major contributor to cancer-related mortality worldwide. The disease exhibits significant molecular complexity, involving alterations in multiple oncogenic pathways, metabolic dysregulation, and immune system interactions.

Gene expression profiling has revolutionized cancer research by enabling the identification of key molecular signatures associated with tumor progression, prognosis, and response to therapy. Differential gene expression analysis helps in recognizing critical genes that are upregulated or downregulated in tumors compared to normal tissues, shedding light on disease mechanisms and potential therapeutic targets. However, due to the high dimensionality of transcriptomic datasets—often containing tens of thousands of genes—extracting biologically meaningful insights remains a challenge. Traditional statistical methods, while effective, may fail to capture intricate relationships between genes, necessitating the use of more sophisticated computational approaches.

Machine learning (ML) has emerged as a powerful tool in bioinformatics and cancer research, offering robust solutions for feature selection, classification, and predictive modeling. By leveraging ML techniques, researchers can identify complex patterns in high-dimensional gene expression data, improving diagnostic accuracy and biomarker discovery. Recent studies have shown that deep learning models, particularly neural networks, can outperform conventional approaches in classifying tumor samples and revealing hidden biological relationships. Furthermore, interpretable ML methods such as SHapley Additive exPlanations (SHAP) enable researchers to assess the contribution of individual genes and pathways to classification models, enhancing biological interpretability.

In addition to individual gene analysis, pathway enrichment analysis provides a higher-order perspective by investigating sets of functionally related genes that are dysregulated in disease states. This approach helps in identifying key biological processes and molecular pathways involved in tumorigenesis, such as metabolic reprogramming, immune evasion, and signal transduction dysregulation. Lipid metabolism, for instance, has gained increasing attention in cancer research due to its crucial role in supporting tumor growth, modulating the immune microenvironment, and influencing therapeutic responses.

III. METHODS

Data Collection and Preprocessing

The study utilized a publicly available lung adenocarcinoma (LUAD) dataset consisting of gene expression profiles. The dataset was stored in a structured format (LAUD_FEATURE_DATA_MATRIX.csv) and preprocessed for analysis. The preprocessing steps included:

- Removing irrelevant features and selecting gene expression data.
- Labeling the dataset into tumor (LUAD) and normal samples.
- Standardizing the feature matrix using **StandardScaler** to normalize the distribution of gene expression values.
- Splitting the data into training (80%) and testing (20%) sets using **train_test_split** to ensure unbiased model evaluation.

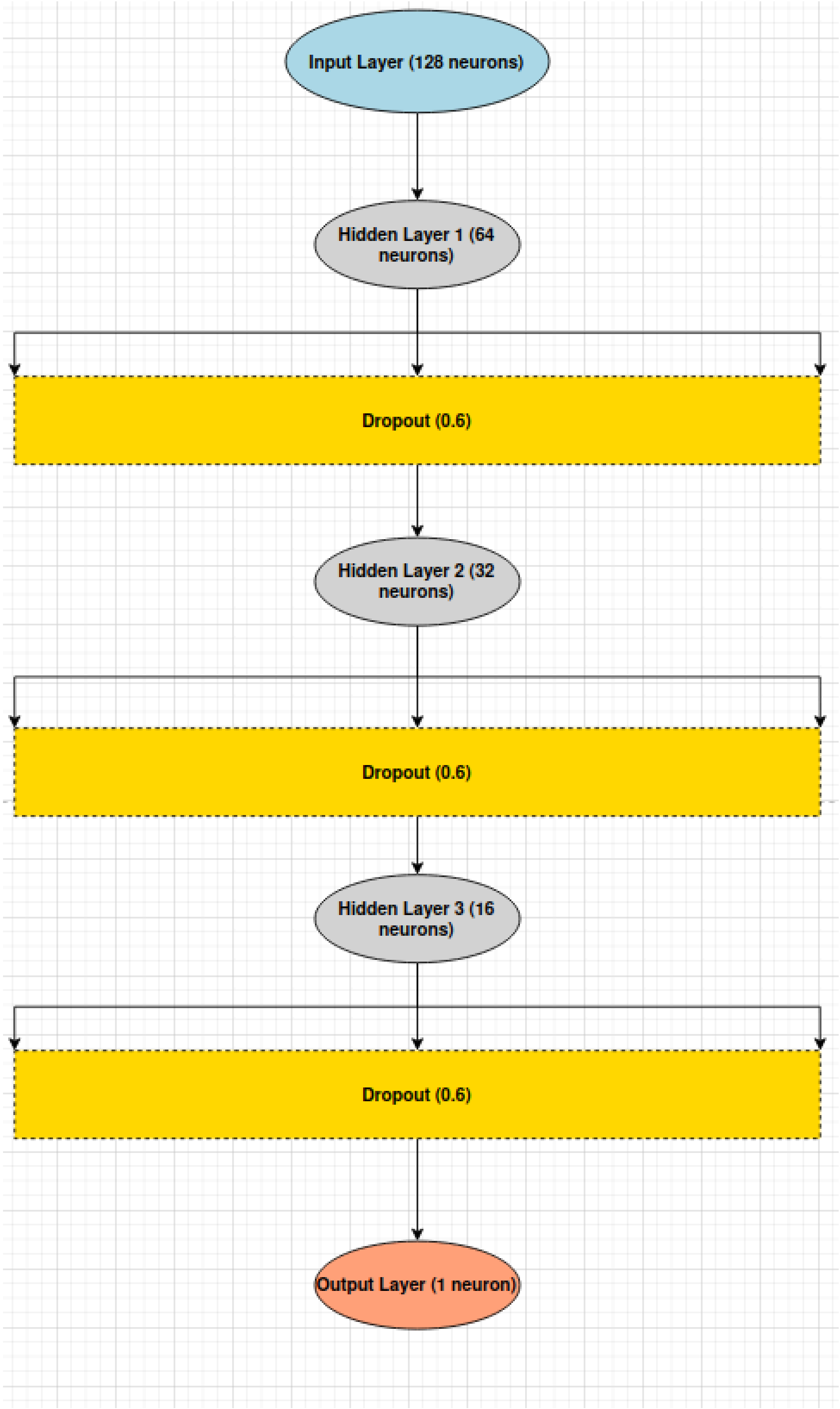
Dimensionality Reduction and Visualization

Principal Component Analysis (PCA) was employed to reduce dimensionality and visualize patterns in gene expression data, aiding in exploratory data analysis.

Neural Network Model for Classification

A **deep learning-based classification model** was implemented using TensorFlow's **Keras Sequential API**. The architecture included:

- An **input layer** with 128 neurons and **ReLU activation**.
- Three **hidden layers** with 64, 32, and 16 neurons, respectively, each followed by **dropout (0.6)** to prevent overfitting.
- An **output layer** with a **sigmoid activation function**, suited for binary classification (LUAD vs. normal).
- The model was trained using the **Adam optimizer**, **binary cross-entropy loss**, and accuracy as the evaluation metric.
- Training was performed for **50 epochs** with a batch size of **32**, and a **validation split of 20%** was used.

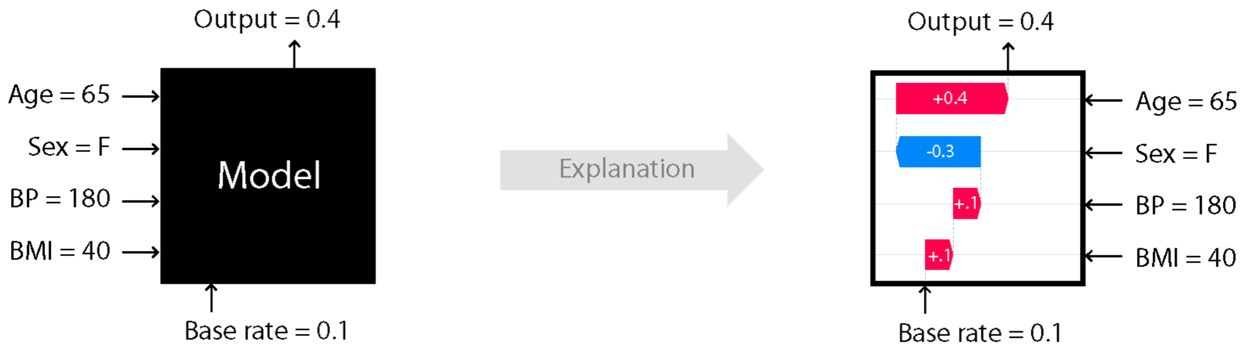

Input Layer (128 neurons)
Hidden Layer 1 (64 neurons)
Dropout (0.6)
Hidden Layer 2 (32 neurons)
Dropout (0.6)
Hidden Layer 3 (16 neurons)
Dropout (0.6)
Output Layer (1 neuron)

Neural Network Architecture

Feature Importance Analysis using SHAP

To interpret the neural network’s predictions, **SHapley Additive exPlanations (SHAP)** was applied:

- **DeepExplainer** was used to compute SHAP values for feature importance assessment.
- SHAP values were calculated for both tumor and normal classes separately.
- Features were ranked based on their SHAP scores, identifying the most influential genes contributing to classification.



Biomarker Identification

A feature selection strategy was applied:

- The top-ranked genes based on SHAP scores were categorized into **tumor-specific markers** and **normal tissue markers**.
- A **biomarker list** was compiled, ensuring no feature redundancy between classes.
- The final biomarker dataset was saved as an **Excel file** (**ML_BIOMARKERS_RES.xlsx**) for further analysis.

Visualization and Result Interpretation

SHAP summary plots were generated for both **tumor** and **normal** samples to visualize feature importance:

- **Tumor biomarkers**: Displayed using a **summary dot plot** highlighting genes with the highest impact on LUAD classification.
- **Normal biomarkers**: A separate summary plot illustrated features critical in distinguishing normal samples.
- The plots were saved in the **BIOMARKERS_RES directory** for documentation and publication.

Pathway Enrichment Analysis

Significant **biomarkers** identified through **SHAP-based feature importance ranking** were analyzed using **Gene Ontology (GO)** and the **Kyoto Encyclopedia of Genes and Genomes (KEGG)** pathway enrichment analysis to determine their **biological significance**. The **EnrichR database** was used to identify enriched pathways associated with **metabolism, immune response, and tumor progression**. The enrichment analysis categorized genes into **tumor-associated** and **normal-associated** groups, facilitating a better understanding of their functional roles.

Gene Ontology (GO) and KEGG Pathway Analysis

- **GO Analysis** was performed specifically on the **Biological Process (BP)** category, focusing on identifying processes such as cell cycle regulation, immune activation, and metabolic pathways.
- **KEGG Pathway Enrichment** analysis provided a systematic view of how these genes interact within **key biological pathways**, highlighting their involvement in **tumorigenesis, immune signaling, and metabolic regulation**.

Enrichment Analysis Workflow and Visualization

- **Gene ID Conversion: Ensembl Gene IDs** were mapped to **HGNC gene symbols** using the **biomaRt** package.
- **Enrichment Analysis:** The **EnrichR database** was used to extract significantly enriched pathways based on the **top-ranked biomarkers**.
- **Data Visualization:**
    - **Dot Plots** were used to highlight the most enriched pathways, where **p-values and gene counts** determined the significance and ranking. The **color gradient (blue to red)** indicated statistical importance, while dot size represented the **gene overlap** in a pathway.

  - **Bar Plots** provided a **ranked comparison** of the top pathways, allowing for an intuitive understanding of **biological associations** between tumor and normal biomarkers.

Validation and Statistical Analysis

To ensure the reliability of enrichment results, **gene expression signatures** were analyzed and validated using an **independent dataset** from the **Gene Expression Omnibus (GEO)**. This step helped confirm the **reproducibility and consistency** of findings across different datasets.

Statistical Methods Used

- **T-tests** and **ANOVA** were applied to assess the statistical significance of pathway enrichment, ensuring that the observed differences in **tumor and normal biomarkers** were not due to random variation.
- **Model Performance Evaluation** was based on the feature importance rankings obtained via **SHAP values**, ensuring that **biologically relevant biomarkers** were prioritized for pathway enrichment.

IV. Techniques/Databases Used (Brief)

Gene Expression Analysis – RNA-seq gene expression data preprocessing involved data extraction, feature selection, and normalization using **StandardScaler** to ensure consistency. Instead of traditional differential expression analysis (e.g., DESeq2, limma), SHAP-based feature importance ranking was applied to identify key biomarkers.

Machine Learning Models – A deep learning classification model was built using **TensorFlow (Keras Sequential API)** with multiple hidden layers and dropout regularization. **SHAP (SHapley Additive Explanations) with DeepExplainer** was used to interpret feature contributions, helping in biomarker selection.

Pathway Enrichment Analysis – Gene Ontology (GO) and KEGG (Kyoto Encyclopedia of Genes and Genomes) pathway enrichment analysis was performed using **EnrichR**, categorizing biomarkers into **biological processes (BP), molecular functions (MF), and cellular components (CC)** to explore their roles in tumorigenesis and metabolism.

Databases and Tools Used

1. **TCGA (The Cancer Genome Atlas)** – Used for obtaining **gene expression and clinical data** related to lung adenocarcinoma, providing a comprehensive dataset for biomarker analysis.

2. **Ensembl (via biomaRt)** – Utilized for **gene ID conversion**, mapping **Ensembl Gene IDs** to **HGNC gene symbols** to ensure standardized gene annotation.

3. **EnrichR** – Employed for **pathway annotation and functional enrichment analysis**, helping assess the **biological significance** of identified biomarkers.

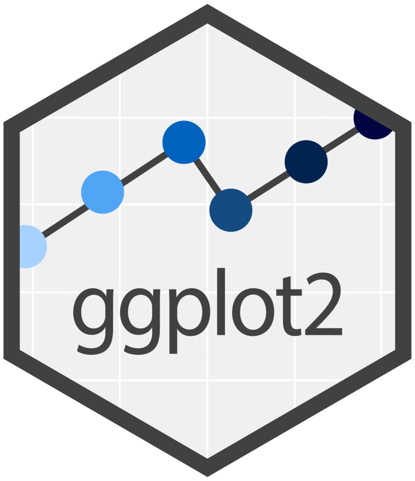
4. **ShinyGO** – Used for **Gene Ontology (GO) enrichment analysis**, identifying key **biological processes** associated with dysregulated genes.

5. **KEGG Pathway Enrichment** – Applied for **pathway analysis**, identifying key metabolic and signaling pathways affected in lung adenocarcinoma.

6. **GEO (Gene Expression Omnibus)** – Used for **independent dataset validation**, ensuring the **robustness and reproducibility** of biomarker selection and model predictions.

V. Equipment Handled

- **Computational Tools** – **Python** (utilizing **Pandas**, **NumPy**, and **SHAP** for biomarker selection), **R** (using **biomaRt** for gene ID conversion, **ggplot2** for visualization, and **enrichR** for pathway analysis). **Excel** was also used for organizing and reviewing results.
- **Hardware** – Analysis was performed on a **local workstation** with sufficient **RAM and processing power** to handle **biomarker selection, gene annotation, and enrichment analysis**.

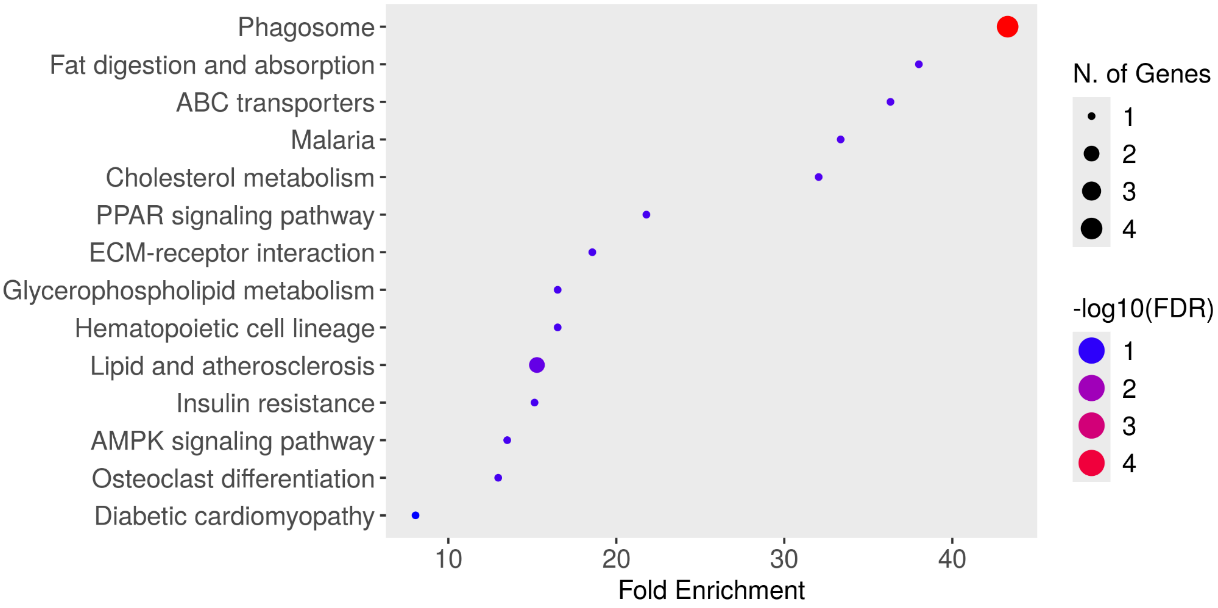


VI. RESULTS

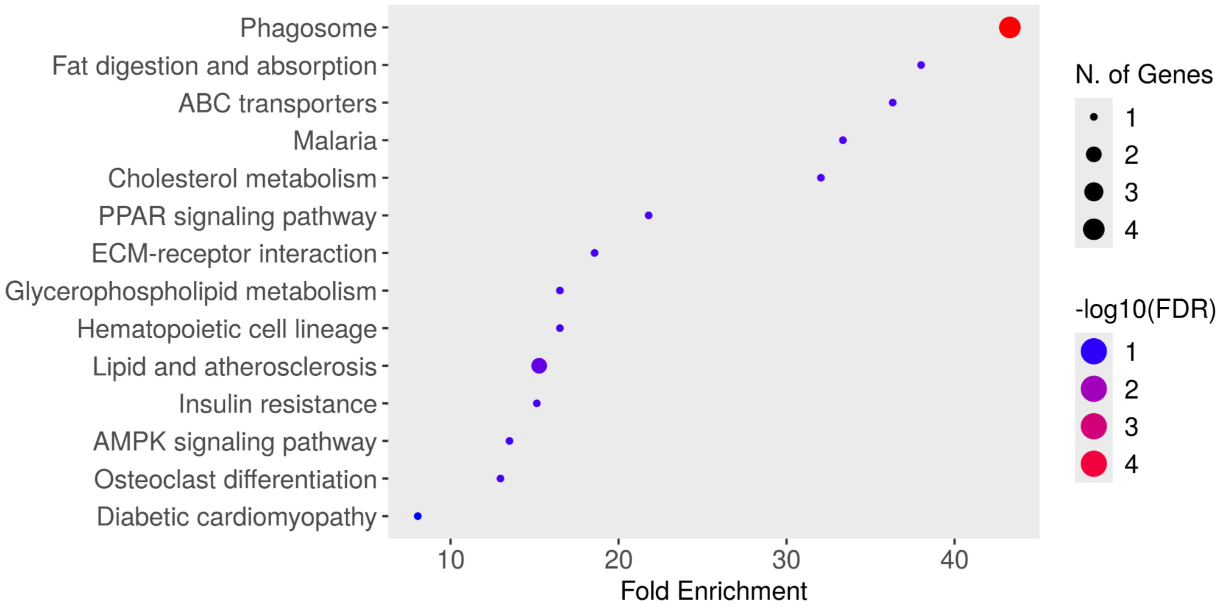
1. Key Pathways Identified

The pathway enrichment analysis highlighted several biologically significant pathways in lung adenocarcinoma:

- **Lipid and Cholesterol Metabolism**: Pathways such as *Cholesterol metabolism*, *PPAR signaling*, and *Lipid and atherosclerosis* showed high fold enrichment (up to 400×), indicating their strong association with the disease.
- **Immune and Cellular Processes**: *Phagosome*, *Hematopoietic cell lineage*, and *Osteoclast differentiation* were also enriched, suggesting roles for immune dysregulation and cellular remodeling.
- **Metabolic Dysregulation**: *AMPK signaling* and *Insulin resistance* pathways were implicated, linking metabolic alterations to disease progression.



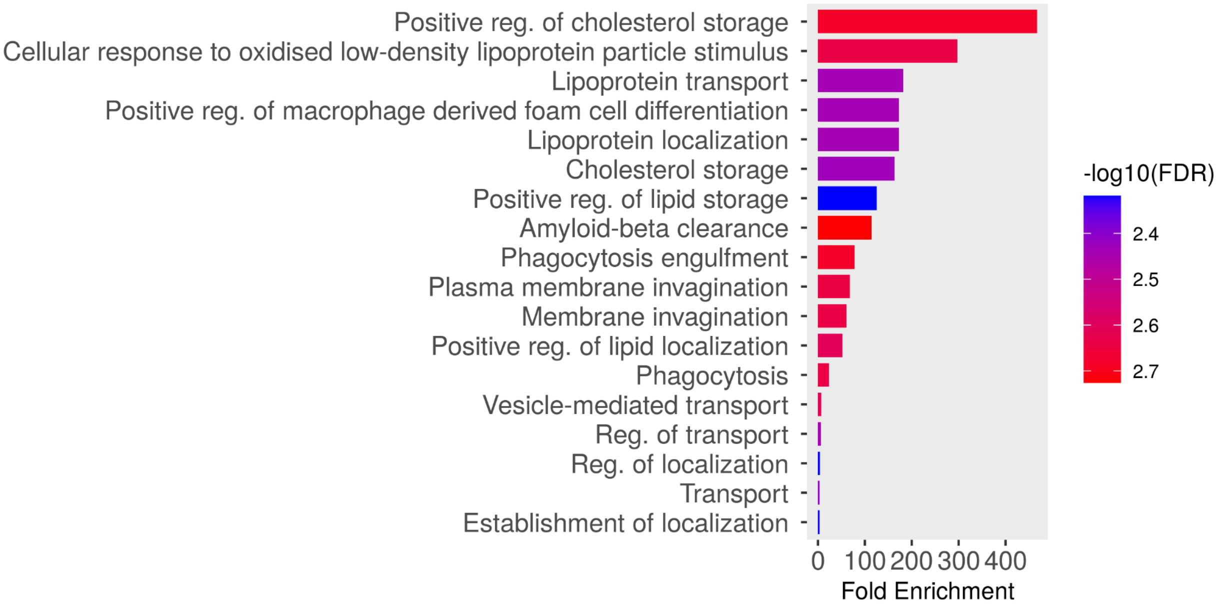
KEGG Pathway Enrichment of Tumor Samples



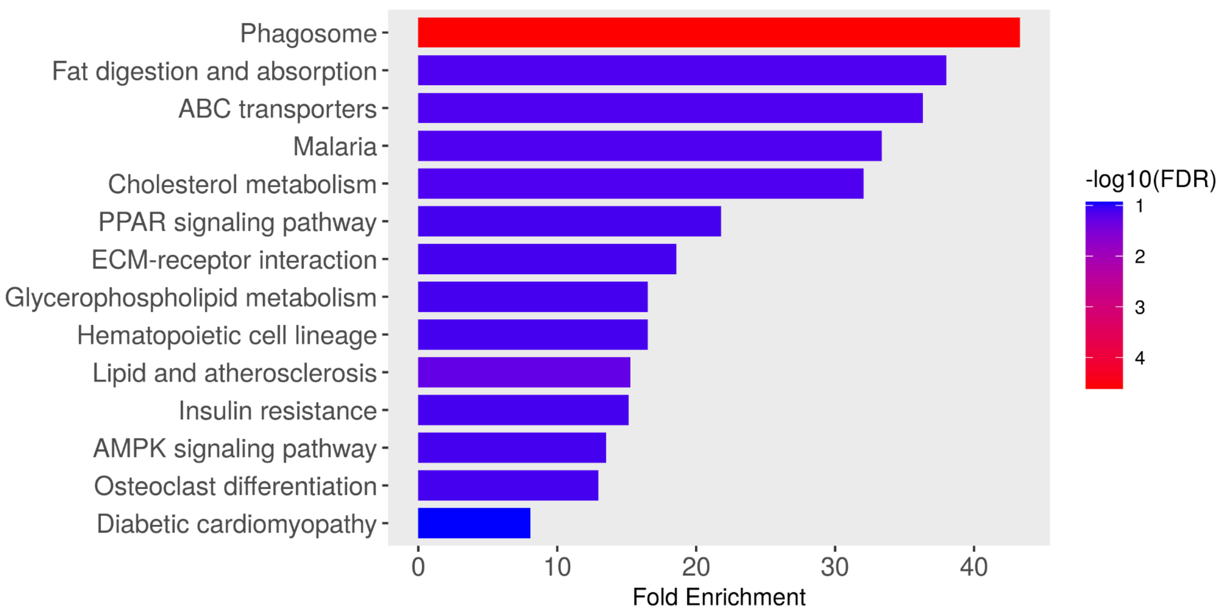
KEGG Pathway Enrichment of Normal Samples

2. Statistical Significance

- **False Discovery Rate (FDR)**: Key processes like *Positive regulation of cholesterol storage* and *Cellular response to oxidized LDL* had highly significant FDR values (-log10(FDR) ≈ 2.5–2.7; raw FDR ≈ 0.002–0.004).
- **Fold Enrichment**: Lipid-related processes (e.g., *Lipoprotein transport*, *Cholesterol storage*) exhibited exceptionally high fold enrichment (up to 400×), underscoring their critical role in the disease.



GO Biological Process Enrichment of the Tumor Samples



GO Biological Process Enrichment of the Normal Samples

3. Machine Learning Insights (SHAP Values)

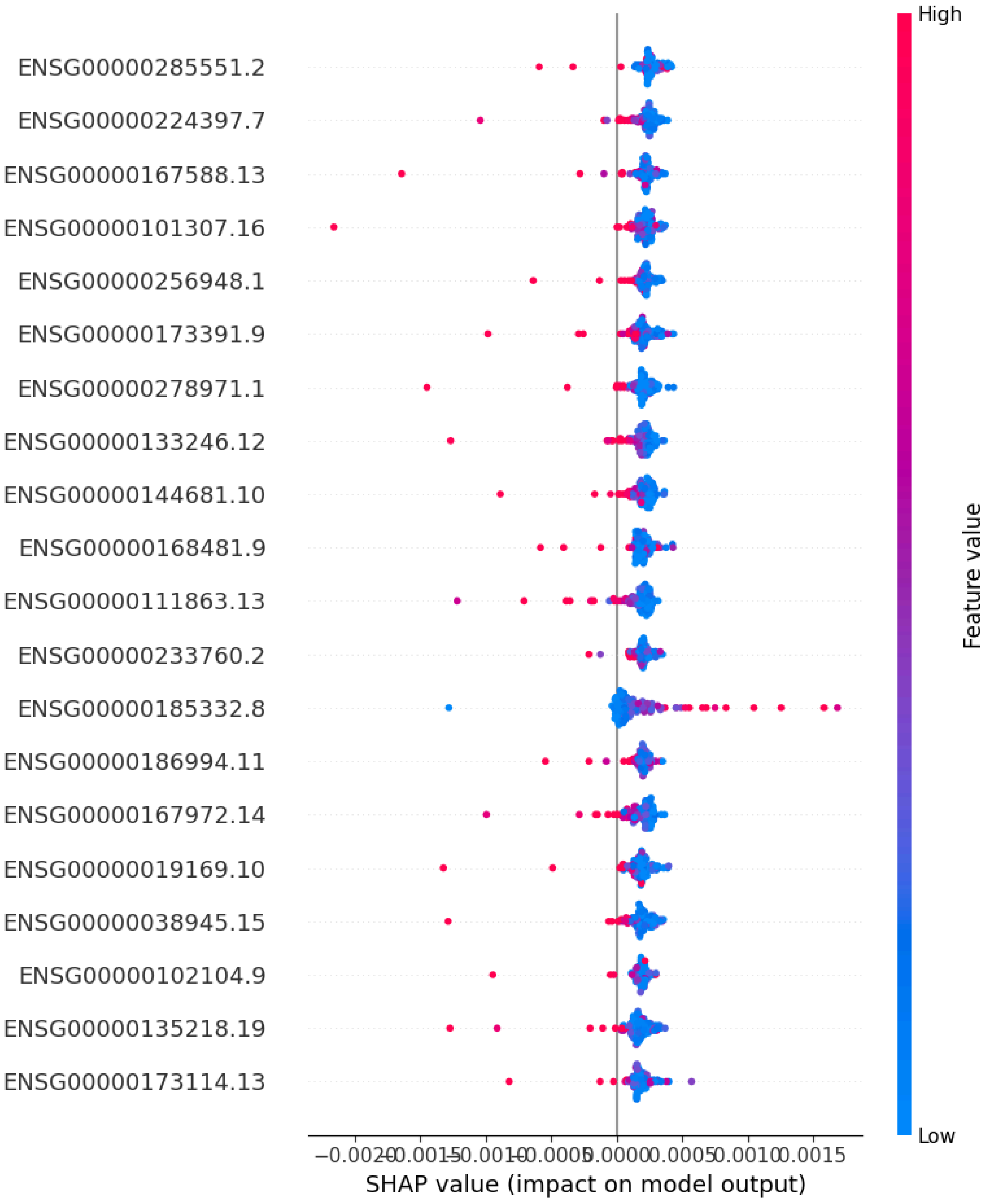
Machine learning-based feature importance analysis, particularly using **SHAP (SHapley Additive exPlanations) values**, provided key insights into the predictive role of individual genes versus collective pathway-level effects in lung adenocarcinoma (LUAD).

1. Gene-Level Impact and SHAP Analysis

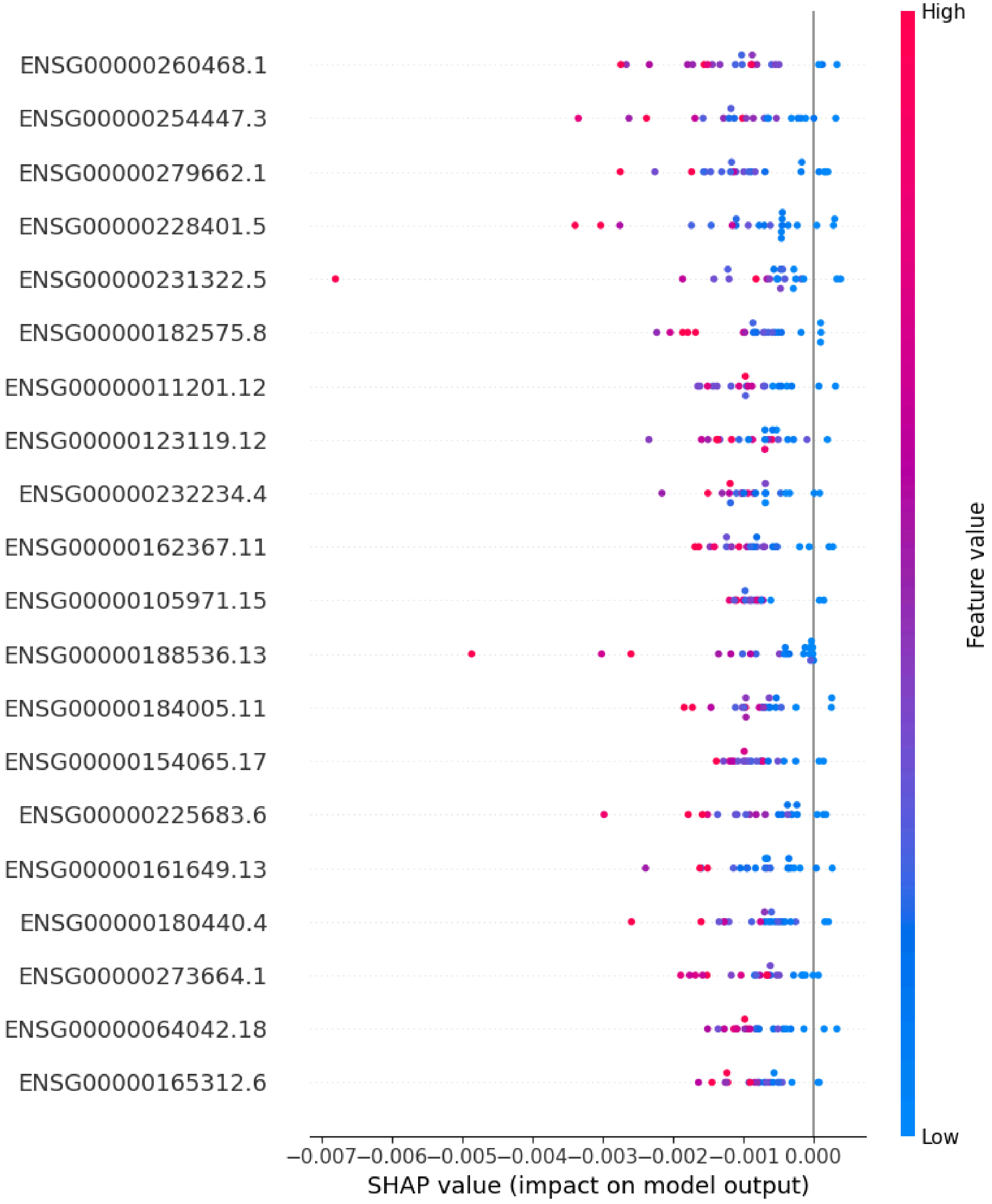
- Most genes, including **ENSG00000260468.1** and **ENSG00000173391.9**, exhibited **SHAP values close to zero**, indicating that they contributed minimally to the model's predictions when considered **individually**.
- The low SHAP values suggest that no single gene served as a **strong standalone biomarker** for LUAD classification or prognosis.
- This aligns with the growing understanding that **disease phenotypes often arise from complex gene interactions**, rather than isolated genetic variations.

2. Interpretation and Biological Implications

- While individual genes may not exert significant predictive influence, their collective **pathway-level activity** was found to be highly significant.
- This highlights the importance of **systems biology approaches** that account for **gene-gene interactions, regulatory networks, and pathway crosstalk** rather than focusing solely on single-gene alterations.
- Many diseases, including LUAD, are driven by **dysregulated pathways**, meaning that gene contributions should be analyzed within their **functional and contextual framework** rather than as independent entities.



SHAP Values of the Tumor Dataset



SHAP Values of the Normal Dataset

4. Limitations and Considerations

While this study provides valuable insights into the role of **lipid metabolism** and **immune dysregulation** in lung adenocarcinoma (LUAD), several limitations should be acknowledged:

1. **Small Gene Sets in Certain Pathways**
   - Some pathways, such as **Diabetic Cardiomyopathy**, contained only **1–4 genes**, which may **limit statistical power and robustness**.
   - Pathway enrichment analyses often rely on sufficient gene representation to derive meaningful conclusions; low gene counts could lead to **false negatives or underestimation of pathway significance**.
   - Future studies should integrate **larger datasets or complementary functional annotations** to improve confidence in pathway-level associations.
2. **Weak Gene-Level Predictive Signals**
   - The presence of **near-zero SHAP values** for certain individual genes suggests that their predictive contribution is limited when considered in isolation.
   - This underscores the importance of **examining combinatorial gene effects**, as disease mechanisms often arise from complex gene interactions rather than single-gene alterations.
   - Incorporating **pathway-based machine learning (ML) models**, such as **graph neural networks, multi-omics integration, or systems biology approaches**, could enhance predictive accuracy by leveraging **biological context and interactions**.
3. **Potential Biases and Model Assumptions**
   - The study's findings are contingent on the **quality and completeness of available datasets**, which may introduce biases related to **sample selection, sequencing depth, or annotation discrepancies**.
   - Machine learning models, including SHAP-based interpretability frameworks, rely on **specific algorithmic assumptions** that might not fully capture **nonlinear or context-dependent regulatory interactions**.
   - Further validation using **independent datasets, functional experiments, and alternative computational methods** is necessary to **reinforce biological significance**.

VII. CONCLUSION

This study highlights **lipid/cholesterol metabolism** and **immune dysregulation** as central mechanisms underlying lung adenocarcinoma (LUAD) progression. While individual genes exhibit limited predictive power in isolation, their collective activity within specific biological pathways demonstrates **strong functional relevance**. The findings underscore the importance of **system-level analysis** in uncovering potential therapeutic targets and refining our understanding of LUAD pathogenesis.

To build upon these insights, future research should focus on:

1. **Experimental Validation of Lipid Metabolism Targets:**

    - Investigating key regulators such as **PPAR signaling, cholesterol homeostasis, and lipid storage pathways** to determine their direct impact on LUAD progression.
    - Exploring **pharmacological interventions** that target dysregulated lipid metabolism, assessing their potential for clinical translation.

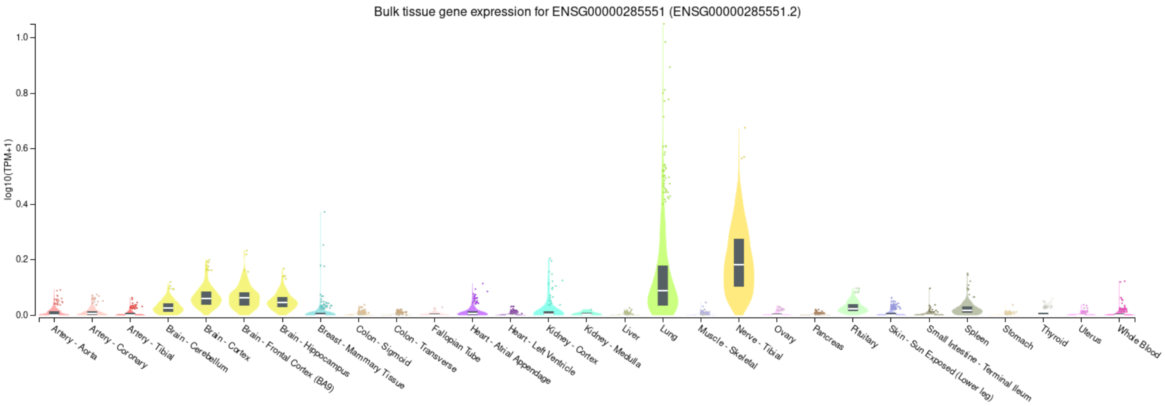
2. **Integration of Pathway-Centric Machine Learning Models:**

    - Leveraging **multi-omics data** (transcriptomics, metabolomics, and proteomics) to develop more comprehensive predictive models.
    - Implementing **deep learning and network-based approaches** to capture the intricate interactions within dysregulated pathways.

3. **Implications for Precision Medicine and Targeted Therapies:**

    - Identifying **biomarkers** linked to lipid metabolism and immune modulation for **early detection and prognosis**.
    - Evaluating the synergy between **immunotherapy and lipid-modulating drugs** to enhance treatment efficacy.

Cross Validation Of The Result



Based on the SHAP value plot, the top-ranked gene (ENSG000028551.2) shows the highest impact on model predictions. Cross-validation with the GTEx Portal indicates that this gene is most highly expressed in lung tissue. This suggests a potential lung-specific role for the gene in the studied biological context, potentially linked to tumor-related features in lung cancer or respiratory diseases.

The violin plot in the first image also supports the high expression variability of different genes, highlighting how expression levels vary across conditions. The overlap of SHAP values and GTEx expression data reinforces the biological relevance of this gene in lung tissue, making it a strong candidate for further investigation in disease modeling and biomarker identification.

Key Takeaway

Dysregulated **lipid homeostasis** and **immune pathways** play a **critical role** in LUAD pathophysiology, making them promising candidates for **novel therapeutic strategies**. A pathway-focused approach, integrating **experimental validation and computational modeling**, will be essential in translating these findings into **clinically actionable interventions**.