Biomarker Discovery for Lung Adenocarcinoma Using the CPTAC-LUAD Multi-omics Dataset

Aishwarya M, Sandhya Elumalai, S Varsha

Department of Biotechnology and Bioinformatics, JSSAHER

Abstract

Lung adenocarcinoma (LUAD) is the most common type of lung cancer and a leading cause of cancer-related deaths worldwide, with risk factors such as smoking and air pollution. In this study, we analyzed a LUAD dataset across transcriptomic, proteomic, and genomic levels to uncover potential biomarkers. In the transcriptomic analysis, RNA-seq data from tumor and normal samples were examined to find differentially expressed genes, and LASSO regression helped narrow down the key biomarkers, with ROC analysis showing high predictive accuracy (AUC > 0.90). For the genomic layer, the top genes were selected using basic ranking of mutation, methylation, and SCNV data. In the proteomic layer, t-tests identified proteins that were significantly altered. By integrating all three omics layers, we highlighted genes that were consistently altered across different molecular levels, resulting in a robust list of candidate biomarkers for early detection, prevention, and potential therapeutic targeting in LUAD. This study shows how a multi-omics approach can provide a more complete understanding of lung adenocarcinoma and aid in biomarker discovery.

Index Terms

Lung Adenocarcinoma, Genomics, Proteomics, Transcriptomics, Multi-Omics, RNA-seq, Differential Expression, Mutation Analysis, Methylation, SCNV, T-test, Biomarkers, LASSO Regression, ROC Analysis, Early Detection, Candidate Genes, Integrative Analysis

I. INTRODUCTION

Transcriptomics is the study of RNA molecules that show which genes are active in a cell, tissue, or organism. It allows us to see how gene activity changes under different conditions. With the growth of RNA sequencing (RNA-seq), it is now possible to measure gene expression on a large scale with good accuracy. Genomics studies all the genes in an organism to understand how DNA changes or mutations can influence disease. Proteomics looks at all the proteins in a cell or tissue to see how their levels and interactions change in health and disease. Multi-omics looks at different types of molecular data together like genes, RNA, and proteins to better understand how diseases work and to find potential biomarkers or targets for treatment. Bioinformatics helps identify and validate biomarkers by analyzing large biological datasets. It can also assist in designing biomarkers and applying them in real-time for early detection, prevention, and monitoring of diseases.

II. METHODOLGY

A. Data Collection

The CPTAC_LUAD dataset, comprising 110 samples, was retrieved from the LinkedOmics portal. This dataset included 14 files: RNAseq (Tumor), RNAseq (Normal), Proteome (Tumor), Proteome (Normal), Phosphoproteome (Tumor), Phosphoproteome (Normal), Acetylproteome (Tumor), Acetylproteome (Normal), Methylation (Tumor), Methylation (Normal), SCNV_LR (Tumor), SCNV_MZD (Tumor), Mutation (Tumor), and the Clinical data file.

B. Genomic Analysis

The genomic data in the CPTAC_LUAD dataset comprised Methylation, Mutation, and SCNV files. These files were loaded into dataframes and analyzed using Python.

Mutation Analysis

The Mutation file was filtered based on the frequency of mutations in each gene across all samples.

Methylation Analysis

The methylation tumor and normal files were filtered based on the mean methylation value for each gene across all samples. The delta methylation difference between tumor mean and normal mean was then calculated to classify genes as hypermethylated or hypomethylated, using a threshold of 0.2. The top 10 genes based on each category were selected.

SCNV Analysis

The SCNV file was filtered based on the mean value for each gene across all samples. Genes were classified as amplified or deleted based on a threshold of 0.05. This yielded the top 10 potential biomarkers based on mutation, methylation, and SCNV values.

Clinical Data Analysis

Using the clinical data, the same analyses were repeated to identify stage-wise top 10 genes. The mutation file was grouped by stage, and the mutation rate percentage was calculated by dividing the frequency of mutation by the number of samples in each stage. The top 10 genes per stage were then retrieved.

Similarly, the methylation files (already classified as hyper- or hypomethylated) were grouped stage-wise, and the top 10 genes for each stage and category were selected. The SCNV data was also grouped by stage to identify the top genes.

Most analyses were performed based on ranking; statistical significance tests were not conducted.

C. Transcriptomic Analysis

Data collection

RNA-seq count data were obtained for 110 lung tumor samples and 101 normal lung samples. The data were stored in tab-delimited text files and imported into R for analysis.

Data processing

Two datasets (tumor and normal) were combined into one count matrix, with genes as rows and samples as columns. A metadata table was created to label each sample as either “Tumor” or “Normal”. This information was required for the statistical analysis.

Differential Expression Analysis

Differential gene expression was analyzed using the **DESeq2** package (Love et al., 2014). The software normalizes the raw count data and applies statistical models to detect genes that are significantly upregulated or downregulated between tumor and normal samples. Genes were ranked based on adjusted *p*-values (False Discovery Rate, FDR) and log2 fold-change.

Feature Selection with LASSO

To reduce the large list of DEGs into smaller biomarker set, we applied LASSO logistic regression using the glmnet package (Friedman et al., 2010). LASSO applies a penalty to shrink regression coefficients, forcing some of them to zero, which effectively selects only the most informative genes. Cross-validation was performed to find the best penalty value (λ).

Model Validation with ROC Curve

The predictive performance of the selected biomarkers was tested using ROC analysis with the pROC package (Robin et al., 2011). Predicted probabilities from the LASSO model were compared to the true sample labels (Tumor or Normal).

The ROC curve was plotted, and the Area Under the Curve (AUC) was calculated to measure accuracy. An AUC value close to 1.0 indicates excellent classification performance.

D. Proteomic Analysis

Data collection

The tumor and normal samples of Lung adenocarcinoma were obtained, and the sample count of the data was 110 and 101 respectively. The data was stored in tab-delimited text files and were uploaded in Google Colab.

Data processing

The two datasets (tumor and normal) contents were read and merged where gene name is in row and sample is in column. The created meta data was normalized and filtered under threshold 0.5. The normalized metadata was used for further analysis.

T-test and Log2FC analysis

The Log2FC analyzed the metadata which helped to measure the protein expression change between to samples and the t test helps to statistically validate the difference between the sample, and the p values helps to numerically understand the differences.

The adjusted p value is used to find potential biomarkers from the sample. It uses a threshold of 0.05 and log2Fc.

Biomarker validation

The biomarkers were validated among known database Human Protein Atlas (HPA). It filtered the biomarkers, considering PTK7. It provided information on gene, protein class, biological information, molecular functions and as so on.

E. Multi Omics Analysis

The top genes across genomic, transcriptomic, and proteomic data were merged to perform an integrated multi-omics analysis. This approach allowed us to identify genes consistently altered across different molecular layers, highlighting strong candidate biomarkers for lung adenocarcinoma.

III. RESULTS

Genomic Level

The top genes were selected from the mutation, methylation, and SCNV data. Mutated genes were used to identify potential driver mutations that may contribute to the disease. Genes with normal methylation levels indicated normal epigenetic regulation, while hypermethylated or hypomethylated genes were considered as possible biomarkers.

Genes with normal copy numbers served as a reference, and genes that were amplified or deleted were identified as potentially affecting gene expression and contributing to tumor development. By comparing genes altered in mutation, methylation, and SCNV data, we identified those consistently changed across all layers, highlighting them as strong candidate biomarkers.

Transcriptomic Level

1. *Differential Expression Gene Analysis*

Using DESeq2, many genes were identified as differentially expressed between tumor and normal samples. Many genes showed high fold changes, and several were statistically significant after adjustment for multiple testing. These genes may represent key molecular changes involved in lung cancer.

2. *LASSO Feature Selection*

The initial DEG list was too large to be used directly as biomarkers. The LASSO regression model reduced the set to a smaller number of genes that had the strongest predictive power. The cross-validation curve showed that the model performed well with a limited number of genes, confirming that feature selection was effective.

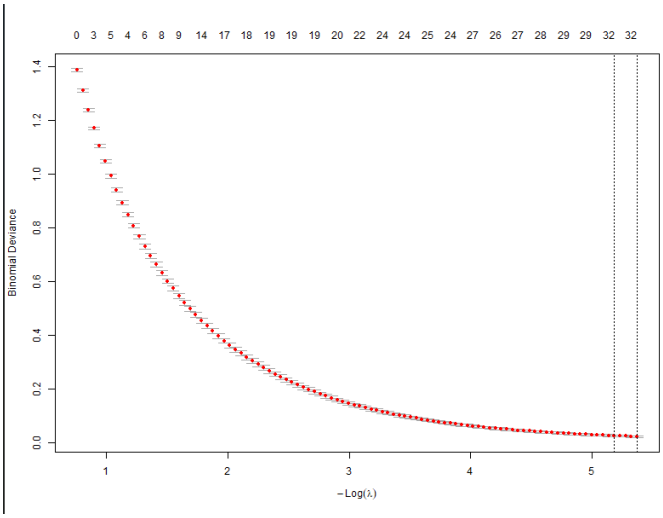

Figure 1: LASSO Feature Selection

3. *ROC Validation*

The ROC curve generated from the LASSO model showed excellent classification of tumor versus normal samples, with an AUC greater than 0.9. this indicates that the selected genes have strong potential as biomarkers for distinguishing diseased from healthy samples.

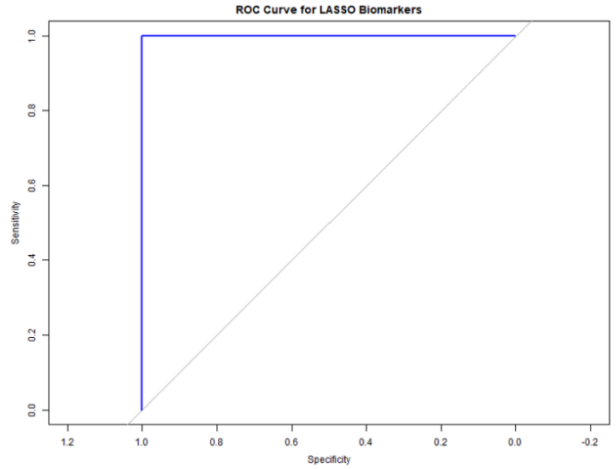

Figure 2: ROC Validation

Proteome Level

With the help of t test, log2fc there were lots of biomarkers obtained and with the help of adjusted p value top 10 differentially expressed gene were listed among them and they were validated with the help of known database. The obtained biomarkers are biologically relevant and must be studied for drug target or mechanistic research.

Gene Symbol	log2FC	adj_p_value
CAVIN2	-0.280714896	2.39E-89
EHD2	-0.25909313	2.58E-88
PALM2-AKAP2	-0.186832919	4.92E-88
SHANK3	-0.17199076	6.90E-86
CAVIN1	-0.252682832	2.29E-85
HSPA12B	-0.231059535	4.63E-84
RASIP1	-0.192001932	2.75E-82
CLIC5	-0.283715784	6.65E-82
STARD13	-0.114131436	1.17E-81
KANK3	-0.22655295	1.54E-81

Table 1: Discovered Potential Biomarkers (Proteome Level)

IV. CONCLUSION

This study identified potential biomarkers across multiple omics levels, but further analysis is still needed. When we integrated the data, no common genes were found across all layers, suggesting that the pipeline should be more robust and ideally integrate transcriptomic, genomic, and proteomic data from the start. In our approach, each omics layer was analyzed separately before being combined. To validate these biomarkers, it will be important to study the associated gene pathways and check if they play a role in cancer-related functions, which can help filter the most relevant genes. Overall, this study is a preliminary investigation, and the approach can be improved for more comprehensive biomarker discovery. In the future, these results should be tested on larger datasets to increase reliability, and the selected genes should be experimentally validated. Combining transcriptomics with genomics and proteomics could provide a clearer understanding of lung adenocarcinoma and aid in discovering useful biomarkers for diagnosis and treatment

SUPPLEMENTARY DATA

All Data related to the paper can be found here: Dataset_Drive_link

ACKNOWLEDGMENT

We sincerely thank the Department of Biotechnology and Bioinformatics, JSSAHER, Mysuru for providing us the opportunity, support, and guidance throughout this hackathon.

DISCOVERED BIOMARKERS						
Genomics					Transcriptomics	Proteomics
Mutation	Hypermethylated	Hypomethylated	SCNV_Amplified	SCNV_Deleted	Transcriptome	Proteome
TP53	TIGD1	IFI6	MBIP	RBMY1E	SAMD11	CAVIN2
MUC16	POLR1C	TIGD5	SFTA3	RBMY1D	ISG15	EHD2
EGFR	RPUSD4	REG1B	NKX2-1	PRORY	XKR8	PALM2-AKAP2
RYR2	ADAP1	HEATR9	NKX2-8	RBMY1F	TTF2	SHANK3
TTN	FOXD4L3	ELF3	PAX9	RBMY1J	EFNA3	CAVIN1
CSMD3	PRR15	OR10J3	EGFR	EIF1AY	IGFN1	HSPA12B
LRP1B	HOXD9	DEFB107A	LANCL2	RPS4Y2	C1orf168	RASIP1
KRAS	POU4F2	-	TERT	RBMY1B	AMY1B	CLIC5
USH2A	OLIG3	-	MYC	RBMY1A1	MSH2	STARD13
ZFHX4	PHOX2B	-	SEC61G	KDM5D	TRAK2	KANK3

Table 2: Discovered Potential Biomarkers across omics levels

We also extend our gratitude to the organizers of the Bio-Hackathon on "Computational Drug Development", including the Indian Institute of Technology, Guwahati, and the University of Engineering and Management, Kolkata, for successfully organizing this event and giving us the chance to participate and apply our skills.

REFERENCES

1. Suhas V Vasaikar, Peter Straub, Jing Wang, Bing Zhang, LinkedOmics: analyzing multi-omics data within and across 32 cancer types, Nucleic Acids Research, Volume 46, Issue D1, 4 January 2018, Pages D956-D963. https://doi.org/10.1093/nar/gkx1090
2. V. Bhaskar, R. Kumar, M. R. Praharaj, S. Gandham, H. K. Maity, U. Sarkar, and B. Dey, "A bovine pulmosphere model and multiomics reveal early host response signature in tuberculosis," Communications Biology, vol. 8, no. 1, p. 559, 2025. doi: 10.1038/s42003-025-07883-6.
3. B. M. F. Nogueira, S. Krishnan, B. Barreto-Duarte, M. Araujo-Pereira, A. T. L. Queiroz, J. J. Ellner, P. Salgame, T. J. Scriba, T. R. Sterling, A. Gupta, and B. B. Andrade, "Diagnostic biomarkers for active tuberculosis: progress and challenges," EMBO Mol. Med., vol. 14, p. e14088, 2022. doi: 10.15252/emmm.202114088.
4. S. Qiu, Y. Cai, H. Yao, C. Lin, Y. Xie, S. Tang, and A. Zhang, "Small molecule metabolites: discovery of biomarkers and therapeutic targets," Signal Transduct. Target. Ther., vol. 8, p. 132, 2023. doi: 10.1038/s41392-023-01399-3.
5. P. Bachanová, A. Cheyne, C. Broderick, S. M. Newton, M. Levin, and M. Kaforou, "Comparative transcriptomic analysis of whole blood mycobacterial growth assays and tuberculosis patients' blood RNA profiles," Sci. Rep., vol. 12, p. 17684, 2022. doi: 10.1038/s41598-022-20409-y.

AUTHORS

First Author – Aishwarya M, M.Sc. Bioinformatics, JSSAHER, aishwaryamkumar9@gmail.com.
Second Author – Sandhya Elumalai, M.Sc. Bioinformatics, JSSAHER, sandhyaelumalai76@gmail.com.
Third Author – S Varsha, M.Sc. Bioinformatics, JSSAHER, varshaadm2020@gmail.com.