American Journal of Applied Bio-Technology Research
(AJABTR)
Vol. 5, July 2026, pp. 26-39 DOI: doi.org/10.15864/ajabtr.514

Research Article

Biomarker Discovery for Lung Adenocarcinoma Using CPTAC-LUAD Multi-Omics Dataset

[1]Aishwarya M, [2]Sandhya Elumalai, [3]S Varsha*

[1-3]Department of Biotechnology and Bioinformatics, JSS Academy of Higher Education and Research, Mysuru - 570015

**Corresponding - sandhyaelumalai76@gmail.com*

Received on 30th September 2025
Revision submitted on February 2026
Published on 28th July 2026

Abstract

An integrative bioinformatics analysis of multi-omics data can help identify reliable biomarkers for lung adenocarcinoma (LUAD). In this study, CPTAC-LUAD datasets were analyzed at genomic, transcriptomic, and proteomic levels to identify potential diagnostic biomarkers. At the genomic level, hypermethylated, hypomethylated, amplified, deleted, and mutated genes were identified. Transcriptomic analysis was performed to detect differentially expressed genes, followed by LASSO-based feature selection to identify key candidate biomarkers. At the proteomic level, differential protein expression and pathway enrichment analyses were carried out to identify significantly altered proteins and associated biological pathways. The biomarkers identified from each omics layer were then integrated to identify potential diagnostic biomarkers associated with LUAD detection.

Keywords: Lung Adenocarcinoma, Genomic, Proteomic, Transcriptomic, Multi-omics, RNA-seq, Differential Expression, Mutation, Methylation, SCNV, Biomarkers, LASSO Regression, Functional Enrichment analysis, ROC Analysis.

Introduction

Lung cancer is the leading cause of cancer-related deaths worldwide, with around 2.5 million new cases and more than 1.8 million deaths reported in 2022. Among the different types of lung cancer, lung adenocarcinoma (LUAD) is the most common subtype [1]. Although there have been improvements in diagnosis and treatment, many patients are still diagnosed at advanced stages, leading to poor survival rates. Smoking is the major risk factor for lung cancer, while air pollution

and environmental pollutants also contribute to the disease. Low-dose computed tomography (LDCT) screening can help detect lung cancer early, but its use is limited due to false-positive results, overdiagnosis, radiation exposure, and high cost. Therefore, identifying reliable molecular biomarkers for early diagnosis is crucial.

Cancer develops because of changes occurring at different biological levels, including genomic, transcriptomic, epigenomic, and proteomic levels. Genomic changes such as mutations and copy number variations can initiate cancer development. Transcriptomic and epigenetic changes affect gene expression and contribute to tumor growth and progression. Proteomic changes reflect the functional activity of cells and the pathways involved in cancer development and treatment response. Since each omics layer provides only partial information about tumor biology, combining multiple omics datasets gives a more complete understanding of the disease and helps identify more reliable biomarkers. In this study, we performed multi-omics biomarker discovery using the CPTAC-LUAD dataset to identify potential diagnostic biomarkers that can distinguish tumor tissues from normal tissues in lung adenocarcinoma.

Data Collection and Dataset Description

The multi omics dataset for LUAD was retrieved from the Linked Omics portal and comprised RNA-seq, Proteome, Methylation, Somatic Copy Number Variation (SCNV) and Mutation data generated by the Clinical Proteomic Tumor Analysis Consortium (CPTAC). Tumor and matched normal samples were available for the RNA-seq, proteome, and methylation datasets, while SCNV and mutation datasets contained only tumor samples.

Data Preprocessing

As the datasets obtained from the source database were already preprocessed and normalized, no additional normalization was performed. Only minimal preprocessing steps, including data type conversion and handling of missing values, were carried out before downstream analysis.

Table 1 CPTAC-LUAD Dataset Summary

Dataset	Unit	Samples	Attributes
RNA-seq (HiSeq, Tumor)	Expression (log2(FPKM))	110	18099
RNA-seq (HiSeq, Normal)	Expression (log2(FPKM)	101	18099
Proteome (TMT, Tumor)	Expression (log2ratio)	110	10316
Proteome (TMT, Normal)	Expression (log2ratio)	101	10316
Methylation (Tumor)	Mean beta value	109	16335
Methylation (Normal)	Mean beta value	97	16335
SCNV-LR (Tumor)	Copy Number Change (log ratio)	109	19267
Mutation (Tumor)	Mutation Status (Binary)	109	11859

For genomic data preprocessing, the methylation datasets contained less than 3% missing values and were retained for analysis while the SCNV dataset had no missing values. For transcriptomic data, missing values were removed, negative values were set to zero, and low-expression features (values less than 1) were filtered out to reduce noise and improve data quality. For proteomic data, proteins with more than 30% missing values were excluded from further analysis. The remaining missing values were imputed using the K-Nearest Neighbors (KNN) method to improve data completeness while preserving the underlying data structure.

Only samples common across all selected omics datasets were retained for further analysis to maintain consistency between different omics layers. After sample matching, a total of 96 LUAD samples were used for downstream analysis.

Genomic analysis

Genomic analyses were performed using the Mutation, Methylation, and SCNV datasets. The methylation dataset contained gene-level, mean beta values ranging from 0 (unmethylated) to 1 (fully methylated), indicating the level of DNA methylation. The SCNV dataset comprised gene-level copy number changes represented as log ratios, where positive values indicate gene amplification, and negative values indicate gene deletion relative to normal samples. The mutation dataset contained binary mutation status for the tumor samples, where 1 indicates the presence of a mutation in a gene and 0 indicates no mutation.

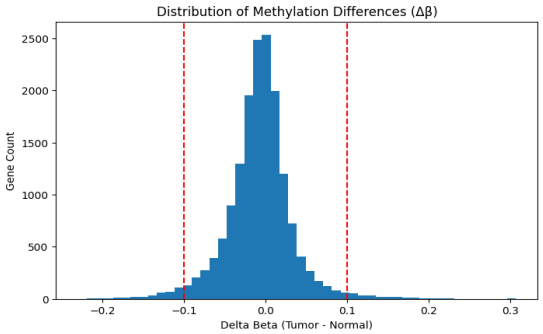
Mutation analysis

Mutation analysis was performed to identify frequently mutated genes. For each gene, the number of mutated samples across all samples was used to calculate mutation frequency and the percentage of mutated samples. Genes were then ranked in decreasing order of their mutation frequency.

Methylation analysis

Gene-level methylation analysis was performed by calculating the average beta value for each gene separately in tumor and normal samples. These values were merged using gene symbols to ensure each gene was correctly matched between the two groups. The difference in mean methylation (Tumor − Normal) was computed for each gene. Genes were classified as hypermethylated or hypomethylated using a predefined cutoff of ±0.1 in delta methylation.

Genes with delta methylation values greater than +0.1 were classified as hypermethylated, while those with values less than −0.1 were classified as hypomethylated. Values within this range were considered to show no significant change in methylation status. This threshold was selected based on the observed distribution of delta beta values (Fig. 1), where most genes clustered near zero, and



±0.1 was used to capture biologically meaningful deviations.

Fig. 1 Distribution of Methylation difference

Welch's t-test was used to compare methylation levels between tumor and normal samples for each gene. The resulting p-values were adjusted for multiple testing using the Benjamini–Hochberg false discovery rate (FDR) method. Genes with an adjusted p-value less than 0.05 and an absolute methylation difference greater than 0.1 were considered significantly differentially methylated and ranked based on adjusted p-value and magnitude of delta methylation.

Somatic Copy Number Variation analysis

Copy number variation (CNV) analysis was performed on tumor samples by calculating the mean CNV value for each gene across all samples. Mean CNV values exhibited a broader distribution with increased variability across genes, reflecting heterogeneous copy number alterations. To distinguish biologically meaningful gains and losses from minor fluctuations around the baseline, a threshold of ±0.2 was used to classify them. Genes with mean CNV values greater than +0.2 were classified as amplified, while those with values less than −0.2 were classified as deleted. Genes within this range were considered to show no significant copy number change. Amplified and deleted genes were then ranked separately based on their mean CNV values to prioritize the most strongly altered genes.

Transcriptomic analysis

RNA-seq analysis was performed using R software. The gene expression data was obtained as normalized expression values. For downstream analysis, tumor and normal samples were merged into a single

expression matrix, with genes arranged as rows and samples as columns. A metadata table was created to define sample groups (tumor and normal), enabling statistical comparison between the two groups

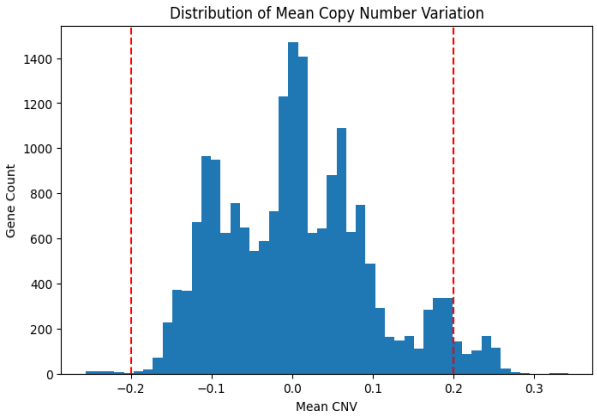


Fig. 2 Distribution of Mean Copy Number Variation

Differential Expression analysis

Differential gene expression was performed using the DESeq2 package. This package applies statistical models to identify genes that are significantly upregulated or downregulated between tumor and normal samples. Results were adjusted for multiple testing and genes with an adjusted p-value (FDR) less than 0.05 and an absolute log2 fold change greater than 1 was considered significantly differentially expressed.

Feature Selection with LASSO

The differentially expressed genes obtained from DESeq2 were further refined using LASSO (Least Absolute Shrinkage and Selection Operator) logistic regression implemented through the glmnet package (Friedman et al., 2010). LASSO applies a penalty to shrink regression coefficients, thereby selecting the most informative genes while reducing less important features. Cross-validation was performed to determine the optimal penalty parameter (lambda). The lambda.min criterion, corresponding to the minimum cross-validation error, was selected for the final model. Genes with non-zero coefficients at lambda.min were retained as the final set of candidate biomarkers. Prior to model fitting, variance stabilizing transformation (VST) from DESeq2 was applied, and feature standardization was performed automatically within glmnet.

Model validation with ROC curve

The predictive performance of the selected biomarkers was evaluated using Receiver Operating Characteristic (ROC) analysis with the pROC package (Robin et al., 2011). Predicted probabilities generated from the LASSO model were compared with the true sample labels (tumor and normal). The ROC curve was plotted to assess the classification performance of the model, and the Area Under the Curve (AUC) was calculated to measure model accuracy, with values closer to 1 indicating better predictive performance. Confidence intervals for the AUC were also computed to evaluate the reliability and robustness of the model.

Proteomic analysis

The proteomic data were provided as log2 expression ratio values. The tumor and normal datasets were merged using gene symbols as common identifiers to generate a unified expression matrix, with genes as

rows and samples as columns.

Differential Expression analysis

Differential expression analysis was performed to identify proteins exhibiting significant differences in abundance between tumor and normal samples. For each protein, expression levels in the two groups were compared using Welch's t-test, which accounts for unequal variances between the two conditions. The log2 fold change was calculated as the difference in mean protein abundance between tumor and normal samples, providing a measure of the magnitude and direction of differential expression.

To control for false positives arising from multiple hypothesis testing, p-values were adjusted using the Benjamini–Hochberg false discovery rate correction method. Potential biomarker candidates were defined as proteins with an adjusted p-value less than 0.05 and an absolute log2 fold change greater than 1, ensuring that selected proteins showed both statistical significance and biologically meaningful changes in expression.

Functional Enrichment analysis

Functional enrichment analysis was performed on the identified biomarker candidates using the GSEApy and Enrichr frameworks. Gene Ontology and KEGG pathway enrichment analyses were carried out to identify significantly enriched biological processes and pathways associated with the selected genes. Multiple testing correction was applied during the enrichment analysis to ensure that the identified enriched terms and pathways were statistically robust and not arising due to random chance.

Multi – Omics analysis

In the preceding sections, independent analyses were performed at the genomic, transcriptomic, and proteomic levels. These results were subsequently integrated to construct a unified multi-omics dataset at the gene level. This integrative framework enables the identification of genes exhibiting multi-layered alterations across different biological levels.

Gene	Mutation Count	Mutation Frequency (%)
TP53	51	53.13
EGFR	36	37.5
MUC16	35	36.46
TTN	34	35.42
RYR2	34	35.42
CSMD3	30	31.25
LRP1B	29	30.21
KRAS	27	28.13
USH2A	26	27.08
ZFHX4	24	25

From the genomic dataset, features including mutation frequency, copy number variation (CNV), and DNA methylation changes were extracted. From the transcriptomic dataset, log2 fold-change values of gene expression were used, and from the proteomic dataset, log2 fold-change values of protein expression were

included. All genes across the individual datasets were merged, and missing values were imputed with zero to maintain consistency across all variables.

Gene-level binary flags were created to represent alterations across different omics layers. A genomic flag was defined based on the presence of mutations or copy number changes such as amplification or deletion. An epigenomic flag was defined using DNA methylation changes, where either hypermethylation or hypomethylation was considered. A transcriptomic flag was defined based on significant gene expression changes, using an absolute RNA log2 fold-change greater than 1. Similarly, a proteomic flag was defined using an absolute protein log2 fold-change greater than 1. These flags were used to indicate whether a gene showed alterations in each omics layer for further integration analysis.

Gene	Delta Methylation	t-statistic	p-value	Adjusted p-value (FDR)
FOXD4L3	0.242968	2.26E+01	4.55E-52	7.44E-48
HOXD10	0.174331	2.06E+01	1.51E-43	1.23E-39
OLIG3	0.211016	2.20E+01	1.72E-41	9.37E-38
ADAP1	0.281479	1.95E+01	3.01E-41	1.23E-37
OTP	0.164047	1.80E+01	1.09E-37	2.22E-34
TBX15	0.161646	1.76E+01	1.36E-36	2.47E-33
SOX17	0.211262	1.80E+01	1.88E-36	3.07E-33
BARHL2	0.162038	1.66E+01	2.68E-36	3.80E-33
NKX6-1	0.137638	1.63E+01	2.79E-36	3.80E-33
PRDM13	0.137917	1.78E+01	3.96E-26	4.98E-33

Results

Genomic level

Mutation results

Table 2 Frequently Mutated genes

Mutation analysis identified several commonly altered genes in LUAD samples. TP53 had the highest mutation frequency (53.13%), followed by EGFR (37.5%), MUC16 (36.46%), TTN (35.42%), and RYR2 (35.42%). These genes are the most frequently mutated in the dataset and may play a role in tumor development and genomic instability.

Methylation results

Delta methylation analysis identified several genes that were significantly hypermethylated in LUAD. Genes such as FOXD4L3, HOXD10, OLIG3, and ADAP1 showed strong positive delta methylation values along with significant t-statistic and adjusted p-values.

Table 3 Top Hypermethylated Genes in LUAD samples

Table 4 Top

Gene	Mean CNV
MBIP	0.342843
SFTA3	0.335551
NKX2-1	0.332085
NKX2-8	0.329785
PAX9	0.323769
TERT	0.284744
RALGAPA1	0.281202
BRMS1L	0.281202
EGFR	0.280099
INSM2	0.277647

Hypomethylated Genes in LUAD samples

The t-statistic indicates the strength of difference in methylation relative to variation in the data. Higher t-values suggest more consistent and significant changes.

A distinct group of genes showed significant hypomethylation in LUAD. HEATR9, LGALS9C, SFN, IFI6, and SLC2A1 exhibited negative delta methylation values, indicating reduced methylation levels in tumor samples. This may suggest possible epigenetic activation or disruption of normal gene regulation in LUAD.

SCNV Results

Gene	Delta Methylation	t-statistic	p-value	Adjusted p-value (FDR)
HEATR9	-2.09E-01	-1.86E+01	8.71E-41	2.76E-37
LGALS9C	-1.44E-01	-1.71E+01	1.18E-39	2.76E-36
SFN	-1.75E-01	-1.63E+01	3.65E-35	2.84E-32
IFI6	-2.02E-01	-1.65E+01	4.74E-35	3.52E-32
OR10S1	-1.98E-01	-1.56E+01	1.32E-33	7.20E-31
ENC1	-1.88E-01	-1.44E+01	5.26E-32	2.26E-29
TIGD5	-2.11E-01	-1.54E+01	3.84E-31	1.43E-28
ALG1L	-1.72E-01	-1.46E+01	5.04E-31	1.79E-28
SLC2A1	-1.09E-01	-1.49E+01	1.19E-30	3.96E-28
CAPSL	-0.135692	-1.49E+01	1.29E-30	4.23E-28

Table 5 Top Amplified genes in LUAD samples

Gene	Mean CNV
RBMY1D	-0.256324
RBMY1E	-0.256324
PRORY	-0.255376
RBMY1F	-0.250754
RBMY1J	-0.250754
CDKN2A	-0.248376
CDKN2B	-0.248376
EIF1AY	-0.246608
RPS4Y2	-0.246608
KDM5D	-0.243318

Table 6 Top Deleted genes in LUAD samples

CNV analysis based on mean CNV values identified several genes with copy number gains. Genes such as MBIP, SFTA3, NKX2-1, EGFR, and TERT showed higher mean CNV values, suggesting they may contribute to tumor growth and cancer development.

In contrast, several tumor suppressor genes showed copy number loss. Genes such as CDKN2A, CDKN2B, KDM5D, and RPS4Y2 showed strong deletions, suggesting a possible loss of growth control and genomic stability.

Transcriptomic level

DESeq2 identified 2030 significantly expressed genes between tumor and normal samples. These genes showed notable fold change differences and statistical significance after multiple testing corrections, suggesting their possible involvement in molecular mechanism associated with lung adenocarcinoma.

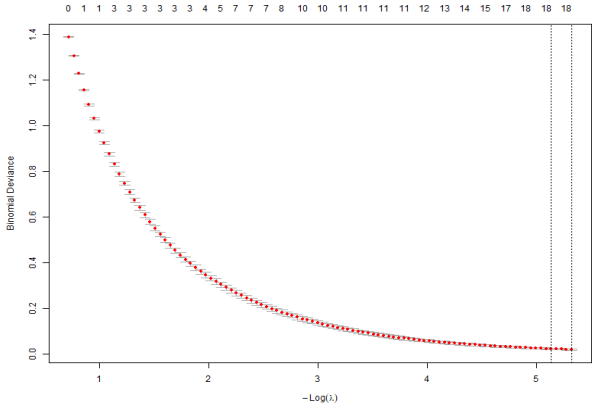


Fig. 3 Cross-validation plot showing optimal lambda selection for LASSO feature selection

The final LASSO model identified 17 candidate biomarker genes with potential to discriminate against LUAD samples from normal samples. The selected biomarkers included ARHGEF16, FAM136A, AGTR1, GPM6A, COL10A1, NECAB1, PKHD1L1, LGI3, PTPRQ, SGCG, GPT2, NSF, ETV4, B3GNT3, CBLC, FERMT1 and MMP11.

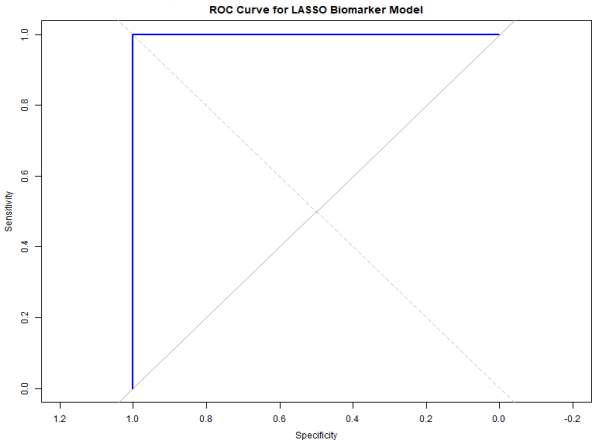


Fig. 4 ROC curve showing classification performance of the LASSO biomarker model

The ROC curve generated from the LASSO model showed good classification performance between tumor and normal samples with an AUC approaching 1. Cross-validation was performed to minimize overfitting and evaluate model robustness.

Proteomic level

ROC curve analysis was performed, and genes with an AUC ≥ 0.75 were shortlisted as potential biomarkers. These genes were further ranked based on their AUC values, and the top 10 genes with significant p-values and FDR-adjusted p values were selected.

Gene	AUC	p value	Adjusted p value
SHANK3	1	1.26E-86	8.52E-84
CLIC5	0.999783	4.31E-83	1.97E-80
ERG	0.999783	2.14E-74	2.40E-72
CAVIN1	0.999783	1.23E-84	6.93E-82
EHD2	0.999674	2.89E-88	3.25E-85
PALM2-AKAP2	0.999674	1.62E-90	5.46E-87
EHD4	0.999566	4.16E-73	3.42E-71
CAVIN2	0.999566	4.34E-90	7.32E-87
RTKN2	0.999566	5.43E-79	1.41E-76
QKI	0.999566	2.72E-73	2.29E-71

Table 7 Protein Biomarker candidate genes

Volcano plot analysis

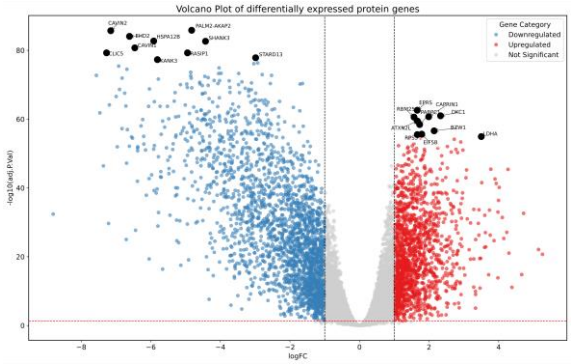


Fig. 5 Volcano plot of differentially expressed protein genes

The volcano plot showed several proteins with significant differences in expression between tumor and normal lung adenocarcinoma samples. Proteins such as CAVIN2, CAV1, EHD2, AGER, CDH5, and SHANK3 showed strong changes in expression along with low adjusted p-values. Among them, CAVIN2, CAV1,

and AGER showed the strongest differences, suggesting that they may serve as potential biomarkers.

Heatmap analysis

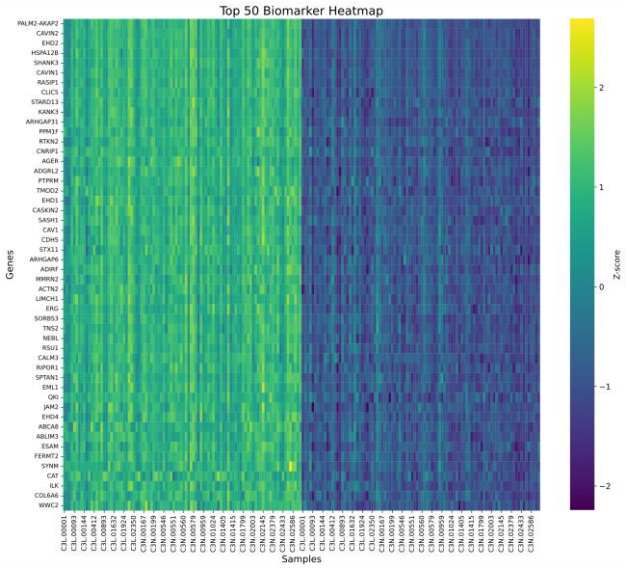


Fig. 6 Heatmap of top differentially expressed protein genes

The heatmap analysis demonstrated clear clustering and separation between tumor and normal lung adenocarcinoma samples based on their protein expression profiles. Proteins such as PALM2-AKAP2, CAVIN2, CAV1, SHANK3, and HSPA12B showed distinct variations in expression across the clustered samples, contributing to the observed separation pattern. These clustering patterns indicate consistent differences in proteomic profiles between tumor and normal conditions.

PCA (Principal Component) analysis

PCA analysis showed clear separation between tumor and normal samples based on global proteomic expression profiles. The tumor and normal samples formed distinct clusters mainly along PC1, which accounted for 25.35% of the total variance, while PC2 accounted for 5.07% of the variance. Normal samples showed tighter clustering, whereas tumor samples appeared more dispersed, indicating greater variability among tumor samples.

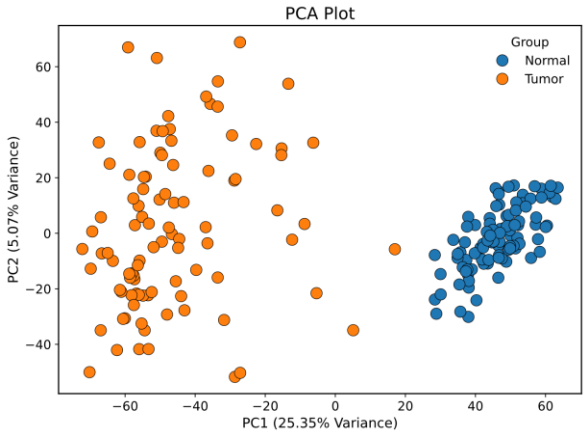


Fig. 7 Principal Component Analysis (PCA) of Proteomic Data

Gene Ontology Enrichment

Gene ontology enrichment analysis revealed that the identified proteins were mainly associated with protein synthesis and ribosomal functions. Most of the potential biomarkers were involved in translation, followed by ribosomal small subunit biogenesis. Enrichment of ribosomal biogenesis pathways suggests dysregulation of ribosome assembly, while enrichment of mitochondrial translation indicates altered protein synthesis activity in lung adenocarcinoma samples.

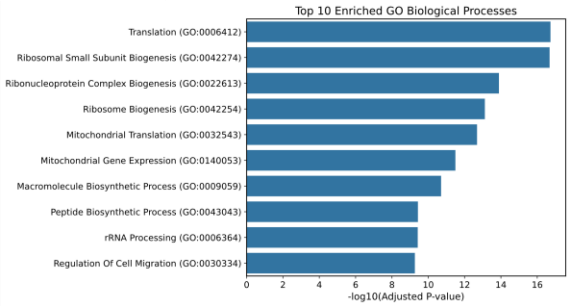


Fig. 8 Gene Ontology (GO) Enrichment Analysis of Protein Candidate Genes

Multi-Omics Results

In the multi-omics integration analysis, a total of 16,519 genes were combined from genomic, transcriptomic and proteomic datasets. Out of these, 369 genes showed changes in at least three omics layers, indicating that they are affected at multiple levels. Five genes (CCDC181, COMP, HSPB6, TBX3, and TBX4) show changes across multiple omics layers, including mutation, methylation, copy number variation, and gene/protein expression. COMP shows increased expression at both transcript and protein levels. HSPB6, TBX3, and TBX4 show decreased expression at both levels. CCDC181 shows reduced protein expression along with other molecular changes. Overall, these genes show consistent changes across different data types, suggesting they may be important in disease-related processes.

Discussion

In this study, a multi-omics approach was used to identify diagnostic biomarkers for lung adenocarcinoma (LUAD) by analyzing genomic, transcriptomic, and proteomic data. A simple scoring method was applied to integrate and rank genes based on multi-omics evidence; however, this approach is preliminary and requires further refinement using better integration methods. The study is limited by the absence of external validation and clinical data, and known LUAD-associated genes were not explored in detail.

Gene	Mutation Frequency	Mutation Percentage (%)	Delta Methylation (Hyper genes)	Delta Methylation (Hypo genes)	Mean CNV (Amplified genes)	Mean CNV (Deleted genes)	Log Fold change (Transcriptomics)	Log Fold Change (Proteomics)
CCDC181	0.0	0.0	0.10	0.0	0.20	0.0	-1.23	-2.83
COMP	2.0	2.08	0.10	0.0	0.00	0.0	3.12	2.45
HSPB6	3.0	3.12	0.19	0.0	0.00	0.0	-2.20	-4.88
TBX3	5.0	5.20	0.11	0.0	0.00	0.0	-1.69	-1.76
TBX4	3.0	3.12	0.12	0.0	0.00	0.0	-1.87	-3.84

Table 8 Genes altered across all Omics levels in LUAD samples

Conclusion

Multi-omics integration can help identify

potential diagnostic biomarkers for LUAD. However, the current approach is preliminary and needs further improvement and validation before clinical use.

Acknowledgment

We extend our gratitude to the organizers of Bio-Hackathon on "Computational Drug Development", Indian Institute of Technology, Guwahati and the University of Engineering and Management, Kolkata for organizing this event and providing us with the opportunity where we could learn and apply our skills.

References

1. Bray, F., Laversanne, M., Sung, H., Ferlay, J., Siegel, R. L., Soerjomataram, I., & Jemal, A. (2024). Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians, 74*(3), 229–263. https://doi.org/10.3322/caac.21834
2. Wang, Y., Zhou, Z., Chen, L., Li, Y., Zhou, Z., & Chu, X. (2021). Identification of key genes and biological pathways in lung adenocarcinoma via bioinformatics analysis. *Molecular and Cellular Biochemistry, 476*, 931–939. https://doi.org/10.1007/s11010-020-03931-6
3. Hsu, Y.-L., Hung, J.-Y., Lee, Y.-L., Chen, F.-W., Chang, K.-F., Chang, W.-A., Tsai, Y.-M., Chong, I.-W., & Kuo, P.-L. (2017). Identification of novel gene expression signature in lung adenocarcinoma by using next-generation sequencing data and bioinformatics analysis. *Oncotarget, 8*(62), 104831–104854. https://doi.org/10.18632/oncotarget.21022
4. Seo, J.-S., Ju, Y. S., Lee, W.-C., Shin, J.-Y., Lee, J. K., Bleazard, T., Lee, J., Jung, Y. J., Kim, J.-O., Shin, J.-Y., Yu, S.-B., Kim, J., Lee, E.-R., Kang, C.-H., Park, I.-K., Rhee, H., Lee, S.-H., Kim, J.-I., Kang, J.-H., & Kim, Y. T. (2012). The transcriptional landscape and mutational profile of lung adenocarcinoma. *Genome Research, 22*(11), 2109–2119. https://doi.org/10.1101/gr.145144.112
5. Vasaikar, S. V., Straub, P., Wang, J., & Zhang, B. (2018). LinkedOmics: Analyzing multi-omics data within and across 32 cancer types. *Nucleic Acids Research, 46*(D1), D956–D963. https://doi.org/10.1093/nar/gkx1090

Data Availability

The CPTAC-LUAD dataset analyzed in this study is publicly available through the LinkedOmics data portal: https://www.linkedomics.org/data_download/CPTAC-LUAD/.

All additional data supporting the findings of this study are available at: https://drive.google.com/drive/folders/1VuXWvsu0Mo5k3nB0pBr7DovCbgZl6iNd.