Pakistan Journal of Bioinformatics (PJB)
Volume 01, Issue 02, Year 2026 — ISSN: 2222-7628

Differential Gene Expression Analysis of Lung Cancer Using Public Data

Rubab Ahmad[1], Muqaddas Shehzadi[2]
Department of Computer Science, University of Agriculture, Faisalabad, Pakistan
[1]rubabahmad6156172@gmail.com, [2]edupak2004@gmail.com

Abstract

Lung adenocarcinoma (LUAD) could demonstrate distinct molecular alterations that appear to depend on the significant smoking status, though the critical transcriptomic differences might provide valuable opportunities for identifying new biomarkers. Moreover, lung cancer may account for 2.21 million cancer cases each year and appears to show the top cause of cancer-related deaths worldwide. Thus, tobacco smoking might indicate the biggest risk factor for lung cancer. However, cases also happen in non-smokers, especially women. Given that molecular processes separate lung adenocarcinomas in smokers from those in non-smokers, these mechanisms are not well understood. Comparative transcriptomic and bioinformatics analyses can systematically identify changes in gene expression and pathways that smoking disrupts. Nevertheless, identifying genes and signaling pathways that smoking affects may help understand disease progress and find treatment targets. Additionally, combining gene expression data with protein–protein interaction (PPI) networks and drug–target mapping can reveal chances for re-purposing drugs for targeted lung cancer treatment. Notwithstanding previous approaches, the study involves identifying differentially expressed genes (DEGs) between smoker and non-smoker lung adenocarcinoma samples using GEO2R. Moreover, the analysis might construct a protein–protein interaction (PPI) network and identify hub genes using STRING. Furthermore, pathway and functional enrichment analysis uses Enrichr. Thus, findings may give insight into molecular targets and repurpose a drug candidate, supporting development of diagnostic biomarkers and targeted therapies. However, bioinformatics integrative analysis might provide insights into smoking-associated molecular mechanisms in LUAD and adds to advancement of precision medicine strategies in lung cancer treatment.

Keywords: Lung adenocarcinoma, Non-small-cell lung carcinoma; differential gene expression; multiomics; protein–protein interactions, Bioinformatics analysis.

1 Introduction

Lung cancer may indicate one of the most significant causes of the important cancer-related morbidity and the substantial mortality worldwide. Moreover, among the various types, non-small cell lung cancer (NSCLC) could account for approximately 85 percent of all cases. Furthermore, NSCLC itself might encompass diverse histological subtypes, with adenocarcinoma and squamous cell carcinoma being predominant forms. However, Adenocarcinoma (LUAD) appears to become the most commonly diagnosed subtype, especially among non-smokers. Given that smoking history shows relevant patterns, squamous cell carcinoma (LUSC) traditionally correlates more with smoking history. Thus, clinical management and outcomes for NSCLC may hinge on early diagnosis and precise prognostication, as advanced-stage diagnosis typically portends poor survival despite advances in therapeutic options. These challenges create an urgent need for effective biomarkers that can help with early detection, subtype differentiation, and prognosis. Identifying molecular signatures linked to tumor behavior, patient survival, and treatment response could change how we provide care. Additionally, understanding the molecular landscape of lung cancer helps in sorting patients for targeted therapies, immunotherapy, and personalized medicine, which are now seen as standard in oncology practice. Therefore, detailed molecular profiling studies, especially those using high-throughput technologies, are essential for discovering new targets and clarifying the mechanisms that cause lung cancers [1], [2], [3]. Microarray technology has transformed genomic research in cancer by enabling the simultaneous evaluation of gene expression levels for thousands of genes in both tumor and normal samples. This technology utilizes hybridization principles to identify messenger RNA transcripts in samples, allowing researchers to pinpoint genes that are differentially expressed (DEGs), which reflect tumor biology or responses to treatment. Public databases such as the Gene Expression Omnibus (GEO) and The Cancer Genome Atlas (TCGA) have compiled a vast collection of datasets related to lung cancer, containing raw gene expression data obtained from bulk tissue microarrays and RNA sequencing. These publicly available datasets offer extensive, accessible resources for integrative analyses and validation studies, which are essential for generating hypotheses and facilitating clinical applications. To effectively analyze microarray data, various bioinformatics tools have been created. GEO2R enables swift identification of differentially expressed genes (DEGs) by comparing experimental groups directly. For more advanced analyses, resources like STRING assist in constructing protein-protein interaction networks. Together, these platforms allow for an in-depth understanding of complex gene expression profiles, helping to identify key molecular drivers of lung cancer [4], [5].Differential gene expression analysis was used to determine whether lung cancer

was upregulated or downregulated in both publicly available datasets and personal data.

2 Methodology

2.1 Data Acquisition

The first step in finding genomic changes in lung cancer is to carefully preprocess microarray data. This includes quality control, background correction, and normalization to cut down on noise and systematic variation. Analytical techniques like GEO2R facilitate differential expression analysis by contrasting tumor and normal tissue samples, utilizing moderated t-tests to determine statistical significance among genes. The Gene Expression Omnibus database provided the publicly available dataset GSE19804, which is part of transcriptomics. The GEO2R web tool was used to analyze this dataset. We used the define feature to label the samples in the dataset, and we checked the grouping before looking at the data. The Affymetrix microarray platform made the GSE19804 dataset, which has expression data from human lung tissue samples. The dataset mainly contains two different biological states: normal lung tissue and lung cancer tissue. The samples were prepared in a controlled laboratory environment to minimize technological discrepancies and facilitate effective data comparison between both states. Including both cancerous and non-cancerous samples in the same dataset makes it possible to find genes that have altered expression patterns linked to the development of lung cancer.

2.2 Differential Gene Expression Analysis Using GEO2R

We used GEO2R, an online tool from the GEO database, to do a differential gene expression analysis. GEO2R uses the limma package from the R/Bioconductor project. This package is very popular for looking at microarray gene expression data. After getting to the GSE19804 dataset, the user chose the option ”Analyze with GEO2R.” The GEO2R interface showed all of the samples, and they were manually sorted into two groups based on their biological condition:
Group 1: Samples that are normal
Group 2: Samples of lung cancer

Since GEO2R compares gene expression levels between specified biological conditions to find statistically significant differences, correctly grouping samples is an essential step.

2.3 Identification of Differentially Expressed Genes (DEGs)

Based on statistical significance, genes with differential expression were found. Genes that had an adjusted p-value of less than 0.05 were chosen as DEGs because they were deemed statistically significant. This cutoff is frequently employed in microarray research to guarantee accurate identification of genes with significant expression variations. For later protein interaction and functional enrichment analyses, the entire list of DEGs was downloaded from GEO2R. Differences between up-regulated and down-regulated genes are also displayed in [Fig. 1], [Fig. 2].

2.4 Protein–Protein Interaction (PPI) Network Construction

2.4.1 STRING Database

The STRING (Search Tool for the Retrieval of Interacting Genes/Proteins) database was used to analyze protein–protein interactions. STRING combines known and anticipated protein interactions from text mining, curated databases, computational prediction techniques, and experimental data.

2.4.2 Network Construction

The ”Multiple proteins” option was used to upload the identified DEGs to STRING, with Homo sapiens selected as the organism. STRING created an interaction network by mapping the gene symbols to the corresponding proteins. Proteins are represented by nodes in the PPI network, and functional or physical interactions between proteins are represented by edges.[Fig. 3] Confidence scores were used to guarantee the dependability of interactions while the network was displayed in network view.[6] text mining and databases.

2.5 Functional Enrichment Analysis

2.5.1 Gene Ontology (GO) Analysis

Gene Ontology enrichment analysis was performed using STRING to understand the biological significance of the DEGs. GO analysis groups gene functions into three main categories: Biological Process, Molecular Function, and Cellular Component. Significant GO terms were identified with an FDR value below 0.05. These pathways together play a role in tumor progression, metastasis, and immune regulation, highlighting their importance for therapy.

2.5.2 KEGG Pathway Analysis

KEGG (Kyoto Encyclopedia of Genes and Genomes) pathway analysis identified biological pathways that are significantly enriched in the DEG list. KEGG pathways offer insight into the molecular interactions and signaling processes involved in lung cancer development. [Fig.6], [9]

3 Results

3.1 Differentially Expressed Genes

Results obtained through differential expression analysis indicated a number of genes that have significant expression differences between lung cancer and normal lung tissue samples. Genes that had significant expression differences were selected based on the criterion that the adjusted p-value is less than 0.05. Genes that had significant expression differences included both down-regulated and up-regulated genes.

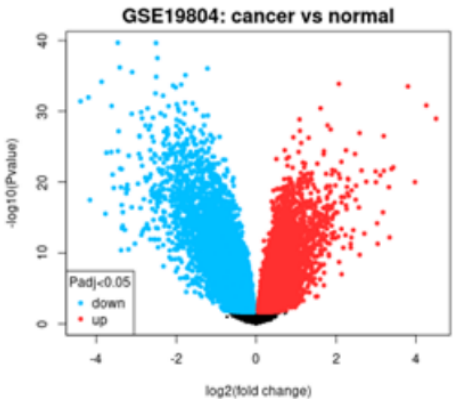

Figure 1: Cancer Vs normal

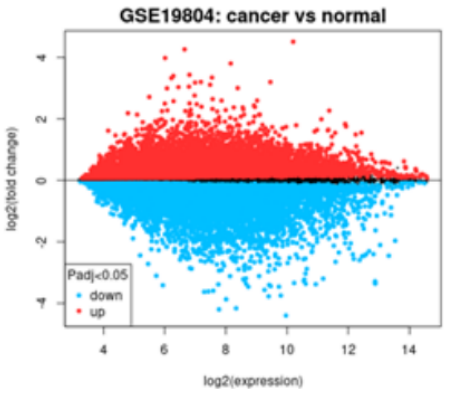

Figure 2: Cancer Vs normal

3.2 Protein–Protein Interaction Network Analysis

The PPI network obtained from the STRING tool showed intricate interactions among the identified DEGs. Many proteins showed highly interconnected clusters on the network map, pointing towards regulated biological processes related to cancer. Highly interconnected proteins can be hub genes acting as potential players in the development of lung cancer.

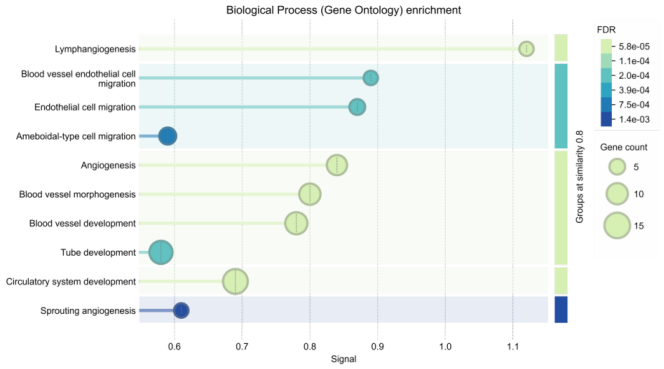

Figure 3: Network

3.3 Gene Ontology (GO) Enrichment Results

3.3.1 Biological Process

Analysis of the GO Biological Process indicated that the significant term was cell cycle regulation. Other significant biological processes were cell proliferation, apoptosis, and signal transduction. Cell cycle regulation is related to the development of cancer. Cell proliferation and apoptosis can be considered mechanisms engaged by the cell to control cell proliferation.

3.3.2 Molecular Function

The Gene Ontology Molecular Function analysis showed an enrichment for protein binding, ATP binding, and enzymatic activity. These functional categories imply different or altered molecular interactions and enzymatic processes in lung cancer cells.

3.3.3 Cellular Component

By performing the GO Cellular Component analysis, it was revealed that a significant portion of these DEGs were functioning in either the nucleus, cytoplasm, or plasma membrane.

Biological Process (Gene Ontology)					
GO-term	description	count in network	strength	signal	false discovery rate
GO:0035295	Tube development	12 of 880	0.73	0.58	0.0022
GO:0003008	System process	16 of 2029	0.49	0.36	0.0196
GO:0048731	System development	25 of 3867	0.41	0.41	0.0022
GO:0002040	Sprouting angiogenesis	4 of 58	1.43	0.61	0.0143
GO:0032501	Multicellular organismal process	35 of 6490	0.33	0.39	0.00059
GO:0001946	Lymphangiogenesis	4 of 13	2.08	1.12	0.00059
GO:0043542	Endothelial cell migration	5 of 69	1.46	0.87	0.0019
GO:0045446	Endothelial cell differentiation	4 of 82	1.28	0.44	0.0452
GO:0032502	Developmental process	28 of 5657	0.29	0.27	0.0362
GO:0072359	Circulatory system development	14 of 901	0.79	0.69	0.00059
GO:0048514	Blood vessel morphogenesis	10 of 419	0.97	0.8	0.00059
GO:0043534	Blood vessel endothelial cell migration	4 of 31	1.71	0.89	0.0022
GO:0001568	Blood vessel development	11 of 505	0.93	0.78	0.00059
GO:0001525	Angiogenesis	9 of 325	1.04	0.84	0.00059
GO:0001667	Ameboidal-type cell migration	6 of 197	1.08	0.59	0.0099

Figure 4: Biological pathway

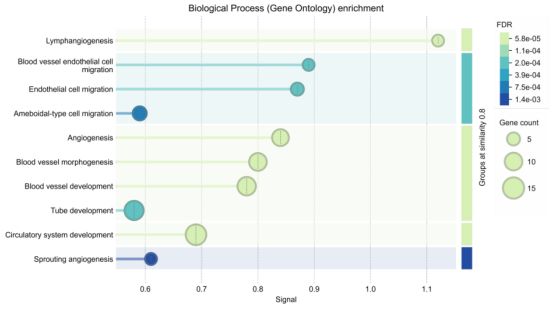

Figure 5: Gene ontology

3.4 KEGG Pathway Enrichment Results

KEGG pathway analysis demonstrated that several cancer-related pathways, including the pathways in cancer, PI3K-Akt signaling pathway, MAPK signaling pathway, and cell cycle pathway were significantly enriched among the DEGs. These pathways regulate cell survival, proliferation, and apoptosis.

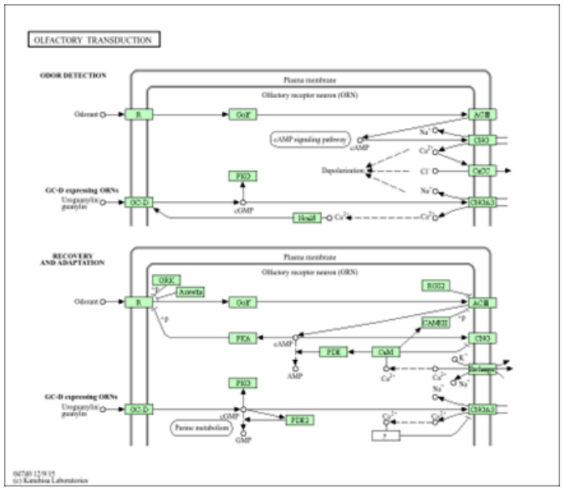

Figure 6: Transduction

4 Discussion

Three methods: differential expression analysis, protein-protein interaction network reconstruction, and functional enrichment analysis, are combined to provide insights into molecular mechanisms of lung cancer. Cancer-related biological processes/pathways enrichment in lung cancer can lead to understanding the significance of improper gene expression in lung cancer development.

5 Conclusion

In conclusion, the analysis of differential gene expression using public microarray resources is a task of inestimable value to the end-goal of finding the molecular basis of lung cancer. Ranging from the identification of core DEGs, pivotal biological pathways, to the creation of PPI networks and survival-prediction gene sets, the above analyses, among others, all bring together our comprehension of the pathogenesis of lung cancer, as it relates to precision approaches, including precision medicine initiatives

References

[1] E. Kettunen, S. Anttila, J. K. Seppänen, A. Karjalainen, H. Edgren, I. Lindström, et al., "Differentially expressed genes in nonsmall cell lung cancer: Expression profiling of cancer-related genes in squamous cell lung cancer," Cancer Genetics and Cytogenetics, vol. 149, no. 2, pp.

[2] Shriwash, N., Singh, P., Arora, S., Ali, S. M., Ali, S., & Dohare, R. (2019). Identification of differentially expressed genes in small and non-small cell lung cancer based on meta-analysis of mRNA. Heliyon, 5(6).

[3] Shriwash, Nitesh, Prithvi Singh, Shweta Arora, Syed Mansoor Ali, Sher Ali, and Ravins Dohare. "Identification of differentially expressed genes in small and non-small cell lung cancer based on meta-analysis of mRNA." Heliyon 5, no. 6 (2019).

[4] M. Liu, X. Yu, C. Qu, and S. Xu, "Predictive value of gene databases in discovering new biomarkers and new therapeutic targets in lung cancer," Medicina, vol. 59, no. 3, Mar. 2023, doi: 10.3390/medicina59030547.

[5] Wang, K., Chen, R., Feng, Z., Zhu, Y. M., Sun, X. X., Huang, W., & Chen, Z. N. (2019). Identification of differentially expressed genes in non-small cell lung cancer. Aging (Albany NY), 11(23), 11170.

[6] D. Szklarczyk, A. L. Gable, D. Lyon, *et al.*, "STRING v11: Protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets," *Nucleic Acids Research*, vol. 47, no. D1, pp. D607–D613, 2019, doi: 10.1093/nar/gky1131.

[7] M. Ashburner, C. A. Ball, J. A. Blake, *et al.*, "Gene ontology: Tool for the unification of biology," *Nature Genetics*, vol. 25, no. 1, pp. 25–29, 2000, doi: 10.1038/75556.

[8] McDoniels-Silvers, A. L., Nimri, C. F., Stoner, G. D., Lubet, R. A., & You, M. (2002). Differential gene expression in human lung adenocarcinomas and squamous cell carcinomas. Clinical cancer research, 8(4), 1127-1138.