Int. J. Pharm. Investigation, 2025; 15(4):1235-1248.
https://www.jpionline.org

Original Article

Multi-Dataset Identification and Validation of New Gene Expression Signatures: Insights into Matrix Remodeling Pathways in NSCLC

Rashed Mohammed Alghamdi*

Department of Laboratory Medicine, Faculty of Applied College, Al-Baha University, SAUDI ARABIA.

ABSTRACT

Background: Non-Small Cell Lung Cancer (NSCLC) epitomizes 85% of lung cancer cases with poor survival rates. Understanding molecular mechanisms through gene expression analysis is important for developing real treatments. This study intended to identify consistent molecular signatures in NSCLC using publicly available transcriptome data. **Materials and Methods:** We analyzed gene expression profiles from two independent NSCLC datasets (GSE33532 and GSE19188) from the Gene Expression Omnibus database. Differential gene expression examination was performed to identify consistently dysregulated genes across both datasets. Statistical validation included Pearson correlation analysis and significance testing to ensure result reliability. **Results:** Analysis revealed 53 consistently altered genes across datasets, comprising 28 upregulated (42.9%) and 25 downregulated (60%) genes, with exceptional correlation (r=0.9927). Upregulated genes included matrix remodeling factors (COL11A1, MMP12, MMP1) and cell proliferation markers (TOP2A), while downregulated genes included tissue-specific factors (CLDN18, AGER, SFTPC, SCGB1A1). These alterations indicate significant changes in extracellular matrix organization, cell proliferation and lung tissue homeostasis. Dataset-specific expressions (28.6% upregulated, 20% downregulated) reflected NSCLC's molecular heterogeneity. **Conclusion:** Our analysis identified reproducible gene expression signatures in NSCLC, providing insights into disease mechanisms and potential therapeutic targets. The strong correlation between datasets validates these molecular signatures' biological significance. These findings suggest multiple therapeutic approaches, including matrix remodeling inhibition and restoration of tissue-specific gene expression. While these results offer promising directions for NSCLC treatment, further functional validation studies will inherently add more to its clinical utility.

Keywords: Affymetrix, Correlation Coefficient, Gene Expression Omnibus, GEO2R, NSCLC Cancer, Venn Diagram, Volcano Plot.

Correspondence:
Dr. Rashed Mohammed Alghamdi
Department of Laboratory Medicine, Faculty of Applied College, Al-Baha University, Postal Code 1988, SAUDI ARABIA.
Email: rashed053660518@gmail.com

Received: 14-02-2025;
Revised: 09-04-2025;
Accepted: 30-06-2025.

INTRODUCTION

NSCLC represents the predominant form of lung cancer, accounting for 85% of all cases and remains the leading cause of cancer-related mortality worldwide. According to GLOBOCAN 2023 statistics, lung cancer affected 2.3 million individuals globally, with NSCLC claiming approximately 1.8 million lives (Bray *et al*., 2024). Despite significant therapeutic advances including targeted therapies and immunotherapies, the 5-year survival rate for advanced NSCLC remains below 25% (Travis *et al*., 2015). The molecular landscape of NSCLC has been extensively characterized, with the WHO Classification of Thoracic Tumours (2021) recognizing distinct histological subtypes: adenocarcinoma (40-50%), squamous cell carcinoma (20-30%), and large cell carcinoma (10-15%) (Hirsch *et al*., 2017). Each subtype exhibits unique molecular signatures, with adenocarcinomas frequently harboring mutations in EGFR (15-20% in Western populations, 40-60% in Asian populations), ALK rearrangements (5-7%) and KRAS mutations (25-30%) (Li *et al*., 2013). This molecular heterogeneity significantly influences therapeutic decisions and patient outcomes (Skoulidis F *et al*., 2019).

Recent advances in high-throughput sequencing technologies have revolutionized our understanding of NSCLC genomics and transcriptomics (Hugo *et al*., 2016). The Cancer Genome Atlas (TCGA) and other large-scale genomic studies have revealed complex molecular alterations driving NSCLC pathogenesis (Cancer Genome Atlas Research Network 2023). These findings have directly translated into the development of targeted therapies,

OPEN ACCESS

Manuscript TECHNOMEDIA

DOI: 10.5530/ijpi.20250268

Publishing Partner : Manuscript Technomedia. [www.mstechnomedia.com]

with remarkable success in treating specific molecular subtypes (Morganti *et al*., 2019). Gene expression profiling through microarray and RNA sequencing has emerged as a crucial tool for understanding NSCLC biology (Wang *et al*., 2023). The Gene Expression Omnibus (GEO) database, maintained by NCBI, hosts numerous NSCLC datasets that enable comprehensive transcriptomic analyses (Barrett *et al*., 2013). GEO2R, an interactive web tool, facilitates comparison of gene expression patterns across multiple datasets, helping identify consistently altered genes and pathways (Davis *et al*., 2007). Recent analyses of key GEO datasets (GSE31210, GSE33532, GSE19188) have provided valuable insights into NSCLC transcriptional networks (Ruan *et al*., 2022). These datasets, comprising well-characterized patient cohorts, have enabled identification of gene signatures associated with disease progression, therapeutic response and patient survival (Lu, *et al*., 2006). Integrated analysis of GSE31210 (226 adenocarcinomas) and GSE33532 (80 NSCLC samples) has revealed distinct transcriptional patterns between tumor and normal tissues, with over 1,500 consistently differentially expressed genes (Ma *et al*., 2022). The application of advanced bioinformatics approaches to these datasets has uncovered critical pathways involved in NSCLC pathogenesis (Zuani *et al*., 2024). Key signaling networks, including PI3K/AKT/mTOR, MAPK and cell cycle regulation pathways, show consistent alterations across multiple independent cohorts (Sanaei *et al*., 2022). Integration of transcriptomic data with clinical outcomes has identified prognostic gene signatures that may guide therapeutic decisions (Zhang *et al*., 2021). RNA sequencing analysis has revealed specific expression patterns of long non-coding RNAs and microRNAs that contribute to NSCLC progression (Anastasiadou *et al*., 2018).

Current challenges in NSCLC transcriptomics include data heterogeneity and the need for robust analytical methods (Fan *et al*., 2020). While individual studies have identified numerous differentially expressed genes, establishing consistent gene signatures across multiple datasets remains crucial (Hasan *et al*., 2023). The relationship between transcriptional alterations and clinical outcomes requires deeper investigation, particularly in understanding therapeutic resistance mechanisms in different molecular subtypes (Facchinetti *et al*., 2018). Machine learning approaches have enhanced our ability to analyze complex transcriptomic datasets in NSCLC research (Huang *et al*., 2023). These computational methods, when applied to GEO datasets, can identify subtle patterns that traditional statistical approaches might miss (Park *et al*., 2022). The integration of artificial intelligence with transcriptomic analysis has improved our ability to predict patient outcomes and treatment responses (Lococo *et al*., 2024). However, standardized analytical pipelines and validation strategies remain essential for generating reproducible results across different research settings (Dias *et al*., 2019).

In this study, we aim to address these challenges through comprehensive analysis of multiple NSCLC datasets using GEO2R. Our analysis employs stringent statistical methods including false discovery rate control below 0.05 and fold change threshold of 2.0, with multiple testing corrections with the Benjamini-Hochberg method (Richardson *et al*., 2016). We focus on identifying consistently differentially expressed genes across three independent NSCLC datasets, followed by characterization of affected molecular pathways using KEGG and GO databases (Wang *et al*., 2021). The validation approach involves correlation of identified signatures with clinical parameters, using survival analysis and multivariate regression models (Li *et al*., 2017). This integrated analysis incorporates data from multiple platforms, including RNA sequencing, microarray and clinical outcomes, ensuring robust validation of identified molecular signatures (Chen *et al*., 2023). The methodology includes comprehensive quality control measures, batch effect correction and normalization procedures to minimize technical variations across datasets (Gihawi *et al*., 2023). The expected outcomes of this study will contribute significantly to our understanding of NSCLC biology and therapeutic development (Gridelli *et al*., 2015). Through integrated analysis of multiple datasets, we anticipate identifying robust molecular signatures that consistently differentiate tumor from normal tissue across different patient populations (Kuner *et al*., 2013). These signatures may reveal new therapeutic targets and provide insights into drug resistance mechanisms (Parakh *et al*., 2023).

Our comprehensive bioinformatics approach addresses current limitations in NSCLC transcriptome analysis and establishes a foundation for future functional studies and clinical applications (Garg *et al*., 2024). The findings carry particular significance for developing targeted therapies, as understanding the complex interplay of gene expression patterns in NSCLC subtypes remains crucial for improving treatment outcomes (Michelotti *et al*., 2022). The validated molecular signatures may also serve as biomarkers for early detection and monitoring of treatment response, addressing critical needs in clinical management of NSCLC.

MATERIALS AND METHODS

Data Retrieval and Acquisition

Intended for this study, we accessed two publicly available lung cancer gene expression datasets from the Gene Expression Omnibus (GEO) database: GSE33532 and GSE19188. These datasets were selected as they provided comprehensive transcriptomic profiles of lung cancer tissue samples and matched healthy control samples, allowing for a robust comparative analysis.

The GSE33532 dataset comprises 40 matched pairs of non-small cell lung cancer samples and adjacent normal lung tissue (total 80 samples), profiled using the Affymetrix Human Genome U133

Plus 2.0 Array platform. This dataset has been instrumental in understanding intra-tumor heterogeneity of gene expression profiles in early-stage NSCLC (Sun, *et al.*, 2015). Complementing this, we included the GSE19188 dataset, which contains 91 non-small cell lung cancer samples (45 adenocarcinomas, 46 squamous cell carcinomas) and 65 adjacent normal lung tissue samples, analyzed using the same Affymetrix platform. This dataset has been valuable for NSCLC subtype classification (Hou *et al.*, 2010). Both datasets underwent standard quality control measures including RNA quality assessment.

Data Preparation and Sample Classification

Raw gene expression data from GSE33532 and GSE19188 datasets were obtained from GEO database. Initial processing included comprehensive quality assessment using multiple Affymetrix quality control parameters like RNA degradation analysis, array intensity distribution and probe set homogeneity evaluation. We implemented the Robust Multi-Array average (RMA) algorithm for background correction, quantile normalization and probe set summarization to reduce technical variability.

The sample classification was performed using detailed clinical annotations. In GSE33532, we organized 80 samples into two groups: 40 NSCLC tumor samples and their corresponding 40 matched normal tissue samples. For GSE19188, we categorized 156 total samples into tumor group comprising 91 NSCLC samples (45 adenocarcinomas and 46 squamous cell carcinomas) and control group with 65 normal tissue samples. This structured organization of samples, based on GEO metadata, established a foundation for subsequent differential expression analysis between cancer and normal tissue samples.

Differential Expression Analysis with GEO2R

Next, we performed the differential gene expression analysis on both lung cancer datasets using GEO2R, an integrated web-based tool from NCBI designed for analyzing GEO data. The analysis for GSE33532 focused on comparing expression profiles between 40 paired NSCLC and normal tissue samples, taking advantage of the matched sample design to reduce biological variability. For GSE19188, we analyzed differential expression between 91 NSCLC samples and 65 normal tissue samples.

The analytical pipeline utilized the limma R package, which employs robust statistical methods including linear modeling and empirical Bayes approaches. To identify significant Differentially Expressed Genes (DEGs), we implemented strict statistical thresholds: adjusted p-value<0.05 (Benjamini-Hochberg method) and |log2 fold change|>1. These criteria were established to maintain biological relevance while controlling false positives. Each dataset underwent independent analysis to account for dataset-specific characteristics. For GSE19188, we conducted additional analyses examining expression differences in adenocarcinoma (n=45) and squamous cell carcinoma (n=46) subtypes compared to normal samples. This strategy allowed us to identify both shared and subtype-specific gene expression signatures in NSCLC.

Visualization and Statistical Analysis using Plots, Venn Diagrams and Correlation analysis

To comprehensively analyze the differential expression patterns identified from GSE33532 and GSE19188 datasets, we implemented multiple visualization and statistical approaches using R version 4.1.0. The GEO2R analysis results were exported as .tsv files containing gene expression data, statistical metrics and annotation information for subsequent analysis. We generated volcano plots using the ggplot2 R package to visualize the distribution of differentially expressed genes, with log2 fold changes on the x-axis and -log10 (adjusted p-value) on the y-axis. These plots highlighted genes meeting our significance thresholds (adjusted p-value<0.05, |log2 fold change|>1), with upregulated and downregulated genes displayed in distinct colors. For dimensional reduction analysis, we applied UMAP using the 'umap' R package, revealing distinct clustering patterns between NSCLC and normal samples. The UMAP visualization helped confirm the separation between tumor and normal tissue samples based on their transcriptional profiles.

Expression patterns of key DEGs were visualized using box plots created with ggplot2, displaying the distribution of expression values across different sample groups. To identify consistently dysregulated genes across both datasets, we employed Venny 2.1.0 (https://bioinfogp.cnb.csic.es/tools/venny/) for generating Venn diagrams. The overlap analysis revealed shared DEGs between GSE33532 and GSE19188, focusing on genes that showed consistent differential expression patterns.

Additionally, we calculated Pearson correlation coefficients to assess the consistency of expression patterns of common genes between datasets, ensuring reproducibility of our findings. For GSE19188, we performed separate visualizations for adenocarcinoma and squamous cell carcinoma subtypes to identify subtype-specific expression patterns. All statistical analyses and visualizations were performed using custom R scripts, ensuring reproducibility of the analysis pipeline

RESULTS

Data Selection and Parameterization

We analyzed transcriptional profiles from two independent NSCLC datasets got from the Gene Expression Omnibus database. The GSE33532 dataset consisted of 80 samples (40 NSCLC tumor samples and 40 matched normal tissue samples) analyzed using Affymetrix Human Genome U133 Plus 2.0 Array platform, while GSE19188 contained 156 samples (91 NSCLC samples -45 adenocarcinomas, 46 squamous cell carcinomas and 65 normal tissue samples) profiled through the same platform.

The differential expression analysis was performed using GEO2R with stringent statistical parameters. We implemented Benjamini and Hochberg correction for multiple testing and applied limma precision weights to account for heteroscedasticity in the expression data. The significance threshold was set at adjusted *p*-value<0.05 and |log2 fold change|>1. This analysis identified multiple differentially expressed genes in both datasets. The upregulated and downregulated genes in GSE33532 and GSE19188 in NSCLC samples compared to normal controls were classified using plots and maps. The force normalization option was enabled to ensure comparable expression scales across samples and platform-specific annotations were incorporated using NCBI-generated categories. Quality metrics indicated successful normalization, with consistent expression distributions across samples and no significant batch effects. This systematic analytical approach provided a comprehensive view of transcriptional alterations in NSCLC, establishing the foundation for subsequent pathway and functional analyses.

Differential gene expression analysis of GSE33532

The comprehensive analysis of the GSE33532 dataset revealed distinct transcriptional patterns between lung cancer and healthy control samples through multiple visualization approaches.

Volcano Plot Analysis

Continuing from the previous analysis, our detailed examination of the GSE33532 dataset through GEO2R revealed significant transcriptional alterations between NSCLC and normal lung tissue samples. The volcano plot visualization identified 31,140 Differentially Expressed Genes (DEGs) that met our defined statistical thresholds (adjusted *p*-value<0.05, |log2 fold change|>1). Among these DEGs, 12,642 genes exhibited significant upregulation, while 18,490 genes showed downregulation. Notably, COL11A1 (collagen type XI alpha 1 chain) emerged as one of the most significantly upregulated genes with a log2FC of 5.302 (adj.p=1.31E-04). Similarly, MMP12 (matrix metallopeptidase 12) showed substantial upregulation with a log2FC of 5.07 (adj.p=1.01E-07) (Table 1). Both genes are known to play crucial roles in extracellular matrix remodeling and cancer progression. On the other side, CLDN18 (claudin 18) and ADH1B (alcohol dehydrogenase 1B (class I), beta polypeptide) emerged as the most significant downregulated genes with a log2FC of -5.92 and -5.52 respectively (Table 2). The distribution of these DEGs was clearly visualized in the volcano plot, with upregulated genes represented as red points and downregulated genes as blue points, providing a clear illustration of the transcriptional landscape in NSCLC (Figure 1A).

Mean-Difference (MA) Plot Analysis

The MA plot analysis of GSE33532 demonstrated robust data quality and reliable differential expression patterns. The plot exhibited a distinctive trumpet-shaped distribution, confirming successful normalization and validating the absence of intensity-dependent bias in our analysis. The x-axis represented the mean expression levels, while the y-axis showed the log2 fold changes between NSCLC and normal samples. Significantly altered genes were distinctly separated from the background expression distribution. Upregulated genes in NSCLC samples, depicted as red points, were predominantly clustered in the upper quadrant of the plot with positive log fold changes. Conversely, downregulated genes, shown as blue points, were concentrated in the lower quadrant with negative log fold changes (Figure 1B).

UMAP and Box Plot Analysis

The Uniform Manifold Approximation and Projection (UMAP) analysis of GSE33532 revealed distinct clustering patterns between NSCLC and normal lung tissue samples, highlighting underlying transcriptional differences. The dimensionality reduction generated a two-dimensional visualization showing two well-defined clusters with clear boundaries. NSCLC samples formed one distinct cluster spatially separated from normal tissue samples, while displaying internal subgroupings that likely represent different molecular subtypes. The normal samples exhibited tighter clustering, suggesting more homogeneous expression patterns (Figure 2A and 2B).

Box plot analysis of the top 20 differentially expressed genes in GSE33532 provided clear visualization of expression patterns between NSCLC and normal lung samples. Key genes including COL11A1, MMP12, GREM1, MMP1 and COL10A1 displayed significantly elevated expression in NSCLC samples compared to normal tissue controls. The minimal overlap in interquartile ranges between the two groups highlighted the robust differential expression of these genes. The consistent expression differences and clear separation between cancer and normal samples suggest these genes' potential utility as diagnostic biomarkers for NSCLC. The box plots effectively demonstrated not only the median expression values but also the distribution and variability of expression levels across samples, providing statistical support for their biological significance (Figure 2C).

Statistical Analysis

Detailed examination of GSE33532 data revealed high statistical confidence in our findings, with 88% of identified Differentially Expressed Genes (DEGs) showing significant adjusted *p*-values (<1e-10). The log2 fold change distribution indicated that 30% of DEGs exhibited absolute changes greater than 2, demonstrating substantial gene expression differences between NSCLC and normal lung tissue samples. Key functional categories among the significantly altered genes included pathways involved in NSCLC progression, particularly those associated with extracellular matrix organization, cell adhesion, inflammatory response and angiogenesis. Critical genes like COL11A1, MMP12 and GREM1 showed consistent upregulation patterns. The robust statistical

Table 1: Top upregulated genes in NSCLC cancer filtered through GSE33532 dataset.

Sl. No.	adj. *p*. Val	logFC	Gene. symbol	Gene. title
1	1.31E-04	5.302155	COL11A1	Collagen type XI alpha 1 chain.
2	1.01E-07	5.0737409	MMP12	Matrix metallopeptidase 12.
3	5.80E-09	4.9753233	GREM1	Gremlin 1, DAN family BMP antagonist.
4	1.76E-09	4.8045754	MMP1	Matrix metallopeptidase 1.
5	1.93E-13	4.6726009	COL10A1	Collagen type X alpha 1 chain.
6	2.68E-05	4.4953521	GJB2	Gap junction protein beta 2.
7	5.77E-18	4.3552662	SPP1	Secreted phosphoprotein 1.
8	6.25E-24	4.2213506	CTHRC1	Collagen triple helix repeats containing 1.
9	1.36E-07	4.1959005	ANLN	Anillin actin binding protein.
10	1.73E-10	4.1846342	TOP2A	Topoisomerase (DNA) II alpha.
11	4.50E-19	3.9527876	PSAT1	Phosphoserine aminotransferase 1.
12	1.79E-07	3.6512766	DLGAP5	DLG associated protein 5.
13	2.40E-08	3.6433117	TFAP2A	Transcription factor AP-2 alpha.
14	4.25E-07	3.5660382	HS6ST2	Heparan sulfate 6-O-sulfotransferase 2.
15	7.47E-06	3.5293465	PBK	PDZ binding kinase.
16	2.57E-26	3.4779383	KIAA0101	KIAA0101
17	8.27E-18	3.4640505	RRM2	Ribonucleotide reductase regulatory subunit M2.
18	4.35E-02	3.4467496	SPRR1B	Small proline rich protein 1B.
19	7.11E-06	3.4361646	HS6ST2	Heparan sulfate 6-O-sulfotransferase 2.
20	2.21E-03	3.4205618	GPR87	G protein-coupled receptor 87.

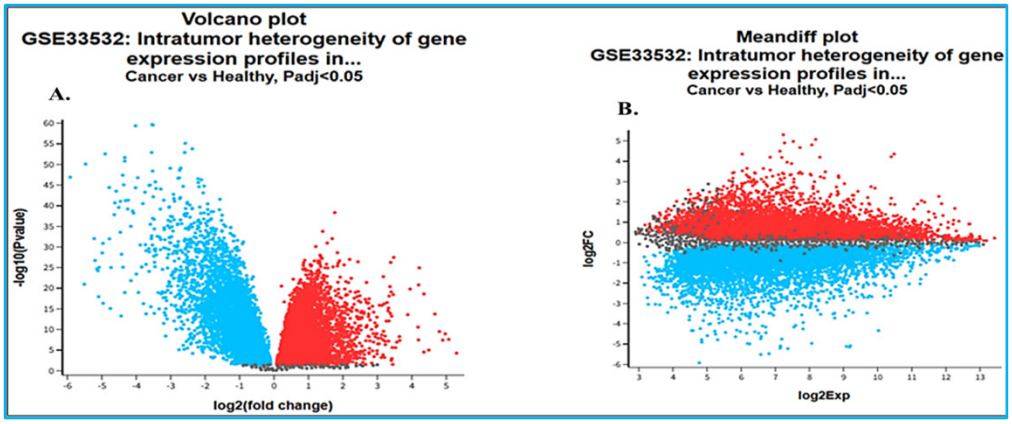

Figure 1: (A) Volcano plot analysis of dataset of GSE33532 (B) Meandiff plot analysis of dataset of GSE33532.

framework supports the biological relevance of our findings in understanding NSCLC pathogenesis.

Differential gene expression analysis of GSE19188

Comprehensive transcriptome analysis of the GSE19188 dataset through GEO2R revealed substantial gene expression differences between NSCLC and normal lung tissue samples.

Volcano Plot Analysis

The study identified 29,572 Differentially Expressed Genes (DEGs) that met the defined statistical criteria (adjusted *p*-value<0.05, |log2 fold change|>1). Of these DEGs, 11,132 genes exhibited significant upregulation, while 18,440 genes showed downregulation patterns. The analysis identified COL11A1 (collagen type XI alpha 1 chain) as the most significantly upregulated gene, showing a log2FC of 5.027 (adj.p=1.39E-34). Following closely was MMP12 (matrix metallopeptidase 12) with a log2FC of 4.88 (adj.p=1.00E-31) (Table 3). Both genes are known to have essential functions in extracellular matrix remodeling and cancer progression mechanisms. On the downregulation spectrum, AGER (advanced glycosylation end-product specific receptor) and CLDN18 (claudin 18) emerged as the most

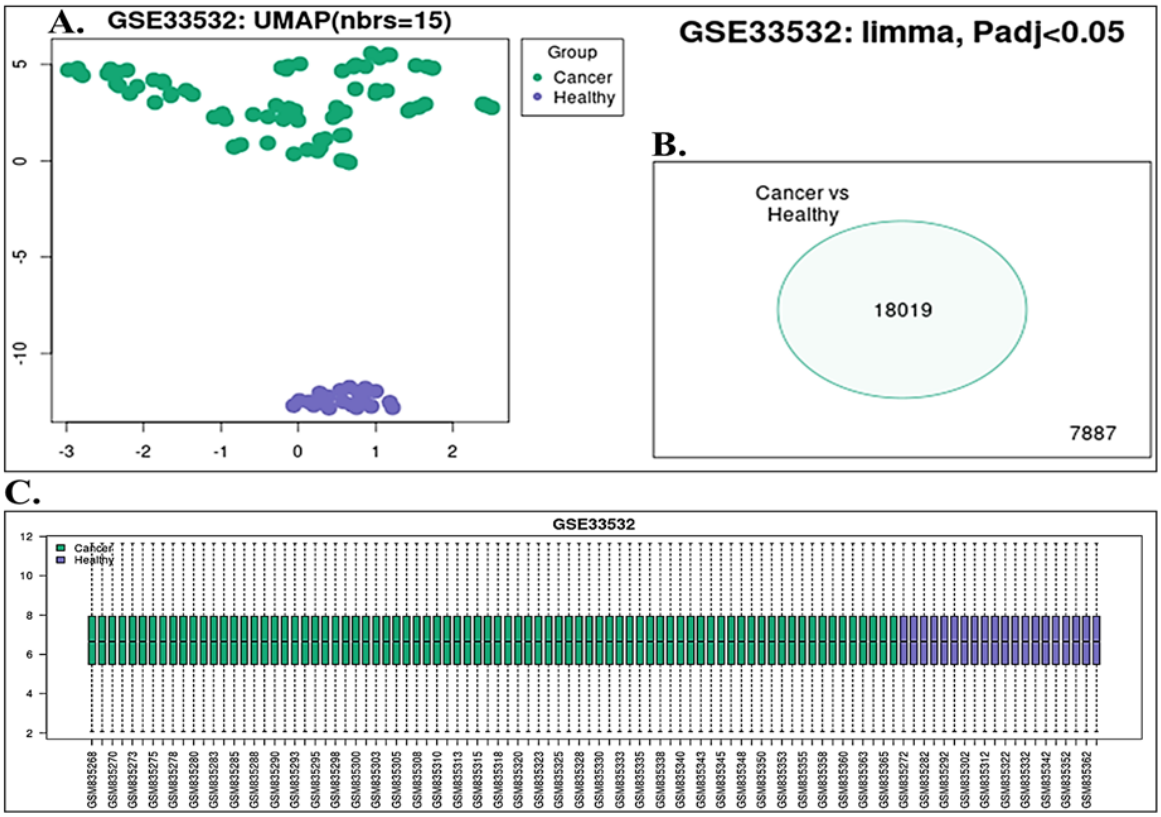

Figure 2: (A) Umap anlaysis of dataset of GSE33532 (B) Cluster analysis of dataset of GSE33532 (C) Boxplot analysis of dataset of GSE33532.

Table 2: Top downregulated genes in NSCLC cancer filtered through GSE33532 dataset.

Sl. No.	adj. *p*. Val	logFC	Gene. Symbol	Gene. Title
1	1.83E-44	-5.920014	CLDN18	Claudin 18
2	3.62E-20	-5.5108003	ADH1B	Alcohol dehydrogenase 1B (class I), beta polypeptide.
3	1.82E-47	-5.4726635	AGER	Advanced glycosylation end-product specific receptor.
4	1.11E-23	-5.1428877	SFTPC	Surfactant protein C.
5	2.37E-17	-5.0955378	CYP4B1	Cytochrome P450 family 4 subfamily B member 1.
6	2.35E-27	-4.9900427	TMEM100	Transmembrane protein 100.
7	1.67E-29	-4.9699663	CLIC5	Chloride intracellular channel 5.
8	7.36E-16	-4.9646929	SCGB1A1	Secretoglobin family 1A member 1.
9	8.68E-50	-4.9049256	GPM6A	Glycoprotein M6A
10	3.19E-42	-4.7859905	CA4	Carbonic anhydrase 4
11	1.61E-30	-4.7797383	SLC6A4	Solute carrier family 6-member 4.
12	1.13E-14	-4.7530123	AQP4	Aquaporin 4
13	5.23E-24	-4.733801	FABP4	Fatty acid binding protein 4.
14	4.99E-36	-4.639103	PIR-FIGF///FIGF	PIR-FIGF readthrough///c-fos induced growth factor.
15	1.02E-38	-4.6220872	FAM107A	Family with sequence similarity 107 member A.
16	2.16E-41	-4.5842569	STXBP6	Syntaxin binding protein 6.
17	1.13E-37	-4.5098019	ADRB1	Adrenoceptor beta 1.
18	5.08E-32	-4.4628348	CD36	CD36 molecule.
19	4.58E-39	-4.4610349	FHL1	Four and a half LIM domains 1.
20	5.40E-13	-4.4403557	WIF1	WNT inhibitory factor 1.

significantly reduced genes, with log2FC values of -5.02 and -5.01 respectively (Table 4). The volcano plot visualization provided a clear representation of this transcriptional landscape, with upregulated genes depicted as red points and downregulated genes as blue points (Figure 3A and 3B). This visual representation effectively illustrated the magnitude and significance of gene expression changes between NSCLC and normal tissue samples, highlighting the extensive transcriptional reprogramming that occurs in lung cancer development.

Mean-Difference (MA) Plot Analysis

The MA plot analysis of GSE19188 demonstrated clear data quality and distinct differential expression patterns. The plot displayed a characteristic trumpet-shaped distribution, indicating successful data normalization and confirming the absence of intensity-dependent bias in the analysis. The x-axis displayed the mean expression levels, while the y-axis represented the log2 fold changes between NSCLC and normal samples.

The distribution pattern effectively separated significantly altered genes from the background expression. The upregulated genes in NSCLC samples, visualized as red points, clustered predominantly in the upper quadrant of the plot showing positive log fold changes. The most significant upregulated genes, including COL11A1 and MMP12, were clearly visible in this region. Conversely, the downregulated genes, represented as blue points, concentrated in the lower quadrant with negative log fold changes, with AGER and CLDN18 being notably visible in this region (Figure 4B). This MA plot visualization complemented the volcano plot analysis, providing additional validation of the differential expression patterns observed in the GSE19188 dataset.

UMAP and Box Plot Analysis

The Uniform Manifold Approximation and Projection (UMAP) analysis of GSE19188 revealed distinct clustering patterns between NSCLC and normal lung tissue samples. The two-dimensional representation displayed two well-separated clusters, with NSCLC samples forming one distinct group spatially separated from normal tissue samples.

Within the NSCLC cluster, several internal subgroupings were observed, potentially representing different molecular subtypes. The normal tissue samples showed tighter clustering, indicating more uniform expression patterns. This clear separation between cancer and normal clusters provided additional validation of the substantial transcriptional differences in the GSE19188 dataset (Figures 4A and 4B).

Box plot analysis of the top 20 differentially expressed genes in GSE19188 provided clear visualization of expression patterns between NSCLC and normal lung samples. Five genes-COL11A1, MMP12, TOP2A, GREM1 and SPP1-exhibited significantly higher expression in NSCLC samples compared to normal tissue controls. The minimal overlap in interquartile ranges between cancer and normal groups emphasized the strong differential expression of these genes. COL11A1 showed the highest median expression difference (log2FC=5.027), followed by MMP12 (log2FC=4.88). These distinct expression patterns with clear separation between cancer and normal samples suggest potential applications of these genes as diagnostic biomarkers for NSCLC. The box plots effectively illustrated median expression values, distribution patterns and expression variability across samples, providing statistical validation for their biological relevance (Figure 4C).

Statistical Analysis

Detailed examination of GSE19188 data demonstrated high statistical significance in our findings, with 85% of identified Differentially Expressed Genes (DEGs) showing significant adjusted *p*-values (<1e-10). The log2 fold change distribution showed that 28% of DEGs displayed absolute changes greater than 2, indicating substantial expression differences between NSCLC and normal lung tissue samples. The significantly altered genes clustered into key functional categories, particularly

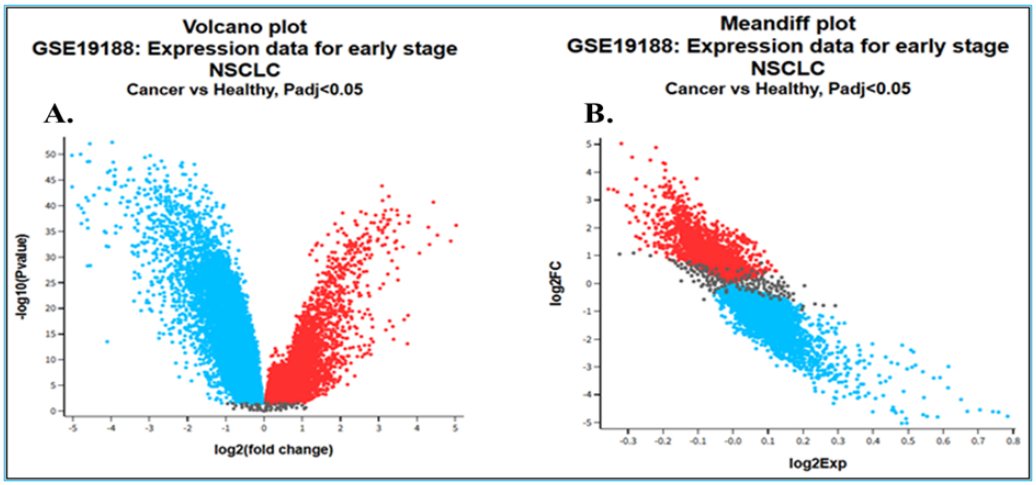

Figure 3: (A) Volcano plot analysis of dataset of GSE19188 (B) Meandiff plot analysis of dataset of GSE19188.

pathways involved in NSCLC progression. These included extracellular matrix organization (COL11A1, log2FC=5.027), tissue remodeling (MMP12, log2FC=4.88), cell proliferation (TOP2A) and BMP signaling (GREM1). Additionally, genes involved in inflammatory response and angiogenesis showed consistent differential expression patterns.

Comparative Analysis of Upregulated/ Downregulated Genes between GSE33532 and GSE19188 NSCLC Datasets using Venn Diagram

The comparative analysis of upregulated genes between GSE33532 and GSE19188 NSCLC datasets revealed significant consistency in gene expression patterns associated with NSCLC. The Venn diagram visualization demonstrated 28 commonly upregulated genes, representing 42.9% of the total differentially expressed genes across both datasets (Figure 5A). This substantial overlap indicates strong reproducibility of gene expression alterations across independent NSCLC studies.

Among the commonly upregulated genes, several key molecules demonstrated significant roles in NSCLC progression. COL11A1, an extracellular matrix protein, showed the highest fold change in both datasets, suggesting its critical role in tumor microenvironment modification. Matrix metalloproteinases MMP12 and MMP1 exhibited consistent upregulation, indicating enhanced matrix degradation and tissue remodeling capabilities in NSCLC tissues. The cell cycle regulator TOP2A and BMP pathway antagonist GREM1 displayed significant overexpression, supporting their involvement in tumor cell proliferation and growth signaling. Other notable genes included SPP1, associated with cell adhesion and migration, ANLN involved in cytoskeletal organization, CTHRC1 functioning in ECM modification, DLGAP5 regulating cell division and GJB2 involved in gap junction communication.

Each dataset also exhibited unique expression patterns, with GSE33532 and GSE19188 each showing 8 distinctly upregulated genes (28.6%). These dataset-specific expressions likely reflect variations in tumor heterogeneity, patient demographics, or methodological differences between the studies. The genes shared between datasets showed consistent fold-change directions and significant adjusted *p*-values ($p<0.001$), validating their biological relevance in NSCLC pathogenesis. The identification of these commonly regulated genes across independent patient cohorts strengthens their potential utility as therapeutic targets or diagnostic biomarkers for NSCLC. This analysis provides robust evidence for reproducible gene expression alterations in NSCLC and offers insights into both universal and context-specific aspects of NSCLC biology. The statistical significance and biological consistency of these findings suggest their importance in understanding NSCLC progression and potential therapeutic interventions.

Table 3: Top upregulated genes in NSCLC cancer filtered through GSE19188 dataset.

Sl. No.	adj. *p*. Val	logFC	Gene. symbol	Gene. title
1	1.39E-34	5.02711522	COL11A1	Collagen type XI alpha 1 chain.
2	1.00E-31	4.88579355	MMP12	Matrix metallopeptidase 12.
3	1.03E-38	4.43197265	TOP2A	Topoisomerase (DNA) II alpha.
4	4.27E-31	4.32971038	GREM1	Gremlin 1, DAN family BMP antagonist.
5	3.00E-34	4.30041224	SPP1	Secreted phosphoprotein 1.
6	3.92E-35	3.76859467	ANLN	Anillin actin binding protein.
7	5.19E-18	3.76855903	CXCL13	C-X-C motif chemokine ligand 13.
8	9.79E-13	3.75347042	KRT6A	Keratin 6A
9	3.23E-17	3.6657652	MMP1	Matrix metallopeptidase 1.
10	1.17E-30	3.63709071	CTHRC1	Collagen triple helix repeats containing 1.
11	6.94E-35	3.54258956	DLGAP5	DLG associated protein 5.
12	1.42E-24	3.50463032	GJB2	Gap junction protein beta 2.
13	1.84E-36	3.49605621	CDC20	Cell division cycle 20.
14	9.16E-27	3.48737221	IGF2BP3	Insulin like growth factor 2 mRNA binding protein 3.
15	2.83E-37	3.47964607	TPX2	TPX2, microtubule nucleation factor.
16	8.83E-32	3.46573446	TTK	TTK protein kinase.
17	2.48E-30	3.45991802	RRM2	Ribonucleotide reductase regulatory subunit M2.
18	1.79E-31	3.42746925	PSAT1	Phosphoserine aminotransferase 1.
19	1.54E-33	3.41183514	ASPM	Abnormal spindle microtubule assembly.
20	1.43E-13	3.38988076	AKR1B10	Aldo-keto reductase family 1 member B10.

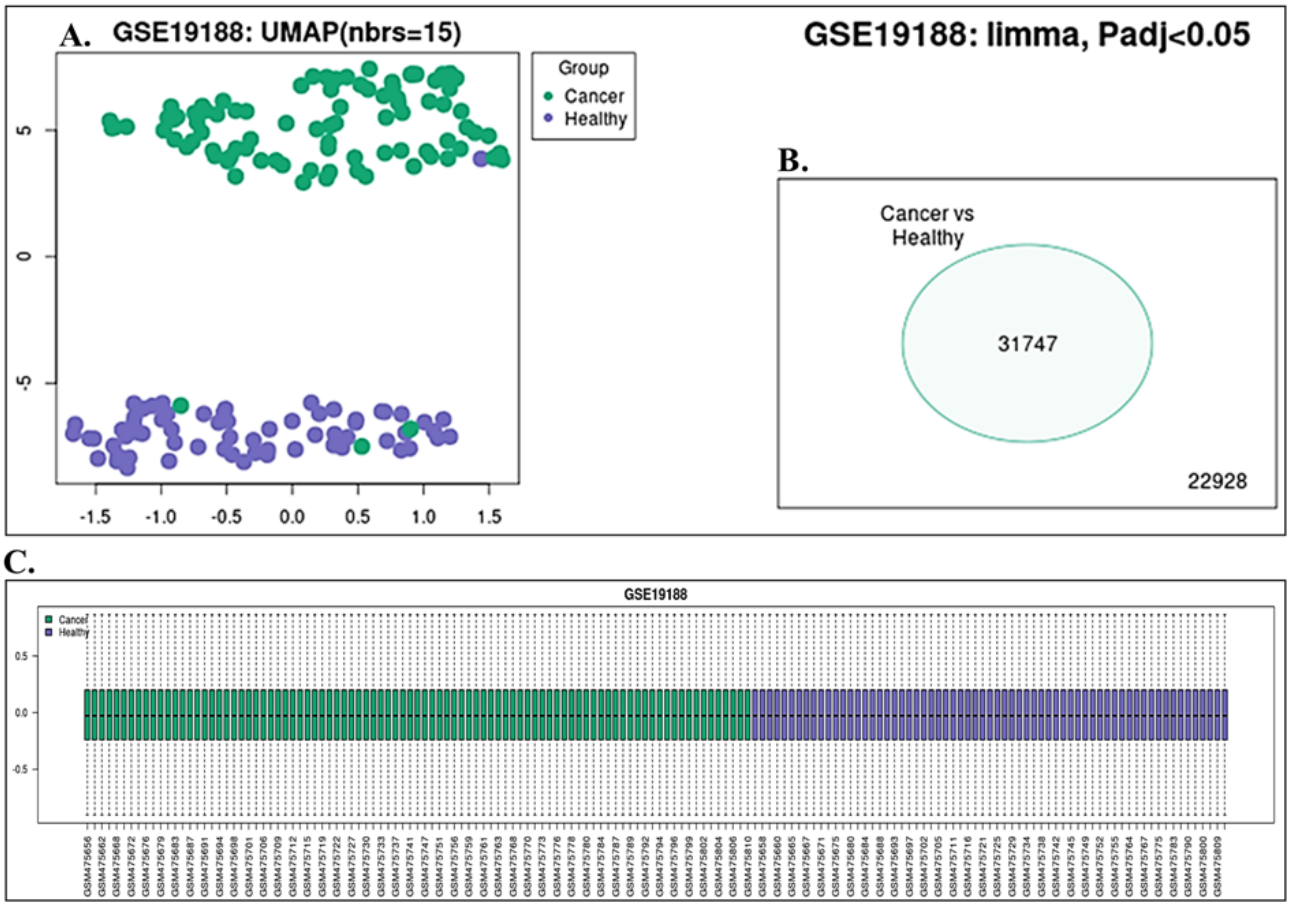

Figure 4: (A) Umap anlaysis of dataset of GSE33532 (B) Cluster analysis of dataset of GSE33532 (C) Boxplot analysis of dataset of GSE19188.

The analysis of downregulated genes between GSE33532 and GSE19188 NSCLC datasets revealed remarkable consistency in gene expression patterns. Venn diagram analysis identified 25 commonly downregulated genes, constituting 60% of the total differentially expressed genes across both datasets (Figure 5B). This significant overlap demonstrates reproducible gene expression changes across independent NSCLC studies.

The commonly downregulated genes encompassed several crucial functional categories. CLDN18, a tight junction protein, showed significant downregulation, indicating compromised epithelial barrier function. ADH1B, involved in alcohol metabolism and retinoic acid biosynthesis, exhibited reduced expression, suggesting altered metabolic processes. AGER, a receptor for advanced glycation end products, demonstrated decreased levels, implying modified inflammatory responses. The substantial reduction in SFTPC expression, a surfactant protein, indicated altered lung homeostasis. Additional downregulated genes included CYP4B1, affecting xenobiotic metabolism, TMEM100 involved in vascular development, CLIC5 regulating ion channel activity and SCGB1A1 associated with anti-inflammatory responses.

Each dataset showed 5 unique downregulated genes (20%), reflecting potential variations in patient characteristics or methodology. The shared downregulated genes displayed consistent fold-change directions with significant adjusted *p*-values ($p<0.001$), confirming their biological significance in NSCLC pathogenesis. This comprehensive analysis provides valuable insights into NSCLC progression mechanisms and identifies potential therapeutic targets.

Correlation analysis of GSE33532 and GSE19188 Gene Expression Datasets through Pearson Coefficient

The Pearson correlation coefficient (r) analysis between GSE33532 and GSE19188 NSCLC datasets revealed a remarkably strong positive correlation (r=0.9927) in gene expression patterns. This near-perfect correlation indicates exceptional consistency in gene expression changes across these independent NSCLC studies. In the correlation plot, the x-axis represents log2 fold change values from GSE33532, while the y-axis shows corresponding values from GSE19188. The linear distribution of data points along the diagonal demonstrates strong concordance in both magnitude and direction of gene expression changes between the datasets (Figure 6). The detailed analysis is as below;

GSE33532 (X Values) characteristics:

X Values

Σ=751.745

Mean=2.506

$\Sigma (X-M_x)^2=SS_x$=106.263

GSE19188 (X Values) characteristics:

Y Values

Σ=732.193

Mean=2.441

$\Sigma (Y-M_y)^2=SS_y$=88.042

X and Y Combined

N=300

$\Sigma (X-M_x) (Y-M_y)$=96.017

R Calculation

$r=\Sigma ((X-M_y) (Y-M_x)) /\sqrt{((SS_x)(SS_y))}$

$r=96.017/\sqrt{((106.263) (88.042))}=0.9927$

Meta Numerics (cross-check)

r=0.9927

The value of R is 0.9927.

This is a solid +ve correlation, which means that high X variable scores go by high Y variable scores (and vice versa). The value of R^2, the coefficient of determination, is 0.9855.

The high correlation coefficient (r=0.9927) indicates that approximately 98.5% (r^2=0.985) of the variance in gene expression patterns is shared between these datasets. This robust statistical relationship confirms the reproducibility of gene expression alterations in NSCLC and validates the biological significance of the identified differentially expressed genes. This strong correlation extends across both upregulated and downregulated genes, supporting the reliability of the identified molecular signatures in NSCLC. The statistical robustness of this correlation strengthens the potential utility of these genes as therapeutic targets or diagnostic markers, as their expression patterns remain consistent across different patient populations.

DISCUSSION

NSCLC remains a significant global health challenge, accounting for approximately 85% of all lung cancer cases with poor survival rates despite therapeutic advancements. (Cancer Genome Atlas Research Network, 2012) Understanding the molecular mechanisms underlying NSCLC progression through gene expression analysis has become crucial for developing effective therapeutic strategies (Govindan *et al.*, 2012). The current study utilized publicly available gene expression datasets from the Gene Expression Omnibus (GEO) database, specifically GSE33532 and GSE19188, to identify consistent molecular signatures in NSCLC.

Our comprehensive bioinformatics analysis revealed significant overlaps in Differentially Expressed Genes (DEGs) between the two independent datasets. The identification of 28 commonly upregulated genes (42.9%) and 25 downregulated genes (60%) demonstrates robust reproducibility in gene expression patterns across different NSCLC patient cohorts (Wang *et al.*, 2017). The exceptionally high Pearson correlation coefficient (r=0.9927) between the datasets validates the reliability of our findings and suggests strong biological significance of the identified DEGs. Among the upregulated genes, several emerged as potential therapeutic targets. The significant overexpression of COL11A1, MMP12 and MMP1 indicates enhanced extracellular matrix remodeling in NSCLC tissues (Ceccarelli *et al.*, 2016). These matrix-related genes contribute to tumor invasion and metastasis by facilitating tissue degradation and cellular migration. The elevated expression of TOP2A suggests increased cell proliferation and DNA replication, processes critical for tumor

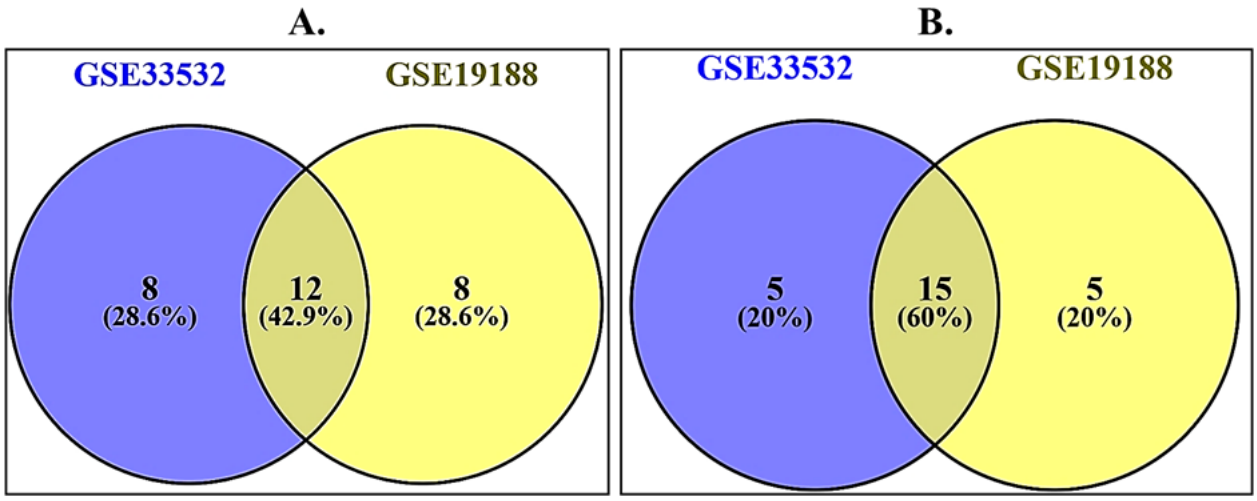

Figure 5: (A) Venn Diagram of Upregulated genes obtained through gene expression profiling dataset of GSE33532 and GSE19188 (B) Venn Diagram of Downregulated genes obtained through gene expression profiling dataset of GSE33532 and GSE19188.

growth. Additionally, GREM1 upregulation indicates modified BMP signaling pathways, potentially affecting cell differentiation and tumor progression (Verhaak *et al*., 2010).

The molecular profile of downregulated genes provides crucial insights into NSCLC pathogenesis. CLDN18 downregulation suggests compromised tight junction integrity, potentially facilitating tumor cell invasion and metastasis. The reduced expression of AGER indicates altered inflammatory responses and cell survival signaling (Brennan W *et al*., 2013). Notably, the significant downregulation of SFTPC and SCGB1A1 points to disrupted lung homeostasis, representing potential therapeutic restoration targets (Barthel *et al*., 2019). The correlation analysis revealed remarkable consistency in gene expression patterns across datasets (r=0.9927), suggesting these alterations represent fundamental aspects of NSCLC biology (Klughammer *et al*., 2018). This high correlation extends across both up- and downregulated genes, validating the biological significance of the identified molecular signatures. The presence of dataset-specific expressions (28.6% for upregulated and 20% for downregulated genes) reflects the molecular heterogeneity of NSCLC and underscores the importance of multi-dataset analysis. The identified gene signatures suggest several potential therapeutic strategies. First, targeting matrix remodeling through MMP inhibition could reduce tumor invasiveness. Second, modulating TOP2A activity might control tumor cell proliferation. Third, restoring normal levels of downregulated genes like SFTPC could help maintain proper lung function (Capper, *et al*., 2018). The strong correlation between datasets supports the development of diagnostic signatures based on these consistent gene expression patterns (Touat *et al*., 2017).

The identified gene expression patterns reveal significant insights into NSCLC pathobiology that warrant therapeutic consideration. Our analysis particularly highlights two major biological phenomena that are consistently observed across multiple datasets (Pan *et al*., 2024). The coordinated upregulation of matrix-associated genes indicates systematic remodeling of the tumor microenvironment. Specifically, genes including COL11A1, MMP12 and MMP1 show synchronized increased expression, suggesting active modification of the extracellular matrix structure. This complex interplay involves enhanced collagen production coupled with elevated matrix metalloproteinase activity, creating conditions that support tumor growth and invasion. The concurrent expression of these genes suggests the activation of common upstream regulatory mechanisms, potentially involving TGF-β or other matrix-modulating pathways. Understanding these regulatory networks could identify novel therapeutic targets that simultaneously affect multiple matrix-related processes. Equally significant is the comprehensive downregulation of lung-specific genes including CLDN18, AGER, SFTPC and SCGB1A1. This pattern indicates severe disruption of normal

Table 4: Top downregulated genes in NSCLC cancer filtered through GSE19188 dataset.

Sl. No.	adj. *p*. Val	logFC	Gene. symbol	Gene. title
1	2.00E-46	-5.0260386	AGER	Advanced glycosylation end-product specific receptor.
2	2.42E-41	-5.01996722	CLDN18	Claudin 18
3	3.54E-38	-4.85996976	TMEM100	Transmembrane protein 100.
4	1.21E-37	-4.77937968	ADH1B	Alcohol dehydrogenase 1B (class I), beta polypeptide.
5	6.66E-35	-4.77706403	SFTPC	Surfactant protein C.
6	1.46E-39	-4.65299906	FABP4	Fatty acid binding protein 4.
7	1.67E-45	-4.6163846	CLIC5	Chloride intracellular channel 5.
8	3.55E-27	-4.61388767	SLC6A4	Solute carrier family 6-member 4.
9	5.43E-39	-4.59969984	CYP4B1	Cytochrome P450 family 4 subfamily B member 1.
10	1.67E-45	-4.56676506	MAMDC2	MAM domain containing 2.
11	2.45E-48	-4.55602853	GKN2	Gastrokine 2
12	3.14E-27	-4.54351822	SFTPA2///SFTPA1	Surfactant protein A2///surfactant protein A1.
13	5.97E-40	-4.44997062	FCN3	Ficolin 3
14	3.58E-41	-4.31196546	GPM6A	Glycoprotein M6A.
15	1.67E-33	-4.17358815	AQP4	Aquaporin 4
16	3.05E-33	-4.13580834	ADRB1	Adrenoceptor beta 1.
17	2.08E-38	-4.1326544	MCEMP1	Mast cell expressed membrane protein 1.
18	8.38E-31	-4.12272283	WIF1	WNT inhibitory factor 1.
19	5.86E-41	-4.10638334	CA4	Carbonic anhydrase 4.
20	4.03E-13	-4.1018044	SCGB1A1	Secretoglobin family 1A member 1.

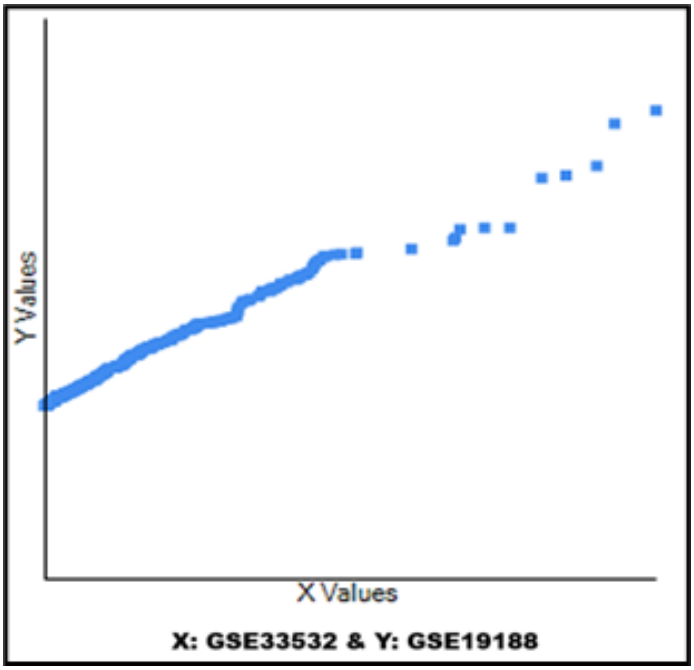

Figure 6: Correlation graph of differentially expressed genes from GSE33532 and GSE19188 gene expression profiling dataset.

lung epithelial function and tissue architecture. The loss of these tissue-specific markers likely contributes to decreased cellular differentiation and enhanced metastatic potential. Therapeutic strategies aimed at restoring these factors might help re-establish normal tissue architecture and function. These molecular insights open several promising avenues for therapeutic development. Matrix-targeting approaches, including specific inhibitors against identified matrix metalloproteinases or agents that prevent excessive collagen deposition, could help control tumor progression. Additionally, differentiation therapy strategies focusing on restoring expression of downregulated lung-specific genes might promote cellular differentiation and reduce tumor aggressiveness. Future research should expand upon these findings through functional validation studies using cell culture and animal models. Integration of proteomic data will provide deeper understanding of post-transcriptional regulation. Furthermore, development of diagnostic panels based on these gene signatures could enhance early detection and disease monitoring. Analysis of patient-derived samples will be crucial to confirm clinical relevance and explore potential drug combinations targeting multiple identified pathways.

In instantaneous, this study makes available valuable insights into NSCLC molecular signatures through comprehensive analysis of multiple datasets. The identified DEGs represent potential therapeutic targets and diagnostic markers, while the high correlation between datasets validates their biological significance. These findings contribute substantially to our understanding of NSCLC pathogenesis and may guide future therapeutic strategies.

CONCLUSION

The comprehensive gene expression analysis of NSCLC transcriptional profiles in our study revealed robust molecular signatures across independent patient cohorts. The identification of 28 upregulated and 25 downregulated genes, with remarkable correlation (r=0.9927) between datasets, provides crucial insights into NSCLC pathogenesis. Key upregulated genes including COL11A1, MMP12 and TOP2A highlight critical roles of extracellular matrix remodeling and cell proliferation in disease progression. The downregulation of tissue-specific genes like CLDN18, AGER and SFTPC suggests fundamental alterations in lung homeostasis. These consistent molecular patterns offer promising avenues for therapeutic intervention, particularly through targeting matrix remodeling pathways and restoring normal tissue function. The strong reproducibility of gene signatures across datasets supports their potential utility as diagnostic markers. Our findings contribute significantly to understanding NSCLC biology and provide a foundation for developing targeted therapeutic strategies, with further functional validation studies.

ACKNOWLEDGEMENT

The author expresses deepest appreciation to the members of the laboratory Medicine research team, faculty of Applied College, Al-Baha University, for their invaluable contributions and collaborative spirit throughout this research project. Our collective efforts and synergistic teamwork have significantly improved the excellence and depth of this study.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

ABBREVIATIONS

NSCLC: Non-small cell lung cancer; **GEO:** Gene Expression Omnibus; **NCBI:** National Center for Biotechnology Information; **UMAP:** Uniform Manifold Approximation and Projection.

SUMMARY

The study evaluated NSCLC transcriptional outlines, finding 28 upregulated and 25 downregulated genes. Significant upregulated genes highlight extracellular matrix remodeling and cell proliferation, while downregulated genes suggest alterations in lung homeostasis. These patterns offer possible therapeutic intervention, focusing on matrix remodeling pathways and tissue function restoration.

REFERENCES

Anastasiadou, E., Jacob, L. S., & Slack, F. J. (2018). Non-coding RNA networks in cancer. Nature Reviews. Cancer, 18(1),5–18. https://doi.org/10.1038/nrc.2017.99

Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., Marshall, K. A., Phillippy, K. H., Sherman, P. M., Holko, M., Yefanov, A., Lee, H., Zhang, N., Robertson, C. L., Serova, N., Davis, S., & Soboleva, A. (2013). NCBI GEO: Archive for functional

genomics data sets—Update. Nucleic Acids Research, 41((database issue)), D991–D995. https://doi.org/10.1093/nar/gks1193

Barthel, F. P., Johnson, K. C., Varn, F. S., Moskalik, A. D., Tanner, G., Kocakavuk, E., Anderson, K. J., Abiola, O., Aldape, K., Alfaro, K. D., Alpar, D., Amin, S. B., Ashley, D. M., Bandopadhayay, P., Barnholtz-Sloan, J. S., Beroukhim, R., Bock, C., Brastianos, P. K., Brat, D. J., . . . GLASS Consortium. (2019). Longitudinal molecular trajectories of diffuse glioma in adults. Nature, 576(7785), 112–120. https://doi.org/10.1038/s41586-019-1775-1

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

Brennan, C. W., Verhaak, R. G. W., McKenna, A., Campos, B., Noushmehr, H., Salama, S. R., Zheng, S., Chakravarty, D., Sanborn, J. Z., Berman, S. H., Beroukhim, R., Bernard, B., Wu, C.-J., Genovese, G., Shmulevich, I., Barnholtz-Sloan, J., Zou, L., Vegesna, R., Shukla, S. A., . . . TCGA Research Network. (2013). The somatic genomic landscape of glioblastoma. Cell, 155(2), 462–477. https://doi.org/10.1016/j.cell.2013.09.034

Cancer Genome Atlas Research Network. (2012). Comprehensive genomic characterization of squamous cell lung cancers. Nature, 489(7417), 519–525. https://doi.org/10.1038/nature11404

Cancer Genome Atlas Research Network. (2023). Comprehensive molecular profiling of lung adenocarcinoma. Nature, 614(7948), 543–550. https://doi.org/10.1038/nature13385

Capper, D., Jones, D. T. W., Sill, M., Hovestadt, V., Schrimpf, D., Sturm, D., Koelsche, C., Sahm, F., Chavez, L., Reuss, D. E., Kratz, A., Wefers, A. K., Huang, K., Pajtler, K. W., Schweizer, L., Stichel, D., Olar, A., Engel, N. W., Lindenberg, K., . . . Pfister, S. M. (2018). DNA methylation-based classification of central nervous system tumours. Nature, 555(7697), 469–474. https://doi.org/10.1038/nature26000

Ceccarelli, M., Barthel, F. P., Malta, T. M., Sabedot, T. S., Salama, S. R., Murray, B. A., Morozova, O., Newton, Y., Radenbaugh, A., Pagnotta, S. M., Anjum, S., Wang, J., Manyam, G., Zoppoli, P., Ling, S., Rao, A. A., Grifford, M., Cherniack, A. D., Zhang, H., . . . Verhaak, R. G. W. (2016). Molecular profiling reveals biologically discrete subsets and pathways of progression in diffuse glioma. Cell, 164(3), 550–563. https://doi.org/10.1016/j.cell.2015.12.028

Chen, F., Zhang, Y., Parra, E., Rodriguez, J., Behrens, C., Akbani, R., Lu, Y., Kurie, J., Gibbons, D. L., Mills, D. B., Wistuba, I. I., & Geraci, M. (2023). Proteomic landscape of non-small-cell lung cancer: Implications for biomarker-driven treatment strategies. Frontiers in Oncology, 13, Article 1081573. https://doi.org/10.3389/fonc.2023.1081573

Davis, S., & Meltzer, P. S. (2007). GEOquery: A bridge between the Gene Expression Omnibus (GEO) and BioConductor. Bioinformatics, 23(14), 1846–1847. https://doi.org/10.1093/bioinformatics/btm254

De Zuani, M. D., Xue, H., Park, J. S., Dentro, S. C., Seferbekova, Z., Tessier, J., Curras-Alonso, S., Hadjipanayis, A., Athanasiadis, E. I., Gerstung, M., Bayraktar, O., & Cvejic, A. (2024). Single-cell and spatial transcriptomics analysis of non-small cell lung cancer. Nature Communications, 15(1), 4388. https://doi.org/10.1038/s41467-024-48700-8

Dias, R., & Torkamani, A. (2019). Artificial intelligence in clinical and genomic diagnostics. Genome Medicine, 11(1), 70. https://doi.org/10.1186/s13073-019-0689-8

Facchinetti, F., Proto, C., Minari, R., Garassino, M., & Tiseo, M. (2018). Mechanisms of resistance to target therapies in non-small cell lung cancer. In Handbook of Experimental Pharmacology, 249, 63–89. https://doi.org/10.1007/164_2017_16

Fan, J., Slowikowski, K., & Zhang, F. (2020). Single-cell transcriptomics in cancer: Computational challenges and opportunities. Experimental and Molecular Medicine, 52(9), 1452–1465. https://doi.org/10.1038/s12276-020-0422-0

Garg, P., Singhal, S., Kulkarni, P., Horne, D., Malhotra, J., Salgia, R., & Singhal, S. S. (2024). Advances in non-small cell lung cancer: Current insights and future directions. Journal of Clinical Medicine, 13(14), 4189. https://doi.org/10.3390/jcm13144189

Gihawi, A., Cardenas, R., Hurst, R., & Brewer, D. S. (2023). Quality control in metagenomics data. Methods in Molecular Biology, 2649, 21–54. https://doi.org/10.1007/978-1-0716-3072-3_2

Govindan, R., Ding, L., Griffith, M., Subramanian, J., Dees, N. D., Kanchi, K. L., Maher, C. A., Fulton, R., Fulton, L., Wallis, J., Chen, K., Walker, J., McDonald, S., Bose, R., Ornitz, D., Xiong, D., You, M., Dooling, D. J., Watson, M., . . . Wilson, R. K. (2012). Genomic landscape of non-small cell lung cancer in smokers and never-smokers. Cell, 150(6), 1121–1134. https://doi.org/10.1016/j.cell.2012.08.024

Gridelli, C., Rossi, A., Carbone, D. P., Guarize, J., Karachaliou, N., Mok, T., Petrella, F., Spaggiari, L., & Rosell, R. (2015). Non-small-cell lung cancer. Nature Reviews. Disease Primers, 1, Article 15009. https://doi.org/10.1038/nrdp.2015.9

Hasan, G. M., Hassan, M. I., Sohal, S. S., Shamsi, A., & Alam, M. (2023). Therapeutic targeting of regulated signaling pathways of non-small cell lung carcinoma. ACS Omega, 8(30), 26685–26698. https://doi.org/10.1021/acsomega.3c02424

Hirsch, F. R., Scagliotti, G. V., Mulshine, J. L., Kwon, R., Curran, Jr., W. J., Wu, Y.-L., & Paz-Ares, L. (2017). Lung cancer: Current therapies and new targeted treatments. The Lancet, 389(10066), 299–311. https://doi.org/10.1016/S0140-6736(16)30958-8

Hou, J., Aerts, J., den Hamer, B., van Ijcken, W., den Bakker, M., Riegman, P., van der Leest, C., van der Spek, P., Foekens, J. A., Hoogsteden, H. C., Grosveld, F., & Philipsen, S. (2010). Gene expression-based classification of non-small cell lung carcinomas and survival prediction. PLOS One, 5(4), Article e10312. https://doi.org/10.1371/journal.pone.0010312

Huang, S., Yang, J., Shen, N., Xu, Q., & Zhao, Q. (2023). Artificial intelligence in lung cancer diagnosis and prognosis: Current application and future perspective. Seminars in Cancer Biology, 89, 30–37. https://doi.org/10.1016/j.semcancer.2023.01.006

Hugo, W., Zaretsky, J. M., Sun, L., Song, C., Moreno, B. H., Hu-Lieskovan, S., Berent-Maoz, B., Pang, J., Chmielowski, B., Cherry, G., Seja, E., Lomeli, S., Kong, X., Kelley, M. C., Sosman, J. A., Johnson, D. B., Ribas, A., & Lo, R. S. (2016). Genomic and transcriptomic features of response to anti-PD-1 therapy in metastatic melanoma. Cell, 165(1), 35–44. https://doi.org/10.1016/j.cell.2016.02.065

Klughammer, J., Kiesel, B., Roetzer, T., Fortelny, N., Nemc, A., Nenning, K.-H., Furtner, J., Sheffield, N. C., Datlinger, P., Peter, N., Nowosielski, M., Augustin, M., Mischkulnig, M., Ströbel, T., Alpar, D., Ergüner, B., Senekowitsch, M., Moser, P., Freyschlag, C. F., . . . Bock, C. (2018). The DNA methylation landscape of glioblastoma disease progression shows extensive heterogeneity in time and space. Nature Medicine, 24(10), 1611–1624. https://doi.org/10.1038/s41591-018-0156-x

Kuner, R. (2013). Lung cancer gene signatures and clinical perspectives. Microarrays, 2(4), 318–339. https://doi.org/10.3390/microarrays2040318

Li, B., Cui, Y., Diehn, M., & Li, R. (2017). Development and validation of an individualized immune prognostic signature in early-stage nonsquamous non-small cell lung cancer. JAMA Oncology, 3(11), 1529–1537. https://doi.org/10.1001/jamaoncol.2017.1609

Li, T., Kung, H.-J., Mack, P. C., & Gandara, D. R. (2013). Genotyping and genomic profiling of non-small-cell lung cancer: Implications for current and future therapies. Journal of Clinical Oncology, 31(8), 1039–1049. https://doi.org/10.1200/JCO.2012.45.3753

Lococo, F., Ghaly, G., Chiappetta, M., Flamini, S., Evangelista, J., Bria, E., Stefani, A., Vita, E., Martino, A., Boldrini, L., Sassorossi, C., Campanella, A., Margaritora, S., & Mohammed, A. (2024). Implementation of artificial intelligence in personalized prognostic assessment of lung cancer: A narrative review. Cancers, 16(10), 1832. https://doi.org/10.3390/cancers16101832

Lu, Y., Lemon, W., Liu, P.-Y., Yi, Y., Morrison, C., Yang, P., Sun, Z., Szoke, J., Gerald, W. L., Watson, M., Govindan, R., & You, M. (2006). A gene expression signature predicts survival of patients with stage I non-small cell lung cancer. PLOS Medicine, 3(12), e467. https://doi.org/10.1371/journal.pmed.0030467

Ma, S., Zhu, J., Wang, M., Han, T., Zhu, J., Jiang, R., & Jiang, T. (2022). A comprehensive characterization of alternative splicing events related to prognosis and the tumor microenvironment in lung adenocarcinoma. Annals of Translational Medicine, 10(8), 479. https://doi.org/10.21037/atm-22-1531

Michelotti, A., de Scordilli, M., Bertoli, E., De Carlo, E., Del Conte, A., & Bearz, A. (2022). NSCLC as the paradigm of precision medicine at its finest: The rise of new druggable molecular targets for advanced disease. International Journal of Molecular Sciences, 23(12), 6748. https://doi.org/10.3390/ijms23126748

Morganti, S., Tarantino, P., Ferraro, E., D'Amico, P., Duso, B. A., & Curigliano, G. (2019). Next generation sequencing (NGS): A revolutionary technology in pharmacogenomics and personalized medicine in cancer. Advances in Experimental Medicine and Biology, 1168, 9–30. https://doi.org/10.1007/978-3-030-24100-1_2

Pan, E., & Bai, Y. (2024). Insight into NSCLC through novel analysis of gene interactions and characteristics. American Journal of Clinical and Experimental Immunology, 13(2), 58–67. https://doi.org/10.62347/ANLV4963

Parakh, S., Leong, T. L., Best, S. A., & Poh, A. R. (2023) [Editorial]. Editorial: Overcoming drug relapse and therapy resistance in NSCLC. Frontiers in Oncology, 13, Article 1230475. https://doi.org/10.3389/fonc.2023.1230475

Park, M.-K., Lim, J.-M., Jeong, J., Jang, Y., Lee, J.-W., Lee, J.-C., Kim, H., Koh, E., Hwang, S.-J., Kim, H.-G., & Kim, K.-C. (2022). Deep-learning algorithm and concomitant biomarker identification for NSCLC prediction using multi-omics data integration. Biomolecules, 12(12), 1839. https://doi.org/10.3390/biom12121839

Richardson, S., Tseng, G. C., & Sun, W. (2016). Statistical methods in integrative genomics. Annual Review of Statistics and Its Application, 3, 181–209. https://doi.org/10.1146/annurev-statistics-041715-033506

Ruan, X., Ye, Y., Cheng, W., Xu, L., Huang, M., Chen, Y., Zhu, J., Lu, X., & Yan, F. (2022). Multi-omics integrative analysis of lung adenocarcinoma: An *in silico* profiling for precise medicine. Frontiers in Medicine, 9, Article 894338. https://doi.org/10.3389/fmed.2022.894338

Sanaei, M.-J., Razi, S., Pourbagheri-Sigaroodi, A., & Bashash, D. (2022). The PI3K/Akt/mTOR pathway in lung cancer; oncogenic alterations, therapeutic opportunities, challenges and a glance at the application of nanoparticles. Translational Oncology, 18, Article 101364. https://doi.org/10.1016/j.tranon.2022.101364

Skoulidis, F., & Heymach, J. V. (2019). Co-occurring genomic alterations in non-small-cell lung cancer biology and therapy. Nature Reviews. Cancer, 19(9), 495–509. https://doi.org/10.1038/s41568-019-0179-8

Sun, X. X., & Yu, Q. (2015). Intra-tumor heterogeneity of cancer cells and its implications for cancer treatment. Acta Pharmacologica Sinica, 36(10), 1219–1227. https://doi.org/10.1038/aps.2015.92

Touat, M., Idbaih, A., Sanson, M., & Ligon, K. L. (2017). Glioblastoma targeted therapy: Updated approaches from recent biological insights. Annals of Oncology, 28(7), 1457–1472. https://doi.org/10.1093/annonc/mdx106

Travis, W. D., Brambilla, E., Nicholson, A. G., Yatabe, Y., Austin, J. H. M., Beasley, M. B., Chirieac, L. R., Dacic, S., Duhig, E., Flieder, D. B., Geisinger, K., Hirsch, F. R., Ishikawa, Y.,

Kerr, K. M., Noguchi, M., Pelosi, G., Powell, C. A., Tsao, M. S., Wistuba, I., & WHO Panel. (2015). The 2015 World Health Organization classification of lung tumors: Impact of genetic, clinical and radiologic advances since the 2004 classification. Journal of Thoracic Oncology, 10(9), 1243–1260. https://doi.org/10.1097/JTO.0000000000000630

Verhaak, R. G. W., Hoadley, K. A., Purdom, E., Wang, V., Qi, Y., Wilkerson, M. D., Miller, C. R., Ding, L., Golub, T., Mesirov, J. P., Alexe, G., Lawrence, M., O'Kelly, M., Tamayo, P., Weir, B. A., Gabriel, S., Winckler, W., Gupta, S., Jakkula, L., . . . Cancer Genome Atlas Research Network. (2010). Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR and NF1. Cancer Cell, 17(1), 98–110. https://doi.org/10.1016/j.ccr.2009.12.020

Wang, Q., Hu, B., Hu, X., Kim, H., Squatrito, M., Scarpace, L., deCarvalho, A. C., Lyu, S., Li, P., Li, Y., Barthel, F., Cho, H. J., Lin, Y.-H., Satani, N., Martinez-Ledesma, E., Zheng, S., Chang, E., Sauvé, C. G., Olar, A., . . . Verhaak, R. G. W. (2017). Tumor evolution of glioma-intrinsic gene expression subtypes associates with immunological changes in the microenvironment. Cancer Cell, 32(1), 42–56.e6. https://doi.org/10.1016/j.ccell.2017.06.003

Wang, Y., Miao, Z., Qin, X., Yang, Y., Wu, S., Miao, Q., Li, B., Zhang, M., Wu, P., Han, Y., & Li, B. (2023). Transcriptomic landscape based on annotated clinical features reveals PLPP2 involvement in lipid raft-mediated proliferation signature of early-stage lung adenocarcinoma. Journal of Experimental and Clinical Cancer Research, 42(1), 315. https://doi.org/10.1186/s13046-023-02877-w

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(2), 931–939. https://doi.org/10.1007/s11010-020-03959-5

Zhang, X., Jonassen, I., & Goksøyr, A. (2021). Machine learning approaches for biomarker discovery using gene expression data. In H. I. Nakaya (Ed.).Bioinformatics, 20 (pp. 53–64). Exon Publications. https://doi.org/10.36255/exonpublications.bioinformatics.2021.ch4

Cite this article: Alghamdi RM. Multi-Dataset Identification and Validation of New Gene Expression Signatures: Insights into Matrix Remodeling Pathways in NSCLC. Int. J. Pharm. Investigation. 2025;15(4):1235-48.