Gene Regulatory Network Based Biomarker Transition from Normalcy to Malignancy: A study of Lung Cancer Using Differential Gene Expression Data

Supriya Mandal[1], Surama Biswas[1,*]

Abstract: Lung cancer is a global threat, and understanding its pathology through gene regulatory network (GRN)-based biomarkers could offer significant insights. In this study, GRN-based biomarkers have been constructed to demonstrate the topology shift of GRN from a normal condition to a lung cancer condition. This study employed three distinct processes to identify significant genes from differential gene expression data related to lung cancer: 1) simple variance-based screening, 2) a recent gene selection algorithm incorporating simulated annealing and a biologically inspired objective function, and 3) a volcano plot comparing log2 fold changes against -log10 P values. These comprehensive methods identified 21 common genes as significant genes, including KRT5, SFTPC, KRT16, and AKR1B10, which have strong associations with lung cancer pathology according to previous literature. To gain deeper insights into the underlying biological mechanisms, co-expression-modelled GRNs have been constructed for both normal and lung cancer conditions based on these significant genes. These networks provide a detailed visual and mathematical representation of gene interactions within the cellular environment, illustrating how these interactions differ between healthy and diseased states. The shift in network biomarker topologies from normal to diseased conditions have been closely analysed using centrality measures, which quantify the importance of genes based on their connectivity and position within the network. Notably, the pronounced change in the degree of centrality for genes such as GJB6 between normal and diseased states suggests their pivotal roles in the transition to lung cancer. Rigorous statistical and machine learning methods validated the significance of the selected genes and their interconnections, ensuring the robustness of the findings.

1. **Introduction**:

One of the most common type of cancer in the world is lung cancer. Of all cancer cases, it makes up 12.3%. The number of new cases of lung cancer reported has doubled, from about 1.2 million in 2000 to over 2.4 million in 2022. Historically, lung cancer is the most frequent cancer in men and it has high morbidity and mortality in both men and women. Its occurrence is significantly impacted by demographic and geographic differences. Lung cancer is highly preventable by reducing smoking habits and making sustainable measures to improve air quality. Though many preventative measures have already been implemented, the risk never goes back to normal and fatal conditions persist with a 90% patient death rate [1]. Current research focus in lung cancer is to study the pathology, early detection and prognosis mechanisms, and understand how different environmental and genetic variables impact the disease [2]. After the COVID-19 pandemic, patients with lung cancer should have poorer prognoses, as they have more risks of death from the virus and may experience severe consequences through contamination. Accordingly, a recent study claimed that individuals with COVID-19 who developed lung cancer had far worse death rates than the COVID-19 community as a whole [3 - 4].

Recent advancements in lung cancer research have significantly improved the understanding of the disease through the application of genomic and proteomic technologies. These technologies offer comprehensive insights beyond single-gene studies, enhance our knowledge of lung carcinogenesis, and have the potential to improve diagnosis, prognosis, and patient management [5]. Genomic research has focused on understanding the histologic differentiation and progression of lung cancer, revealing critical genetic alterations that drive the disease [6]. The emergence of individualized diagnostics and treatments has revolutionized lung cancer care with specific therapies targeting genetic alterations. Such targeted alterations are found in Estimated Glomerular Filtration Rate (EGFR)-mutant adenocarcinoma that leads to more effective treatments compared to traditional chemotherapy [7]. Genomic testing has become integral in identifying mutations and resistance mechanisms in lung cancer, aiding in the development of precision medicine approaches that deliver treatments to individual genetic profiles. This testing is now a part of national guidelines due to its demonstrated benefits in improving patient outcomes [8]. Furthermore, studies on intra-tumour heterogeneity (ITH) have shown that transcriptomic diversity within

tumours contributes to immune evasion and therapy resistance. By analysing paired whole-exome and RNA sequencing data, researchers have linked gene expression patterns to tumour evolution and metastasis, highlighting the complex interplay between the genome and transcriptome in lung cancer progression [9].

Gene regulatory network (GRN) reconstruction is pivotal in systems biology, providing insights into gene interactions and regulatory mechanisms within cells. Utilizing correlation and regression analyses of multifactorial perturbation data, a method decomposes the GRN inference problem into different regression tasks, predicting target gene expression levels from potential regulator genes. This approach, integrating correlation coefficients and sum of squared residuals, has shown promising results in accurately estimating gene regulations [10]. GRNs serve as comprehensive maps reflecting genetic and epigenetic influences, aiding in understanding the functions of individual molecules and overall cellular organization [11]. Differential gene regulatory networks capture variable gene expression under different conditions, crucial for studying development, organ formation, and disease progression through high-throughput sequencing data and gene-gene expression correlation analysis [12]. Evaluations of GRN inference algorithms from single-cell transcriptomic data, like BEELINE, highlight the superior performance of methods not requiring pseudo time ordering of cells, improving the precision and applicability of GRN inference [13]. Co-expression clusters, indicating genes involved in common functions, are integral to understanding regulatory networks, though the direct regulatory proximity assumption needs further validation [14]. Enhancing GRN inference involves considering network structure, with co-expression-based methods suited for inferring function-specific regulons and regulatory interaction-focused methods excelling in global GRN inference. Effective data integration requires balancing dataset size with potential noise inclusion [15]. Advanced models like DeepRIG, employing graph auto-encoder techniques to embed global regulatory information, demonstrate superior performance in reconstructing accurate GRNs from single-cell RNA-seq data, identifying novel regulatory elements specific to different cell types and disease conditions, showcasing the potential of graph-based approaches in GRN inference [16].

GRN analysis has become instrumental in identifying biomarkers for various diseases, including sepsis and cancer. By focusing on microRNA (miRNA) regulatory networks, researchers have identified novel miRNA biomarkers associated with the early diagnosis of sepsis, enhancing the sensitivity and specificity of sepsis diagnostics [17]. Similarly, GRN-based approaches have revealed critical biomarkers for ovarian and lung cancers, such as RAD51AP1, which is overexpressed in these cancers and correlates with poorer overall survival. This finding, supported by gene suppression studies, underscores the role of RAD51AP1 in cancer proliferation and highlights its potential as a prognostic biomarker [18]. In bladder cancer, the construction of a competing endogenous RNA (ceRNA) regulatory network, involving lncRNA-miRNA-mRNA interactions, has identified specific gene signatures that can predict patient prognosis. This ceRNA network, developed using data from The Cancer Genome Atlas (TCGA), has provided insights into the molecular mechanisms underlying bladder cancer and has established a model for prognostic prediction [19].

Network-based approaches have significantly advanced the identification and understanding of biomarkers for cancer which facilities researchers to explore gene interactions and regulatory mechanisms of the disease. These methodologies integrate human transcriptome and interactome data and prioritize genes in cancer-specific networks that lead to the identification of cancer genes crucial for tumorigenesis and progression. These genes, showing extensive perturbations, contribute to accurate tumor diagnosis and patient survival prediction [20]. Utilizing microarray gene expression profiles and protein-protein interaction information, network-based biomarkers have been developed to explore lung carcinogenesis mechanisms, highlighting pathways and potential therapeutic targets [21]. The Gene Sub-Network-based Feature Selection (GSNFS) approach, incorporating novel searching and scoring algorithms which has further refined subnetwork identification, improve classification performance and outperform across datasets as compared to traditional greedy algorithms [22]. In lung adenocarcinoma (LUAD), Weighted Gene Coexpression Network Analysis (WGCNA) has identified key gene modules and hub genes related to cancer stem cell features, with prognostic signatures constructed from these hub genes demonstrating high predictive value [23]. Moreover, transcriptional regulatory networks have been constructed to identify dysregulated transcription factors in lung cancer, revealing co-expressed and co-regulated patterns linked to critical biological processes and cancer hallmarks. These dysregulated transcription factors,

validated through enrichment analysis and survival correlation studies, offer potential for early diagnosis and personalized therapies [24].

This research introduces the development of gene regulatory network (GRN)-based biomarkers aimed at illustrating the transformation of network biomarkers between normal and lung cancer conditions. The study leveraged data on differential gene expression in lung cancer to identify three sets of significant genes using distinct methodologies: 1) a straightforward variance-based screening approach, 2) a novel gene selection algorithm that optimizes the mean collective difference between healthy and diseased states through simulated annealing and a biologically inspired objective function [25], and 3) analysis employing a volcano plot (log2 fold change vs. -log10 P values). Through these approaches, 21 common genes, including KRT5, SFTPC, KRT16, and AKR1B10, emerged as pertinent markers, each previously associated with lung cancer pathology in existing literature. These identified genes were employed to construct co-expression-modeled gene regulatory networks for both normal and lung cancer scenarios. The shift in network biomarker topologies from health to disease states was scrutinized using centrality measurements. Comparative analysis revealed substantial differences in gene interaction patterns and centrality between healthy and diseased states. Notably, genes such as KRT5 and KRT16 exhibited higher centrality degrees within the lung cancer network, indicating heightened regulatory roles in diseased conditions. Conversely, genes like MT1M and CLDN18 consistently displayed low centrality scores across both conditions, suggesting a limited involvement in disease progression. The notable shift in centrality measures between the healthy and sick states for genes, such as GJB6, that are not thought to be significant in relation to lung cancer was of special interest. This change suggests that these genes may have some role in the development or progression of lung cancer. Rigorous statistical and machine learning analyses were conducted to ensure the robustness and significance of the identified genes and their interrelationships within the constructed networks. This comprehensive approach underscores the potential of GRN-based biomarkers in enhancing our understanding of lung cancer pathogenesis and in identifying novel therapeutic targets and diagnostic markers.

2. **Methodologies**:

Gene regulatory networks (GRNs) in healthy and diseased lung tissues are compared and analysed in this study in an effort to improve our understanding of lung cancer. Data collection and pre-processing, identification of significant genes, GRN inference, centrality analysis, and statistical validation are some of the crucial processes in the proposed work. By identifying important regulatory genes and the ways in which they interact in healthy vs. diseased conditions, the proposed methodology aims to provide insights that may be used to guide future developments in lung cancer diagnosis and treatment.

2.1. Data Collection and pre-processing: Differential gene expression datasets contain gene expression profiles that highlight differences in gene activity such as amount of protein coding between various biological conditions, for example, healthy versus diseased tissues. These datasets are available from various public database like National Center for Biotechnology Information (NCBI)'s Gene Expression Omnibus (GEO) Database [26 – 27] and provide a strong basis for studying differential gene regulation since they include gene expression profiles from diseased (e.g. lung cancer) and healthy tissue samples. To guarantee correctness and dependability, a thorough pre-processing pipeline is applied to the raw data. In order to reduce potential sources of mistake, data cleaning first entails deleting unnecessary columns, such as "Probe Id," and rows that lack gene names. After that, to account for technological differences and guarantee sample comparability, normalization techniques are applied to the gene expression data to map quantities between [0 – 1]. Finally, a transpose operation has been performed and as an outcome the data matrix is obtained for which the rows represent normal or diseased samples, the columns represent genes and elements are expression levels (numeric quantities between [0 – 1]).

2.2. Identification of Significant Genes: Three separate approaches have been used in parallel to identify a small number of significant genes from the massive dataset that contains close to 50,000 genes: 1) screening of genes based on variance, 2) a recent optimization method of gene selection known as Differential Gene Expression based Simulated Annealing (DGESA) [25], and 3) a Volcano plot method that applied statistical basis of significant gene identification.

2.2.1. *Screening of Genes based on Variance*: Variance-based gene identification is a crucial step in understanding the differential expression of genes across samples. The variance of each gene's expression levels is calculated to measure how much these levels deviate from their mean value. This involves computing the mean expression level for each gene (see Eq. 1) and then determining the variance using Eq. 2, which quantifies the spread of expression values around this mean. Genes are then sorted in descending order based on their variance. Those with the highest variance are selected for further analysis, as they are likely to exhibit significant differences in expression between healthy and lung cancer tissues. This method ensures that the focus remains on genes with the most substantial variation, which are potentially the most biologically relevant for distinguishing between the different tissue types.

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad (1)$$

$$\sigma^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu)^2 \qquad (2)$$

Here, μ is the mean expression value, n is the total number of healthy and diseased samples respectively, x_i is expression value of i-th sample of current gene and σ^2 is the computed variance of current gene across all healthy and diseased sample.

2.2.2. *Gene selection using Differential Gene Expression based Simulated Annealing (DGESA)*: A recent algorithm, named Differential Gene Expression based Simulated Annealing (DGESA) [25] has been implemented here for significant gene selection of lung cancer. DGESA is based on the famous meta-heuristics algorithm, called Simulated Annealing (SA) where a concept of metallurgy referred as annealing is simulated to provide heuristic information traverse and assess the points in a search space of an optimization process against a properly designed objective function. In DGESA, this objective function is formulated in an efficient way (see Eq. 3) to maximize cumulative mean difference of differential expression profiles (in healthy vs. lung cancer samples) across all genes in a set (i.e. candidate solution comprising a set of genes index selected from the dataset). DGESA starts with setting few algorithmic parameters like temperature (T) set to a very high value, a cooling rate (C) and a randomly selected set of gene index into the initial candidate solution (s) of a fixed length. The solution s is evaluated by Eq. 3 and undergone few iterations. In each iteration, few neighbourhood search steps followed by a temperature reduction step using Eq. 4. In a neighbourhood search step, a new solution s' is generated from s using perturbation. If objective function value of s' is better than that of s, s' replaces s, called a movement. Otherwise, s' has a probability (p) dependent scope to replace s. DGESA terminates if maximum of iterations exceeded or over iterations the objective function value of successive candidate solution achieve convergence.

$$f(s) = \sum_{i=1}^{n}\left|\overline{LC(x_i)} - \overline{H(x_i)}\right| \qquad (3)$$

$$p = e^{-\Delta f/T} \qquad (4)$$

Here, $f(s)$ is the objective function value calculated upon s, $\left|\overline{LC(x_i)} - \overline{H(x_i)}\right|$ is the absolute value of the quantity, $\overline{LC(x_i)} - \overline{H(x_i)}$ where $LC(x_i)$ is the mean of all lung cancer expression values of current gene in s and $\overline{H(x_i)}$ is the mean of all healthy expression values of the same.

2.2.3. *Gene Identification through Volcano Plot*: For identification of lung cancer genes, volcano plot has been generated here. A volcano plot is a type of scatter plot used in genetics and genomics to visualize the results of a differential expression analysis. It's called a volcano plot because the shape of the plot resembles a volcano, with a large number of genes or transcripts at the base and a smaller number of highly differentially expressed genes at the top. The x-axis typically represents the fold change or log2 fold change of gene expression between two conditions (e.g., lung cancer and healthy states), while the y-axis represents the statistical significance of the difference, and has been expressed as a -log10 (p-value). Genes with large fold changes and high statistical significance appear in the top left and right corners of the plot, making it easy to identify the most differentially expressed genes. Volcano plots are useful for identifying genes that are significantly up- or down-regulated between different conditions.

2.3. GRN Inference: Using the selected genes, GRNs have been constructed for both healthy and lung cancer tissue samples through a co-expression model. This process begins with a correlation analysis, where Pearson correlation coefficients are calculated (see Eq. 3) for each pair of genes to assess the strength and direction of their linear relationships. Significant correlations are identified and used to define edges in the networks, indicating potential regulatory interactions between the genes. These GRNs are then visualized using network analysis tools, which illustrate the complex web of interactions and regulatory pathways. By comparing the GRNs of healthy and lung cancer tissues, differences in gene interactions and regulatory mechanisms can be identified, providing insights into how gene regulation is altered in disease states. This approach helps to uncover key regulatory genes and pathways that may contribute to disease development and progression, facilitating the identification of potential targets for therapeutic intervention.

2.4. Centrality Analysis: Centrality measures are metrics used to identify the most important nodes within a network. Here, it is performed on the GRNs to determine each gene's regulatory importance. Centrality measures that are considered here include:

a) Degree Centrality: This calculates the number of direct connections each gene has, indicating its immediate influence within the network.
b) Closeness Centrality: This assesses how quickly a gene can influence the entire network, calculated as the reciprocal of the average shortest path length from a gene to all others, reflecting its overall reach.
c) Betweenness Centrality: This evaluates how often a gene acts as a bridge within the network, determined by the number of shortest paths between gene pairs that pass through it, highlighting its role in facilitating interactions.

By computing these centrality measures, key regulatory genes with significant influence, reach, and bridging roles in the GRNs of healthy and diseased tissues are identified, aiding in the discovery of potential therapeutic targets and biomarkers.

2.5. Comparative Analysis between Healthy and Diseased Network Biomarkers: In the Comparative Analysis, centrality measures of genes in both healthy and diseased Gene Regulatory Networks (GRNs) are compared to identify significant differences in gene regulation. This process involves identifying key regulatory genes, specifically those with high centrality measures in the diseased network and notable changes from their centrality in the healthy network. These genes are highlighted as potential key regulators of disease. Additionally, a comparison of the overall network topology is conducted to examine the structure and connectivity patterns of the GRNs. This helps to understand how lung cancer impacts gene regulation by revealing alterations in network organization and interactions. Through this comparative analysis, crucial network biomarkers can be identified, providing insights into the underlying mechanisms of lung cancer and potentially uncovering targets for therapeutic intervention. All steps of the methodology are comprehensively summarized in Figure 1.

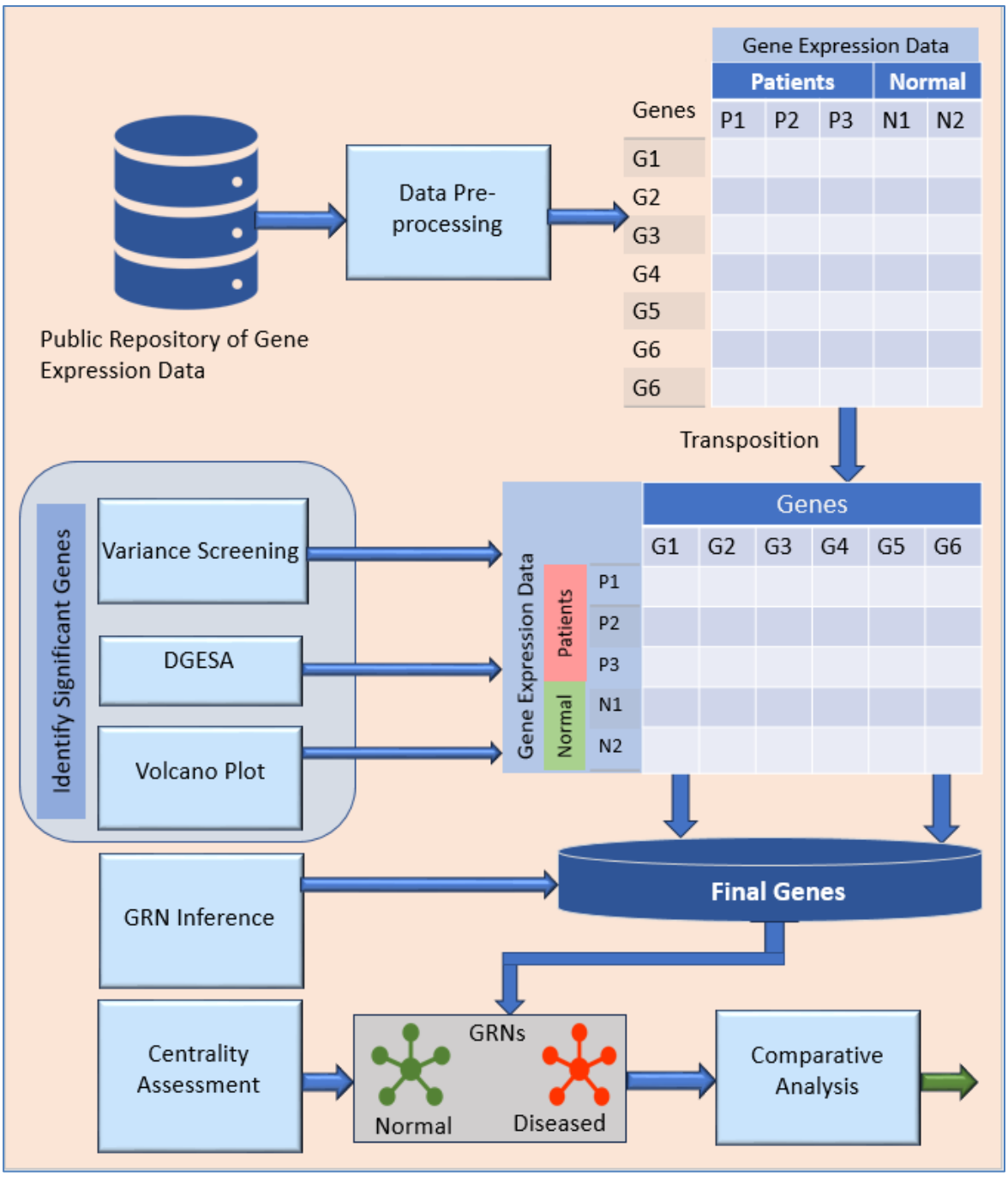

Figure 1. ***Overall Framework for Comparative Analysis of Gene Regulatory Networks (GRNs) in Lung Cancer****. The framework involves five key steps: (1) Data Collection and Pre-processing of gene expression datasets from public databases like GEO, followed by normalization; (2) Identification of Significant Genes using variance-based screening, DGESA, and Volcano plots; (3) GRN Inference using co-expression models to build networks for healthy and lung cancer tissues; (4) Centrality Analysis to identify key regulatory genes based on degree, closeness, and betweenness; and (5) Comparative Analysis to highlight differences in network structure and regulatory roles between healthy and diseased conditions.*

3. **Results**:

3.1. Data and materials

We collected two datasets of non-small cell lung cancer from the GEO database, NCBI, with accession numbers GSE18842 and GSE19188, will be referred as first and the second dataset hereafter. The first dataset is of human lung cancer, specifically non-small cell lung cancer (NSCLC), using microarray

technology. It includes expression profiles from 91 samples—46 tumours and 45 controls. The second dataset corresponds to a genome-wide gene expression acquisition of 91 patients with non-small cell lung cancer (NSCLC), analysing 91 tumour samples and 65 adjacent normal lung samples. The datasets undergo the pre-processing and transposition processes as outlined in Section 2.1. Python 3.12.2 and libraries including Pandas, NumPy, Scikit-learn, NetworkX, Statsmodels, Plotly, Dash, Venn, Matplotlib, and Seaborn have been employed for all analyses in this research. Detailed mappings of each library to its corresponding analysis can be found in the supplementary materials.

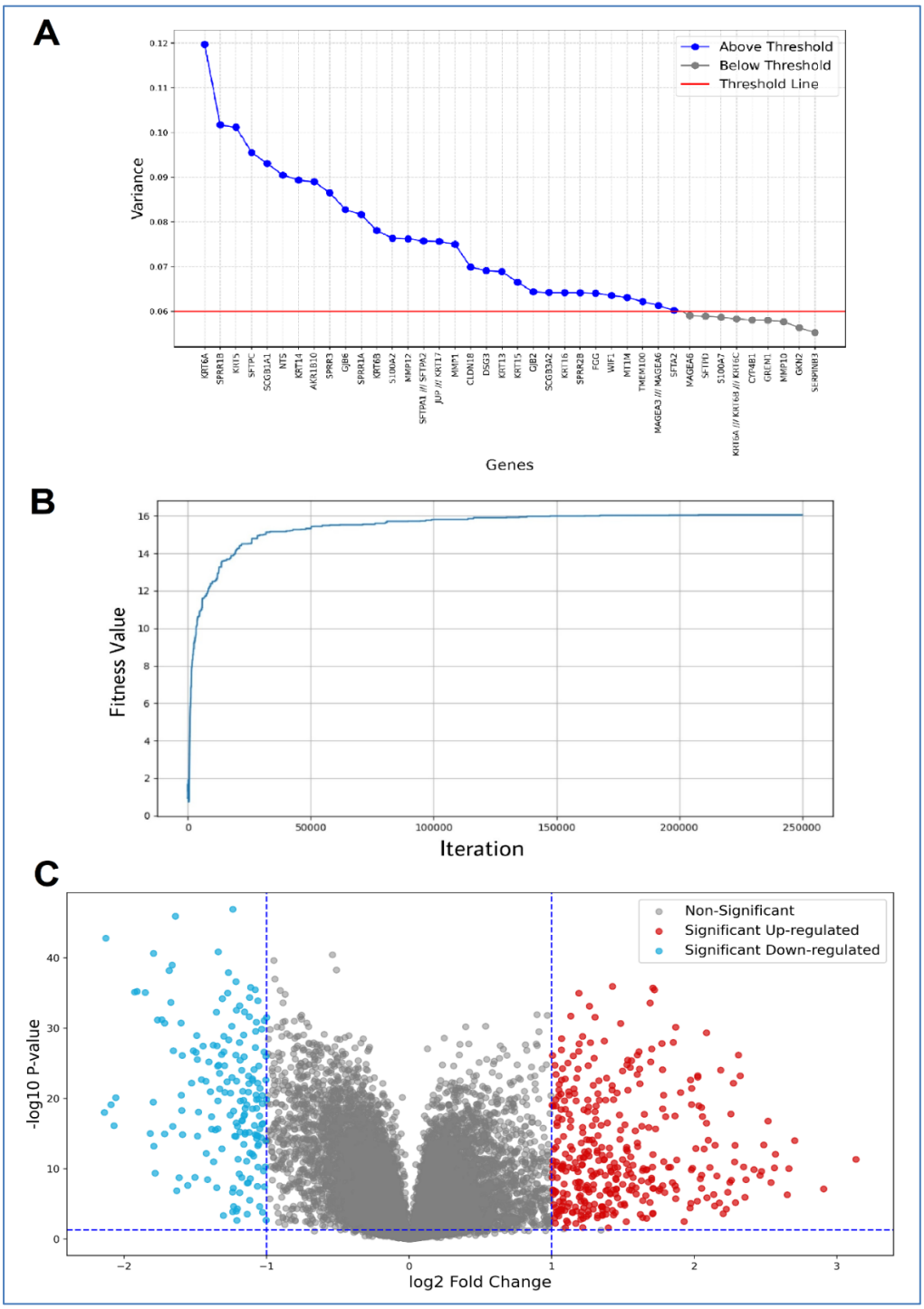

Figure 2. Gene Selection Methodologies: Variance Graph, Optimization Curve, and Volcano Plot. Panel A illustrates the variance graph, showcasing the distribution of gene expression levels across samples. Panel B displays the optimization curve from the Differential Gene Expression based Simulated Annealing (DGESA) algorithm, highlighting the convergence towards

an optimal solution over iterations. Panel C presents the volcano plot, visualizing the log2 fold change against the -log10 p-values of gene expression, identifying 551 significantly upregulated and downregulated genes.

3.2. Gene Selection

Here, we present the results of gene selection, network inference, and analysis. Through three different gene selection methodologies applied to both datasets (GSE18842 and GSE19188)—variance-based screening, a simulated annealing-based algorithm (DGESA), and volcano plot analysis—we identified significant genes associated with lung cancer. These genes were used to construct gene regulatory networks (GRNs) for both healthy and diseased conditions. Comparative analysis of the GRNs revealed key differences in gene interactions and regulatory roles, highlighting potential biomarkers and targets for lung cancer diagnosis and treatment.

In variance-based gene prioritization, we selected the top 40 genes and visualized the data using a line diagram. Upon analysing the graph, we established a threshold value of 0.06, with genes exhibiting values above this threshold categorized as the first set of significant genes (see panel A of Figure 2). This approach resulted in the identification of 31 genes in the first dataset and 33 in the second dataset that meet the criteria for further analysis. The variance-based method effectively highlights genes with substantial differential expression, providing a robust foundation for subsequent.

The DGESA algorithm has been implemented following the tuning of its parameters. The main loop of the algorithm iterates 250,000 times, with the temperature gradually decreasing according to a cooling rate of 0.95. This systematic reduction in temperature facilitates convergence towards an optimal solution while still allowing exploration of different gene sets. Throughout the iterations, fitness values have been recorded, enabling the visualization of the optimization process (see panel B of Figure 2). For this optimization, a total of 40 genes have been selected, yielding a final fitness value of 16.0471. The algorithm and fitness function utilized in this study are consistent with those described in reference [25].

The implementation of the volcano plot has been carried out using Python, providing a clear visualization of differentially expressed genes in the context of lung cancer. By utilizing libraries such as Matplotlib and Seaborn, the plot was generated to illustrate the relationship between the log2 fold change and the -log10 p-values of gene expression. This graphical representation effectively highlights genes that are significantly upregulated or downregulated, with distinct markers for each category (see panel C of Figure 2). In this analysis, a total of 551 genes were identified as significantly regulated from the first dataset (98 in case of the second dataset), facilitating further exploration of their roles in lung cancer pathology.

The analysis of gene selection across two datasets involved three methodologies: variance-based screening, DGESA, and volcano plot analysis. In the first dataset, a total of 31, 40, and 551 genes were identified through these approaches, leading to the discovery of 21 common genes, as illustrated in the Venn diagram in Panel A (see Figure 3 and Table 1). According to published research, 17 of these 21 genes have a substantial relationship with lung cancer. Similarly, in the second dataset, 33, 40, and 98 genes were identified, resulting in 20 common genes, depicted in Panel B (see Figure 3 and Table 2). Panel C (see Figure 3 and Table 3) examines the overlap between genes from both datasets by DGESA, revealing 19 shared genes from the total of 40 genes identified. Finally, Panel D (see Figure 3 and Table 4) illustrates the overlap of common genes identified in Panels A and B, indicating 5 genes that are consistently present across both datasets. All other relevant genes may be obtained in Supplement 1.

The clustered expression heat map provides a comprehensive visual representation of the selected genes from the first dataset, showcasing their expression levels in normal and diseased conditions. This heat map includes genes identified through variance-based screening, DGESA, and volcano plot analysis, allowing for easy comparison of expression profiles. Each row represents a significant gene, while each column corresponds to individual samples, illustrating distinct expression patterns associated with healthy and lung cancer conditions. This approach enhances the visual understanding of how selected genes behave across different conditions, facilitating insights into their potential roles in lung cancer biology. For clarity, all subsequent analyses will focus on this first dataset.

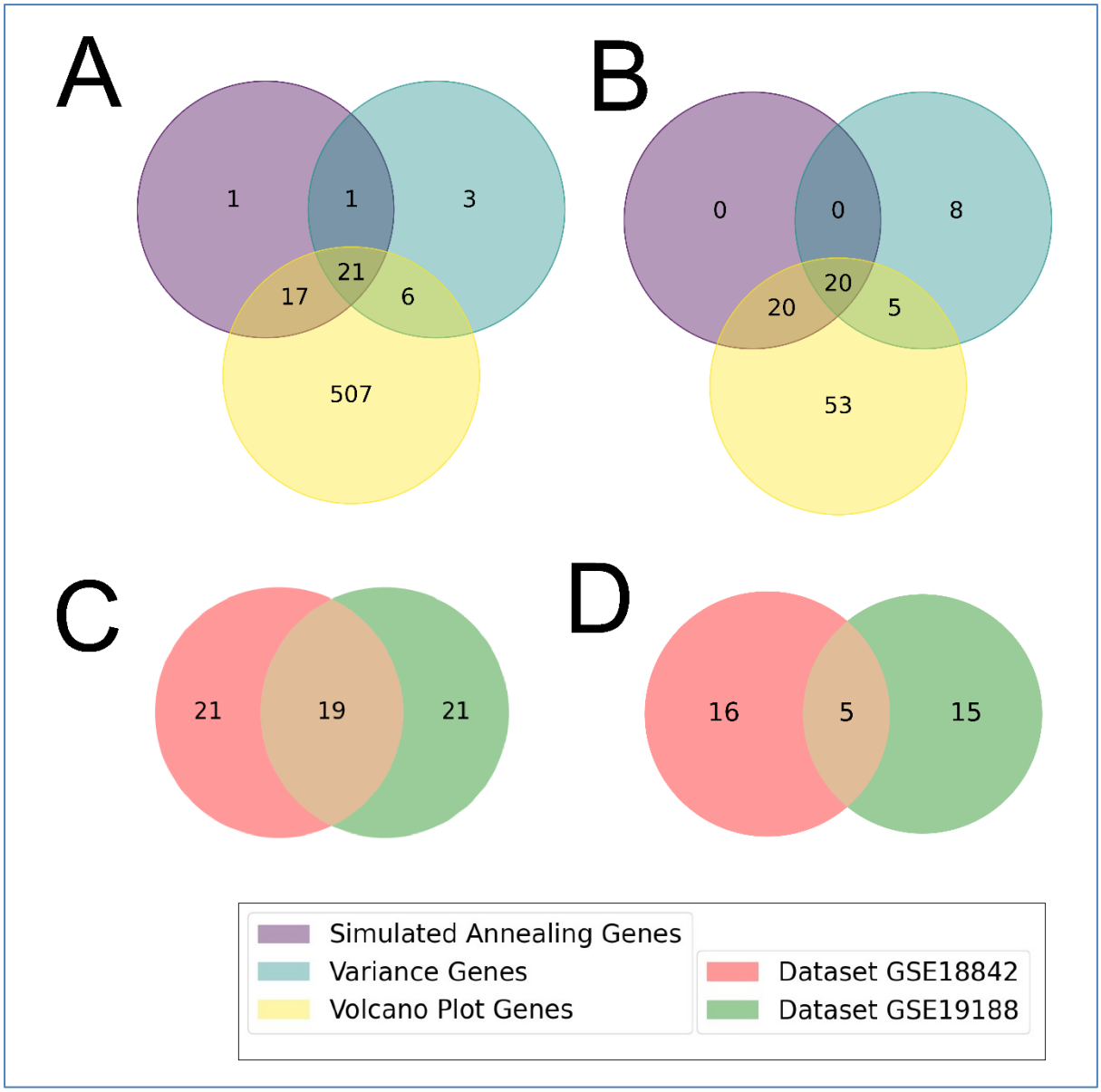

Figure 3: Overlap and Identification of Significant Genes from Two Datasets Using Multiple Gene Selection Methodologies. Gene Selection Analysis and Overlap Visualization. Panel A presents a Venn diagram illustrating the intersection of 31, 40, and 551 genes identified through variance-based screening, DGESA, and volcano plot analysis in the first dataset, revealing 21 common genes. Panel B shows a similar analysis for the second dataset, where 33, 40, and 98 genes were obtained, leading to the identification of 20 common genes. Panel C displays the overlap of genes obtained from both datasets, highlighting 19 shared genes from the 40 identified. Finally, Panel D depicts the overlap of common genes between Panels A and B, showcasing 5 genes in common.

Table 1. Common Genes Identified by Three Gene Selection Methods in the First Dataset. This table lists the genes (21) that were identified as common across all three gene selection methodologies—variance-based screening, the DGESA algorithm, and volcano plot analysis—applied to the first dataset.

SFTPC [28]	DSG3	KRT14
KRT6B	MMP12	MMP1
GJB6	KRT5	SPRR1B
SPRR3 [29]	S100A2	AKR1B10
TMEM100 [30]	MT1M	SPRR1A
KRT16	WIF1	CLDN18
GJB2	JUP /// KRT17	KRT6A

Table 2. Common Genes Identified by Three Gene Selection Methods in the Second Dataset. This table lists the genes (20) that were identified as common across all three gene selection methodologies—variance-based screening, the DGESA algorithm, and volcano plot analysis.

SCGB1A1	SFTA3	SFTPA1 /// SFTPA2	WIF1
TNNC1	MMP12	MCEMP1	SLC6A4
CLDN18	GKN2	CPB2	SFTPC
AGER	TMEM100	SCGB3A2	MAMDC2
FCN3	CYP4B1	SFTPD	FIGF /// PIR-FIGF

Table 3. Common Genes Identified by DGESA between the Two Datasets. This table lists the 19 genes that were identified as common between the two datasets using the DGESA algorithm. These genes represent key markers selected for further analysis in the study.

CLEC3B /// EXOSC7	TNNC1	SFTPC	FIGF /// PIR-FIGF
TMEM100	MMP12	SFTPA1 /// SFTPA2	SFTPD
CLDN18	MCEMP1	FCN3	WIF1
AGER	CYP4B1	FAM107A	CPB2
PEBP4	GKN2	LRRK2	

The clustered expression heat map provides a comprehensive visual representation of the selected genes from the first dataset, showcasing their expression levels in normal and diseased conditions (panel A and B of Figure 4 respectively). For clarity and better visual representation, all subsequent analyses focus on this first dataset, as the heat map enables a more detailed exploration of gene expression differences between normal and diseased states. Clustered heat maps like this are commonly used in gene expression studies to highlight how groups of genes behave across different conditions, facilitating insights into co-regulated genes or those with related functions. This type of heat map includes hierarchical clustering, where both the x-axis and the labels on the right side represent gene names, and a dendrogram on the left side visualizes the clustering of genes. Hierarchical clustering groups genes with similar expression profiles, making it easier to identify patterns and relationships among them. This allows for the identification of potential subgroups of genes that might share biological functions or contribute to disease progression. Clustered heat maps are commonly used in gene expression studies to highlight how groups of genes behave similarly across different conditions, facilitating insights into co-regulated genes or those with related functions. For clarity and better visual representation, all subsequent analyses will focus on this first dataset, as the heat map enables a more detailed exploration of gene expression differences between normal and diseased states (see panel A and B of Figure 4).

Table 4. Final 5 Genes Identified from the Overlap of Common Genes in Table 1 and Table 2. This table presents the 5 genes that were identified as common between the genes listed in Table 1 and Table 2, representing key markers for further analysis in the study.

WIF1	MMP12	TMEM100	CLDN18	SFTPC

The clustered expression heat map provides a comprehensive visual representation of the 21 identified genes from the first dataset, showcasing their expression levels in normal and lung tumour conditions. This type of heat map includes hierarchical clustering, where both the x-axis and the labels on the right side represent gene names, and a dendrogram on the left side visualizes the clustering of genes. Hierarchical clustering groups genes with similar expression profiles, making it easier to identify patterns and relationships among them. This approach allows for the identification of potential subgroups of genes that may share biological functions or contribute to disease progression. In the healthy state, the heat map reveals three to four distinct regions with high correlation (indicated by red clusters), showing groups of genes with similar expression patterns. However, in the lung tumour

condition, a significant shift occurs, with a single large highly correlated region forming from genes KRT14 to KRT6A. This region includes 13 genes belonging to the KRT, SPRR, and GJB families.

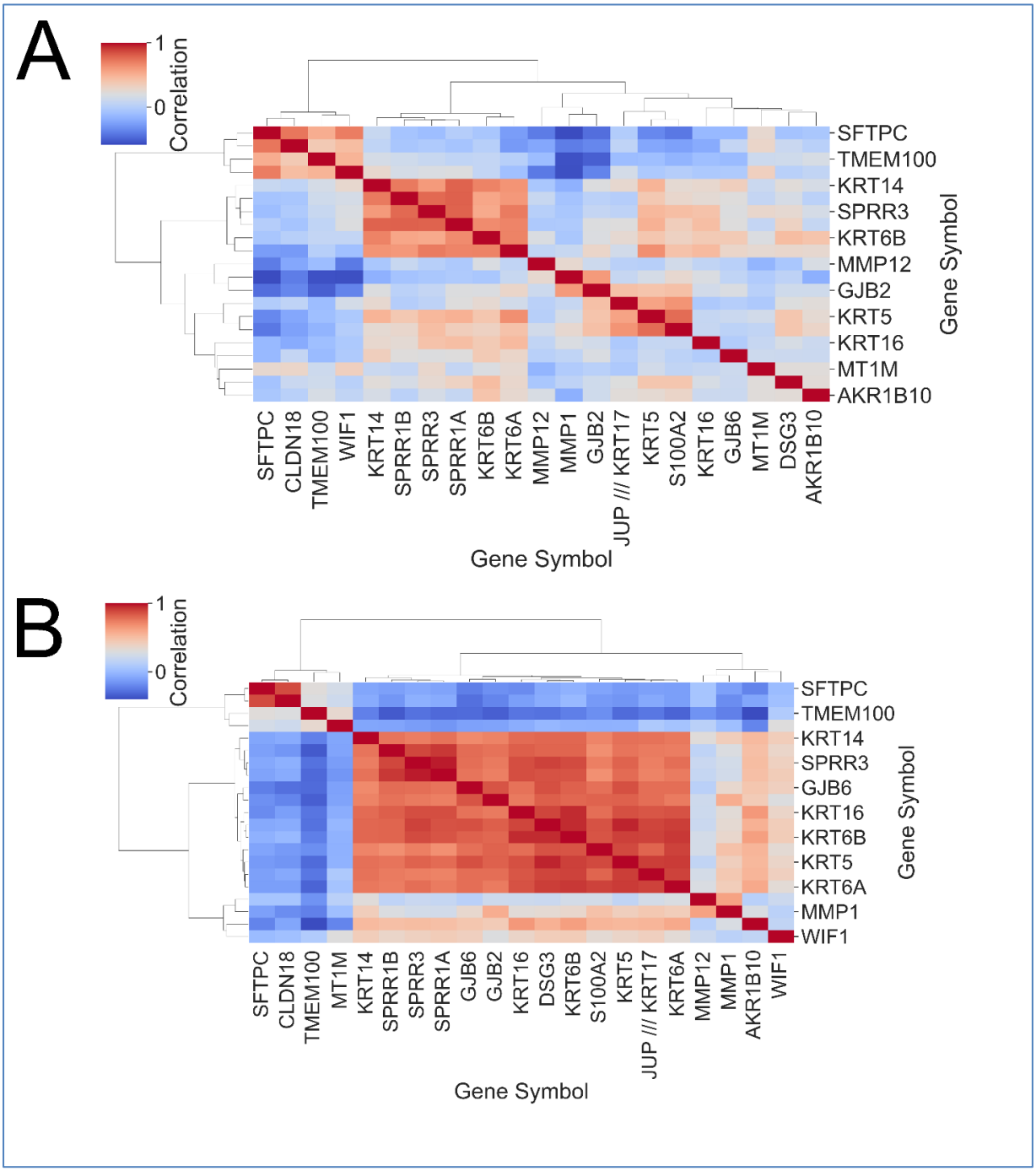

Figure 4: Clustered Expression Heat map of Selected Genes from the First Dataset. This heat map displays the expression profiles of significant genes identified through variance-based screening, DGESA, and volcano plot analysis in both normal and diseased conditions. Each row represents a selected gene, while each column corresponds to a sample, enabling the visualization of differential expression patterns between healthy (panel A) and lung cancer samples (panel B).

Gene Regulatory Networks (GRNs) were constructed for both healthy and lung tumour states using the 21 identified genes. The networks were designed based on co-expression relationships derived from correlation analysis and a predefined threshold of 0.3. Each pair of genes in the networks is connected by edges, where the edge properties (colour and thickness) represent the strength and direction of their correlation. For negative

correlations, edges are coloured blue. Strong negative correlations (values from -1 to -0.75) are represented by thick blue edges, while moderate negative correlations (values from -0.75 to -0.5) are denoted by thin blue edges. Weaker negative correlations (values from -0.5 to -0.3) are shown as light sea green with thin edges. Positive correlations are visualized distinctly for healthy and disease states. In the lung tumour network, weak positive correlations (values from 0.3 to 0.5) are represented by light coral with thin edges, while moderate positive correlations (values from 0.5 to 0.75) use red with thin edges. Strong positive correlations (values from 0.75 to 1) are shown as Indian red with thick edges. In the healthy network, weak positive correlations (values from 0.3 to 0.5) are shown in light green with thin edges, while moderate positive correlations (values from 0.5 to 0.75) appear in green with thin edges. Strong positive correlations (values from 0.75 to 1) are represented by thick green edges. This systematic construction allows for a clear comparison of co-expression patterns between the normal and lung tumour states, highlighting how gene interactions shift with disease progression.

A comparative analysis between the gene regulatory networks (GRNs) of healthy and lung tumour states reveals a marked difference in the strength of gene interactions. In the lung tumour GRN, genes exhibit stronger co-expression bonds, as evidenced by a greater number of thick edges, indicating higher correlation values between gene pairs. This suggests that the regulatory interactions among the genes become more tightly connected in the tumour condition. In contrast, the healthy GRN displays more dispersed and weaker interactions, with a higher prevalence of thin edges, indicating lower correlation strengths. This shift towards stronger co-expression in the lung tumour state may reflect a more synchronized or coordinated gene regulation pattern, potentially contributing to the altered cellular behaviour and disease progression associated with lung cancer.

For a systematic comparison of the gene regulatory networks (GRNs) between healthy and lung tumour states, key centrality measures were computed, including Degree Centrality, Closeness Centrality, and Betweenness Centrality, with the results detailed in Tables X and Y. These metrics provide insights into the importance and influence of individual genes within the networks. The analysis reveals a significant increase in centrality measures when transitioning from the healthy to the diseased state. Specifically, genes in the lung tumour GRN exhibit higher Degree Centrality, suggesting they have more connections with other genes, and increased Closeness Centrality, indicating a shorter path to other nodes in the network. The elevated Betweenness Centrality in the tumour state also highlights the enhanced role of certain genes as key intermediaries in gene regulation pathways. These changes suggest a shift in the network topology, where certain genes become more central in the regulatory dynamics associated with lung cancer.

Table 5. ***Centrality Measures of Identified Genes in the Healthy State****. This table presents the degree centrality, closeness centrality, and betweenness centrality of the identified genes in the healthy state. These measures reflect the network connectivity and the relative influence of each gene within the co-expression network, providing a basis for comparison with their roles in the lung tumour state.*

Gene	Degree Centrality	Closeness Centrality	Betweenness Centrality
SFTPC	0.40	0.57	0.24
DSG3	0.20	0.47	0.00
KRT14	0.35	0.50	0.03
KRT6B	0.55	0.56	0.19
MMP12	0.10	0.38	0.00
MMP1	0.25	0.42	0.00
GJB6	0.10	0.37	0.00
KRT5	0.55	0.67	0.30
SPRR1B	0.30	0.49	0.00
SPRR3	0.40	0.51	0.02
S100A2	0.45	0.63	0.19
AKR1B10	0.05	0.36	0.00
TMEM100	0.25	0.42	0.00
MT1M	0.05	0.31	0.00

SPRR1A	0.45	0.53	0.03
KRT16	0.20	0.38	0.00
WIF1	0.35	0.43	0.11
CLDN18	0.25	0.42	0.00
GJB2	0.40	0.57	0.19
JUP /// KRT17	0.15	0.48	0.00
KRT6A	0.40	0.51	0.02

Table 6. Centrality Measures of Identified Genes in the Diseased State. This table presents the degree centrality, closeness centrality, and betweenness centrality of the identified genes in the diseased state. These measures reflect the network connectivity and the relative influence of each gene within the co-expression network, providing a basis for comparison with their roles in the lung tumour state.

Gene	**Degree Centrality**	**Closeness Centrality**	**Betweenness Centrality**
SFTPC	0.10	0.41	0.10
DSG3	0.75	0.71	0.00
KRT14	0.80	0.74	0.07
KRT6B	0.75	0.71	0.00
MMP12	0.10	0.44	0.00
MMP1	0.55	0.61	0.03
GJB6	0.80	0.80	0.03
KRT5	0.80	0.80	0.03
SPRR1B	0.75	0.77	0.02
SPRR3	0.70	0.69	0.00
S100A2	0.75	0.71	0.00
AKR1B10	0.70	0.74	0.02
TMEM100	0.45	0.63	0.21
MT1M	0.10	0.49	0.00
SPRR1A	0.75	0.77	0.02
KRT16	0.75	0.71	0.00
WIF1	0.65	0.67	0.06
CLDN18	0.05	0.29	0.00
GJB2	0.75	0.77	0.03
JUP /// KRT17	0.75	0.71	0.00
KRT6A	0.80	0.80	0.03

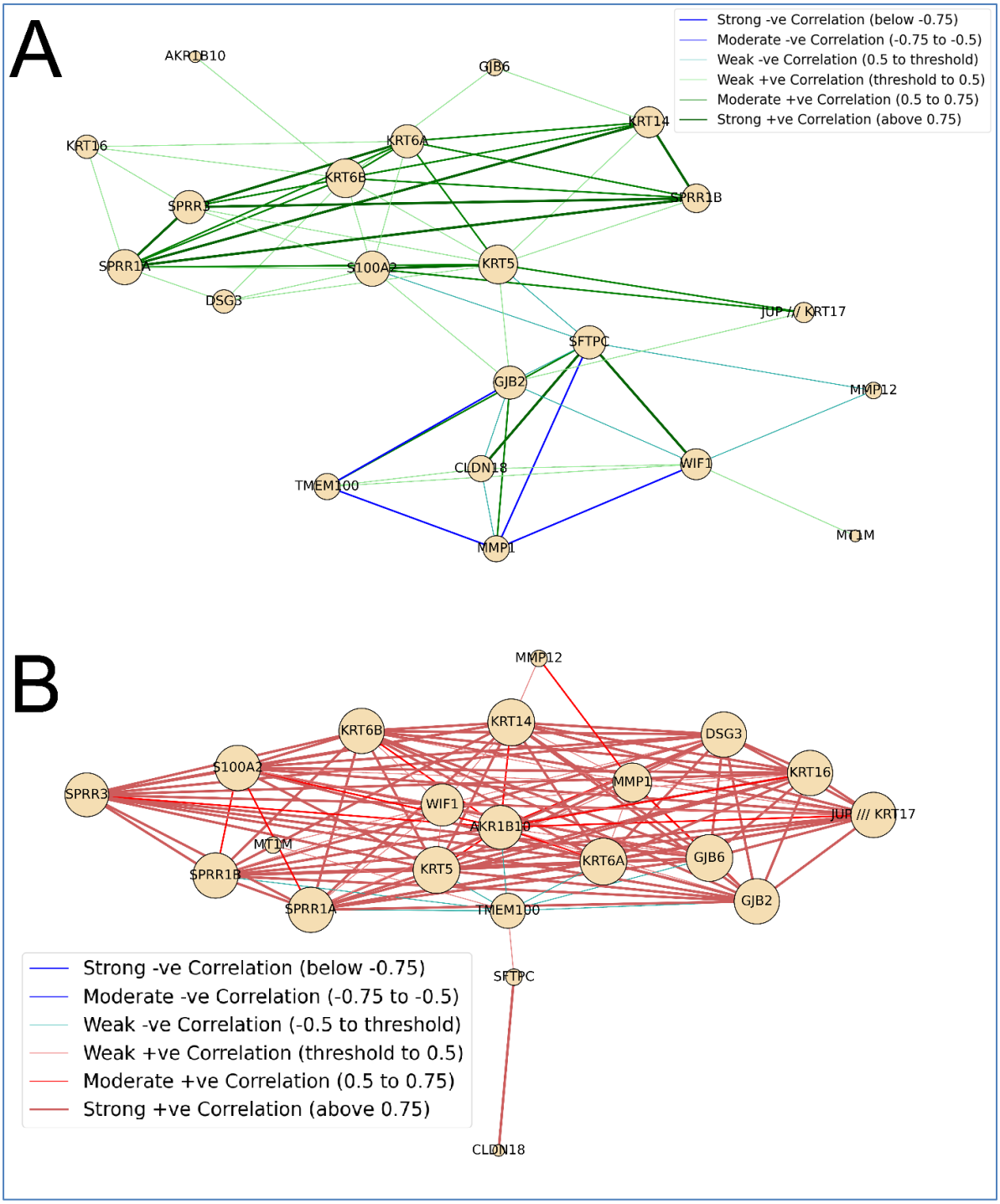

Figure 5. : Gene Regulatory Networks (GRNs) of Healthy and Lung Tumour States. This figure displays the co-expression-based GRNs constructed for both healthy and lung tumour conditions using the 21 identified genes. The edges between gene pairs indicate correlation strength and direction, with negative correlations represented in shades of blue and positive correlations distinguished by colour variations in each state. In the lung tumour network, weak to strong positive correlations are shown in light coral, red, and Indian red, while in the healthy network, they appear in light green and green shades. Stronger correlations are denoted with thicker edges, providing insights into the differences in gene interaction patterns between the two states.

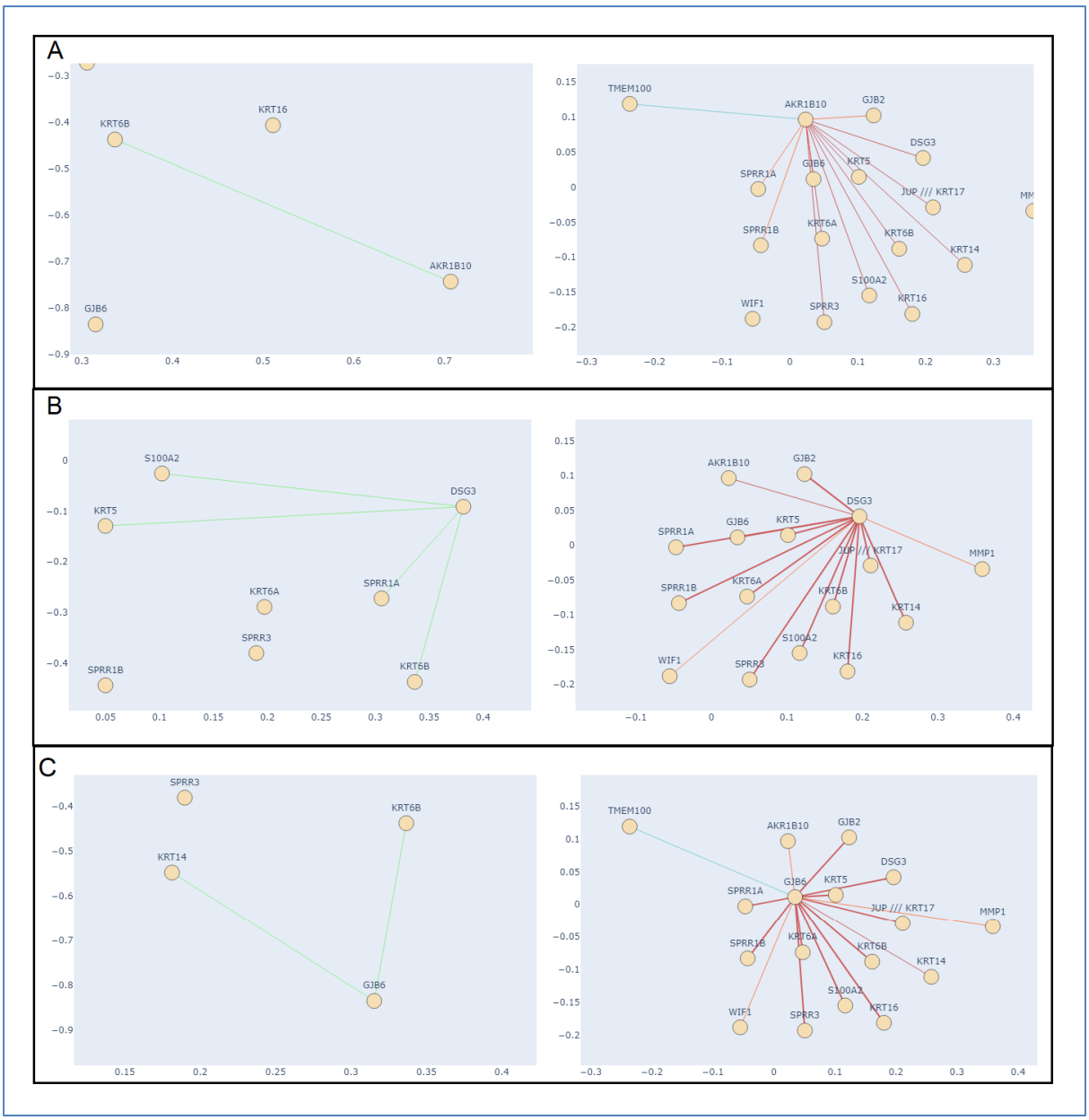

Figure 6. Increased Node Degrees of AKR1B10, DSG3, and GJB6 in Healthy vs. Lung Tumour States. This figure illustrates the differences in connectivity for three genes—AKR1B10, DSG3, and GJB6—in the gene regulatory networks (GRNs) of healthy and lung tumor conditions. The node sizes represent the degree of each gene, with larger nodes indicating a higher degree of connectivity. The selected genes show a significant increase in their node degrees in the lung tumour network, indicating a more central role in the altered gene regulatory interactions in the diseased state.

To highlight the differences in gene interactions between healthy and lung tumour conditions, we focused on three genes—AKR1B10, DSG3, and GJB6—that showed a substantial increase in their node degrees when transitioning from healthy to diseased states (see Figure 6). Node degree reflects the number of direct interactions a gene has with others in the network, indicating its regulatory influence. In the lung tumour state, these genes displayed a higher degree of connectivity, suggesting an enhanced role in maintaining the altered network structure associated with cancer progression. All other genes showing similar property are placed in the supplements.

In contrast to the genes that gained connectivity in the transition from healthy to lung tumour states, some genes exhibited a reduction in their node degrees, suggesting a diminished role in the diseased condition. Notably, CLDN18 and SFTPC showed a decrease in their node degrees when comparing the gene regulatory networks

(GRNs) of healthy and lung cancer states. This reduction indicates a loss of their direct interactions with other genes, potentially reflecting a reduced regulatory influence in the altered biological environment of lung cancer.

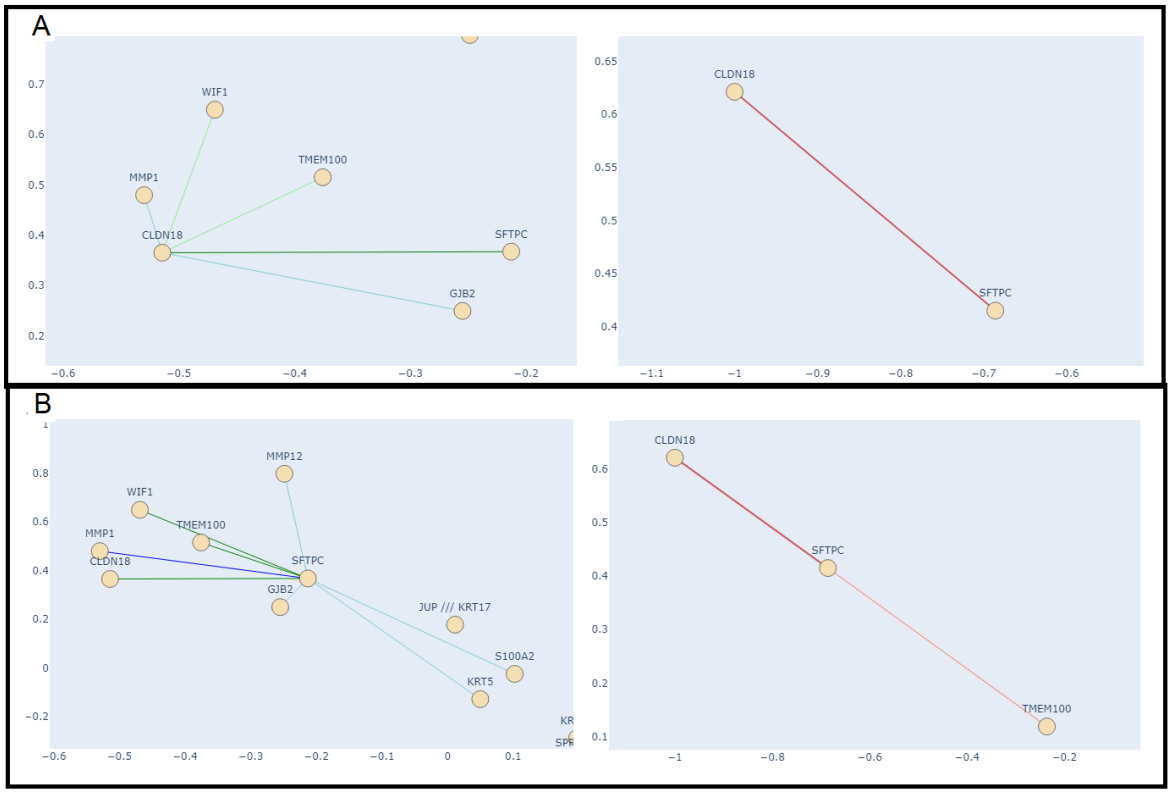

Figure 7. Decreased Node Degrees of CLDN18 and SFTPC in Healthy vs. Lung Tumour States. This figure displays the differences in connectivity for CLDN18 and SFTPC within the gene regulatory networks (GRNs) of healthy and lung tumour conditions. Node sizes correspond to the degree of each gene, with smaller nodes indicating a reduced degree of connectivity. Both CLDN18 and SFTPC show a notable decrease in their node degrees in the lung tumour network, suggesting a diminished regulatory role in the network interactions associated with the diseased state.

4. **Discussion**:

This study presents a comprehensive analysis of differential gene selection for lung cancer by employing three distinct methodologies across two datasets. By implementing variance-based screening, a simulated annealing-based algorithm (DGESA), and volcano plot analysis, we successfully identified common genes that exhibit significant differential expression patterns. Through overlapping these gene sets, we further examined gene expression patterns using clustered heat maps, which provided valuable insights into the co-expression networks of the overlapped genes in both healthy and diseased states. The findings reveal interesting correlations and interactions among the selected genes, highlighting potential biomarkers for lung cancer progression.

One of the notable aspects of this research is the identification of a core set of common genes across the methodologies, emphasizing the reliability and robustness of the selected gene sets. The application of clustered heat maps allowed for the visualization of distinct expression patterns between healthy and lung tumor conditions, revealing the presence of multiple highly correlated gene clusters in normal samples. In contrast, a single large correlated region comprising 13 genes from the KRT, SPRR, and GJB families emerged in the lung tumor condition. This indicates a shift in regulatory relationships and suggests that these gene families may play a significant role in lung cancer pathology.

The co-expression networks constructed for both healthy and diseased states further illustrate the differential relationships among the identified genes. Notably, genes in the lung tumor GRNs exhibited stronger

connections, suggesting enhanced regulatory interactions that may contribute to malignant transformation. Centrality measures computed for these networks provided additional insights into the importance of specific genes within the networks. Increased centrality scores in the tumor state for genes such as AKR1B10, DSG3, and GJB6 indicate their pivotal roles in lung cancer progression, while decreased centrality for genes like CLDN18 and SFTPC may suggest a loss of regulatory influence in the malignant state.

Among the 21 common genes identified in this study, 17 have been previously confirmed in the literature as having significant associations with lung cancer. This correlation underscores the relevance and reliability of the gene selection methodologies employed. Genes such as KRT5, KRT16, and AKR1B10 have been documented in numerous studies for their roles in tumorigenesis and progression in lung cancer. The existing literature highlights these genes' involvement in crucial biological processes, including cell proliferation, differentiation, and apoptosis, all of which are vital in the context of malignancy. This strong alignment with established findings not only reinforces the validity of our results but also suggests that the remaining four genes warrant further investigation to elucidate their potential roles in lung cancer.

Despite these strengths, this research is not without limitations. The study relies on publicly available datasets, which may introduce variability due to differences in sample collection, processing, and data quality. Furthermore, while the analysis successfully identified significant gene alterations, the biological relevance of these genes needs to be validated through functional studies and clinical trials. Additionally, the generalizability of the findings may be constrained by the specific populations represented in the datasets. Addressing these limitations in future work could enhance the understanding of lung cancer biology and the clinical utility of the identified biomarkers.

In conclusion, this study highlights the potential of gene regulatory networks and differential gene expression analysis as powerful tools for identifying biomarkers of lung cancer. The comprehensive approach taken here has provided valuable insights into the transition from normalcy to malignancy, revealing crucial gene interactions and regulatory mechanisms. These findings lay the groundwork for further investigation into the functional roles of these genes in lung cancer and their potential as therapeutic targets or diagnostic markers.

5. **Conclusion**:

This research presents a significant advancement in lung cancer diagnostics by developing gene regulatory network (GRN)-based biomarkers that elucidate the transformation of network biomarkers from normal to malignant conditions. Utilizing three distinct methodologies—variance-based screening, a novel simulated annealing-based gene selection algorithm, and volcano plot analysis—this study identified 21 common significant genes associated with lung cancer across two independent datasets. Notable markers, including KRT5, SFTPC, KRT16, and AKR1B10, were recognized for their established relevance in lung cancer pathology. The construction of co-expression-modelled GRNs for healthy and diseased states revealed critical differences in gene interaction patterns, emphasizing the dynamic nature of gene regulation during tumour progression. The analysis of centrality measures illuminated how specific genes like KRT5 and KRT16 exhibit significantly higher centrality in the lung cancer network, indicating their pivotal roles in disease progression. In contrast, genes such as MT1M and CLDN18 showed consistently low centrality scores, suggesting limited involvement in lung cancer. Interestingly, the significant shifts observed in genes like GJB6, previously not strongly associated with lung cancer, point toward their potential involvement in tumorigenesis. This comprehensive approach, integrating rigorous statistical and machine learning techniques, provides a robust framework for understanding molecular mechanisms underlying lung cancer. Furthermore, the identification of 17 out of the 21 common genes confirmed in existing literature reinforces their clinical relevance, suggesting potential utility in patient stratification and personalized treatment. Overall, this research not only sheds light on the intricate relationships between gene expression and lung cancer but also lays a solid foundation for future studies. Continued exploration into the molecular intricacies of lung cancer will pave the way for advancements in diagnostics, therapeutic strategies, and ultimately improved patient outcomes.

Functional validation of identified genes is crucial for establishing their roles in lung cancer progression and exploring their therapeutic potential. Longitudinal studies could provide insights into how these biomarkers' expression changes over time, aiding personalized treatment strategies. Additionally, investigating interactions within broader regulatory networks may uncover novel pathways involved in lung cancer, ultimately contributing to improved diagnostic and prognostic tools in clinical settings.

Reference

1. Minna, J. D., Roth, J. A., & Gazdar, A. F. (2002). Focus on lung cancer. Cancer cell, 1(1), 49-52.
2. Tao, M. H. (2019). Epidemiology of lung cancer. Lung Cancer and Imaging, 4-1.
3. Rogado, J., Pangua, C., Serrano-Montero, G., Obispo, B., Marino, A. M., Pérez-Pérez, M., ... & Lara, M. Á. (2020). Covid-19 and lung cancer: A greater fatality rate? Lung cancer, 146, 19-22.
4. Mohamed, S., Casiraghi, M., Spaggiari, L., & Bertolaccini, L. (2024). Lung cancer in the era of COVID-19. Management, Body Systems, and Case Studies in COVID-19, 211-219.
5. Granville, C. A., & Dennis, P. A. (2005). An overview of lung cancer genomics and proteomics. American journal of respiratory cell and molecular biology, 32(3), 169-176.
6. Borczuk, A. C., Toonkel, R. L., & Powell, C. A. (2009). Genomics of lung cancer. Proceedings of the American Thoracic Society, 6(2), 152-158.
7. Buettner, R., Wolf, J., & Thomas, R. K. (2013). Lessons learned from lung cancer genomics: the emerging concept of individualized diagnostics and treatment. Journal of Clinical Oncology, 31(15), 1858-1865.
8. Parikh, A. R. (2019). Lung Cancer Genomics. Acta Medica Academica, 48(1).
9. Martínez-Ruiz, C., Black, J. R., Puttick, C., Hill, M. S., Demeulemeester, J., Larose Cadieux, E., ... & McGranahan, N. (2023). Genomic–transcriptomic evolution in lung cancer and metastasis. Nature, 616(7957), 543-552.
10. Xiong, J., & Zhou, T. (2012). Gene regulatory network inference from multifactorial perturbation data using both regression and correlation analyses.
11. Zheng, G., & Huang, T. (2018). The reconstruction and analysis of gene regulatory networks. Computational Systems Biology: Methods and Protocols, 137-154.
12. Singh, A. J., Ramsey, S. A., Filtz, T. M., & Kioussi, C. (2018). Differential gene regulatory networks in development and disease. Cellular and Molecular Life Sciences, 75, 1013-1025.
13. Pratapa, A., Jalihal, A. P., Law, J. N., Bharadwaj, A., & Murali, T. M. (2020). Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. Nature methods, 17(2), 147-154.
14. Yin, W., Mendoza, L., Monzon-Sandoval, J., Urrutia, A. O., & Gutierrez, H. (2021). Emergence of co-expression in gene regulatory networks. PloS one, 16(4), e0247671.
15. Escorcia-Rodríguez, J. M., Gaytan-Nuñez, E., Hernandez-Benitez, E. M., Zorro-Aranda, A., Tello-Palencia, M. A., & Freyre-González, J. A. (2023). Improving gene regulatory network inference and assessment: The importance of using network structure. Frontiers in Genetics, 14, 1143382.
16. Wang, J., Chen, Y., & Zou, Q. (2023). Inferring gene regulatory network from single-cell transcriptomes with graph autoencoder model. PLoS Genetics, 19(9), e1010942.
17. Huang, J., Sun, Z., Yan, W., Zhu, Y., Lin, Y., Chen, J., ... & Wang, J. (2014). Identification of microRNA as sepsis biomarker based on miRNAs regulatory network analysis. BioMed research international, 2014(1), 594350.
18. Chudasama, D., Bo, V., Hall, M., Anikin, V., Jeyaneethi, J., Gregory, J., ... & Karteris, E. (2018). Identification of cancer biomarkers of prognostic value using specific gene regulatory networks (GRN): a novel role of RAD51AP1 for ovarian and lung cancers. Carcinogenesis, 39(3), 407-417.
19. Jiang, J., Bi, Y., Liu, X. P., Yu, D., Yan, X., Yao, J., ... & Li, S. (2020). To construct a ceRNA regulatory network as prognostic biomarkers for bladder cancer. Journal of cellular and molecular medicine, 24(9), 5375-5386.
20. Zhang, L., Fan, S., Vera, J., & Lai, X. (2023). A network medicine approach for identifying diagnostic and prognostic biomarkers and exploring drug repurposing in human cancer. Computational and structural biotechnology journal, 21, 34-45.

21. Wang, Y. C., & Chen, B. S. (2011). A network-based biomarker approach for molecular investigation and diagnosis of lung cancer. BMC medical genomics, 4, 1-15.
22. Doungpan, N., Engchuan, W., Chan, J. H., & Meechai, A. (2016). GSNFS: Gene subnetwork biomarker identification of lung cancer expression data. BMC medical genomics, 9, 241-250.
23. Liao, Y., Wang, Y., Cheng, M., Huang, C., & Fan, X. (2020). Weighted gene coexpression network analysis of features that control cancer stem cells reveals prognostic biomarkers in lung adenocarcinoma. Frontiers in Genetics, 11, 311.
24. Otálora-Otálora, B. A., López-Kleine, L., & Rojas, A. (2023). Lung cancer gene regulatory network of transcription factors related to the hallmarks of cancer. Current Issues in Molecular Biology, 45(1), 434-464.
25. Sinha, K., Chakraborty, S., Bardhan, A., Saha, R., Chakraborty, S., & Biswas, S. (2024). A New Differential Gene Expression Based Simulated Annealing for Solving Gene Selection Problem: A Case Study on Eosinophilic Esophagitis and Few Other Gastro-Intestinal Diseases. medRxiv, 2024-05.
26. https://www.ncbi.nlm.nih.gov/geo/
27. Barrett, T., Suzek, T. O., Troup, D. B., Wilhite, S. E., Ngau, W. C., Ledoux, P., ... & Edgar, R. (2005). NCBI GEO: mining millions of expression profiles—database and tools. Nucleic acids research, 33(suppl_1), D562-D566.
28. Li, Q., Wang, Y., Hu, R., & Yang, G. (2020). Dysregulation of SPRR3/miR-876-3p axis contributes to tumorigenesis in non-small-cell lung cancer. OncoTargets and therapy, 2411-2419.
29. He, Q., Dong, Y., Zhu, Y., Ding, Z., Zhang, X., Wang, Z., ... & He, Y. (2021). TMEM100 induces cell death in non-small cell lung cancer via the activation of autophagy and apoptosis. Oncology Reports, 45(5), 1-11.
30. Yuanhua, L., Pudong, Q., Wei, Z., Yuan, W., Delin, L., Yan, Z., ... & Bo, S. (2019). TFAP2A induced KRT16 as an oncogene in lung adenocarcinoma via EMT. International journal of biological sciences, 15(7), 1419.