biology

MDPI

Article

A Systems Biology and LASSO-Based Approach to Decipher the Transcriptome–Interactome Signature for Predicting Non-Small Cell Lung Cancer

Firoz Ahmed [1,*], Abdul Arif Khan [2], Hifzur Rahman Ansari [3] and Absarul Haque [4,5]

[1] Department of Biochemistry, College of Science, University of Jeddah, P.O. Box 80327, Jeddah 21589, Saudi Arabia
[2] Department of Pharmaceutics, College of Pharmacy, King Saud University, P.O. Box 2457, Riyadh 11451, Saudi Arabia
[3] King Abdullah International Medical Research Center (KAIMRC), King Saud Bin Abdulaziz University for Health Sciences, King Abdulaziz Medical City, Ministry of National Guard Health Affairs, P.O. Box 9515, Jeddah 21423, Saudi Arabia
[4] King Fahd Medical Research Center, King Abdulaziz University, P.O. Box 80216, Jeddah 21589, Saudi Arabia
[5] Department of Medical Laboratory Sciences, Faculty of Applied Medical Sciences, King Abdulaziz University, P.O. Box 80216, Jeddah 21589, Saudi Arabia
* Correspondence: fahmed1@uj.edu.sa

Simple Summary: Non-small cell lung cancer (NSCLC) is a serious public health issue due to its high mortality rate. To improve the survival rate of NSCLC with better treatment, it is imperative to develop a biomarker-based prediction tool that can accurately identify NSCLC at a very early stage. Cancer development initiates due to aberrations in gene expression and the regulatory networks; therefore, these features hold a great potential to diagnose cancer at an early stage compared with the visible morphological and pathological changes. In this study, we integrated gene expression and interactome data to identify candidate genes altered in NSCLC compared with normal samples. We then used a machine learning technique to identify a signature of 17 genes and developed a model for predicting NSCLC. Interestingly, our model predicted NSCLC across different independent test datasets with high accuracy. Finally, the model was implemented to create a user-friendly web tool, *NSCLCpred*, to predict NSCLC using the expression profile of 17 genes. We expect that our findings will guide the identification of NSCLC patients and provide more insight into the understanding of disease development.

Abstract: The lack of precise molecular signatures limits the early diagnosis of non-small cell lung cancer (NSCLC). The present study used gene expression data and interaction networks to develop a highly accurate model with the least absolute shrinkage and selection operator (LASSO) for predicting NSCLC. The differentially expressed genes (DEGs) were identified in NSCLC compared with normal tissues using TCGA and GTEx data. A biological network was constructed using DEGs, and the top 20 upregulated and 20 downregulated hub genes were identified. These hub genes were used to identify signature genes with penalized logistic regression using the LASSO to predict NSCLC. Our model's development involved the following steps: (i) the dataset was divided into 80% for training (TR) and 20% for testing (TD1); (ii) a LASSO logistic regression analysis was performed on the TR with 10-fold cross-validation and identified a combination of 17 genes as NSCLC predictors, which were used further for development of the LASSO model. The model's performance was assessed on the TD1 dataset and achieved an accuracy and an area under the curve of the receiver operating characteristics (AUC-ROC) of 0.986 and 0.998, respectively. Furthermore, the performance of the LASSO model was evaluated using three independent NSCLC test datasets (GSE18842, GSE27262, GSE19804) and achieved high accuracy, with an AUC-ROC of >0.99, >0.99, and 0.95, respectively. Based on this study, a web application called *NSCLCpred* was developed to predict NSCLC.

check for updates

Citation: Ahmed, F.; Khan, A.A.; Ansari, H.R.; Haque, A. A Systems Biology and LASSO-Based Approach to Decipher the Transcriptome–Interactome Signature for Predicting Non-Small Cell Lung Cancer. *Biology* **2022**, *11*, 1752. https://doi.org/10.3390/biology11121752

Academic Editors: Chung-Der Hsiao and Tzong-Rong Ger

Received: 31 October 2022
Accepted: 28 November 2022
Published: 30 November 2022

Keywords: non-small cell lung cancer; LASSO model; artificial intelligence; gene expression; biological networks; hub genes

1. Introduction

Despite all the rapid advancements in the development of anticancer therapy, lung cancer is still a significant contributor to cancer-associated deaths, with almost 1.8 million worldwide mortalities recorded in the year 2020 [1]. Non-small cell lung cancer (NSCLC) is a major contributor to lung cancer cases, making up almost 85% of cases of primary lung cancer [2]. The high mortality rate associated with NSCLC is because the disease is often diagnosed late in most patients, resulting in a poor prognosis, even with the availability of advanced treatment modalities [3]. Furthermore, nearly half of the initially diagnosed early-stage tumors (Stage I or II) in patients eventually proceed to the late stage, resulting in metastatic NSCLC. Advanced NSCLC is generally categorized into Stage IIIB or IV tumors, and its current treatments options include immunotherapy, systemic chemotherapy, and targeted drug therapy, which often lead to imposing remarkably great adverse impacts not only on the lives of patients but also at the socio-economic level [4–6], especially the family and friends providing informal care to the NSCLC patient. The continuous increase in the economic burden has posed a great financial challenge to society, as the number of advanced NSCLC cases is increasing. Thus, it appears that the patients and their caregivers seem to be affected by the due course of stage progression. As a result, it directly influences the increase in direct and indirect costs as the stages of the disease advance. It has been shown that the overall economic burden of the management of lung cancer in Europe is considerably surprising because the direct costs of caring for such NSCLC patients amount to more than EUR 3 billion annually [7]. As is quite evident from several studies, the clinical outcome for NSCLC patients is directly dependent on the stage of the tumor when it is diagnosed [8,9]. So far, screening for NSCLC patients relies on using chest radiographs or sputum cytologic profile analyses, which remain adequate and have failed to provide a mortality benefit in many studies of clinical trials [8,9]. Therefore, there is a need to focus on discovering and validating a set of biomarkers with high sensitivity and specific discriminatory power that might be utilized in early screening programs along with having diagnostic and prognostic significance that would allow the accurate detection of such patients in the early stages of the disease, consequently enabling the clinician to reduce the mortality rate of NSCLC patients.

However, imaging techniques, including X-ray, CT, MRI, PET scans, and tissue biopsy, are routine practices in diagnosing lung cancer and are generally used when the cancer is at an advanced stage [10]. The powerful imaging techniques can detect a tumor only when its size is at least 7 mm with billions of cells [11]. Unfortunately, only 16% of cases are detected before the spread of lung cancer to other organs [12]. However, the recent advancement of high-throughput sequencing technology has improved the understanding of the underlying pathological changes and identified the genomic and environmental factors involved in lung cancer [13,14]. The accumulated knowledge is being utilized for advancing diagnostic accuracy, and a significant improvement has been gained in the treatment outcomes in several cases. In addition to these technological improvements, evidence has indicated that utilizing a growing number of molecular-level approaches could be beneficial for improving the early diagnosis and treatment outcomes of lung cancer. Hence, there is a great need to adopt advanced molecular techniques for early diagnosis, better management, and treatment of cancer patients in super-specialized hospitals [15]. In the past few decades, studies have used different approaches to identify the genes and mutation signatures underpinning lung cancer, leading to new therapeutic targets for better treatment. A previous study used an integrative systems biology approach and revealed a driver network that promotes cell proliferation in NSCLC, which could be a promising therapeutic target [16]. Interestingly, the study found that the driver

network consisted of 26 upregulated genes associated with spindles, kinetochores, nuclear division, chromosome segregation, and the cell cycle G2/M transition and their upstream regulators, FOXM1 and MYBL2 [16]. A recent study used gene expression data from the TGF-β-induced epithelial–mesenchymal transition in NSCLC cells and identified a cluster of differentially expressed genes associated with specific metabolic processes such as glycolysis, pyruvate metabolism, and the tricarboxylic acid cycle [17]. Interestingly, the same study elucidated the potential links in the regulation of NSCLC's progression and found 10 genes as prognostic biomarkers associated with a decrease in the overall survival of NSCLC patients. The prognostic markers might be helpful for evaluating treatment outcomes and monitoring and selecting suitable therapeutic strategies in NSCLC [17]. Integrated bioinformatic analysis of differentially expressed genes was successfully used to identify the potential prognostic gene signatures in other cancers, including esophageal squamous cell carcinoma and cervical cancer [18,19].

Furthermore, several models have been developed for predicting the risk score for lung cancer [20–23]. However, these studies mainly focused on utilizing epidemiological factors, symptoms, and clinical assessments as features, not gene expression features, for development of the model; therefore, these models have been suggested to have certain limitations due to their low accuracy. Another study developed a prognostic model for NSCLC patients using immune-specific transcriptomic and clinicopathological data, and achieved an area under the curve (AUC) of 0.673 [24]. Failure to diagnose lung cancer at an early stage, resulting in metastasis to other organs, has posed a significant challenge in treating cancer thoroughly. Therefore, an early-stage detection method with the highest possible accuracy is essential for better treatment and prognosis of NSCLC.

The recent advancement and breakthroughs in high-throughput sequencing technology have enabled the rapid growth of transcriptomic and other omics data from cancer samples, thus providing an excellent opportunity to improve deep insights and early diagnosis [25,26]. However, identifying the signature genes for early cancer diagnoses and the interpretation of the underlying mechanisms in high-dimensional and complex data remains a great challenge [27,28]. Artificial intelligence (AI) and machine learning techniques (MLT) have successfully been used in biomedicine and crop improvement [29–33]. The application of AI and MLT has also shown promising results in cancer diagnosis and drug discovery, where a predictive model can be built by learning and generalizing from the training data. The model is applied to new data to make predictions [34–36].

This work combined transcriptome–interactome signatures to develop an efficient model for predicting NSCLC. Briefly, we used the gene expression profile to identify the differentially expressed genes (DEGs), followed by finding the hub genes in biological networks. These hub genes were used to select the signature genes that best discriminated NSCLC from normal samples by the least absolute shrinkage and selection operator (LASSO). The highly accurate predictive LASSO model was developed by using selected features, and then the results were interpreted for a mechanical understanding. In feature selection, LASSO penalizes the regression variable coefficient and shrinks them to zero. After that, it selects the variables with non-zero coefficients for constructing the model. The larger the parameter λ, the greater the number of coefficients that shrink to zero. Therefore, we have to tune and select the minimum value of the parameter λ to obtain a sufficient number of coefficients. The present work demonstrated the potential use of this approach for developing predictive models for the early diagnosis of other cancers.

2. Materials and Methods

The experimental workflow of our study is given in Figure 1A and consisted of two parts. The first part of the work identified the biologically important genes associated with NSCLC using DEGs and their interaction network. The next part used the relevant genes, identified in the first part of the work as input for feature selection and model development with the LASSO. Finally, the model's performance was evaluated on independent test datasets and the 17-gene signature was validated in lung cancer data and literature.

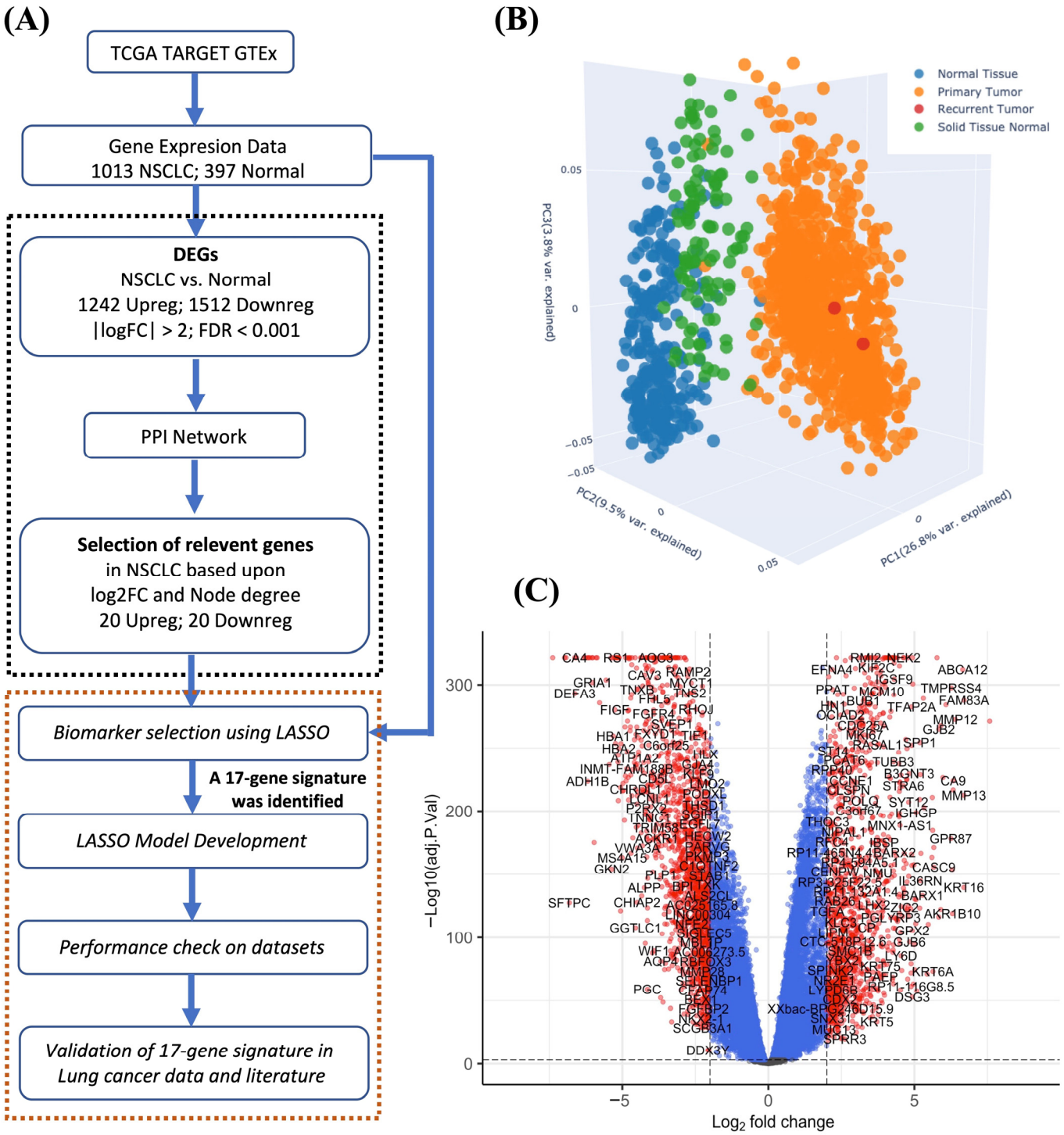

Figure 1. (**A**) Workflow of our study. (**B**) The PCA plot for samples using 2500 genes with the most significant variance. Each point represents the gene expression of a sample. Samples with similar gene expression profiles are closer in the three-dimensional space. (**C**) Volcano plot of DEGs in lung cancer compared with normal samples. The DEGs with |log2FC| > 2.0 and adj.p.val < 0.001 are shown in red.

2.1. Identification of DEGs

The gene expression data of the NSCLC and normal samples were obtained from TCGA TARGET GTEx using the UCSC Xena application (https://xena.ucsc.edu/, accessed on 7 August 2021). The gene expression dataset was in "RSEM norm_count"

format. The NSCLC data consisted of 1013 samples of lung adenocarcinoma and lung squamous cell carcinoma taken from The Cancer Genome Atlas (TCGA) [37]. The normal sample included 397 samples of "lung solid tissue—normal" collected from TCGA and "normal lung tissue" taken from the Genotype-Tissue Expression (GTEx) database [38]. TCGA provides gene expression and other omics data as well as clinical data from primary cancer and matched normal samples across 33 cancer types. GTEx provides tissue-specific gene expression and regulation data from nearly 1000 non-diseased individuals. The DEGs in lung cancer compared with normal samples were analyzed using the Xena application adapted from the Appyter RNA-seq analysis pipeline from Ma'ayan lab (https://github.com/MaayanLab/appyter-catalog, accessed on 7 August 2021). The RNA-seq data underwent quantile normalization, and the DEGs were identified. A gene was considered to be upregulated when log2FC > 2 and adj.p.value < 0.001, but deemed to be downregulated when log2FC < −2 and adj.p.value < 0.001. The volcano plot was made using the *EnhancedVolcano* tool in R version 4.1.2 [39].

2.2. Construction of the Interaction Network

The biological interactions data of humans were screened from BioGRID version 4.4.205 (last modified 29 December 2021) [40]. The interaction data were filtered to screen the interactions of the identified DEGs in lung cancer. The DEGs data were added to the interaction network, and the sub-network was prepared on the basis of the log2FC value. Cytoscape version 3.8 was used to visualize the interaction network of biologically important genes [41].

2.3. Identification of Biologically Important Nodes in the Network

The network topology was calculated using Cytoscape's built-in *NetworkAnalyzer* tool. The nodes' sizes were arranged according to the value of their degree in the original BioGRID human interaction network. In contrast, the color of the nodes was set as per their log2FC value in cancer and normal samples in the lung. The interaction network was further filtered based on the log2FC value in lung samples.

2.4. Training and Testing Dataset

The final dataset consisted of the expression values in the RSEM of 40 genes identified by analyzing the DEGs and interactomes from 1013 samples of lung cancer and 397 samples as the control. This dataset was divided into two parts: an 80% training dataset (TR, with 1128 samples) and a 20% test dataset 1 (TD1, with 282 samples). The model was developed using 10-fold cross-validation (cv) on the TR dataset, and the performance was checked on TD1 dataset (Table 1). To further validate the accuracy and robustness of the LASSO model, we used additional validation test dataset 2 (TD2), which contained the gene expression data of lung cancer obtained from three microarrays: GSE18842 [42], GSE27262 [43], and GSE19804 [44]. These raw microarray data were normalized using the RMA of the *Oligo package* in R. The number of samples of lung cancer and the adjacent non-tumor tissues in TD2 is provided in Table S1.

Table 1. The number of TCGA and GTEx lung samples used for DEGs analysis and development of the LASSO logistic regression model.

Sample Type	Disease Class	Training Dataset (TR)	Test Dataset 1 (TD1)
Primary tumor	Lung cancer (positive = 1)	802	209
Recurrent tumor		2	0
Normal solid tissue	Normal lung (negative = 0)	91	18
Normal tissue (GTEx)		233	55
Total		1128	282

2.5. Construction of the LASSO Model

We used the R package *glmnet* version 4.1-3 to develop a penalized logistic regression LASSO model on the TR dataset with 10-fold cv. We divided the data randomly into 10 sets; of these, we used 9 sets for training and the remaining set for testing. First, the penalty regularization parameter lambda was determined by 10-fold cv with the *cv.glmnet* module. Next, the final model was developed by *glmnet* with a lambda value which maximized the value of AUC (lambda.min). The expression data of genes with non-zero coefficients were used to create the final LASSO model.

2.6. Performance of the Models

The performance of the LASSO models was evaluated with the following parameters.

(1) *Sensitivity*, also called the recall or true positive rate, which indicates the percentage of correctly predicted cancer samples.

$$Sensitivity = \frac{TP}{TP + FN}$$

(2) *Specificity*, which indicates the percentage of correctly predicted normal samples.

$$Specificity = \frac{TN}{TN + FP}$$

(3) *Accuracy* is the percentage of correct predictions overall.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

(4) *Positive predictive value (PPV)*, also called the precision.

$$PPV = \frac{TP}{TP + FP}$$

(5) *Negative predictive value (NPV)*

$$NPV = \frac{TN}{TN + FN}$$

where *TP* stands for true positive, *TN* stands for true negative, *FP* stands for false positive, *FN* stands for false negative.

(6) *Area Under the Curve (AUC)*. The performance was tested at various thresholds using the receiver operating characteristics (ROC) to plot a graph of the true positive rate (sensitivity on the *y*-axis) versus the false positive rate (1 – specificity on the *x*-axis). The higher the mean AUC-ROC values, the better the model was for distinguishing between lung cancer and normal samples. In addition, we used precision–recall (PRC), which is a plot of the precision (positive predictive value on the *y*-axis) versus the recall (sensitivity or true positive rate on the *x*-axis) for all possible thresholds. The larger the value of AUC-PRC, the better the model's performance. If the positive and negative data were imbalanced, the PRC curve was preferred for checking the model's performance.

2.7. Functional Enrichment of Key Genes Obtained by the LASSO Model

Lung cancer signature genes identified by the LASSO were used to construct a separate interaction network of nodes and their first neighbors as per BioGRID. The signature genes were further analyzed for functional enrichment using Gene Ontology (biological processes) and for pathway enrichment using KEGG with DAVID version 6.8 [45].

3. Results

3.1. Identification of DEGs

In order to find the genes associated with lung cancer, we performed a differential gene expression analysis using the limma-voom tool [46]. The expression profiles of the 2500 genes with the greatest variance were used for PCA analysis. Three principal components comprising 40.1% (26.8%, 9.5%, and 3.8%) of the total variance showed that the lung cancer samples were clustered apart from the normal samples (Figure 1B). Furthermore, we obtained 2754 DEGs, including 1242 upregulated and 1512 downregulated genes in lung cancer compared with normal samples with $|\log2FC| > 2$, and adj.p.value < 0.001 (Figure 1C). According to the log2FC values, the top five upregulated genes were *CST1*, *FAM83A*, *KRT16*, *MMP13*, and *MMP12*, whereas the top five downregulated genes were *DEFA1B*, *SLC6A4*, *DEFA1*, *SFTPC*, and *CA4* (Table S2). The complete list of upregulated and downregulated genes is provided in Supplementary Tables S3 and S4, respectively.

3.2. Identification of the Relevant Interacting Genes

Human-related biological interaction data were obtained from BioGRID, which contains 40,843 nodes and 977,146 edges. This data in BioGRID also included some interactions involving species other than humans; therefore, the human-specific interactions were filtered, and we obtained 33,235 nodes with 909,098 edges. The network parameters, including the degree of the nodes, were calculated for this interaction network. The DEGs data were integrated into this interaction network, and the sub-network was filtered, as presented in Figure 2. The top 20 nodes according to the degree were further sorted on the basis of their log2FC value, thus revealing the top 20 upregulated hub genes and the top 20 downregulated hub genes (Table 2).

Table 2. The top 20 upregulated and 20 downregulated hub genes from the DEGs' interaction network.

Upregulate Genes			Downregulated Genes		
Log2FC	**Degree**	**Name**	**Log2FC**	**Degree**	**Name**
4.76	698	*SOX2*	−5.16	279	*GPR17*
4.33	804	*CDC20*	−5.06	297	*ZBTB16*
4.19	1143	*ANLN*	−4.13	278	*CMTM5*
4.08	1063	*KIF20A*	−3.66	723	*ACTC1*
3.73	1834	*KIF14*	−3.55	294	*USHBP1*
3.51	817	*AURKB*	−2.87	429	*TRIM63*
3.29	550	*MKI67*	−2.84	404	*ADRB2*
3.24	635	*CDK1*	−2.82	715	*LRRK2*
3.14	577	*RAD51*	−2.70	270	*NR4A1*
3.12	1207	*MCM2*	−2.70	764	*MEOX2*
3.10	695	*PLK1*	−2.69	843	*CAV1*
2.86	529	*CDKN2A*	−2.59	315	*CLEC4D*
2.62	1032	*KIF23*	−2.46	411	*CLEC4E*
2.58	934	*ECT2*	−2.43	455	*GPR182*
2.50	986	*PRC1*	−2.41	433	*SYNE3*
2.50	1465	*RECQL4*	−2.33	342	*CRYAB*
2.37	553	*KRT31*	−2.27	294	*KANK2*
2.35	1354	*EGLN3*	−2.19	297	*ALB*
2.31	849	*CDH1*	−2.09	367	*LMO2*
2.14	1189	*AGR2*	−2.07	348	*HECW2*

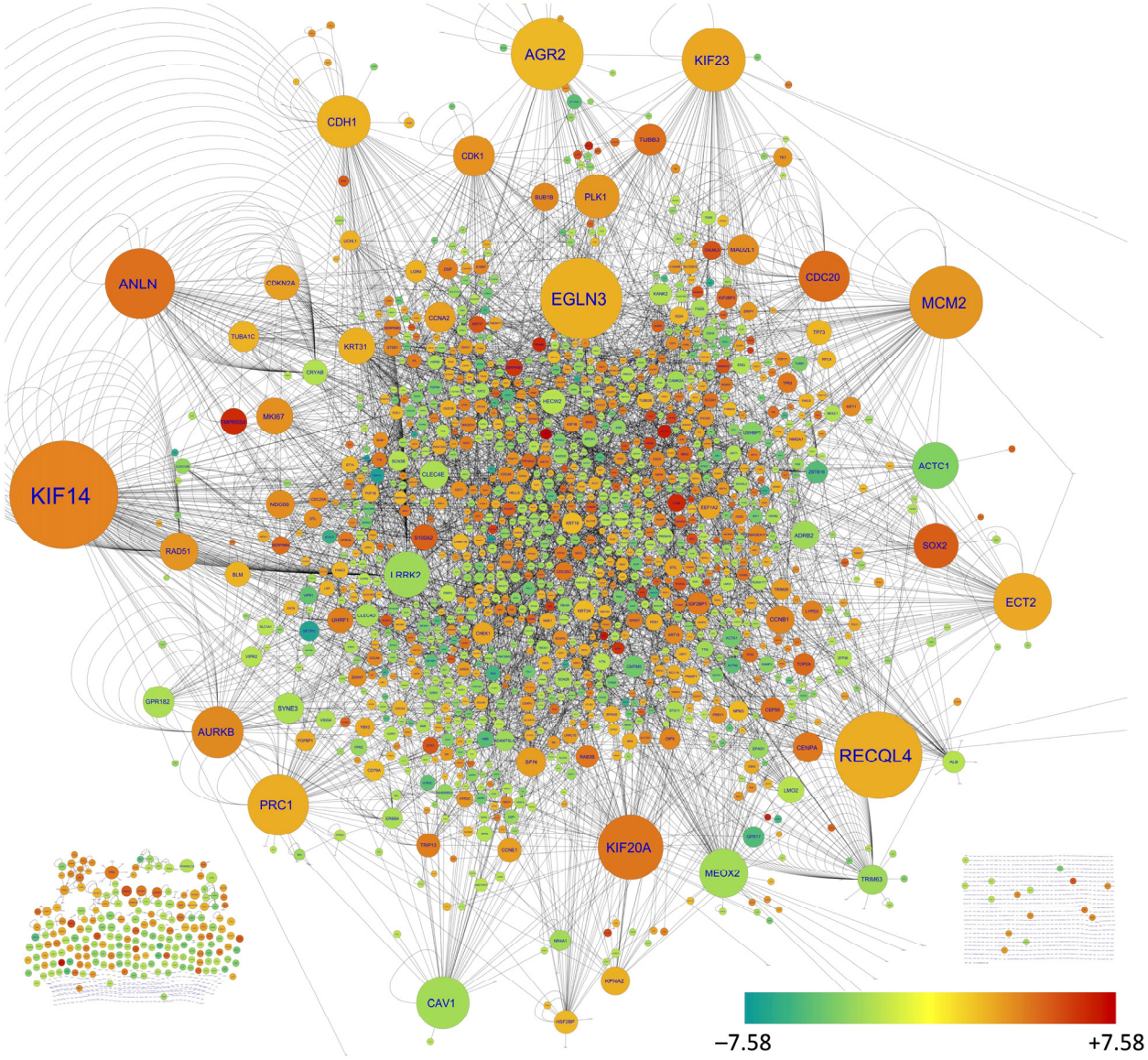

Figure 2. Biological interaction network of DEGs identified in NSCLC samples according to BioGRID v 4.4.205. The DEGs' nodes were filtered from all interactions available in BioGRID. The node sizes are arranged as per their degree in the original human interaction network and therefore indicate their central involvement in human cellular interactions. The colors of the nodes were determined by their log2FC value, where green to red represents negative to positive log2FC values.

3.3. Development of the LASSO Model

The gene expression profiles of 40 hub genes in lung cancer and control samples were used to build a classifier in order to predict lung cancer. First, we fitted the LASSO logistic regression model and plotted the coefficients at different log lambda values (Figure 3A). The plot displays the behavior of coefficients at different values of lambda. After that, we used 10-fold cv to find the best value of lambda that maximized the AUC curve (Figure 3B). We selected lambda.min (0.0005101641) as the best lambda and identified the 17 important genes with non-zero coefficients (Figure 3C, Table S5). Furthermore, the gene expression patterns of these 17 important genes were extracted from the TR dataset, and a heatmap was plotted, revealing that their expression patterns in lung cancer and the normal sample were distinct (Figure 3D). The set of 17 genes was further explored to find the gene family in the Molecular Signatures Database (MsigDB v7.5.1: http://www.gsea-msigdb.org/gsea/

msigdb/index.jsp, accessed on 15 June 2022) [47,48]. Based upon the protein homology or biochemical activity, we found that the dysregulated genes belonged to tumor suppressors, oncogenes, translocated cancer genes, and transcription factors (Table S6). Finally, the LASSO model was developed by using the expression profiles of 17 genes in the TR dataset at lambda.min (0.0005101641) for predicting lung cancer. The lung cancer signature genes in the LASSO model were present in the order of *KIF14, RAD51, CDKN2A, KIF23, RECQL4, EGLN3, CDH1, ZBTB16, CMTM5, ACTC1, ADRB2, NR4A1, CLEC4D, CLEC4E, SYNE3, CRYAB,* and *KANK2.* The model was constructed using the expression values of the 17 genes and their coefficients, and the risk score for lung cancer was calculated as follows.

$$\text{Risk score for NSCLC} = -3.207 + (-1.016 * KANK2) + (-0.929 * CLEC4D) + (-0.64 * ADRB2) + (-0.533 * CRYAB) + (-0.322 * NR4A1) + (-0.297 * CMTM5) + (-0.174 * ZBTB16) + (-0.12 * ACTC1) + (-0.118 * RAD51) + (-0.117 * KIF23) + (-0.087 * SYNE3) + (0.136 * CLEC4E) + (0.403 * CDKN2A) + (0.459 * EGLN3) + (0.675 * KIF14) + (1.372 * RECQL4) + (1.457 * CDH1)$$

where the gene name indicates its expression value in "RSEM norm_count". A gene is associated with a lower risk of lung cancer if its coefficient is less than zero (0). On the contrary, a gene is associated with a higher risk of lung cancer if its coefficient is greater than zero (0).

3.4. Performance of the LASSO Model on Independent Datasets

We evaluated the reliability of the LASSO model on the TD1 dataset, and found that the model achieved an accuracy, specificity, and sensitivity of 0.986, 0.959, and 0.995, respectively, at the 0.5 threshold (Table 3). The performance of the model showed an AUC-ROC and AUC-PRC of 0.9988 and 0.999, respectively, on the TD1 dataset (Figure 4A). Furthermore, we evaluated its performance on the TD2 dataset (GSE18842, GSE27262, and GSE19804) containing the gene expression patterns of NSCLC. On GSE18842, the LASSO model achieved an accuracy, specificity, and sensitivity of 1, 1, and 1, respectively, at the 0.5 threshold (Table S7). Furthermore, on the same data, the model showed an AUC-ROC and AUC-PRC of >0.99 and >0.99, respectively (Figure 4B). However, the performance of the LASSO model on GSE27262 achieved an accuracy, specificity, and sensitivity of 0.980, 1, and 0.960, respectively, at the 0.5 threshold (Table S8). Notably, the same model achieved 100% accuracy, specificity, and sensitivity when the threshold value was decreased to 0.4 as compared with 0.5 in the GSE27262 dataset (Table S8). The model showed an AUC-ROC and AUC-PRC of >0.99 and >0.99, respectively, on GSE27262 (Figure 4C). However, on the dataset of GSE19804, the model's performance was reduced, having an accuracy, AUC-ROC, and AUC-PRC of 0.725, 0.95, and 0.96, respectively (Figure 4D and Table S9).

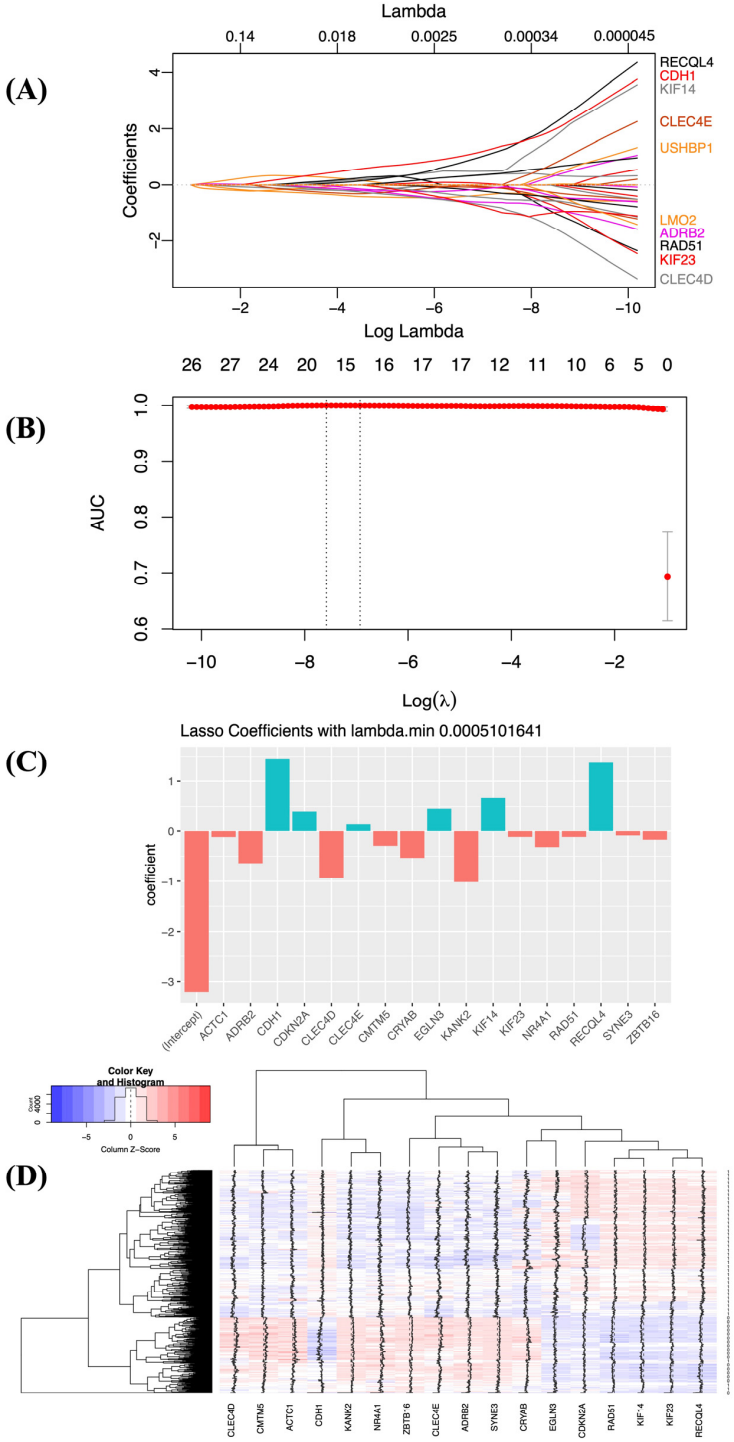

Figure 3. Construction of the risk score model for lung cancer prediction using LASSO logistic regression with 10-fold cv using *glmnet*. (**A**) LASSO regression coefficient profiles of 40 genes associated with lung cancer at different values of log lambda. Each curve indicates a gene and the path of its coefficient against the different values of log lambda. (**B**) This plot displays the AUC value (in red) with varying values of log lambda. The vertical dotted line at the left indicates the value of λ lambda.min that gives the maximum average AUC. The vertical dotted line at the right shows the largest value of λ lambda.1se; the performance is within one standard error of the maximum average AUC. The numbers across the top are the nonzero coefficient estimates. (**C**) Bar graph representing the regression coefficients for the most relevant genes (17 genes) at lambda.min = 0.0005101641. The blue-green bar represents positive coefficients; the red bar represents negative coefficients. (**D**) Heatmap of the expression patterns of relevant genes (17 genes) from the TR dataset.

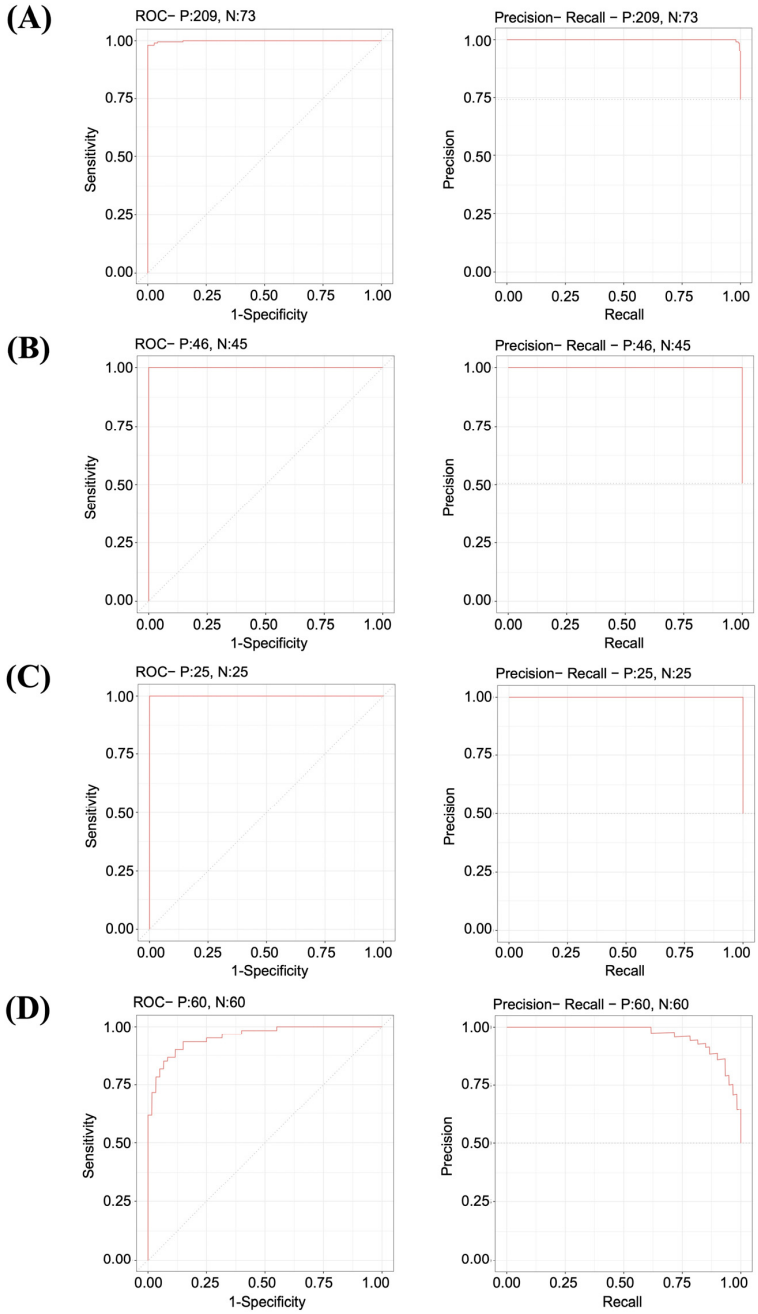

Figure 4. Performance of the LASSO model on the independent test datasets. (**A**) Performance on the TD1 dataset that contained 209 cancer and 73 normal samples, with the ROC curve showing an AUC of 0.9988 and the PRC curve showing an AUC of 0.999. (**B**) Performance on the GSE18842 dataset that contained 46 cancer and 45 normal samples, with the ROC curve showing an AUC of >0.99 and the PRC curve showing an AUC of >0.99. (**C**) Performance on the GSE27262 dataset that contained 25 cancer and 25 normal samples, with the ROC curve showing an AUC of >0.99 and the PRC curve showing an AUC of >0.99. (**D**) Performance on the GSE19804 dataset that contained 60 cancer and 60 normal samples, with the ROC curve showing an AUC of 0.95 and the PRC curve showing an AUC of 0.96. The ROC graphs plot the true positive rate (sensitivity on the *y*-axis) versus the false positive rate (1-specificity on the *x*-axis) for all possible thresholds. The value of the AUC varies from 0 to 1. The larger the value of the AUC, the better the model can differentiate between lung cancer and normal samples. The diagonal dashed line represents an AUC of 0.5, which indicates random prediction by the model. The PRC plots the precision (positive predictive value on the *y*-axis) versus the recall (sensitivity or true positive rate on the *x*-axis) for all possible thresholds. The larger the AUC, the better the model's performance. The ROC and PRC curves were built with the R package *precrec*.

Table 3. Performance of the LASSO model on the test dataset TD1.

Threshold	Accuracy	Specificity	Sensitivity	TN	TP	FN	FP	NPV	PPV
0	0.741	0.000	1.000	0	209	0	73	NA	0.741
0.1	0.982	0.945	0.995	69	208	1	4	0.986	0.981
0.2	0.982	0.945	0.995	69	208	1	4	0.986	0.981
0.3	0.982	0.945	0.995	69	208	1	4	0.986	0.981
0.4	0.982	0.945	0.995	69	208	1	4	0.986	0.981
0.5	**0.986**	**0.959**	**0.995**	**70**	**208**	**1**	**3**	**0.986**	**0.986**
0.6	0.982	0.959	0.990	70	207	2	3	0.972	0.986
0.7	0.986	0.973	0.990	71	207	2	2	0.973	0.990
0.8	0.982	0.973	0.986	71	206	3	2	0.959	0.990
0.9	0.986	1.000	0.981	73	205	4	0	0.948	1.000
1	0.259	1.000	0.000	73	0	209	0	0.259	NA

NA means not available.

3.5. Comparative Analysis of Logistic Regression Models

Furthermore, we also examined the performance of models developed via logistic regression using *glm* from the R package. The logistic regression models were developed on the TR dataset, and their performance was assessed using the test dataset TD1. We found that the logistic regression model developed from 40 genes identified on the basis of the log2FC and node degree achieved an AUC-ROC of 0.9828. In comparison, the performance was slightly reduced (AUC-ROC: 0.9789) when the model was developed by using the 17-gene signature (Figure S1). On the basis of the ROC curve, we concluded that the signatures of the 17 genes achieved better performance with the LASSO model (AUC: 0.9988) compared with logistic regression (AUC: 9789). The LASSO regression selected the important features by shrinking the coefficient towards zero, which also had the advantage of avoiding model overfitting, and interpreting the possible roles of the features in lung cancer.

3.6. Interaction Network and Functional Enrichment Analysis of Genes from the LASSO Model

To understand the biological function of the signature genes, we performed a functional enrichment analysis with the DAVID bioinformatics tool (version 6.8). Figure 5A represents the interaction network of the nodes and their first neighbors. Functional enrichment analysis of 17 genes revealed that the signature genes were significantly involved in the cancer pathway and apoptotic process (Figure 5B).

3.7. Validation of the 17-Gene Signature in Lung Cancer Data

We validated our identified 17-gene signature for NSCLC in various experimental studies using the Expression Atlas (release 38; https://www.ebi.ac.uk/gxa/home, accessed on 7 August 2022). We took the gene name, selected *Homo sapiens* as the species, and lung cancer as the biological condition, and submitted these to the database. Next, the differential expression data were downloaded with "diseases" as the experimental variables, and were considered only the data with a comparison between cancer vs. normal with $|\log2FC| > 2$. The result identified that 14 genes were differentially expressed out of 17 across 32 studies (Table S10). We found positive LASSO coefficients for the genes *CDKN2A, EGLN3, KIF14*, and *RECQL4* that were upregulated in cancer compared with normal samples. On the contrary, negative LASSO coefficients were found for the genes *ADRB2, CRYAB, NR4A1, CMTM5, ZBTB16, SYNE3*, and *RAD51*, which were downregulated in cancer compared with normal samples. Negative LASSO coefficients were found for the genes *KANK2* and *CLEC4D* that were downregulated in cancer compared with normal samples, but their

log2FC was between −1 and −2. Thus, the results above support the efficient selection of genes by the LASSO for predicting NSCLC. However, we also found positive LASSO coefficients for the genes *CLEC4E*, and *CDH1* that were downregulated, and a negative LASSO coefficient for the gene *KIF23*, which was both upregulated and downregulated in different studies.

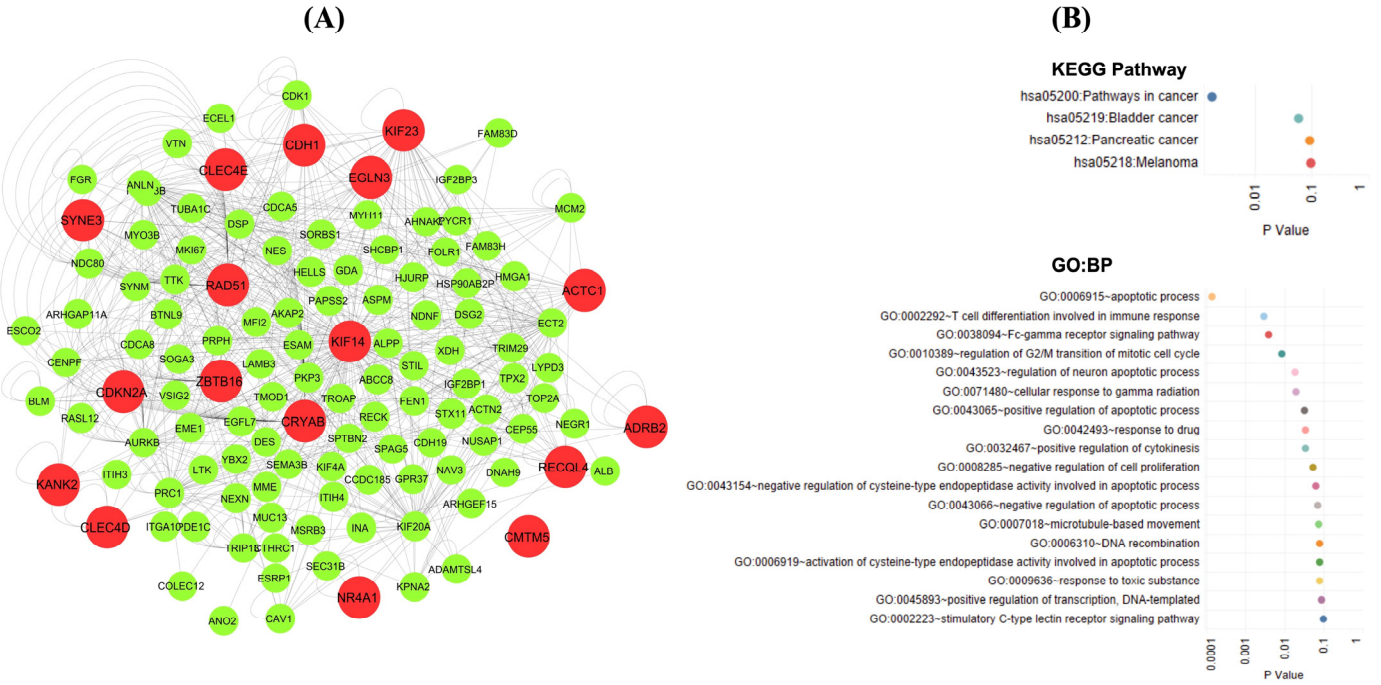

Figure 5. Sub-network of the signature genes and their functional enrichment identified through the LASSO. (**A**) Interaction network of genes identified through transcriptome–interactome signatures and their first neighbors. The genes are represented as red nodes, while their first neighbors are shown in green. (**B**) Functional enrichment analysis of 17 targets against the pathway database KEGG and biological process in the Gene Ontology database.

4. Discussion

Detecting cancer at its early stage is the foremost goal of preventing cancer's proliferation and metastasis. Therefore, developing highly accurate and reliable molecular diagnostic tools for cancer, including predictive and prognostic models, is indispensable for diagnosis at the early stage and finding suitable treatment modalities [49]. Lung cancer is a significant reason for cancer-associated fatality [1], which is mainly categorized into small cell lung cancer (~15% cases) and non-small cell lung cancer (NSCLC, ~85% cases). The main histological types of NSCLC are adenocarcinoma and squamous cell carcinoma [50,51]. It has been suggested that early surgical resection of NSCLC can increase the 5-year survival by up to 70%; unfortunately, almost 75% of cases are detected at the time of advanced disease (Stages III/IV), making it difficult to manage the disease despite significant modern advancements in oncology practice [51]. Furthermore, past works have identified various diagnostic and prognostic signatures in cancers for patient stratification using gene expression and omics data; however, this approach has failed to capture the synergistic effects of gene expression [17–19]. Therefore, applying AI and MLT to omics data could be a promising approach for identifying and developing better prognostic and diagnostic models in cancers. The uses of AI in lung cancer detection have been a focus of research and gained significant scientific attention during the recent SARS-CoV-2 pandemic. The application of AI-based models using chest X-rays, CT scans, and PET scans have been suggested as important methods for detecting lung cancer [52,53]. Though the approach of image-based lung cancer detection is important for tackling this challenge, the use of

molecular signatures can greatly surpass AI-based detection of the morphological changes, which become apparent after a long carcinogenic molecular transformation.

A previous study identified 17 candidate genes in lung adenocarcinoma for predicting survival in non-smoking patients [54]. The study used weighted gene co-expression network analysis (WGCNA) and LASSO Cox regression to identify the prognostic signature; however, the model achieved an AUC-ROC of 0.736 on the training dataset [54]. Another study also used WGCNA and LASSO Cox regression, and identified four genes that predicted high and low overall survival in lung adenocarcinoma with an AUC-ROC of 0.71 on the training dataset [55]. Most of the previous studies focused on developing prognostic models for lung cancer; however, it is imperative to develop a model for predicting lung cancer with the high accuracy required for early detection and better management of patients. Furthermore, integrating the gene expression and interaction data has huge potential to identify the crucial genes associated with disease initiation. Therefore, this study implemented and identified the molecular transcriptome–interactome signatures for developing a LASSO-based machine learning model for predicting NSCLC. First, we identified the DEGs in lung cancer compared with normal samples using the lung-associated TCGA and GTEx data. Next, the human-specific interaction data were downloaded from BioGRID version 4.4.205, a continuously updated, large biomedical interaction repository currently holding almost 2.3 million proteins and genetic interactions from more than 78,000 publications [40]. The human interactions in BioGRID are also available with relevant literature references and therefore represent high-quality interaction data. In addition, the Cytoscape tool was used for reconstruction, visualization, and analysis of the biological network [41]. The important nodes from the DEGs' interaction network were constructed on the basis of their degree, reflecting each node's centrality in a particular interaction network. Hence, we identified the important nodes on the basis of their degree [56], indicating the identification of crucial nodes that may be involved in the proliferation of lung cancer. However, identifying the crucial genes that are relevant to detecting lung cancer is challenging. Therefore, we used the LASSO for feature selection and development of a model that identified a combined expression pattern of 17 genes (*KANK2*, *CLEC4D*, *ADRB2*, *CRYAB*, *NR4A1*, *CMTM5*, *ZBTB16*, *ACTC1*, *RAD51*, *KIF23*, *SYNE3*, *CLEC4E*, *CDKN2A*, *EGLN3*, *KIF14*, *RECQL4*, and *CDH1*) and their associated coefficients as a robust predictor of NSCLC. The performance of our developed LASSO model was highly accurate, with an AUC-ROC greater than 0.99 on most of the independent datasets of NSCLC, indicating that the selected 17-gene signature might be crucial for developing NSCLC (Figure 4). These genes belong to various categories, including tumor suppressors, oncogenes, translocated cancer genes, and transcription factors (Table S6). Furthermore, we validated our 17-gene signature across several studies and found most of these genes were differentially expressed; thus, our finding is supported by other studies (Table S10).

Among the 17 signature genes, *KANK2*, *CLEC4D*, *ADRB2*, *CRYAB*, *NR4A1*, *CMTM5*, *ZBTB16*, *ACTC1*, and *SYNE3* showed downregulation in NSCLC with negative LASSO coefficients. Genes with negative coefficients indicate a lower risk of lung cancer if their expression is upregulated. *KANK2* gene encoding protein, also known as SRC interacting protein, is involved in transcription regulation and caspase-independent apoptosis. It is a tumor suppressor gene, and its downregulation is associated with NSCLC [57]. The mRNA expression level of *CLEC4D* was reported to be significantly lower in hepatocellular carcinoma [58]. According to GTEx V8 (https://gtexportal.org/, accessed on 15 October 2022), the lung is one of the tissues with high expression of *CLEC4D* mRNA. The gene *ADRB2* codes for the beta-2-adrenergic receptor, and its downregulation and polymorphisms are associated with lung cancer [59–61]. The alpha B-crystallin (encoded by *CRYAB*) is a molecular chaperon that binds to avert the aggregation of misfolded proteins and to inhibit apoptosis [62,63]. Studies have shown that the high expression of *CRYAB* is associated with tumor development and is a marker of poor prognosis for head and neck cancer [64], and breast cancer [65]. On the contrary, the role of *CRYAB* in lung cancer is

controversial and needs more study [66]. *CMTM5* acts as a tumor-suppressor gene, and it is downregulated in several cancers, such as myeloid leukemia, ovarian cancer, prostate cancer, cervical carcinoma, and pancreatic cancer [67]. *ZBTB16* encodes for a zinc finger TF and is associated with the progression of the cell cycle. *ZBTB16* is underexpressed in multiple cancer types, including lung cancer [68]. Therefore, the selection and inclusion of downregulated genes with a negative coefficient in our LASSO model justified its high predictive accuracy, warranting further in vitro experiments to understand the mechanism(s) of NSCLC development.

The *CDKN2A, EGLN3, KIF14, RECQL4,* and *CDH1* genes showed upregulation in NSCLC and had positive LASSO coefficients. Genes with positive coefficients increase the risk of lung cancer if their expression is upregulated. EGLN3 is a member of *Caenorhabditis elegans* gene *egl-9* (*EGLN*) family of oxygen- and α-ketoglutarate dependent prolyl hydroxylases. EGLN3 catalyzes the hydroxylation of extracellular signal-regulated kinase 3 (Erk3) and increased its stability, which is recognized as a strong potent driver of cancers [69]. Thus, our finding has been validated by other studies where the *EGLN3* was reported to be vital for the growth of numerous cancers, including lung cancer [69]. The upregulated kinesin family member gene *KIF14* is a mitotic kinesin and plays an essential role in tumor development. Similar to lung cancer, overexpression of *KIF14* was also reported in several cancers, and the upregulation of this kinesin family member gene has been associated with poor prognosis [70]. RECQL4, a helicase known as a molecular motor, is involved in unwrapping the DNA, an essential event during DNA replication and DNA repair. Notably, the chromosomal site of the RECQL4 gene is considered as a hot-spot position for frequent mutation often highly detected in sporadic breast cancers [71]. Furthermore, Arora et al. demonstrated that the depletion of RECQL4 levels led to weakening of the DNA duplication rate and increased chemosensitivity in cultured breast cancer cells. Thus, their study confirmed that RECQL4 upregulation is linked with tumor progression in breast cancers [71]. Furthermore, another study showed that a high expression of the *BLM* gene, a paralog of RECQL4, was associated with poor prognosis in lung cancer [72]. Hence, we anticipate that further study of these genes in a model of a lung cancer cell line will eventually shed some more light on their involvement in NSCLC's development and progression.

These feature genes are involved in several aspects of cancer progression as documented, and the important role of these targets in NSCLC indicates the importance of their detection by the LASSO model, which was also evident while performing a functional enrichment analysis of these target genes with DAVID Bioinformatics Resource 6.8. (Figure 5B) [73]. The functional overrepresentation analysis against the KEGG pathway database revealed that these targets are associated with cancer pathways. Moreover, analysis against Gene Ontology biological process terms indicated that these targets are involved in regulation of apoptosis, a crucial pathway dysregulated during cancer development. The inclusive picture involving the feature genes and their functional overrepresentation analyses revealed the importance of these factors in developing the MLT-based model.

The LASSO-based model has been used to diagnose other diseases, indicating its potential for detecting cancers [74,75]. As noted previously, early cancer detection is key to preventing several cancer-associated complications. In addition, this can also reduce the significant economic burden on the healthcare system by reducing the chance of metastasis and mortality. Our study used a systems biology and LASSO-based approach, and identified the transcriptome–interactome signatures that achieved high accuracy in predicting NSCLC. Thus, evidence of the high accuracy of our model indicated that the strategy of integrating transcriptome–interactome signatures has enormous potential to develop better models for predicting other diseases, including various cancers. In addition to the late diagnosis, other obstacles preventing the long-term survival of NSCLC patients include a lack of advanced treatment and an accurate prognosis model due to the disease's heterogeneity, as well as differences in cancer care facilities across the world.

Our study has a few limitations, including the following. Firstly, the publicly available data are imbalanced and contain many NSCLC cancer samples compared with normal samples. Therefore, we used the AUC-ROC and AUC-PRC curves to check the performance of the model developed on imbalanced data at different threshold values. Second, our model did not include gene mutations, intra-tumoral heterogeneity, and other clinical features associated with cancer. Third, we identified a 17-gene signature that needs to be further validated using qRT-PCR in several clinical samples of NSCLC. Finally, tissue biopsies are needed to quantify the genes' expression levels, which are invasive, costly, and time-consuming.

5. Conclusions

In summary, we conducted an integrative approach to identify the transcriptome and interactome signatures for discriminating NSCLC from normal samples. We then applied LASSO logistic regression to find a 17-gene signature and developed a model for predicting NSCLC. The performance of our model showed high accuracy across several independent datasets of NSCLC. Finally, we developed a web application, *NSCLCpred* (https://hifzuransari.shinyapps.io/NSCLCpred/, accessed on 31 October 2022), for detecting NSCLC using the expression profile of 17 genes. Our findings could be helpful in creating a new strategy for diagnosing NSCLC patients. Furthermore, we expect our identified gene signature to provide novel insights and therapeutic targets for NSCLC.

Supplementary Materials: The following supporting information can be downloaded at https://www.mdpi.com/article/10.3390/biology11121752/s1. Figure S1: Performance of logistic regression models on the test dataset 1 (TD1). Table S1: The number of samples in the test dataset 2 (TD2) used to assess the performance of the LASSO model. Table S2: List of the top 20 DEGs based on log2FC values. Table S3: Upregulation of genes in lung cancer compared with normal samples. A gene was considered to be upregulated at log2FC > 2 and adj.p.value < 0.001. Table S4: Downregulation of genes in lung cancer compared with normal samples. A gene was considered to be downregulated at log2FC < −2 and adj.p.value < 0.001. Table S5: Genes with non-zero coefficients selected by using LASSO. Table S6: The 17-gene signature of the LASSO model is associated with different gene families according to the Molecular Signature Database (MSigDB). Table S7: Performance of the LASSO model on the independent TD2 dataset GSE18842. Table S8: Performance of the LASSO model on the independent TD2 dataset GSE27262. Table S9: Performance of the LASSO model on the independent TD2 dataset GSE19804. Table S10: Expression patterns of the 17-gene signature across various studies on lung cancer. The data were extracted from Expression Atlas release 38 (https://www.ebi.ac.uk/gxa/home, accessed on 7 August 2022).

Author Contributions: F.A. is the PI who conceived the idea and designed the project; collected, analyzed, and interpreted the data; developed the LASSO model; and wrote and revised the manuscript. A.A.K. refined the project; collected, analyzed, and interpreted the data; generated the biological networks; and wrote and revised the manuscript. H.R.A. checked the data's accuracy and analyzed the data, developed the model and R/Shiny web tool, and wrote and revised the manuscript. A.H. interpreted the results, and wrote and revised the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded by the Deanship of Scientific Research (DSR), University of Jeddah, Jeddah, under grant No. (UJ-02-019-DR).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The computer code, LASSO Model, and supporting data of *NSCLCpred* are available at the GitHub repository (https://github.com/firozimtech/NSCLCpred, accessed on 31 October 2022). The R/Shiny web application of *NSCLCpred* is available at https://hifzuransari.shinyapps.io/NSCLCpred/ (accessed on 31 October 2022).

Acknowledgments: This work was funded by the Deanship of Scientific Research (DSR), University of Jeddah, Jeddah, under grant No. (UJ-02-019-DR). The authors, therefore, acknowledge with thanks to DSR, the University of Jeddah, for its technical and financial support.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Sung, H.; Ferlay, J.; Siegel, R.L.; Laversanne, M.; Soerjomataram, I.; Jemal, A.; Bray, F. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA A Cancer J. Clin.* **2021**, *71*, 209–249. [CrossRef] [PubMed]
2. Remon, J.; Hendriks, L.E.L. Targeted therapies for unresectable stage III non-small cell lung cancer. *Mediastinum* **2021**, *5*, 22. [CrossRef] [PubMed]
3. Beckett, P.; Tata, L.J.; Hubbard, R.B. Risk factors and survival outcome for non-elective referral in non-small cell lung cancer patients–analysis based on the National Lung Cancer Audit. *Lung Cancer* **2014**, *83*, 396–400. [CrossRef] [PubMed]
4. Iyer, S.; Taylor-Stokes, G.; Roughley, A. Symptom burden and quality of life in advanced non-small cell lung cancer patients in France and Germany. *Lung Cancer* **2013**, *81*, 288–293. [CrossRef] [PubMed]
5. Walker, M.S.; Wong, W.; Ravelo, A.; Miller, P.J.E.; Schwartzberg, L.S. Effectiveness outcomes and health related quality of life impact of disease progression in patients with advanced nonsquamous NSCLC treated in real-world community oncology settings: Results from a prospective medical record registry study. *Health Qual. Life Outcomes* **2017**, *15*, 160. [CrossRef]
6. Grant, M.; Sun, V.; Fujinami, R.; Sidhu, R.; Otis-Green, S.; Juarez, G.; Klein, L.; Ferrell, B. Family caregiver burden, skills preparedness, and quality of life in non-small cell lung cancer. *Oncol. Nurs. Forum* **2013**, *40*, 337–346. [CrossRef]
7. Gibson, G.J.; Loddenkemper, R.; Lundbäck, B.; Sibille, Y. Respiratory health and disease in Europe: The new European Lung White Book. *Eur. Respir. J.* **2013**, *42*, 559–563. [CrossRef]
8. Soda, H.; Tomita, H.; Kohno, S.; Oka, M. Limitation of annual screening chest radiography for the diagnosis of lung cancer. A retrospective study. *Cancer* **1993**, *72*, 2341–2346. [CrossRef]
9. Prorok, P.C.; Andriole, G.L.; Bresalier, R.S.; Buys, S.S.; Chia, D.; Crawford, E.D.; Fogel, R.; Gelmann, E.P.; Gilbert, F.; Hasson, M.A.; et al. Design of the Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Trial. *Control. Clin. Trials* **2000**, *21*, 273S–309S. [CrossRef]
10. Prabhakar, B.; Shende, P.; Augustine, S. Current trends and emerging diagnostic techniques for lung cancer. *Biomed. Pharmacother.* **2018**, *106*, 1586–1599. [CrossRef]
11. Rodríguez, J.; Avila, J.; Rolfo, C.; Ruíz-Patiño, A.; Russo, A.; Ricaurte, L.; Ordóñez-Reyes, C.; Arrieta, O.; Zatarain-Barrón, Z.L.; Recondo, G.; et al. When Tissue is an Issue the Liquid Biopsy is Nonissue: A Review. *Oncol. Ther.* **2021**, *9*, 89–110. [CrossRef]
12. Goebel, C.; Louden, C.L.; McKenna, R., Jr.; Onugha, O.; Wachtel, A.; Long, T. Diagnosis of Non-small Cell Lung Cancer for Early Stage Asymptomatic Patients. *Cancer Genom. Proteom.* **2019**, *16*, 229–244. [CrossRef]
13. Wang, J.; Liu, Q.; Yuan, S.; Xie, W.; Liu, Y.; Xiang, Y.; Wu, N.; Wu, L.; Ma, X.; Cai, T.; et al. Genetic predisposition to lung cancer: Comprehensive literature integration, meta-analysis, and multiple evidence assessment of candidate-gene association studies. *Sci. Rep.* **2017**, *7*, 8371. [CrossRef]
14. Walser, T.; Cui, X.; Yanagawa, J.; Lee, J.M.; Heinrich, E.; Lee, G.; Sharma, S.; Dubinett, S.M. Smoking and lung cancer: The role of inflammation. *Proc. Am. Thorac. Soc.* **2008**, *5*, 811–815. [CrossRef]
15. Dietel, M.; Bubendorf, L.; Dingemans, A.M.; Dooms, C.; Elmberger, G.; Garcia, R.C.; Kerr, K.M.; Lim, E.; Lopez-Rios, F.; Thunnissen, E.; et al. Diagnostic procedures for non-small-cell lung cancer (NSCLC): Recommendations of the European Expert Group. *Thorax* **2016**, *71*, 177–184. [CrossRef]
16. Ahmed, F. Integrated Network Analysis Reveals FOXM1 and MYBL2 as Key Regulators of Cell Proliferation in Non-small Cell Lung Cancer. *Front. Oncol.* **2019**, *9*, 1011. [CrossRef]
17. Giannos, P.; Kechagias, K.S.; Gal, A. Identification of Prognostic Gene Biomarkers in Non-Small Cell Lung Cancer Progression by Integrated Bioinformatics Analysis. *Biology* **2021**, *10*, 1200. [CrossRef]
18. Feng, Z.; Qu, J.; Liu, X.; Liang, J.; Li, Y.; Jiang, J.; Zhang, H.; Tian, H. Integrated bioinformatics analysis of differentially expressed genes and immune cell infiltration characteristics in Esophageal Squamous cell carcinoma. *Sci. Rep.* **2021**, *11*, 16696. [CrossRef]
19. Giannos, P.; Kechagias, K.S.; Bowden, S.; Tabassum, N.; Paraskevaidi, M.; Kyrgiou, M. PCNA in Cervical Intraepithelial Neoplasia and Cervical Cancer: An Interaction Network Analysis of Differentially Expressed Genes. *Front. Oncol.* **2021**, *11*, 779042. [CrossRef]
20. Cassidy, A.; Duffy, S.W.; Myles, J.P.; Liloglou, T.; Field, J.K. Lung cancer risk prediction: A tool for early detection. *Int. J. Cancer* **2007**, *120*, 1–6. [CrossRef]
21. Gray, E.P.; Teare, M.D.; Stevens, J.; Archer, R. Risk Prediction Models for Lung Cancer: A Systematic Review. *Clin. Lung Cancer* **2016**, *17*, 95–106. [CrossRef] [PubMed]
22. Ahmad, A.S.; Mayya, A.M. A new tool to predict lung cancer based on risk factors. *Heliyon* **2020**, *6*, e03402. [CrossRef] [PubMed]
23. Yeh, M.C.; Wang, Y.H.; Yang, H.C.; Bai, K.J.; Wang, H.H.; Li, Y.J. Artificial Intelligence-Based Prediction of Lung Cancer Risk Using Nonimaging Electronic Medical Records: Deep Learning Approach. *J. Med. Internet Res.* **2021**, *23*, e26256. [CrossRef] [PubMed]

24. Yang, D.; Ma, X.; Song, P. A prognostic model of non small cell lung cancer based on TCGA and ImmPort databases. *Sci. Rep.* **2022**, *12*, 437. [CrossRef] [PubMed]
25. Niu, B.; Li, J.; Li, G.; Poon, S.; Harrington, P.B. Analysis and Modeling for Big Data in Cancer Research. *BioMed Res. Int.* **2017**, *2017*, 1972097. [CrossRef]
26. Shait Mohammed, M.R.; Zamzami, M.; Choudhry, H.; Ahmed, F.; Ateeq, B.; Khan, M.I. The Histone H3K27me3 Demethylases KDM6A/B Resist Anoikis and Transcriptionally Regulate Stemness-Related Genes. *Front. Cell Dev. Biol.* **2022**, *10*, 780176. [CrossRef]
27. Chin, L.; Hahn, W.C.; Getz, G.; Meyerson, M. Making sense of cancer genomic data. *Genes Dev.* **2011**, *25*, 534–555. [CrossRef]
28. Karimi, M.R.; Karimi, A.H.; Abolmaali, S.; Sadeghi, M.; Schmitz, U. Prospects and challenges of cancer systems medicine: From genes to disease networks. *Brief. Bioinform.* **2022**, *23*, bbab343. [CrossRef]
29. Ahmed, F.; Kumar, M.; Raghava, G.P. Prediction of polyadenylation signals in human DNA sequences using nucleotide frequencies. *Silico Biol.* **2009**, *9*, 135–148. [CrossRef]
30. Ahmed, F.; Raghava, G.P. Designing of highly effective complementary and mismatch siRNAs for silencing a gene. *PLoS ONE* **2011**, *6*, e23443. [CrossRef]
31. Ahmed, F.; Senthil-Kumar, M.; Dai, X.; Ramu, V.S.; Lee, S.; Mysore, K.S.; Zhao, P.X. pssRNAit: A Web Server for Designing Effective and Specific Plant siRNAs with Genome-Wide Off-Target Assessment. *Plant Physiol.* **2020**, *184*, 65–81. [CrossRef]
32. Ahmed, F.; Ansari, H.R.; Raghava, G.P. Prediction of guide strand of microRNAs from its sequence and secondary structure. *BMC Bioinform.* **2009**, *10*, 105. [CrossRef]
33. Ahmed, F.; Kaundal, R.; Raghava, G.P. PHDcleav: A SVM based method for predicting human Dicer cleavage sites using sequence and secondary structure of miRNA precursors. *BMC Bioinform.* **2013**, *14* (Suppl. 14), S9. [CrossRef]
34. Elemento, O.; Leslie, C.; Lundin, J.; Tourassi, G. Artificial intelligence in cancer research, diagnosis and therapy. *Nat. Rev. Cancer* **2021**, *21*, 747–752. [CrossRef]
35. Arjmand, B.; Hamidpour, S.K.; Tayanloo-Beik, A.; Goodarzi, P.; Aghayan, H.R.; Adibi, H.; Larijani, B. Machine Learning: A New Prospect in Multi-Omics Data Analysis of Cancer. *Front. Genet.* **2022**, *13*, 824451. [CrossRef]
36. Liu, R.; Rizzo, S.; Whipple, S.; Pal, N.; Pineda, A.L.; Lu, M.; Arnieri, B.; Lu, Y.; Capra, W.; Copping, R.; et al. Evaluating eligibility criteria of oncology trials using real-world data and AI. *Nature* **2021**, *592*, 629–633. [CrossRef]
37. Weinstein, J.N.; Collisson, E.A.; Mills, G.B.; Shaw, K.R.; Ozenberger, B.A.; Ellrott, K.; Shmulevich, I.; Sander, C.; Stuart, J.M.; Network, C.G.A.R. The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.* **2013**, *45*, 1113–1120. [CrossRef]
38. Carithers, L.J.; Ardlie, K.; Barcus, M.; Branton, P.A.; Britton, A.; Buia, S.A.; Compton, C.C.; DeLuca, D.S.; Peter-Demchok, J.; Gelfand, E.T.; et al. A Novel Approach to High-Quality Postmortem Tissue Procurement: The GTEx Project. *Biopreserv. Biobank.* **2015**, *13*, 311–319. [CrossRef]
39. Blighe, K.; Rana, S.; Lewis, M. *EnhancedVolcano: Publication-Ready Volcano Plots with Enhanced Colouring and Labeling*. 2018. Available online: https://github.com/kevinblighe/EnhancedVolcano (accessed on 10 November 2021).
40. Oughtred, R.; Rust, J.; Chang, C.; Breitkreutz, B.J.; Stark, C.; Willems, A.; Boucher, L.; Leung, G.; Kolas, N.; Zhang, F.; et al. The BioGRID database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions. *Protein Sci. A Publ. Protein Soc.* **2021**, *30*, 187–200. [CrossRef]
41. Killcoyne, S.; Carter, G.W.; Smith, J.; Boyle, J. Cytoscape: A community-based framework for network modeling. *Methods Mol. Biol.* **2009**, *563*, 219–239. [CrossRef]
42. Sanchez-Palencia, A.; Gomez-Morales, M.; Gomez-Capilla, J.A.; Pedraza, V.; Boyero, L.; Rosell, R.; Fárez-Vidal, M.E. Gene expression profiling reveals novel biomarkers in nonsmall cell lung cancer. *Int. J. Cancer* **2011**, *129*, 355–364. [CrossRef] [PubMed]
43. Wei, T.Y.; Juan, C.C.; Hisa, J.Y.; Su, L.J.; Lee, Y.C.; Chou, H.Y.; Chen, J.M.; Wu, Y.C.; Chiu, S.C.; Hsu, C.P.; et al. Protein arginine methyltransferase 5 is a potential oncoprotein that upregulates G1 cyclins/cyclin-dependent kinases and the phosphoinositide 3-kinase/AKT signaling cascade. *Cancer Sci.* **2012**, *103*, 1640–1650. [CrossRef] [PubMed]
44. Lu, T.P.; Tsai, M.H.; Lee, J.M.; Hsu, C.P.; Chen, P.C.; Lin, C.W.; Shih, J.Y.; Yang, P.C.; Hsiao, C.K.; Lai, L.C.; et al. Identification of a novel biomarker, SEMA5A, for non-small cell lung carcinoma in nonsmoking women. *Cancer Epidemiol. Biomark. Prev.* **2010**, *19*, 2590–2597. [CrossRef] [PubMed]
45. Sherman, B.T.; Hao, M.; Qiu, J.; Jiao, X.; Baseler, M.W.; Lane, H.C.; Imamichi, T.; Chang, W. DAVID: A web server for functional enrichment analysis and functional annotation of gene lists (2021 update). *Nucleic Acids Res.* **2022**, *50*, W216–W3221. [CrossRef] [PubMed]
46. Law, C.W.; Chen, Y.; Shi, W.; Smyth, G.K. voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.* **2014**, *15*, R29. [CrossRef]
47. Liberzon, A.; Birger, C.; Thorvaldsdóttir, H.; Ghandi, M.; Mesirov, J.P.; Tamayo, P. The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst.* **2015**, *1*, 417–425. [CrossRef]
48. Subramanian, A.; Tamayo, P.; Mootha, V.K.; Mukherjee, S.; Ebert, B.L.; Gillette, M.A.; Paulovich, A.; Pomeroy, S.L.; Golub, T.R.; Lander, E.S.; et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 15545–15550. [CrossRef]
49. Hawkes, N. Cancer survival data emphasise importance of early diagnosis. *Bmj* **2019**, *364*, l408. [CrossRef]
50. Herbst, R.S.; Heymach, J.V.; Lippman, S.M. Lung cancer. *N. Engl. J. Med.* **2008**, *359*, 1367–1380. [CrossRef]
51. Blandin Knight, S.; Crosbie, P.A.; Balata, H.; Chudziak, J.; Hussell, T.; Dive, C. Progress and prospects of early detection in lung cancer. *Open Biol.* **2017**, *7*, 170070. [CrossRef]

52. Goncalves, S.; Fong, P.C.; Blokhina, M. Artificial intelligence for early diagnosis of lung cancer through incidental nodule detection in low- and middle-income countries-acceleration during the COVID-19 pandemic but here to stay. *Am. J. Cancer Res.* **2022**, *12*, 1–16.
53. Joshi, S.; Pandit, S.V.; Shukla, P.K.; Almalki, A.H.; Othman, N.A.; Alharbi, A.; Alhassan, M. Analysis of Smart Lung Tumour Detector and Stage Classifier Using Deep Learning Techniques with Internet of Things. *Comput. Intell. Neurosci.* **2022**, *2022*, 4608145. [CrossRef]
54. Mao, Q.; Zhang, L.; Zhang, Y.; Dong, G.; Yang, Y.; Xia, W.; Chen, B.; Ma, W.; Hu, J.; Jiang, F.; et al. A network-based signature to predict the survival of non-smoking lung adenocarcinoma. *Cancer Manag. Res.* **2018**, *10*, 2683–2693. [CrossRef]
55. Wang, H.; Lu, D.; Liu, X.; Jiang, J.; Feng, S.; Dong, X.; Shi, X.; Wu, H.; Xiong, G.; Cai, K. Survival-related risk score of lung adenocarcinoma identified by weight gene co-expression network analysis. *Oncol. Lett.* **2019**, *18*, 4441–4448. [CrossRef]
56. Batada, N.N.; Hurst, L.D.; Tyers, M. Evolutionary and physiological importance of hub proteins. *PLoS Comput. Biol.* **2006**, *2*, e88. [CrossRef]
57. Zhang, D.L.; Qu, L.W.; Ma, L.; Zhou, Y.C.; Wang, G.Z.; Zhao, X.C.; Zhang, C.; Zhang, Y.F.; Wang, M.; Zhang, M.Y.; et al. Genome-wide identification of transcription factors that are critical to non-small cell lung cancer. *Cancer Lett.* **2018**, *434*, 132–143. [CrossRef]
58. Zhang, Y.; Wei, H.; Fan, L.; Fang, M.; He, X.; Lu, B.; Pang, Z. CLEC4s as Potential Therapeutic Targets in Hepatocellular Carcinoma Microenvironment. *Front. Cell Dev. Biol.* **2021**, *9*, 681372. [CrossRef]
59. Zheng, Q.; Min, S.; Zhou, Q. Identification of potential diagnostic and prognostic biomarkers for LUAD based on TCGA and GEO databases. *Biosci. Rep.* **2021**, *41*, BSR20204370. [CrossRef]
60. Mei, L.; Huang, C.; Wang, A.; Zhang, X. Association between ADRB2, IL33, and IL2RB gene polymorphisms and lung cancer risk in a Chinese Han population. *Int. Immunopharmacol.* **2019**, *77*, 105930. [CrossRef]
61. Tian, Z.Q.; Li, Z.H.; Wen, S.W.; Zhang, Y.F.; Li, Y.; Cheng, J.G.; Wang, G.Y. Identification of Commonly Dysregulated Genes in Non-small-cell Lung Cancer by Integrated Analysis of Microarray Data and qRT-PCR Validation. *Lung* **2015**, *193*, 583–592. [CrossRef]
62. Treweek, T.M.; Meehan, S.; Ecroyd, H.; Carver, J.A. Small heat-shock proteins: Important players in regulating cellular proteostasis. *Cell Mol. Life Sci.* **2015**, *72*, 429–451. [CrossRef] [PubMed]
63. Kamradt, M.C.; Lu, M.; Werner, M.E.; Kwan, T.; Chen, F.; Strohecker, A.; Oshita, S.; Wilkinson, J.C.; Yu, C.; Oliver, P.G.; et al. The small heat shock protein alpha B-crystallin is a novel inhibitor of TRAIL-induced apoptosis that suppresses the activation of caspase-3. *J. Biol. Chem.* **2005**, *280*, 11059–11066. [CrossRef] [PubMed]
64. Mao, Y.; Zhang, D.W.; Lin, H.; Xiong, L.; Liu, Y.; Li, Q.D.; Ma, J.; Cao, Q.; Chen, R.J.; Zhu, J.; et al. Alpha B-crystallin is a new prognostic marker for laryngeal squamous cell carcinoma. *J. Exp. Clin. Cancer Res.* **2012**, *31*, 101. [CrossRef] [PubMed]
65. Chan, S.K.; Lui, P.C.; Tan, P.H.; Yamaguchi, R.; Moriya, T.; Yu, A.M.; Shao, M.M.; Hliang, T.; Wong, S.I.; Tse, G.M. Increased alpha-B-crystallin expression in mammary metaplastic carcinomas. *Histopathology* **2011**, *59*, 247–255. [CrossRef] [PubMed]
66. Campbell-Lloyd, A.J.; Mundy, J.; Deva, R.; Lampe, G.; Hawley, C.; Boyle, G.; Griffin, R.; Thompson, C.; Shah, P. Is alpha-B crystallin an independent marker for prognosis in lung cancer? *Heart Lung Circ.* **2013**, *22*, 759–766. [CrossRef]
67. Xu, G.; Dang, C. CMTM5 is downregulated and suppresses tumour growth in hepatocellular carcinoma through regulating PI3K-AKT signalling. *Cancer Cell Int.* **2017**, *17*, 113. [CrossRef]
68. He, J.; Wu, M.; Xiong, L.; Gong, Y.; Yu, R.; Peng, W.; Li, L.; Li, L.; Tian, S.; Wang, Y.; et al. BTB/POZ zinc finger protein ZBTB16 inhibits breast cancer proliferation and metastasis through upregulating ZBTB28 and antagonizing BCL6/ZBTB27. *Clin. Epigenetics* **2020**, *12*, 82. [CrossRef]
69. Jin, Y.; Pan, Y.; Zheng, S.; Liu, Y.; Xu, J.; Peng, Y.; Zhang, Z.; Wang, Y.; Xiong, Y.; Xu, L.; et al. Inactivation of EGLN3 hydroxylase facilitates Erk3 degradation via autophagy and impedes lung cancer growth. *Oncogene* **2022**, *41*, 1752–1766. [CrossRef]
70. Qiu, H.L.; Deng, S.Z.; Li, C.; Tian, Z.N.; Song, X.Q.; Yao, G.D.; Geng, J.S. High expression of KIF14 is associated with poor prognosis in patients with epithelial ovarian cancer. *Eur. Rev. Med. Pharmacol. Sci.* **2017**, *21*, 239–245.
71. Arora, A.; Agarwal, D.; Abdel-Fatah, T.M.; Lu, H.; Croteau, D.L.; Moseley, P.; Aleskandarany, M.A.; Green, A.R.; Ball, G.; Rakha, E.A.; et al. RECQL4 helicase has oncogenic potential in sporadic breast cancers. *J. Pathol.* **2016**, *238*, 495–501. [CrossRef]
72. Alzahrani, F.A.; Ahmed, F.; Sharma, M.; Rehan, M.; Mahfuz, M.; Baeshen, M.N.; Hawsawi, Y.; Almatrafi, A.; Alsagaby, S.A.; Kamal, M.A.; et al. Investigating the pathogenic SNPs in BLM helicase and their biological consequences by computational approach. *Sci. Rep.* **2020**, *10*, 12377. [CrossRef]
73. Huang, D.W.; Sherman, B.T.; Tan, Q.; Collins, J.R.; Alvord, W.G.; Roayaei, J.; Stephens, R.; Baseler, M.W.; Lane, H.C.; Lempicki, R.A. The DAVID Gene Functional Classification Tool: A novel biological module-centric algorithm to functionally analyze large gene lists. *Genome Biol.* **2007**, *8*, R183. [CrossRef]
74. Chen, Y.; Chu, C.W.; Chen, M.I.C.; Cook, A.R. The utility of LASSO-based models for real time forecasts of endemic infectious diseases: A cross country comparison. *J. Biomed. Inform.* **2018**, *81*, 16–30. [CrossRef]
75. Meng, Z.; Wang, M.; Guo, S.; Zhou, Y.; Zheng, M.; Liu, M.; Chen, Y.; Yang, Z.; Zhao, B.; Ying, B. Development and Validation of a LASSO Prediction Model for Better Identification of Ischemic Stroke: A Case-Control Study in China. *Front. Aging Neurosci.* **2021**, *13*, 630437. [CrossRef]