Exploring the Genetic Differences Between Small Cell and Non-small Cell Lung Cancer Using Bioinformatic Approaches: A Preliminary Study

Muhammad Amru Nazri[1], Faris Aizat Ahmad Fajri[1], Muhamad Harith Zulkifli[1], and Fazlin Mohd Fauzi[1,2](✉)

[1] Faculty of Pharmacy, Universiti Teknologi MARA Selangor, Puncak Alam Campus, 42300 Selangor, Bandar Puncak Alam, Malaysia
fazlin5465@uitm.edu.my

[2] Collaborative Drug Discovery Research, Faculty of Pharmacy, Universiti Teknologi MARA Selangor, Puncak Alam Campus, 42300 Selangor, Bandar Puncak Alam, Malaysia

Abstract. Lung cancer can be categorized into two types, which are Non Small Cell Lung Carcinoma (NSCLC) and Small Cell Lung Carcinoma (SCLC). 85% of lung cancer cases are NSCLC, although SCLC is the more aggressive. Mutation of EGFR, ALK and KRAS are characteristics of NSCLC and these findings have led to the discovery of targeted therapy for NSCLC. Targeted therapy for SCLC is lagging as identifying its genetic markers is complicated by the molecular complexity of its pathophysiology. Hence, in this study, genetic differences between SCLC and NSCLC were explored using bioinformatics approaches. The study was divided into two parts where the first involves feature selection and principal component analysis to differentiate the two lung cancer types based on mRNA gene expression. Additionally, top 20 mutated genes for each type were determined using odds ratio (OR). In the second phase, a predictive model was built using outcome of the mRNA gene expression analysis. The results showed that the mRNA expression of 20 identified genes could differentiate the two lung cancer types. This was further corroborated by the predictive model where a sensitivity and specificity of 1.0 was achieved. However, with the small number of data, further analyses are warranted. The OR and protein–protein interaction (PPI) showed that KRAS, NFE2L2, MUC6 and ARHGAP35 genes to be potential biomarkers for NSCLC as well as potential pathway for its progression. This preliminary study shows that bioinformatics approach could aid in understanding SCLC and NSCLC, which could lead to discovery of novel targeted therapy and potential biomarkers.

Keywords: Non small cell carcinoma · Small cell lung carcinoma · principal component analysis · Random Forest · feature selection

A. Sri Wahyuni et al. (Eds.): ICB-Pharma 2022, AHCPS 3, pp. 389–418, 2023.
https://doi.org/10.2991/978-94-6463-050-3_34

1 Introduction

According to the GLOBOCAN 2020 report published by the World Health Organisation (WHO), 18% of cancer related deaths in 2020 were due to lung cancer, which was the highest among all cancer types [1]. Furthermore, lung cancer recorded the second highest number of new cases among all cancer types with over 2 million cases, just behind breast cancer [1]. It is also the leading cause of cancer-related mortality in male worldwide. Smoking is the most common cause of lung cancer, although there are a small percentage of non-smokers being diagnosed with lung cancer.

Lung cancer can be classified into two main types, which are Small Cell Lung Carcinoma (SCLC) and Non Small Cell Lung Carcinoma (NSCLC). Approximately 85% of lung cancer cases are NSCLC, although SCLC is the more aggressive of the two [2]. SCLC is characterised by rapid growth, high tumour burden and early metastasis [3]. It originates in the bronchi while NSCLC originates in lung tissues and composed of larger cells than SCLC when observed under the microscope. Both types are tested in a similar manner involving conventional chest radiography, computed tomography (CT), magnetic resonance (MR) and positron emission tomography (PET) [4]. In majority of the cases, both types of lung cancer are diagnosed in the later stage due to complex diagnostic work up, which could decrease the chance of survival [5].

Mutation of EGFR, KRAS and ALK are characteristics of NSCLC patients, which have led to genetic testing to diagnose NSCLC as well as the discovery and introduction of target therapy as part of the treatment regimen for NSCLC [6]. EGFR (Epidermal Growth Factor Receptor) mutations have been observed in a significant percentage NSCLC case involving non-smokers and female of Asian descent [7]. Administration of Tyrosine Kinase Inhibitors (TKIs) such as gefitinib and erlotinib showed high response rates in patients with EGFR somatic mutations particularly exon 21 L858R, exon 18 G719X and exon 19 deletions [8]. In contrast, mutation of exon 20 T790M mutation is linked to acquired TKI resistance [9]. Mutation of KRAS (Kirsten rat sarcoma virus) is largely seen in adenocarcinomas, in approximately 25% of the case. However, KRAS mutation is commonly seen in smokers but less so in Asians [10]. Fusion between ALK (Anaplastic Lymphoma Kinase) between and EML4 (Echinoderm Microtubule-Associated Protein Like 4) was observed in around 7% NSCLC adenocarcinoma patients, and common in non- or light smokers [11]. Patients with this fusion protein can be treated with ALK inhibitors such as brigatinib [12], ceritinib [13] and crizotinib [14]. In contrast, the development of novel drugs for SCLC is lagging even though TP53 and RB1 mutations have been found in 75–90% of SCLC patients, as the mutations are primarily loss of function [15, 16]. Additionally, several of the therapeutic targets in SCLC such as amplification of MYCs are not ‘druggable’ [17], where it is unlikely to bind to small molecules with high affinity.

Several studies utilising computational approaches have been reported in unveiling the genetics of lung cancer for the purpose of diagnosis as well as discovery of novel therapy. Baoshan et al., [18] identified 16 potential prognostic markers of lung adenocarcinoma (LUAD), a subtype of SCLC through the use of predictive model. The computational model was built using LUAD RNA-Seq data and clinical data from the Cancer Genome Atlas (TCGA) where random survival forest and forward selection were used as machine learning algorithm. External validation in three different data sets showed a C-index values between 0.656 to 0.672. Pathway analysis revealed that eleven of the genes were linked to Nicotine addiction pathway. Lai et al., [19] employed deep neural network and used gene expression and clinical data of NSCLC patients to predict 5-year survival of NSCLC patients. The model showed high accuracy, with an AUC of 0.81 and 75% accuracy. Li et al.,[20] built a risk prediction model for LUAD using gene expression profile obtained from TCGA and Gene Expression Omnibus (GEO). The genes were first evaluated for its prognostic relevance using three algorithms, which were Random Forest, sigFeature and univariate Cox regression [20]. 16 potential genes were identified and used to build a predictive model using the least absolute shrinkage and selection operator (LASSO) algorithm. The model showed good performance in classifying high and low risk patients with C-index of 0.7, 0.689, 0.696, 0.682 and 0.794 in the training set, internal testing set, entire TCGA set, external testing set, and external validation set respectively [20].

As lung cancer is the leading cause of cancer-related deaths worldwide, understanding the full extent of the disease is crucial. Hence, the aim of this study is to explore the genetic differences between SCLC and NSCLC through bioinformatics approaches, which could further aid in the diagnosis of the disease as well as discovery of novel targeted therapy.

2 Materials and method

2.1 Design of Study

This study is divided into two phases (see Fig. 1*)*. In the first phase, Principal Component Analysis (PCA) and feature selection were employed to first mine important genes that differentiates NSCLC and SCLC. Odds ratio was also conducted to analyse significantly mutated genes in both NSCLC and SCLC. In the second phase, predictive models were built using information determined previously to classify SCLC and NSCLC.

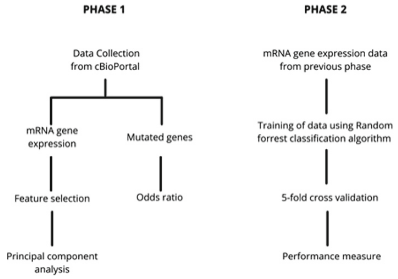
PHASE 1
PHASE 2
Data Collection from cBioPortal
mRNA gene expression
Mutated genes
Feature selection
Odds ratio
Principal component analysis
mRNA gene expression data from previous phase
Training of data using Random forrest classification algorithm
5-fold cross validation
Performance measure

Fig. 1. Design of the study. The study is divided into two parts where the first involves the use of PCA on mRNA gene expression data, and Odds Ratio on mutated genes data. The second phase involves the building of a prediction model based on the result of the mRNA gene expression from the first phase

2.2 Dataset

All the data was obtained from cBioPortal, a library of genomic information developed by Memorial Sloan-Kettering Cancer Centre (MSKCC) [21]. This platform contains more than 40 datasets collected from The Cancer Genome Atlas (TCGA) and hence provides various data and samples on patient case set, cancer study and cancer genomic profiles such as gene expression and mutation [21]. mRNA gene expression and gene mutation data was collected for NSCLC and SCLC cases. The details of the data collected can be found in Tables 1 and 2. The data from SCLC and NSCLC were combined based on mutual variables and any duplicates were removed.

2.3 Feature Selection

Feature selection (FS) is a technique that is used to reduce the number of variables before it is run through a machine learning algorithm. The objective of this feature selection is to reduce large number of variables based on its importance and redundancies. This reduction will be beneficial in reducing computational power and process time during the building of a predictive model. The FS used here was the Tree Based Feature Selection (TBFS) method, which measured the impurity-based feature importance of each variable, hence can be used to remove irrelevant data in the sample [31]. TBFS is based on random forest which comprises of many decision trees. In the tree, there will be nodes that represent variables of the dataset, and this node then will be branching out into several

Table 1. SCLC dataset used which includes their origin, data type and number of data

Dataset	Number of data	Data type(s) available	References
Small Cell Lung Cancer (CLCGP, Nat Genet 2012)	29	Gene mutation	[22]
Small Cell Lung Cancer (Johns Hopkins, Nat Genet 2012)	80	Gene mutation	[16]
Small Cell Lung Cancer (U Cologne, Nature 2015)	120	Gene mutation mRNA Expression	[15]
Small-Cell Lung Cancer (Multi-Institute, Cancer Cell 2017)	20	Gene mutation	[23]
Thoracic PDX (MSK, Provisional)	21	Gene mutation	Data generated in Charles Rudin Lab
Lung Cancer (SMC, Cancer Research 2016)	4	Gene mutation mRNA expression	[24]
Total	274		

Table 2. NSCLC dataset used which includes their origin, data type and number of data

Dataset	Number of data	Data type(s) available	References
Non-Small Cell Lung Cancer (MSK, Cancer Cell 2018)	75	Gene mutation	[25]
Non-Small Cell Lung Cancer (MSKCC, J Clin Oncol 2018)	240	Gene mutation	[26]
Non-Small Cell Lung Cancer (TRACERx, NEJM & Nature 2017)	447	Gene mutation	[27]
Non-Small Cell Lung Cancer (University of Turin, Lung Cancer 2017)	41	Gene mutation	[28]
Non-small cell lung cancer (MSK, Science 2015)	16	Gene mutation	[29]
Pan-Lung Cancer (TCGA, Nat Genet 2016)	1144	Gene Mutation	[30]
Thoracic PDX (MSK, Provisional)	100	Gene mutation CNA	Data generated in Charles Rudin Lab
Lung Cancer (SMC, Cancer Research 2016)	18	Gene mutation mRNA expression	[24]
Total	2081		

other nodes [31]. The node's importance can be calculated by:

$$ni_A = w_A C_A - w_{left(A)} C_{left(A)} - w_{right(A)} C_{right(A)} \tag{1}$$

where ni is the node importance of node A. w_j indicate the weighted sample reaching node A and C_A indicate the impurity value of node A. $left(A)$ and $right(A)$ indicate branches node in the left and the right respectively.

The feature importance can be calculated by:

$$fi_A = \frac{\sum_{A:node\,A\,splits\,on\,feature\,i} ni_A}{\sum_{k \in all\,nodes} ni_k} \tag{2}$$

which indicate the feature importance of variable and ni_A is the importance node of A.

The value of feature importance can be normalized to a value between 0 and 1. This can be achieved by the following formula application:

$$norm\,fi_A = \frac{fi_A}{\sum_{j \in all\,features} fi_A} \tag{3}$$

The average of the feature importance in all the trees in Random Forest is calculated and this value will be the final feature importance. The formula for average feature importance is;

$$RFfi_i = \frac{\sum_{j \in all\ tress} normfi_{iA}}{T} \tag{4}$$

where is $RFfi_i$ the Random Forest feature importance and T is the total tree in the Random Forest. The feature importance value will be assigned to a value between 0 and 1. A relative value of the feature importance is calculated by dividing the feature importance and the highest feature importance in the Random Forest and multiply it by 100. This is performed as there is a large number of variables and calculating the relative value would make it easier to interpret the results. The relative feature importance is calculated as such:

$$Relative\,feature\ importance_i = \frac{RFfi_i}{RFfi_{max}} \times 100 \tag{5}$$

2.4 Principle Component Analysis

Principal component analysis (PCA) is a method of reducing the dimensionality of robust datasets, increasing its interpretability while preserving as much variability and minimizing information loss [32]. This statistical technique creates new uncorrelated variables or principal components, that successively maximize variance. The PCA was performed using the scikit-learn package through the 'PCA' function in Python and plotted using the *ggplot* package in Rstudio [33].

Given a data matrix, $\boldsymbol{X}$, of $n \times p$, where n is the number of rows of instances and p is the number of features, the principal component for each variable, $\boldsymbol{x}$, is calculated as the weighted average of the original variables. The matrix containing the principal components of the data is referred to as matrix Y and can thus be calculated as:

$$\mathrm{Y} = \mathrm{W} \cdot \mathrm{X} \tag{6}$$

where $\boldsymbol{W}$ is a matrix of coefficients that is obtained from the calculation of covariance, eigenvalues and eigenvector. Eigenvalues and eigenvectors are the linear algebra concepts that needed to be computed from the covariance matrix in order to determine the principal components of the data [34]:

$$y_{ij} = W_{li}X_{lj} + W_{2i}X_{2j} + \ldots + W_{pi}X_{pj} \tag{7}$$

The covariance between two variables, $\mathrm{x_i}$ and $\mathrm{x_j}$ can be calculated as:

$$Cov(xi, xj) = \frac{1}{n-1}\sum_{i=1}^{n}\left(xi - \overline{xi}\right)\left(xj - \overline{xj}\right) \tag{8}$$

The eigenvalues and eigenvectors are then determined from the covariance matrix. The eigenvectors (principal components) determine the directions of the new feature space, and the eigenvalues determine their magnitude.

2.5 Odds Ratio

Odds ratio (OR) is defined as a measurement of association between the exposure and the outcome [35]. In this study, the OR is used in the gene mutation data to measure the association of the genes with the types of lung cancer.

An OR of a value of > 1 signify a high association of the exposure and the outcome, while a value of < 1 signify otherwise. An OR value of 1 signify that there is no association between the exposure and the outcome. In this study, the odds ratio was calculated for both SCLC over NSCLC and vice versa. OR is calculated as such:

$$OR = \frac{\left(\frac{nA}{tA}\right)}{\left(\frac{nB}{tB}\right)} \tag{9}$$

where nA is the total amount of occurrence/frequency of a specific gene mutation in NSCLC and tA is the total amount of the occurrence/frequency of all of the genes occurs in NSCLC. While nB is the total amount of occurrence/frequency of a specific gene mutation in SCLC and tB is the total amount occurrence/frequency of all the genes occurs in SCLC.

2.6 Predictive Model

Predictive model is constructed from first learning the functional relationships between variables and outcome in a training set using a classification algorithm. Then, prediction of a possible outcome from variables of new instances can be performed.

2.6.1 Training Set

The training set here contains the mRNA expression of NSCLC and SCLC patients containing 20 genes identified in the previous phase.

2.6.2 Random Forest Classification Algorithm

Random forest is a technique for classification based on an ensemble, or forest, of decision tree [36]. As the name suggest, a prediction will be made using tree-based algorithm method by constructing a forest from the production of several or large number of trees (known as decision trees). The trees were built using training sets consisting of multiple feature or variables for each of the instance in the training set. Then, output results were produced from the variables of the training set of interest. The result was obtained by aggregating all the outputs from different trees. There are two stages in Random Forest which are: (i) random forest creation and (ii) prediction from the random forest classifier created in the first stage [36].

Firstly, the algorithm will build m amount of decision trees. Each of the decision trees will be initiated with a single node where a number of randomly selected samples will serve as the data set. Then, a bootstrap sample of n number of variables of the training data were drawn and selected at random.From the random selected subset, the variable that provides the best split, measured using the Gini index, will split the node into two daughter nodes, specifying possible outcomes [36]. The tree was further split until a maximum size is reached without pruning. Gini index (S) is calculated as follow:

$$\mathrm{Gini(s)} = 1 - \sum_{j}^{2} P. \tag{10}$$

where P is the relative frequency of class j in S. Each time, the split then was divided into two subsets of S_1 and S_2 in which gini (S) data was divided into:

$$Gini_{split}(s) = \frac{n1}{n}\mathrm{gini}(s1) + \frac{n2}{n}gini(s_2). \tag{11}$$

This process will repeat until the tree has reached a specified number of branches and assigned a terminal leaf node. At the end of the tree, class probability will be calculated. In this study, m was set at 100 and n was set as the square root of total number of variables [36]. The outcome was calculated as the mean of class probability from each decision trees. The algorithm was written in Python and using the *scikit-learn* package.

2.6.3 Internal Validation

Fivefold cross validation was employed to internally validate the prediction model. Here, the training set was divided into 5 folds of equal number of data. One fold will serve as the test set while the rest serves as the training set. The process repeats until each fold has served as the test set. At each iteration, the performance of the model will be determined which will be compiled at the end.

2.6.4 Performance Measure

Sensitivity and specificity were used to measure the performance of the model. Sensitivity is used evaluate the model's ability to predict true positive and can be calculated as such:

$$Sensitivity = \frac{TP}{TP + FN} \tag{12}$$

where TP is true positives and FN is false negatives.

Specificity evaluates the model's ability to predict the proportion of actual negative cases and can be calculated as such:

$$Specificity = \frac{TN}{TN + FP} \tag{13}$$

where TN is true negatives and FP is false positives.

2.7 Protein–Protein Interaction Using STRING

Protein–Protein Interaction (PPI) prediction using STRING was employed to see whether two proteins may interact. STRING measures both direct (physical) and indirect (functional) interactions between two proteins, based on experimental data of protein–protein interactions [37].

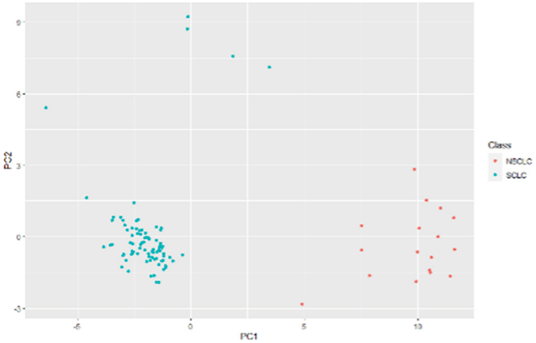

Fig. 2. PCA plot of mRNA expression of NSCLC and SCLC patients

A score is provided for each protein–protein association. The scores represent confidence scores, ranging from 0 to 1, indicating estimated likelihood that the association is biological significant, given the supporting evidence. The supporting evidence is based on seven factors, which are neighbourhood in genome, gene fusions, cooccurrence across genomes, co-expression, experimental/biochemical data, association in curated databases and co-mentioned in PubMed abstracts. These factors are represented by colour coded edges. Based on the seven factors, a combined and final confidence score is computed. A good interaction should not only have a high combined score, but also having more than one factor contributing to the score.

3 Results

3.1 PCA Profile of mRNA Expression of SCLC and NSCLC

The mRNA variables were reduced from 17,793 to 20 using feature selection to reduce overfitting, complexity and the curse of dimensionality. Table 3 shows the summary of the 20 genes used in the PCA. The data were then analysed using PCA, where PC1 and PC2 were plotted (see Fig. 2).

From Fig. 2 and Table 3, several observations can be made. Firstly, there is a clear separation between the two lung cancer subtypes from the PCA plot. This suggests that the subtypes could be differentiated by looking at their mRNA expression in the 20 genes. Secondly, as can be seen in Table 3, majority of the genes are linked to lung cancer, where some can be directly linked to the specific subtypes, or to just lung cancer in general. An example of this is Recoverin (RCVRN), which is found to be a paraneoplastic antigen for lung cancer. The gene is found to be expressed in both SCLC and NSCLC. RCVRN have also been suggested to be a candidate for targeted therapy (42). Another example is Cholecystokinin (CCK), where several studies showed CCK to have an association with lung cancer. Han et al. [38], showed that CCK inhibit the P53 gene transcription, which is involved in cells apoptosis, and also a candidate for targeted drug therapy [38]. The subtypes of the genes, CCK-A and CCK-B were found to be mainly expressed in SCLC [39]. BARHL2 gene has been found to be associated with DNA methylation in Squamous Cell Carcinoma (SCC), a subtype of SCLC where this methylation process is not found in any normal lung cell [40].

Table 3. The details of the 20 genes identified through feature selection to construct the PCA between SCLC and NSCLC. RFI refers to relative feature importance

Gene name	Gene symbol	Description	RFI
brain protein 44-like protein 2	LOC347411	This mRNA shows an association with lung cancer. Study by Pang et al. [41] demonstrated that LOC347411 is among the many targets for miRNA hsa-miR-106b-3p, which exhibit tumour suppression properties	100
Transcription Factor 23	TCF23	No information or study that link TCF23 to any type of lung cancers	86.34
E3 ubiquitin-protein ligase TRIM63	TRIM63	TRIM63 is involved in the de novo inhibition of skeletal muscle protein synthesis under amino acid starvation	70.07
Transition Protein 1	TNP1	TNP1 is involved in the normal spermatogenesis process. The gene is found to be heavily deregulated in patient with smoking habit [42]. Hence, it might have an indirect association to lung cancer	67.92
Chromosome 1 open reading frame 185	C1orf185	This is a protein coding gene. No information or study that link C1orf185to any type of lung cancers	64.12
Glucose-6-phosphatase 2	G6PC2	Downregulation of G6PC2 was observed in SCC although no further information was provided. [43]	62.65
Defensin Beta 134	DEFB134	No information or study that link DEFB134 to any type of lung cancers	58.76
Golgin A6 Family Member B	GOLGA6B	No information or study that link GOLGA6B to any type of lung cancers	58.76
Homeobox protein Nkx-6.2	NKX6-2	NKX6-2 is from the family of homeodomain transcription factors where members of this family have association with various cancer such as lung and thyroid cancer. The family of gene is important in normal development of lung, heart, prostate, thyroid and CNS [44] but no direct associations are found between NKX6-2 and lung cancer type	58.64
BarH-like 2 homeobox	BARHL2	BARHL2 is associated with DNA methylation in SCC where this methylation process is not found in any normal lung cell [40]	58.64

(*continued*)

Table 3. (*continued*)

Gene name	Gene symbol	Description	RFI
Somatostatin receptor type 4	SSTR4	The expression of this gene transcript is found to be high in inflamed tissue of the lung [45]. SSTR4 also is found to be highly expressed in pulmonary carcinoid tumours [40, 46]	54.14
Transmembrane Protein 235	TMEM235	No information or study that link TMEM235 to any type of lung cancers	54.11
Schwannomin Interacting Protein 1	SCHIP1	No information or study that link SCHIP1 with any type of lung cancers. However, SCHIP1 is found to be associated with asthmatic in paediatrics [47]	54.11
Recoverin	RCVRN	This gene is found to be a paraneoplastic antigen for lung cancer. The gene is found expressed in both SCLC and NSCLC [48]	54.11
Oxytocin-neurophysin 1	OXT	Péqueux et al. [49] found that OXT is expressed in $\geq 50\%$ in both lung cancer types, indicating its possibility as a potential new target for targeted therapy	54.11
Mas-related G-protein coupled receptor member E	MRGPRE	This gene is found mainly in the brain region and no information or study that link MRGPRE to any type of lung cancers	54.11
Cholecystokinin	CCK	Several studies show the association of CCK n with lung cancer. Han et al. [38], showed that CCK inhibit P53 gene transcription, which is involve in cells apoptosis, hence the cancer cells can replicate and grow[38]. The subtypes of the genes, CCK-A and CCK-B were found to be mainly expressed in SCLC [39]	54.11
Chromosome 9 open reading frame 170	C9orf170	No information or study that link C9orf170 to type of lung cancers	54.11
Homeobox protein prophet of PIT-1	PROP1	This gene seems to be involved in lung atelectasis [50]. Although atelectasis can be linked with lung cancer pathophysiology [51], further studies are needed to support this	51.15
chromosome 17 open reading frame 102	C17orf102	No information or study that link C17orf102 to any type of lung cancers	50.17

3.2 Odds Ratio Profile of Gene Mutation of SCLC and NSCLC

Table 4 shows the top 20 significantly mutated genes of NSCLC against SCLC and their details. A high odds ratio indicates that the gene mutation is more prominent in NSCLC than SCLC. Several of the genes listed in Table 4 have already shown association

Table 4 Details of the top 20 mutated genes of NSCLC.

Gene name	Gene symbol	Log OR	Description
Nuclear Pore associated Protein 1	NPAP1	1.45	Gene encoded for protein associated with nuclear pore complex. Jiang et al. [52] found that NPAP1 mutation is among the most common mutation in lung cancer, accounting for 17.7% among all the other gene mutation
Leucine Rich Repeat Transmembrane Neuronal 4	LRRTM4	1.27	No information or study that link LRRTM4 to any type of lung cancers
BMP/retinoic acid-inducible neural-specific protein 2	BRINP2	1.21	No direct studies on this gene and lung cancer were found. But, an association link between low expression of the gene in the fibrosis causes by inflammation of the lung such as in COPD and asthma [59], which, can increased risk of getting lung cancer [60]
Protein Phosphatase 1 Regulatory Subunit 3A	PPP1R3A	1.21	The gene transcript are present in some of the human cancer lines including SCLC and NSCLC [61]. Mutation of the genes were also found in both lung cancer types [62]. PP1R3A is found to be downregulated in SCLC, specifically Squamous Cell Carcinoma (SCC) [63]
Transmembrane O-Mannosyltransferase Targeting Cadherins 1	TMTC1	1.20	TMTC1 transports mannosyl residue to hydroxyl group of serine or threonine residue. TMTC1 is found to be involved in lung function [64] and a study showed that a down regulation of TMTC1 is present in NSCLC, specifically LUAD [65]

(*continued*)

Table 4 (*continued*)

Gene name	Gene symbol	Log OR	Description
Myosin Heavy Chain 11	MYH11	1.19	MYH11 as a major contractile protein. Several studies demonstrated that down-regulation of MYH11 gene are associated with lung cancer. [66–68]
Disks large-associated protein 2	DLGAP2	1.18	DLGAP2 is a signalling molecule in postsynaptic neuronal cells. DLGAP2 seems to have an association with Central Nervous System (CNS) disorder such ss Alzheimer and autism. Krishnan et a., [69], showed that DLGAP2 is uniquely found in Asians with NSCLC. In addition, DLGAP seems to be involved in highly in epigenetic modification (DNA methylation), which leads to LUAD [70]
Nuclear factor erythroid 2-related factor 2	NFE2L2	1.16	NFE2L2 regulate antioxidant protein expression, which protect against any oxidative damage induced by inflammation or injury. Several studies have shown the gene to be associated with NSCLC. Goeman et al. [53] demonstrated that Kelch-like ECH-associated protein-1 (KEAP1) and NFE2L2 mutation can be defined as a molecular subtype of LUAD. Furthermore, Jessica et al. [71] showed that KEAP1-NFE2L2 pathway is among the most common mutation pathway in NSCLC. Paul et al. [72] demonstrated that NFE2L2 is a frequently mutated oncogene which can drive NSCLC

(*continued*)

Table 4 (*continued*)

Gene name	Gene symbol	Log OR	Description
Tenascin-X	TNXB	1.15	TNXB is an extracellular-matrix glycoprotein and have anti-adhesive properties, which can cause matrix maturation in wound healing process. Low TNXB expression is found is some cancers including lung and breast, which suggests its possibility as a potential biomarker for NSCLC [73]
Mucin 6 oligomeric mucus/gel-forming	MUC6	1.11	MUC6 provide protective mechanism of epithelial cells. This gene is suggested to be involved in the progression of LUAD [74]. Kishikawa et al. [75] demonstrated the presence of MUC6 on unique clinicopathological subset of Invasive Mucinous Adenocarcinoma (IMA), a subtype of LUAD. Another study showed that the expression of this gene alongside MUC2 are associated with the prognosis of lung cancer and lymph node metastasis [76]
SLIT and NTRK like family member 4	SLITRK4	1.09	No direct association was found between the gene and lung cancer. However, mutation of SLITRK4 have been associated with the metastasis of cancerous cell from colon cancer to the lung [77]
SLIT and NTRK like family member 5	SLITRK5	1.08	SLITRK5 is involved in neurite-modulating activity. This gene is associated with neuronal disorder [78, 79] but a study by Jiang et al. [80] found a mutation of SLITRK5 with TP53 in post-treatment with carboplatin

(*continued*)

Table 4 (*continued*)

Gene name	Gene symbol	Log OR	Description
POTE ankyrin domain family, member G	POTEG	1.08	POTEG is encoded for cancer-testis antigens (CTAs) or cancer-germline genes. Study by Qiu et al. [81] revealed potential association between POTEG and SCLC. This gene has been suggested to have biomarker properties for lung cancer as it is tumour specific [82]
Kell antigen system	KEL	1.07	No study shows direct association between the gene and lung cancer. However, mutation of KEL has been seen in NSCLC [83]
Solute carrier family 12-member 5	SLC12A5	1.07	SLC12A5 is involved in electroneutral potassium-chloride cotransport. SLC12A5 is linked to high proliferation rate, metastasis rate, and G1-S cycle transition in lung Xia et al. [84] discovered the presence of SLC12A5 in LUAD, where it is linked to a poor prognosis
Protocadherin Gamma Subfamily A, 3	PCDHGA3	1.05	PCDHGA3 is involved in immunoglobulin regulation expression. No information or study that link PCDHGA3 with any type of lung cancers
Rho GTPase Activating Protein 35	ARHGAP35	1.05	ARHGAP35 is involved in cell differentiation, cell adhesion as well as cell migration. Héraud et al. [85] demonstrated the association between ARHGAP35 and NSCLC and suggested its potential as an oncogene.. Ouyang et al. [86] showed that the mutation spectrum of ARHGAP35 display tumour suppressing properties

(*continued*)

Table 4 (*continued*)

Gene name	Gene symbol	Log OR	Description
NOD-like receptor family pyrin domain containing 10	NLRP10	1.04	NLRP10 is involved in apoptosis and immune system of the mammal. No information or study that link NLRP10 with any type of lung cancers
Kirsten rat sarcoma	KRAS	1.03	KRAS is involved in cell communication, which include signals for cell growth, proliferation, maturation and differentiation. KRAS is among the most commonly mutated gene found in NSCLC specifically LUAD [87–89]. This mutation has never been seen in SCLC, hence, it can serve as a specific biomarker for NSCLC [87]. This mutation also is found frequently in lung cancer patient with history of smoking compared to non-smoker [87]. Many of the point mutation in KRAS affect codon 12 of the proteins involve in NSCLC [89]
Vav Guanine Nucleotide Exchange Factor 3	VAV3	1.03	No direct association between VAV3 and lung cancer types are found. However, Chen et al., [90] showed that the activation of VAV3 by up-regulation of LINC01234 is important in NSCLC metastasis

to NSCLC such as NPAP1, PPP1R3A, TMTC1, DLGAP2, NFE2L2, TNXB, MUC6, KEL, SLC12A5, KRAS and VRAV3. Several genes listed such as PPP1R3A and POTEG are found in SCLC.

KRAS was among the 20 genes summarised in Table 4 and as previously mentioned, it is found to be significantly mutated in NSCLC. NPAP1, which has the highest odds ratio might be a potential biomarker for NSCLC. Several studies have shown that NPAP1 is the most mutated gene found in NSCLC. Jiang et al. [52] demonstrated that NPAP1 has a prevalence of 17.7% in the tissue samples tested. Another gene worth mentioning here is NFE2L2, which is involved in the Kelch like ECH associated protein-1 and Nuclear Factor Erythroid 2 like-2 (KEAP1/NFE2L2) stress response pathway. KEAP1 is involved in the degradation mediation of NFE2L2, while NFE2L2 provides cytoprotective mechanism through transcription of genes encoded for antioxidant proteins and detoxifying enzyme. The mutation of both KELCH and NFE2L2 are significant in LUAD and SCC respectively [53]. NFE2L2 does not directly cause lung cancer but rather its mutation drive lung cancer progression [54] by promoting the cancerous cell survival and its drug resistance [55]. Lung cancer patients with mutation and co-mutation of EGFR and KEAP1/NFE2L2 were significantly correlated to failure when treated with EGFR Tyrosine Kinase [56]. Hence, NFE2l2 has been deemed to have poor prognosis in lung cancer.

Table 5 shows the top 20 significantly mutated genes of SCLC against NSCLC and their details. Unlike NSCLC, the role of the listed genes to SCLC is not as clear or direct based on literature search. However, one particular gene, HNRNPAB, have been found to be significantly mutated in SCLC. The mRNA of Heterogeneous nuclear ribonucleoprotein A/B (HNRNPAB) is found to be altered in both lung cancer types. The HNRNPAB transcription is found to be most highly expressed in SCLC rather than NSCLC [57]. This is further supported by Ocak et al. [58], which found that HNRNPAB is overexpressed in three samples of SCLC. Thus, HNRNPAB may be a potential biomarker for SCLC.

3.3 Predictive Models Based on mRNA Expression

Table 6 shows the result of the classification model using mRNA data of NSCLC and SCLC. The 20 genes filtered using feature selection previously were used as variables for the model. The model showed both sensitivity and specificity values of 100% in its performance measurement. This indicates that the predictive model is able to predict NSCLC and SCLC based on mRNA gene expression. However, the high performance may be due to the low number of data and hence the results should not be inflated.

3.4 Decision Tree of mRNA Gene Expression

Figure 3 shows a single Random Forest tree of mRNA data. Note that this is only an example of single decision tree, and a random forest contains hundreds of predictive trees (here it is set at 100). Here, the gene HIST1H4A is at the root node (uppermost node) of the decision tree, which is the most important feature for that decision tree. Gini value indicate probability of misclassifying an instance and a lower value indicates a better split. Value indicates the number of data sampled at particular node. From Fig. 3, if the z-score of HIST1H4A is less or equal to 1.22, it will be classified as NSCLC. Here, the gini value is 0 which indicate that this is a terminal node. If HIST1H4A is more than 1.22, it will reach another node where now the expression of SCHIP1 gene will be questioned. The branching of the decision tree will proceed until it has reached the terminal node or the specified number of branches.

Table 5. Top 20 mutated genes of SCLC

Gene name	Gene symbol	Log OR	Description
Interleukin 1 Receptor Associated Kinase 1 Binding Protein 1	IRAK1BP1	1.91	No information or study that link IRAK1BP1 with any type of lung cancers were found
ATP synthase lipid-binding protein	ATP5G1	1.81	No information or study that link ATP5G1 with any type of lung cancers were found
GDF5OS	GDF5OS	1.81	This DNA methylation of this gene is found to be associated in asthmatic patient [91] but no information or study that link GDF5OS with any type of lung cancers
Golgin subfamily A member 8A	GOLGA8A	1.81	This gene is found to be up regulated in LUAD, a subtype of NSCLC [92]
Glycoprotein Ib Platelet Subunit Alpha	GP1BA	1.81	GP1BA is involve in blood clotting process. No information or study that link GP1BA with any type of lung cancers
Heterogeneous nuclear ribonucleoprotein A/B	HNRNPAB	1.81	HNRNPAB is involved in pre-mRNA processing and mRNA metabolism. This gene and its transcript (mRNA) is found in both lung cancer types, but a higher expression is seen in SCLC compared to NSCLC [57, 92]
lung carcinoma-associated 10	LCA10	1.81	No information or study that link LCA10 with any type of lung cancers are found
Membrane Bound O-Acyltransferase Domain Containing 4	MBOAT4	1.81	MBOAT4 is crucial in growth-hormone release. The protein is expressed in various organ including lung [93]. No information or study that link MBOAT4 with any type of lung cancers, but, this gene is found to be over-expressed in metastasized cancer [94]

(*continued*)

Table 5. (*continued*)

Gene name	Gene symbol	Log OR	Description
MicroRNA 146a	MIR146A	1.81	MIR146A regulates the expression of COX-2, which is found to be overexpressed in lung cancer. Deregulation of MIR146A was linked to overexpression of COX-2 in lung cancer specifically NSCLC [95]
Ras-related protein Rab-41	RAB41	1.81	RAB41is involved in autophagy pathway. This gene is found to expressed in LUAD and associated with poor outcome of LUAD [96]
RAS Like Family 10 Member A	RASL10A	1.81	This gene is found to be involved in the EGFR-TKI resistant in NSCLC [97]
RP1-202O8.2 (Clone-based (Vega) gene)	RP1-202O8.3	1.81	No information or study that link RP1-202O8.3 with any type of lung cancers are found
S100 Calcium Binding Protein P	S100P	1.81	S100P is involved in cell cycle progression as well as differentiation. This gene is overexpressed in LUAD, specifically in stage T1 of cancer but not in more advanced stages such as T2 [98]
Sperm acrosome associated 6	SPACA6P	1.81	Downregulation of SPACA6 is induced by XAV939, which promote apoptosis in NSCLC[99]
T Cell Receptor Alpha Variable 10	TRAV10/ Vα24	1.81	TRV10 is found to be associated with invariant Natural Killer T cells (iNKT) receptor [100], which is involved in anti-tumour activity. iNKT have been deemed to be targeted therapy for NSCLC [101]

(*continued*)

Table 5. (*continued*)

Gene name	Gene symbol	Log OR	Description
T Cell Receptor Alpha Variable 8–2	TRAV8-2	1.81	No information or study that link TRAV8-2 with any type of lung cancers are found
AC092653.5 (Clone-based (Vega) gene)	AC092653.5	1.75	No information or study that link AC092653.5 with any type of lung cancers are found
ATP synthase, H+ transporting, mitochondrial F1 complex, epsilon subunit pseudogene 2	ATP5EP2	1.75	This gene is a pseudogene which is found to be one of the top gene found in various cancer line, but the gene expression in lung cancer types such as LUAD and SCC is found to be moderately expressed [102]
DKFZP667F0711	DKFZP667F0711	1.75	No information or study that link DKFZP667F0711 with any type of lung cancers are found
Dickkopf-3	DKK3	1.75	This gene is believed to be a tumour suppressor gene but further study has shown that this gene has other function that may be responsible for LUAD [103]. This gene is also downregulated by DNA methylation, leading to Docetaxel-resistant NSCLC and demethylation of the gene can induce apoptosis in NSCLC [104]

Table 6. Performance of classification model using mRNA expression

TP	FP	TN	FN	Sensitivity	Specificity
15	0	85	0	1	1

3.5 Protein–Protein Interaction (PPI)

3.5.1 *PPI of* mRNA *Gene Expression of Lung Cancer*

Figure 4 shows the potential protein–protein interaction (PPI) between the 20 genes used in the PCA plot previously using STRING. The genes are represented by nodes and if an

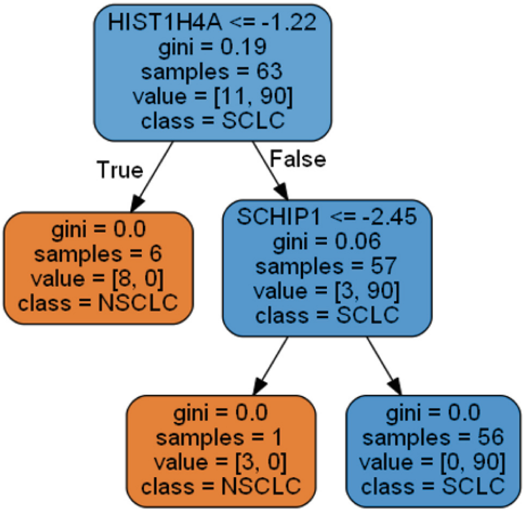

Fig. 3. One of the decision trees of the random forest generated in the study. Here, the most important feature is the gene HIST1H4A

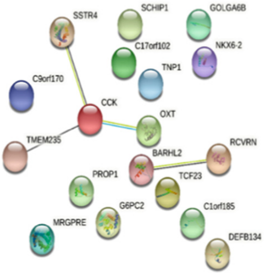

Fig. 4. PPI interaction between the 20 genes identified through feature selection

interaction is predicted between two genes, an edge connects the nodes. Three genes are connected to CCK, which are SSTR4, OXT and TMEM235. Both of SSTR4 and OXT genes are associated with lung cancer, but no literature support could be found linking TMEM235 to lung cancer currently. RCVRN was predicted to interact with BARHL2, where one is associated with SCLC and another with NSCLC respectively. This may indicate potential pathways that both types may share in lung cancer progression.

3.5.2 PPI of Top 20 Mutated Genes of NSCLC

Figure 5 shows the STRING plot of mutated genes of NSCLC. In this plot, KRAS gene may be central in lung NSCLC progression. Three genes are linked to KRAS, which are NFE2L2, MUC6 and ARHGAP35. All three are associated with NSCLC such as NFE2L2, which was explained previously. The presence of MUC6 on unique clinicopathological subset of Invasive Mucinous Adenocarcinoma (IMA), a subtype of

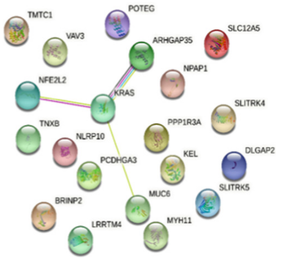

Fig. 5. PPI interaction between the top 20 mutated genes of NSCLC

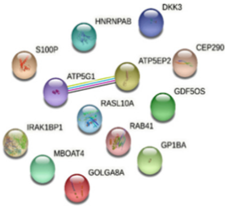

Fig. 6. PPI interaction between the top 20 mutated genes of NSCLC

LUAD have been recently demonstrated [75]. ARHGAP35 has been suggested to be a potential oncogene for NSCL [85, 86]. Thus, these protein–protein interaction plot might indicate that these genes can be a combined biomarker for NSCLC.

3.5.3 PPI of Top 20 Mutated Genes of SCLC

Figure 6 shows the STRING plot of mutated genes of SCLC. The only association is between ATP5G1 and ATP5EP2, where ATPS5EP2 is associated with lung cancer type, albeit with NSCLC.

4 Discussion

Based on the results, several key findings were identified. Firstly, mRNA gene expression of the 20 genes identified through feature selection could be used in differentiating the two lung cancers. This is due to distinct pattern of mRNA profile of SCLC and NSCLC observed in the PCA plot. The clear separation in the PCA based on the mRNA expression was further corroborated by the predictive model where it showed a sensitivity and specificity of 1. However, given the small number of samples, the result of the predictive model should not be inflated. Further studies involving more data and external validation are needed to corroborate this and to improve on the model before it can be an option as a diagnostic tool. The distinction between both is supported by Shriwash et al. [105], where gene expression between NSCLC and SCLC are well characterise. In the study, there were 489 under expressed and 440 over expressed genes in NSCLC, and 525 under expressed 489 over expressed genes in SCLC, and some gene expression was found to overlap between the two lung cancer types [105]. Similar results were observed by Watanabe et al. [106], where PCA analysis of 12 genes showed good characterisation of four histopathological subtypes of lung cancer cells. Hence, this highlights that gene expression may be used as a diagnostic tool as well as reveal potential biomarkers for different subtypes of lung cancer.

Following up on the previous point, odds ratio analysis of the two subtypes showed no overlap where no genes were listed in both Tables 4 and 5. This could point to potential biomarkers for both subtypes. However, from surveying the literature it is evident that the genes identified for NSCLC are much better substantiated compared to SCLC. Additionally, several of the genes identified for SCLC were found to be linked to NSCLC. This highlights the challenge of identifying the genetic landscape of SCLC compared to NSCLC. This sentiment is shared in several studies such as Kim et al. [107] where recurrently mutated genes in SCLC sample such as COL4A2 and COL22A1 were difficult to be associated with the pathogenesis of SCLC, probably due to the molecular complexity of the SCLC pathophysiology. The complexity of SCLC may be further complicated by the presence of passenger genes [108]. Passenger genes are mutated genes that does not contribute to disease progression, which makes identifying specific gene biomarkers for SCLC difficult, thus, leading to a poor diagnosis and prognosis of SCLC [107]. In the case of NSCLC, several genes warrant further validation, in particular NFE2L2, MUC6 and ARHGAP35. All three were connected to KRAS in the PPI and have supporting evidence in their involvement in NSCLC. Individually, each gene could be a potential biomarker for NSCLC but when considered together, these genes may unveil a new pathway in NSCLC that may provide further understanding of the disease.

Lastly, epigenetic may play a role in lung cancer progression. This study does not incorporate epigenome profile, however, mutated genes found in lung cancer, specifically in NSCLC have shown association with epigenetic mechanism. The main epigenetic mechanism found to be most associated with the genes identified in this study is DNA methylation. DNA methylation involves the binding of methyl molecule in a certain region of the nitrogenous bases, mainly cytosine in mammal organism [109]. This epigenetic mechanism can act as ‘on–off’ switch that regulates gene expression and hence may contribute to lung cancer progression. One of the genes connected to epigenetic mechanism is BARHL2, which is associated with NSCLC, specifically SCC. This gene has a 12 CpG methylation and this epigenetic modification are not seen in any normal lung cancer tissue or cell [40]. Another gene with epigenetic mechanism in this study is DKK3, a tumour suppressor gene, where its hypermethylation of DKK3 is found in NSCLC[110]. DNA methylation DKK3 have been shown to lead to its downregulation, rendering it incapable of inducing apoptosis in Docetaxel-resistant NSCLC cell [111]. Another gene worth mentioning that participate in epigenetic modification is DLGAP2, which is associated with KEAP1 mutation, a candidate for tumour suppressor gene in NSCLC. DLGAP2 experience hypomethylation when KEAP1 is mutated in NSCLC, hence impairing cancer suppression function of mutated KEAP1 gene in NSCLC [70]. Most of the epigenetic modification of genes in this study seem to be only associated with NSCLC.

5 Conclusion

This study looked at the genetic profile of NSCLC and SCLC lung cancers using unsupervised and supervised machine learning methods. Several key findings were determined, which were: (i) mRNA expression can be used to differentiate between the two subtypes, (ii) genetic biomarkers for SCLC are more challenging to be identified compared to NSCLC, (iii) the KRAS-NFE2L2-MUC6-ARHGAP35 axis should be further explored as both biomarkers as well as potential pathway in NSCLC progression, and (iv) epigenetic mechanism may play a role in lung cancer progression in particular NSCLC. Future studies warrant the in vitro and in vivo validation of the genes identified here, as well as using more data in the predictive model. One limitation of this study is that a general mutation analysis using odds ratio was performed. A detailed analysis incorporating the type of mutation as well as its location would provide more information. Nevertheless, as this is a preliminary study, the results shown were corroborated by scientific literature and could serve as the foundation for further studies.

Acknowledgements. This research was supported by Universiti Teknologi MARA through the Global Research Reputation Grant (600-RMC 5/3/GRR (003/2020))

Authors' Contributions. MAN, FAAF and FMF are responsible for the design of the work. MAN is responsible for data collection, analysis and interpretation. All authors are responsible for critical revision of the article, with MAN being responsible for drafting the article.

References

1. H. Sung *et al.*, Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries, *CA: A Cancer Journal for Clinicians,* 2021, vol. 71, no. 3, pp. 209–249.
2. R. A. El-Zein, S. Abdel-Rahman, K. J. Santee, R. Yu, and S. Shete, Identification of Small and Non-Small Cell Lung Cancer Markers in Peripheral Blood Using Cytokinesis-Blocked Micronucleus and Spectral Karyotyping Assays, *Cytogenetic and Genome Research,* 2017, vol. 152, no. 3, pp. 122-131, doi: https://doi.org/10.1159/000479809.
3. C. M. Rudin and J. T. Poirier, Shining light on novel targets and therapies, *Nature Reviews Clinical Oncology,* 2017, vol. 14, no. 2, pp. 75-76, doi: https://doi.org/10.1038/nrclinonc.2016.203.
4. W. De Wever, J. Coolen, and J. A. Verschakelen, Imaging techniques in lung cancer, *Breathe,* 2011, vol. 7, no. 4, pp. 338-346, doi: https://doi.org/10.1183/20734735.022110.
5. T. Stokstad, S. Sørhaug, T. Amundsen, and B. H. Grønberg, Medical complexity and time to lung cancer treatment – a three-year retrospective chart review, *BMC Health Services Research,* 2017, vol. 17, no. 1, p. 45, doi: https://doi.org/10.1186/s12913-016-1952-y.
6. F. R. Hirsch, K. Suda, J. Wiens, and P. A. Bunn, Jr., New and emerging targeted treatments in advanced non-small-cell lung cancer, (in eng), *Lancet,* 2016, vol. 388, no. 10048, pp. 1012-24, doi: https://doi.org/10.1016/s0140-6736(16)31473-8.
7. X. Chen *et al.*, Genetic profile of non-small cell lung cancer (NSCLC): A hospital-based survey in Jinhua, (in eng), *Molecular genetics & genomic medicine,* 2020, vol. 8, no. 9, pp. e1398-e1398, doi: https://doi.org/10.1002/mgg3.1398.

8. T. J. Lynch *et al.*, Activating mutations in the epidermal growth factor receptor underlying responsiveness of non-small-cell lung cancer to gefitinib, (in eng), *New England Journal of Medicine,* 2004, vol. 350, no. 21, pp. 2129-39, doi: https://doi.org/10.1056/NEJMoa040938.
9. W. Pao *et al.*, Acquired resistance of lung adenocarcinomas to gefitinib or erlotinib is associated with a second mutation in the EGFR kinase domain, (in eng), *PLoS Medicine,* 2005, vol. 2, no. 3, pp. e73-e73, doi: https://doi.org/10.1371/journal.pmed.0020073.
10. P. J. Roberts, T. E. Stinchcombe, C. J. Der, and M. A. Socinski, Personalized medicine in non-small-cell lung cancer: is KRAS a useful marker in selecting patients for epidermal growth factor receptor-targeted therapy?, (in eng), *Journal of Clinical Oncology,* 2010, vol. 28, no. 31, pp. 4769-77, doi: https://doi.org/10.1200/jco.2009.27.4365.
11. E. L. Kwak *et al.*, Anaplastic lymphoma kinase inhibition in non-small-cell lung cancer, (in eng), *The New England journal of medicine,* 2010, vol. 363, no. 18, pp. 1693-1703, doi: https://doi.org/10.1056/NEJMoa1006448.
12. D. R. Camidge *et al.*, Brigatinib versus Crizotinib in ALK-Positive Non–Small-Cell Lung Cancer, *New England Journal of Medicine,* 2018, vol. 379, no. 21, pp. 2027-2039, doi: https://doi.org/10.1056/NEJMoa1810171.
13. A. T. Shaw *et al.*, Ceritinib in ALK-Rearranged Non–Small-Cell Lung Cancer, *New England Journal of Medicine,* 2014, vol. 370, no. 13, pp. 1189-1197, doi: https://doi.org/10.1056/NEJMoa1311107.
14. L. Friboulet *et al.*, The ALK inhibitor ceritinib overcomes crizotinib resistance in non-small cell lung cancer, (in eng), *Cancer Discov,* 2014, vol. 4, no. 6, pp. 662-673, doi: https://doi.org/10.1158/2159-8290.Cd-13-0846.
15. J. George *et al.*, Comprehensive genomic profiles of small cell lung cancer, *Nature,* 2015, vol. 524, no. 7563, pp. 47-53.
16. C. M. Rudin *et al.*, Comprehensive genomic analysis identifies SOX2 as a frequently amplified gene in small-cell lung cancer, *Nature Genetics,* 2012, vol. 44, no. 10, pp. 1111-1116.
17. H. Taniguchi, T. Sen, and C. M. Rudin, Targeted Therapies and Biomarkers in Small Cell Lung Cancer, (in English), *Frontiers in Oncology,* Mini Review 2020, vol. 10, no. 741, doi: https://doi.org/10.3389/fonc.2020.00741.
18. B. Ma, Y. Geng, F. Meng, G. Yan, and F. Song, Identification of a sixteen-gene prognostic biomarker for lung adenocarcinoma using a machine learning method, *Journal of Cancer,* 2020, vol. 11, no. 5, p. 1288.
19. Y.-H. Lai, W.-N. Chen, T.-C. Hsu, C. Lin, Y. Tsao, and S. Wu, Overall survival prediction of non-small cell lung cancer by integrating microarray and clinical data with deep learning, *Scientific reports,* 2020, vol. 10, no. 1, pp. 1-11.
20. Y. Li, D. Ge, J. Gu, F. Xu, Q. Zhu, and C. Lu, A large cohort study identifying a novel prognosis prediction model for lung adenocarcinoma through machine learning strategies, *BMC Cancer,* 2019, vol. 19, no. 1, p. 886, doi: https://doi.org/10.1186/s12885-019-6101-7.
21. J. Gao *et al.*, “The cBioPortal for cancer genomics and its application in precision oncology,” ed: AACR, 2016.
22. M. Peifer *et al.*, Integrative genome analyses identify key somatic driver mutations of small-cell lung cancer, *Nature Genetics,* 2012, vol. 44, no. 10, pp. 1104-1110.
23. E. E. Gardner *et al.*, Chemosensitive relapse in small cell lung cancer proceeds through an EZH2-SLFN11 axis, *Cancer Cell,* 2017, vol. 31, no. 2, pp. 286-299.
24. S.-W. Um *et al.*, Molecular evolution patterns in metastatic lymph nodes reflect the differential treatment response of advanced primary lung cancer, *Cancer Research,* 2016, vol. 76, no. 22, pp. 6568-6576.
25. M. D. Hellmann *et al.*, Genomic features of response to combination immunotherapy in patients with advanced non-small-cell lung cancer, *Cancer Cell,* 2018, vol. 33, no. 5, pp. 843–852. e4.

26. H. Rizvi *et al.*, Molecular determinants of response to anti–programmed cell death (PD)-1 and anti–programmed death-ligand 1 (PD-L1) blockade in patients with non–small-cell lung cancer profiled with targeted next-generation sequencing, *Journal of Clinical Oncology,* 2018, vol. 36, no. 7, p. 633.
27. M. Jamal-Hanjani *et al.*, Tracking the evolution of non–small-cell lung cancer, *New England Journal of Medicine,* 2017, vol. 376, no. 22, pp. 2109-2121.
28. T. Vavalà *et al.*, Precision medicine in age-specific non-small-cell-lung-cancer patients: Integrating biomolecular results into clinical practice—A new approach to improve personalized translational research, *Lung Cancer,* 2017, vol. 107, pp. 84-90.
29. N. A. Rizvi *et al.*, Mutational landscape determines sensitivity to PD-1 blockade in non–small cell lung cancer, *Science,* 2015, vol. 348, no. 6230, pp. 124-128.
30. J. D. Campbell *et al.*, Distinct patterns of somatic genome alterations in lung adenocarcinomas and squamous cell carcinomas, *Nature Genetics,* 2016, vol. 48, no. 6, pp. 607-616.
31. R. Shaikh. "Feature selection techniques in machine learning with python." https://towardsdatascience.com/feature-selection-techniques-inmachine-learning-with-python-f24e7da3f36e (accessed 20 December, 2021).
32. I. Jolliffe, *Principal component analysis*. New York: Wiley Online Library, 2005.
33. H. Wickham and M. H. Wickham, "The ggplot package," ed: Google Scholar, 2007.
34. Z. Jaadi, A step-by-step explanation of Principal Component Analysis (PCA), *Retrieved June,* 2021, vol. 7, p. 2021.
35. M. Szumilas, Explaining odds ratios, *Journal of the Canadian academy of child and adolescent psychiatry,* 2010, vol. 19, no. 3, p. 227.
36. L. Breiman, Random Forests, *Machine Learning,* 2001, vol. 45, no. 1, pp. 5-32, doi: https://doi.org/10.1023/A:1010933404324.
37. D. Szklarczyk *et al.*, STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets, *Nucleic Acids Research,* 2019, vol. 47, no. D1, pp. D607-D613.
38. Y. Han *et al.*, Cholecystokinin attenuates radiation-induced lung cancer cell apoptosis by modulating p53 gene transcription, *American Journal of Translational Research,* 2017, vol. 9, no. 2, p. 638.
39. T. Sethi, T. Herget, S. V. Wu, J. H. Walsh, and E. Rozengurt, CCKA and CCKB receptors are expressed in small cell lung cancer lines and mediate Ca2+ mobilization and clonal growth, *Cancer Research,* 1993, vol. 53, no. 21, pp. 5208-5213.
40. T. A. Rauch, Z. Wang, X. Wu, K. H. Kernstine, A. D. Riggs, and G. P. Pfeifer, DNA methylation biomarkers for lung cancer, *Tumor Biology,* 2012, vol. 33, no. 2, pp. 287-296.
41. A. L. Y. Pang and O. M. Rennert, Modulation of microRNA expression in human lung cancer cells by the G9a histone methyltransferase inhibitor BIX01294, *Oncology letters,* 2014, vol. 7, no. 6, pp. 1819-1825.
42. H. Amor, A. Zeyad, and M. E. Hammadeh, Tobacco smoking and its impact on the expression level of sperm nuclear protein genes: H2BFWT, TNP1, TNP2, PRM1 and PRM2, *Andrologia,* p. e13964.
43. F. Tian, J. Zhao, X. Fan, and Z. Kang, Weighted gene co-expression network analysis in identification of metastasis-related genes of lung squamous cell carcinoma based on the Cancer Genome Atlas database, (in eng), *Journal of thoracic disease,* 2017, vol. 9, no. 1, pp. 42-53, doi: https://doi.org/10.21037/jtd.2017.01.04.
44. J. K. Pedersen *et al.*, Endodermal expression of Nkx6 genes depends differentially on Pdx1, *Developmental Biology,* 2005, vol. 288, no. 2, pp. 487-501.
45. Z. Varecza *et al.*, Expression of the somatostatin receptor subtype 4 in intact and inflamed pulmonary tissues, *Journal of Histochemistry and Cytochemistry,* 2009, vol. 57, no. 12, pp. 1127-1137.

46. T. Vesterinen *et al.*, Somatostatin receptor expression is associated with metastasis and patient outcome in pulmonary carcinoid tumors, *The Journal of Clinical Endocrinology & Metabolism,* 2019, vol. 104, no. 6, pp. 2083-2093.
47. L. P. Hayden, M. H. Cho, B. A. Raby, T. H. Beaty, E. K. Silverman, and C. P. Hersh, Childhood asthma is associated with COPD and known asthma variants in COPDGene: a genome-wide association study, *Respiratory Research,* 2018, vol. 19, no. 1, pp. 1-11.
48. A. V. Bazhin *et al.*, Recoverin as a paraneoplastic antigen in lung cancer: the occurrence of anti-recoverin autoantibodies in sera and recoverin in tumors, *Lung Cancer,* 2004, vol. 44, no. 2, pp. 193-198.
49. C. Péqueux, C. Breton, M.-T. Hagelstein, V. Geenen, and J.-J. Legros, Oxytocin receptor pattern of expression in primary lung cancer and in normal human lung, *Lung Cancer,* 2005, vol. 50, no. 2, pp. 177-188.
50. I. O. Nasonkin *et al.*, Pituitary hypoplasia and respiratory distress syndrome in Prop1 knockout mice, *Human Molecular Genetics,* 2004, vol. 13, no. 22, pp. 2727-2735.
51. Y. Bulbul *et al.*, Pulmonary atelectasis and survival in advanced non-small cell lung carcinoma, *Upsala Journal of Medical Sciences,* 2010, vol. 115, no. 3, pp. 176-180.
52. J. Jiang *et al.*, Concordance of Genomic Alterations by Next-Generation Sequencing in Tumor Tissue versus Cell-Free DNA in Stage I–IV Non–Small Cell Lung Cancer, *The Journal of Molecular Diagnostics,* 2020, vol. 22, no. 2, pp. 228-235.
53. F. Goeman *et al.*, Mutations in the KEAP1-NFE2L2 pathway define a molecular subset of rapidly progressing lung adenocarcinoma, *Journal of Thoracic Oncology,* 2019, vol. 14, no. 11, pp. 1924-1934.
54. X. Xu *et al.*, NFE2L2/KEAP1 Mutations Correlate with Higher Tumor Mutational Burden Value/PD-L1 Expression and Potentiate Improved Clinical Outcome with Immunotherapy, *The oncologist,* 2020, vol. 25, no. 6, p. e955.
55. P. Huppke *et al.*, Activating de novo mutations in NFE2L2 encoding NRF2 cause a multisystem disorder, *Nature Communications,* 2017, vol. 8, no. 1, pp. 1-10.
56. J. A. Hellyer *et al.*, Impact of KEAP1/NFE2L2/CUL3 mutations on duration of response to EGFR tyrosine kinase inhibitors in EGFR mutated non-small cell lung cancer, *Lung Cancer,* 2019, vol. 134, pp. 42-45.
57. I. Pino *et al.*, Altered patterns of expression of members of the heterogeneous nuclear ribonucleoprotein (hnRNP) family in lung cancer, *Lung Cancer,* 2003, vol. 41, no. 2, pp. 131-143.
58. S. Ocak *et al.*, Discovery of new membrane-associated proteins overexpressed in small-cell lung cancer, *Journal of Thoracic Oncology,* 2014, vol. 9, no. 3, pp. 324-336.
59. S. Bazan-Socha *et al.*, Reticular Basement Membrane Thickness Is Associated with Growth-and Fibrosis-Promoting Airway Transcriptome Profile-Study in Asthma Patients, *International Journal of Molecular Sciences,* 2021, vol. 22, no. 3, p. 998.
60. Y. Sekine, H. Katsura, E. Koh, K. Hiroshima, and T. Fujisawa, Early detection of COPD is important for lung cancer surveillance, *European Respiratory Journal,* 2012, vol. 39, no. 5, pp. 1230-1240.
61. M. Montori-Grau, M. Guitart, C. García-Martínez, A. Orozco, and A. M. Gómez-Foix, Differential pattern of glycogen accumulation after protein phosphatase 1 glycogen-targeting subunit PPP1R6 overexpression, compared to PPP1R3C and PPP1R3A, in skeletal muscle cells, *BMC Biochemistry,* 2011, vol. 12, no. 1, pp. 1-13.
62. T. Kohno, S. Takakura, T. Yamada, A. Okamoto, T. Tanaka, and J. Yokota, Alterations of the PPP1R3 gene in human cancer, *Cancer Research,* 1999, vol. 59, no. 17, pp. 4170-4174.
63. W. Huang *et al.*, Validation and target gene screening of hsa-miR-205 in lung squamous cell carcinoma, *Chinese Medical Journal,* 2014, vol. 127, no. 2, pp. 272-278.

64. T.-C. Yao *et al.*, Genome-wide association study of lung function phenotypes in a founder population, *Journal of Allergy and Clinical Immunology,* 2014, vol. 133, no. 1, pp. 248–255. e10.
65. M. Lu *et al.*, Identification of significant genes as prognostic markers and potential tumor suppressors in lung adenocarcinoma via bioinformatical analysis, *BMC Cancer,* 2021, vol. 21, no. 1, pp. 1-13.
66. R. S. Stearman *et al.*, Analysis of orthologous gene expression between human pulmonary adenocarcinoma and a carcinogen-induced murine model, *The American journal of pathology,* 2005, vol. 167, no. 6, pp. 1763-1775.
67. M. J. Nie *et al.*, Clinical and prognostic significance of MYH11 in lung cancer, *Oncology letters,* 2020, vol. 19, no. 6, pp. 3899-3906.
68. D. G. Beer *et al.*, Gene-expression profiles predict survival of patients with lung adenocarcinoma, *Nature Medicine,* 2002, vol. 8, no. 8, pp. 816-824.
69. V. G. Krishnan *et al.*, Whole-genome sequencing of asian lung cancers: second-hand smoke unlikely to be responsible for higher incidence of lung cancer among Asian never-smokers, *Cancer Research,* 2014, vol. 74, no. 21, pp. 6071-6081.
70. M. Elshaer, A. I. ElManawy, A. Hammad, A. Namani, X. J. Wang, and X. Tang, Integrated data analysis reveals significant associations of KEAP1 mutations with DNA methylation alterations in lung adenocarcinomas, *Aging,* 2020, vol. 12, no. 8, p. 7183.
71. J. A. Hellyer, S. K. Padda, M. Diehn, and H. A. Wakelee, Clinical implications of KEAP1-NFE2L2 mutations in non-small cell lung cancer, *Journal of Thoracic Oncology,* 2020.
72. P. Paik, L. Ahn, M. Ginsberg, D. Mcfarland, L. Doyle, and C. Rudin, P2. 13–44 Targeting NFE2L2 Mutations in Advanced Squamous Cell Lung Cancers with the TORC1/2 Inhibitor TAK-228, *Journal of Thoracic Oncology,* 2018, vol. 13, no. 10, p. S816.
73. S. Liot *et al.*, Loss of Tenascin-X expression during tumor progression: A new pan-cancer marker, *Matrix Biology Plus,* 2020, vol. 6, p. 100021.
74. H. Awaya, Y. Takeshima, M. Yamasaki, and K. Inai, Expression of MUC1, MUC2, MUC5AC, and MUC6 in atypical adenomatous hyperplasia, bronchioloalveolar carcinoma, adenocarcinoma with mixed subtypes, and mucinous bronchioloalveolar carcinoma of the lung, *American Journal of Clinical Pathology,* 2004, vol. 121, no. 5, pp. 644-653.
75. S. Kishikawa *et al.*, Diffuse expression of MUC6 defines a distinct clinicopathological subset of pulmonary invasive mucinous adenocarcinoma, *Modern Pathology,* 2021, vol. 34, no. 4, pp. 786-797.
76. A. Hamamoto *et al.*, Aberrant expression of the gastric mucin MUC6 in human pulmonary adenocarcinoma xenografts, *International Journal of Oncology,* 2005, vol. 26, no. 4, pp. 891-896.
77. L. T. Fang *et al.*, Comprehensive genomic analyses of a metastatic colon cancer to the lung by whole exome sequencing and gene expression analysis, *International Journal of Oncology,* 2014, vol. 44, no. 1, pp. 211-221.
78. S. V. Shmelkov *et al.*, Slitrk5 deficiency impairs corticostriatal circuitry and leads to obsessive-compulsive–like behaviors in mice, *Nature Medicine,* 2010, vol. 16, no. 5, p. 598.
79. C. C. Proenca, K. P. Gao, S. V. Shmelkov, S. Rafii, and F. S. Lee, Slitrks as emerging candidate genes involved in neuropsychiatric disorders, *Trends in Neurosciences,* 2011, vol. 34, no. 3, pp. 143-153.
80. J. Jiang *et al.*, Plasma-based longitudinal mutation monitoring as a potential predictor of disease progression in subjects with adenocarcinoma in advanced non-small cell lung cancer, *BMC Cancer,* 2020, vol. 20, no. 1, pp. 1-9.
81. Z. Qiu *et al.*, A novel mutation panel for predicting etoposide resistance in small-cell lung cancer, *Drug Design, Development and Therapy,* 2019, vol. 13, p. 2021.
82. F. A. M. Maggiolini *et al.*, Evolutionary Dynamics of the POTE Gene Family in Human and Nonhuman Primates, *Genes,* 2020, vol. 11, no. 2, p. 213.

83. E. Mascarenhas *et al.*, Comprehensive genomic profiling of Brazilian non-small cell lung cancer patients (GBOT 0118/LACOG0418), *Thoracic cancer,* 2021, vol. 12, no. 5, pp. 580-587.

84. W. Xia *et al.*, The TWIST1-centered competing endogenous RNA network promotes proliferation, invasion, and migration of lung adenocarcinoma, *Oncogenesis,* 2019, vol. 8, no. 11, pp. 1-15.

85. C. Héraud, M. Pinault, V. Lagrée, and V. Moreau, p190RhoGAPs, the ARHGAP35-and ARHGAP5-encoded proteins, in health and disease, *Cells,* 2019, vol. 8, no. 4, p. 351.

86. H. Ouyang, P. Luong, M. Frödin, and S. H. Hansen, p190A RhoGAP induces CDH1 expression and cooperates with E-cadherin to activate LATS kinases and suppress tumor cell growth, *Oncogene,* 2020, vol. 39, no. 33, pp. 5570-5587.

87. P. M. Westcott and M. D. To, The genetics and biology of KRAS in lung cancer, *Chinese journal of cancer,* 2013, vol. 32, no. 2, p. 63.

88. N. Guibert *et al.*, KRAS mutations in lung adenocarcinoma: molecular and epidemiological characteristics, methods for detection, and therapeutic strategy perspectives, *Current Molecular Medicine,* 2015, vol. 15, no. 5, pp. 418-432.

89. H. Yang, S.-Q. Liang, R. A. Schmid, and R.-W. Peng, New horizons in KRAS-mutant lung cancer: dawn after darkness, *Frontiers in oncology,* 2019, vol. 9, p. 953.

90. Z. Chen *et al.*, Up-regulated LINC01234 promotes non-small-cell lung cancer cell metastasis by activating VAV3 and repressing BTG2 expression, *Journal of Hematology & Oncology,* 2020, vol. 13, no. 1, pp. 1-14.

91. T. T. Hoang *et al.*, Epigenome-wide association study of DNA methylation and adult asthma in the Agricultural Lung Health Study, *European Respiratory Journal,* 2020, vol. 56, no. 3.

92. W. Tian, X. Yang, H. Yang, and B. Zhou, GINS2 Functions as a Key Gene in Lung Adenocarcinoma by WGCNA Co-Expression Network Analysis, *OncoTargets and therapy,* 2020, vol. 13, p. 6735.

93. C. T. Lim, B. Kola, A. Grossman, and M. Korbonits, The expression of ghrelin O-acyltransferase (GOAT) in human tissues, *Endocrine Journal,* 2011, vol. 58, no. 8, pp. 707-710.

94. F. Chen, Y. Zhang, S. Varambally, and C. J. Creighton, Molecular correlates of metastasis by systematic pan-cancer analysis across The Cancer Genome Atlas, *Molecular Cancer Research,* 2019, vol. 17, no. 2, pp. 476-487.

95. A. L. Cornett and C. S. Lutz, Regulation of COX-2 expression by miR-146a in lung cancer cells, *RNA,* 2014, vol. 20, no. 9, pp. 1419-1430.

96. F. Deng, L. Shen, H. Wang, and L. Zhang, Classify multicategory outcome in patients with lung adenocarcinoma using clinical, transcriptomic and clinico-transcriptomic data: machine learning versus multinomial models, *American journal of cancer research,* 2020, vol. 10, no. 12, p. 4624.

97. M. Serizawa, T. Takahashi, N. Yamamoto, and Y. Koh, Genomic aberrations associated with erlotinib resistance in non-small cell lung cancer cells, *Anticancer Research,* 2013, vol. 33, no. 12, pp. 5223-5233.

98. G. Rehbein, A. Simm, H.-S. Hofmann, R.-E. Silber, and B. Bartling, Molecular regulation of S100P in human lung adenocarcinomas, *International Journal of Molecular Medicine,* 2008, vol. 22, no. 1, pp. 69-77.

99. H. Yu, Z. Han, Z. Xu, C. An, L. Xu, and H. Xin, RNA sequencing uncovers the key long non-coding RNAs and potential molecular mechanism contributing to XAV939-mediated inhibition of non-small cell lung cancer, *Oncology letters,* 2019, vol. 17, no. 6, pp. 4994-5004.

100. F. Cortesi, G. Delfanti, G. Casorati, and P. Dellabona, The pathophysiological relevance of the iNKT cell/mononuclear phagocyte crosstalk in tissues, *Frontiers in immunology,* 2018, vol. 9, p. 2375.

101. J. B. Altman, A. D. Benavides, R. Das, and H. Bassiri, Antitumor responses of invariant natural killer T cells, *Journal of immunology research,* 2015, vol. 2015.
102. Y. Li *et al.*, A comprehensive genomic pan-cancer classification using The Cancer Genome Atlas gene expression data, *BMC Genomics,* 2017, vol. 18, no. 1, pp. 1-13.
103. Y. Komori, J. Kano, N. Nakano, S. Sakashita, N. Sakamoto, and M. Noguchi, Dickkopf-related protein 3 promotes cell adhesion and invasion during progression of lung adenocarcinoma, *Pathology International,* 2019, vol. 69, no. 11, pp. 646-654.
104. L. Tao, G. Huang, Y. Chen, and L. Chen, DNA methylation of DKK3 modulates docetaxel chemoresistance in human nonsmall cell lung cancer cell, *Cancer Biotherapy and Radiopharmaceuticals,* 2015, vol. 30, no. 2, pp. 100-106.
105. [105]N. Shriwash, P. Singh, S. Arora, S. M. Ali, S. Ali, and R. Dohare, Identification of differentially expressed genes in small and non-small cell lung cancer based on meta-analysis of mRNA, *Heliyon,* 2019, vol. 5, no. 6, p. e01707.
106. T. Watanabe *et al.*, Comparison of lung cancer cell lines representing four histopathological subtypes with gene expression profiling using quantitative real-time PCR, *Cancer cell international,* 2010, vol. 10, no. 1, pp. 1-12.
107. K.-B. Kim, C. T. Dunn, and K.-S. Park, Recent progress in mapping the emerging landscape of the small-cell lung cancer genome, *Experimental & Molecular Medicine,* 2019, vol. 51, no. 12, pp. 1-13.
108. L. A. Byers and C. M. Rudin, Small cell lung cancer: where do we go from here?, *Cancer,* 2015, vol. 121, no. 5, pp. 664-672.
109. L. D. Moore, T. Le, and G. Fan, DNA methylation and its basic function, *Neuropsychopharmacology,* 2013, vol. 38, no. 1, pp. 23-38.
110. T. Hayashi *et al.*, DNA methylation status of REIC/Dkk-3 gene in human malignancies, *Journal of Cancer Research and Clinical Oncology,* 2012, vol. 138, no. 5, pp. 799-809.
111. L. Hamzehzadeh, M. Caraglia, S. L. Atkin, and A. Sahebkar, Dickkopf homolog 3 (DKK3): A candidate for detection and treatment of cancers?, *Journal of Cellular Physiology,* 2018, vol. 233, no. 6, pp. 4595-4605.