



Article

# Uncovering Prognosis-Related Genes and Pathways by Multi-Omics Analysis in Lung Cancer

Ken Asada <sup>1,2</sup>, Kazuma Kobayashi <sup>1,2</sup>, Samuel Joutard <sup>1,2</sup>, Masashi Tubaki <sup>3</sup>, Satoshi Takahashi <sup>1,2</sup>, Ken Takasawa <sup>1,2</sup>, Masaaki Komatsu <sup>1,2</sup>, Syuzo Kaneko <sup>2</sup>, Jun Sese <sup>2,4</sup> and Ryuji Hamamoto <sup>1,2,\*</sup>

- <sup>1</sup> Cancer Translational Research Team, RIKEN Center for Advanced Intelligence Project, 1-4-1 Nihonbashi, Chuo-ku, Tokyo 103-0027, Japan; ken.asada@riken.jp (K.A.); kazumakob@ncc.go.jp (K.K.); samuel.joutard@kcl.ac.uk (S.J.); sing.monotonyflower@gmail.com (S.T.); ktakazaw@ncc.go.jp (K.T.); maskomat@ncc.go.jp (M.K.)
- <sup>2</sup> Division of Molecular Modification and Cancer Biology, National Cancer Center Research Institute, 5-1-1 Tsukiji, Chuo-ku Tokyo 104-0045, Japan; sykaneko@ncc.go.jp (S.K.); sesejun@humanome.jp (J.S.)
- <sup>3</sup> National Institute of Advanced Industrial Science and Technology, Artificial Intelligence Research Center, 2-3-26, Aomi, Koto-ku, Tokyo 135-0064, Japan; tsubaki.masashi@aist.go.jp
- <sup>4</sup> Humanome Lab, 2-4-10, Tsukiji, Chuo-ku, Tokyo 104-0045, Japan
- \* Correspondence: rhamamot@ncc.go.jp; Tel.: +81-3-3547-5271

Received: 26 January 2020; Accepted: 27 March 2020; Published: 30 March 2020



**Abstract:** Lung cancer is one of the leading causes of death worldwide. Therefore, understanding the factors linked to patient survival is essential. Recently, multi-omics analysis has emerged, allowing for patient groups to be classified according to prognosis and at a more individual level, to support the use of precision medicine. Here, we combined RNA expression and miRNA expression with clinical information, to conduct a multi-omics analysis, using publicly available datasets (the cancer genome atlas (TCGA) focusing on lung adenocarcinoma (LUAD)). We were able to successfully subclass patients according to survival. The classifiers we developed, using inferred labels obtained from patient subtypes showed that a support vector machine (SVM), gave the best classification results, with an accuracy of 0.82 with the test dataset. Using these subtypes, we ranked genes based on RNA expression levels. The top 25 genes were investigated, to elucidate the mechanisms that underlie patient prognosis. Bioinformatics analyses showed that the expression levels of six out of 25 genes (*ERO1B*, *DPY19L1*, *NCAM1*, *RET*, *MARCH1*, and *SLC7A8*) were associated with LUAD patient survival ( $p < 0.05$ ), and pathway analyses indicated that major cancer signaling was altered in the subtypes.

**Keywords:** multi-omics analysis; lung cancer; survival-associated genes

## 1. Introduction

Lung cancer is one of the leading causes of death worldwide, mostly due to a late diagnosis. In fact, an estimated nearly 136,000 patients are expected to die from lung cancer in 2020 in the United States [1]. Even though it only contains 9% of the world's population, Europe accounts for 25% of the global cancer burden, with an estimated 3.9 million new cancer cases and 1.9 million expected cancer deaths in 2018 [2]. Within these cases, the most common cause of cancer death was lung cancer, and 280,000 are expected to die from lung cancer in 2019 [3]. In Asia, and especially in Japan, the number of new cases of lung cancer in 2018 was 118,971 (13.5%), which is the worst number of cases among all cancers. The same was true for the risk of death; 81,820 (20.0%), as indicated by the statistics summarized by the World Health Organization (WHO) [4].

Lung cancer can be classified into two major types: small-cell lung cancer (SCLC), which accounts for approximately 15% of cases and non-small-cell lung cancer (NSCLC), which accounts for approximately 85%. Therefore, NSCLC involves the majority of the lung cancer population, and adenocarcinoma is the most common type of NSCLC. Multiple mutations have been reported to occur in NSCLC, but needless to say, the spectrum of mutations is different between different subtypes [5,6]. Thus, knowing the clinical, pathological, and molecular biological outcomes in diverse aspects is quite important to achieve an improvement in the quality of life of cancer patients.

Recently, in the medical field, deep-learning-driven classification of cancer showed a great success [7]. After that, many images-based machine-learning and deep-learning studies demonstrated their use for cancer prediction, prognosis, or even to assess treatment response in lung cancer [8–10]. However, single-level omics data have limitations, particularly because cancer is a heterogeneous disease, so relying on results obtained from single-level omics data may be risky and misleading; thus, it could affect the understanding of cancer as a whole and possibly negatively affect patients.

One of the proposed approaches to overcome this problem is a multi-omics analysis, an approach that has rapidly emerged in disease-related biology. A new cancer subtyping method, with the integration of multi-omics data, has already been used to reveal molecular subtypes of cancer with TCGA dataset. Multi-omics analysis, using integrated TCGA data of RNA expression, DNA methylation, point mutations, and copy number variation, demonstrated a prediction capability for poor patient outcomes [11]. Multi-omics analysis with a TCGA hepatocellular carcinoma (LIHC) dataset was also performed, using a deep-learning-based and machine-learning-based pipeline to predict patient survival, using RNA expression, DNA methylation, and miRNA expression [12]. The authors implemented an autoencoder to reduce the dimension of multi-omics features as an unsupervised approach, and then, the reduced features were further analyzed via the Cox proportional hazards (Cox-PH) model, to select survival-associated features. A similar approach was applied by using gene expression and copy number variants to classify poor or good subtypes in neuroblastoma [13].

Here, to develop a classifier for the prediction of lung-cancer-patient prognosis and to investigate a patient risk-dependent analysis, we applied a deep-learning- and machine-learning-based pipeline for multi-omics analysis of lung cancer data. We chose data of RNA expression and miRNA expression as input data, so that the result we received could be interpretable, since RNA expression is regulated by miRNA by functional duplexes. Firstly, we developed an SVM that was able to distinguish prognosis-related subtypes from the TCGA LUAD. Secondly, we performed a risk-dependent pathway analysis that can give us relevant information and knowledge about potential mechanisms related to the different subtypes. Lastly, using differentially expressing RNAs in the subtypes, we found novel genes that are associated with patient survival, and we demonstrated that newly identified genes were associated with prognosis.

## 2. Materials and Methods

### 2.1. TCGA Set

We downloaded multi-omics LUAD data from the Genomic Data Commons (GDC) TCGA data portal (<https://portal.gdc.cancer.gov>), using TCGA Assembler 2 (<https://github.com/compgenome365/TCGA-Assembler-2>; [14] with R package (R version 3.5.1). A total of 384 patients with RNA sequencing data (RNA-seq; normalized data) and miRNA sequencing data (miRNA-seq; defined using human reference genome 19 and miRBase version 20 (<http://www.mirbase.org/>)) were assembled into one multi-omics dataset, in the last step of the procedure. Patients' clinical data were manually downloaded from the GDC data portal, and a total of 364 patients were available for the next analysis step. Data were preprocessed by following previous reports to deal with zero values [12]. In the last step, zero values were removed and RNA-seq data, and miRNA-seq data were standardized against patients, followed by reassembling, to make a multi-omics dataset before being fed into an autoencoder.

## 2.2. Clustering to Obtain Inferred Labels from LUAD Multi-Omics Dataset

We basically followed a pipeline and the previously published autoencoder hyper parameter settings [11]. As previously described, we implemented the autoencoder with three hidden layers (500, 100, and 500 nodes) with Keras (<https://keras.io>; version 2.2.4). For the hyper parameter settings, L1 and L2 regulation weights were set at 0.001 and 0.0001, respectively. Learning rate was set at 0.01, with a decay of  $1e-6$ , and epochs were set at 150, with a dropout rate of 0.5. Stochastic gradient descent (SGD) was used as an optimizer. A bottleneck feature space of dimension 100 for each patient was extracted for further analysis.

To obtain clinically associated features from the bottleneck feature space of dimension 100, we built a univariate Cox-PH model, using a survival package in R. A log-rank  $p$ -value of less than 0.05 was considered as significant to select the clinical associated features.

To cluster the survival-associated features and to obtain the inferred labels, we first performed the elbow method [15], to determine the optimal clustering number in a range from one to ten. Based on the result depicted by the elbow method, we performed further analysis, to obtain the optimal number of clusters, using the Silhouette index [16] and Calinski–Harabasz criterion [17]. In the last step, based on the above results, we performed a K-means clustering, using the K, and previously determined and visualized the result with a t-Distributed Stochastic Neighbor Embedding (t-SNE) [18]. We used the scikit-learn library to perform the aforementioned clustering, and the obtained inferred labels were used to draw a Kaplan–Meier plot and then develop the classifiers described in Sections 2.3 and 2.4.

## 2.3. Kaplan–Meier Analysis

Inferred labels obtained at clustering were used for the Kaplan–Meier analysis, to evaluate the prognosis significance of LUAD patients. Survival analysis was performed by using the R survival package, and the survival curve was drawn by using the R survminer package.

## 2.4. ANOVA Feature Ranking of miRNA and RNA Expression to Develop SVM Classifier and LUAD Prognosis-Dependent Classifiers

The multi-omics data used to draw a Kaplan–Meier plot were split into 60% for a training dataset and 40% for a test dataset. Analysis of Variance (ANOVA) method was applied to 60% training dataset to rank miRNA and RNA contributing to the subtypes. ANOVA method with the inferred labels was conducted by using the R limma package [19].

Ranked miRNAs from 5 to 20 and ranked RNAs from 5 to 30 were systematically used to develop SVM. A fixed number of miRNA and RNA were then applied to develop another three classifiers (k-nearest neighbors (KNN), Random Forest (RF), and Logistic Regression (LR)), to compare the accuracy with SVM.

## 2.5. Clinical Characterization

Two distinct populations clustered by K-means algorithm were estimated, using their prognosis with their clinical information by Kaplan–Meier analysis. Clinical data used were obtained from previous reports [20], LUAD data were extracted from TCGA-CDR-Supplementary Table S1, and smoking history indicator was downloaded from the GDC data portal website [21] by selecting the “bcr biotab” option on the “Data Format” list, under the “Files” tab. On the “Cases” tab, we selected the Exposures Environmental Tobacco Smoke Exposure with the project TCGA-LUAD (file name; nationwidechildren.org\_clinical\_patient\_luad.txt).

## 2.6. Somatic Mutation Analysis (SNPs and Small Indels)

LUAD somatic mutation data were downloaded from University of California, Santa Cruz (UCSC) Xena server (<https://xenabrowser.net/datapages/>) and analyzed for mutations occurring in each patient.

### 2.7. Copy Number Analysis

Log<sub>2</sub> transformed LUAD copy number dataset was downloaded from UCSC Xena server (<https://xenabrowser.net/datapages/>). Copy number variant was analyzed with the platform of Affymetrix SNP 6.0 platform and assembled by hg38.

### 2.8. Pathway Analysis Enrichment in the Poor Prognosis Subtype

Gene set enrichment analysis (GSEA), Kyoto Encyclopedia of Genes and Genomes (KEGG) analysis, and Gene Ontology (GO) analysis were performed by using DESeq2, fgsea, and tidyquant packages in R, to analyze enriched pathways in the poor survival subtype.

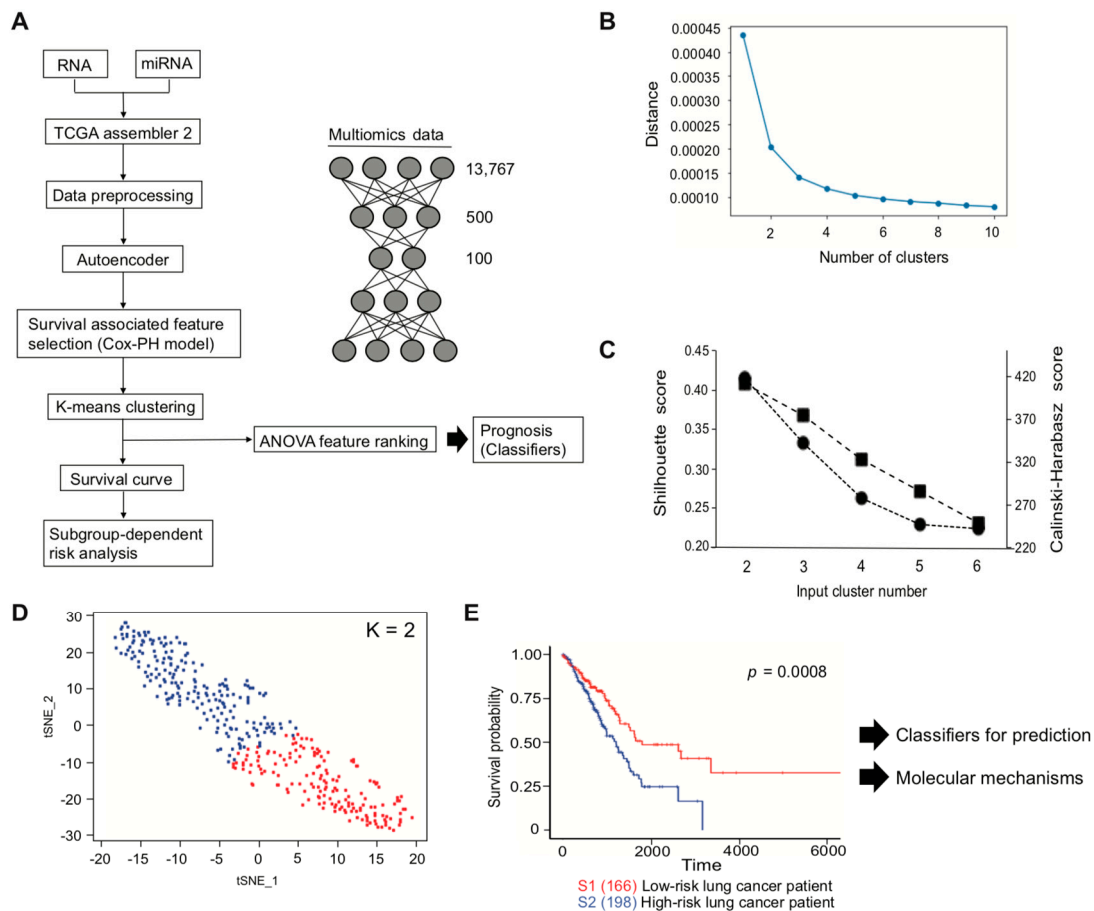
### 2.9. Identification of the Novel Genes Associated with LUAD Patient Survival

Expression analysis was performed between the two subtypes, using the R DESeq2 package. Then, the top 25 statistically significant RNAs between the subtypes obtained from the analysis were used to analyze whether one was associated with patient survival. To draw the Kaplan–Meier plot and to obtain *p*-values from each analysis, OncoLnc web server (<http://www.oncolnc.org>) was used [22]. Then, 25% from the high-expression subgroup (*n* = 123) and 25% from the low-expression subgroup (*n* = 123) were used to compute prognosis. The remaining 50%, forming the intermediate-expression subgroup was excluded from the analysis.

## 3. Results

### 3.1. Subtypes Obtained from Unsupervised Approach

We used TCGA LUAD data for multi-omics analysis, to identify prognosis-related genes. Multi-omics data were generated by TCGA assembler 2, and the data were preprocessed before conducting the omics analysis described in the Materials and Methods section. A total of 13,767 features from RNA-seq and miRNA-seq data were used as an input, which were then encoded to bottleneck feature space of dimension 100 through the autoencoder (Figure 1A). To select the clinically associated features from the bottleneck features, univariate Cox-PH model was performed. In total, 33 out of 100 features showed a statistical significance by log-rank test (*p* < 0.05, Supplementary Table S1). Therefore, the 33 features were further interrogated, to determine whether they could be subcategorized depending on survival outcome. We first roughly estimated the number of clusters through the elbow method (Figure 1B) and then refined the result with more precise analyses, using the Silhouette index and Calinski–Harabasz criterion (Figure 1C; black circle (Silhouette index) and black square (Calinski–Harabasz criterion)). Both of them indicated cluster number two as an optimal number for clustering. Thus, K-means clustering was conducted with *K* = 2, which showed a reasonable clustering result, using t-SNE for visualization (Figure 1D). The inferred labels obtained from K-means clustering were applied to estimate patient survival, and patients were successfully sub-classed into either a poor (high-risk) or a good (low-risk) survival subtype (Figure 1E).



**Figure 1.** Overall workflow for classification of lung-cancer subtypes. (A) Multi-omics analysis pipeline. (B) Clustering result of elbow method. (C) Clustering results of the Silhouette index and Calinski–Harabasz criterion. (D) Clustering result of K-means clustering. Red dot represents S1, and blue dot represents S2 subtype in Figure 1E. (E) Kaplan–Meier plot using patient labels obtained from Figure 1D.

### 3.2. Performance of Four Classifiers Using Inferred Labels

To predict lung-cancer-patient survival, we developed several supervised classifiers for which inputs were obtained from the unsupervised autoencoder. We first considered developing an SVM model because of the previously reported prediction success using multi-omics data from TCGA LIHC [12] and the neuroblastoma project combined from Therapeutically Applicable Research to Generate Effective Treatment (TARGET) with Sequencing Quality Control [13].

For a rigorous evaluation, the multi-omics data we used to draw the Kaplan–Meier plot in Figure 1E were split into a training dataset and a test dataset. With the inferred labels, an ANOVA method was applied to the training dataset, to rank the miRNA and RNA that contribute to the subtypes (Supplementary Table S2, ranked top 20 miRNAs and top 30 RNAs). As we expected and initially speculated on obtaining the interpretable results, top ranked miRNAs were matched to the sequences of top ranked genes analyzed by TargetScanHuman web server [23] ([http://www.targetscan.org/vert\\_72/](http://www.targetscan.org/vert_72/)). We then systematically used high-ranking miRNAs and RNAs, to build the SVM. The developed SVM was evaluated by using the test dataset, to estimate its accuracy. The combination of the top 20 miRNA and top 25 RNA expressions gave the best prediction results and had an accuracy of 0.82 with the test dataset (Table 1). The result of confusion matrix is as shown in Table 2.

**Table 1.** Evaluation of SVM model performance.

Number of Features (miRNA Expression + RNA Expression)	Train Score Accuracy	Test Score Accuracy
10 (5 + 5)	0.61	0.57
20 (10 + 10)	0.81	0.66
30 (15 + 15)	0.89	0.71
40 (20 + 20)	0.94	0.82
45 (20 + 25)	0.95	0.82
50 (20 + 30)	0.97	0.80

**Table 2.** Confusion matrix of SVM.

	Predicted Positive	Predicted Negative
Positive class	62	10
Negative class	17	57

Although the combination of the feature selection by ANOVA, followed by the development of an SVM model, gave the best performance of cancer-patient-survival prediction [13]. In this case, we investigated three additional classifiers, KNN performed with either a hyperparameter of Manhattan or Euclidean distance, RF with either a hyperparameter of Entropy or Gini impurity, and LR with either L1 or L2 regression. The best test score of KNN was 0.76, 0.67 for RF and 0.75 for LR (Table 3). As we expected, these results suggested that SVM is the best classifier if we follow the multi-omics-autoencoder-clinical-associated feature selection by Cox-PH pipeline.

**Table 3.** Evaluation of KNN, RF, and LR performance.

Class	KNN		RF			LR		
	Manhattan	Euclidean	Tree	Entropy	Gini	C	L1	L2
1	0.72	0.70	1	0.54	0.54	1	0.75	0.74
2	0.71	0.68	2	0.64	0.64	5	0.72	0.75
3	0.76	0.73	3	0.64	0.66	10	0.71	0.74
4	0.73	0.74	4	0.64	0.67	50	0.70	0.71
5	0.74	0.75	5	0.66	0.67	100	0.70	0.71
6	0.71	0.72	6	0.66	0.66	500	0.70	0.69
7	0.73	0.75	7	0.67	0.65	1000	0.70	0.70
8	0.71	0.75	8	0.67	0.65			
9	0.73	0.75	9	0.67	0.65			
10	0.73	0.75	10	0.67	0.65			

### 3.3. Insight into the Genes that Are Associated with Patient Prognosis

Identifying the types of biological features is of interest, and thus, we first investigated the clinical data in the different subtypes. Table 4 shows that there were more new tumor events in the high-risk group (41.9%; (83/198)), as compared with the low-risk group (28.9%; (48/166)) (Fisher test  $p = 0.01$ ), and that female patients tended to be in the high-risk group (58.0%; (115/198) versus 50.6%; (84/166) in the low-risk group, Fisher test  $p = 0.15$ ). On the other hand, ages at diagnosis, tumor stages, and smoking history indicator seem to be similar in percentage in the two subtypes.



**Table 4.** Clinical characterization in LUAD low-risk and high-risk subtypes.

Low-Risk ( <i>n</i> = 166)		High-Risk ( <i>n</i> = 198)	
Age at initial pathologic diagnosis (age)	65.5 ± 9.8	Age at initial pathologic diagnosis (age)	65.6 ± 10.2
Tumor stage *	(No.)	Tumor stage *	(No.)
Discrepancy	1	Discrepancy	2
Stage I	2	Stage I	2
Stage IA	47	Stage IA	50
Stage IB	48	Stage IB	53
Stage II	0	Stage II	1
Stage IIA	20	Stage IIA	19
Stage IIB	14	Stage IIB	29
Stage IIIA	21	Stage IIIA	31
Stage IIIB	3	Stage IIIB	4
Stage IV	10	Stage IV	7
Gender	(No.)	Gender	(No.)
Male	82	Male	79
Female	84	Female	119
Vital state	(No.)	Vital state	(No.)
Alive	116	Alive	117
Dead	50	Dead	81
Overall survival time (days)	996.0 ± 967.2	Overall survival time (days)	730.5 ± 560.0
New tumor event	(No.)	New tumor event	(No.)
Yes	48	Yes	83
No	118	No	115
Days to event	588.9 ± 539.0	Days to event	503.7 ± 444.4
Progression-free interval Available	(No.)	Progression-free interval Available	(No.)
	59		88
Progression-free interval time (days)	836.9 ± 874.2	Progression-free interval time (days)	605.8 ± 518.3
Smoking history indicator	(No.)	Smoking history indicator	(No.)
1	22	1	29
2	39	2	45
3	46	3	48
4	54	4	69
5	0	5	4
NA or unknown	5	NA or unknown	3

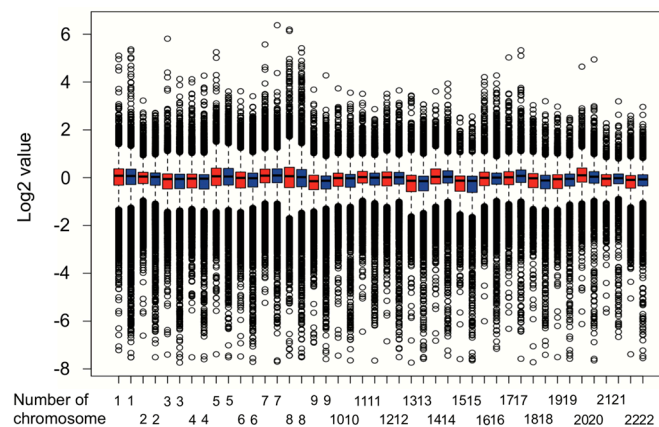
\* American Joint Committee on Cancer (AJCC) pathology states.

Next, we aimed to examine if the subtypes have well-investigated gene mutations, and if so, whether these vary between subtypes. We decided to analyze 18 gene mutations that were found through a comprehensive molecular profiling of TCGA LUAD [6,24]. The result is summarized in Table 5. Our findings indicate that *NF1*, a tumor-suppressor gene that negatively regulates the RAS signaling pathway was more often mutated in the high-risk subtype (14.1% versus 6.0% in the low-risk subtype, Fisher test  $p = 0.01$ ). However, other genes, such as *TP53*, which is frequently mutated in human cancers [25], or *EGFR*, *KRAS*, and *BRAF*, which are mutations that often inform patient therapy [26], were not highly mutated in the high-risk subtype, suggesting that there may be other factors that can distinguish the different subtypes.

**Table 5.** Gene mutations analysis of 18 genes reported as having a statistically significant mutation in the LUAD dataset. Gene names and number of mutations (number of patients) are summarized.

Genes	Low-Risk	High-Risk
<i>TP53</i>	64 (63)	83 (80)
<i>KRAS</i>	42 (41)	40 (38)
<i>KEAP1</i>	26 (26)	27 (26)
<i>STK11</i>	20 (18)	17 (16)
<i>EGFR</i>	17 (13)	17 (13)
<i>NF1</i>	12 (10)	29 (28)
<i>BRAF</i>	12 (10)	8 (8)
<i>SETD2</i>	11 (10)	12 (10)
<i>RBM10</i>	10 (9)	11 (10)
<i>MGA</i>	10 (9)	17 (14)
<i>MET</i>	4 (4)	5 (5)
<i>ARID1A</i>	9 (8)	10 (7)
<i>PIK3CA</i>	8 (8)	9 (8)
<i>SMARCA4</i>	12 (11)	18 (18)
<i>RB1</i>	7 (6)	6 (6)
<i>CDKN2A</i>	7 (5)	6 (6)
<i>U2AF1</i>	0 (0)	0 (0)
<i>RIT1</i>	4 (3)	2 (2)

Therefore, we carried out a copy number variation analysis. Results from the copy number variation analysis were shown as Figure 2. Chromosome 2, 6, 8, 10, 11, 17, 18, 19, 20, and 22 had a different copy numbers ( $p < 0.05$ , Mann–Whitney U-test).



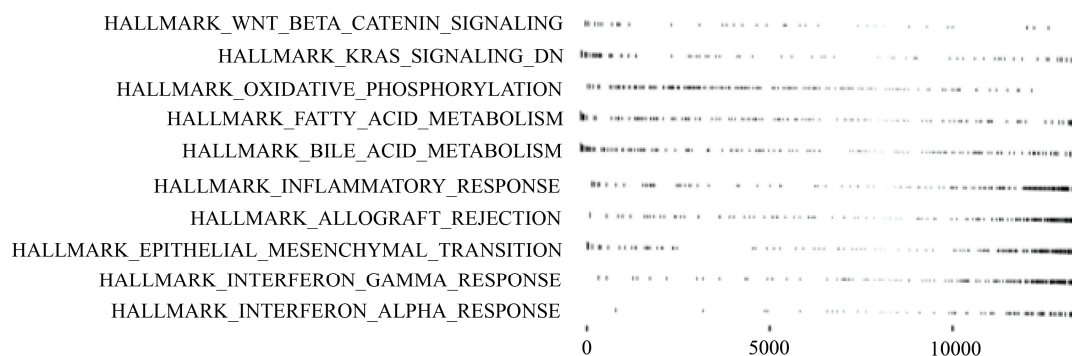
**Figure 2.** Copy number variation analysis in two subtypes. Red box represents low-risk subtype, and blue box represents high-risk subtype.

### 3.4. GSEA, KEGG Pathway Analysis, and GO Analysis

Next, to explore the molecular mechanism that underlies the subtypes, differentially expressing RNA was extracted, using DESeq2 [27]. We performed a GSEA, KEGG pathway analysis, and GO analysis [28–31] to elucidate enriched pathways in the subtypes. The result of GSEA is shown in Figure 3, and the results of the KEGG and GO analyses are summarized in Table 6.

GSEA revealed that Wnt/ $\beta$ -catenin-signaling KRAS-signaling genes downregulated by KRAS activation that could be regulated by NF1 (Table 5), oxidative phosphorylation, and fatty acid metabolism were upregulated. Meanwhile, epithelial-mesenchymal transition (EMT) and inflammatory response, such as interferon gamma response and interferon alpha response, were downregulated in the high-risk subtype, as compared with the low-risk subtype (Figure 3). Notably, above pathways are one of the typical pathways of cancer [32,33].





**Figure 3.** Subtype-specific signaling pathways obtained from GSEA. The left represents pathway names, and the right represents gene ranks.

**Table 6.** Summary of KEGG pathway, miRNA, and GO analysis.

KEGG Pathway	<i>p</i> -Value	Adjusted <i>p</i> -Value
Fatty acid metabolism	$1.64 \times 10^{-3}$	$2.00 \times 10^{-3}$
Oxidative phosphorylation	$1.49 \times 10^{-3}$	$2.00 \times 10^{-3}$
Valine, leucine and isoleucine degradation	$1.61 \times 10^{-3}$	$2.00 \times 10^{-3}$
Arachidonic acid metabolism	$1.69 \times 10^{-3}$	$2.00 \times 10^{-3}$
Pyruvate metabolism	$1.64 \times 10^{-3}$	$2.00 \times 10^{-3}$
KEGG miRNA	<i>p</i> -Value	Adjusted <i>p</i> -Value
miR-501_AAAGGAT	$1.45 \times 10^{-3}$	0.319
miR-26a/miR-26b_TACTTGA	$8.44 \times 10^{-3}$	0.481
miR-507_GTGCAAA	$8.71 \times 10^{-3}$	0.481
miR-33_CAATGCA	$6.05 \times 10^{-3}$	0.481
miR-200b/miR-200c/miR-429_CAGTATT	$2.03 \times 10^{-2}$	0.660
GO Analysis	<i>p</i> -Value	Adjusted <i>p</i> -Value
Spinal cord development	$1.61 \times 10^{-3}$	$5.97 \times 10^{-2}$
Neuromuscular junction development	$1.64 \times 10^{-3}$	$5.97 \times 10^{-2}$
Cytoplasmic translation	$3.08 \times 10^{-3}$	$5.97 \times 10^{-2}$
Positive regulation of calcium ion transport	$2.87 \times 10^{-3}$	$5.97 \times 10^{-2}$
Regulation of antigen receptor mediated signaling pathway	$2.62 \times 10^{-3}$	$5.97 \times 10^{-2}$

KEGG pathway analysis showed that fatty acid metabolism, oxidative phosphorylation, valine, leucine, and isoleucine degradation pathways were significantly different. For the miRNA, miR-501 that activates Wnt/ $\beta$ -catenin signaling in gastric cancer and colorectal cancer [34,35] and tumor suppressor miR-26 that has been reported to regulate the Wnt/ $\beta$ -catenin signaling in prostate cancer and cholangiocarcinoma [36] were significance between subtypes. Intriguingly, miR-26 is also known to contribute TGF- $\beta$ -induced EMT [37] and inflammation response [38], which could be associated with the low-risk subtype. Furthermore, miR-507 targets *KDR* (kinase insert domain receptor or VEGF receptor), and VEGF receptor is regulated by Wnt/ $\beta$ -catenin signaling and *KRAS* pathways [39]. The VEGF receptor is required in response to VEGF-dependent cell survival via EMT in colon carcinoma cell lines [40,41]. Additionally, miR-200 families are well-known miRNAs that regulate Wnt/ $\beta$ -catenin signaling [34] and also directly regulate EMT by targeting transcriptional repressors of *ZEB1* and *ZEB2*, which regulate *CDH1* expression [42,43]. These results indicate that the miRNAs we identified may play an important role in both high- and low-risk subtypes.

The top five GO were summarized in Table 6. Spinal cord development (GO:0021510): the spinal cord primarily conducts sensory and motor nerve impulses in the central nervous systems. The spinal cord development is co-annotated with cell–cell signaling by Wnt (GO:0198738) that is in the 6th place of 1505 co-occurring terms. Neuromuscular junction development (GO:0007528): the neuromuscular

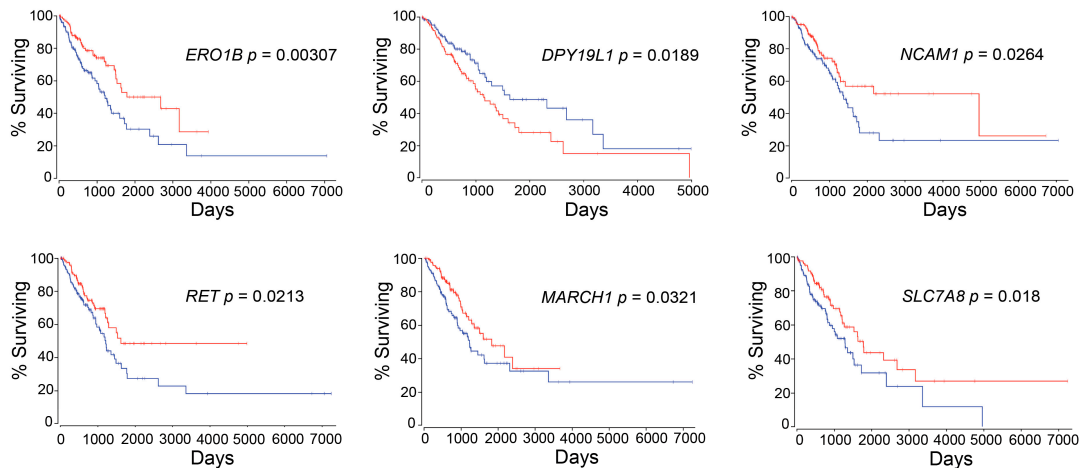
junction is the process to organize a neuromuscular junction at the cellular level. The neuromuscular junction development is co-annotated with blood vessels morphogenesis (GO:0048514) that could be associated with tumor angiogenesis in the 9th place of 1341 co-occurring terms and Wnt signalosome (GO:1990909) is also co-annotated in the 16th place of the same co-occurring terms. Cytoplasmic translation (GO:0002181); cytoplasmic translation is linked with translation (GO:0006412) and the translation is further linked with gene expression (GO:0010467), which is concordant with widely accepted knowledge that gene expression and protein synthesis are upregulated in cancer. Positive regulation of calcium ion transport (GO:0051928) is any process that activates calcium ion efflux. Cytosolic calcium ion concentration is well known to be associated with cellular functions such as gene expression, proliferation, differentiation, migration, metabolism, apoptosis, and angiogenesis [44]. Regulation of antigen receptor-mediated signaling pathway (GO:0050854) is any process that regulates signaling pathways by the cross-linking of an antigen receptor on immune cells. In particular, the relation between neuromuscular junction and cancer development has been previously demonstrated. Yes-associated protein (YAP) and  $\beta$ -catenin regulate synaptic differentiation and the YAP activation induced by the suppression of Hippo pathways promotes liver cancer development [45].

### 3.5. Identification of the Novel Genes Associated with LUAD Patient Survival

We focused on the top 25 differentially expressing RNAs that were extracted by using DESeq2, as shown in Table 7 and the RNA expression levels of the top 25 genes were investigated by using OncoLnc, whether they were associated with LUAD patient survival or not. Interestingly, six out of 25 genes, which are *ERO1B* (endoplasmic reticulum oxidoreductase 1 beta), *DPY19L1* (dpy-19 like C-mannosyltransferase 1), *NCAM1* (neural cell adhesion molecule 1), *RET* (ret proto-oncogene), *MARCH1* (membrane associated ring-CH-type finger 1), and *SLC7A8* (solute carrier family 7 member 8), were identified as survival-associated genes that can affect patient prognosis (Figure 4).

**Table 7.** Top 25 RNAs with statistical significance between the two subtypes.

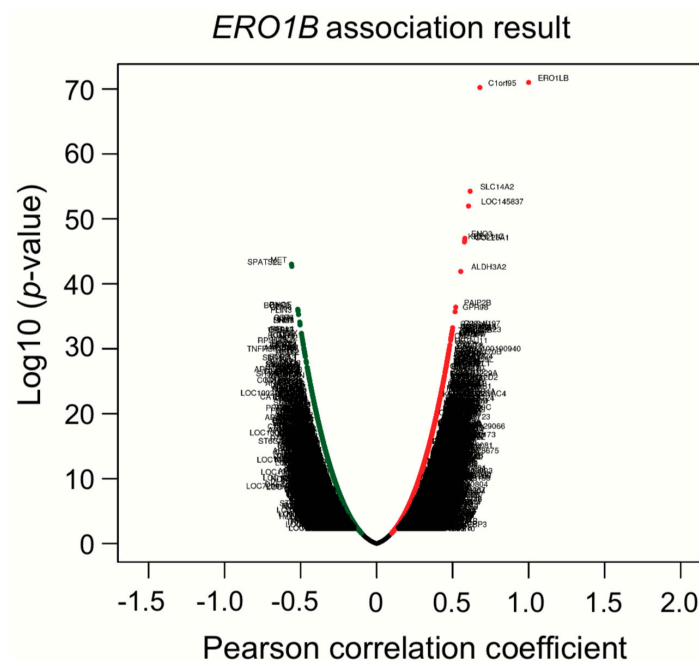
Rank	Gene Name	Log2 Fold Change	<i>p</i> -Value	Adjusted <i>p</i> -Value
1	<i>H19</i>	3.45	$3.01 \times 10^{-40}$	$4.08 \times 10^{-36}$
2	<i>CBR1</i>	1.71	$3.98 \times 10^{-27}$	$2.69 \times 10^{-23}$
3	<i>ENO3</i>	1.93	$1.47 \times 10^{-20}$	$6.64 \times 10^{-17}$
4	<i>POLR3H</i>	0.89	$1.22 \times 10^{-18}$	$4.12 \times 10^{-15}$
5	<i>GREB1</i>	1.49	$2.39 \times 10^{-18}$	$6.47 \times 10^{-15}$
6	<i>ERO1B</i>	1.18	$2.20 \times 10^{-16}$	$4.97 \times 10^{-13}$
7	<i>KCNE4</i>	1.40	$2.81 \times 10^{-15}$	$5.44 \times 10^{-12}$
8	<i>ODC1</i>	1.32	$5.12 \times 10^{-15}$	$8.68 \times 10^{-12}$
9	<i>DPY19L1</i>	-0.72	$2.38 \times 10^{-14}$	$3.58 \times 10^{-11}$
10	<i>WNT4</i>	1.22	$5.19 \times 10^{-14}$	$7.03 \times 10^{-11}$
11	<i>NCAM1</i>	1.37	$1.48 \times 10^{-13}$	$1.82 \times 10^{-10}$
12	<i>RET</i>	1.82	$3.39 \times 10^{-13}$	$3.83 \times 10^{-10}$
13	<i>ESR1</i>	-1.17	$1.93 \times 10^{-12}$	$2.02 \times 10^{-9}$
14	<i>MARCH1</i>	0.94	$2.54 \times 10^{-12}$	$2.45 \times 10^{-9}$
15	<i>SLIT1</i>	1.28	$2.98 \times 10^{-12}$	$2.70 \times 10^{-9}$
16	<i>ZNF710</i>	0.82	$4.35 \times 10^{-12}$	$3.68 \times 10^{-9}$
17	<i>GID8</i>	0.38	$7.25 \times 10^{-12}$	$5.78 \times 10^{-9}$
18	<i>CLU</i>	1.06	$9.22 \times 10^{-12}$	$6.94 \times 10^{-9}$
19	<i>AREG</i>	-1.37	$1.07 \times 10^{-11}$	$7.66 \times 10^{-9}$
20	<i>ALDH3A2</i>	0.72	$1.15 \times 10^{-11}$	$7.80 \times 10^{-9}$
21	<i>MMP11</i>	-1.30	$1.33 \times 10^{-11}$	$8.61 \times 10^{-9}$
22	<i>FAM105A</i>	0.98	$2.33 \times 10^{-11}$	$1.43 \times 10^{-8}$
23	<i>SLC7A8</i>	0.88	$3.19 \times 10^{-11}$	$1.88 \times 10^{-8}$
24	<i>BATF2</i>	-0.68	$4.27 \times 10^{-11}$	$2.41 \times 10^{-8}$
25	<i>ST3GAL3</i>	0.53	$4.99 \times 10^{-11}$	$2.71 \times 10^{-8}$



**Figure 4.** Newly identified survival-associated genes. The red line represents high-expression subtype, and the blue line represents low-expression subtype.

### 3.6. Co-Expression Analysis Reveals *ERO1B*, *ENO3*, and *KCNE4* Genes Are Directed to Upregulate

To further investigate whether the identified six genes are associated with, or potentially show, synergistic effects, a co-expression analysis was conducted with the TCGA LUAD dataset, using LinkedOmics, an interactive web-based tool [46] (<http://www.linkedomics.org/login.php>). Intriguingly, *ERO1B* was co-expressed with *ENO3* that is in the third place of top 25 genes, *KCNE4* that is in the seventh place, and *RET* that is in the 12th place (Figure 5; Tables 7 and 8). Although we do not know the detailed mechanisms behind why these genes are co-expressed, epigenetic regulations or even miRNAs that can regulate multiple target genes, even with one miRNA only, may be involved. To address the abovementioned hypothesis, a miRNA target gene search was performed to find out miRNAs with a sequence that matches to the *ERO1B*, *ENO3*, *KCNE4*, and *RET* transcripts. The TargetScanHuman web server was used for the analysis, and we found that miR-6838 had a predicted to consequential pairing of *ENO3* and *KCNE4*. This suggests that the mechanism of co-expression regulation of *ENO3* and *KCNE4* may be related to the miRNA expression.



**Figure 5.** Co-expression analysis of *ERO1B*. Correlation analysis was performed with Pearson correlation test against *ERO1B* gene.

Table 8. Summary of co-expression genes.

Rank	Target Gene	Pearson Correlation	p-Value	FDR *
1	<i>C1orf95</i>	0.679	$6.11 \times 10^{-71}$	$6.12 \times 10^{-67}$
2	<i>SLC14A2</i>	0.615	$5.63 \times 10^{-55}$	$3.75 \times 10^{-51}$
3	<i>LOC145837</i>	0.605	$1.07 \times 10^{-52}$	$5.36 \times 10^{-59}$
4	<i>ENO3</i>	0.581	$9.89 \times 10^{-48}$	$3.95 \times 10^{-44}$
5	<i>SEC11C</i>	0.578	$2.53 \times 10^{-47}$	$7.47 \times 10^{-44}$
6	<i>KIT</i>	0.578	$2.67 \times 10^{-47}$	$7.47 \times 10^{-44}$
7	<i>COL25A1</i>	0.578	$3.55 \times 10^{-47}$	$8.86 \times 10^{-44}$
8	<i>MET</i>	-0.559	$9.74 \times 10^{-44}$	$2.16 \times 10^{-40}$
9	<i>SPATS2L</i>	-0.558	$2.06 \times 10^{-43}$	$4.11 \times 10^{-40}$
10	<i>ALDH3A2</i>	0.553	$1.28 \times 10^{-42}$	$2.33 \times 10^{-39}$
46	<i>RET</i>	0.482	$2.91 \times 10^{-31}$	$1.24 \times 10^{-28}$
77	<i>KCNE4</i>	0.462	$1.50 \times 10^{-28}$	$3.85 \times 10^{-26}$

\* False discovery rate (Benjamini–Hochberg procedure). Red arrows indicate ones of top 25 RNA shown in Table 7.

#### 4. Discussion

Here, we developed a pipeline, using a TCGA LUAD dataset, with the aim of efficiently identifying genes of interest that are associated with the lung cancer patients survival. Pipeline development started with multi-omics data to implement an autoencoder, followed by clinical associated feature selection by Cox-PH. Selected features were then labeled depending on the result of K-means clustering, which is later demonstrated to be associated with patient survival. The inferred labels, or two subtypes classed by K-means clustering, were applied to plot a Kaplan–Meier survival estimation, to visualize whether the labels were associated with a poor or a good patient survival subtype and used to develop an SVM that can successfully predict patient prognosis.

During autoencoder optimization, batch size, epochs, and activation function varied. Based on our results, a batch size of 1 and epochs of 150, or even between 100 and 150, gave reasonable results, while avoiding overfitting by early stopping [47] and/or Rectified Linear Unit (ReLU) function replacing tanh function at the last layer [48] did not work well in our autoencoder. Clustering analyses applied with clinically associated features demonstrated that  $K = 2$  was the optimal number, and this is concordant with the previous report performing with the 10 TCGA cancer dataset [49].

The multi-omics analysis with TCGA LIHC showed more *TP53* gene mutations in the high-risk subtype (Fisher test  $p = 0.042$ ), but unfortunately, other genes such as *EGFR* were not investigated [12]. In our case, *TP53* was slightly more mutated in the high-risk subtype (0.42%), compared with the low-risk subtype (0.39%), but not significance (Fisher test  $p = 0.633$ ). Whole-exome sequencing data of LUAD were analyzed independently in the oncogene-positive subset (*KRAS*, *EGFR*, *ERBB2*, *BRAF*, *MET*, *ALK*, *RET*, *ROS*, *HRAS*, *NRAS*, and *MAP2K1* driver mutations) and the oncogene-negative subset [24]. The authors found that *TP53* and *NF1* co-mutations were enriched in the oncogene-negative subset. Additionally, RNA profiling provided new subtypes that the proximal-inflammatory subtype (formerly squamoid) was co-mutated with *TP53* and *NF1* [6,24]. In our analysis, we found *NF1* mutations were more enriched in the high-risk subtype, suggesting that the high-risk subtype we identified might correspond to the subset that has *TP53* and *NF1* co-mutations in [24].

*ERO1B* was first reported as an endoplasmic reticulum disulfide oxidase [50]. Later, additional biological functions, such as insulin biogenesis and glucose homeostasis, were demonstrated [51]. In relation to lung cancer, *ERO1B* has been recently identified as a gene that, together with an additional three genes identified using TCGA LUAD dataset, is able to predict patient prognosis [52] and has been suggested to be a biomarker for pancreatic cancer [53,54]. *DPY19L1* was firstly identified as an unclassified gene from human brain cDNA libraries in 1998 [55]. Still, its function remains unknown, and no evidence has been reported so far on the link between *DPY19L1* and cancer prognosis. Therefore, to the best of our knowledge, this is the first report to reveal the association between

*DPY19L1* expression and the prognosis in lung cancer patients. *NCAM1* or *CD56* is a member of the immunoglobulin superfamily involved in cell–cell interaction and cell–matrix interactions during the development. Additionally, it plays a fundamental role in processes such as cell migration and cell survival, in specific phenotypes of neural cells [56]. *NCAM1* may play an important role in EMT not only in intrahepatic cholangiocarcinoma but also in lung cancer via miR-200 (Table 6 and [57]). Recently, antibody-based anticancer treatment was analyzed with the expression levels of *NCAM1*. The phase 1/2 study is ongoing, since *NCAM1* is expressed on several malignancies, including SCLC [58–60], or could be available to predict prognosis in adult acute lymphoblastic leukemia patient [61]. *RET* was identified in 1985. *RET* is a receptor-type tyrosine kinase with multiple domains. *RET* was first discovered in papillary thyroid carcinoma, and later in sporadic tumors, neurodegenerative diseases, and Hirschsprung’s disease [62]. *RET* can be found in the rearrangement of genes generating *RET* fusion proteins in many cancers, including lung cancer, and thus an inhibitor was recently approved by the FDA for cancer therapy [63]. It is important to note that, not only genetic factors, but also epigenetic factors, affect *RET* expression that influences the probability of patient survival [64]. It suggests that multi-omics analysis, including epigenetic data, could improve availability of output, in terms of precision medicine or personalized medicine, as we recently reported [65]. The E3 ubiquitin ligase *MARCH1* plays an important role in immunology [66], although only a few publications have focused on *MARCH1* in the context of cancer [67,68]. Therefore, further studies in this area are required and could have the potential to contribute to the field of cancer research, and more particularly lung cancer. *SLC7A8* or *LAT-2* is an L-type amino acid transporter-2 protein that binds and regulates mechanistic target of rapamycin kinase (mTOR) activation in pancreatic cancer [69]. L-type amino acid transporters are known to be novel targets for cancer therapy [70,71]. However, as is the case for *DPY19L1* and *MARCH1*, no publications have demonstrated the link between lung cancers.

We identified six genes with expression levels that were associated with patient survival, using the autoencoder, followed by bioinformatics analysis. The practice guidelines in oncology illustrate a strategy of patient treatment based on the result of gene mutations, such as *EGFR*, *ALK*, *ROS1*, and *PD-L1* [72], but not considering RNA or miRNA expression levels. It might be of great help to estimate survival outcome and to make treatment strategy for patients if several RNA-expression levels, such as *ERO1B*, *DPY19L1*, *NCAM1*, *RET*, *MARCH1*, and/or *SLC7A8*, are also examined at the time when patients are diagnosed.

To elucidate whether six genes were only associated with LUAD patient prognosis or whether these genes were key regulators of other types of NSCLC prognosis, survival analysis against TCGA lung squamous cell carcinoma (LUSC) was performed. The *p*-values for high expression and low expression of genes of interest were from 0.106 to 0.674, suggesting that the genes we identified were LUAD-specific survival-related genes. This result gave us confidence that the multi-omics analysis we developed truly identified input-data-specific survival-associated features. In other words, if we would like to identify genes of interest that are associated with LUSC patient survival, we need to use a LUSC dataset as an input.

Co-expression analysis showed that *ERO1B*, *ENO3*, *RET*, and *KCNE4* were co-upregulated. Later, we showed that *ENO3* and *KCNE4* have a target sequence for miR-6838. The functional role of miR-6838 has been recently investigated, showing that miR-6838 regulates EMT in triple-negative breast cancer by inhibiting the Wnt pathway [73]. KEGG miRNA target analysis in Table 6 indicated that miR-26 families were enriched in the high-risk subtype. Based on the TargetScanHuman analysis, miR-26 is one of four miRNAs that was predicted to bind to the *ERO1B* transcript and suppress gene expression. As we mentioned in Section 3.4, KEGG miRNA analysis revealed that miR-501, miR-26, miR-507, miR-33, and miR-200/miR-429 were involved in lung-cancer subtypes. The miRNAs we identified have been previously reported as regulating Wnt/ $\beta$ -catenin signaling and/or contributing EMT signaling. Taken together, not only KEGG analysis, but also co-expression analysis, gave us insight into the molecular mechanisms that underlie patient prognosis.



A limitation of this study is the difficulty with preparing the validation dataset. The SVM model we developed uses 20 miRNA and 25 RNA expressions. Thus, we need a validation dataset that includes miRNA expression, RNA expression, and clinical information. There are datasets available that include miRNA expression (GSE63805) and RNA expression (GSE63459), together with clinical information. However, some of the miRNA expression and RNA expression for the top 20 miRNA and top 25 RNAs used to develop the SVM model were missing, and therefore we were not able to evaluate the SVM with the abovementioned publicly available dataset. This constitutes a technical limitation of the study, since it makes it difficult to assess the robustness of the developed classifier. Therefore, we decided to use the TCGA dataset again, for the validation. All data (364 patients) were randomly split into 75% and 25%, and the 25% of patient data were used for validation. Result of the accuracy score of the developed SVM model was 0.92.

The second limitation of this study is the fact the frequency of certain gene mutations can vary depending on the patients' race. For example, EGFR mutation is more often found in Asian American patients than Caucasian or African American patients [74]. Therefore, the SVM model we developed may not be able to distinguish a high-risk subtype from a low-risk subtype if the model is applied to a different distributed dataset such as on containing an Asian population. In that case, the SVM model will need to be redeveloped.

## 5. Conclusions

Lung cancer is one of the leading causes of death worldwide. Understanding the factors that are linked with patient prognosis is essential to enhance the effectiveness of patient therapy. Recently, multi-omics analysis has emerged, allowing to classify groups of patients based on prognosis and at a more individual scale, in the context of precision medicine. Here, we only combined RNA expression, miRNA expression, and clinical information, to develop an SVM to predict patient survival in lung cancer. This enables us to significantly reduce the input omics data size, since DNA methylation data are by far bigger than other omics data; it also enables us to become interpretable.

Using bioinformatics, we established that (1) the *NF1* gene was more mutated, and (2) Wnt/ $\beta$ -catenin, as well as KRAS signaling pathways, can occur in the high-risk subtype. On the other hand, (3) pathways of KRAS, Wnt/ $\beta$ -catenin, and/or TGF- $\beta$  derived EMT pathways, together with the combination of miRNA expression, could be the ones associated with low-risk subtype.

**Supplementary Materials:** The following are available online at <http://www.mdpi.com/2218-273X/10/4/524/s1>. Table S1: Survival associated 33 features extracted by log-rank test. Table S2: ANOVA ranked miRNA and genes (top 20 and 25, respectively) used for SVM. Computer code: Keras code we implemented an autoencoder algorithm.

**Author Contributions:** K.A. performed and analyzed the experiments and wrote the manuscript; K.A., K.K., S.J., M.T., S.T., K.T., M.K., S.K., J.S., and R.H. discussed data; R.H. supervised the experiments and edited the manuscript. All authors read and approved the final manuscript.

**Funding:** This work was supported by JST CREST (Grant Number JPMJCR1689), JSPS Grant-in-Aid for Scientific Research on Innovative Areas (Grant Number JP18H04908), and the Public/Private R&D Investment Strategic Expansion Program (PRISM), Cabinet Office, Japan.

**Acknowledgments:** We would like to thank Daisuke Okanohara and Kouya Shiraishi for their helpful feedback. We would also like to thank all members of Hamamoto laboratory for their kind help.

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

## References

1. Siegel, R.L.; Miller, K.D.; Jemal, A. Cancer statistics, 2020. *CA Cancer J. Clin.* **2020**, *70*, 7–30. [[CrossRef](#)] [[PubMed](#)]
2. Ferlay, J.; Colombet, M.; Soerjomataram, I.; Dyba, T.; Randi, G.; Bettio, M.; Gavin, A.; Visser, O.; Bray, F. Cancer incidence and mortality patterns in Europe: Estimates for 40 countries and 25 major cancers in 2018. *Eur. J. Cancer* **2018**, *103*, 356–387. [[CrossRef](#)] [[PubMed](#)]



3. Malvezzi, M.; Carioli, G.; Bertuccio, P.; Boffetta, P.; Levi, F.; La Vecchia, C.; Negri, E. European cancer mortality predictions for the year 2019 with focus on breast cancer. *Ann. Oncol.* **2019**, *30*, 781–787. [[CrossRef](#)] [[PubMed](#)]
4. WHO. The Global Cancer Observatory. Available online: <https://gco.iarc.fr/today/data/factsheets/populations/392-japan-fact-sheets.pdf> (accessed on 28 March 2020).
5. Mok, T.S. Personalized medicine in lung cancer: What we need to know. *Nat. Rev. Clin. Oncol.* **2011**, *8*, 661–668. [[CrossRef](#)]
6. Inamura, K. Lung Cancer: Understanding Its Molecular Pathology and the 2015 WHO Classification. *Front. Oncol.* **2017**, *7*, 193. [[CrossRef](#)] [[PubMed](#)]
7. Esteva, A.; Kuprel, B.; Novoa, R.A.; Ko, J.; Swetter, S.M.; Blau, H.M.; Thrun, S. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **2017**, *542*, 115–118. [[CrossRef](#)]
8. Ardila, D.; Kiraly, A.P.; Bharadwaj, S.; Choi, B.; Reicher, J.J.; Peng, L.; Tse, D.; Etemadi, M.; Ye, W.; Corrado, G.; et al. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nat. Med.* **2019**, *25*, 954–961. [[CrossRef](#)]
9. Xu, Y.; Hosny, A.; Zeleznik, R.; Parmar, C.; Coroller, T.; Franco, I.; Mak, R.H.; Aerts, H. Deep Learning Predicts Lung Cancer Treatment Response from Serial Medical Imaging. *Clin. Cancer Res.* **2019**, *25*, 3266–3275. [[CrossRef](#)]
10. Hosny, A.; Parmar, C.; Coroller, T.P.; Grossmann, P.; Zeleznik, R.; Kumar, A.; Bussink, J.; Gillies, R.J.; Mak, R.H.; Aerts, H. Deep learning for lung cancer prognostication: A retrospective multi-cohort radiomics study. *PLoS Med.* **2018**, *15*, e1002711. [[CrossRef](#)]
11. Ramazzotti, D.; Lal, A.; Wang, B.; Batzoglou, S.; Sidow, A. Multi-omic tumor data reveal diversity of molecular mechanisms that correlate with survival. *Nat. Commun.* **2018**, *9*, 4453. [[CrossRef](#)]
12. Chaudhary, K.; Poirion, O.B.; Lu, L.; Garmire, L.X. Deep Learning-Based Multi-Omics Integration Robustly Predicts Survival in Liver Cancer. *Clin. Cancer Res.* **2018**, *24*, 1248–1259. [[CrossRef](#)] [[PubMed](#)]
13. Zhang, L.; Lv, C.; Jin, Y.; Cheng, G.; Fu, Y.; Yuan, D.; Tao, Y.; Guo, Y.; Ni, X.; Shi, T. Deep Learning-Based Multi-Omics Data Integration Reveals Two Prognostic Subtypes in High-Risk Neuroblastoma. *Front. Genet.* **2018**, *9*, 477. [[CrossRef](#)] [[PubMed](#)]
14. Wei, L.; Jin, Z.; Yang, S.; Xu, Y.; Zhu, Y.; Ji, Y. TCGA-assembler 2: Software pipeline for retrieval and processing of TCGA/CPTAC data. *Bioinformatics* **2018**, *34*, 1615–1617. [[CrossRef](#)] [[PubMed](#)]
15. Yuan, C.; Yang, H. Research on K-Value Selection Method of K-Means Clustering Algorithm. *J.—Multidiscip. Sci. J.* **2019**, *2*, 16. [[CrossRef](#)]
16. Rousseeuw, P.J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **1987**, *20*, 13. [[CrossRef](#)]
17. Calinski, T.; Harabasz, J. A dendrite method for cluster analysis. *Commun. Stat. - Theory Methods* **1974**, *3*, 27. [[CrossRef](#)]
18. Laurens van der Maaten, G.H. Visualizing Data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.
19. Ritchie, M.E.; Phipson, B.; Wu, D.; Hu, Y.; Law, C.W.; Shi, W.; Smyth, G.K. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **2015**, *43*, e47. [[CrossRef](#)]
20. Liu, J.; Lichtenberg, T.; Hoadley, K.A.; Poisson, L.M.; Lazar, A.J.; Cherniack, A.D.; Kovatich, A.J.; Benz, C.C.; Levine, D.A.; Lee, A.V.; et al. An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics. *Cell* **2018**, *173*, 400–416 e11. [[CrossRef](#)]
21. GDC data portal website. Available online: <https://portal.gdc.cancer.gov/repository> (accessed on 28 March 2020).
22. Anaya, J. OncoLnc: Linking TCGA survival data to mRNAs, miRNAs, and lncRNAs. *Peer J. Comput. Sci.* **2016**, *2*, 11. [[CrossRef](#)]
23. Agarwal, V.; Bell, G.W.; Nam, J.W.; Bartel, D.P. Predicting effective microRNA target sites in mammalian mRNAs. *Elife* **2015**, *4*. [[CrossRef](#)] [[PubMed](#)]
24. Cancer Genome Atlas Research, N. Comprehensive molecular profiling of lung adenocarcinoma. *Nature* **2014**, *511*, 543–550. [[CrossRef](#)] [[PubMed](#)]
25. Petitjean, A.; Achatz, M.I.; Borresen-Dale, A.L.; Hainaut, P.; Olivier, M. TP53 mutations in human cancers: Functional selection and impact on cancer prognosis and outcomes. *Oncogene* **2007**, *26*, 2157–2165. [[CrossRef](#)]
26. Shaw, A.T.; Engelman, J.A. ALK in lung cancer: Past, present, and future. *J. Clin. Oncol.* **2013**, *31*, 1105–1111. [[CrossRef](#)] [[PubMed](#)]

27. Love, M.I.; Huber, W.; Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **2014**, *15*, 550. [[CrossRef](#)] [[PubMed](#)]
28. Ashburner, M.; Ball, C.A.; Blake, J.A.; Botstein, D.; Butler, H.; Cherry, J.M.; Davis, A.P.; Dolinski, K.; Dwight, S.S.; Eppig, J.T.; et al. Gene ontology: Tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **2000**, *25*, 25–29. [[CrossRef](#)] [[PubMed](#)]
29. The Gene Ontology Consortium. The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Res.* **2019**, *47*, D330–D338.
30. Sergushichev, A.A. An algorithm for fast preranked gene set enrichment analysis using cumulative statistic calculation. *bioRxiv* **2016**.
31. Kanehisa, M.; Goto, S. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **2000**, *28*, 27–30. [[CrossRef](#)] [[PubMed](#)]
32. Hanahan, D.; Weinberg, R.A. The hallmarks of cancer. *Cell* **2000**, *100*, 57–70. [[CrossRef](#)]
33. Hanahan, D.; Weinberg, R.A. Hallmarks of cancer: The next generation. *Cell* **2011**, *144*, 646–674. [[CrossRef](#)] [[PubMed](#)]
34. Fan, D.; Ren, B.; Yang, X.; Liu, J.; Zhang, Z. Upregulation of miR-501-5p activates the wnt/beta-catenin signaling pathway and enhances stem cell-like phenotype in gastric cancer. *J. Exp. Clin. Cancer Res.* **2016**, *35*, 177. [[CrossRef](#)] [[PubMed](#)]
35. Wu, F.; Xing, T.; Gao, X.; Liu, F. miR5013p promotes colorectal cancer progression via activation of Wnt/betacatenin signaling. *Int. J. Oncol.* **2019**, *55*, 671–683. [[PubMed](#)]
36. Peng, Y.; Zhang, X.; Feng, X.; Fan, X.; Jin, Z. The crosstalk between microRNAs and the Wnt/beta-catenin signaling pathway in cancer. *Oncotarget* **2017**, *8*, 14089–14106. [[PubMed](#)]
37. Cai, Z.G.; Wu, H.B.; Xu, X.P.; Li, W. Down-regulation of miR-26 plays essential roles in TGFbeta-induced EMT. *Cell Biol. Int.* **2018**.
38. Kwon, Y.; Kim, Y.; Eom, S.; Kim, M.; Park, D.; Kim, H.; Noh, K.; Lee, H.; Lee, Y.S.; Choe, J.; et al. MicroRNA-26a/-26b-COX-2-MIP-2 Loop Regulates Allergic Inflammation and Allergic Inflammation-promoted Enhanced Tumorigenic and Metastatic Potential of Cancer Cells. *J. Biol. Chem.* **2015**, *290*, 14245–14266. [[CrossRef](#)]
39. Zhang, X.; Gaspard, J.P.; Chung, D.C. Regulation of vascular endothelial growth factor by the Wnt and K-ras pathways in colonic neoplasia. *Cancer Res.* **2001**, *61*, 6050–6054.
40. Jia, L.; Liu, W.; Cao, B.; Li, H.; Yin, C. MiR-507 inhibits the migration and invasion of human breastcancer cells through Flt-1 suppression. *Oncotarget* **2016**, *7*, 36743–36754. [[CrossRef](#)]
41. Bates, R.C.; Goldsmith, J.D.; Bachelder, R.E.; Brown, C.; Shibuya, M.; Oettgen, P.; Mercurio, A.M. Flt-1-dependent survival characterizes the epithelial-mesenchymal transition of colonic organoids. *Curr. Biol.* **2003**, *13*, 1721–1727. [[CrossRef](#)]
42. Korpala, M.; Kang, Y. The emerging role of miR-200 family of microRNAs in epithelial-mesenchymal transition and cancer metastasis. *RNA Biol.* **2008**, *5*, 115–119. [[CrossRef](#)]
43. Park, S.M.; Gaur, A.B.; Lengyel, E.; Peter, M.E. The miR-200 family determines the epithelial phenotype of cancer cells by targeting the E-cadherin repressors ZEB1 and ZEB2. *Genes Dev.* **2008**, *22*, 894–907. [[CrossRef](#)]
44. Berridge, M.J.; Lipp, P.; Bootman, M.D. The versatility and universality of calcium signalling. *Nat. Rev. Mol. Cell Biol.* **2000**, *1*, 11–21. [[CrossRef](#)]
45. Xiong, W.C.; Mei, L. Agrin to YAP in Cancer and Neuromuscular Junctions. *Trends Cancer* **2017**, *3*, 247–248. [[CrossRef](#)]
46. Vasaikar, S.V.; Straub, P.; Wang, J.; Zhang, B. LinkedOmics: Analyzing multi-omics data within and across 32 cancer types. *Nucleic Acids Res.* **2018**, *46*, D956–D963. [[CrossRef](#)]
47. Raskutti, G.; Wainwright, M.J.; Yu, B. Early stopping for non-parametric regression: An optimal data-dependent stopping rule. In Proceedings of the IEEE 2011 49th Annual Allerton Conference, Monticello, IL, USA, 28–30 September 2011.
48. Li, P.; Nguyen, P.-M. On Random Deep Weight-Tied Autoencoders: Exact Asymptotic Analysis, Phase Transitions, and Implications to Training. In Proceedings of the ICLR 2019 Conference, New Orleans, LA, USA, 6–9 May 2019.
49. Rappoport, N.; Shamir, R. Multi-omic and multi-view clustering algorithms: Review and cancer benchmark. *Nucleic Acids Res.* **2018**, *46*, 10546–10562. [[CrossRef](#)]

50. Pagani, M.; Fabbri, M.; Benedetti, C.; Fassio, A.; Pilati, S.; Bulleid, N.J.; Cabibbo, A.; Sitia, R. Endoplasmic reticulum oxidoreductin 1-beta (ERO1-Lbeta), a human gene induced in the course of the unfolded protein response. *J. Biol. Chem.* **2000**, *275*, 23685–23692. [[CrossRef](#)]
51. Zito, E.; Chin, K.T.; Blais, J.; Harding, H.P.; Ron, D. ERO1-beta, a pancreas-specific disulfide oxidase, promotes insulin biogenesis and glucose homeostasis. *J. Cell. Biol.* **2010**, *188*, 821–832. [[CrossRef](#)]
52. Zhang, W.; Shen, Y.; Feng, G. Predicting the survival of patients with lung adenocarcinoma using a four-gene prognosis risk model. *Oncol. Lett.* **2019**, *18*, 535–544. [[CrossRef](#)]
53. Xie, J.; Zhu, Y.; Chen, H.; Shi, M.; Gu, J.; Zhang, J.; Shen, B.; Deng, X.; Zhan, X.; Peng, C. The Immunohistochemical Evaluation of Solid Pseudopapillary Tumors of the Pancreas and Pancreatic Neuroendocrine Tumors Reveals ERO1Lbeta as a New Biomarker. *Medicine (Baltimore)* **2016**, *95*, e2509. [[CrossRef](#)]
54. Zhu, T.; Gao, Y.F.; Chen, Y.X.; Wang, Z.B.; Yin, J.Y.; Mao, X.Y.; Li, X.; Zhang, W.; Zhou, H.H.; Liu, Z.Q. Genome-scale analysis identifies GJB2 and ERO1LB as prognosis markers in patients with pancreatic cancer. *Oncotarget* **2017**, *8*, 21281–21289. [[CrossRef](#)]
55. Nagase, T.; Ishikawa, K.-i.; Suyama, M.; Kikuno, R.; Hirose, M.; Miyajima, N.; Tanaka, A.; Kotani, H.; Nomura, N.; Ohara, O. Prediction of the coding sequences of unidentified human genes. XII. The complete sequences of 100 new cDNA clones from brain which code for large proteins in vitro. *DNA Res.* **1998**, *5*, 10. [[CrossRef](#)]
56. Maness, P.F.; Schachner, M. Neural recognition molecules of the immunoglobulin superfamily: Signaling transducers of axon guidance and neuronal migration. *Nat. Neurosci.* **2007**, *10*, 19–26. [[CrossRef](#)]
57. Oishi, N.; Kumar, M.R.; Roessler, S.; Ji, J.; Forgues, M.; Budhu, A.; Zhao, X.; Andersen, J.B.; Ye, Q.H.; Jia, H.L.; et al. Transcriptomic profiling reveals hepatic stem-like gene signatures and interplay of miR-200c and epithelial-mesenchymal transition in intrahepatic cholangiocarcinoma. *Hepatology* **2012**, *56*, 1792–1803. [[CrossRef](#)]
58. Socinski, M.A.; Kaye, F.J.; Spigel, D.R.; Kudrik, F.J.; Ponce, S.; Ellis, P.M.; Majem, M.; Lorigan, P.; Gandhi, L.; Gutierrez, M.E.; et al. Phase 1/2 Study of the CD56-Targeting Antibody-Drug Conjugate Lorvotuzumab Mertansine (IMGN901) in Combination With Carboplatin/Etoposide in Small-Cell Lung Cancer Patients With Extensive-Stage Disease. *Clin. Lung Cancer* **2017**, *18*, 68–76 e2. [[CrossRef](#)]
59. Yoshida, T.; Ri, M.; Kinoshita, S.; Narita, T.; Totani, H.; Ashour, R.; Ito, A.; Kusumoto, S.; Ishida, T.; Komatsu, H.; et al. Low expression of neural cell adhesion molecule, CD56, is associated with low efficacy of bortezomib plus dexamethasone therapy in multiple myeloma. *PLoS ONE* **2018**, *13*, e0196780. [[CrossRef](#)]
60. Crossland, D.L.; Denning, W.L.; Ang, S.; Olivares, S.; Mi, T.; Switzer, K.; Singh, H.; Huls, H.; Gold, K.S.; Glisson, B.S.; et al. Antitumor activity of CD56-chimeric antigen receptor T cells in neuroblastoma and SCLC models. *Oncogene* **2018**, *37*, 3686–3697. [[CrossRef](#)]
61. Aref, S.; Azmy, E.; El-Bakry, K.; Ibrahim, L.; Mabed, M. Prognostic impact of CD200 and CD56 expression in adult acute lymphoblastic leukemia patients. *Hematology* **2018**, *23*, 263–270. [[CrossRef](#)]
62. Jhiang, S.M. The RET proto-oncogene in human cancers. *Oncogene* **2000**, *19*, 5590–5597. [[CrossRef](#)]
63. Anonymous. First RET Inhibitor on Path to FDA Approval. *Cancer Discov.* **2019**, *9*, 1476–1477.
64. Griseri, P.; Garrone, O.; Lo Sardo, A.; Monteverde, M.; Rusmini, M.; Tonissi, F.; Merlano, M.; Bruzzi, P.; Lo Nigro, C.; Ceccherini, I. Genetic and epigenetic factors affect RET gene expression in breast cancer cell lines and influence survival in patients. *Oncotarget* **2016**, *7*, 26465–26479. [[CrossRef](#)]
65. Hamamoto, R.; Komatsu, M.; Takasawa, K.; Asada, K.; Kaneko, S. Epigenetics Analysis and Integrated Analysis of Multiomics Data, Including Epigenetic Data, Using Artificial Intelligence in the Era of Precision Medicine. *Biomolecules* **2019**, *10*, 62. [[CrossRef](#)] [[PubMed](#)]
66. Liu, H.; Mintern, J.D.; Villadangos, J.A. MARCH ligases in immunity. *Curr. Opin. Immunol.* **2019**, *58*, 38–43. [[CrossRef](#)] [[PubMed](#)]
67. Meng, Y.; Hu, J.; Chen, Y.; Yu, T.; Hu, L. Silencing MARCH1 suppresses proliferation, migration and invasion of ovarian cancer SKOV3 cells via downregulation of NF-kappaB and Wnt/beta-catenin pathways. *Oncol. Rep.* **2016**, *36*, 2463–2470. [[CrossRef](#)] [[PubMed](#)]
68. Xie, L.; Dai, H.; Li, M.; Yang, W.; Yu, G.; Wang, X.; Wang, P.; Liu, W.; Hu, X.; Zhao, M. MARCH1 encourages tumour progression of hepatocellular carcinoma via regulation of PI3K-AKT-beta-catenin pathways. *J. Cell Mol. Med.* **2019**, *23*, 3386–3401. [[CrossRef](#)]

69. Feng, M.; Xiong, G.; Cao, Z.; Yang, G.; Zheng, S.; Qiu, J.; You, L.; Zheng, L.; Zhang, T.; Zhao, Y. LAT2 regulates glutamine-dependent mTOR activation to promote glycolysis and chemoresistance in pancreatic cancer. *J. Exp. Clin. Cancer Res.* **2018**, *37*, 274. [[CrossRef](#)]
70. Hayashi, K.; Anzai, N. Novel therapeutic approaches targeting L-type amino acid transporters for cancer treatment. *World J. Gastrointest. Oncol.* **2017**, *9*, 21–29. [[CrossRef](#)]
71. Wang, Q.; Holst, J. L-type amino acid transport and cancer: Targeting the mTORC1 pathway to inhibit neoplasia. *Am. J. Cancer Res.* **2015**, *5*, 1281–1294.
72. Ettinger, D.S.; Wood, D.E.; Aisner, D.L.; Akerley, W.; Bauman, J.; Chirieac, L.R.; D’Amico, T.A.; DeCamp, M.M.; Dilling, T.J.; Dobelbower, M.; et al. Non-Small Cell Lung Cancer, Version 5.2017, NCCN Clinical Practice Guidelines in Oncology. *J. Natl. Compr. Canc. Netw.* **2017**, *15*, 504–535. [[CrossRef](#)]
73. Liu, G.; Wang, P.; Zhang, H. MiR-6838-5p suppresses cell metastasis and the EMT process in triple-negative breast cancer by targeting WNT3A to inhibit the Wnt pathway. *J. Gene. Med.* **2019**, e3129. [[CrossRef](#)]
74. Bauml, J.; Mick, R.; Zhang, Y.; Watt, C.D.; Vachani, A.; Aggarwal, C.; Evans, T.; Langer, C. Frequency of EGFR and KRAS mutations in patients with non small cell lung cancer by racial background: Do disparities exist? *Lung Cancer* **2013**, *81*, 347–353. [[CrossRef](#)]



© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).