


Article

Gene Screening in High-Throughput Right-Censored Lung Cancer Data

Chenlu Ke ¹, Dipankar Bandyopadhyay ^{2,*} , Mario Acunzo ³ and Robert Winn ⁴

¹ Department of Statistical Sciences and Operations Research, Virginia Commonwealth University, Richmond, VA 23284, USA

² Department of Biostatistics, Virginia Commonwealth University, Richmond, VA 23284, USA

³ Department of Internal Medicine, Virginia Commonwealth University, Richmond, VA 23284, USA

⁴ Massey Cancer Center, Virginia Commonwealth University, Richmond, VA 23284, USA

* Correspondence: dbandyop@vcu.edu; Tel.: +1-804-827-2058

Abstract: Background: Advances in sequencing technologies have allowed collection of massive genome-wide information that substantially advances lung cancer diagnosis and prognosis. Identifying influential markers for clinical endpoints of interest has been an indispensable and critical component of the statistical analysis pipeline. However, classical variable selection methods are not feasible or reliable for high-throughput genetic data. Our objective is to propose a model-free gene screening procedure for high-throughput right-censored data, and to develop a predictive gene signature for lung squamous cell carcinoma (LUSC) with the proposed procedure. Methods: A gene screening procedure was developed based on a recently proposed independence measure. The Cancer Genome Atlas (TCGA) data on LUSC was then studied. The screening procedure was conducted to narrow down the set of influential genes to 378 candidates. A penalized Cox model was then fitted to the reduced set, which further identified a 6-gene signature for LUSC prognosis. The 6-gene signature was validated on datasets from the Gene Expression Omnibus. Results: Both model-fitting and validation results reveal that our method selected influential genes that lead to biologically sensible findings as well as better predictive performance, compared to existing alternatives. According to our multivariable Cox regression analysis, the 6-gene signature was indeed a significant prognostic factor (p -value < 0.001) while controlling for clinical covariates. Conclusions: Gene screening as a fast dimension reduction technique plays an important role in analyzing high-throughput data. The main contribution of this paper is to introduce a fundamental yet pragmatic model-free gene screening approach that aids statistical analysis of right-censored cancer data, and provide a lateral comparison with other available methods in the context of LUSC.



Citation: Ke, C.; Bandyopadhyay, D.; Acunzo, M.; Winn, R. Gene Screening in High-Throughput Right-Censored Lung Cancer Data. *Onco* **2022**, *2*, 305–318. <https://doi.org/10.3390/onco2040017>

Academic Editor: Fred Saad

Received: 29 July 2022

Accepted: 14 October 2022

Published: 17 October 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 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 (<https://creativecommons.org/licenses/by/4.0/>).

Keywords: high dimensional data; lung cancer; right-censored; survival; sure independence screening; TCGA

1. Introduction

Lung cancer is the leading cause of cancer-related death worldwide, with an estimated 1.8 million [1] deaths (from GLOBOCAN 2020 estimates). Among these, the non-small cell lung cancer (NSCLC) is the most common histological cell type and often presents in an advanced stage [2]. NSCLC is classified into adenocarcinoma, and squamous cell carcinoma (LUSC) subtypes, where LUSC comprises approximately 30% of all lung cancers [3]. Although molecular targeted therapy has been developed to significantly improve patient survival [4], tumour heterogeneity have been found to render the therapy ineffective [5]. For many patients the therapeutic options are still limited, especially for the LUSC subtype. The identification of biomarkers that contribute to early detection and effective treatment of LUSC is a vital yet ongoing research task, which is characterized by high-throughput data generated in a massive and fast manner by 'omics' technologies, such as transcriptomics, metabolomics and proteomics. As it is commonly believed that only a small portion of the

clinical and genetic features are related to a certain endpoint of interest, a key aspect of the related statistical analysis is to extract core information by identifying low-dimensional sparse presentations of the predictive features, which is like finding a needle in a haystack for high-dimensional data. Traditional modeling techniques are handicapped, if the dimension (the number of variables) exceeds sample size. For example, the proportional hazards (PH) model has been widely used for predicting time-to-event outcomes, but the partial likelihood estimation is not appropriate for studying the simultaneous relationship of the high-throughput microarray data with the outcomes [6]. Hence, variable selection [7] becomes an indispensable part of the statistical analysis pipeline. However, when the number of variables is much larger than sample size, exact variable selection is often beyond the hope to achieve. Univariate analysis is commonly used to select significant biomarkers for downstream analysis without multiple testing correction [8–10], which accumulates false discoveries [11]. For large-scale multiple testing, the power to reject a non-null hypothesis while controlling for the family wise error rate through, for example, the Bonferroni adjustment, is greatly reduced as the number of tests increases [12]. Regularization methods such as LASSO [13] have also been applied to conduct gene selection for LUSC [14], but they can suffer from both statistical and computational issues if the number of features far exceeds sample size [15].

Recent years have seen rising attention to variable screening as a less ambitious yet efficient way to reduce dimension for ultrahigh dimensional data. Variable screening was first [15] introduced for linear model to quickly filter out redundant features through marginal independence learning based on the Pearson correlation. In other words, features are ranked based on their marginal associations with the outcome variable and unimportant genes are removed from the bottom of an ordered list. The screening mechanism asymptotically almost surely identifies all important predictors, and thus is called ‘sure independence screening’ (SIS). Since conjecturing about underlying model structure is presumably challenging in high dimensional spaces, more flexible approaches have emerged to avoid model specifications [15–19]. Screening has found applications ranging from quality control in the data processing step for genetic studies [20] to identifying predictive biomarkers for understanding biological mechanisms [21]. Notwithstanding the vast literature in feature screening for fully observed outcomes, the development of screening procedures to accommodate censoring has been less fruitful. Model-based methods include SIS for Cox PH model [22], the principled Cox sure screening [23] and the feature aberration at survival times screening [24], among others. In particular, SIS for Cox PH model [22] has been employed to discover prognostic gene signatures in breast cancer [25,26] and lung cancer [27,28]. However, if the association between the biomarkers and the survival outcomes cannot be well captured by the Cox model, which is rather difficult to check in practice for high dimensional data (due to the somewhat restrictive PH assumptions), SIS may fail to detect significant markers. Therefore, we argue that screening procedures that requires no model specification should be promoted, when there is insufficient information about data distribution and the underlying model structure. Existing model-free approaches include the quantile adaptive SIS [29], the censored rank independence screening (CRIS [30]), the survival impact index screening [31], the integrated powered density screening (IPO D [32]), and the robust screening via distance correlation [33]. Although the effectiveness of these aforementioned methods have been established through simulation studies [34], they have not been examined in the context of gene screening for cancer survival data. As appealing as the idea of variable screening is, the lack of application and dissemination hinders practical usage, which can benefit researchers from a wide biomedical domain.

In this paper, we proposed a model-free gene screening procedure for high-throughput right-censored cancer data based on the expected conditional characteristic function-based independence criterion (ECCFIC [35]). The ECCFIC correlation can be viewed as a nonlinear generalization of the classical coefficient of determination R^2 since it requires no linearity or distributional assumptions and therefore can be used to achieve model-free screening. We applied the screening procedure to the TCGA LUSC dataset and identified a novel 6-gene signature for prognosis of LUSC patients. The performance of the screening procedure was evaluated via comparing and contrasting to existing alternatives.

2. Materials and Methods

2.1. Data Description

Gene expression data and clinical data for patients with LUSC were acquired from TCGA (<https://cancergenome.nih.gov/> accessed on 30 September 2022) for model training and testing. In addition, information obtained from the Gene Expression Omnibus (GEO) database (GSE37745 [36] and GSE30219 [37]; <https://www.ncbi.nlm.nih.gov/geo/> (accessed on 30 September 2022)) was used for external validation. For a patient without an event (death), the overall survival time from first diagnosis was censored by the last follow-up date. Disease-free survival is defined as time to new tumor event after the initial treatment. Aside from 17,557 common genes in all datasets, 5 clinical covariates were also included in the analysis: age at diagnosis, gender, smoking history, metastasis and tumor stage. In total, 473 and 127 cases with completed data were extracted from TCGA and GEO datasets, respectively; 760 genes were excluded due to complete missing or low expression (with an interquartile range of 0). Table 1 summarizes the clinical and pathological characteristics of the TCGA patients. The majority of the patients were older than 60 at first diagnosis (82.6%) and had smoking history within 15 years (78.0%). Additionally, 81.6% of the patients had stage I and II squamous cell carcinoma, with only 1.5% of the patients presenting with stage IV carcinoma. 207 patients died during follow-up. The survival times range from 0.03 to 173.69 months, with a median of 21.19 months. The recurrence rate was 34.1%.

2.2. Sure Independence Screening for Right-Censored Data

We first introduce the concept of sure independence screening. Let T denote the time to event with respect to a certain cancer type, C denote the censoring time, $Y := \min(T, C)$ denote the observed time and $\delta := I(T \leq C)$ denote the failure indicator, where $I(\cdot)$ is the indicator function. Let $\mathbf{X} \in \mathbb{R}^p$ be the vector of all genes. Throughout the paper, we assume independent censoring, that is, $(T, \mathbf{X}) \perp\!\!\!\perp C$. Let \mathcal{A} denote the index set of the influential genes, that is,

$$\mathcal{A} := \{1 \leq j \leq p : P(T > t | \mathbf{X}) \text{ functionally depends on } X_j\}.$$

Our goal is to achieve gene screening, that is, to find a reduced index set that covers \mathcal{A} with cardinality smaller than n . Note that gene screening is less ambitious than exact gene selection that recovers \mathcal{A} precisely, but employed to quickly eliminate the majority of irrelevant genes and reduce the high dimensional data to a manageable subset.

Table 1. Summary of clinical and pathological characteristics.

Variables	Frequency (Percent)
Age	
Less than 50	15 (3.2%)
50–59	67 (14.2%)
60–69	178 (37.6%)
70–79	186 (39.3%)
80 or greater	27 (5.7%)
Gender	
Female	125 (26.4%)
Male	348 (73.6%)
Smoking History	
Current reformed smoker for ≤ 15 years	236 (49.9%)
Current reformed smoker for > 15 years	81 (17.1%)
Current reformed smoker, duration not specified	5 (1.1%)
Current smoker	133 (28.1%)
Lifelong non-smoker	18 (3.8%)
Lymph Node Metastasis	
N0	302 (63.8%)
N1, N2, N3	165 (34.9%)
NX	6 (1.3%)
Distant Metastasis	
M0	386 (81.6%)
M1, M1a, M1b	7 (1.5%)
MX	80 (16.9%)
Pathological Stage	
I	236 (49.9%)
II	150 (31.7%)
III	80 (16.9%)
IV	7 (1.5%)

2.3. The Screening Index

Before we introduce the procedure of gene screening, we briefly review the measure that will be used to assess the dependence between the survival time and each candidate gene. Let U and V be two random variables. The generalized ECCFIC [35] for testing $U \perp\!\!\!\perp V$ is defined as

$$\mathcal{H}_K^2(U|V) := E_V E_{U|V,U'|V} K(U, U') - E_{U,U'} K(U, U'),$$

for a characteristic [38] positive definite kernel $K : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$, where $E_{U|v,U'|v}$ denotes $E(\cdot|V = v, V' = v)$ and (U', V') is an independent and identically distributed copy of (U, V) . Examples of characteristic kernels include Gaussian, Laplacian, inverse multi-quadratics, and distance-induced kernels [39]. A corresponding correlation measure is then defined as

$$\rho_K(U|V) := \frac{\mathcal{H}_K^2(U|V)}{\mathcal{H}_K^2(U|U)},$$

where $\mathcal{H}_K^2(U|U) = E_U K(U, U) - E_{U,U'} K(U, U')$. It can be showed that $0 \leq \rho_K(U|V) \leq 1$, where $\rho_K(U|V) = 0$ if and only if U and V are independent and $\rho_K(U|V) = 1$ if and only if U is a function of V . The ECCFIC correlation can capture nonlinear dependence with the kernel trick and thus, is more general than the coefficient of determination or the Pearson correlation coefficient.

The Nadaraya-Watson estimator of $\mathcal{H}_K^2(U|V)$ relying on a selected smoothing kernel $G : \mathbb{R} \rightarrow \mathbb{R}$ and a tuning bandwidth $h := h(n) \in \mathbb{R}$ is given by

$$\mathcal{H}_{K,G,n}^2(U|V) := \frac{1}{n^5} \sum_{t_1,t_2,t_3,t_4,t_5=1}^n \frac{G_{t_1t_2}G_{t_1t_3}d_{t_2t_3t_4t_5}}{\frac{1}{n^2} \sum_{s_1,s_2=1}^n G_{t_1s_1}G_{t_1s_2}},$$

where $G_{ts} := G_h(V_t - V_s)$, $G_h(v) := h^{-q}G(v/h)$, $d_{t_2t_3t_4t_5} := K_{t_2t_3} - K_{t_2t_4} - K_{t_3t_5} + K_{t_4t_5}$ and $K_{ts} := K(U_t, U_s)$. Furthermore, a natural estimator of $\mathcal{H}_K^2(U|U)$ is given by

$$\mathcal{H}_{K,n}^2(U|U) := \frac{1}{n} \sum_{i=1}^n K(U_i, U_i) - \frac{1}{n^2} \sum_{i_1,i_2=1}^n K(U_{i_1}, U_{i_2}).$$

Then the ECCFIC correlation can be estimated by

$$\rho_{K,G,n}(U|V) := \frac{\mathcal{H}_{K,G,n}^2(U|V)}{\mathcal{H}_{K,n}^2(U|U)}.$$

In practice, the bandwidth h is often set to $1.06\tilde{\sigma}n^{-1/5}$, where $\tilde{\sigma}$ is estimated by the sample standard deviation of V [40].

2.4. The Screening Algorithm

We now provide an algorithm to achieve gene screening for high-throughput right-censored cancer data. The ECCFIC correlation between $U_T := F_T(T)$ and $U_{X_j} := F_{X_j}(X_j)$ is adopted to quantify the importance of the individual gene X_j ($j = 1, \dots, p$), where $F_T(\cdot)$ is the cumulative distribution function (CDF) of T and $F_{X_j}(\cdot)$ is the CDF of X_j . Note that $T \perp\!\!\!\perp X_j$ if and only if $U_T \perp\!\!\!\perp U_{X_j}$, but we choose to work with the later condition, since (1) T is not observable but U_T can be easily estimated by the well-known Kaplan–Meier estimator, and (2) U_{X_j} 's provide robustness to heavy tails or outliers of the gene expression.

For a characteristic kernel K of choice, let $w_j := \rho_K(U_T|U_{X_j})$. Given the observed data $\{X_i, Y_i, \delta_i\}_{i=1}^n$, the steps of our algorithm are as follows:

1. Estimate the survival function by the Kaplan–Meier estimator as

$$\hat{F}_T(t) := 1 - \prod_{i=1}^n \left(1 - \frac{1}{\sum_{l=1}^n I\{Y_l \geq Y_i\}} \right)^{\delta_i I\{Y_i \leq t\}}$$

and compute the empirical CDF of X_j as $\hat{F}_{X_j}(x) = \frac{1}{n} \sum_{i=1}^n I\{X_{ij} \leq x\}$;

2. Treat $\{\hat{F}_{X_j}(X_{ij}), \hat{F}_T(Y_i)\}_{i=1}^n$ as the observed data of (U_{X_j}, U_T) and compute the sample correlation $\hat{w}_j := \rho_{K,G,n}(U_T|U_{X_j})$ for $j = 1, \dots, p$.
3. Let $\hat{\mathcal{A}} := \{1 \leq j \leq p : \hat{w}_j \text{ is among the first } d \text{ largest of all}\}$.

We henceforth refer to our procedure as the ECCFIC-based sure independence screening, or ESIS for short. In practice, common choices of d are $\lfloor n/\log(n) \rfloor$, $2\lfloor n/\log(n) \rfloor$, $3\lfloor n/\log(n) \rfloor$, and $n - 1$ [15,16]. Once the dataset is sufficiently downsized by ESIS, traditional lower dimensional methods can be used afterwards for gene selection and statistical inference (Figure 1). It is noteworthy to point out that ESIS does not impose any model assumptions on the distribution of $T|X$. The R code to implement the proposed algorithm is available at <https://github.com/cke23/GeneScreeningDemo1> (accessed on 30 September 2022).

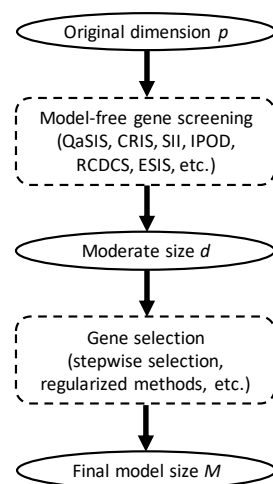


Figure 1. Overall diagram of gene screening and selection procedures.

2.5. Application

The TCGA data were divided into a training set and a testing set in a ratio of 4:1 by stratified randomization based on censoring. The training set was comprised of 379 samples and the testing set was comprised of 94 samples. We first performed ESIS on the training set and pre-selected $379 - 1 = 378$ genes. The characteristic kernel as well as the smoothing kernel were both chosen to be the Gaussian kernel. A Penalized Cox model with LASSO regularization (abbreviated as PenCox henceforth) was then applied to the reduced training data for further gene selection and prognosis simultaneously via R package glmnet. The optimal tuning parameter was determined through 10-fold cross validation. A patient's risk score was calculated as the linear predictor of the fitted PenCox model. Patients were classified as having a high-risk gene signature or a low-risk gene signature, with the median risk score of the training group being the cutoff. The same cutoff value was also applied when assigning the test samples (TCGA LUSC data) and the external validation samples (GEO datasets) into two risk groups. To evaluate the predictive performance of the PenCox model built upon the ESIS-selected genes, the Kaplan–Meier curves of the two risk groups for both overall survival and disease-free survival were compared using the log-rank tests. Moreover, the time-dependent receiver operating characteristic (ROC) curve along with the area under the curve (AUC) were calculated. Finally, a Cox model was fitted to the entire TCGA dataset to make inference about independent prognostic factors associated with survival, and the selected gene signature, age, gender, tumor stage, and smoking history were used as covariates. The same analysis preceded by two existing screening methods, namely CRIS [30] and IPOD [32], were also conducted, respectively, for comparisons. As a baseline model, we performed a naive screening procedure followed by PenCox. That is, we ranked the genes by their variations and select the top 700 for downstream analysis [41]. The purpose of the baseline model was to evaluate the classical regularization method with relatively high dimensional data and the naive screening procedure assisted to reduce the computational cost. In total, four models were included for comparisons: Naive+PenCox, CRIS+PenCox, IPOD+PenCox, and ESIS+PenCox.

3. Results

Table 2 lists the influential genes selected by each of the four competing models. All models successfully distinguished the two risk groups for the training data with p -values < 0.001 . For the testing samples, the ESIS+PenCox model also led to a separation between the two groups (p -value = 0.078). Patients with a high-risk gene signature had a shorter median overall survival than those with a low-risk gene signature (34.7 months vs. 71.3 months). Moreover, patients with a high-risk gene signature were associated with a shorter disease-free survival than patients with a low-risk gene signature (29.7 months vs. not reached for median survival, p -value = 0.041). The same observation held for the

subgroup of patients with metastasis (34.4 months vs. not reached for median disease-free survival; p -value = 0.010). For the external validation samples, the ESIS+PenCox provided the best stratification among the four models in terms of overall survival (p -value = 0.016) and disease-free survival (p -value = 0.005 for all patients and p -value = 0.083 for patients with metastasis). The prognostic indices based on the genes selected by the other screening methods were less informative, leading to insignificant discrepancies between the two risks groups in the validation data. The PenCox model with naive screening suffered from the high dimensionality (700 genes) and failed to predict overall survival and disease-free survival effectively. Figure 2 shows the overall survival curves for high-and-low risk groups in the testing and external validation cohorts, while results for disease-free survival are presented in Figure 3. Figure 4 displays the ROC curves at 1, 3, 5 and 10 years for the competing models on the external validation data. The results also suggest that the ESIS+PenCox model provided the best predictions.

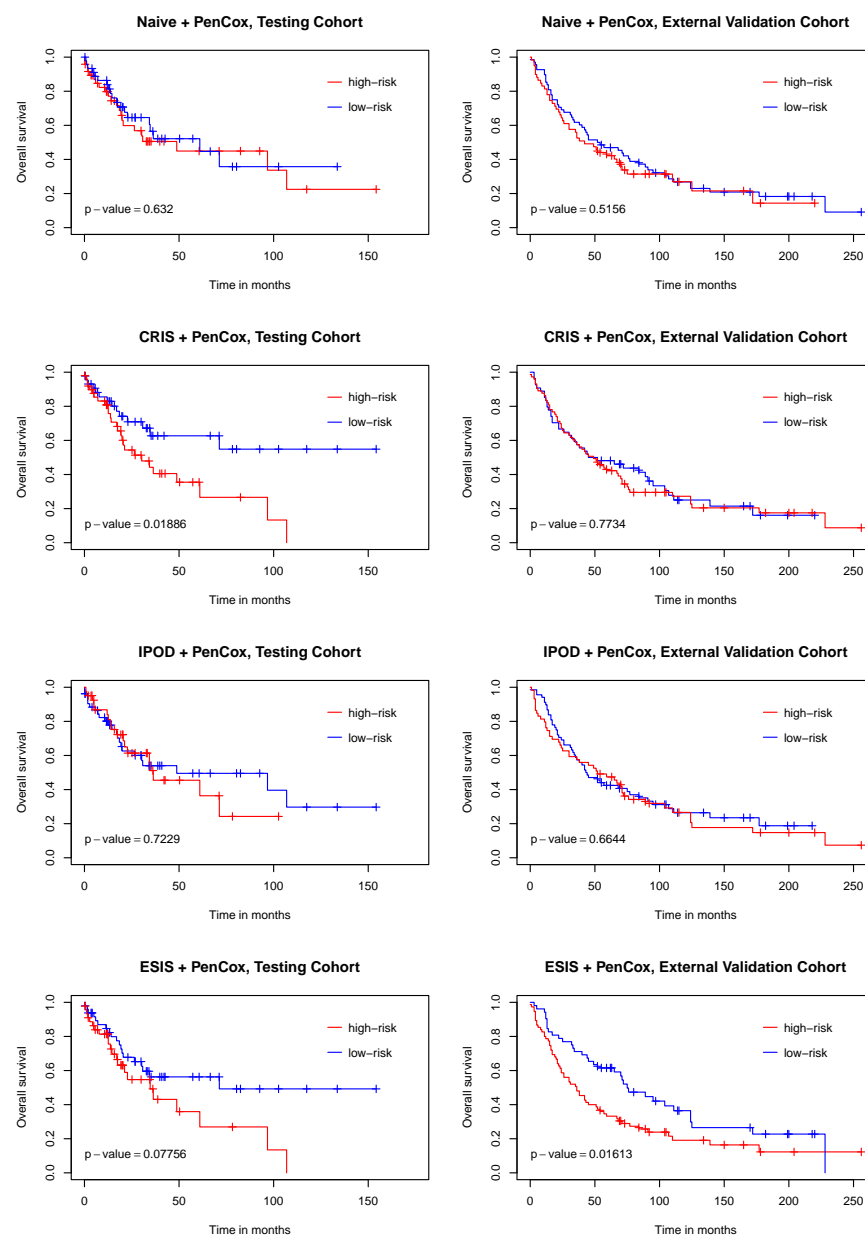


Figure 2. Kaplan–Meier curves of overall survival for test (TCGA) and validation (GEO) cohorts varying with the risk level determined by the four competing models. p -values were obtained from the log-rank tests contrasting the two risk groups.

Table 2. Genes selected by the four competing models. A risk gene with a positive coefficient from the fitted PenCox model is denoted by “+”, while a protective gene with a negative coefficient is denoted by “-”.

Model (No. of Genes Selected)	Gene Names
Naive + PenCox (6)	PCDHA5(+), C9ORF131(+), PM20D1(+), PCDHA3(+), FAM196B(+), PITX3(-)
CRIS + PenCox (10)	CCDC79(+), LCN1(+), GPR78(+), SSX1(+), CCKAR(+), SLC10A2(+), STARD6(-), GUCY2F(-), DPPA2(+), LINC00628(+)
IPOD + PenCox (4)	TRIM58(+), C9ORF131(+), PKNX2(+), PCDHGA11(+)
ESIS + PenCox (6)	NACC2(+), FAM65A(+), LOC641845(-), MON1B(+), IBTK(+), SDHAF3(-)

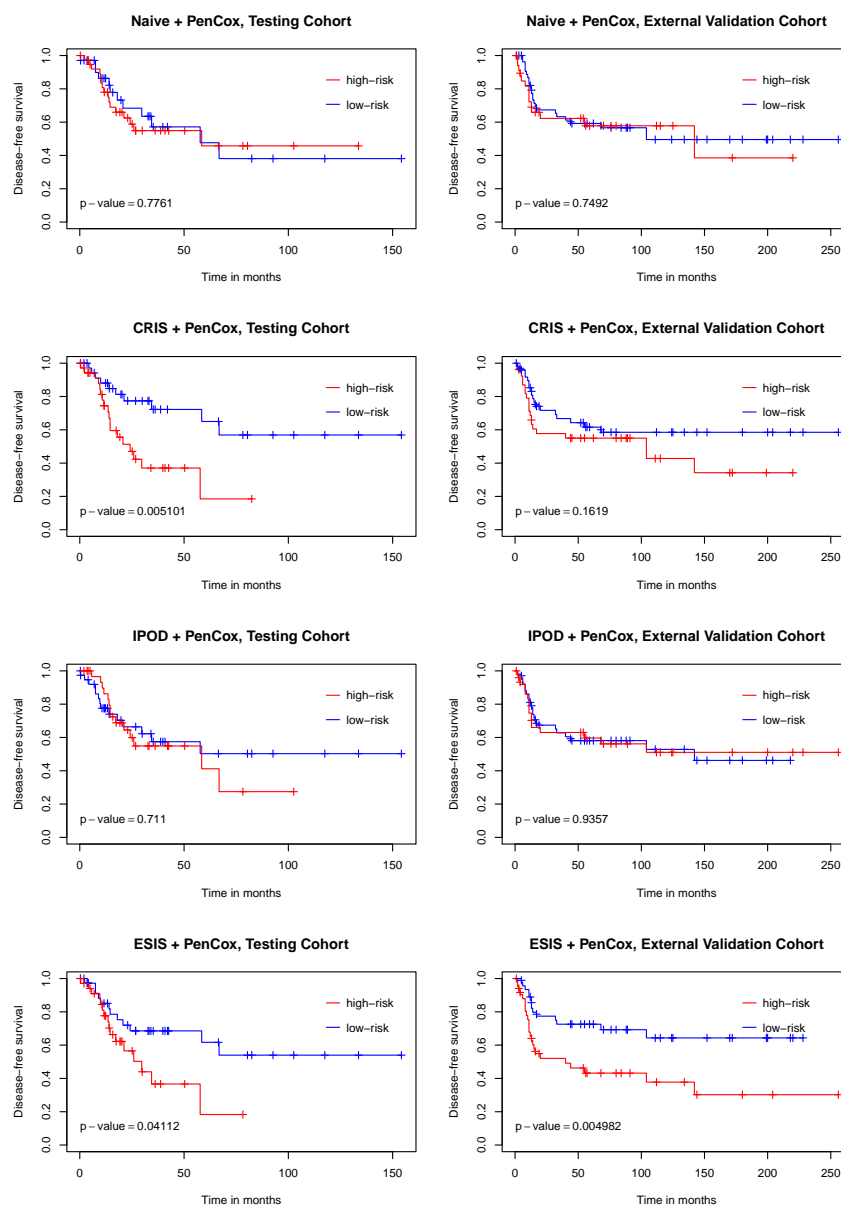


Figure 3. Kaplan–Meier curves of disease-free survival for test (TCGA) and validation (GEO) cohorts varying with the risk level determined by the four competing models. *p*-values were obtained from the log-rank tests contrasting the two risk groups.

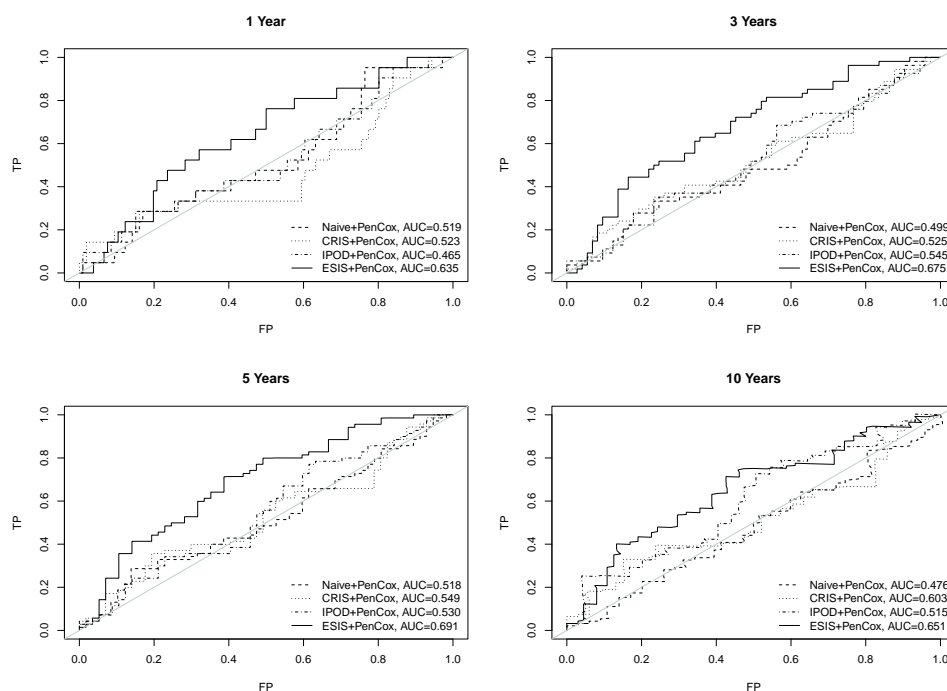


Figure 4. ROC curves of overall survival predicted by the four competing models on the external validation data.

From the multivariable Cox regression (Table 3), the 6-gene signature selected by ESIS+PenCox was a strong predictor with an hazard ratio of 12.59 (p -value < 0.001), adjusted for other clinical covariates. There was a 2% increase in the expected hazard relative to a one year increase in age (p -value = 0.008). Subjects with first or the second stage of cancer experienced reduction of hazard by 41% (p -value = 0.003) and 37% (p -value = 0.018), respectively, compared to those in later stages. Current smokers were associated with worse prognosis (p -value = 0.005).

Table 3. Multivariable Cox regression analysis of the risk of death against the 6-gene signature identified by ESIS+PenCox and other clinical covariates. CI denotes confidence interval.

Variable	Hazard Ratio (95% CI)	p -Value
6-gene signature	12.59 (4.11, 38.56)	<0.001
Age	1.02 (1.01, 1.04)	0.008
Gender		
Male	0.92 (0.67, 1.28)	0.629
Female	-	-
Tumor stage		
I	0.59 (0.42, 0.83)	0.003
II	0.63 (0.43, 0.92)	0.018
III or IV	-	-
Smoking history		
Lifelong non-smoker	1.94 (0.83, 4.54)	0.126
Current smoker	1.54 (1.14, 2.07)	0.005
Current reformed smoker	-	-

Finally, we highlight some biological insights associated with the genes selected by ESIS+PenCox. The protective gene SDHAF3 has been found to be involved in the maturation of succinate dehydrogenase (SDH) genes, which are known as classical tumor suppressors [42]. Following the suppression of SDH genes, an accumulation of succinate

results in stabilizing HIF- α , thereby promoting angiogenesis and ROS production [43]. In particular, inhibition of SDHB induces the transition to anaerobic metabolism, better known as the Warburg effect, which is widely observed in human cancers [44]. Single nucleotide polymorphisms (SNPs) in SDH genes have been associated with the clinical outcome of NSCLC patients [45]. Studies have shown that the indication of IBTK may be expanded beyond hematological malignancies [46]. In several cancers, IBTK functions to sustain tumorigenesis and cell survival [47]. For instance, IBTK has been identified as a risk gene of NSCLC, owing to its association with KRAS, AKT1, BRAF and MAPK1 [48]. It has also been revealed that FAM65A binds to Rho GTPases that regulate cancer cell migration [49,50]. FAM65A is a component of the gene expression profiles for atopy [51] and pulmonary function impairment [52]. Mon1 mediates the transition from early-to-late endosome in metazoa by switching Rab5 for Rab7 via guanine nucleotide exchange factors [53]. Mon1b, the mammalian homolog of Mon1, interacts with Numb for docking of early endosomes [54]. Mon1b is elevated in colon cancer, with its knockdown in vitro leading to a reduction of proliferation, migration, and invasion [55]. There is evidence that NACC2/RBB inhibits cell cycle progression and promotes apoptosis by enhancing the p53 pathway [56]. NACC2 has also been identified as an NTRK fusion protein, specifically in pilocytic astrocytoma [57,58]. NTRK gene fusions lead to constitutive activation of TRK kinases in multiple cancers, thereby making them promising candidates for chemotherapeutic drug development [59]. The protective gene LOC641845/STMP1 is a short trans-membrane mitochondrial protein that participates in the regulation of cellular respiration [60]. Although this gene has not been widely studied, it appears to have a role in Paget's disease of the bone [61].

Thanks to the ENCODE transcription factor target datasets [62,63] that are available on the Harmonizome database [64], we identified two transcription factors, E2F4 and ELF1, which regulate five out of the six genes selected by the ESIS+PenCox model: NACC2, FAM65A, MON1B, IBTK, and SDHAF3. E2F4 is a member of the E2F family of transcription factors which regulate the expression of key genes implicated in cell division [65]. In particular, E2F4 belongs to a subclass of repressive E2Fs that play a role in cell cycle exit and terminal differentiation [65]. ELF1 belongs to the E26 transformation specific (ETS) family of transcription factors which regulate the expression of genes involved in several processes that are considered the hallmarks of cancer [66,67]. ELF1 binds to the HER2 promoter and is upregulated in several cancers (prostate, ovarian, breast, leukemia, lymphoma) [66].

4. Discussion

Finding prognostic gene signatures for cancer survival is a vital task in biomedical research. Since it is commonly believed that only a small portion of genes are related to a certain outcome, how to recover the most influential subset from massive data becomes a challenge in related statistical analysis. Traditional variable selection methods such as stepwise selection can only be applied when the number of variables is smaller than the sample size. Researchers often use prior knowledge or univariate analysis to select genes for downstream analysis, which lacks quantitative justification and could hinder the discovery of novel gene markers. Although regularization methods have also been widely used, they can be unstable for high-throughput data (the number of genes far exceeds the number of samples). Fast and effective variable screening tools for high dimensional survival data have been emerging in the past decade in the statistics literature. However, the dissemination of these attractive methods to biomedical fields is limited. In this paper, we proposed a novel sure independence screening procedure for identifying prognostic genes in LUSC. Our approach was able to reduce the dimension efficiently while preserving influential genes that lead to biologically sensible findings and provide better prognosis for LUSC in comparison with competing methods. The proposed gene screening tool is fundamental and general, and thus can be readily applicable to other cancer databases with

right censored survival. Classical gene selection and prognostic modeling can be conducted subsequently after the dataset is downsized through screening.

Admittedly, this paper poses some open questions besides what it solves. Our method allows a variety of kernels for detecting important genes involved with different types of model structure, but kernel selection is commonly challenging and requires a large amount of practical experience for the researchers. As future work, we plan to develop a composite algorithm integrating results of distinct kernels. Besides, in many applications, researchers know from previous investigations that certain features are responsible for the survival outcomes or should be controlled for in the studies. Examples include TNM clinical stage, pathological stage, metastasis, age, gender, smoking history, and known gene markers. Although some of the covariates were included in the final prognostic modeling stage in our study, they may also assist in the selection of important genes while being shielded in the screening procedure. In future work, we also plan to investigate such conditional screening procedures that can incorporate prior information to improve the screening power.

5. Conclusions

We developed a novel and powerful model-free gene screening approach that aids statistical analysis of high-throughput right-censored data. The application to TCGA LUSC data provided a paradigm of its implementation combining classical gene selection and prognostic modeling. As a result, we discovered a novel and effective six-gene model to predict the prognosis of patients with LUSC. It is expected that this presented work will be a desired addition to a cancer epidemiologist's toolbox.

Author Contributions: Conceptualization, C.K. and D.B.; methodology, C.K. and D.B.; software, C.K.; validation, C.K.; formal analysis, C.K.; investigation, C.K. and D.B.; resources, R.W.; data curation, C.K.; writing—original draft preparation, C.K.; writing—review and editing, C.K., D.B. and M.A.; visualization, C.K. and D.B.; supervision, D.B.; project administration, D.B.; funding acquisition, R.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by grants P20CA252717, P20CA264067, R01DE031134, R21DE031879 from the United States National Institutes of Health (NIH) and the VCU Quest fund. Services and products in support of this research project were also generated by the VCU Massey Cancer Center Biostatistics Shared Resource, supported, in part, with funding from NIH-NCI Cancer Center Support Grant P30CA016059.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The LUSC data used in this paper is available at the TCGA repository <https://cancergenome.nih.gov/> (accessed on 30 September 2022) and the GEO database <https://www.ncbi.nlm.nih.gov/geo/> (accessed on 30 September 2022). The R code for the proposed algorithm is available at <https://github.com/cke23/GeneScreeningDemo1> (accessed on 30 September 2022).

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Sung, H.; Ferlay, J.; Siegel, R.L.; Laversanne, M.; Soerjomataram, I.; Jemal, A.; Bray, F. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* **2021**, *71*, 209–249. [[CrossRef](#)] [[PubMed](#)]
2. Custodio, A.; Méndez, M.; Provencio, M. Targeted therapies for advanced non-small-cell lung cancer: Current status and future implications. *Cancer Treat. Rev.* **2012**, *38*, 36–53. [[CrossRef](#)]
3. The Cancer Genome Atlas Research Network. Comprehensive genomic characterization of squamous cell lung cancers. *Nature* **2012**, *489*, 519–525. [[CrossRef](#)]
4. Suda, K.; Mitsudomi, T. Successes and limitations of targeted cancer therapy in lung cancer. *Successes Limitations Target. Cancer Ther.* **2014**, *41*, 62–77.

5. Lee, Y.T.; Tan, Y.J.; Oon, C.E. Molecular targeted therapy: Treating cancer with specificity. *Eur. J. Pharmacol.* **2018**, *834*, 188–196. [[CrossRef](#)] [[PubMed](#)]
6. Pi, L.; Halabi, S. Combined performance of screening and variable selection methods in ultra-high dimensional data in predicting time-to-event outcomes. *Diagn. Progn. Res.* **2018**, *2*, 21. [[CrossRef](#)] [[PubMed](#)]
7. Chandrashekar, G.; Sahin, F. A survey on feature selection methods. *Comput. Electr. Eng.* **2014**, *40*, 16–28. [[CrossRef](#)]
8. Larsen, J.E.; Pavay, S.J.; Passmore, L.H.; Bowman, R.; Clarke, B.E.; Hayward, N.K.; Fong, K.M. Expression profiling defines a recurrence signature in lung squamous cell carcinoma. *Carcinogenesis* **2007**, *28*, 760–766. [[CrossRef](#)]
9. Skrzypski, M.; Jassem, E.; Taron, M.; Sanchez, J.J.; Mendez, P.; Rzyman, W.; Gulida, G.; Raz, D.; Jablons, D.; Provencio, M.; et al. Three-gene expression signature predicts survival in early-stage squamous cell carcinoma of the lung. *Clin. Cancer Res.* **2008**, *14*, 4794–4799. [[CrossRef](#)]
10. Xu, F.; Lin, H.; He, P.; He, L.; Chen, J.; Lin, L.; Chen, Y. A TP53-associated gene signature for prediction of prognosis and therapeutic responses in lung squamous cell carcinoma. *Oncoimmunology* **2020**, *9*, 1731943. [[CrossRef](#)] [[PubMed](#)]
11. Qu, H.Q.; Tien, M.; Polychronakos, C. Statistical significance in genetic association studies. *Clin. Investig. Med. Med. Clin. Exp.* **2010**, *33*, E266. [[CrossRef](#)] [[PubMed](#)]
12. Benjamini, Y.; Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B (Methodol.)* **1995**, *57*, 289–300. [[CrossRef](#)]
13. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B (Methodol.)* **1996**, *58*, 267–288. [[CrossRef](#)]
14. Chen, J.W.; Dhahbi, J. Lung adenocarcinoma and lung squamous cell carcinoma cancer classification, biomarker identification, and gene expression analysis using overlapping feature selection methods. *Sci. Rep.* **2021**, *11*, 13323. [[CrossRef](#)] [[PubMed](#)]
15. Fan, J.; Lv, J. Sure independence screening for ultrahigh dimensional feature space. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)* **2008**, *70*, 849–911. [[CrossRef](#)] [[PubMed](#)]
16. Li, R.; Zhong, W.; Zhu, L. Feature screening via distance correlation learning. *J. Am. Stat. Assoc.* **2012**, *107*, 1129–1139. [[CrossRef](#)]
17. Balasubramanian, K.; Sriperumbudur, B.; Lebanon, G. *Ultrahigh Dimensional Feature Screening via RKHS Embeddings*; Artificial Intelligence and Statistics: Scottsdale, AZ, USA, 2013; pp. 126–134.
18. Mai, Q.; Zou, H. The fused Kolmogorov filter: A nonparametric model-free screening method. *Ann. Stat.* **2015**, *43*, 1471–1497. [[CrossRef](#)]
19. Cui, H.; Li, R.; Zhong, W. Model-free feature screening for ultrahigh dimensional discriminant analysis. *J. Am. Stat. Assoc.* **2015**, *110*, 630–641. [[CrossRef](#)] [[PubMed](#)]
20. Beyene, J.; Atenafu, E.G.; Hamid, J.S.; To, T.; Sung, L. Determining relative importance of variables in developing and validating predictive models. *BMC Med. Res. Methodol.* **2009**, *9*, 64. [[CrossRef](#)] [[PubMed](#)]
21. Heinzl, A.; Perco, P.; Mayer, G.; Oberbauer, R.; Lukas, A.; Mayer, B. From molecular signatures to predictive biomarkers: Modeling disease pathophysiology and drug mechanism of action. *Front. Cell Dev. Biol.* **2014**, *2*, 37. [[CrossRef](#)] [[PubMed](#)]
22. Fan, J.; Feng, Y.; Wu, Y. High-dimensional variable selection for Cox’s proportional hazards model. In *Borrowing Strength: Theory Powering Applications—A Festschrift for Lawrence D. Brown*; Institute of Mathematical Statistics: Hayward, CA, USA, 2010; pp. 70–86.
23. Zhao, S.D.; Li, Y. Principled sure independence screening for Cox models with ultra-high-dimensional covariates. *J. Multivar. Anal.* **2012**, *105*, 397–411. [[CrossRef](#)] [[PubMed](#)]
24. Gorst-Rasmussen, A.; Scheike, T. Independent screening for single-index hazard rate models with ultrahigh dimensional features. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)* **2013**, *75*, 217–245. [[CrossRef](#)]
25. Iuliano, A.; Occhipinti, A.; Angelini, C.; De Feis, I.; Liò, P. Combining pathway identification and breast cancer survival prediction via screening-network methods. *Front. Genet.* **2018**, *9*, 206. [[CrossRef](#)]
26. Tschodu, D.; Ulm, B.; Bendrat, K.; Lippoldt, J.; Gottheil, P.; Käs, J.A.; Niendorf, A. Comparative analysis of molecular signatures reveals a hybrid approach in breast cancer: Combining the Nottingham Prognostic Index with gene expressions into a hybrid signature. *PLOS ONE* **2022**, *17*, e0261035. [[CrossRef](#)]
27. Zhang, R.; Chen, C.; Dong, X.; Shen, S.; Lai, L.; He, J.; You, D.; Lin, L.; Zhu, Y.; Huang, H.; et al. Independent validation of early-stage non-small cell lung cancer prognostic scores incorporating epigenetic and transcriptional biomarkers with gene-gene interactions and main effects. *Chest* **2020**, *158*, 808–819. [[CrossRef](#)] [[PubMed](#)]
28. Zhao, K.; Li, Z.; Tian, H. Twenty-gene-based prognostic model predicts lung adenocarcinoma survival. *Oncotargets Ther.* **2018**, *11*, 3415. [[CrossRef](#)]
29. He, X.; Wang, L.; Hong, H.G. Quantile-adaptive model-free variable screening for high-dimensional heterogeneous data. *Ann. Stat.* **2013**, *41*, 342–369. [[CrossRef](#)]
30. Song, R.; Lu, W.; Ma, S.; Jessie Jeng, X. Censored rank independence screening for high-dimensional survival data. *Biometrika* **2014**, *101*, 799–814. [[CrossRef](#)] [[PubMed](#)]
31. Li, J.; Zheng, Q.; Peng, L.; Huang, Z. Survival impact index and ultrahigh-dimensional model-free screening with survival outcomes. *Biometrics* **2016**, *72*, 1145–1154. [[CrossRef](#)]
32. Hong, H.G.; Chen, X.; Christiani, D.C.; Li, Y. Integrated powered density: Screening ultrahigh dimensional covariates with survival outcomes. *Biometrics* **2018**, *74*, 421–429. [[CrossRef](#)] [[PubMed](#)]
33. Chen, X.; Chen, X.; Wang, H. Robust feature screening for ultra-high dimensional right censored data via distance correlation. *Comput. Stat. Data Anal.* **2018**, *119*, 118–138. [[CrossRef](#)]

34. Hong, H.G.; Li, Y. Feature selection of ultrahigh-dimensional covariates with survival outcomes: A selective review. *Appl. Math.* **2017**, *32*, 379–396. [[CrossRef](#)] [[PubMed](#)]
35. Ke, C.; Yin, X. Expected Conditional Characteristic Function-based Measures for Testing Independence. *J. Am. Stat. Assoc.* **2020**, *115*, 985–996. [[CrossRef](#)]
36. Botling, J.; Edlund, K.; Lohr, M.; Hellwig, B.; Holmberg, L.; Lambe, M.; Berglund, A.; Ekman, S.; Bergqvist, M.; Pontén, F.; et al. Biomarker Discovery in Non-Small Cell Lung Cancer: Integrating Gene Expression Profiling, Meta-analysis, and Tissue Microarray Validation Gene Expression-Based Biomarker Discovery in NSCLC. *Clin. Cancer Res.* **2013**, *19*, 194–204. [[CrossRef](#)] [[PubMed](#)]
37. Rousseaux, S.; Debernardi, A.; Jacquiau, B.; Vitte, A.L.; Vesin, A.; Nagy-Mignotte, H.; Moro-Sibilot, D.; Brichon, P.Y.; Lantuejoul, S.; Hainaut, P.; et al. Ectopic activation of germline and placental genes identifies aggressive metastasis-prone lung cancers. *Sci. Transl. Med.* **2013**, *5*, 186ra66. [[CrossRef](#)]
38. Fukumizu, K.; Gretton, A.; Lanckriet, G.R.; Schölkopf, B.; Sriperumbudur, B.K. Kernel Choice and Classifiability for RKHS Embeddings of Probability Distributions. In *Advances in Neural Information Processing Systems 22*; Bengio, Y., Schuurmans, D., Lafferty, J.D., Williams, C.K.I., Culotta, A., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2009; pp. 1750–1758.
39. Sejdinovic, D.; Sriperumbudur, B.; Gretton, A.; Fukumizu, K. Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *Ann. Stat.* **2013**, *41*, 2263–2291. [[CrossRef](#)]
40. Silverman, B.W. *Density Estimation for Statistics and Data Analysis*; CRC Press: Boca Raton, FL, USA, 1986.
41. Ren, J.; Du, Y.; Li, S.; Ma, S.; Jiang, Y.; Wu, C. Robust network-based regularization and variable selection for high-dimensional genomic data in cancer prognosis. *Genet. Epidemiol.* **2019**, *43*, 276–291. [[CrossRef](#)]
42. Dwight, T.; Na, U.; Kim, E.; Zhu, Y.; Richardson, A.L.; Robinson, B.G.; Tucker, K.M.; Gill, A.J.; Benn, D.E.; Clifton-Bligh, R.J.; et al. Analysis of SDHAF3 in familial and sporadic pheochromocytoma and paraganglioma. *BMC Cancer* **2017**, *17*, 497. [[CrossRef](#)]
43. Moreno, C.; Santos, R.M.; Burns, R.; Zhang, W.C. Succinate Dehydrogenase and Ribonucleic Acid Networks in Cancer and Other Diseases. *Cancers* **2020**, *12*, 3237. [[CrossRef](#)]
44. Tseng, P.L.; Wu, W.H.; Hu, T.H.; Chen, C.W.; Cheng, H.C.; Li, C.F.; Tsai, W.H.; Tsai, H.J.; Hsieh, M.C.; Chuang, J.H.; et al. Decreased succinate dehydrogenase B in human hepatocellular carcinoma accelerates tumor malignancy by inducing the Warburg effect. *Sci. Rep.* **2018**, *8*, 3081. [[CrossRef](#)]
45. Guo, X.; Li, D.; Wu, Y.; Chen, Y.; Zhou, X.; Wang, X.; Huang, X.; Li, X.; Yang, H.; Xing, J. Genetic variants in genes of tricarboxylic acid cycle key enzymes are associated with prognosis of patients with non-small cell lung cancer. *Lung Cancer* **2015**, *87*, 162–168. [[CrossRef](#)] [[PubMed](#)]
46. Campbell, R.; Chong, G.; Hawkes, E.A. Novel indications for Bruton’s tyrosine kinase inhibitors, beyond hematological malignancies. *J. Clin. Med.* **2018**, *7*, 62. [[CrossRef](#)] [[PubMed](#)]
47. Albano, F.; Chiurazzi, F.; Mimmi, S.; Vecchio, E.; Pastore, A.; Cimmino, C.; Frieri, C.; Iaccino, E.; Pisano, A.; Golino, G.; et al. The expression of inhibitor of bruton’s tyrosine kinase gene is progressively up regulated in the clinical course of chronic lymphocytic leukaemia conferring resistance to apoptosis. *Cell Death Dis.* **2018**, *9*, 13. [[CrossRef](#)] [[PubMed](#)]
48. Tian, S. Identification of monotonically differentially expressed genes for non-small cell lung cancer. *BMC Bioinform.* **2019**, *20*, 177. [[CrossRef](#)] [[PubMed](#)]
49. Mardakheh, F.K.; Self, A.; Marshall, C.J. RHO binding to FAM65A regulates Golgi reorientation during cell migration. *J. Cell Sci.* **2016**, *129*, 4466–4479. [[CrossRef](#)] [[PubMed](#)]
50. Ridley, A. RhoA, RhoB and RhoC have different roles in cancer cell migration. *J. Microsc.* **2013**, *251*, 242–249. [[CrossRef](#)]
51. Howrylak, J.A.; Moll, M.; Weiss, S.T.; Raby, B.A.; Wu, W.; Xing, E.P. Gene expression profiling of asthma phenotypes demonstrates molecular signatures of atopy and asthma control. *J. Allergy Clin. Immunol.* **2016**, *137*, 1390–1397. [[CrossRef](#)]
52. Kachuri, L.; Johansson, M.; Rashkin, S.R.; Graff, R.E.; Bossé, Y.; Manem, V.; Caporaso, N.E.; Landi, M.T.; Christiani, D.C.; Vineis, P.; et al. Immune-mediated genetic pathways resulting in pulmonary function impairment increase lung cancer susceptibility. *Nat. Commun.* **2020**, *11*, 27. [[CrossRef](#)]
53. Poteryaev, D.; Datta, S.; Ackema, K.; Zerial, M.; Spang, A. Identification of the switch in early-to-late endosome transition. *Cell* **2010**, *141*, 497–508. [[CrossRef](#)]
54. Shao, X.; Liu, Y.; Yu, Q.; Ding, Z.; Qian, W.; Zhang, L.; Zhang, J.; Jiang, N.; Gui, L.; Xu, Z.; et al. Numb regulates vesicular docking for homotypic fusion of early endosomes via membrane recruitment of Mon1b. *Cell Res.* **2016**, *26*, 593–612. [[CrossRef](#)]
55. Jiang, L.; Qian, J.; Yang, Y.; Fan, Y. Knockdown of MON1B Exerts Anti-Tumor Effects in Colon Cancer In Vitro. *Med. Sci. Monit. Int. Med. J. Exp. Clin. Res.* **2018**, *24*, 7710–7718. [[CrossRef](#)] [[PubMed](#)]
56. Xuan, C.; Wang, Q.; Han, X.; Duan, Y.; Li, L.; Shi, L.; Wang, Y.; Shan, L.; Yao, Z.; Shang, Y. RBB, a novel transcription repressor, represses the transcription of HDM2 oncogene. *Oncogene* **2013**, *32*, 3711–3721. [[CrossRef](#)] [[PubMed](#)]
57. Kheder, E.S.; Hong, D.S. Emerging Targeted Therapy for Tumors with NTRK Fusion Proteins Novel Targeted Therapy for NTRK-Rearranged Tumors. *Clin. Cancer Res.* **2018**, *24*, 5807–5814. [[CrossRef](#)] [[PubMed](#)]
58. Jones, D.T.; Hutter, B.; Jäger, N.; Korshunov, A.; Kool, M.; Warnatz, H.J.; Zichner, T.; Lambert, S.R.; Ryzhova, M.; Quang, D.A.K.; et al. Recurrent somatic alterations of FGFR1 and NTRK2 in pilocytic astrocytoma. *Nat. Genet.* **2013**, *45*, 927–932. [[CrossRef](#)] [[PubMed](#)]
59. Amatu, A.; Sartore-Bianchi, A.; Siena, S. NTRK gene fusions as novel targets of cancer therapy across multiple tumour types. *ESMO Open* **2016**, *1*, e000023. [[CrossRef](#)]

60. Zhang, D.; Xi, Y.; Coccimiglio, M.L.; Mennigen, J.A.; Jonz, M.G.; Ekker, M.; Trudeau, V.L. Functional prediction and physiological characterization of a novel short trans-membrane protein 1 as a subunit of mitochondrial respiratory complexes. *Physiol. Genom.* **2012**, *44*, 1133–1140. [[CrossRef](#)]
61. Mullin, B.H.; Zhu, K.; Brown, S.J.; Mullin, S.; Tickner, J.; Pavlos, N.J.; Dudbridge, F.; Xu, J.; Walsh, J.P.; Wilson, S.G. Genetic regulatory mechanisms in human osteoclasts suggest a role for the STMP1 and DCSTAMP genes in Paget's disease of bone. *Sci. Rep.* **2019**, *9*, 1052. [[CrossRef](#)]
62. Feingold, E.; Pachter, L. The ENCODE (ENCyclopedia of DNA elements) project. *Science* **2004**, *306*, 636–640.
63. Consortium, E.P. A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biol.* **2011**, *9*, e1001046.
64. Rouillard, A.D.; Gundersen, G.W.; Fernandez, N.F.; Wang, Z.; Monteiro, C.D.; McDermott, M.G.; Ma'ayan, A. The harmonizome: A collection of processed datasets gathered to serve and mine knowledge about genes and proteins. *Database* **2016**, *2016*, baw100. [[CrossRef](#)]
65. Trimarchi, J.M.; Lees, J.A. Sibling rivalry in the E2F family. *Nat. Rev. Mol. Cell Biol.* **2002**, *3*, 11–20. [[CrossRef](#)] [[PubMed](#)]
66. Kar, A.; Gutierrez-Hartmann, A. Molecular mechanisms of ETS transcription factor-mediated tumorigenesis. *Crit. Rev. Biochem. Mol. Biol.* **2013**, *48*, 522–543. [[CrossRef](#)] [[PubMed](#)]
67. Hanahan, D.; Weinberg, R.A. Hallmarks of cancer: The next generation. *Cell* **2011**, *144*, 646–674. [[CrossRef](#)] [[PubMed](#)]