Cancer Genetics 235–236 (2019) 1–12

Cancer Genetics

ORIGINAL ARTICLE

A comparison of survival analysis methods for cancer gene expression RNA-Sequencing data

Pichai Raman[a,b,h], Samuel Zimmerman[c], Komal S. Rathi[b,h], Laurence de Torrenté[c,1], Mahdi Sarmady[d,i], Chao Wu[d], Jeremy Leipzig[d,e], Deanne M. Taylor[b,j], Aydin Tozeren[a], Jessica C. Mar[c,f,g,*]

[a] *School of Biomedical Engineering, Sciences and Health Systems, Drexel University, Philadelphia, PA, United States;* [b] *Department of Biomedical and Health Informatics, The Children's Hospital of Philadelphia, Philadelphia, PA, United States;* [c] *Department of Systems and Computational Biology, Albert Einstein College of Medicine, Bronx, NY, United States;* [d] *Division of Genomic Diagnostics, Department of Pathology and Laboratory Medicine, The Children's Hospital of Philadelphia, Philadelphia, PA, United States;* [e] *College of Computing and Informatics, Drexel University, Philadelphia, PA, United States;* [f] *Department of Epidemiology and Population Health, Albert Einstein College of Medicine, Bronx, NY, United States;* [g] *Australian Institute for Bioengineering and Nanotechnology, The University of Queensland, Brisbane, Australia;* [h] *Center for Data-Driven Discovery in Biomedicine, Children's Hospital of Philadelphia, Philadelphia, PA, United States;* [i] *Department of Pathology and Laboratory Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA;* [j] *The Department of Pediatrics, The University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, USA*

Abstract

Identifying genetic biomarkers of patient survival remains a major goal of large-scale cancer profiling studies. Using gene expression data to predict the outcome of a patient's tumor makes biomarker discovery a compelling tool for improving patient care. As genomic technologies expand, multiple data types may serve as informative biomarkers, and bioinformatic strategies have evolved around these different applications. For categorical variables such as a gene's mutation status, biomarker identification to predict survival time is straightforward. However, for continuous variables like gene expression, the available methods generate highly-variable results, and studies on best practices are lacking. We investigated the performance of eight methods that deal specifically with continuous data. K-means, Cox regression, concordance index, D-index, 25th–75th percentile split, median-split, distribution-based splitting, and KaplanScan were applied to four RNA-sequencing (RNA-seq) datasets from the Cancer Genome Atlas. The reliability of the eight methods was assessed by splitting each dataset into two groups and comparing the overlap of the results. Gene sets that had been identified from the literature for a specific tumor type served as positive controls to assess the accuracy of each biomarker using receiver operating characteristic (ROC) curves. Artificial RNA-Seq data were generated to test the robustness of these methods under fixed levels of gene expression noise. Our results show that methods based on dichotomizing tend to have consistently poor performance while C-index, D-index, and k-means perform well in most settings. Overall, the Cox regression method had the strongest performance based on tests of accuracy, reliability, and robustness.

Keywords Survival analysis, Kaplan–Meier, TCGA, Cancer, Gene expression.
© 2019 Elsevier Inc. All rights reserved.

List of Abbreviations: TCGA, The Cancer Genome Atlas; OV, ovarian serous cystadenocarcinoma; HNSC, head and neck squamous cell carcinoma; PRAD, prostate adenocarcinoma; KIRC, kidney renal clear cell carcinoma; expo, Expression Oncology Project; AUC, Area Under the Curve; FPR, False positive rate; TPR, True positive rate; ROC, Receiver operating characteristic.

Received November 15, 2018; received in revised form March 19, 2019; accepted April 9, 2019

2210-7762/$ - see front matter © 2019 Elsevier Inc. All rights reserved.
https://doi.org/10.1016/j.cancergen.2019.04.004

* Corresponding author at: Australian Institute for Bioengineering and Nanotechnology, The University of Queensland, Brisbane, Australia.

E-mail addresses: ramanp@email.chop.edu, sezimmer@einstein.yu.edu, rathik@email.chop.edu, ldetorrente@nygenome.org, sarmadym@email.chop.edu, wuc8@email.chop.edu, leipzig@panoramamedicine.com, taylordm@email.chop.edu, at62@drexel.edu, j.mar@uq.edu.au

[1] Present address: New York Genome Center, New York, NY, United States.

Introduction

All large-scale cancer profiling studies consistently share a common feature that support the notion that cancer is a complex, heterogeneous genetic disease. Collections of genome-wide transcriptional profiling datasets captured using technologies like microarrays and RNA-sequencing (RNA-seq) have improved our insight into the nature of this heterogeneity for many different tumor types. Through efforts such as the Cancer Genome Atlas (TCGA) [1], the Expression Oncology Project (expo) [2], and projects stemming from the International Cancer Genome Consortium (ICGC) [3], gene expression datasets are publicly accessible with some collections even containing thousands of patient-derived tumor samples. From these vast resources, we have come to recognize that despite the heterogeneity that exists in cancer profiles, dominant features can be extracted from the data and represent diagnostic capacity in the form of gene signatures and biomarkers. To patient care, these gene sets are valuable because they can be used to predict key properties of a tumor, such as grade and molecular subtype, or the expected length of survival time of a patient. Therefore, the most significant goal for bioinformatics, as a field, is to develop clear practices for how to best identify those genes that are predictive markers of survival status of cancer patients.

Studies of genetic variation in tumors have identified specific genes and their associated lesions that have been used to delineate distinct patient sub-groups with prior success. Some examples include, the PTCH1 inactivating mutation in medulloblastoma [4], MYCN amplification in neuroblastoma [5], or KRAS mutation status in non-small cell lung cancer [6]. While these lesions have been valuable for identifying genetic regulators of cancer, it is becoming increasingly clear that given the complex and graded nature in which tumors are controlled by genes, there are advantages to expanding our focus to gene expression-based markers as well. Instead of dichotomous or binary associations that are represented by genetic variants or copy number aberrations, identifying predictors of patient survival using gene expression may result in markers that can predict change between these two variables with greater sensitivity or more subtle degrees of detection.

To identify predictive biomarkers, a host of statistical methods have been adapted from survival analysis techniques, including the Kaplan-Meier estimator, the log-rank test, or the Cox regression model. In general, these methods are geared towards binary inputs like the presence or absence of a mutation, lesion, gene fusion, translocation or other genetic event that is discrete. Standard workflows exist for handling this binary data, and identification of these biomarkers are considered routine. On the other hand, the equivalent framework for continuous inputs, like data based on gene expression is not so well-established. This is because it is not often clear what the optimal choice is for estimating the relationship between a gene's expression profile and survival status in a patient cohort. For instance, a standard survival analysis model could be used to assess the degree to which changes are occurring between gene expression and patient survival time, but general assumptions regarding regulatory conditions of the data may not always hold, especially those based on the distribution of the data, such as a Normal distribution. Alternatively, a continuous variable may be transformed into a binary one using a threshold, and standard workflows for dichotomizing data can be applied. Although these approaches are attractive because of their simplicity, the specification of an appropriate threshold is not a trivial question to solve. Fundamentally, the issue that complicates biomarker prediction for patient survival, and prevents its straightforward expansion from discrete variables to continued-based ones, like gene expression, is the degree of variability in the data. Simply put, a question that remains unaddressed is whether it is more effective to derive biomarkers based on approaches that dichotomize a gene's expression profile, and if so, how can an optimal breakpoint be identified to facilitate this dichotomization?

To our knowledge, a comprehensive investigation to address this important question has not been performed, and therefore is the focus of this study. The performance of different statistical methods that estimate the effect of gene expression from RNA-seq data and survival status were compared to determine what is the optimal strategy for identifying predictive markers of cancer patient survival. RNA-seq datasets from four TCGA studies - ovarian serous cystadenocarcinoma (OV), prostate adenocarcinoma (PRAD), kidney renal clear cell carcinoma (KIRC), and head and neck cancer squamous cell carcinoma (HNSC) – were used to test the reliability and accuracy of eight competing survival analysis methods. The methods selected were based on the Cox regression model, k-means, the concordance index (C-index), the D-index, and dichotomization using the median, distributional shapes, KaplanScan, and the 25th–75th percentile split. The four cancers were selected as a representative panel to evaluate the eight survival analysis methods. Given that there are over 30 tumor types represented in TCGA now, a large number of combinations could have been sourced. However, the ovarian and prostate cancers were selected to represent two different kinds of sex-specific tumors. The head and neck and kidney cancers were also included because they represent less common tumor types (∼4% of all new cancer cases in the United States) [7].

As with any applied statistical approach, the eight methods selected each bring their own set of advantages and limitations for finding biomarkers of survival based on gene expression data (Fig. 1). Cox regression is a flexible, well-established method that allows for the inclusion of multiple covariates to adjust for explanatory variables. This provides a way to further improve the accuracy of the estimate between patient survival and gene expression by accounting for other contributing factors, including batch effects, biometric or clinical variables. K-means has been borrowed from exploratory data analysis methods, where k-means clustering is used to split the gene expression data into two patient groups through an unsupervised, non-parametric approach that iterates until the method arrives at convergence. A log-rank test assesses the difference in survival for these two patient groups to determine a gene's utility as a prognostic biomarker. The KaplanScan method identifies the optimal breakpoint by considering multiple candidates, and chooses the one that creates the most significant separation between the two patient groups. While the KaplanScan method is advantageous because it avoids relying on an arbitrary threshold for dichotomization, it suffers from an increased rate of false positives, and correction methods to adjust for multiple hypothesis must be used [8].

In our study, two of the methods that were selected are those based on quantile dichotomization. The simplest

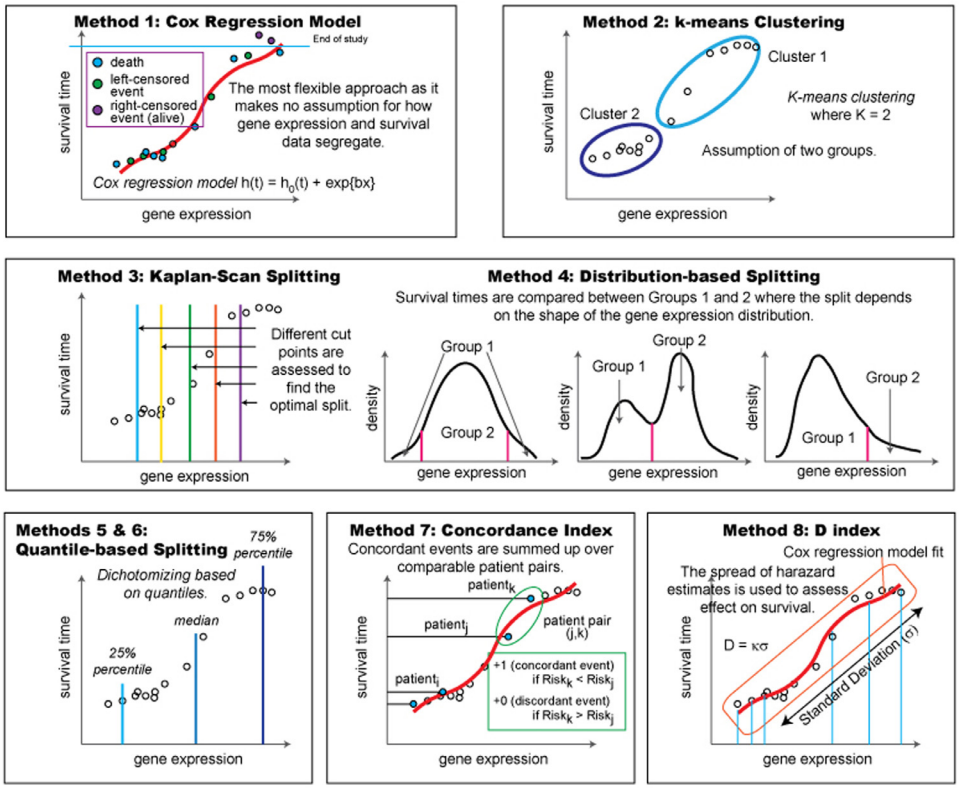

Fig. 1 Schematic outline of the eight methods used to identify gene expression-based cancer biomarkers.

version uses the median as the breakpoint to create the two patient groups, and a related second version of this approach is instead based on a 25th, 75th percentile split [9]. To account for genes that may have a diverse range of distributions that are not necessarily Normally-distributed, we also included a method based on the shape of the gene expression distribution, termed, distribution-based splitting. For instance, if a gene's expression profile is Normally-distributed, then patient sub-groups are instinctively formed based on the most extreme 5th percentiles of the gene expression distribution. Alternatively, if the gene is asymmetric or bimodally-distributed then the sub-groups are determined based on only one side of the distribution, or using the two separate modes, respectively. This method emerged from a novel investigation into the prevalence of different distribution shapes in cancer RNA-seq data [10].

In addition to statistical measures of performance, a worthwhile consideration of one method over another relates to its popularity and ease of application. Statistical software suites such as SPSS or SAS accommodate Cox regression or a median cut-off for survival analysis as standard inputs for generic data. Standalone tools, such as Cutoff Finder [11], an online suite of optimization and visualization routines for cutoff determination, are also available and offer a variety of survival analysis implementations. Both the k-means and distribution-based splitting approach are examples of methods that have been developed for other purposes, with their application to biomarker prediction for patient survival being a novel implementation.

To evaluate the performance of the eight methods, we focused on three aspects of performance: reliability, accuracy, and robustness. Reliability of a survival analysis method was assessed by splitting each cancer dataset in a stratified manner into two half-groups, performing survival analysis, and computing the correlation of results between the two data halves. For every gene, the P-value measuring the association between expression and patient survival was calculated using each of the eight methods. We computed the correlation between the negative $\log_{10}$(P-value) between the two partitions of a cancer dataset to assess reliability of each method. For instance, a dataset of 5000 genes and 150 samples (100 alive, 50 deceased) would be divided into two sets of 100 samples (each with 75 alive and 25 deceased). Survival analysis was performed for each of the 5000 genes on the first dataset and the second dataset independently. The correlation of each gene between the two sets was then examined to quantify the reliability of the method for all genes in the cancer dataset.

Accuracy of a survival analysis method was assessed using analysis of ROC curves based on gene sets that were identified from GeneSigDB [12] as relevant, tumor type-specific positive controls. The ROC curve describes the relationship between the proportion of true positives (sensitivity) and false positives (1 - specificity) resulting from each possible decision threshold value in a two-class classification problem. The overall curve formed from the different classification points is informative of the accuracy of the method, specifically the area under this curve which is used as a quantitative

measure of how sensitive a method is at detecting true positives while avoiding as many false positives as possible.

Robustness of a survival method was assessed by generating *in silico* data where fixed amounts of noise were added and accuracy assessed using ROC curves. The *in silico* datasets were generated with controlled levels of noise and the same set of positive controls. The AUC was used to assess the accuracy of a method at differing levels of noise to investigate how effective a method was at identifying the positive controls in the data in the presence of increasing noise.

Materials and methods

Gene expression datasets

RNA-sequencing datasets were obtained from the Cancer Genome Atlas (TCGA) using the Broad GDAC FIREHOSE utility [13], and the corresponding patient survival data was obtained for kidney cancer, ovarian cancer, prostate cancer, and head and neck cancers. Genes that served as positive controls for assessing accuracy across the eight methods were obtained from the database genesigdb [12]. This resource contained curated sets of gene signatures from the literature, and gene lists that were relevant to survival in the specific type of tumor were included in our study. To ensure the strongest relevance and specificity to biomarkers of survival, signatures in genesigdb that were associated with classifying cancer molecular subtypes, or were specific biomarkers of a molecular subtype, were discarded. The gene signature that we enlisted for the accuracy tests for ovarian cancer was a set of 115 genes from Spentzos et al. that is referred to as the Ovarian Cancer Prognostic Profiles (OCPP) [14]. This list was derived from computational analysis on gene expression data. For kidney cancer, a list of 259 genes that correlated with survival (post-surgery), from Zhao et al [15] was used. For head and neck cancer, a list of 42 genes from Chung et al. was derived computationally from gene expression data, that compared high-risk and low-risk patients [16]. Finally, for prostate cancer we used a list of 50 genes, from Henshall et al., which was obtained from a survival analysis of gene expression microarray data on a patient cohort [17]. It is important to highlight that the signatures that were included in our study were identified because they were found to be the most relevant one to the tumor type and patient survival, as opposed to a signature that were simply the most effective at discriminating between specific subtypes. ROC curves were visualized using the *ggplot2* R package [18] and Area Under the Curve (AUC) values were obtained using the *AUC* R package [19].

Simulated gene expression datasets

Artificial RNA-Seq count datasets were generated using the 'SimSeq' Package in R [20]. This package requires a set of source RNA-Seq data to model and mimic the distribution of counts. We used the head and neck cancer dataset as our source model as it had the largest number of patient samples, and hence represented the most reliable and robust source of data. Noise was generated from a Normally distribution and added to the simulated profile of each gene as a percentage of the mean of that gene across all patient samples. A total of 7 datasets were generated, consisting of 150 patient samples and 5000 genes, with controlled levels of simulated noise, ranging from 0 to 1.5 times the values of the mean. 250 positive controls were also generated by taking a random sample of genes and adding a random multiplier (sampled from a uniform distribution with the range defined from values 1 to 8) to all of those samples with simulated patient survival events for that gene. Specifically, for any given gene in the set of 250 positive controls, we took the samples with an event and multiplied them all by an integer sampled uniformly from the set [21]. The probability of being multiplied by one of these integers was set as uniform for this simulation.

Survival analysis methods

The KaplanScan method

Given the continuous nature of gene expression data, a scanning approach to the Kaplan–Meier method, called KaplanScan has been applied in many genomics-based applications [22–24] and is presented on the R2 Genomics web portal [25]. It is based on moving samples between two designated groups to identify the breakpoint or separation point that gives rise to the optimal P-value. For instance, for a given gene, the first step in this method requires the expression values to be sorted. Upon sorting, the bottom 5% were assigned to group 1, and the top 95% were assigned to group 2. This step serves to convert the continuous expression profile into a binary one for subsequent survival analysis. The log-rank test is then applied to evaluate the difference in survival between the two groups (group 1 and group 2), and a P-value is calculated. In the subsequent step, the smallest sample from group 2 is transferred to group 1 and the log-rank test is run again. This transfer of samples continues iteratively until group 1 encapsulates the bottom 95% of values and group 2 holds the top 5% of values. The lowest P-value was then chosen as the optimal breakpoint of the two groups and represents the output of this iterative part of the analysis. A Benjamini–Hochberg correction is then performed on all the P-values generated from this scanning approach to account for the multiple hypotheses tests incurred in this scheme. Both the original and corrected P-values can be returned at the optimal breakpoint (lowest P-value) for each gene.

Dichotomizing based on quantiles

Two versions were tested, one using median-based dichotomization, and two, using breakpoints determined by the 25th percentile and 75th percentile, creating three partitions where the middle partition is not included in the survival analysis.

k-means

k-means is a standard clustering method that partitions the data points into K groups where K is a pre-specified number. There are different ways to implement k-means, and for our particular implementation, each iteration of the algorithm, involved assigned every patient sample to one of two clusters based on whose mean or centroid minimizes the within-cluster sum of squares. This is followed by an update on the cluster centroids as the memberships of the samples change. The algorithm finally converges when the membership of

genes in the two clusters is stable and no longer changes. We specify K=2 to dichotomize the continuous variable into 2 separate patient groups. Standard survival analysis is then run on the binary, transformed gene expression data where the covariate represents the cluster membership identified by the aforementioned k-means algorithm.

Cox regression

A common method that does not require dichotomizing a variable *a priori* is Cox regression. This model is one of the most commonly used statistical methods for survival analysis. This model provides an estimate of treatment effect on survival after adjustment for other explanatory variables. In addition, it allows for the estimation of the risk (or hazard) of death of an individual given their prognostic variables. The model is written as: $h(t)=h0(t) \times \exp(b1 \times 1+b2 \times 2+\cdots+bnxn)$ where $h(t)$ is the hazard function that estimates the risk at any given time t and is determined by a set of n covariates ($x1, x2, \ldots, xn$). The regression coefficients ($b1, b2, \ldots, bn$) represent the amount of adjustment in the proportional change in hazard due to the covariates. $h0$ is the baseline hazard function that corresponds to the probability of hazard when all covariates take on value equal to zero.

Distribution dichotomization method

The premise of this method is that genes may have different expression distributions in a patient cohort, and that testing for differences in patient survival time based on gene expression should therefore accommodate the shape or type of the distribution. For example, if a gene is symmetrically distributed, where it follows a Normal distribution then it would be a natural choice to compare the survival times of patients in both upper and lower tails of this distribution using a quantile cutoff. Alternatively, if the distribution is asymmetric, then a more sensible comparison may be to compare the survival times of the patients falling in the tail versus non-tail regions. Finally, if the gene's distribution is not unimodal but instead bimodal, then a much more natural comparison is between the patients in each mode of the distribution. We used a computational scheme that assesses the most likely distribution of a gene's expression profile by first considering bimodality through the Bimodal Index (BI) [26]. If the BI $>$ 1.1, the gene is designated bimodal and survival time is tested between the patients classified in one group/mode versus another. If the gene is not bimodal, the expression distribution is simultaneously tested for belonging to the Normal, Lognormal, Pareto, Gamma and Cauchy distribution and the gene is assigned to the distribution with the most significant P-value. For genes with a distribution that is either Normal or Cauchy, survival is tested for patients in the upper and lower tails versus the patients in the non-tail region. For the Gamma, Pareto and Lognormal distributions, survival is compared between the tail and non-tail regions. If all distributions tested were not significant then the gene was listed as having an unknown distribution and survival analysis was not performed.

For testing Normality and Lognormality, the Shapiro test from the R package stats (version 3.2.2) was used with a threshold of 0.01, and log transformation of the data respectively [27]. For Pareto, Gamma and Cauchy, the Kolmogorov-Smirnov test was applied [28]. For this test, the parameters were estimated with the Maximum Likelihood Estimates (MLE). For the MLE of Gamma, we used the rGammaGamma R package (version 1.0.12.). For Cauchy, the two parameters were set as the median and the interquartile range. As the parameters were estimated directly from the data, we applied a parametric bootstrap to estimate the final P-value. This idea of resampling to find the null distribution of the test statistics when estimating the parameters is based on the Lilliefors test [29]. The significance threshold for the final P-value was set to 0.01. For testing Bimodality, the Bimodality Index was computed using the R package ClassDiscovery (version 3.0.0.).

Concordance index (C-index) method

The C-index is an accuracy measure that is based on the probability of concordance between the predicted and observed responses [30,31]. Specifically, for a pair of patients chosen at random, the C-index determines the probability that the patient sample with the higher risk prediction will experience an event before the patient with the lower risk in the pair. This measure is computed for all pairs of observed responses and the number of times the predictions are concordant is summarized as a statistic from which the P-value is derived under assumptions of asymptotic Normality [32].

D-index method

The D-index is a statistical metric that assesses prognostic ability of a potential biomarker by measuring the degree of separation between two Kaplan-Meier curves constructed by splitting the patient population into two groups based on the biomarker [33]. Under assumptions made by the Cox proportional hazard model, the D-index is calculated as an estimate of the log hazard ratio between the two populations being compared. The C-index and D-index functions were computed using the survcomp package (https://github.com/bhklab/survcomp/tree/master/man) [34].

Results

The eight survival analysis methods were evaluated based on performance according to three criteria, reliability, accuracy, and robustness. First, reliability was assessed by dividing each cancer dataset into two sets, running the survival analysis method on both datasets to identify biomarkers, and the consistency of results between the two half datasets was compared. Second, accuracy of a method was assessed by comparing the results of each method to a gold standard list of known prognostic expression markers that were specific to each tumor type. We then computed ROC curves to compare the relative false positive rate (FPR) to true positive rate (TPR) of the eight methods. Third, robustness for each method was tested by generating *in silico* data with controlled levels of noise, and a set of known "positive controls". In this way, the performance of the eight methods was assessed in the presence of increasing amounts of noise in the data.

Assessment of reliability identified the k-means and Cox regression methods as those with the strongest performance.

Reliability was assessed by investigating whether a method would give similar results if only half of the data was presented to the method. Two sets of gene-specific $\log_{10}$-transformed P-values were obtained from applying the same method to two halves of the cancer datasets, and a correlation coefficient was used to evaluate the similarity in results

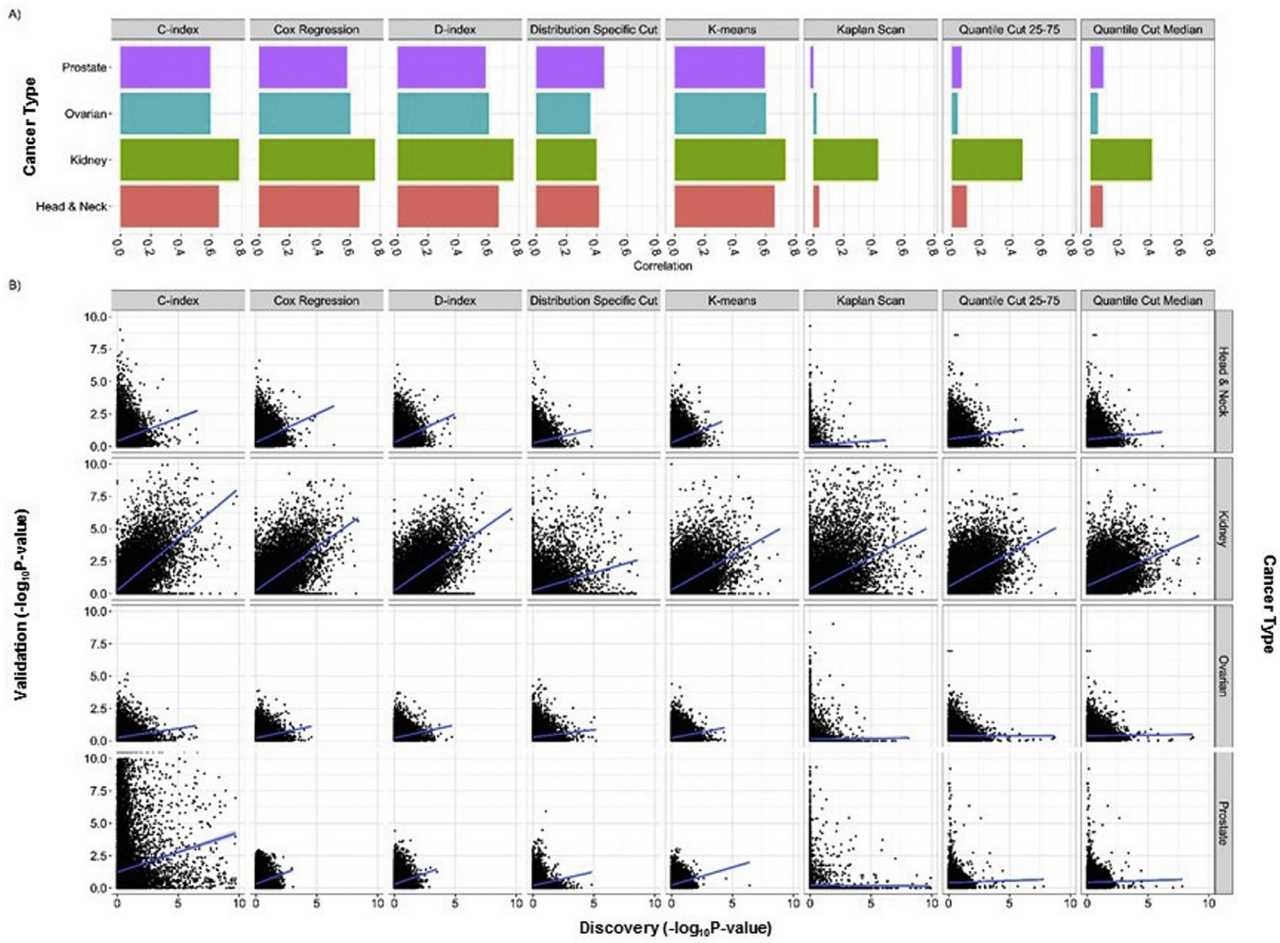

Fig. 2 Assessing the reliability of the eight methods. A. Bar chart displaying the Spearman correlation coefficients between the two data halves for each method demonstrate which methods are more reliable. **B.** Scatter plot of -$\log_{10}$(P-values) for all genes in one data half versus the other data half for each method indicate which methods identified that same set of significant genes in each data half. Blue lines indicate equality between the two halves.

to score the reliability of each method. Given that the distribution of P-values can vary between datasets, a Spearman's rank correlation coefficient (ρ) was used to quantify the similarity of results from the split datasets of each cancer type. Visualization of this correlation using a bar chart (Fig. 2A) presented a macroscopic view across the four cancer types, and a trellised scatter plot (Fig. 2B) demonstrated how the distribution of P-values aligned between the two halves of the datasets. For the four cancer types, the Cox regression, C-index, D-index, and k-means all had the strongest measures of reliability compared with the other methods ($0.6 < \rho < 0.8$).

The set of methods that were based on dichotomizing a variable, either using the median, the 25th–75th percentile, or KaplanScan fared the worst ($-0.02 < \rho < 0.43$). The reason for this clear contrast in the reliability of these methods may lie in the fact that in general, the median is not the most natural cut-point for gene expression profiles and splitting at the quartiles reduces power of detection leading to a higher rate of false positives. In a similar vein, the KaplanScan algorithm is likely to generate a larger number of false positives because it attempts to find an optimal cut-point based on the P-value of the difference of the dichotomized groups as opposed to the data itself. Performance based on reliability for the distribution-based splitting method was in between the two extremes of strong and weak performance ($0.36 < \rho < 0.45$). Given that this method is based on modeling that is more attuned to the shape of the gene expression profile, it avoided some of the issues of KaplanScan and dichotomized-based methods that were previously mentioned, and therefore had stronger performance in reliability. However, the distribution-based splitting method is not optimized for identifying survival-based biomarkers, and as a result, its performance based on reliability was sub-standard when compared to Cox regression, C-index or D-index. Furthermore, the flexibility of the k-means method, which makes no assumptions about distributions or associated parameters, is likely to be the factor underlying the stronger performance in reliability of this method versus the distribution-based splitting method.

Inspection of the genes that were identified as statistically significant by each method, showed that, k-means, Cox re-

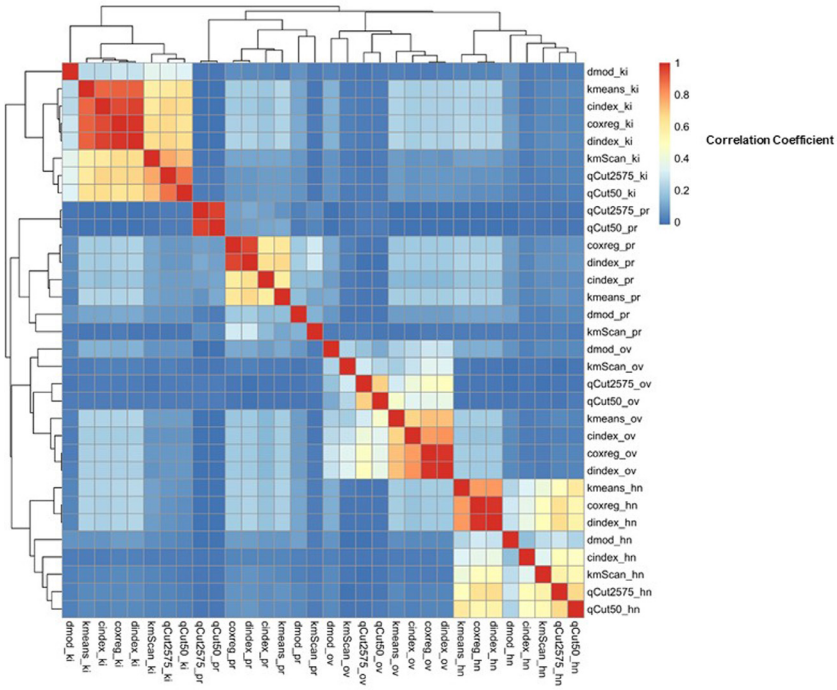

Fig. 3 Investigating similarities between the eight methods based on the set of significant genes that were identified for each cancer type. Heatmap showing the hierarchical clustering of the correlation coefficient of each method and cancer type based on -$\log_{10}$(P-values) to one another. The grouping indicates some methods identified similar sets of significant genes as gene expression-based biomarkers.

gression, C-index and D-index, had a similar set of significant genes (Fig. 3). The heatmap shows the correlation of the results of survival analysis across each of the cancer datasets (unsplit) for each method. While the groups initially partition by cancer type, the D-index and Cox regression results were always immediate neighbors in the same pair of the dendrogram (Fig. 3). Similarly, the two quantile-based methods (median, 25th–75th) paired together for all four cancer types. For ovarian cancer and prostate cancers, the distribution-specific method was paired with KaplanScan, and in kidney and head and neck cancers, the distribution-specific method was clustered in its own branch of the dendrogram (Fig. 3). More broadly, a clustering pattern was noted where the four methods, k-means, Cox regression, C-index, and D-index grouped together, whereas the KaplanScan, median and 25th–75th percentile-based methods tracked together separately in a different group.

Unsurprisingly, the type of cancer also seemed to have an effect on the reliability of the eight methods. Kidney cancer had the best reliability measures compared to all tumor types that were included in this study. This could be due to heterogeneity in the data where a specific tumor type may have a different number of molecular subtypes than other types of tumors. It is possible that for the four different tumor types that were tested in our study, a variable number of molecular subtypes were represented. A PCA using all available RNA-seq data indicated that at the genome-wide level, kidney cancer had a similar degree of heterogeneity since the spread between patient samples was comparable for all types of cancer (Supplemental Fig. 1). However, at the gene-specific level, results from the k-means method indicated that kidney cancer had greater variability in gene expression between the two clusters identified by k-means. Specifically, when the mean of expression for each cluster was calculated, the fold-change between these two means was largest in value for clusters that captured 50% of all patients (Fig. 4). For other tumor types, the largest fold-change values in mean cluster expression occurred for more skewed percentages, where clusters represented less than 25% of all patients. It seemed likely therefore that for the kidney cancer dataset, separation into more dominant pairs of clusters occurred with genes that showed a more extreme degree of different expression as compared to other tumor types that were included in this study. For survival analysis, when the expression data separates more definitively into groups with and without a patient survival event, it would certainly be easier to detect such a gene, independent of the method applied.

Accuracy assessment based on tumor type-specific positive controls demonstrated that Cox regression outperformed other survival analysis methods

The accuracy of the eight methods was assessed based on their ability to identify a set of tumor type-specific gene signatures which served as a surrogate for positive controls. ROC curves for each cancer dataset demonstrated highly variable performance (Fig. 5) and the area under the curve (AUC) values in Table 1 ranged from poor to good performance (0.479 to

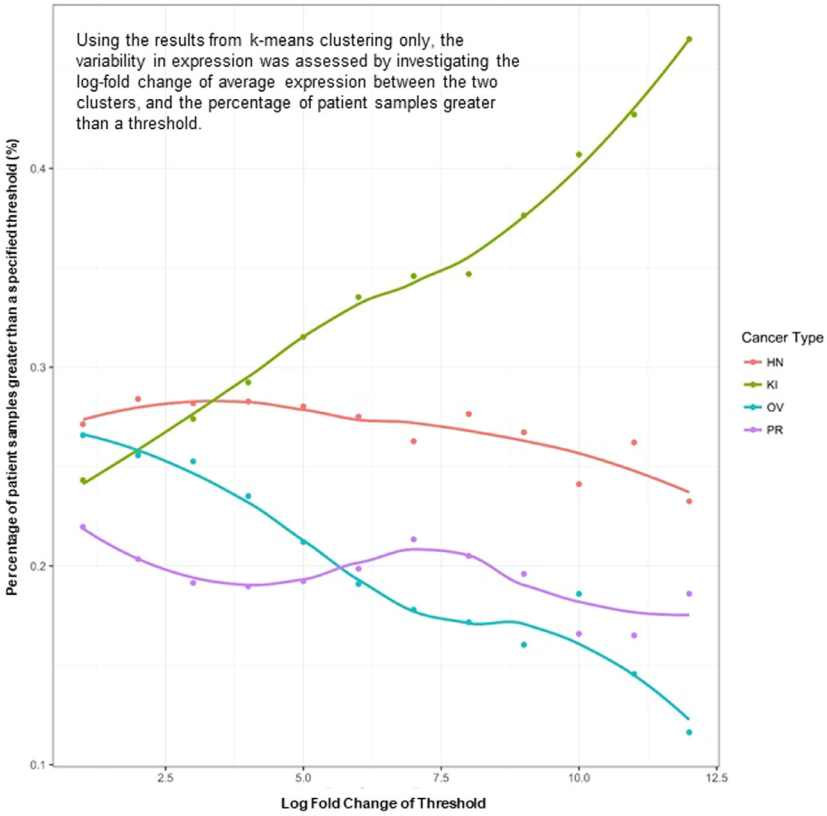

Fig. 4 Investigating the degree of variability in the different tumor types tested in this study. Using the results from k-means clustering, the variability in expression was assessed by investigating the log-fold change of average expression between the two clusters, and the percentage of patient samples greater than a threshold.

Table 1 Ranking of the eight methods based on accuracy. AUC values based on each method's ability to identify a set of positive controls that were tumor type-specific and derived from the literature, over a range of thresholds.

Method	Head & Neck	Kidney	Ovarian	Prostate	Mean	Rank Based on Mean
Cox Regression	7.0693E−01	6.9476E−01	6.7370E−01	6.1466E−01	6.7251E−01	1
K-Means	7.1827E−01	6.9364E−01	6.6247E−01	6.0593E−01	6.7008E−01	2
D-index	6.9920E−01	6.9004E−01	6.7344E−01	6.0891E−01	6.6790E−01	3
C-index	5.7950E−01	6.9928E−01	6.5399E−01	6.5675E−01	6.4738E−01	4
Median-Split	5.5054E−01	6.2663E−01	5.0335E−01	5.3615E−01	5.5417E−01	5
25th–75th Percentile Split	5.3204E−01	6.2769E−01	5.7235E−01	4.7884E−01	5.5273E−01	6
KaplanScan	5.3171E−01	6.1959E−01	5.0674E−01	5.3606E−01	5.4853E−01	7
Distribution-based Splitting	5.6100E−01	4.9976E−01	5.8779E−01	5.3929E−01	5.4696E−01	8

0.78). On average, the Cox regression method had the highest AUC value across all four cancer datasets (Table 1). For ovarian cancer, the Cox regression method had the highest AUC value, and the second-highest AUC value for the remaining three cancer types. Although the Cox regression had the highest average AUC value, the k-means, C-index, and the D-index also had similar performance in accuracy. Notably, the D-index had the highest AUC values for two cancer datasets (kidney and prostate cancers). In fact, the gains observed in AUC value with one method over another were generally quite small between Cox regression, k-means, C-index, and D-index (Table 1).

Indeed, an ANOVA on the AUC values shows that the type of method had a significant effect on the values ($P\text{-value} = 7.47 \times 10^{-5}$) and a post-hoc test using Tukey's HSD test confirmed that the k-means, C-index, D-index and Cox regression methods have significantly higher AUC values when compared individually to most other methods (4 out of 7 pairwise comparisons were statistically significant for each method, adjusted P-value < 0.05, Supplemental Table 1). In contrast, the KaplanScan, median-split, and 25th–75th percentile split did not show a significant difference in AUC values when compared with one another. Similarly, Cox regression and k-means themselves were not significantly different from one another in terms of their AUC, and the same was observed for C-index compared with D-index. This could have been because these methods generally yielded very similar

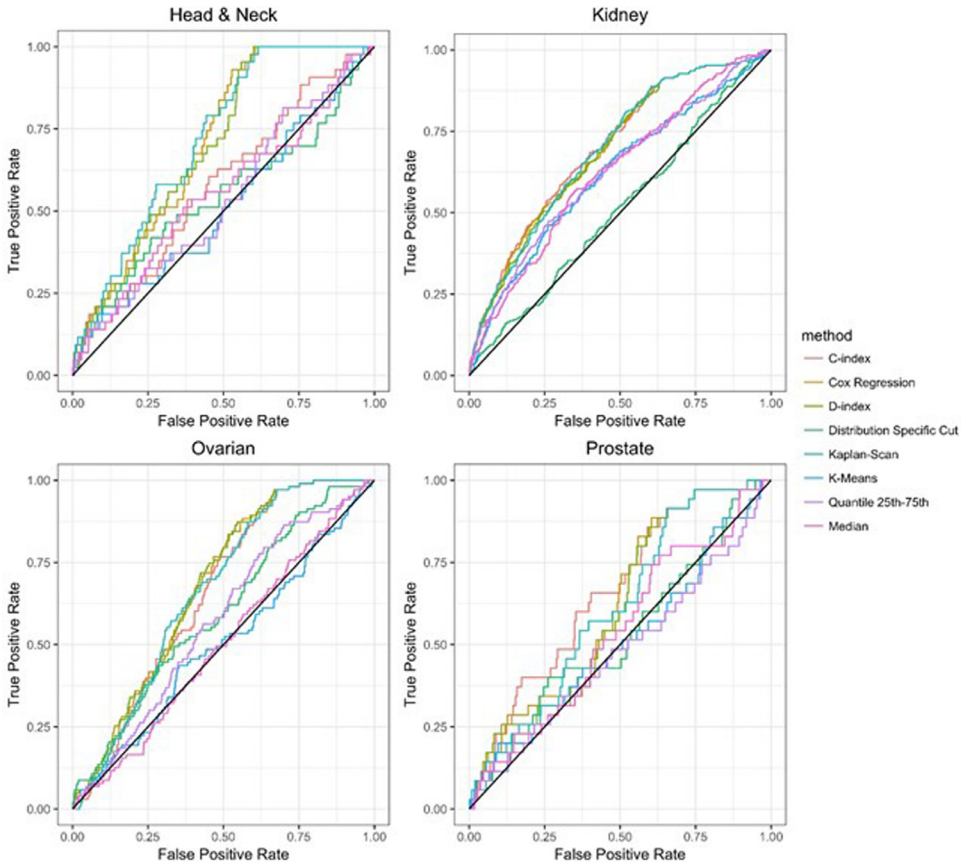

Fig. 5 Assessing the accuracy of the eight methods. ROC curves were used to evaluate the ability of each method to identify a set of positive controls that were tumor type-specific and derived from the literature.

results based on the genes that they had identified (Supplemental Table 1).

As to why the Cox regression method of survival analysis performed much better than the other methods it could be primarily due to the genes chosen by this method. While it is certainly possible that this could have been influenced by the "gold standard" or known true positive lists chosen, it seems unlikely that each of the cancer lists favor this method more strongly, with good performance also observed for the k-means, C-index, and D-index methods. Thus, it seems probable that the Cox regression method, and to a lesser extent, the k-means, C-index, and D-index methods simply perform better than the other methods because of their ability to identify true positives and false positives more robustly in the presence of noisy data.

Testing for robustness using *in silico* data identified the D-index as the method that was the least sensitive to different levels of noise

Overall, our tests identified D-index as the method that was the most robust to noise using the *in silico* datasets. As we varied the noise incrementally in our artificially simulated dataset, both the median and 25th–75th percentile-split methods increasingly lost their ability to detect signal in the data, and the distribution-specific method also dropped precipitously in performance (Fig. 6). The Cox regression method had the second most robust performance overall with the C-index coming third. The D-index had superior performance that was above the C-index and Cox regression for all noise levels, with AUC > 0.9 for maximum noise level of 1.5. This observation was consistent with the results that assessed accuracy using the ROC curves that were based on positive controls using real RNA-seq data. It is possible this result was due to the ability of the Cox regression method to model continuous data more robustly against noise than the other methods. In contrast to its fair performance in the previous evaluations, reliability and accuracy, k-means was not found to be very robust and did not perform well in the presence of noise, despite being a non-parametric approach. The method of creating positive controls, generating data, and even adding noise are certainly factors that could have caused or influenced this result.

Comparing the performance of the survival analysis methods using well-known cancer genes

It is useful to examine the results from this study through the lens of several well-known cancer genes for each tumor type. Anoctamin 1 (*ANO1*) expression has been implicated previously as a marker of poor prognosis and a potential driver of metastasis in HNSC [21,35]. When examining the TCGA HNSC dataset, we find that while almost all methods do detect *ANO1* as being significantly related to survival (P-value < 0.05); with Cox regression, C-index, and D-index assigned

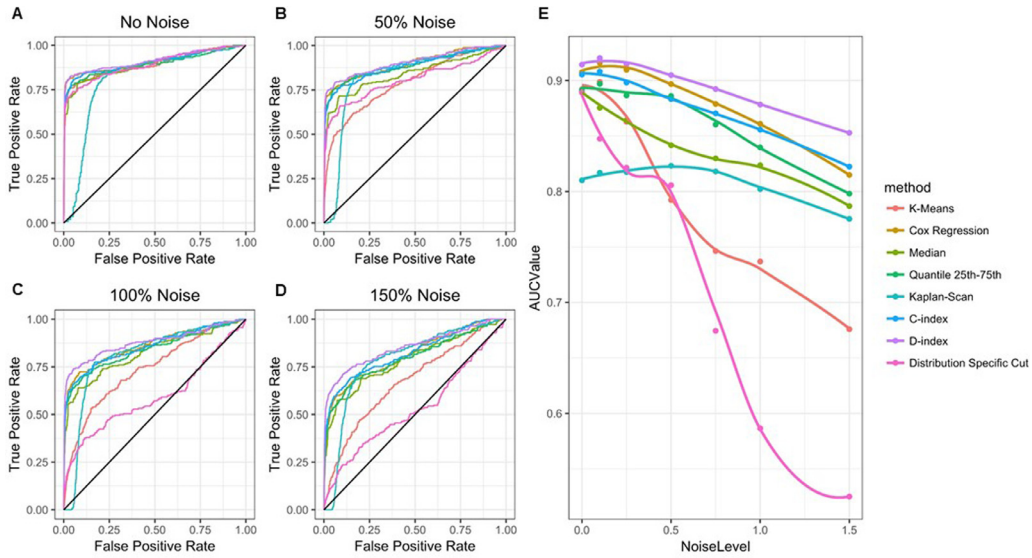

Fig. 6 Assessing the robustness of the eight methods. A - D. ROC curves demonstrate variable performance of the methods as the noise level in the simulated data increases from 0 to 150%. **E.** The trend in AUC values across increasing levels of noise indicate that the D-index has the most robust performance of all eight methods, followed by Cox regression.

to *ANO1* the most significant P-values (P-value < 0.005). In contrast, the dichotomization methods yielded much higher P-values, both approximately at P-value ~ 0.03. Since this is close to the significance threshold that is typically chosen in most studies (P-value < 0.05), it stands to reason that if the sample size was smaller it may not have even been detected by these methods.

A similar example was seen in the kidney cancer data, with the well-known *FLT1* gene, which plays a role in a number of tumorigenic processes including angiogenesis, proliferation, and metastasis, and has been associated with survival [15,36]. In the TCGA kidney cancer dataset, while all methods do detect this gene as being significantly associated with survival, dichotomization-based methods had the highest P-values (approximately 100-fold).

A more extreme case can be observed in the TCGA ovarian cancer dataset, when examining the epigenetic regulator *KDM5A*. This gene is known to be amplified and overexpressed in a number of different cancer types, including serous ovarian cancer, and has been implicated in EMT, poor survival, and suppression of apoptosis [37,38]. In the ovarian cancer dataset we note that Cox regression, C-index, and D-index again find that this gene is significantly associated with survival (P-value < 0.01) whereas the dichotomization-based methods do not detect this gene as associated with survival at an alpha level of 0.05. Similarly, the C-X-C Motif Chemokine Ligand 12 (*CXCL12*) gene plays a role in the metastasis of prostate carcinomas, which may indicate that expression of this gene is associated with poor survival [39,40]. Here again we find that this gene is not detected as associated with survival using the dichotomization-based methods, yet identified using Cox regression, C-index, and D-index. Overall, these results point to the value of moving away from the dichotomization-based methods for adopting rational strategies for survival analysis using continuous variables like gene expression data.

Discussion

While the eight methods were selected to conduct the same task – to identify gene-expressed based prognostic markers –mechanistically, the methods vary in their approach and the performance of these methods could largely be explained by the degree to which assumptions were made by each method. The method that made the least assumptions and therefore had the greatest flexibility, was the Cox regression method, which overall, had the strongest performance of the eight methods. The Cox regression is a framework that is commonly adopted as the most general method in survival analysis. In our study that focused on identifying prognostic predictors using gene expression, this method is a natural choice for modeling this continuous variable using a single covariate. The use of k-means as an alternative method could be considered the second least restrictive method because it assumed only the existence of two patient sub-groups, and beyond this, did not make any further assumptions regarding how these groups should be structured.

The class of methods tested that were based around a dichotomized value, such as the KaplanScan method, were more restrictive than the Cox regression and k-means methods, because they assumed that an inherent ordering existed amongst the patients. The KaplanScan method had the additional set of iterative steps that sought to infer the boundary that would split the population into the two most optimal groups but retained some flexibility in that this boundary was derived from the data. Moreover, given that the number of multiple hypotheses to be tested is directly linked to the number of genes considered in the analysis, the false positive rate grows linearly with more genes. Consider, for example, 100 samples of data, one would end up running 90 different log-rank tests for a given gene using KaplanScan. This therefore represents a clear limitation to this method.

The distribution-based splitting method used parametric assumptions to define a set of rules for dichotomization based on quantiles that were estimated directly from a gene's expression distribution in the patient population. While this method allowed for flexibility in that the dichotomization rule would adapt given the shape of the distribution, it also carried an additional layer of assumptions regarding the parametrization of these distributions that may not always hold true for each gene. Finally, the two methods based on dichotomizing using a quantile value were the simplest approaches to implement. However, they represented the most restrictive and least flexible of the eight methods because they did not account for the shape of the distribution, the data structure amongst patients, or the existence of alternative candidates for the breakpoint location. It was therefore unsurprising that the methods from this category had the poorest performance overall in all tests of reliability, accuracy, and robustness.

A standard way to turn a continuous variable into a binary or categorical one is simply to dichotomize the variable by finding an appropriate breakpoint or threshold. Since the range and distribution of gene expression profiles vary, especially for methods applied transcriptome-wide, quantile-based dichotomization schemes are commonly used. A natural choice for selecting a breakpoint for data-derived quantiles is to use the median. This breakpoint is often used in situations that involve identifying diagnostic markers for clinical tests. An alternative is to use quartiles, such as the 25th percentile and 75th percentile. An advantage of this approach is that patient groups are identified by more extreme changes in gene expression and therefore it may be easier to detect changes in survival time. However, 50% of patient samples are not included in the derivation of the predictive biomarker so for datasets with small sample sizes, this form of dichotomization may be more unstable. In either case, when relying on the median and quantiles as a threshold for dichotomization, one of the key concerns is the relative stability of these statistics, especially when highly-variable data such as RNA-seq data are used.

To investigate this issue in the context of this study, a gene was selected at random from TCGA as a representative example, and 500 bootstrap sets of 50 samples (gene expression measures) were generated for each of the four cancer types. The median, first quartile (Q1) and third quartile (Q3) were calculated from these bootstrapped datasets as a surrogate for having multiple datasets from which to investigate the stability of these threshold-based statistics (Supplemental Fig. 2). The distribution of these threshold statistics showed considerable variability that was statistically significant across the four cancer types (Levene's test, P-value $< 10^{-25}$). Given the significant degree of variability, these results demonstrated that the thresholds based on these statistics were not always robust between different types of cancers. Certainly, it was evident that the variance of the median, Q1 and Q3 were larger in the head and neck cancer dataset and ovarian cancer dataset, compared with the prostate and kidney cancer datasets. These results provide further support for the poor performance observed with the threshold-based dichotomization methods and suggest that it is potentially erroneous to apply them without thoroughly investigating the data *a priori*.

This study set out to determine which method had the most optimal performance for identifying gene expression-based prognostic biomarkers from cancer RNA-sequencing data. Out of eight methods, and using three different sets of assessments for accuracy, reliability, and robustness, the Cox regression had the best overall performance. For accuracy, the Cox regression, k-means, C-index and D-index had the strongest performance. The Cox regression had the most reliable performance, followed in second place by the k-means method. For robustness, the D-index had the strongest performance and the Cox regression method was the second most robust. A conclusion of this study is the recommendation against the use of methods that involve dichotomizing the gene expression data based on quantiles or the Kaplan-Scan method as these both performed poorly on our tests. It should be highlighted that the testing framework that was designed for our study was motivated by the task of performing an unbiased discovery or identification of candidate biomarkers from large-scale datasets. If a marker is identified from non-bioinformatics evidence, then it certainly may be reasonable to then determine an appropriate cutoff to guide binary treatment decisions in a clinical setting. We have also shown that the number of highly differentially expressed genes in a cancer greatly influences the ability to predict markers, independent of the method employed. In the future, we hope to investigate robust methods and more sensitive approaches to handle the problem of identifying markers of survival in the presence of heterogeneous cancer data. Improving these kinds of techniques will hopefully pave the way to develop personalized and more accurate cancer diagnostic tests that have widespread generalizability to other patient populations.

Conflict of interest

The authors state that they have no conflict of interest.

Acknowledgements

The authors would like to thank Jim Zhang and Eric Lim at Children's Hospital of Philadelphia and Stoke Therapeutics, respectively, that offered invaluable advice. J. C. Mar and S. Zimmerman are supported by a grant from NYSTEM (C029154). J.C.M. is supported by an Australian Research Council Future Fellowship (FT170100047) and by a Metcalf Prize from the National Stem Cell Foundation of Australia.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.cancergen.2019.04.004.

References

[1] Cancer Genome Atlas Research, N. The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet 2013;45(10):1113–20.
[2] *The International Genomics Consortium*. 2019]; Available from: https://www.ncbi.nlm.nih.gov/bioproject/PRJNA91763.

[3] Zhang J, et al. International Cancer Genome Consortium Data Portal–a one-stop shop for cancer genomics data. Database (Oxford) 2011;2011:bar026.
[4] Samkari A, White J, Packer R. SHH inhibitors for the treatment of medulloblastoma. Expert Rev Neurother 2015;15(7):763–70.
[5] Crabbe DC, Peters J, Seeger RC. Rapid detection of MYCN gene amplification in neuroblastomas using the polymerase chain reaction. Diagn Mol Pathol 1992;1(4):229–34.
[6] Riely GJ, Marks J, Pao W. KRAS mutations in non-small cell lung cancer. Proc Am Thorac Soc 2009;6(2):201–5.
[7] Siegel RL, Miller KD, Jemal A. Cancer statistics, 2018. CA Cancer J Clin 2018;68(1):7–30.
[8] Contal C, O'Quigley J. An application of changepoint methods in studying the effect of age on survival in breast cancer. Comput Stat Data Anal 1999;30(3):253–70.
[9] Rao M, Klein J, Moeschberger M. Survival analysis techniques for censored and truncated data. Technometrics 1998;40:159.
[10] de Torrente, L., et al., The shape of gene expression distributions matter: how incorporating distribution shape improves the interpretation of cancer transcriptomic data. 2019: bioRvix. doi:https://doi.org/10.1101/572693.
[11] Budczies J, et al. Cutoff Finder: a comprehensive and straightforward Web application enabling rapid biomarker cutoff optimization. PLoS One 2012;7(12):e51862.
[12] Culhane AC, et al. GeneSigDB: a manually curated database and resource for analysis of gene expression signatures. Nucleic Acids Res 2012;40(Database issue):D1060–6.
[13] Broad GDAC Firehose. 2019]; Available from: http://gdac.broadinstitute.org/.
[14] Spentzos D, et al. Gene expression signature with independent prognostic significance in epithelial ovarian cancer. J Clin Oncol 2004;22(23):4700–10.
[15] Zhao H, et al. Gene expression profiling predicts survival in conventional renal cell carcinoma. PLoS Med 2006;3(1):e13.
[16] Chung CH, et al. Gene expression profiles identify epithelial–to-mesenchymal transition and activation of nuclear factor-kappaB signaling as characteristics of a high-risk head and neck squamous cell carcinoma. Cancer Res 2006;66(16):8210–18.
[17] Henshall SM, et al. Survival analysis of genome-wide gene expression profiles of prostate cancers identifies new prognostic targets of disease relapse. Cancer Res 2003;63(14):4196–203.
[18] Wickham H. ggplot2: elegant graphics for data analysis. New York: Springer-Verlag; 2016.
[19] Ballings, M. and D. Van den Poel, *AUC: Threshold independent performance measures for probabilistic classifiers*. 2013. http://cran.r-project.org/web/packages/AUC/AUC.pdf.
[20] Benidt S, Nettleton D. SimSeq: a nonparametric approach to simulation of RNA-sequence datasets. Bioinformatics 2015;31(13):2131–40.
[21] Ayoub C, et al. ANO1 amplification and expression in HNSCC with a high propensity for future distant metastasis and its functions in HNSCC cell lines. Br J Cancer 2010;103(5):715–26.
[22] Molenaar JJ, et al. Sequencing of neuroblastoma identifies chromothripsis and defects in neuritogenesis genes. Nature 2012;483(7391):589–93.
[23] Choo S, et al. Reactivation of TWIST1 contributes to Ewing sarcoma metastasis. Pediatr Blood Cancer 2018;65(1).
[24] Somarelli JA, et al. Mesenchymal-Epithelial Transition in Sarcomas is Controlled by the Combinatorial Expression of MicroRNA 200s and GRHL2. Mol Cell Biol 2016;36(19):2503–13.
[25] *R2: Genomics Analysis and Visualization Platform* http://r2.amc.nl/.
[26] Wang J, et al. The bimodality index: a criterion for discovering and ranking bimodal signatures from cancer gene expression profiling data. Cancer Inform 2009;7:199–216.
[27] Shapiro S, Wilk M. An analysis of variance test for normality (Complete Samples). Biometrika 1965;52:591.
[28] Massey FJ. The Kolmogorov–Smirnov test for goodness of fit. J Am Stat Assoc 1951;46(253):68–78.
[29] Lilliefors H. On the Kolmogorov–Smirnov test for normality with mean and variance unknown. J Am Stat Assoc 1967;62:399–402.
[30] Gerds TA, et al. Estimating a time-dependent concordance index for survival prediction models with covariate dependent censoring. Stat Med 2013;32(13):2173–84.
[31] Brentnall AR, Cuzick J. Use of the concordance index for predictors of censored survival data. Stat Methods Med Res 2018;27(8):2359–73.
[32] Harrell FE Jr, et al. Evaluating the yield of medical tests. JAMA 1982;247(18):2543–6.
[33] Royston P, Sauerbrei W. A new measure of prognostic separation in survival data. Stat Med 2004;23(5):723–48.
[34] Schroder MS, et al. survcomp: an R/Bioconductor package for performance assessment and comparison of survival models. Bioinformatics 2011;27(22):3206–8.
[35] Ruiz C, et al. Enhanced expression of ANO1 in head and neck squamous cell carcinoma causes cell migration and correlates with poor prognosis. PLoS One 2012;7(8):e43265.
[36] Yamashita M, et al. Soluble Flt-1 Has Cytotoxic Effects on BeWo Choriocarcinoma Cells. Reprod Sci 2018;25(6):830–6.
[37] Feng T, et al. KDM5A promotes proliferation and EMT in ovarian cancer and closely correlates with PTX resistance. Mol Med Rep 2017;16(3):3573–80.
[38] Hu D, et al. KDM5A regulates a translational program that controls p53 protein expression. iScience 2018;9:84–100.
[39] Singh S, et al. CXCL12-CXCR4 interactions modulate prostate cancer cell migration, metalloproteinase expression and invasion. Lab Invest 2004;84(12):1666–76.
[40] Vaday GG, et al. CXCR4 and CXCL12 (SDF-1) in prostate cancer: inhibitory effects of human single chain Fv antibodies. Clin Cancer Res 2004;10(16):5630–9.