Abstract. Feature selection in the framework of meta-analyses (meta feature selection), combines meta-analysis with a feature selection process and thus allows meta-analysis feature selection across multiple datasets. In the present study, a meta feature selection procedure that fitted a multiple Cox regression model to estimate the effect size of a gene in individual studies and to identify the overall effect of the gene using a meta-analysis model was proposed. The method was used to identify prognostic gene signatures for lung adenocarcinoma and lung squamous cell carcinoma. Furthermore, redundant gene elimination (RGE) is of crucial importance during feature selection, and is also essential for a meta feature selection process. The current study demonstrated that the proposed meta feature selection procedure with RGE outperforms that without RGE in terms of predictive ability, model parsimony and biological interpretation.

Introduction

Lung adenocarcinoma (AC) and lung squamous cell carcinoma (SCC) are two major histological subtypes of non-small cell lung cancer (NSCLC) and accounted for ~70% of lung cancer (LC) cases worldwide in 2010 (1). For patients with NSCLC, the five-year survival rate is less than 15% (2). At present, the most promising strategy is early diagnosis followed by surgical resection of the tumors (3). Postoperative adjuvant chemotherapy may improve the survival rate of patients with a poor prognosis. However, it is not recommended for patients with stage IA NSCLC, whose five-year survival rate is approximately 70% (2). Therefore, using biomarkers to identify patients with NSCLC who may benefit from adjuvant chemotherapy is of clinical importance.

A biomarker is a measurable indicator of a biological state or condition (4). At present, biomarkers are used in numerous scientific fields. Previous studies have investigated the development of novel technologies for the accurate and easy detection and measurement of potential biomarkers (5-7). Microarray technology may be used to monitor thousands of genes and measure their expression values simultaneously. Previous studies have demonstrated that the signatures obtained using gene expression values from microarray experiments may distinguish between AC and SCC with perfect accuracy (8-13), and may determine the prognosis of patients with NSCLC (14-16). The identification of such gene signatures is generally accomplished with the aid of a feature selection algorithm, which reduces the number of genes under consideration, speeds up the learning process and improves the biological interpretation of resulting models (17).

RNA-sequencing (RNA-seq) technology has notable advantages over microarray technology, including increased precision for identification of differentially expressed genes (DEGs) (18), and it has replaced microarray technology as the first choice for gene expression profiling (19). However, the vast majority of existing statistical methods, including those used to identify the differentially expressed genes or select the relevant genes associated with the phenotypes of interest, were designed for continuous gene expression measures obtained from microarray experiments. The introduction of the R function voom (20) has allowed the count values of RNA-Seq data to be transformed into continuous values that follow approximately normal distributions. Consequently, current statistical methods used to analyze data obtained from microarray experiments may be directly applied to RNA-Seq.
data. This allows the investigation of the generalization of a gene signature trained from one platform to another platform.

Regardless of the technology used, it is frequently observed that the resulting prognostic gene signatures rarely overlap when the same analytic procedure is applied to different datasets. The accumulation of NSCLC gene expression data and the integration of experiments using a specific statistical method, for example a meta-analysis (21) or an integrative analysis (22,23), allow for one unique gene signature across multiple studies. In the present study, the overall prognostic value of a gene across three NSCLC microarray datasets was estimated using a meta-analysis model, with the aims of identifying subtype-specific prognostic signatures for AC and SCC and identifying the patients who may benefit from adjuvant treatment. Previous studies in the framework of meta-analyses on NSCLC for the purpose of prognosis have mainly focused on either the identification of prognostic genes for one specific subtype (24) or both AC and SCC patients (25). By contrast, the present study takes into account that the genes associated with the survival time of patients with AC and SCC may differ (26-28), with the aid of a feature selection algorithm capable of identifying subtype-specific prognostic gene signatures known as the Cox-filter method (27). The resulting subtype-specific prognostic gene signatures were verified on a RNA-seq dataset and an independent microarray dataset, and the biological relevance of those genes was investigated using the GeneCards database (www.genecards.org) (29).

Another feature of the present study was the control over redundant genes by evaluating the Pearson's correlation coefficients (PCCs) of a specific gene with all selected genes in a forward stepwise regression manner. The term ‘redundancy’ refers to the hidden associations or grouping structures that exist among genes, and therefore a gene may be erroneously included in the final gene signature due to its high correlation with the true relevant gene/s (30). Redundant genes do not contribute to the discriminative ability of a final model, and in numerous cases they hinder this ability. Therefore, the inclusion of redundant genes substantially influences the quality of a final gene signature (30).

Materials and methods

Experimental data. The Gene Expression Omnibus (GEO; www.ncbi.nlm.nih.gov/geo) repository of the National Institutes of Health was searched for the potential microarray experiments using the following keywords: ‘lung cancer’, ‘adenocarcinoma’, ‘squamous cell carcinoma’, ‘survival’ and ‘Affymetrix chip’. Subsequently, the selected datasets were further examined to identify whether patients with AC and SCC were included and whether survival information was available. The studies that did not include both were excluded. In total, four experiments were selected for inclusion in the present study: GSE3141 (31), GSE37745 (32), GSE30219 (33) and GSE50081 (16). The characteristics of each dataset are summarized in Table I.

The RNA-Seq data were downloaded from the lung Adenocarcinoma (LUAD; for AC subtype) and lung squamous cell carcinoma (LUSC; for SCC subtype) cohorts on The Cancer Genome Atlas (TCGA; level 3, tCGA-data.ncbi.nlm.nih.gov/tcga). The data of patients at the early stages of the disease that had not undergone any adjuvant treatment and that included survival information were selected. A total of 70 patients with AC and 55 patients with SCC were identified.

Pre-processing procedures. The raw data (CEL files) of the three microarray data sets, GSE37745, GSE30219 and GSE50081, served as the training set. They were downloaded from the GEO repository. The expression values were obtained using the frozen robust multiarray analysis (frma) algorithm (34), and subsequently normalized using quantile normalization using R frma package (34). In cases where multiple probe sets matched to one specific gene, the probe set with the largest fold change was selected. Subsequently, the pre-processed gene expression matrix of GSE3141 was downloaded from the GEO repository, and was considered as one of the test sets. This test set was used to investigate the effects that different pre-processing procedures may have on the downstream analyses.

For the second test set, the RNA-seq data, the counts-per-million values were calculated and log2 transformed by the Voom function (20) in R limma package (35). The purpose of having this test set was to examine the applicability of a gene signature trained on one platform to a different platform. The downstream analysis was performed on the 14,573 genes in the microarray data and the RNA-seq data.

Statistical methods

Cox-filter. The Cox-filter method (27) was used to identify genes associated with the survival rates of patients with the AC/SCC histology subtype. In this method, a Cox model is fitted on each gene, and the hazard function of patient i for gene g (g=1,...,p) is calculated as follows:

$$\lambda_g(t) = \lambda_{0g} \exp(\beta_2g I(j=1) + \beta_{1g} X_{ig} + \beta_{0g}(j=1) X_{ig})$$

Where X_g=(X_{g1},...,X_{gn}) represents actual expression values for the gene g under consideration. \(\lambda_{0g}(t)=\lambda_{0SCC}(t)\) is the baseline hazard function for the SCC group. The two groups have different baseline hazard functions, with \(\beta_{0g}\) representing the difference between the SCC and AC groups in terms of log baseline hazard function. I(j=1) is an indicator, taking the value of 1 if the histology subtype j of patient i is SCC, or the value of 0 if patient i has AC. Both \(\beta_{2g}\) and \(\beta_{3g}\) are the parameters of interest, with \(\beta_{2g}\) representing the change in log hazard rate associated with 1-unit increase in the actual expression value of gene g among AC and \(\beta_{3g}\) representing the additional change in log hazard rate associated with the SCC subtype. The values of \(\beta_{ACg}\), i.e., \(\beta_{1g}\) and \(\beta_{2g}\), determine whether subtype-specific prognostic genes exist. For example, \(\beta_{ACg}=0\) and \(\beta_{SCCg}=0\) correspond to an AC-specific gene, and \(\beta_{ACg}=0\) and \(\beta_{SCCg}\neq0\) correspond to an SCC-specific gene.

Overall effect size estimation using meta-analysis. The overall effect sizes were calculated using a meta-analysis model. The general model in a meta-analysis setting is written as follows:

$$Y_{ig} = \theta_{ig} + \varepsilon_{ig}, \quad \varepsilon_{ig} \sim N(0, \sigma_{ig}^2)$$

$$\theta_{ig} = \mu_{ig} + \delta_{ig}, \quad \delta_{ig} \sim N(0, \tau_{ig}^2)$$

where Y is the effect size, \(\theta\) is the true effect size, \(\mu\) is the mean effect size, \(\delta\) is the within-study variation, \(\sigma\) is the within-study standard deviation, \(\tau\) is the between-study standard deviation, and \(\varepsilon\) is the within-study error term.
Where \( Y_{ij} \) represents the estimated \( \beta \) coefficient (either \( \beta_{\text{AC}} \) or \( \beta_{\text{SCC}} \)) for study \( j \) \((j=1,\ldots,J)\) for a specific gene \( g \). \( \theta_{gi} \) is the study-specific hazard ratio for gene \( g \), and \( \varepsilon_{gi} \) is an error term which is assumed to follow a normal distribution \( \sigma_{gi}^2 \) represents the within-study variance of gene \( g \) for study \( j \). \( \gamma_{gi} \) and \( \delta_{gi} \) were estimated by the Cox-filter models for each study and thus considered to be known or more precisely observed. Furthermore, \( \theta_{gi} \) is assumed to be drawn from a superpopulation with an overall mean of \( \mu_{g} \) and a variance of \( \tau_{g}^2 \). Of note, \( \mu_{g} \) is the average of the hazard ratio over all studies for gene \( g \), which is the parameter of interest. \( \delta_{gi} \) is the error term for a superpopulation. \( \tau_{g}^2 \) is the between-study variance, which represents the variability between studies and is estimated by the DerSimonian and Laird method (36). Under a fixed-effect model, \( \tau_{g}^2=0 \), and \( \tau_{g}^2>0 \) corresponds to a random-effect model. The Cochran's Q statistic that follows a \( \chi^2 \) distribution under the null hypothesis \((H_0: \tau_{g}^2=0 \text{ vs. } H_1: \tau_{g}^2>0)\) was used to determine whether a fixed-effect or a random-effect model was more appropriate (37).

If a fixed-effect model was selected, the estimated \( \mu_{g} \) was standardized by its standard errors to obtain the Z-score, and the Z-score was the ratio of the estimated \( \mu_{gi} \), (represented by \( \mu_{gi}^* \)), to the square root of \( \tau_{gi}^2+\text{se}(\mu_{gi}^*)^2 \) in the random-effect model. The Z-score was assumed to follow a standard normal distribution and the adjusted P-values of the Z-score determined whether the specific gene was associated with the survival rate. In the present study, a gene with an adjusted P-value<0.05, where the Benjamini and Hochberg procedure (38) was used for the multiple comparison correction, and an integrated effect size (log hazard ratio)>0.5 was considered to indicate a statistically significant difference.

### Redundant gene elimination (RGE)

The expression values of genes are not independent from one another as there are relationships or grouping structures among genes. The correlations among genes result in numerous redundant genes and the removal of the redundant genes may lead to improved prediction accuracy and model stability (30). In the present study, in order to eliminate redundant genes in the resulting gene lists identified using the Cox-filter meta-analysis, the genes were arranged in an ascending order according to their adjusted P-values in the Cox-filter models, and a null set \( S \) was defined. The gene with the most significant P-value was placed into the newly-defined set \( S \) and the PCCs of the kth gene \((k=2,\ldots,p\text{, where } p \text{ is the number of genes in the list})\) with the genes inside set \( S \) were subsequently calculated for each study. If the absolute PCCs of one specific gene with the genes inside \( S \) for all studies were <0.4 (based on the sensitivity analysis), this gene was placed into \( S \), otherwise the gene was omitted. The PCCs were calculated for each gene in the list. The final prognostic gene signature with RGE was the resulting set \( S \).

The proposed procedures are referred to as the meta Cox-filter method with RGE (for the procedures with the add-on step of redundant gene elimination, the cut-off of PCCs is set at 0.4 based on a sensitivity analysis over values from 0.2 to 0.5, with an increment of 0.1 and 0.1). A multiple Cox regression model was fitted with all identified genes as covariates when the sizes of the resulting gene signatures were small enough, otherwise a multiple Cox regression model was fitted with the first five PCs of all identified genes as covariates and the meta Cox-filter method without RGE hereafter.

### Pathway enrichment analysis

The Search Tool for the Retrieval of Interacting Genes/Proteins (STRING, version 11.0, www.string-db.org) (39) was used to search for the Gene Ontology (GO: geneontology.org) terms (40) and the Kyoto Encyclopedia of Genes and Genomes (KEGG; https://www.genome.jp/kegg/) pathways (41) that were enriched by the STRING software (39) is a stand-alone online server used to construct the gene-to-gene interaction networks based on the data of gene fusion, co-occurrence, co-expression, experiments, various curated pathway databases and text mining.

### Performance statistics

The censoring-adjusted C-statistic (42) over the follow-up period \((0, \tau)\) was used to evaluate the performance of a resulting prognostic gene signature. The C-statistic is defined as follows:

\[
C_{i}(\beta) = P(g(X_{i}) > g(X_{j}) | T_i < T_j, T_i < \tau)
\]

Where \( g(X_{i}) \) is the risk score for patient \( i \) with predictor vector \( X_{i} \), representing the expression values of the selected prognostic genes. The risk scores for patients were constructed by fitting an extra multiple Cox regression model with either all genes or the first five principal components (PCs) of all identified genes as covariates, with \( \beta \) representing the coefficients before the covariates. \( T_i \) and \( T_j \) were the survival/censoring time for patient \( i \) and patient \( j \), respectively. A C-index value between

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Study</th>
<th>Raw data</th>
<th>Platform</th>
<th>Normalization method</th>
<th>Number of events/Number of AC patients</th>
<th>Number of events/Number of SCC patients</th>
<th>(Refs.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GSE3141</td>
<td>Bild et al, 2006</td>
<td>No</td>
<td>HGU 133 Plus 2</td>
<td>MAS5</td>
<td>32/58</td>
<td>26/53</td>
<td>(31)</td>
</tr>
<tr>
<td>GSE30219</td>
<td>Rousseaux et al, 2013</td>
<td>Yes</td>
<td>HGU 133 Plus 2</td>
<td>FRMA</td>
<td>43/85</td>
<td>19/22</td>
<td>(33)</td>
</tr>
<tr>
<td>GSE37745</td>
<td>Botling et al, 2013</td>
<td>Yes</td>
<td>HGU 133 Plus 2</td>
<td>FRMA</td>
<td>27/40</td>
<td>20/24</td>
<td>(32)</td>
</tr>
<tr>
<td>GSE50081</td>
<td>Der et al, 2014</td>
<td>Yes</td>
<td>HGU 133 Plus 2</td>
<td>FRMA</td>
<td>51/127</td>
<td>16/42</td>
<td>(16)</td>
</tr>
</tbody>
</table>

Yes, the corresponding data/information were available; FRMA, frozen robust multiarray analysis. *, dataset used as a test set.
Table II. Sensitivity analysis to determine the cut-off value of average absolute Pearson’s correlation coefficients over three microarray studies.

<table>
<thead>
<tr>
<th>Cut-off value</th>
<th>Size</th>
<th>GSE30219</th>
<th>GSE37745</th>
<th>GSE50081</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AC</td>
<td>SCC</td>
<td>AC</td>
<td>SCC</td>
</tr>
<tr>
<td>0.2</td>
<td>2</td>
<td>3</td>
<td>0.708</td>
<td>0.665</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.3</td>
<td>9</td>
<td>6</td>
<td>0.628</td>
<td>0.593</td>
</tr>
<tr>
<td>0.4</td>
<td>24</td>
<td>12</td>
<td>0.804</td>
<td>0.903</td>
</tr>
<tr>
<td>0.5a</td>
<td>54</td>
<td>28</td>
<td>0.638</td>
<td>0.541</td>
</tr>
<tr>
<td>1a</td>
<td>131</td>
<td>203</td>
<td>0.605</td>
<td>0.652</td>
</tr>
</tbody>
</table>

*Since the size of resulting gene signature is increasing, resulting in a multiple Cox regression model fitting problem, the first five principal components were used as covariates to fit multiple Cox models. Based on this, the cut-off value of absolute Pearson correlation coefficients is set as 0.4. AC, adenocarcinoma; SCC, squamous cell carcinoma.

0.6 and 0.7 indicates a prognostic signature has satisfactory performance (43).

Furthermore, using the median of those risk scores, that is g(X), as a cut-off, the patients were classified into either a low-or high-risk group. The Kaplan-Meier curves for these two groups were obtained, and log-rank tests were used to compare these two curves. A smaller P-value suggested a more significant difference between the survival curves of the low-and the high-risk groups.

Statistical language and packages. All statistical analyses were conducted in R language (version 3.3; www.r-project.org).

Results and Discussion

In the present study, the data of three microarray experiments (GSE37745, GSE30219 and GSE50081) were used as the training sets, and the GSE3141 dataset and RNA-seq data from the TCGA database were used as the test sets to validate the performance of the resulting prognostic gene signatures.

First, a sensitivity analysis was conducted in order to determine the optimal cut-off for the absolute PCCs, which identifies the genes that are regarded as redundant genes and may thus be excluded from the final gene lists. The C-statistics of the resulting gene signatures were calculated for each study and are presented in Table II. Based on these statistics and the size of the final gene lists, the cut-off was set at 0.4, which corresponded to the largest C-statistics for all studies taken together and the smallest number of selected genes. Furthermore, it was observed that the proportion of redundant genes existing within a gene list selected by a filter method, such as the Cox-filter method, was substantial. Following RGE, the size of the AC-specific prognostic signature was reduced from 131 to 24, indicating that approximately 80% of identified genes were redundant. Similarly, the size of the SCC-specific prognostic signature was reduced from 203 to 12, indicating that the percentage of redundant genes for the SCC subtype was even higher than the AC subtype. The 131-gene signature for AC and the 203-gene signature for SCC identified by the proposed method without RGE are listed in Table SI, along with the significance levels of those genes and the labels for redundancy.

The meta Cox-filter method with RGE identified a 24-gene AC-specific prognostic signature and a 12-gene SCC-specific prognostic signature. The Kaplan-Meier plots of the two signatures on the three training sets are presented in Fig. 1 and those on the two test sets are presented in Fig. 2. Using two independent studies, the GSE3141 dataset (31) and the TCGA RNA-Seq data (under the cohorts of LUAD and LUSC), the resulting gene lists were demonstrated to have a good predictive performance. Therefore, the results of the present study have good generalization. The genes of the 24-gene AC-specific prognostic signature and the 12-gene SCC-specific prognostic signature are presented in Fig. 3 with those directly associated with lung cancer underlined. Using these two gene lists, clinicians may design corresponding diagnostic kits to calculate the risk score for a patient with NSCLC and predict the prognosis, and ultimately allow for the possibility of a more ‘personalized’ treatment. Therefore, the results of the present study are clinically important.

The Venn-diagram in Fig. 3 indicates there five genes [retinol dehydrogenase 13 (RDH13), zinc finger protein 24 (ZNF24), LSM11 U7 small nuclear RNA associated (LSM11), down-regulator of transcription 1 (DR1) and zinc finger protein 385D overlapping between the two sets of signatures. According to the GeneCards database (29), none of the five overlapped genes are directly associated with lung cancer. However, all of them are indirectly associated with lung cancer as shown by the GeneCards database (29). For example, RDH13 and ZNF24 interact with a well-known cancer-associated gene tumor protein 53 (TP53) which encodes a tumor suppressor protein that contains transcriptional activation, DNA binding and oligomerization domains. Mutations in TP53 are associated with a several types of of human cancer. Furthermore, the GeneCards database (29) indicates these five genes are associated with other well-known cancer-associated genes. Specifically, LSM11 is associated with KRAS proto-oncogene GTPase (KRAS), epidermal growth factor receptor (EGFR) and signal transducer and activator of transcription 3. RDH13 interacts with EGFR and MET proto-oncogene, receptor tyrosine kinase,
and ZNF24 interplays with KRAS and vascular endothelial growth factor (VEGF) A and phosphatase and tensin homolog, and DR1 is targeted by Jun proto-oncogene, AP-1 transcription factor subunit. All these well-known genes are associated with lung cancer. For example, EGFR was revealed to be involved in the development and progression of lung cancer (44) and VEGF gene polymorphism serves a role in the development of lung cancer (45).

Choi et al (46) conducted a multivariate analysis that demonstrated a significant correlation between strong transglutaminase 2 (TGM2) expression and shorter disease-free survival in patients with NSCLC and the non-adenocarcinoma subtype, and the correlation in the patients with the adenocarcinoma was not significant. However, the present study identified TGM2 as a hazardous gene for the AC subtype by the meta Cox-filter methods with and without RGE. The forest plot for this gene is presented in Fig. 4 and it demonstrates that the hazard ratios of TGM2 in all individual studies were positive, but only that of the GSE30029 dataset was significant (P<0.05). A meta-analysis model increases statistical power, so that consistently hazardous but not statistically significant effects across studies become statistically meaningful when taken together. Further investigation on the prognostic value of TGM2 for the AC subtype is required.

In the present study, STRING software (39) was used to search for the GO terms (40) and the KEGG pathways (41) that were enriched by the AC-and the SCC-specific prognostic signatures. The results are presented in Fig. 5. The figure

![Figure 1. Kaplan-Meier plots of the AC-and SCC-specific prognostic signatures for the training sets. The training sets included three microarray studies: GSE30029, GSE37745 and GSE50081. Based on the risk scores, patients were divided into two categories (a low-risk group and a high-risk group) using the medians of the risk scores as cut-offs. The P-values of the log-rank tests comparing the survival curves of the low-and high-risk groups are presented in each plot. AC, adenocarcinoma; SCC, squamous cell carcinoma.](image)
shows that there is no overlap between the two sets of enriched gene sets for GO terms and KEGG pathways, respectively. Therefore, the pathways enriched by these two gene signatures are subtype-specific.
Figure 4. Forest plot for one specific gene, TGM2. The meta Cox-filter method indicated that TGM2 is a hazardous gene, as a higher expression value was associated with a short survival time of patients with AC. TGM2, transglutaminase 2.

Figure 5. Venn-diagrams of the enriched GO terms and the KEGG pathways in the 24-gene AC-specific prognostic signature and the 12-gene SCC-specific prognostic signature using the Search Tool for the Retrieval of Interacting Genes/Proteins database. GO, Gene Ontology; KEGG, Kyoto Encyclopedia of Genes and Genomes; AC, adenocarcinoma; SCC, squamous cell carcinoma; CC, cellular component; BP, biological process.
Feature selection in the framework of meta-analysis combines meta-analysis with the feature selection process and thus performs meta-analysis feature selection for multiple datasets. It has a notable advantage over the methods of implementing a specific feature selection algorithm for individual studies and then taking the intersection of the gene signatures given by individual studies. Namely, it can select the same set of genes across multiple experiments. The proposed procedure in the present study and the meta threshold gradient descent regularization (MTGDR) method (47) belong to the meta feature selection category. However, the proposed procedure, a simple combination of the Cox-filter method and a meta-analysis on the summary statistics (i.e., β coefficients), is not as complicated as the MTGDR method. Compared with the MTGDR method, therefore, the proposed procedure is easier to understand and

Table III. Performance statistics of the proposed procedure on two independent test sets.

<table>
<thead>
<tr>
<th>Method</th>
<th>C-index</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>GSE3141</td>
<td>RNA-seq</td>
</tr>
<tr>
<td></td>
<td>AC</td>
<td>SCC</td>
</tr>
<tr>
<td>mCox-filter with RGE</td>
<td>0.857</td>
<td>0.659</td>
</tr>
<tr>
<td>mCox-filter without RGE</td>
<td>0.752</td>
<td>0.714</td>
</tr>
</tbody>
</table>
| mCox-filter with RGE corresponds to the meta Cox-filter method with RGE, where the cut-off value of absolute Pearson's correlation coefficient between two genes is set at 0.4. mCox-filter without RGE corresponds to the meta Cox-filter method. RGE, redundant gene elimination; RNA-seq, RNA-sequencing; AC, adenocarcinoma; SCC, squamous cell carcinoma.

Figure 6. Gene-to-gene interaction networks of the AC-and SCC-specific prognostic signatures constructed using the Search Tool for the Retrieval of Interacting Genes/Proteins database. Edges with different colors indicated that the evidence of interaction was from different sources. For example, a light blue edge is predicted from a curated database and a purple edge is based on experimentally determined interactions. These two networks demonstrated that the identified genes are rarely connected to each other, indicating that the redundant gene elimination screened out highly associated genes from the final gene lists. The overlapped genes that are in both signatures are circled in red. AC, adenocarcinoma; SCC, squamous cell carcinoma.
implement. The proposed method is potentially a superior choice when a researcher aims to identify the subtype-specific prognostic genes using multiple related gene expression datasets. Therefore, the results of the present study have potential for clinical application.

RGE is an important aspect of the feature selection process (48). Consistent with previous studies (49-51), the present study demonstrated that the RGE step is beneficial by improving the predictive performance, downsampling the sizes of the final gene signatures and increasing model parsimony, thus facilitating the experimental validations. Therefore, the additional consideration of deleting redundant genes is highly recommended, particularly when a filter method is utilized to perform feature selection. This is because filters generally screen genes one by one according to their relevance scores with the outcome of interest and thus lead to a high false positive rate by including numerous redundant genes (52).

Acknowledgements
Not applicable.

Funding
The current study was supported by the National Natural Science Foundation of China (grant no. 31401123).

Availability of data and materials
The datasets used in the present study are available in the GEO database (www.ncbi.nlm.nih.gov/geo) or the TCGA database (tcga-data.nci.nih.gov/tcga).

Authors' contributions
ST conceived the study. LW, CL, TW and ST performed the data analyses and interpreted the results. ST, CL, TW and ST wrote the manuscript.

Ethics approval and consent to participate
Not applicable.

Patient consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

References


This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International (CC BY-NC-ND 4.0) License.