An 8-Gene Signature for Classifying Major Subtypes of Non-Small-Cell Lung Cancer

Cancer Informatics
Volume 21: 1–12
© The Author(s) 2022
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/11769351221100718
SAGE

Mehdi Hamaneh and Yi-Kuo Yu

National Center for Biotechnology Information, National Library of Medicine,
National Institutes of Health, Bethesda, MD, USA.

ABSTRACT

MOTIVATION: The precise diagnosis of the major subtypes, lung adenocarcinoma and lung squamous cell carcinoma, of non-small-cell lung cancer is of practical importance as some treatments are subtype-specific. However, in some cases diagnosis via the commonly-used method, that is staining the specimen using immunohistochemical markers, may be challenging. Hence, having a computational method that complements the diagnosis is desirable. In this paper, we propose a gene signature for this purpose.

RESULTS: We developed an expression-based method that systematically suggests a huge set of candidate gene signatures and finds the best candidate. By applying this method to a training set, the optimal gene signature was found by considering close to 765 billion candidate signatures. The 8-gene signature found for classifying the 2 aforementioned subtypes comprises TP63, CALML3, KRT5, PKP1, TESC, SPINK1, C9orf152, and KRT7. The signature achieved a high overall prediction accuracy of 0.936 when tested using 34 independent gene expression datasets obtained using different technologies and comprising 2556 adenocarcinoma and 1630 squamous cell carcinoma samples. Additionally, the signature performed well in clinically challenging cases, that is poorly differentiated tumors and specimens obtained from biopsies. In comparison with 2 previously reported signatures, our signature performed better in terms of overall accuracy and especially accuracy of classifying lung squamous cell carcinoma.

CONCLUSIONS: Our signature is easy to use and accurate regardless of the technology used to obtain the gene expression profiles. It performs well even in clinically challenging cases and thus can assist pathologists in diagnosis of the ambiguous cases.

KEYWORDS: Non-small-cell lung cancer, subtype classification, gene expression

RECEIVED: January 21, 2022. **ACCEPTED:** April 9, 2022.

TYPE: Original Research

FUNDING: The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the Intramural Research Program of the National Library of Medicine at the National Institutes of Health grant number: ZIA LM092404.

DECLARATION OF CONFLICTING INTERESTS: The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

CORRESPONDING AUTHOR: Yi-Kuo Yu, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, 8600 Rockville Pike, Bethesda, MD 20894, USA. Email: yyu@ncbi.nlm.nih.gov

Introduction

Lung cancer has been historically divided into 2 main types: small-cell lung carcinoma and non-small-cell lung carcinoma (NSCLC), which accounts for approximately 80% of lung cancers.[1] NSCLC has 2 major subtypes, that are lung adenocarcinoma (LUAD) and squamous cell carcinoma (LUSC).[2] Over the past decade or so several therapeutic options have been discovered that can only be used to treat patients with specific NSCLC subtypes.[3] This has elevated the importance of precise diagnosis of LUAD and LUSC. Such diagnoses are usually done using immunohistochemistry.[3,4] However, since in most cases a small specimen obtained using a biopsy is used for diagnosis,[2] precise classification of these subtypes may be difficult in some patients, especially in poorly differentiated cases.[5] Thus, having a computational tool to complement diagnosis through immunohistochemistry can be beneficial. Additionally, such a computational method may be helpful in discovering new markers for immunohistochemical diagnosis of LUAD/LUSC[6] and possibly even shed light on the differences of the 2 subtypes at the molecular level. For these reasons, several expression-based methods have recently been proposed to find gene signatures that can be used to distinguish LUAD and LUSC samples. In the following paragraphs we briefly summarize some of the previously published studies on this topic.

Employing differential gene expression analysis, Hou et al[7] identified a 50-gene signature for LUAD/LUSC classification and tested it on one independent dataset. Using a method based on the nearest shrunken centroid classification algorithm, Charkiewicz et al[8] discovered a 53-gene signature for LUAD/LUSC classification and tested it on a single dataset. Huang et al[9] employed 3 different methods and reported high classification accuracies for the 3 methods. Employing a machine learning algorithm and using microarray data for training, Wu et al[10] discovered a 5-gene signature and tested their method/signature on RNA-Seq data from The Cancer Genome Atlas (TCGA). Su et al[11] used a few machine learning algorithms and discovered different signatures with different numbers of genes. They showed that a 13-gene signature, found using a random forest algorithm, performed best and had a good performance when applied to the TCGA data.

The methods mentioned above tested their signatures on 1 or 2 independent datasets. However, a signature obtained using a particular dataset may not perform well when applied to others. This is because, due to batch effects, the expression data obtained using different platforms/technologies are generally not comparable to each other. Fortunately, the within-sample relative gene expression orderings are not expected to be affected by the batch effects.[6] Thus, one can discover a gene signature with a robust performance consisting of a pair of genes whose expression levels switch ordering between LUAD and LUSC. Using this idea, Li et al[6] found a signature comprising 2 genes KRT5 and AGR2, and classified a sample as LUAD (LUSC) if the expression level of KRT5 was lower (higher) than that of AGR2. Li et al discovered their signature by considering sets of 2 oppositely regulated genes, that is 1 significantly up- and 1 significantly down-regulated genes. (Here, and in the rest of the paper, down/up-regulated means lower/higher expression level in LUAD in comparison with LUSC). Girard et al[5] also proposed a signature consisting of 21 pairs of oppositely regulated genes. However, they used a correlation-based scoring system to classify LUAD versus LUSC, that is they classified a sample as LUAD if its 42-gene expression profile had a higher correlation with the mean expression profile of LUAD than with that of LUSC (a nearest neighbor approach). The 42-gene signature of Girard et al performed well when tested on many datasets.

Some of the aforementioned methods have achieved high classification accuracy (ACC), defined as the number of correct diagnoses divided by the total number of patients. However, improvement is still desirable and helpful as even a modest increase in diagnosis accuracy could have major implications for some patients in terms of treatment. On the other hand, signatures with fewer genes (but with the same high accuracy) are more helpful for finding new drug targets because they narrow down the candidate genes. Therefore, it is desirable to find a new signature with high accuracy and a smaller number of genes.

In this paper we report a simple gene selection approach for discovering a more accurate signature that comprises the same number of highly up- and down-regulated genes. The method of discovery, which is explained in detail in Methods, is summarized as follows. First, using a training set, we identified the list of highly differentially expressed genes (HDEGs) between LUAD and LUSC. We then ranked the N identified HDEGs, in ascending order, based on their false discovery rates (FDRs). To find the optimal $2n$-gene signature, that is consisting of n up- and n down-regulated genes, the top n genes in the list were then chosen as the first half of the signature. To identify the second half, we considered the set of $N-n$ genes ranked lower than nth, and found all of its n-gene subsets whose members were oppositely regulated with respect to the top n genes. Each of these n-gene subsets were then added to the top n genes, creating a candidate $2n$-gene signature. Using a correlation-based scoring system equivalent to that of Girard et al[5] (see Methods for details), we then calculated the ACC for all the resulting candidate $2n$-gene signatures and identified the one with maximum ACC. The most accurate $2n$-gene signature was identified for each possible value of n. We found that, among these signatures, the one corresponding to $n=4$ gave the best results while having the lowest number of genes, that is we found an 8-gene signature.

We tested our 8-gene signature using 34 additional datasets (including the TCGA data; see Table 1) comprising 2556 (1630) LUAD (LUSC) samples and observed an overall high classification accuracy of 0.936. The signature performed well in cases that might be clinically difficult to classify, that is poorly differentiated samples and specimens obtained from biopsies. A comparison between our signature with those of Girard et al[5] and Li et al,[6] indicated that our signature outperforms both signatures in terms of overall prediction accuracy and especially prediction accuracy of LUSC, while having a smaller number of genes in comparison with the signature of Girard et al. Based on these results we believe our signature can assist pathologists in diagnosis of LUAD/LUSC.

Methods

Experimental data

TCGA[12,13] data (raw counts) for LUAD and LUSC primary tumors were downloaded from the data portal of the Genomic Data Commons (https://portal.gdc.cancer.gov/) of the National Cancer Institute.[14] For very rare cases in which multiple tumor samples were available for the same patient, we used only one of the tumor samples. Genes with zero counts across all samples were removed. However, since the data still contained many zeros, a pseudocount of 1 was added to the raw counts before normalization (using DESeq2[15]) and log-transformation.

Additional LUAD/LUSC datasets were found by either searching the Gene Expression Omnibus (GEO) database[16] directly, or by searching the previously published literature on this topic. Initially, we searched for datasets including at least 10 samples of each of the 2 subtypes. However, since the identified datasets included insufficient poorly differentiated samples to draw a reliable conclusion, datasets containing poorly differentiated samples from only one of the subtypes were also included in the study. Our search identified 34 GEO datasets that are given in Table 1. For these datasets, the normalized expression levels were downloaded from the GEO. If not already in log scale, the expression levels were log-transformed. For each platform, the mapping between probe-set IDs and gene IDs (or symbols) were performed using the corresponding annotation file in GEO. If more than 1 probe-sets mapped to a gene, the expression level of the gene was computed by averaging those of the corresponding probe-sets. Also, to be able to compare the TCGA data with those from GEO, Ensembl IDs were mapped to Entrez gene IDs using BioMart[17] as implemented in R. If more than 1 Ensembl IDs were mapped to an Entrez ID, their expression levels were averaged.

The dataset GSE41271[18] was employed for training and the rest of the datasets were used for testing. We chose GSE41271 as the training set because it is the largest dataset (containing

Table 1. Datasets used in this study.

	Dataset	N^{a}_{LUAD}	N^{a}_{LUSC}	N^{a}
1	GSE10245	40	18	58
2	GSE115457	40	30	70
3	GSE11969[c]	90	35	125
4	GSE14814	71	52	123
5	GSE16534	21	22	43
6	GSE17710[c]	0	56	56
7	GSE18842[c]	14	31	45
8	GSE19188	45	27	72
9	GSE2109[c]	35	39	74
10	GSE21933	11	10	21
11	GSE26939[c]	116	0	116
12	GSE28571	50	28	78
13	GSE29013[e]	30	25	55
14	GSE29016	38	12	50
15	GSE30219	85	61	146
16	GSE31546[c]	16	0	16
17	GSE31547[c]	30	0	30
18	GSE31799	29	20	49
19	GSE33532	40	16	56
20	GSE37745	106	66	172
21	GSE41271[b]	183	80	263
22	GSE42127	133	43	176
23	GSE43580[c]	77	73	150
24	GSE44170[c,e]	0	38	38
25	GSE4573[c]	0	130	130
26	GSE50081	127	42	169
27	GSE5828[c]	0	59	59
28	GSE5843[c]	48	0	48
29	GSE58661[d]	38	36	74
30	GSE60644	77	22	99
31	GSE68465[c]	443	0	443
32	GSE8569[c]	30	36	66
33	GSE8894	61	72	133
34	GSE94601[c]	102	30	132
35	TCGA	513	501	1014

[a] N_{LUAD} and N_{LUSC} are respectively the numbers of the LUAD and LUSC samples. N denotes the total number of samples.
[b] GSE41271 was used for training.
[c] These datasets include information regarding the degree of differentiation of at least some of the samples.
[d] Samples obtained from biopsies.
[e] Formalin fixed paraffin-embedded (FFPE) samples.

both LUAD and LUSC samples) among the GEO datasets we found. Since, among the identified 34 datasets, TCGA is the only one containing samples obtained from RNA-Seq, we did not use it as the training set to be able to test our signature on RNA-Seq as well as microarray data. Consisting of 183 LUAD and 80 LUSC samples, our training set may seem too small compared to the testing set (comprising 2556 LUAD and 1630 LUSC samples). However, classifiers trained on too-small datasets are usually overtrained, that is their prediction accuracy for the testing set is significantly lower than that for the training set. As we show in the Results section, we observed almost equal accuracies for the training and testing sets, indicating the appropriateness of the size of the training set. To provide further support for the suitability of our training set, we also note that the previously published signatures (that have been tested on a large number of samples) have used the same,[5] or a smaller training set.[6]

Sample scoring and classification

This section describes how we score and classify samples provided a gene signature is *given*. The gene selection procedure, that is the method we used for discovering the optimal $2n$-gene signature, is explained in the next section. Suppose a $2n$-gene signature, containing n up- and n down-regulated genes, and a training dataset containing LUAD/LUSC samples are given. Let $\mathbf{E}_i$ be a $2n \times 1$ vector whose jth element, E_{ij}, is the log-transformed expression value of the jth gene of the signature in the sample i. The score of sample i is defined as $S_i = r(\mathbf{E}_i, \mathbf{E}_0^{AD}) - r(\mathbf{E}_i, \mathbf{E}_0^{SC})$.[5] Here $\mathbf{E}_0^{AD}$ and $\mathbf{E}_0^{SC}$ are respectively the average expression vectors of the $2n$ genes in LUAD and LUSC calculated using the training dataset, and $r(X,Y)$ denotes the correlation between X and Y. We classify sample i as LUSC if $S_i < 0$, otherwise the sample is classified as LUAD. In other words, to classify the sample, we use a nearest neighbor approach.

The score defined above can be calculated in a simpler manner. To this end, let us also introduce the centered expression vector $\tilde{\mathbf{E}}_i$ whose elements are given by $\tilde{E}_{ij} = E_{ij} - E_i^{av}$, where E_i^{av} denotes the average of expression levels of the $2n$ genes in the sample i. We note that $r(\mathbf{E}_i, \mathbf{E}_0^{AD}) = \langle \tilde{\mathbf{E}}_i, \tilde{\mathbf{E}}_0^{AD} \rangle / (\| \tilde{\mathbf{E}}_i \| \| \tilde{\mathbf{E}}_0^{AD} \|)$. Here $\|\bullet\|$ denotes the norm of $\bullet$, $\langle \tilde{\mathbf{E}}_i, \tilde{\mathbf{E}}_0^{AD} \rangle$ is the dot product of $\tilde{\mathbf{E}}_i$ and $\tilde{\mathbf{E}}_0^{AD}$, and $\tilde{\mathbf{E}}_0^{AD}$ is the centered vector corresponding to $\mathbf{E}_0^{AD}$. Considering a similar relation for $r(\mathbf{E}_i, \mathbf{E}_0^{SC})$, one can show the alternative score s_i, defined below, is proportional to S_i.

$$s_i = r(\mathbf{E}_i, \mathbf{E}_0) = \frac{1}{\| \mathbf{E}_0 \|} S_i, \qquad (1)$$

where

$$\mathbf{E}_0 = \frac{\tilde{\mathbf{E}}_0^{AD}}{\| \tilde{\mathbf{E}}_0^{AD} \|} - \frac{\tilde{\mathbf{E}}_0^{SC}}{\| \tilde{\mathbf{E}}_0^{SC} \|}. \qquad (2)$$

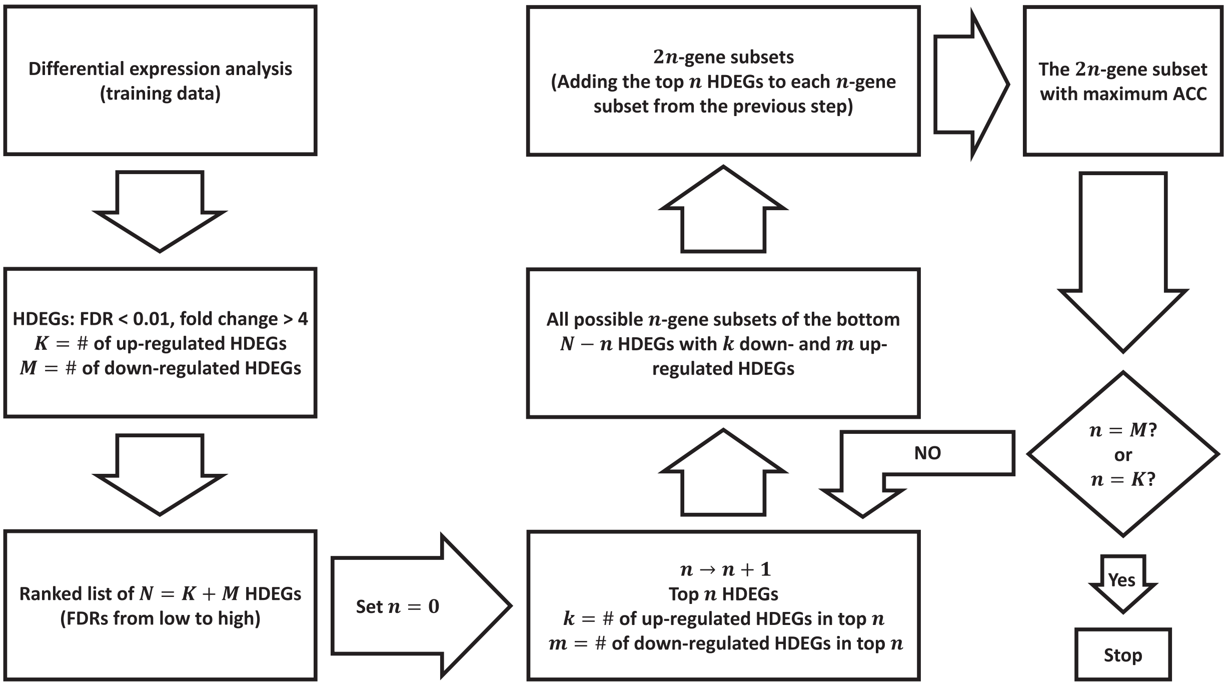
Differential expression analysis
(training data)
HDEGs: FDR < 0.01, fold change > 4
K = # of up-regulated HDEGs
M = # of down-regulated HDEGs
Ranked list of $N = K + M$ HDEGs
(FDRs from low to high)
Set $n = 0$
$n \rightarrow n + 1$
Top n HDEGs
k = # of up-regulated HDEGs in top n
m = # of down-regulated HDEGs in top n
All possible n-gene subsets of the bottom
$N - n$ HDEGs with k down- and m up-regulated HDEGs
$2n$-gene subsets
(Adding the top n HDEGs to each n-gene
subset from the previous step)
The $2n$-gene subset
with maximum ACC
$n = M$?
or
$n = K$?
NO
Yes
Stop

Figure 1. The signature discovery procedure. The steps taken to find the best performing $2n$-gene sets are depicted. The optimal gene signature is then identified by finding the n for which ACC is maximized.

In other words, equation (1) is equivalent to the scoring method of Girard et al.[5] However, it has the advantage of introducing only one reference expression ($\mathbf{E}_0$) rather than 2 ($\mathbf{E}_0^{\mathrm{AD}}$ and $\mathbf{E}_0^{\mathrm{SC}}$).

Finding the gene signature

A 2-step gene selection procedure was employed to find the optimal gene signature. The required steps, which are explained below, are depicted in Figure 1.

Identifying and ranking HDEGs. Given the expression profiles of LUAD and LUSC samples in the training dataset, a gene was considered an HDEG if: (1) there was a statistically significant difference between the log-transformed expression levels of the gene in LUAD and LUSC and (2) the absolute value of this difference was larger than 2 (ie, there was a larger than 4 fold change). Specifically, for each gene, we first used the Wilcoxon rank-sum test to assess the statistical significance of differential expression (comparing the log-transformed expression levels of the gene in the LUAD samples with those in LUSC samples). The Benjamini-Hochberg[19] procedure was then employed to correct for multiple hypotheses testing with an FDR cutoff of 0.01. For each gene, we then calculated $\bar{E}^{\mathrm{AD}}$ ($\bar{E}^{\mathrm{SC}}$) that is the average log-transformed expression level of the gene in LUAD (LUSC) in the training set. Finally, genes satisfying both $\mathrm{FDR} < 0.01$ and $|\bar{E}^{\mathrm{AD}} - \bar{E}^{\mathrm{SC}}| > 2$ were considered HDEGs (here $|\bullet|$ denotes the absolute value of $\bullet$). The N identified HDEGs were then ranked, in ascending order, based on their FDRs. In rare cases 2 or more genes had the same FDRs. To break the tie, we used the fold change, that is, we ranked the tied genes, in descending order, based on their fold changes.

Searching for the $2\mathrm{n}$ *-gene signature with the highest ACC.* The most statistically significant HDEGs are expected to have the most discriminative power. Thus, we chose the top n genes in the ranked list of HDGEs as the first half of the signature. To find the second half, we considered all possible subsets of the remaining $N-n$ genes that contained n genes oppositely regulated relative to the top n. In other words, if the first half of the signature (the top n) contained k up- and m down-regulated genes, among the bottom $N-n$ we found all sets of n genes comprising m up- and k down-regulated genes. We then added the top m genes to each of these subsets to find all possible $2n$-gene signatures containing the top n genes, which resulted in $\frac{(K-k)!}{(K-k-m)!m!}\frac{(M-m)!}{(M-m-k)!k!}$ candidate signatures. Here K is the number of up-regulated and M is the number of down-regulated HDEGs ($K+M=N$). Using each candidate signature and employing Eq. 1, the samples in the training set were then scored and classified as LUAD (LUSC) if their scores were non-negative (negative). Finally, the ACC for each candidate signature was computed and the best-performing $2n$-gene signature was identified as the one with the highest ACC. This procedure was performed for all possible values of n, that is $n = 1, 2, \ldots, \min(K, M)$, where $\min(K, M)$ is the smaller of K and M. The best performing $2n$-gene signatures were then compared to each other to find the optimal n and so the optimal gene signature.

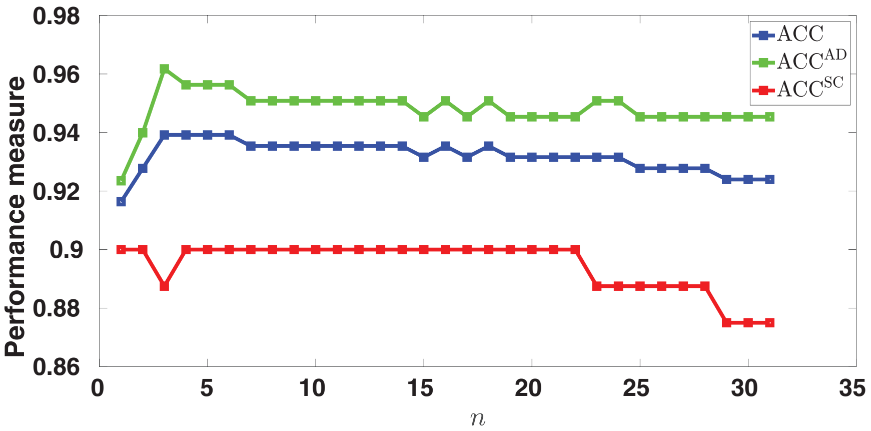

Figure 2. Finding the signature. The maximum achieved ACC is plotted as a function of n that is the number of pairs in the signature. Also shown in the figure are plots of ACC^{AD} and ACC^{SC} as functions of n. The optimal point is $n = 4$ that is the lowest n for which the ACC is maximized and ACC^{AD} – ACC^{SC} is minimized.

Results

Finding the gene signature

We employed the approach described in Methods for discovering a gene signature comprising equal numbers of up- and down-regulated genes. Using GSE41271 dataset as the training set, we first identified $N = 121$ HDEGs consisting of $M = 90$ down- and $K = 31$ up-regulated genes. The HDEGs were then ranked based on their FDRs, generating a ranked list of genes (Supplemental Table S1). All possible $2n$-gene signatures ($n = 1, 2, \ldots 31$) were then identified (close to 765 billion candidate gene signatures). Using each candidate signature the samples in the training set were classified and ACC, sensitivity (fraction of LUAD samples classified correctly; denoted by ACC^{AD}), and specificity (fraction of LUSC samples classified correctly; denoted by ACC^{SC}) were calculated. Given the large number of gene combinations considered when $n > 5$, we noticed that more than 1 $2n$-gene signatures achieved the maximum ACC, that is some signatures were tied. For a given $n > 5$, among the tied $2n$-gene signatures, we picked the one with the lowest difference between ACC^{AD} and ACC^{SC} to have a more balanced prediction accuracy. If there were still tied signatures, we chose the one whose members had the lowest sum of ranks, that is we summed the ranks (Supplemental Table S1) of the genes in each signature and picked the one corresponding to the smallest sum.

The maximum achieved ACC (for the training dataset) is plotted as a function of n in Figure 2. Also shown in the figure are plots of ACC^{SC} and ACC^{AD}. In the case of ACC, the figure shows a flat maximum at $n = 3, 4, 5$ and 6. However, at $n = 3$ the figure indicates a larger difference between ACC^{SC} and ACC^{AD} in comparison with the cases of $n = 4, 5, 6$. We thus concluded that signatures with $n = 4, 5$, or 6 were superior to the 6-gene signature ($n = 3$). On the other hand, we were looking for the smallest signature that had the best performance. Therefore, since the 8-, 10-, and 12-gene signatures were tied, we picked the 8-gene signature as our LUAD/LUSC classifier. The 8 genes in the signature and their "weights," that is the elements of the vector $\mathbf{E}_0$ (see equation (1)), are given in Table 2. Note that a negative weight for a gene means that the gene is down-regulated. The biological relevance of these 8 genes are discussed in the next section.

Biological relevance of the signature

To provide evidence supporting the relevance of the identified 8 genes to lung cancer, we first conducted a literature search trying to find publications relating these genes to this cancer. We found that 5 out of the 8 genes (TP63, KRT5, CALML3, PKP1, and SPINK1) were among the list of known lung cancer genes reported by Girard et al.[5] On other hand, KRT7 has been found to be a biomarker of LUAD.[20] Additionally, it has been reported that elevation of TESC in NSCLC intensifies cancer stem cell properties and inactivation of TESC has been suggested as a way to improve current therapeutic strategies for lung cancer.[21] Also, 2 of the genes (KRT5 and TP63) are routinely used for immunohistochemical diagnosis of LUAD/LUSC,[5] and CALML3 has been recently shown to perform as well as the commonly used immunohistochemical markers.[4] We did not find literature support for involvement of C9orf152 in NSCLC. However, this does not mean that C9orf152 has no role in this cancer. Our method/signature suggests C9orf152 as a potentially new lung cancer gene to be investigated experimentally.

As an indirect evidence of biological relevance of our signature, we looked at the relation between the scores of the samples and their degrees of differentiation (tumor grades). Among datasets used in this study (Table 1), there are 15 datasets reporting tumor grades for some or all samples. Since these datasets have too few samples with certain grades, to increase statistical power, we pooled the data together and formed 2 combined datasets one for LUAD and one for LUSC. Numerical values were then assigned to the tumors based on their grades, that is 1 for "poorly differentiated," 2 for

Table 2. The 8 genes in the signature and their corresponding weights, FDRs, and fold changes.

Gene symbol	Weight	FDR	Fold change
TP63	−0.662	1.8×10^{-24}	44.2
CALML3	−0.683	2.5×10^{-24}	126.1
KRT5	−0.872	2.5×10^{-24}	58.3
PKP1	−0.138	2.5×10^{-24}	5.6
TESC	0.597	6.7×10^{-18}	10.2
SPINK1	0.768	5.6×10^{-17}	8.1
C9orf152	0.534	3.0×10^{-16}	21.2
KRT7	0.456	5.8×10^{-12}	32.0

The top-4 (bottom-4) genes are down-regulated (up-regulated).

"moderate to poorly differentiated," 3 for "moderately differentiated," 4 for "moderate to well differentiated," and 5 for "well differentiated." For LUAD and LUSC respectively the number of samples in each category is given in Supplemental Tables S2 and S3. We then calculated the rank correlation, measured by Kendall's τ, between the sample scores and their numerical grades for each subtype. The results, $\tau = 0.15$ for LUAD ($p = 2\times10^{-9}$) and $\tau = -0.10$ for LUSC ($p = 7\times10^{-3}$), indicated weak, but statistically significant, correlations. Note that $\tau > 0$ for LUAD and $\tau < 0$ for LUSC, indicating that in both subtypes the correctly classified samples with higher scores, in terms of absolute value, are more likely to have higher degrees of differentiation, an observation also made by Girard et al.[5]

The overlap between our signature and those reported in the literature (see Introduction) was also investigated. We found that our signature shares KRT5/TP63/PKP1/CALML3/TESC/SPINK1 with the 42-gene signature of Girard et al,[5] KRT5 with the 2-gene signature of Li et al,[6] KRT5/TESC/KRT7 with the 53-gene signature proposed by Charkiewicz et al,[8] KRT5 with the 5-gene signature of Wu et al,[10] CALML3 with the 13-gene signature discovered by Su et al,[11] and CALML3/PKP1/TP63 with the 50-gene signature of Hou et al.[7]

Testing the gene signature

The 8-gene signature was tested on 34 independent datasets, that is all datasets in Table 1 except for GSE41271 that was used for training. Specifically, we employed our signature (Table 2) in conjunction with equation (1) to score all samples in the 34 testing datasets. In some samples, the expression values for some of the 8 genes were missing. Thus, we ignored such genes when calculating the score of samples with missing expression values. In other words, the correlations were calculated using only genes in the signature that were also present in the sample being classified. The scores were then used to classify the samples and the 3 performance measures, that is ACC, ACC^{AD}, and ACC^{SC} were calculated. In this section we first discuss how our signature did overall and then focus on its performance in special cases, that is poorly differentiated samples, specimens obtained using biopsies, and formalin fixed paraffin-embedded (FFPE) specimens.

Overall performance. The 3 performance measures, plotted in Figure 3, and given in Supplemental Tables S4-S6, indicate overall good performance across all datasets, although some variability in the performance measures is observed (we address this issue below). To get a better sense of the overall performance of the signature, we pooled the scores of all samples from all testing datasets. Such integration, which has also been done by Girard et al[5] and Li et al,[6] is possible because the scores are correlation-based, and so they are comparable across different datasets. The distributions of LUAD and LUSC scores of all samples in the testing data are shown in Figure 4. The figure shows that a threshold of zero generally works well although the 2 distributions have long tails. The pooled scores were used to compute the overall performance measures that were $\text{ACC}_{\text{all}} = 0.936$, $\text{ACC}^{\text{AD}}_{\text{all}} = 0.951$, and $\text{ACC}^{\text{SC}}_{\text{all}} = 0.912$, demonstrating overall good performance. In comparison, for the training set we obtained $\text{ACC}_{\text{train}} = 0.939$, $\text{ACC}^{\text{AD}}_{\text{train}} = 0.956$, and $\text{ACC}^{\text{SC}}_{\text{train}} = 0.900$.

Although Figure 3 shows good performance across all datasets, some variability in ACC^{AD} and ACC^{SC} is observed from dataset to dataset. To argue that this level of variability is acceptable we note that:

1. Variability in the results has also been reported by other studies that have tested their signature on multiple datasets.[5,6] In fact, the variability in our results is smaller than those reported in these studies (see Girard et al[5] and Li et al[6] and the "Performance comparison" section).
2. All ACC^{AD} s are larger than 0.80 and the vast majority of them ($83\,\%$) are larger than $\text{ACC}^{\text{AD}}_{\text{all}} - 0.05$, where $\text{ACC}^{\text{AD}}_{\text{all}} = 0.951$ is the overall ACC^{AD} reported above. Similarly, all ACC^{SC} s are larger than 0.80 and $79\,\%$ of them are larger than $\text{ACC}^{\text{SC}}_{\text{all}} - 0.05$, indicating that in most cases the variability is small.
3. For datasets containing a small number of samples, low accuracy may be just due to chance. For example, GSE21933 contains only 11 LUAD samples. Our signature correctly classified 9 out of 11 resulting in $\text{ACC}^{\text{AD}} = 0.818$, one of the lowest reported in Figure 3. However, if $\text{ACC}^{\text{AD}}_{\text{all}} = 0.951$ is truly the prediction accuracy, there is an 8.4% chance (calculated using binomial distribution) that randomly picking 11 LUAD samples will result in $\text{ACC}^{\text{AD}} = 0.818$. Thus, we cannot reject the hypothesis that the low accuracy in this case is due to chance.

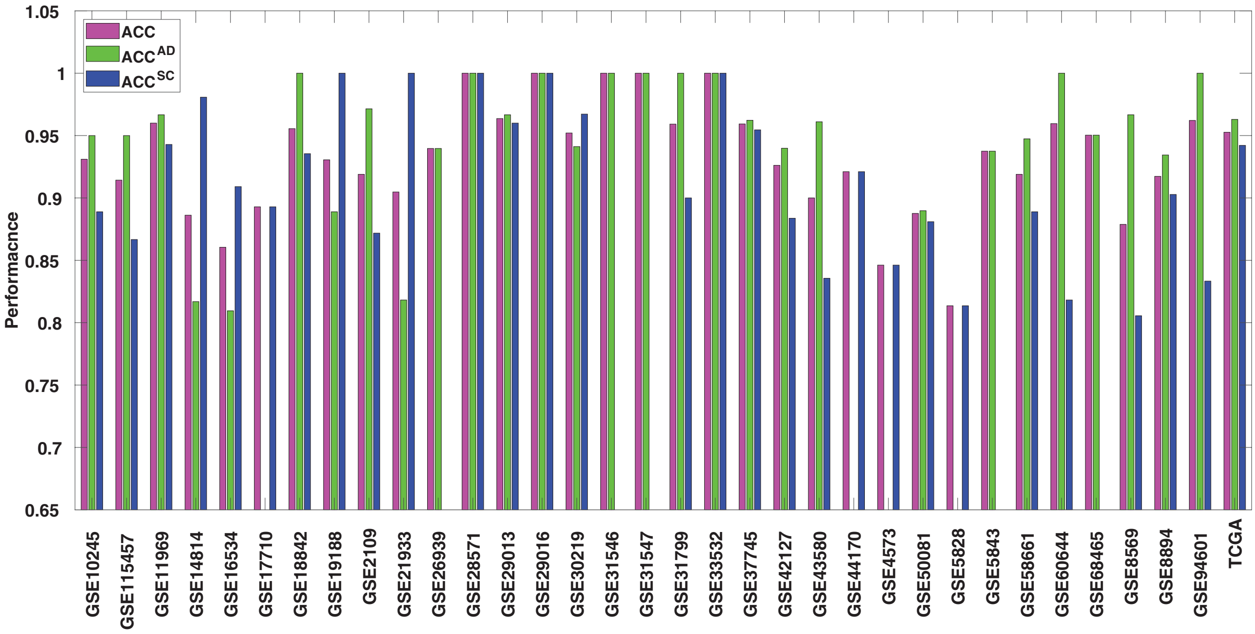

Figure 3. Testing the signature. The performance of the 8-gene signature, measured by ACC, ACC^{AD}, and ACC^{SC}, is shown for the 34 testing datasets. Some datasets do not contain samples from both subtypes, and so the corresponding bars are missing in the figure.

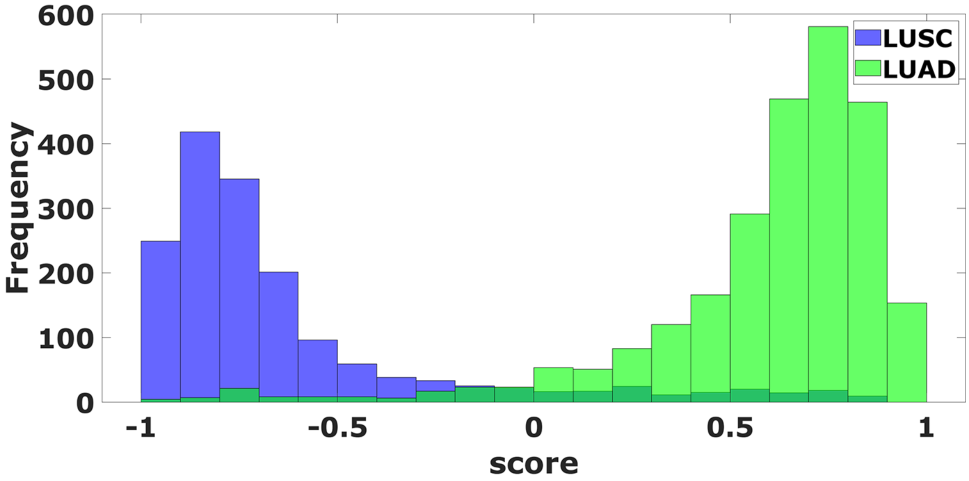

Figure 4. Distribution of scores. For the 2 subtypes, the distributions of the scores of the samples in the testing set are shown. The cutoff of zero works well for the vast majority of samples.

Poorly differentiated cases. Poorly differentiated samples are sometimes difficult to classify using immunohistochemistry, and so it is particularly important to assess the performance of our signature when these samples are concerned. We used the 15 aforementioned datasets containing tumor grade information for this assessment (see Table 1 and Supplemental Tables S2 and S3). Specifically, we partitioned the samples in these datasets into 2 groups with the poorly differentiated samples in the first group (containing 326/174 LUAD/LUSC samples), and the rest of the samples in the second group (containing 560/268 LUAD/LUSC samples). The proposed signature was then employed to classify the samples in both groups. The results, shown in Figure 5, indicate lower ACC^{AD} and ACC^{SC} in poorly differentiated samples compared to those in not-poorly differentiated samples. However, the differences are not large (only a 3% decrease in ACC^{AD} and a 9% decrease in ACC^{SC}) and the overall accuracy of $ACC = 0.890$ still indicates good performance.

Among the poorly differentiated samples, the 23 (19 LUAD and 4 LUSC) samples from GSE94601 are of particular interest because they were reclassified (from large cell lung carcinoma) as LUAD/LUSC according to WHO2015 criteria,[22] whereas the rest of the samples were classified using WHO2004 criteria. Thus, the performance of the signature was also assessed specifically for these 23 samples. The signature correctly classified all 19 LUAD samples ($ACC^{AD} = 1.000$), but misclassified 2 of the 4 LUSC samples ($ACC^{SC} = 0.500$). The low ACC^{SC}, however, may be due to chance because

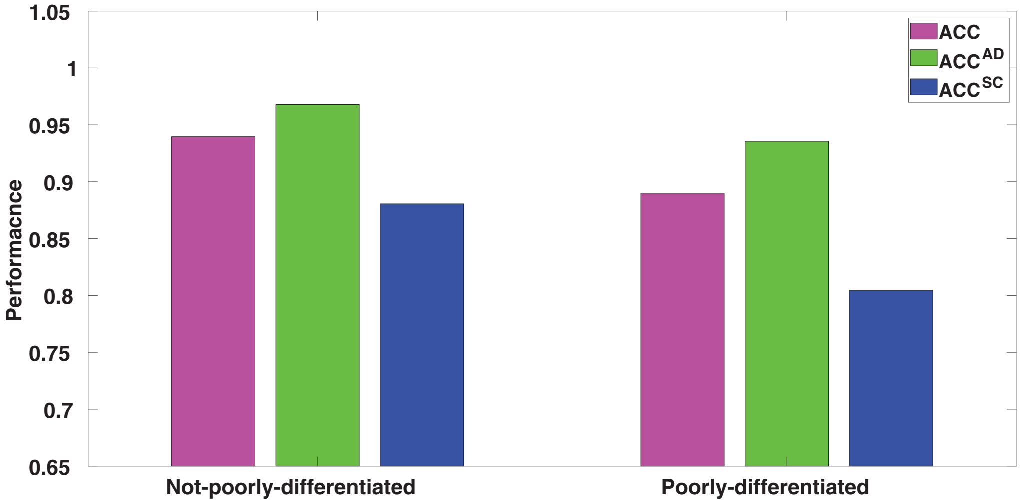

Figure 5. Poorly differentiated samples vs. not-poorly differentiated samples. The figure compares the performances of the signature in the 2 cases of poorly and not-poorly differentiated samples. The ACC is slightly lower for the poorly differentiated samples, but is still good.

the number of LUSC samples (4) is too small. To examine this hypothesis, we calculated the probability of obtaining $ACC^{SC} = 0.5$ if 4 samples were randomly picked from the 170 LUSC samples classified using WHO2004 criteria (all poorly differentiated LUSC samples but the 4 from GSE94601). This probability ($p = 0.14$) was well above the commonly used thresholds (0.01 or 0.05) for statistical significance assessment, indicating that the hypothesis that the bad performance was due to chance cannot be rejected. Simply, a larger number of samples is needed to draw a conclusion regarding the performance of our method when it comes to poorly differentiated LUSC samples classified according to WHO2015.

Specimens obtained from biopsies. Specimens obtained from biopsies are sometimes difficult to classify clinically. We found only one publicly available dataset containing such samples that is GSE58661. The rest of datasets have been obtained using surgical resection. As seen from Figure 3 and Supplemental Tables S4-S6, the signature performed well in classifying these samples with $ACC^{AD} = 0.947$ and $ACC^{SC} = 0.889$ that are only slightly lower than the overall values $ACC^{AD}_{all} = 0.951$ and $ACC^{SC}_{all} = 0.912$ mentioned above. In fact, in comparison with GSE58661, many datasets used for testing have lower ACC^{AD} and/or ACC^{SC} s (Supplemental Tables S5 and S6). Thus we conclude that the performance of the signature does not significantly depend on whether the specimen has been obtained using a biopsy or surgical resection.

FFPE specimens. Two of the datasets included in this study contain FFPE samples: GSE29013 and GSE44170. Our signature performed very well when applied to GSE29013 with $ACC = 0.964$, $ACC^{AD} = 0.967$, and $ACC^{SC} = 0.960$. For GSE44170, containing only LUSC samples, we obtained $ACC^{SC} = 0.921$ that is larger than the overall $ACC^{SC}_{all} = 0.912$. (Figure 3 and Supplemental Tables S4-S6).

Discrepant cases

Our 8-gene signature achieved an overall high ACC when applied to the testing set. However, there were some discrepant cases for which classification using our signature did not agree with the reported classification by pathologists. To understand how the discrepant (misclassified) samples differ from the correctly classified ones, we looked at the expression levels of the LUAD markers NKX2-1 and NAPSA as well as those of the LUSC markers KRT5 and TP63. These genes were chosen because they are among the best markers for immunohistochemical diagnosis of LUAD/LUSC.[5,22,23] Note that since the expression levels of a gene in different datasets may not be comparable, the samples from different datasets were not pooled for this analysis, that is each dataset was analyzed separately. For each of the 34 testing datasets, the samples were grouped in 4 sets: (1) correctly classified LUAD samples (LUAD-confirmed), (2) misclassified LUAD samples (LUAD-missed), (3) correctly classified LUSC samples (LUSC-confirmed), and (4) misclassified LUSC samples (LUSC-missed). In all datasets (except TCGA) at least one of the misclassified sets included too few (fewer than 10) samples for a reliable analysis. Thus, to have a reliable statistical significance assessment, we limited this analysis to the TCGA data with 20 LUAD and 29 LUSC misclassified samples. Specifically, the expression levels of the aforementioned 4 marker genes were compared in the 4 groups formed from the TCGA data. The results of these comparisons are given in Figure 6 and the corresponding *p*-values, calculated using the Wilcoxon rank-sum test, are given in Supplemental Table S7.

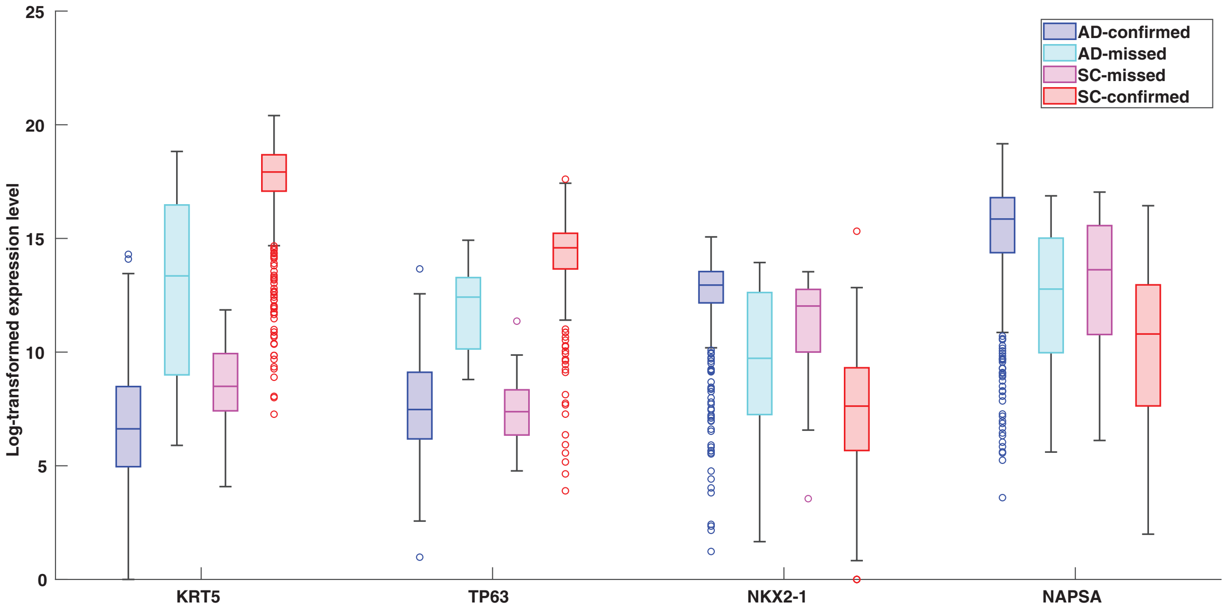

Figure 6. Boxplots comparing the expression levels of LUAD and LUSC markers in the TCGA data. The expression levels of KRT5, TP63, NKX2-1, and NAPSA are compared among 4 sets of samples: LUAD-confirmed, LUAD-missed, LUSC-missed, and LUSC-confirmed. The circles show the outliers.

As expected, Figure 6 shows that the expression levels of LUAD markers (NKX2-1, NAPSP) are higher in the LUAD-confirmed set than those in the LUSC-confirmed set. Conversely, the figure indicates LUSC markers (KRT5, TP63) have higher expression levels in the LUSC-confirmed set compared with those in the LUAD-confirmed set. Expectedly, these differences are highly statistically significant with the P-values being smaller than 10^{-100} (Supplemental Table S7). If the samples in the LUAD-missed and LUSC-missed sets are truly misclassified by our signature, one expects to observe significantly higher expression levels of LUAD (LUSC) markers in the LUAD-missed (LUSC-missed) set. However, the figure shows the opposite pattern, that is, it indicates significantly higher expression levels for LUSC markers (KRT5 and TP63) in the LUAD-missed set when compared with those in the LUSC-missed set ($p(\text{KRT5}) = 1\times10^{-3}$ and $p(\text{TP63}) = 5\times10^{-8}$; see Supplemental Table S7). Given these results and the fact that these 2 genes are among the 8 genes in the signature, it is not surprising that our method has classified the LUAD-missed samples as LUSC and the LUSC-missed samples as LUAD. The figure also indicates unexpected higher expression levels of LUAD markers (NKX2-1, NAPSP) in the LUSC-missed set compared with the LUAD-missed set, although these differences are not statistically significant (Supplemental Table S7). These observations suggest that *some* of the discrepant samples *may* have been misdiagnosed by the pathologists and that the ACC of our signature *may* be higher than what was reported in the previous section. The possibility of misdiagnosis by pathologists[5] is exactly why a signature like ours may be helpful.

It is important to note that, although KRT5, TP63, NKX2-1, and NAPSP are among the best markers for immunohistochemical diagnosis, the expression level of any one of these genes cannot be individually used as a computational tool for reliable classification. This is because: (1) as shown in the figure, there are many outliers (circles in the figure), and more importantly (2) classification requires a cutoff, and a universal cutoff that works well for different datasets is not possible to find because of batch effects. That is why our approach uses a signature comprising multiple pairs of oppositely regulated genes.

Performance comparison

We compared the performance of our 8-gene signature with those of the signatures proposed by Girard et al[5] and Li et al.[6] These 2 signatures were chosen for a couple of reasons. First, they are the only ones among signatures cited in this paper that have been tested on multiple datasets. Second, the signature (and classification method) of Girard et al[5] is the most accurate we have seen and that of Li et al[6] uses the lowest number of genes reported in the literature. Finally, like our signature, these 2 contain the same number of up- and down-regulated genes. We used the 34 testing datasets and computed the ACCs, ACC^{AD} s, and ACC^{SC} s, for the 3 signatures. The results, given in Supplemental Tables S4-S6, show that none of the 3 signatures outperforms the others in all datasets. Thus, to evaluate the overall performance of the 3 signatures, we decided to compare them using the previously mentioned pooled scores. To assess the statistical significance of the differences between the performance measures the McNemar's test was used, that is all p-values mentioned in this section were calculated using this test.

Comparison with the signature of Girard et al. The performance measures for both signatures were calculated for all samples as well as all poorly differentiated samples using the pooled scores

Table 3. Comparison with the signature of Girard et al.

	Signature	ACC	ACC^{AD}	ACC^{SC}
All (2556, 1630)	Our sig.	0.936	0.951	0.912
	Girard	0.927	0.951	0.888
PD (326, 174)	Our sig.	0.890	0.936	0.805
	Girard	0.854	0.896	0.776
Biopsy (38, 36)	Our sig.	0.919	0.947	0.889
	Girard	0.919	0.974	0.861
FFPE (30, 63)	Our sig.	0.946	0.967	0.936
	Girard	0.935	0.933	0.936

Abbreviation: PD, poorly differentiated.
For each category, the number of LUAD and LUSC samples are respectively given in parentheses.

and the results are given in Table 3. As the table indicates, when considering all samples, our signature performs better than Girard's in terms of ACC^{SC} ($p = 4.3\times10^{-5}$) and ACC ($p = 5.6\times10^{-3}$), while having the same ACC^{AD}. In the clinically important case of poorly differentiated samples, on the other hand, we achieve higher ACC^{AD} ($p = 1.6\times10^{-2}$) and ACC ($p = 2.0\times10^{-2}$). In this case our ACC^{SC} is also higher, but the difference is not statistically significant ($p = 3.3\times10^{-1}$). For completeness, in Table 3 we have also compared the performance measures for the pooled FFPE samples (GSE29013 GSE44170) and biopsy samples (GSE58661). Although there are some differences between some of the performance measures, they are not statistically significant (all $p \geq 0.5$), indicating that the 2 signatures perform comparably in these 2 cases (FFPE and biopsy samples).

The improvements reported in Table 3 are modest, but one should keep in mind that: (1) given the large number of lung cancer patients, even a modest improvement can help some patients in terms of getting the right treatment, and (2) our signature includes significantly fewer genes than that of Girard's (8 vs 42). Additionally, when applied to individual datasets, our signature performs more robustly than Girard's, resulting in less variable ACC^{SC} (Supplemental Tables S4-S6). To demonstrate this, we calculated the standard deviations of the 3 performance measures and found $\sigma(\text{ACC}_{\text{G}}^{\text{SC}}) = 0.121$, $\sigma(\text{ACC}_{\text{G}}^{\text{AD}}) = 0.047$, $\sigma(\text{ACC}_{\text{G}}) = 0.108$, $\sigma(\text{ACC}^{\text{SC}}) = 0.062$, $\sigma(\text{ACC}^{\text{AD}}) = 0.055$, and $\sigma(\text{ACC}) = 0.045$. Here σ denotes standard deviation, and subscript $_{\text{G}}$ shows that the value has been calculated using the Girard signature. These values indicate a comparable σ for ACC^{AD}, but a larger variability in ACC^{SC} (and consequently in ACC) when the Girard's signature is used.

Comparison with the signature of Li et al. We followed the same procedure to compare the performance of our signature with that of Li et al. However, since GSE30219 and GSE18842 were used by Li *et al.* to train their model, these datasets were excluded. Three additional samples were excluded because Li's signature was not able to classify them due to missing expression values. The performance comparison results are given in Table 4. As indicated in the table, there are no differences in the performance measures for FFPE and biopsy samples. Considering all samples, however, the signature proposed in this paper outperforms Li's signature in terms of ACC^{SC} ($p = 4.2\times10^{-10}$) and ACC ($p = 1.0\times10^{-6}$), while having a comparable ACC^{AD}. In the case of poorly differentiated samples, Table 4 again indicates our signature achieves significantly higher ACC^{SC} ($p = 1.2\times10^{-2}$) and ACC ($p = 1.4\times10^{-2}$). (In this case our ACC^{AD} is also slightly higher, but the difference is not statistically significant with $p = 3.4\times10^{-1}$). Additionally, our signature is more robust when applied to individual datasets (Supplemental Tables S4-S6) with $\sigma(\text{ACC}^{\text{SC}}) = 0.062$, $\sigma(\text{ACC}^{\text{AD}}) = 0.055$, and $\sigma(\text{ACC}) = 0.045$ compared with $\sigma(\text{ACC}_{\text{L}}^{\text{SC}}) = 0.115$, $\sigma(\text{ACC}_{\text{L}}^{\text{AD}}) = 0.080$, and $\sigma(\text{ACC}_{\text{L}}) = 0.107$. Here, the subscript $_{\text{L}}$ refers to Li's signature.

Discussion

Although in most cases pathologists can easily distinguish LUAD from LUSC, the classification is sometimes difficult especially in poorly differentiated tumors and when biopsy is used to obtain the specimen. As a result many patients are not diagnosed with a specific subtype or are misdiagnosed. The fact that re-examination of the specimens sometimes results in reclassification provides evidence for possible misdiagnoses.[5] Since proper treatment depends on precise classification, such misdiagnoses can have important implications for many patients. A computational method to help pathologists better classify the NSCLC subtypes (in challenging cases) is thus beneficial.

To accurately distinguish LUAD from LUSC, in this paper we propose an 8-gene signature identified using a simple gene selection method in conjunction with a correlation-based nearest neighbor classification approach. The approach is based on the observation that the relative orderings of expression levels

Table 4. Comparison with the signature of Li et al.

	Signature	ACC	ACC^{AD}	ACC^{SC}
All (2456, 1536)	Our sig.	0.935	0.951	0.910
	Li	0.918	0.948	0.869
PD (321, 159)	Our sig.	0.888	0.935	0.792
	Li	0.850	0.919	0.712
Biopsy (38, 36)	Our sig.	0.919	0.947	0.889
	Li	0.919	0.947	0.889
FFPE (30, 63)	Our sig.	0.946	0.967	0.936
	Li	0.946	0.967	0.936

Abbreviation: PD, poorly differentiated.
For each category, the number of LUAD and LUSC samples are respectively given in parentheses.

are not affected by batch effects, which suggests a signature containing both highly up- and down-regulated genes in conjunction with the correlation-based scoring is likely to perform well across different datasets.[6] On the other hand, genes with more statistically significant fold changes are expected to be more discriminative for classifying LUAD versus LUSC. Thus, we devised a discovery method in which the signature has $2n$ genes, half of which are the top n genes from a list of HDEGs, ranked based on the statistical significance of their fold changes. The second half is found by examining a huge number of sets of n genes that are oppositely regulated with respect to the genes in the first half. For a given n, the optimal gene signature is then chosen among these candidate $2n$-gene sets based on their calculated ACCs.

In this approach the two-halves of the signature are treated differently, that is the first half is fixed while the second half is chosen assuming the first half is present in the signature. To explain why this approach works, we note that the expression level of each high-ranking gene in a ranked list of HDEGs can be used to classify the samples (by ranking the samples using the expression levels, and applying an appropriate cutoff). With this approach even a single high-ranking gene can have a high discriminative power.[4] However, such a 1-gene signature is not ideal, because in practice the cutoff is generally dataset-dependent due to batch effects. The remedy is to add a highly oppositely regulated gene, making a 2-gene signature, and to use correlation-based scores rather than expression levels to rank the samples. However, one should note that in this scenario the second gene has a secondary role of introducing a natural cutoff of zero that is not expected to be dataset-dependent.

The generalization of the approach mentioned in the previous paragraph is to use a $2n$-gene signature by choosing the top n genes (in the ranked HDEG list) as the first half. An alternative approach to discovering such $2n$-gene signatures is to consider all possible combinations of n up- and n down-regulated genes. However, this approach may result in signatures that are overtrained. (Additionally, this approach is prohibitively computationally expensive except for small values of n). To demonstrate this problem, we considered all possible combinations of 4 up- and 4 down-regulated genes in our HDEG list, treating the two-halves equally. We then found the 8 genes that maximized ACC in our training dataset (using the same tie breaker as before). The resulting 8-gene set had higher performance measures (when applied to the training set) than those of our 8-gene signature given in Table 2 ($ACC = 0.951$, $ACC^{AD} = 0.962$, $ACC^{SC} = 0.925$ compared with $ACC_{train} = 0.939$, $ACC^{AD}_{train} = 0.956$, $ACC^{SC}_{train} = 0.900$). However, when this 8-gene set was applied to the testing data, the performance measures were lower than the ones obtained using our signature ($ACC = 0.905$, $ACC^{AD} = 0.948$, $ACC^{SC} = 0.839$ compared with $ACC_{all} = 0.936$, $ACC^{AD}_{all} = 0.951$, and $ACC^{SC}_{all} = 0.912$). Our results show that this overtraining can be avoided by applying the constraint that, in a $2n$-gene signature, the first half must be the top n genes in the ranked HDEGs list, which guarantees the highest-ranking, most discriminative genes are included in the signature.

Our signature is easy to use. Given a sample, the correlation between its 8-gene expression profile and the vector $\mathbf{E}_0$ (Table 2) is calculated and the sample is classified as LUAD (LUSC) if the correlation is non-negative (negative). In addition to its ease of use, the extensive testing of our 8-gene signature demonstrated good performances across many datasets, obtained using different technologies/platforms. For example, although we used microarray data for training and discovery, good performance measures were observed when our signature was applied to RNA-seq (TCGA) data. Our signature performed well in clinically challenging cases including poorly differentiated samples and specimens obtained from biopsies. Most of the genes in our signature turned out to be known lung cancer genes and almost all of them were also included in other gene signatures published previously. However, in comparison with 2 of the most accurate signatures previously proposed, our signature had a better performance.

A limitation of this study is that it focuses on only the 2 major subtypes of NSCLC. Of course, there are other NSCLC subtypes and also small-cell lung carcinoma, which have not been considered in this study and cannot be classified using the 8-gene signature. (Any sample from subtypes other than LUAD and LUSC would be classified as either LUAD or LUSC). As an important note, we emphasize that the signature proposed here is not meant to replace pathologists, that is it is to be used as a tool to assist pathologists classify challenging cases that are known to be either LUAD or LUSC. Additionally, it should be noted that all studies cited in this paper have also limited their scope to these subtypes. Given the prevalence of LUAD and LUSC, gene signatures focusing on only these subtypes can still be helpful to pathologists to achieve a more accurate classification. Another limitation of this paper, and others cited here, is that most samples used for testing the signature have been classified according to WHO2004 criteria. This is because few samples classified according to WHO2015 criteria are publicly available. We tested our signature using the samples (19 LUAD and 4 LUSC) in GSE94601 that were reclassified (according to WHO2015) to either LUAD or LUSC. Our signature was 100% accurate in classifying the 19 LUAD samples. However, the number of LUSC samples is just too small to draw a conclusion in the case of the LUSC (only 4 samples). Despite these limitations, which are shared by other studies cited, the good results obtained here suggest our signature can be useful to pathologists.

Acknowledgements

This study utilized the computational resources of the Biowulf system at the National Institutes of Health, Bethesda, MD (http://biowulf.nih.gov). The results published here are in part based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga.

Author Contributions

MH and YKY designed research; MH performed research; MH and YKY analyzed data and wrote the paper.

ORCID iD

Yi-Kuo Yu https://orcid.org/0000-0002-6213-7665

Supplemental Material

Supplemental material for this article is available online.

References

1. The American Thoracic Society and The European Respiratory Society. Pretreatment evaluation of non-small-cell lung cancer. *Am J Respir Crit Care Med*. 1997;156:320-332.
2. Travis WD. Pathology of lung cancer. *Clin Chest Med*. 2011;32:669-692.
3. Travis WD, Brambilla E, Nicholson AG, et al. The 2015 world health organization classification of lung tumors: impact of genetic, clinical and radiologic advances since the 2004 classification. *J Thorac Oncol*. 2015;10:1243-1260.
4. Zhan C, Yan L, Wang L, et al. Identification of immunohistochemical markers for distinguishing lung adenocarcinoma from squamous cell carcinoma. *J Thorac Dis*. 2015;7:1398-1405.
5. Girard L, Rodriguez-Canales J, Behrens C, et al. An expression signature as an aid to the histologic classification of non-small cell lung cancer. *Clin Cancer Res*. 2016;22:4880-4889.
6. Li X, Shi G, Chu Q, et al. A qualitative transcriptional signature for the histological reclassification of lung squamous cell carcinomas and adenocarcinomas. *BMC Genomics*. 2019;20:881.
7. Hou J, Aerts J, den Hamer B, et al. Gene expression-based classification of non-small cell lung carcinomas and survival prediction. *PLoS One*. 2010;5:e10312.
8. Charkiewicz R, Niklinski J, Claesen J, et al. Gene expression signature differentiates histology but not progression status of early-stage NSCLC. *Transl Oncol*. 2017;10:450-458.
9. Huang Z, Chen L, Wang C. Classifying lung adenocarcinoma and squamous cell carcinoma using RNA-seq data. *Cancer Stud Mol Med Open J*. 2017;3:27-31.
10. Wu X, Wang L, Feng F, Tian S. Weighted gene expression profiles identify diagnostic and prognostic genes for lung adenocarcinoma and squamous cell carcinoma. *J Int Med Res*. 2020;48:0300060519893837.
11. Su R, Zhang J, Liu X, Wei L. Identification of expression signatures for non-small-cell lung carcinoma subtype classification. *Bioinformatics*. 2020; 36:339-346.
12. Cancer Genome Atlas Research Network. Comprehensive genomic characterization of squamous cell lung cancers. *Nature*. 2012;489:519-525.
13. The Cancer Genome Atlas Research Network. Comprehensive molecular profiling of lung adenocarcinoma. *Nature*. 2014;511:543-550.
14. Grossman RL, Heath AP, Ferretti V, et al. Toward a shared vision for cancer genomic data. *N Engl J Med*. 2016;375:1109-1112.
15. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with deseq2. *Genome Biol*. 2014;15:550.
16. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI. Gene expression and hybridization array data repository. *Nucleic Acids Res*. 2002; 30:207-210.
17. Durinck S, Spellman PT, Birney E, Huber W. Mapping identifiers for the integration of genomic datasets with the r/bioconductor package biomart. *Nat Protoc*. 2009;4:1184-1191.
18. Sato M, Larsen JE, Lee W, et al. Human lung epithelial cells progressed to malignancy through specific oncogenic manipulations. *Mol Cancer Res*. 2013; 11:638-650.
19. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Ser B*. 1995;57:289-300.
20. Chen Z, Fillmore CM, Hammerman PS, Kim CF, Wong KK. Non-small-cell lung cancers: a heterogeneous set of diseases. *Nat Rev Cancer*. 2014;14: 535-546.
21. Lee JH, Choi SI, Kim RK, Cho EW, Kim IG. Tescalcin/c-Src/IGF1Rβ-mediated stat3 activation enhances cancer stemness and radioresistant properties through aldh1. *Sci Rep*. 2018;8:1-13.
22. Karlsson A, Brunnström H, Micke P, et al. Gene expression profiling of large cell lung cancer links transcriptional phenotypes to the new histological who 2015 classification. *J Thorac Oncol*. 2017;12:1257-1267.
23. Kim MJ, Shin HC, Shin KC, Ro JY. Best immunohistochemical panel in distinguishing adenocarcinoma from squamous cell carcinoma of lung: tissue microarray assay in resected lung cancer specimens. *Ann Diagn Pathol*. 2013; 17:85-90.