Identification of tumor-educated platelet biomarkers of non-small-cell lung cancer

Meiling Sheng¹,*
Zhaohui Dong²,*
Yanping Xie³

¹Department of Respiration, Jinhua People’s Hospital, Jinhua, Zhejiang 321000, China; ²Department of Intensive Care Unit, First Hospital of Huzhou, First Affiliated Hospital of Huzhou University, Huzhou, Zhejiang 313000, China; ³Department of Respiratory Medicine, First Hospital of Huzhou, First Affiliated Hospital of Huzhou University, Huzhou, Zhejiang 313000, China

*These authors contributed equally to this work

Background: Lung cancer is a severe cancer with a high death rate. The 5-year survival rate for stage III lung cancer is much lower than stage I. Early detection and intervention of lung cancer patients can significantly increase their survival time. However, conventional lung cancer-screening methods, such as chest X-rays, sputum cytology, positron-emission tomography (PET), low-dose computed tomography (CT), magnetic resonance imaging, and gene-mutation, -methylation, and -expression biomarkers of lung tissue, are invasive, radiational, or expensive. Liquid biopsy is non-invasive and does little harm to the body. It can reflect early-stage dysfunctions of tumorigenesis and enable early detection and intervention.

Methods: In this study, we analyzed RNA-sequencing data of tumor-educated platelets (TEPs) in 402 non-small-cell lung cancer (NSCLC) patients and 231 healthy controls. A total of 48 biomarker genes were selected with advanced minimal-redundancy, maximal-relevance, and incremental feature-selection (IFS) methods.

Results: A support vector-machine (SVM) classifier based on the 48 biomarker genes accurately predicted NSCLC with leave-one-out cross-validation (LOOCV) sensitivity, specificity, accuracy, and Matthews correlation coefficients of 0.925, 0.827, 0.889, and 0.760, respectively. Network analysis of the 48 genes revealed that the WASF1 actin cytoskeleton module, PRKAB2 kinase module, RSRC1 ribosomal protein module, PDHB carbohydrate-metabolism module, and three intermodule hubs (TPM2, MYL9, and PPP1R12C) may play important roles in NSCLC tumorigenesis and progression.

Conclusion: The 48-gene TEP liquid-biopsy biomarkers will facilitate early screening of NSCLC and prolong the survival of cancer patients.

Keywords: tumor-educated platelet, TEP, liquid biopsy, minimal redundancy, maximal relevance, MRMR, incremental feature selection, IFS, non-small-cell lung cancer, NSCLC

Introduction

Lung cancer is a severe cancer with a high death rate.¹,² Early detection of lung cancer is the most effective way to increase survival time, since survival time is directly associated with lung cancer stage and early-treatment patients will have better diagnoses.³ The 5-year survival rates for stage I and stage III lung cancer patients are 67% and 23%, respectively.³ The survival difference between early-stage and late-stage lung cancer is huge. Therefore, early screening of lung cancer is the key to lung cancer prevention and therapy.

Conventionally, lung cancer is detected through chest X-rays, sputum cytology, positron-emission tomography (PET), low-dose computed tomography (CT), and magnetic resonance imaging.⁴ However, many diagnosed patients are already in late stages.⁵ Although PET and CT are developing progressively higher resolutions and can detect smaller tumors, they are radiational and expensive.
In recent years, sequencing technologies have developed rapidly. It has been found that tumor tissue can release small numbers of tumor cells, DNA, RNA, or exosomes into blood. These tumor cells in blood are called circulating tumor cells (CTCs). Nowadays, CTCs can be isolated and DNA and RNA with CTCs sequenced accurately. Other types of liquid-biopsy components include ctDNA, ctRNA, exosomes, and tumor-educated platelets (TEPs). Tumor-derived exosomes contain various molecules, such as dsDNA and small RNA, and can reflect the status of tumor cells. TEPs are blood platelets that contain tumor RNAs. They are a great source of tumor-derived RNAs. There have been several studies showing that TEP RNAs can be cancer biomarkers. Liquid biopsy has become ever more important in early lung cancer detection and is the one of the foundations of personalized medicine. It can reflect early-stage dysfunctions of tumorigenesis and enable early detection and intervention.

In this study, we analyzed RNA-sequencing data of TEPs in 402 non-small-cell lung cancer (NSCLC) patients and 231 healthy controls. By comparing their expression differences with the minimal redundancy, maximal relevance (MRMR) method, differentially expressed genes were ranked. Then, with incremental feature selection (IFS), optimal biomarkers were selected. Finally, a support vector machine (SVM) classifier based on the optimal biomarkers was constructed and evaluated. TEP biomarkers could be a useful way to enable early intervention in lung cancer patients and prolong their survival.

**Methods**

**Blood gene-expression profiles of NSCLC**

Blood gene-expression profiles of NSCLC patients were downloaded from the Gene Expression Omnibus with accession number GSE89843 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE89843). There were 402 NSCLC samples and 231 healthy control samples. Samples with chronic pancreatitis, epilepsy, multiple sclerosis, insignificant atherosclerosis, pulmonary hypertension, stable angina pectoris, and unstable angina pectoris were excluded. Expression levels of 4,722 genes in TEPs were measured using RNA sequencing. We considered the 402 NSCLC samples as positive samples, the 231 healthy control samples as negative samples, and the expression levels of the 4,722 genes as classification features. The goal was to identify the differentially expressed genes between NSCLC and healthy controls and construct an effective TEP-biomarker-based NSCLC classifier. The workflow of TEP-biomarker-based NSCLC-classifier construction is shown in Figure 1. First, TEP data were preprocessed as a matrix with rows of samples and columns of genes. Then, genes were ranked with the MRMR method. After MRMR, genes were all ranked. Then, with the ranked-gene list, incremental feature selection was adopted to optimize the biomarker-gene set. Finally, biomarkers were determined and the final SVM classifier constructed.

**Figure 1** Workflow of TEP biomarker-based NSCLC classifier construction.

**Notes:** First, TEP data were preprocessed as a matrix with rows of samples and columns of genes. Then, genes were ranked with the MRMR method. After MRMR, genes were all ranked. Then, with the ranked-gene list, incremental feature selection was adopted to optimize the biomarker-gene set. Finally, biomarkers were determined and the final SVM classifier constructed.

**Abbreviations:** TEP, tumor-educated platelet; NSCLC, non-small-cell lung cancer; MRMR, minimal redundancy, maximal relevance; SVM, support vector machine; LOOCV, leave-one-out cross-validation; MCC, Matthews correlation coefficient.
and columns with genes. Then, genes were ranked with the MRMR method. After MRMR, the genes were all ranked. Then, with the ranked-gene list, the IFS method was used to optimize the biomarker-gene set. Finally, biomarkers were determined and the final SVM classifier constructed. Each step is illustrated in the following sections.

**Biomarker-gene selection based on MRMR and IFS methods**

We used the MRMR method to rank the genes based on their relevance with sample labels (NSCLC or healthy controls) and redundancy between genes. To illustrate this method clearly, let us use \( \Omega_s \), \( \Omega_r \), and \( \Omega_m \) to represent the complete set of candidate genes for biomarker ranking, the selected \( m \) biomarker genes, and the to-be-selected \( n \) genes, respectively. The relevance of gene \( g \) from \( \Omega_s \) with sample type \( t \) can be measured with mutual information \( (D) \):\(^{16,17} \)

\[
D = I(g, t) 
\]

After we defined mutual information, the redundancy \( (R) \) of the gene \( g \) with the selected biomarker genes in \( \Omega_s \) can be calculated:

\[
R = \frac{1}{m} \left( \sum_{g,i} I(g, g_i) \right) 
\]

To select the best gene \( g_k \) from \( \Omega_s \) that can maximize its relevance with sample type \( t \) and minimize its redundancy with the selected biomarker genes in \( \Omega_s \), we need to maximize the MRMR function:

\[
\max_{g_i \in \Omega_r} \left[ I(g_j, t) - \frac{1}{m} \sum_{g,i} I(g_j, g_i) \right] (j = 1, 2, ..., n) 
\]

After \( n \) rounds of evaluation, a ranked-gene list can be obtained:

\[
S = \{ g'_1, g'_2, ..., g'_k, ..., g'_n \} 
\]

The position of a gene in this ranked list \( (h) \) reflects the trade-off between relevance with sample classes, ie, whether a sample is NSCLC, and redundancy with selected biomarker genes, ie, genes with smaller index values. The genes on the top are better than the genes on the bottom.

To reduce computation complexity, we analyzed only the top 500 MRMR genes. To determine how many genes should be selected to form the optimal biomarkers, we adopted the IFS method\(^{18–23} \) and constructed 500 SVM classifiers. In this study, we used the SVM function with default parameters from R package e1071 (https://cran.r-project.org/web/packages/e1071) to build the SVM classifier. Each time, a candidate gene set \( S_k = \{ g'_1, g'_2, ..., g'_k \} (1 \leq k \leq 500) \) of the top \( k \) genes in the MRMR list was used to build the SVM classifier. The performance of the top \( k \)-gene classifier was evaluated with leave-one-out cross-validation (LOOCV).

Finally, an IFS curve was plotted, with the top genes used as the \( x \)-axis and the LOOCV Matthews correlation coefficients (MCCs) of classifiers as the \( y \)-axis. Based on the IFS curve, we can decide how many genes should be used to build a classifier with great performance and little complexity. Usually, the peak or the change point of the IFS curve was chosen.

**Prediction-performance evaluation of SVM classifier**

As mentioned, LOOCV, also known as jackknife testing, was used to evaluate the prediction performance of each SVM classifier. LOOCV continues for \( n \) rounds to test all samples one by one. In each round of LOOCV, one sample was tested while the other samples were trained. After \( n \) rounds, all samples were tested one at a time. LOOCV is widely used to evaluate prediction performance.\(^{24,25} \) Although the independent test has also been widely used, the selection of independent-test samples is arbitrary, and sometimes the choice of different validation cohorts may lead to totally different conclusions, as the validation samples may have different distributions from the training samples. Cross-validation can overcome these problems.\(^{26} \)

By comparing the predicted sample classes with the actual sample classes, sensitivity \( (Sn) \), specificity \( (Sp) \), accuracy \( (ACC) \), and MCC were calculated to evaluate prediction performance:

\[
Sn = \frac{TP}{TPFN} 
\]

\[
Sp = \frac{TN}{TNFP} 
\]

\[
ACC = \frac{TPTN}{TPTNFPFN} 
\]

\[
MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TPFP)(TPFN)(TNFP)(TNFN)}} 
\]

where TP, TN, FP, and FN stand for true positive (NSCLC), true negative (healthy control), false positive (NSCLC), and false negative (healthy control), respectively. Since the
sizes of positive (NSCLC) and negative (healthy control) samples were imbalanced in this study, MCC was a better measurement than ACC. MCC considered both sensitivity and specificity.  

Results and discussion
Genes showing different expression patterns between NSCLC and healthy controls
We obtained the top 500 most discriminative genes of NSCLC and healthy control samples using the MRMR method. The MRMR method is based on information theory. Mutual information is used to measure relevance and redundancy. It has been widely used in the bioinformatics field.  

We used a C/C++ version of MRMR software (http://home.penglab.com/proj/mRMR) to apply the gene-ranking process. Unlike statistical test methods, such as the t-test for case-control experiment design and ANOVA for multiple-group design, MRMR not only considers the relevance between genes and sample classes but also redundancy between genes.

Optimal biomarkers identified from MRMR gene list with IFS methods
After MRMR analysis, we applied the IFS procedure to select the optimal number of top MRMR genes to form the biomarker-gene set. The relationship between the number of genes and prediction MCCs was plotted as an IFS curve (Figure 2). It can be seen that when 266 genes were used, the LOOCV MCC was the highest \(0.764\), but even early, when only 48 genes were used, the MCC was 0.760. To consider both using fewer genes and achieving higher prediction MCC, we chose the 48 genes as the optimal biomarker-gene set, since increasing the number of genes beyond 48 would not significantly increase the MCC any more. The 48 genes are shown in Table 1.

Prediction performance of the 48-gene classifier
The 48 genes were chosen based on MRMR and IFS methods. To evaluate their prediction power objectively, we calculated LOOCV sensitivity, specificity, accuracy, and MCC. The confusion matrix of predicted sample classes and actual sample classes is shown in Table 2. LOOCV sensitivity, specificity, accuracy and MCC of the 48-gene classifier were 0.925, 0.827, 0.889, and 0.760, respectively.

To demonstrate more intuitively the discriminative power of these 48 genes for NSCLC and healthy control samples, we draw a heat map using these 48 genes (Figure 3). It can be seen that even without an advanced machine-learning algorithm, such as SVM, the simple hierarchical clustering can group most NSCLC and healthy control samples into the right clusters. Upregulation and downregulation patterns of these 48 genes were very clear between NSCLC and healthy control samples.

Biological significance of the 48 biomarker genes
To explore the regulatory mechanisms of the 48 genes, we mapped them onto Search Tool for the Retrieval of Interacting Genes/Proteins (STRING), a comprehensive and widely used protein functional association network. The subnetwork of these 48 genes extracted from STRING is shown in Figure 4, with selected genes highlighted in red. It can be seen that there were several modules on the network that were circled together.

On the bottom left is the WASF1 module which included MYO5A and WASF1. These two genes both interacted with
Table 1 The 48 genes selected by MRMR and IFS methods

<table>
<thead>
<tr>
<th>Order</th>
<th>Gene ID</th>
<th>Symbol</th>
<th>Score</th>
<th>Order</th>
<th>Gene ID</th>
<th>Symbol</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>ENSG0000008154</td>
<td>PCNP</td>
<td>0.104</td>
<td>25</td>
<td>ENSG00000101335</td>
<td>MYL9</td>
<td>-0.007</td>
</tr>
<tr>
<td>2</td>
<td>ENSG00000238683</td>
<td>SNRD13</td>
<td>0.047</td>
<td>26</td>
<td>ENSG00000120963</td>
<td>ZNF706</td>
<td>-0.008</td>
</tr>
<tr>
<td>3</td>
<td>ENSG00000163703</td>
<td>CRELD1</td>
<td>0.045</td>
<td>27</td>
<td>ENSG00000106034</td>
<td>CPED1</td>
<td>-0.007</td>
</tr>
<tr>
<td>4</td>
<td>ENSG00000111371</td>
<td>SLC38A1</td>
<td>0.034</td>
<td>28</td>
<td>ENSG000000227165</td>
<td>WDR11-A51</td>
<td>-0.011</td>
</tr>
<tr>
<td>5</td>
<td>ENSG00000198627</td>
<td>TPM2</td>
<td>0.034</td>
<td>29</td>
<td>ENSG00000167100</td>
<td>SAMD14</td>
<td>-0.01</td>
</tr>
<tr>
<td>6</td>
<td>ENSG00000101849</td>
<td>TBL1X</td>
<td>0.026</td>
<td>30</td>
<td>ENSG00000103316</td>
<td>CRIM</td>
<td>-0.01</td>
</tr>
<tr>
<td>7</td>
<td>ENSG00000130948</td>
<td>HSD17B3</td>
<td>0.027</td>
<td>31</td>
<td>ENSG00000154146</td>
<td>NRGN</td>
<td>-0.013</td>
</tr>
<tr>
<td>8</td>
<td>ENSG00000142089</td>
<td>IFITM3</td>
<td>0.021</td>
<td>32</td>
<td>ENSG00000168291</td>
<td>PDHB</td>
<td>-0.013</td>
</tr>
<tr>
<td>9</td>
<td>ENSG00000156738</td>
<td>MS4A1</td>
<td>0.021</td>
<td>33</td>
<td>ENSG00000067836</td>
<td>ROGD1</td>
<td>-0.014</td>
</tr>
<tr>
<td>10</td>
<td>ENSG00000151575</td>
<td>TEX9</td>
<td>0.024</td>
<td>34</td>
<td>ENSG00000112290</td>
<td>WASF1</td>
<td>-0.015</td>
</tr>
<tr>
<td>11</td>
<td>ENSG00000125534</td>
<td>PPP1F</td>
<td>0.022</td>
<td>35</td>
<td>ENSG00000266356</td>
<td>AC090615.1</td>
<td>-0.013</td>
</tr>
<tr>
<td>12</td>
<td>ENSG00000125503</td>
<td>PPP1R12C</td>
<td>0.018</td>
<td>36</td>
<td>ENSG00000128731</td>
<td>HERC2</td>
<td>-0.017</td>
</tr>
<tr>
<td>13</td>
<td>ENSG00000248242</td>
<td>AC004053.1</td>
<td>0.019</td>
<td>37</td>
<td>ENSG00000156639</td>
<td>ZFAND3</td>
<td>-0.017</td>
</tr>
<tr>
<td>14</td>
<td>ENSG00000173598</td>
<td>NUDT4</td>
<td>0.019</td>
<td>38</td>
<td>ENSG00000136108</td>
<td>CKA2</td>
<td>-0.017</td>
</tr>
<tr>
<td>15</td>
<td>ENSG00000172037</td>
<td>LAMB2</td>
<td>0.012</td>
<td>39</td>
<td>ENSG00000213465</td>
<td>AR2</td>
<td>-0.017</td>
</tr>
<tr>
<td>16</td>
<td>ENSG00000185909</td>
<td>KLHC4B</td>
<td>0.006</td>
<td>40</td>
<td>ENSG00000174891</td>
<td>RSC1</td>
<td>-0.019</td>
</tr>
<tr>
<td>17</td>
<td>ENSG00000248309</td>
<td>MEF2C-A51</td>
<td>0.007</td>
<td>41</td>
<td>ENSG00000173083</td>
<td>HPSE</td>
<td>-0.019</td>
</tr>
<tr>
<td>18</td>
<td>ENSG00000163359</td>
<td>COL6A3</td>
<td>0.006</td>
<td>42</td>
<td>ENSG00000119684</td>
<td>MLH3</td>
<td>-0.02</td>
</tr>
<tr>
<td>19</td>
<td>ENSG00000197535</td>
<td>MYOSA</td>
<td>0.005</td>
<td>43</td>
<td>ENSG00000172543</td>
<td>CTSW</td>
<td>-0.02</td>
</tr>
<tr>
<td>20</td>
<td>ENSG00000166311</td>
<td>SMPD1</td>
<td>0.001</td>
<td>44</td>
<td>ENSG00000131791</td>
<td>PRKAB2</td>
<td>-0.022</td>
</tr>
<tr>
<td>21</td>
<td>ENSG00000162664</td>
<td>ZNF326</td>
<td>-0.001</td>
<td>45</td>
<td>ENSG00000125821</td>
<td>DTD1</td>
<td>-0.022</td>
</tr>
<tr>
<td>22</td>
<td>ENSG00000113761</td>
<td>ZNF346</td>
<td>-0.001</td>
<td>46</td>
<td>ENSG00000137312</td>
<td>FLOT1</td>
<td>-0.023</td>
</tr>
<tr>
<td>23</td>
<td>ENSG00000108010</td>
<td>GLRX3</td>
<td>-0.004</td>
<td>47</td>
<td>ENSG00000259483</td>
<td>AL355073.1</td>
<td>-0.022</td>
</tr>
<tr>
<td>24</td>
<td>ENSG00000108679</td>
<td>LGALS3BP</td>
<td>-0.003</td>
<td>48</td>
<td>ENSG00000076043</td>
<td>REXO2</td>
<td>-0.025</td>
</tr>
</tbody>
</table>

Abbreviations: MRMR, minimal redundancy, maximal relevance; IFS, incremental feature selection.

Table 2 Confusion matrix of predicted sample classes and actual sample classes using 48 genes

<table>
<thead>
<tr>
<th></th>
<th>Actual NSCLC</th>
<th>Actual healthy controls</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predicted NSCLC</td>
<td>372</td>
<td>30</td>
</tr>
<tr>
<td>Predicted healthy controls</td>
<td>40</td>
<td>191</td>
</tr>
</tbody>
</table>

Abbreviation: NSCLC, non-small-cell lung cancer.
metabolism. MLH3 and SLC38A1 were less connected with these carbohydrate metabolism genes than PDHB. Also, it has been reported that the haplotype MSH3 was associated with lung cancer and SLC38A1 significantly overexpressed in NSCLC. At the top middle was the RSRC1 module, which included RSRC1 and FLOT1. Within this module, eight genes (RPS11, RPS14, RPS15, RPS26, RPS28, RPS3, RPS3A, and RPS9) that RSRC1 interacted with were ribosomal protein genes. Ribosome is important for protein biosynthesis, and there have been several reports that downregulation of ribosomal protein can inhibit or attenuate NSCLC growth and migration. Also, they have been considered oncogenes of NSCLC.

Another gene was FLOT1. It has been reported that in NSCLC, the expression of FLOT1 was abnormal and correlated with tumor progression and poor survival.

To summarize, the possible biological mechanism of the NSCLC TEP biomarkers is shown in Figure 5. The intermodule hub genes, including TPM2, MYL9, and PPP1R12C, stitched together the WASF1 module, which regulated actin cytoskeleton, the PRKAB2 module, which was involved in the AMPK–EGFR pathway, and the PDHB module, which was involved in carbohydrate metabolism. The PDHB module interacted with the RSRC1 module, which was associated with protein biosynthesis, growth, and migration.

**Conclusion**

Early detection of lung cancer is critical for NSCLC patients, since early-stage patients have much longer survival than late-stage patients. Unfortunately, conventional lung cancer screening, such as chest X-rays, sputum cytology, PET, CT, and magnetic resonance imaging, are invasive, radiational, or expensive. Liquid biopsy makes early detection possible, since CTC, ctDNA, ctRNA, exosomes, and TEP reflect early changes during tumorigenesis. By analyzing TEP RNA-sequencing data of NSCLC patients and healthy controls, we identified 48 TEP biomarkers. These biomarkers can accurately predict NSCLC. In-depth biological network analysis suggested that there were four modules and three intermodule hubs that may trigger NSCLC. Our results provided novel insights into tumorigenesis and a useful tool for early detection and treatment of NSCLC.
Figure 5 Possible biological mechanism of the NSCLC TEP biomarkers. 
Notes: Intermodule-hub genes, including TPM2, MYL9, and PPP1R12C, stitched together the WASF1 module, which regulated actin cytoskeleton, the PRKAB2 module, which was involved in the AMPK–EGFR pathway, and the PDHB module, which was involved in carbohydrate metabolism. The PDHB module interacted with the RSRC1 module, which was associated with protein biosynthesis, growth, and migration.
Abbreviations: NSCLC, non-small-cell lung cancer; TEP, tumor-educated platelet.
Acknowledgments

This study was supported by Science Technology Department of Zhejiang Province (2017C37103). Meiling Sheng and Zhaohui Dong are co-first authors for this study.

Disclosure

The authors report no conflicts of interest in this work.

References


