Research Article

A Five-Gene Expression Signature Predicts Clinical Outcome of Ovarian Serous Cystadenocarcinoma

Li-Wei Liu, 1 Qiuhao Zhang, 1 Wenna Guo, 2 Kun Qian, 1 and Qiang Wang 1

1 State Key Laboratory of Pharmaceutical Biotechnology, School of Life Sciences, Nanjing University, Nanjing 210023, China
2 School of Life Sciences, Shanghai University, Shanghai 200444, China

Correspondence should be addressed to Qiang Wang; wangq@nju.edu.cn

Received 16 April 2016; Accepted 25 May 2016

Academic Editor: Jialiang Yang

Copyright © 2016 Li-Wei Liu et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Ovarian serous cystadenocarcinoma is a common malignant tumor of female genital organs. Treatment is generally less effective as patients are usually diagnosed in the late stage. Therefore, a well-designed prognostic marker provides valuable data for optimizing therapy. In this study, we analyzed 303 samples of ovarian serous cystadenocarcinoma and the corresponding RNA-seq data. We observed the correlation between gene expression and patients’ survival and eventually established a risk assessment model of five factors using Cox proportional hazards regression analysis. We found that the survival time in high-risk patients was significantly shorter than in low-risk patients in both training and testing sets after Kaplan-Meier analysis. The AUROC value was 0.67 when predicting the survival time in testing set, which indicates a relatively high specificity and sensitivity. The results suggest diagnostic and therapeutic applications of our five-gene model for ovarian serous cystadenocarcinoma.

1. Introduction

Ovarian serous cystadenocarcinoma is a common female genital cancer that causes more deaths than any other cancer of the female reproductive system. According to Global Cancer Statistics, approximately 230,000 women are diagnosed with ovarian cancer every year, and an estimated 150,000 women die of this disease annually [1]. Ovarian serous cystadenocarcinoma, a type of epithelial ovarian cancer, accounts for about 90% of all ovarian cancers [2]. Studies suggest that the risk factors for the disease include nulliparity, early menarche, late menopause, and family history [3]. Since the disease is often asymptomatic, the majority of patients are diagnosed at an advanced stage, with tumor invasion. Studies showed that the 5-year survival of stage I patients is greater than 90%, while that of patients in stages III to IV is less than 20% [4, 5]. The recent increase in the incidence of ovarian cancer has attracted the interest and attention of researchers worldwide.

With the development of sequencing technology, the research focus has been on the study of signature analysis for prognostic monitoring of ovarian cancer [6–12]. Microarray studies require precise design of probes despite the currently available and well-studied biomarkers for ovarian cancers. Other studies using miRNAs as biomarkers also suggest the limited value for clinical application, and miRNA therapy is still not clinically feasible. Compared with the foregoing methods, gene expression markers not only possess higher practical value, but also yield higher accuracy.

Here, we analyzed 303 clinical samples of ovarian serous cystadenocarcinoma and the corresponding RNA-seq data. We determined the relationship between gene expression data and survival time, in an effort to develop effective and accurate biomarkers for outcome prediction and personalized treatment.

2. Materials and Methods

2.1. Patient Samples and Gene Expression Data. We collected data from a total of 587 samples of serous cystadenocarcinoma (April 2016) from TCGA (http://cancergenome.nih.gov/) and finally used 303 samples (Table S1, at Supplementary Material available online at http://dx.doi.org/10.1155/2016/6945304) in this study after excluding 284 samples with unknown survival time or insufficient gene expression data. The 303 samples were assigned into 13 batches
and randomly allocated to training and testing sets. The prognostic marker model was established with a training set containing 8 batches (batches 9, 11–15, and 17–18) with 168 samples and validated using a testing set, comprising 5 batches (batches 19–22, 24, and 409) with 135 samples.

2.2. Statistical Analysis. Initially, we screened the samples by excluding those with unclear survival time or status. We retained only those genes expressed in more than half of the samples for further analysis. The expression level was then determined by logarithmic transformation and univariate Cox regression analysis. The significance of genes with $p$ value less than 0.001 was evaluated using random forests. We selected 100 genes of the largest importance to perform multivariate Cox’s analysis. Considering the practicality of clinical testing, we established 75,287,520 models with variables ranging from one to five genes using Cox proportional hazards regression analysis [35]. Further, all the 75,287,520 models were subjected to Receiver Operating Characteristic (ROC) analysis and the model with the largest area was selected.

Kaplan-Meier analysis was then conducted in both training and testing groups to validate the efficiency of the model. In order to test the independence and reproducibility of our model, we divided the samples into different datasets according to their ages and disease stages. We then performed Kaplan-Meier analyses and ROC analyses in each condition with IBM SPSS Statistics 22 (http://www.ibm.com/analytics/us/en/technology/spss/).

3. Results

3.1. Sample Characteristics. According to the screening criteria described, we randomly allocated the 303 samples with explicit survival time, survival state, and expression data into training and testing sets for modeling and validation, respectively. The median age of diagnosis in the selected patients was 58 years, the median survival time was 949 days, and the median survival of late-stage patients was 1069 days. A single patient was found in clinical stage I and 21, 241, and 38 patients were in stages II, III, and IV, respectively. The clinical stages of two patients were unknown (Table 1).

3.2. Obtain Genes Associated with Survival Time. Subsequently, we constructed 75,287,520 models comprising factors from 1 to 5 based on the 100 genes with the highest significance in the random forest method. The survival risk score of each patient was calculated according to the corresponding risk formula in each model, and the ROC curves were drawn. We extracted a batch of 5 genes (GPR128, AGXT, CYTH3, C10orf76, and TSPAN9) (Table 2) with the largest AUROC using the following formula: risk score = $(0.0796 \times \text{expression point of GPR128}) + (0.3451 \times \text{expression point of AGXT}) + (0.3402 \times \text{expression point of CYTH3}) + (0.6198 \times \text{expression point of C10orf76}) + (0.3402 \times \text{expression point of TSPAN9})$. All of these genes were reported previously (Table 3). The CYTH3 gene was expressed in the liver alone, playing a key role in regulating protein sorting and membrane trafficking [21]. Its use as a prognostic molecular marker in liver disease is also discussed. TSPAN9 is probably directly related to the proliferation of cancer cells. Other genes not directly correlated with the development of cancer may affect metabolism via signal transduction and indirectly affect the development of cancer.

3.3. Test the Predictive Ability of the Constructed Model Using Testing Set. After constructing the five-variable model with training set, we performed a Kaplan-Meier survival analysis of both training and testing sets to determine its prognostic value. In the training set, by calculating each patient’s risk score using the model, we divided the patients into two groups, designated as high-risk ($n = 84$) and low-risk groups ($n = 84$), based on their risk scores. The average survival time of patients in the low-risk group was 1,443 days, longer than in the high-risk group, which was 892 days. Kaplan-Meier analysis indicated a significant difference ($p < 0.001$) between the high-risk and low-risk groups in survival time [Figure 1(a)]. The prognosis of high-risk group appeared worse than that of the low-risk group, indicating that our model successfully distinguished the risk pattern. The higher risk tended to result in shorter survival time. Similar results of Kaplan-Meier analysis were found in the test group [Figure 1(b)], suggesting that our model was universally applicable in determining the risk level and predicting the survival of patients.

In order to further confirm the prognostic value of our model in predicting the survival time, we performed ROC analysis of the test group, setting 3 years as the cut-off, and calculated the risk score as the variable. The AUROC value of 0.670 (Figure 2) indicated a relatively high specificity and sensitivity.
Table 2: Five genes strongly correlated with patients’ survival time in training set.

<table>
<thead>
<tr>
<th>Gene name</th>
<th>𝑝 value</th>
<th>Hazard ratio</th>
<th>Coefficient</th>
<th>Variable importance</th>
<th>Relative importance</th>
</tr>
</thead>
<tbody>
<tr>
<td>GPR128</td>
<td>0.00092</td>
<td>1.0828</td>
<td>0.0796</td>
<td>0.0009</td>
<td>0.2478</td>
</tr>
<tr>
<td>AGXT189</td>
<td>0.00038</td>
<td>1.4121</td>
<td>0.345</td>
<td>0.0005</td>
<td>0.1442</td>
</tr>
<tr>
<td>CYTH39265</td>
<td>0.00048</td>
<td>1.4052</td>
<td>0.3402</td>
<td>0.0005</td>
<td>0.1432</td>
</tr>
<tr>
<td>C10orf76</td>
<td>0.00037</td>
<td>1.8585</td>
<td>0.6198</td>
<td>0.0009</td>
<td>0.2446</td>
</tr>
<tr>
<td>TSPAN9</td>
<td>0.0008</td>
<td>1.2884</td>
<td>0.2534</td>
<td>0.0002</td>
<td>0.0518</td>
</tr>
</tbody>
</table>

Table 3: Five-gene functions in previous research.

<table>
<thead>
<tr>
<th>Chromosomal</th>
<th>Start site</th>
<th>End site</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>GPR128</td>
<td>chr3</td>
<td>100328433 100414323</td>
<td>Playing important role in the transduction of intercellular signals across the plasma membrane; related to weight gain and intestinal contraction frequency in mouse [13–16].</td>
</tr>
<tr>
<td>AGXT</td>
<td>chr2</td>
<td>240868479 240880502</td>
<td>Expressing proteins involved in glyoxylate detoxification in the peroxisomes; its mutation causes primary hyperoxaluria type I, a severe inborn error of metabolism [17–20].</td>
</tr>
<tr>
<td>CYTH3</td>
<td>chr7</td>
<td>6161776 6272644</td>
<td>Mediating the regulation of protein sorting and membrane trafficking; related to HCC (hepatocellular carcinoma) tissues and could serve as prognostic factor [21–24].</td>
</tr>
<tr>
<td>C10orf76</td>
<td>chr10</td>
<td>101845599 102056193</td>
<td>Currently unknown; a recent study suggested the loss of C10orf76 resulted in the upregulation of several genes [25–29].</td>
</tr>
<tr>
<td>TSPAN9</td>
<td>chr12</td>
<td>3077355 3286564</td>
<td>Mediating signal transduction events that play a role in the regulation of cell development, activation, growth, and motility; associated with adhesion receptors of the integrin family and regulates integrin-dependent cell migration [30–34].</td>
</tr>
</tbody>
</table>

Figure 1: Kaplan-Meier curves with two-sided log rank test show correlation between five-gene model and survival time in both training set and testing set. (a) In training set, by calculating each patient’s risk score out of the model, we divided the patients into two groups, named as high-risk group (𝑛= 84) and low-risk group (𝑛= 84), based on their risk scores. Kaplan-Meier analysis was then performed and significant difference (𝑝< 0.001) was found between high-risk and low-risk group in the level of survival time. (b) Similar process and results are showed in testing set.
Table 4: Cox proportional hazard regression analyses in training and testing sets.

<table>
<thead>
<tr>
<th>Variables</th>
<th>Univariable model</th>
<th>Multivariable model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training group</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Five-gene model</td>
<td>2.672 (1.801–3.965)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Age</td>
<td>1.683 (1.153–5.457)</td>
<td>0.007</td>
</tr>
<tr>
<td>Testing group</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Five-gene model</td>
<td>2.248 (1.397–3.620)</td>
<td>0.001</td>
</tr>
<tr>
<td>Age</td>
<td>1.224 (0.772–1.941)</td>
<td>0.389</td>
</tr>
</tbody>
</table>

Low-risk patients survived longer than in the high-risk group ($p \leq 0.001$). Similar results were obtained with the groups of patients at different disease stages (stages I and II were merged because of limited specimen) except stage IV (Figure S1), which may be attributed to the relatively small sample size. However, the AUROC of this group was rather high. These analyses established that our model was independent of other risk factors and successfully distinguished low risk from high risk in each dataset.

4. Discussion

Ovarian serous cystadenocarcinoma is a common female genital cancer. Due to the absence of early-stage clinical symptoms and effective diagnosis, most patients were diagnosed with advanced disease. Further, due to the lack of effective treatment, the management of epithelial ovarian cancer is passive. Developing reliable prognostic molecular markers provides meaningful guidance for a reasonable and effective management program.

In this study, we analyzed 303 clinical samples of ovarian serous cystadenocarcinoma and the corresponding RNA-seq data, observed the correlation between gene expression and
survival time, and eventually established a risk assessment model based on five factors. Two of these genes (TSPAN9 [30–34], CYTH3 [21–24]) were directly correlated with cancer, with CYTH3 identified as a biomarker in liver cancer.

By calculating each patient's risk score, we found that each set showed significant differences in survival time between low-risk and high-risk groups, indicating that the model accurately predicted the mortality risk. The AUROC value in testing group is 0.670, representing a relatively high specificity and sensitivity.

In conclusion, our gene expression biomarkers can be used for accurate patient risk assessment, demonstrating practical value in predicting clinical outcomes. Our results are based on the samples derived from 303 individuals. Expanding sample size, especially including early-stage cancer patients, will further improve the prognostic value of the model.

Competing Interests
The authors declare that there is no conflict of interests regarding the publication of this paper.

Authors’ Contributions
Li-Wei Liu and Qiuhao Zhang contributed equally to this work.

Acknowledgments
This work was supported by the National Natural Science Foundation of China (31471200). The authors are grateful to the High Performance Computing Center (HPCC) of Nanjing University for doing the numerical calculations in this paper on its IBM Blade cluster system.

References
[22] A. W. Malaby, B. van den Berg, and D. G. Lambrecht, “Structural basis for membrane recruitment and allosteric activation of cytohesin family Arf GTPase exchange factors,” Proceedings of...


