A genomewide transcriptomic approach identifies a novel gene expression signature for the detection of lymph node metastasis in patients with early stage gastric cancer

Izumi, Daisuke; Gao, Feng; Toden, Shusuke; Sonohara, Fuminori; Kanda, Mitsuro; Ishimoto, Takatsugu; Kodera, Yasuhiro; Wang, Xin; Baba, Hideo; Goel, Ajay

Published in:
EBioMedicine

Published: 01/03/2019

Document Version:
Final Published version, also known as Publisher’s PDF, Publisher’s Final version or Version of Record

License:
CC BY-NC-ND

Publication record in CityU Scholars:
Go to record

Published version (DOI):
10.1016/j.ebiom.2019.01.057

Publication details:

Citing this paper
Please note that where the full-text provided on CityU Scholars is the Post-print version (also known as Accepted Author Manuscript, Peer-reviewed or Author Final version), it may differ from the Final Published version. When citing, ensure that you check and use the publisher's definitive version for pagination and other details.

General rights
Copyright for the publications made accessible via the CityU Scholars portal is retained by the author(s) and/or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights. Users may not further distribute the material or use it for any profit-making activity or commercial gain.

Publisher permission
Permission for previously published items are in accordance with publisher’s copyright policies sourced from the SHERPA RoMEO database. Links to full text versions (either Published or Post-print) are only available if corresponding publishers allow open access.

Take down policy
Contact lbscholars@cityu.edu.hk if you believe that this document breaches copyright and provide us with details. We will remove access to the work immediately and investigate your claim.
A genomewide transcriptomic approach identifies a novel gene expression signature for the detection of lymph node metastasis in patients with early stage gastric cancer

Daisuke Izumi,a,b,c,1, Feng Gao,d,1, Shusuke Toden,a Fuminori Sonohara,a,e, Mitsuro Kanda,e, Takatsugu Ishimoto,b,f, Yasuhiro Kodera,e, Xin Wang,g,d,**, Hideo Baba,b, Ajay Goel,a,**

1. Introduction

Lymph node (LN) metastasis is one of the major factors which influences poor prognosis in gastric cancer (GC) patients [1]. Therefore, accurate identification of LN status prior to treatment, particularly in early stages (mucosal and submucosal), is considered critical for improving treatment strategies and survival outcomes in these patients. Currently, diagnosis for LN metastasis is primarily made through various imaging modalities such as computed tomography (CT) and positron emission tomography with CT (PET-CT). However, these imaging-based
These data underscore the need to develop more sensitive and specific histopathological risk-assessment criteria currently used in the clinic. Overtreatment and unnecessary gastrectomies due to the inadequacy of current clinical practice, patients that are considered following endoscopic resections, we also included T2 lesions considered during endoscopy. Unfortunately however, pathological evaluation of the post-surgical gastrectomy tissues, and data for serum levels of carcinoembryonic antigen (CEA) and cancer antigen (CA) 19–9, were collected from each participating institution.

2.2. Ethics statement

Written informed consent was obtained from all patients, and the study was approved by the institutional review boards of all the participating institutions.

2.3. Study design and participants

Our study design included the following two major phases: a biomarker discovery and a clinical validation phase. Based on RNA-Seq data for T1 patient specimens in the TCGA dataset, we first prioritized 15 genes differentially expressed between 5 LNP and 13 LNN patients with GC. Using the same set of specimens for training, we built a multivariate logistic regression model using the 15 genes as covariates, and
subsequently derived a LN risk scoring formula. To demonstrate the robustness of this panel as a diagnostic marker, and its applicability to patients with T2 stage GC, we first evaluated its performance in the training set of T1 patients. This was followed by in silico validation in an expanded TCGA dataset involving 96 T1/T2 patients (LNP 49, LNN 47), and another independent set of 188 T2 patients (LNP 157, LNN 131) from the ACRG cohort.

In the clinical validation phase, two large, independent patient cohorts were analyzed to validate the 15-gene signature identified during the discovery phase. Using qRT-PCR data derived from 101 T1/T2 patient (LNP 24, LNN 77) specimens in the clinical cohort-1 as the testing set, we conducted a multivariate logistic regression analysis for qRT-PCR, from which a LN risk scoring formula was derived. The diagnostic performance of the 15-gene signature subsequently evaluated using an independent qRT-PCR dataset from 147 (LNP 26, LNN 121) T1 specimens from the clinical validation cohort-2. To demonstrate the clinical significance of our data, we benchmarked our gene signature against the conventional tumor markers, CEA and CA19-9. Computed tomography (CT) was performed before surgery in all patients belonging to the clinical cohort-2, and the imaging results were evaluated by board certified radiologists. When the size on the short axis of the regional LN was >10 mm, clinical LN status was deemed to be positive.

### 2.4. RNA isolation and quantitative reverse-transcription PCR

Total RNA extraction from tissue specimens was performed using miRNeasy RNA isolation kits (Qiagen, Hilden, Germany). Synthesis of complementary DNA (cDNA) was conducted on 1 μg of total RNA using the High Capacity cDNA Reverse Transcription Kit (Invitrogen, Carlsbad, CA, USA). Quantitative real-time reverse transcription analysis (qRT-PCR) was performed using the SensiFAST™ SYBR® Lo-ROX Kit (Bioline, London, UK) on the Quantstudio 7 Real Time PCR System (Life Technologies, Carlsbad, CA, USA). The average expression levels of target genes were normalized against beta-actin using the comparative CT method [14]. To ensure consistent measurements throughout all assays, for each PCR amplification reaction, three independent cDNA samples were loaded as internal controls to account for any plate-to-plate variation, and the results from each plate were normalized against internal normalization controls.

### 2.5. Statistical analysis

Wilcoxon’s signed-rank tests, Mann-Whitney U tests and Kruskal-Wallis tests were used to analyze gene expression data, as appropriate. The Benjamini-Hochberg method was used to correct for multiple hypotheses testing, wherever applicable. Risk scores derived from the 15-gene multivariate logistic regression model were used to plot receiver-operating-characteristic (ROC) curves and calculate area under the curves (AUCs). Confidence intervals for the ROC curves were calculated using the method of DeLong [15] as well as the statistical significance of comparison two ROC curves. Univariate and multivariate logistic regression models were employed to evaluate the statistical significance of clinicopathological variables and the 15-gene model in diagnosing LN metastasis status. All statistical analyses were performed using Medcalc V12.3.0 (Broekstraat 52, 9030; Mariakerke, Belgium), the GraphPad Prism V5.0 (GraphPad Software, San Diego, California, USA) and R (3.3.3, R Development Core Team, https://cran.r-project.org/).

## 3. Results

### 3.1. Genome-wide discovery of a novel gene expression signature to detect lymph node metastasis in early stage gastric cancer

To identify a panel of genes that can help diagnose patients with lymph node metastasis, we first analyzed RNA-seq expression profiling data from 18 patients with early stage T1 cancers, which were either LN metastasis positive or negative. Among a total of 20,531 genes, 84 genes were differentially expressed between 5 lymph node positive (LNP) and 13 negative (LNN) patients ($P < .01$ [Wilcoxon signed-rank test], log2 fold change $>1.5$; Fig. 1a and S2). To identify a robust candidate gene signature, we further narrowed down the gene list to 15 by filtering out lowly expressed genes (average expression level $<3$ log2-transformed TPM). Using multivariate logistic regression analysis, we found the 15 candidate genes were able to successfully distinguish LNP from LNN GC patients in the training set ($AUC = 1.000$, 95% CI 1.000–1.000; Fig. 1b).

In view of the availability of multiple public datasets consisting of T2 GC patients, we next investigated whether our T1 lymph node metastasis GC gene signature could also identify LN status in these additional patient cohorts. Intriguingly, our genes were able to distinguish LNP from LNN patients in an expanded set of 96 T1 and T2 patients in the TCGA cohort ($AUC = 0.839$, 95% CI 0.757–0.921; Fig. 1c), as well as in the ACRG cohort of 188 T2 patients ($AUC = 0.829$, 95% CI 0.752–0.906; Fig. 1d). These data highlight the diagnostic potential of our novel 15-gene in identifying LN metastasis in early-stage gastric cancer patients.

### 3.2. Validation and establishment of the 15-gene signature for detecting Lymph node status in gastric cancer patients

Next, we assessed the diagnostic accuracy of the 15 gene-panel by qRT-PCR in 24 LNP and 77 LNN tissue specimens in the clinical cohort-1 ($n = 101$). While individual genes had limited predictive power (AUCs varying between 0.506 and 0.605), the combination of...
various genes demonstrated significant detecting power in identifying LN metastasis in GC patients. Using a multivariate logistic regression analysis, we obtained a risk scoring formula for the 15-gene signature as follows, risk score = 1.300 - (0.233 × CSAR1) - (0.274 × CD83) + (0.048 × ET4V) - (0.213 × FAM13A) + (0.207 × FKBP10) - (0.227 × GPRIN3) + (0.052 × LCK) + (0.124 × NRRA2) + (0.402 × PRSS21) + (0.035 × RGPD1) - (0.196 × SLCA23) - (0.212 × SLF2) + (0.174 × TMEM86B) - (0.118 × TRPV4) - (0.075 × Y622). Although the performance of each individual genes was not significant, risk scores for LN metastasis determined using this formula for patients in this testing cohort demonstrated a very encouraging and robust diagnostic performance (AUC = 0.765, 95% confidence interval [CI], 0.667–0.862; OR = 23.61, 95% CI, 3.034–183.6; Fig. 2a and Table 2).

Of interest, our 15-gene signature was significantly superior compared to the conventional tumor markers, CEA (P = 0.033 [DeLong]) and CA19–9 (P = 0.044 [DeLong]; Fig. 2b) in identifying LNP patients. To further validate the diagnostic efficiency of this 15-gene signature, we next examined its performance in an independent validation cohort comprising of 26 LNP and 121 LNN T1 GC patients using the multivariate logistic regression analysis. In line with results from our testing cohort, the 15-gene signature was once again able to robustly distinguish LNP from LNN early stage GC patients (AUC = 0.742, 95% CI, 0.631–0.852; OR = 6.563, 95% CI, 2.585–16.66; Fig. 2b and Table 2).

3.3. The 15-gene signature outperformed conventional diagnostic approaches for LN metastasis detection in gastric cancer patients

Using multivariate analysis, we demonstrated that our 15-gene signature was able to successfully detect LN metastasis, independent of pre-operative clinical factors such as age, gender, tumor markers and clinical LN status determined by computed tomography (Table 3). To further evaluate the significance of this signature, we next compared the diagnostic potential of our gene signature versus various pre-operative clinical factors including tumor markers and clinical LN status. We found the diagnostic value of our 15-gene signature was significantly superior via a vis conventional tumor markers including the levels of circulating CEA (AUC = 0.520, 95% CI, 0.429–0.610; P = 0.044 [DeLong]) and CA19–9 (AUC = 0.518, 95% CI, 0.427–0.608; P = 0.0047 [DeLong]; Fig. 2b and Table 2).

In addition, to evaluate the performance of the 15-gene signature, we compared its performance against the clinical N stage determined by preoperative diagnostic CT scans in the patients from the validation cohort. Interestingly, our gene expression signature demonstrated a significantly superior accuracy (0.803, 95% CI, 0.510–0.905) compared to CT imaging data for identifying the presence of LN metastasis (AUC = 0.742; P = 0.038 [DeLong]; Fig. 2c and Table 2), which only achieved an AUC of 0.595 (95% CI, 0.511–0.675).

3.4. A combination of the 15-gene signature together with other clinicopathological features further improves the diagnostic accuracy for LN metastasis detection in gastric cancer patients

We next asked whether a combination of our 15-gene signature together with currently used clinicopathological factors (e.g. age, gender, tumor markers, and clinical N stage using multivariate logistic regression analysis) might further enhance the diagnostic accuracy of our panel. It was interesting to observe that indeed integration of our gene expression signature with the clinical N stage, significantly improved the discriminative accuracy of our biomarker panel further in identifying LNP gastric cancer cases (AUC = 0.789, 95% CI, 0.706–0.857) compared to use of the 15-gene signature alone (Fig. 2d and Table 2).

4. Discussion

As we usher into the new era of precision-medicine, tailoring individualized treatments are definitely going to serve as cornerstones for a more effective cancer care. Currently early stage GC patients, which are deemed to be high-risk for lymph node (LN) metastasis based upon various pre-surgical histopathological features are frequently over-treated, due to the lack of availability of adequate molecular markers that can more robustly identify such metastasis prior to the surgery. In this study, we undertook a systematic and comprehensive, genome-wide transcriptomic biomarker discovery, and developed a panel of genes for the identification of LN metastasis in patients with early stage gastric cancers (T1 and T2), using independent, publicly-available gene expression datasets. Subsequently, a 15-gene signature was optimized for qRT-PCR based analysis using the clinical testing cohort-1 by logistic regression analysis, followed by validation in an independent patient cohort. Finally, we performed a head-to-head comparison between the 15-gene signature with the conventional tumor markers (CEA and CA19–9) as well as CT-based imaging, and demonstrated its superiority for identifying GC patients with LN metastasis.

Recent advancements in high-throughput sequencing technologies have resulted in comprehensive molecular characterization of GC [16]. Similar to other major malignancies, multiple molecular subtypes of GC have been proposed based on integrative analysis of transcriptome-wide gene expression profiles [17]. Accordingly, several gene expression-based cancer biomarkers utilizing multiple genes have been suggested over the years [7, 9]. Because RNA-sequencing provides molecular insights into tumor heterogeneity and the disease process, in this study we focused on establishing a gene expression based-signature for the diagnosis of LN metastasis in early stage GC patients using a transcriptomic-wide analysis of T1 tumors from GC patients. We identified a cluster of 15 highly expressed genes, several of which are functionally relevant and GC-associated genes including CSAR1, CD83, NR4A2, ET4V, and TRPV4. In gastric cancer, CSAR1 has been shown to promote motility and invasiveness of cancer by activating RhoA, and its expression is reported to be associated with prognosis of GC patients [18]. CD83 is a molecular marker for mature dendritic cells. In gastric cancer, decreased density of CD83 (+) dendritic cells and increased density of FOXP3 (+) regulatory T cells, are observed in the primary tumor and metastatic lymph nodes of GC, and has been shown to inversely correlate with prognosis of GC patients [19]. An in vitro study has shown that the expression of orphan nuclear receptor NR4A2 in GC cells attenuates 5-fluorouracil-induced apoptosis and affects chemoresistance, and predicts an unfavorable postoperative survival of GC patients with chemotherapy [20]. Another putative oncogene, PEA3/ET4V has been shown to be upregulated at both mRNA and protein levels in GC tissue and the increased expression correlates with the expression of their downstream metastasis associated target gene, MMP-1 and high expression of PEA3/ET4V was associated with poor prognosis in GC [21]. Similarly, TRPV4 was shown to be a gene required for cancer cell invasion and trans-endothelial migration and its expression in GC correlated with poor clinical outcomes [22]. Furthermore, as a gene signature, three genes, NR4A2, FAM13A and PRSS21 had a significant contribution to our gene signature, suggesting that these genes play an important mechanistic role in GC LN metastasis.
In the past, a few studies have attempted to identify gene-expression-based biomarkers that may facilitate identification of LN metastasis in GC patients using cDNA microarrays [23–25]. However, to the best of our knowledge, ours is the first study to perform a systematic and comprehensive biomarker discovery from multiple RNA-Seq based datasets. Second, we performed validation of our discovered biomarkers in multiple, independent, datasets from publicly available resources, followed by confirmation of our results in in-house, clinical patient cohorts. Third, we focused our biomarker discovery and validation effort specifically in early-stage cancers, because excessive surgical treatment in these individuals have long-term consequences with adverse quality of life. Fourth, we compared the performance of our biomarkers with various tumor markers and CT imaging results, and successfully demonstrated the superiority of our signature over these with currently used modalities in the clinical settings.

One of the potential limitations of our study is that retrospective clinical cohorts were used for the development of the gene panel. In addition, one of the limitations of the present study is that we used frozen tissue and FFPE-derived RNA from resected tissues. Considering that this gene-signature will be examined in pre-surgical biopsy specimens in a clinical setting, further prospective trials are required to examine the robustness and performance of our 15-gene signature in fresh biopsy tissues. Furthermore, another limitation was that the sample size for biomarker discovery was limited. Since one of the primary objectives of our study was to identify biomarkers for early-stage gastric cancers, we focused on patients with T1 cases with LN metastasis, which further reduced the total number of patients during the discovery phase. Consequently, we were limited to deriving our gene-signature with the limited sample size, and could not fully utilize appropriate power calculations for biomarker discovery. Therefore, we would like to acknowledge this potential limitation of our study, that our effort would have been more comprehensive, if we had an access to larger cohorts of patients with T1 LN metastasis, which will likely require a multi-institutional effort given the rarity of this disease. Nevertheless, the reassuring aspect of our study is that regardless of this concern, 15-gene signature was successfully able to identify LN in GC patients, and was superior to both currently used tumor markers (CEA and CA-19-9) as well as CT imaging. Although the further clinical validation is required using a large prospective cohort, our gene signature was able to discriminate LNP patients from LNN patients using surgically resected sample.

In conclusion, we have developed a novel 15-gene signature, which can potentially be used for the identification of LN metastasis in patients with early stage GC. Pending further validation in prospective clinical cohorts, these markers offer promising potential for the identification of LN metastasis in gastric cancer patients.

**Funding source**

The present work was supported by the grants CA72851, CA181572, CA184792, CA202797 and CA187956 from the National Cancer Institute; a grant (RP140784) from the Cancer Prevention Research Institute of Texas (CPRIT), pilot grants from the Baylor Sammons Cancer Center, as
Table 2
Statistical evaluation of the performance of individual signature genes in the clinical cohorts

<table>
<thead>
<tr>
<th>Gene</th>
<th>Odds ratio (95% CI)</th>
<th>AUC (95% CI)</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Odds ratio (95% CI)</th>
<th>AUC (95% CI)</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>CSAR1</td>
<td>0.848 (0.661–1.088)</td>
<td>(0.424–0.626)</td>
<td>0.583</td>
<td>0.533</td>
<td>0.899 (0.784–1.031)</td>
<td>(0.457–0.624)</td>
<td>0.541</td>
<td>0.269</td>
</tr>
<tr>
<td>CD83</td>
<td>1.081 (0.814–1.435)</td>
<td>(0.420–0.623)</td>
<td>0.458</td>
<td>0.701</td>
<td>0.827 (0.583–1.173)</td>
<td>(0.460–0.627)</td>
<td>0.545</td>
<td>0.346</td>
</tr>
<tr>
<td>ETV4</td>
<td>0.994 (0.792–1.246)</td>
<td>(0.413–0.616)</td>
<td>0.417</td>
<td>0.727</td>
<td>1.057 (0.918–1.216)</td>
<td>(0.518–0.681)</td>
<td>0.602</td>
<td>0.615</td>
</tr>
<tr>
<td>FAM13A</td>
<td>1.167 (0.850–1.601)</td>
<td>(0.450–0.652)</td>
<td>0.958</td>
<td>0.208</td>
<td>1.079 (0.978–1.190)</td>
<td>(0.526–0.689)</td>
<td>0.610</td>
<td>0.731</td>
</tr>
<tr>
<td>FKBP10</td>
<td>0.992 (0.793–1.241)</td>
<td>(0.405–0.607)</td>
<td>0.833</td>
<td>0.312</td>
<td>1.070 (0.860–1.330)</td>
<td>(0.447–0.614)</td>
<td>0.769</td>
<td>0.405</td>
</tr>
<tr>
<td>GRIN3</td>
<td>0.942 (0.727–1.219)</td>
<td>(0.450–0.632)</td>
<td>0.875</td>
<td>0.247</td>
<td>1.096 (0.767–1.566)</td>
<td>(0.442–0.609)</td>
<td>0.526</td>
<td>0.423</td>
</tr>
<tr>
<td>LCK</td>
<td>0.911 (0.722–1.150)</td>
<td>(0.488–0.687)</td>
<td>0.458</td>
<td>0.779</td>
<td>1.007 (0.901–1.126)</td>
<td>(0.455–0.622)</td>
<td>0.596</td>
<td>0.308</td>
</tr>
<tr>
<td>NRA2</td>
<td>0.923 (0.683–1.248)</td>
<td>(0.444–0.645)</td>
<td>0.417</td>
<td>0.792</td>
<td>0.878 (0.782–0.989)</td>
<td>(0.512–0.676)</td>
<td>0.346</td>
<td>0.385</td>
</tr>
<tr>
<td>PRSS21</td>
<td>0.756 (0.627–0.985)</td>
<td>(0.503–0.701)</td>
<td>0.625</td>
<td>0.623</td>
<td>0.983 (0.910–1.063)</td>
<td>(0.426–0.593)</td>
<td>0.509</td>
<td>0.346</td>
</tr>
<tr>
<td>RGDPI1</td>
<td>0.946 (0.711–1.259)</td>
<td>(0.424–0.626)</td>
<td>0.208</td>
<td>0.974</td>
<td>0.899 (0.767–1.055)</td>
<td>(0.543–0.705)</td>
<td>0.626</td>
<td>0.577</td>
</tr>
<tr>
<td>SLC2A3</td>
<td>1.116 (0.932–1.337)</td>
<td>(0.495–0.693)</td>
<td>0.542</td>
<td>0.714</td>
<td>0.973 (0.792–1.195)</td>
<td>(0.469–0.635)</td>
<td>0.553</td>
<td>0.846</td>
</tr>
<tr>
<td>SLN2F</td>
<td>1.040 (0.775–1.395)</td>
<td>(0.404–0.607)</td>
<td>0.833</td>
<td>0.260</td>
<td>0.975 (0.886–1.073)</td>
<td>(0.420–0.587)</td>
<td>0.504</td>
<td>0.846</td>
</tr>
<tr>
<td>TMEM86B</td>
<td>0.839 (0.689–1.021)</td>
<td>(0.435–0.638)</td>
<td>0.333</td>
<td>0.896</td>
<td>0.915 (0.734–1.141)</td>
<td>(0.446–0.613)</td>
<td>0.530</td>
<td>0.962</td>
</tr>
<tr>
<td>TRPV4</td>
<td>1.041 (0.814–1.311)</td>
<td>(0.427–0.629)</td>
<td>0.375</td>
<td>0.831</td>
<td>1.062 (0.985–1.146)</td>
<td>(0.520–0.683)</td>
<td>0.603</td>
<td>0.846</td>
</tr>
<tr>
<td>YBX2</td>
<td>1.048 (0.911–1.207)</td>
<td>(0.444–0.646)</td>
<td>0.500</td>
<td>0.675</td>
<td>1.005 (0.929–1.087)</td>
<td>(0.461–0.627)</td>
<td>0.545</td>
<td>0.846</td>
</tr>
<tr>
<td>Risk score</td>
<td>14.77 (8.676–2514)</td>
<td>(0.670–0.844)</td>
<td>0.958</td>
<td>0.506</td>
<td>602.2 (29.48–12.298)</td>
<td>(0.663–0.810)</td>
<td>0.742</td>
<td>0.865</td>
</tr>
</tbody>
</table>

* AUC, area under the ROC curve.

Table 3

<table>
<thead>
<tr>
<th>MicroRNA</th>
<th>Univariate analysis</th>
<th>Multivariate analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>OR (95% CI)</td>
<td>P value</td>
</tr>
<tr>
<td></td>
<td>OR (95% CI)</td>
<td>P value</td>
</tr>
</tbody>
</table>

Abbreviation: OR, odds ratio.

Acknowledgements

We thank Preethi Ravidranathan for a critical reading of the manuscript.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ebiom.2019.01.057.

References