LASSO-based Cox-PH model identifies an 11-lncRNA signature for prognosis prediction in gastric cancer

YONGHONG ZHANG1*, HUAMIN LI2*, WENYONG ZHANG1, YA CHE3, WEIBING BAI4 and GUANGLIN HUANG5

1Department of General Surgery, Shangluo Central Hospital, Shangluo, Shaanxi 726000; 2Department of Pathology, Weinan Central Hospital, Weinan, Shaanxi 714000; 3Department of Medical Oncology, Shangluo Central Hospital, Shangluo, Shaanxi 726000; 4Department of General Surgery, Yulin Xingyuan Hospital, Yulin, Shaanxi 719000, P.R. China

Received December 23, 2017; Accepted September 13, 2018

DOI: 10.3892/mmr.2018.9567

Abstract. The present study aimed to identify a long non-coding (lnc) RNAs-based signature for prognosis assessment in gastric cancer (GC) patients. By integrating gene expression data of GC and normal samples from the National Center for Biotechnology Information Gene Expression Omnibus, the EBI ArrayExpress and The Cancer Genome Atlas (TCGA) repositories, the common RNAs in Genomic Spatial Event (GSE) 65801, GSE29998, E-MTAB-1338, and TCGA set were screened and used to construct a weighted correlation network analysis (WGCNA) network for mining GC-related modules. Consensus differentially expressed RNAs (DERs) between GC and normal samples in the four datasets were screened using the MetaDE method. From the overlapped lncRNAs shared by preserved WGCNA modules and the consensus DERs, an lncRNAs signature was obtained using L1-penalized (lasso) Cox-proportional hazard (PH) model. LncRNA-mRNA networks were constructed for these signature lncRNAs, followed by functional annotation. A total of 14,824 common mRNAs and 2,869 common lncRNAs were identified in the 4 sets and 5 GC-associated WGCNA modules were preserved across all sets. MetaDE method identified 1,121 consensus DERs. A total of 50 lncRNAs were shared by preserved WGCNA modules and the consensus DERs. Subsequently, an 11-lncRNA signature was identified by LASSO-based Cox-PH model. The lncRNAs signature-based risk score could divide patients into 2 risk groups with significantly different overall survival and recurrence-free survival times. The predictive capability of this signature was verified in an independent set. These signature lncRNAs were implicated in several biological processes and pathways associated with the immune response, the inflammatory response and cell cycle control. The present study identified an 11-lncRNA signature that could predict the survival rate for GC.

Introduction

Gastric cancer (GC) is the fifth leading cause of malignancy worldwide, with a 5-year survival rate of <10% (1,2). In China, it is the second most commonly diagnosed cancer in men and the third most commonly diagnosed cancer in women (3). The poor prognosis is primarily attributable to patients being frequently identified at an advanced stage and therefore difficult to cure (4). Early detection is key to improving survival rate of GC patients. Therefore, discovery of valuable molecular biomarkers is of significance for the facilitation of early diagnosis and effective prediction of prognosis and thereby contributing to improved outcomes in GC patients.

Long noncoding RNAs (lncRNAs) are defined as a group of non-protein-coding transcripts of greater than 200 nucleotides in length, which are characterized by tissue-specific expression patterns (5,6). With the number of lncRNAs being triple the number of protein-coding genes, lncRNAs are predicted to exhibit a more important role in basic, translational and clinical oncology than protein-coding genes (7). Several lncRNAs have been demonstrated in GC, including H19 (8-10), HOTAIR (11,12) and ANRIL (13). However, the association of lncRNAs with GC prognosis has not been fully elucidated. Although a recent study by Miao et al (14) reported a 4-lncRNA signature of prognostic value for GC patients, the association of lncRNAs with GC prognosis has not been fully elucidated. A comprehensive analysis of gene expression data of GC patients from more databases is required for acquiring a more convincing prognostic lncRNAs signature.

In contrast with the study of Miao et al (14), the present study performed an integrated analysis on GC gene expression data mined in the National Center for Biotechnology Information (NCBI), Gene Expression Omnibus (GEO), EBI...
ArrayExpress and TCGA repositories. The present study was mainly focused on revealing the critical lncRNAs involved in GC pathogenesis and the roles of the critical lncRNAs in the molecular mechanisms of GC. An 11-lncRNA signature was identified for prognostic risk assessment of GC patients using weighted correlation network analysis (WGCNA) network, the MetaDE method and a LASSO-based Cox-proportional hazard (PH) model. In addition, the prognostic significance of this signature was validated in an independent set. In order to reveal the molecular mechanisms of these critical lncRNAs, the lncRNA-mRNA interaction network was constructed for functional and pathway enrichment analysis. The results revealed that these critical lncRNAs can regulate the associated mRNAs to influence the immune response, inflammatory response and cell cycle in the pathogenesis of GC.

Materials and methods

Data resource and preprocessing. Gene expression profiles for GC were searched in publicly accessible GEO at the NCBI (http://www.ncbi.nlm.nih.gov/geo/) and EBI ArrayExpress (https://www.ebi.ac.uk/arrayexpress/). Inclusion criteria were: Human gene expression data; gastric cancer specimens and paired normal specimens; total count of specimens ≥50. Finally, Genomic Spatial Event (GSE) (15) 6580 and GSE29998 downloaded from NCBI GEO and E-MTAB-1338 from EBI ArrayExpress were selected in the present study (Table I).

Raw data (TXT) in GSE6580, GSE29998 and E-MTAB-1338 were subject to log2 transformation by limma (version 3.4.0) software (16) (https://bioconductor.org/packages/release/bioc/html/limma.html). Subsequently, the data were transformed from a skewed distribution to normal distribution, followed by median normalization. Based on the platform annotation files (Table I), probe sets that were assigned with a RefSeq transcript ID and/or Ensembl gene ID were obtained, of which the probe sets labeled as ‘NR’ (non-coding RNA in the Refseq database) were selected. In addition, platform sequencing data was aligned with human genome (GRCh38) (17,18) using Clustal 2 (http://www.clustal.org/clustal2/) (19). The resulting lncRNAs and the above-mentioned lncRNAs annotated in Refseq database were combined and used in further analysis.

The present study also acquired mRNA-seq data of 384 GC samples and 26 normal controls from TCGA portal (https://gdc-portal.nci.nih.gov/), which did not require preprocessing. Common RNAs of the GSE6580, GSE29998, E-MTAB-1338 and TCGA were used for further analysis.

WGCNA network analysis. WGCNA (20) is a bioinformatics tool used to build a gene co-expression networks to mine network modules closely associated with diseases. Based on the common RNAs identified, WGCNA package (21) (version 1.61) in R 3.4.1 language was applied to identify GC-associated RNA modules (https://cran.r-project.org/web/packages/WGCNA/index.html) in the present study. The TCGA set was used as the training set, while GSE6580, GSE29998 and E-MTAB-1338 were selected as testing sets. Comparability of these 4 sets were assessed by correlation analysis of RNA expression levels. A weighted gene co-expression network was built as previously described (20).

Briefly, the soft threshold power of β was determined using scale-free topology criterion. Following the removal of RNAs with coefficients of variation <0.1, the weighted adjacency matrix was then developed. A dynamic tree cut algorithm was used to mine modules with a module size ≥30 and a minimum cut height of 0.95. In addition, preservation of modules in all 4 datasets was examined using the module preservation function of the WGCNA package. In addition, functional annotation of the modules identified was investigated using the userListEnchment function of WGCNA package.

Identification of consensus differentially expressed RNAs. Consensus differentially expressed RNAs (DERs) between GC specimens and normal control specimens across the 4 datasets (GSE6580, GSE29998, E-MTAB-1338 and TCGA) were identified with metaDE package (22,23) (https://cran.r-project.org/web/packages/MetaDE/) in R language version 3.4.1. The cutoff was set at tau2=0, Qpval>0.05, P<0.05 and false discovery rate (FDR)<0.05. tau2 denotes the amount of heterogeneity while Qpval denotes heterogeneity of a dataset. The common lncRNAs shared by the list of consensus DERs and the RNAs in the preserved WGNCA modules were selected for further analysis.

Development of a prognostic risk scoring system for GC. L1-penalized (lasso) characterized by simultaneous variable selection and shrinkage is a useful method for determining interpretable prediction rules in high-dimensional data (24). In order to determine an lncRNA signature for prognosis, the penalized package (24) in R language (version 3.4.1) was applied to fit a lasso Cox-PH (25) to the overlapped lncRNAs. Based on the optimal lambda value that was selected through a 1,000 cross-validations, a panel of prognostic lncRNAs was determined. An equation for calculating risk score was generated based on the expression levels of these prognostic lncRNAs and their regression coefficients from the Cox-PH model as follows:

\[
\text{Risk score} = \beta_1 \text{lncRNA}_1 \times \text{exprlncRNA}_1 + \beta_2 \text{lncRNA}_2 \times \text{exprlncRNA}_2 + \cdots + \beta_n \text{lncRNA}_n \times \text{exprlncRNA}_n
\]

Risk score was calculated and assigned to each patient in the training set (TCGA set, Table II). With the median risk score as cutoff, all patients in the training set were split into a high-risk group and a low-risk group. Overall survival (OS) time and recurrence-free survival (RFS) time of the two risk groups were analyzed and compared by Kaplan-Meier survival analysis and the logrank test.

The robustness of the risk scoring system was validated in an independent dataset (GSE62254) (26) downloaded from NCBI GEO (platform: GPL570, Affymetrix Human Genome U133 Plus 2.0 Array). GSE62254 included the gene expression data of 300 GC tissue samples (Table II). Raw data was preprocessed using an oligo (27) package in R language (version 3.4.1). Risk score and risk groups were determined similarly for the GSE62254 dataset. Discrepancies in OS time and RFS time between the risk groups were analyzed using Kaplan-Meier survival analysis and the log rank test.

Functional analysis of prognostic lncRNAs. To investigate the biological function of these prognostic lncRNAs identified...
above in GC tumorigenesis, lncRNA-mRNA networks were constructed for them based on the correlation coefficients between RNAs from WGCNA modules. Gene ontology (GO; http://www.geneontology.org/) function and Kyoto Encyclopedia of Genes and Genomes (KEGG; https://www.kegg.jp/) pathway enrichment analysis was performed for the RNAs in these lncRNA-mRNA networks by DAVID Bioinformatics Tool (28,29) (version 6.8; https://david-d.ncifcrf.gov/).

Results

RNA expression data. Following data preprocessing, the present study identified 17,693 common RNAs in the GSE6580, GSE29998, E-MTAB-1338 and TCGA sets, including 14,824 mRNAs and 2,869 lncRNAs (Table III).

WGCNA network and modules. Based on these common RNAs, WGCNA was used to mine GC-associated modules, with TCGA set as the training set and GSE6580, GSE29998, E-MTAB-1338 as validation sets. The correlation of gene expression between these sets was in the range of 0.4-1 with P<1x10\(^{-200}\) (Fig. 1), indicating good comparability between the sets. For adjacencies calculation, the soft threshold power of \(\beta\) was determined to be 5 when the scale-free topology fit (scale-free R\(^2\)) achieved 0.9 (Fig. 2).

A total of 11 modules (black, blue, brown, turquoise, yellow and purple) were mined with WGCNA for the TCGA dataset. In the resulting dendrogram (Fig. 3A), these modules were represented by branches in different colors. Module mining was also conducted in GSE29998, GSE6580 and E-MTAB-1338. The gene dendrograms are presented in Fig. 3B-D.

As illustrated in a gene multi-dimensional scaling (MDS) plot (Fig. 4A), RNAs of the same module were prone to cluster together, suggesting similar expression patterns of RNAs in the same module. A hierarchical clustering analysis of the 11 modules identified that the associated modules clustered together, such as the black module and the yellow module, the pink module and the purple module, the magenta module and the red module, and the grey module and the turquoise module (Fig. 4B). Not unexpectedly, these modules were also close to each other in the module MDS plot (Fig. 4C).

In addition, out of the 11 modules, black, blue, brown, turquoise and yellow modules with Z-score >5 were identified to be well preserved across the GSE6580, GSE29998, E-MTAB-1338 and TCGA sets (Table IV). Functional
annotation of the 5 modules was performed using WGCNA package (Table IV). The black module was associated with digestion. The blue module was associated with immune response. The brown module was correlated with cell cycle. The turquoise module was associated with cell adhesion. The yellow module was linked to protein amino acid glycosylation (Table IV).

Consensus DERs. The metaDE package identified 1,121 consensus DERs in the GSE6580, GSE29998, E-MTAB-1338 and TCGA sets, of which 255 were lncRNAs. A heatmap of these consensus DERs was generated by heatmap.sig.genes function in MetaDE package (Fig. 5). Clearly, expression patterns of these consensus DERs were similar in 4 datasets. Furthermore, 288 RNAs were overlapped between the 5
Table IV. Characteristics of WGCNA network modules.

| TCGA   | GSE29998 | GSE65801 | E-MTAB-133 | Color | Module size | Module preservation (Z-score) | Module characterization |
|--------|----------|----------|-------------|-------|-------------|-------------------------------|------------------------|
| D1M1   | D2M1     | D3M1     | D4M1        | Black | 59          | 28.06                         | Digestion              |
| D1M2   | D2M2     | D3M2     | D4M2        | Blue  | 417         | 31.59                         | Immune response        |
| D1M3   | D2M3     | D3M3     | D4M3        | Brown | 411         | 26.25                         | Cell cycle             |
| D1M4   | D2M4     | D3M4     | D4M4        | Green | 111         | 6.41                          | -                      |
| D1M5   | D2M5     | D3M5     | D4M5        | Grey  | 1,097       | 4.90                          | -                      |
| D1M6   | D2M6     | D3M6     | D4M6        | Nagenta| 38          | 10.21                         | -                      |
| D1M7   | D2M7     | D3M7     | D4M7        | Pink  | 56          | 22.08                         | -                      |
| D1M8   | D2M8     | D3M8     | D4M8        | Red   | 78          | 17.64                         | -                      |
| D1M9   | D2M9     | D3M9     | D4M9        | Turquoise| 564      | 29.46                         | Cell adhesion          |
| D1M10  | D2M10    | D3M10    | D4M10       | Yellow| 215         | 14.37                         | Protein amino acid glycosylation |
| D1M11  | D2M11    | D3M11    | D4M11       | Purple | 35         | 8.30                          | -                      |

WGCNA, weighted correlation network analysis; TCGA, The Cancer Genome Atlas; GSE, Genomic Spatial Event.
Figure 5. A heatmap of consensus RNAs identified by MetaDE. RNAs expression patterns are similar in the TCGA, GSE29998, GSE65801 and E-MTAB-1338 sets. TCGA, The Cancer Genome Atlas; GSE, Genomic Spatial Event.

Figure 6. Analysis of overlapped RNAs. (A) Venn diagram displaying the overlapped RNAs between the preserved WGCNA modules and the consensus DERs identified by MetaDE. (B) Distribution of overlapped mRNAs (upper) and IncRNAs (lower) in the 5 preserved WGCNA modules (black, blue, brown, turquoise and yellow). Inc, long non-coding; WGCNA, weighted correlation network analysis; DERs, differentially expressed RNAs.
Among these overlapped RNAs, 50 were lncRNAs, of which 32 were included in the blue module, 14 in the brown module, 3 in the turquoise module and 1 in the yellow module (Fig. 6B).

**Table V. The 11 prognostic lncRNAs identified by LASSO-based Cox-proportion hazard model.**

| lncRNA         | Coefficient | HR      | 95% CI          |
|----------------|-------------|---------|-----------------|
| ARHGAP5-AS1    | 0.0124      | 1.1907  | 0.8259-1.7166   |
| FLVCR1-AS1     | -0.1191     | 0.6610  | 0.4916-0.8886   |
| H19            | 0.9171      | 1.0497  | 0.9390-1.1735   |
| HOTAIR         | -0.4973     | 0.8970  | 0.6584-1.2222   |
| LINC00221      | 1.1799      | 1.9190  | 1.2021-3.0633   |
| MCF2L-AS1      | -0.7009     | 0.7785  | 0.6053-1.0014   |
| MUC2           | -0.0902     | 0.9516  | 0.8631-1.0492   |
| PRSS30P        | 0.2572      | 1.1254  | 0.8263-1.5329   |
| SCARNA9        | -0.8615     | 0.7383  | 0.5449-1.0004   |
| TP53TG1        | 0.1493      | 1.1386  | 0.8808-1.4720   |
| XIST           | -0.9235     | 0.5469  | 0.1926-1.5527   |

lnc, long non-coding; HR, hazard ratio; CI, confidence interval.

Figure 7. Kaplan-Meier curves for OS time (left) and RFS time (right) of patients in (A) TCGA and (B) GSE62254 sets. Patients of each set are divided by risk score into a high-risk group and a low-risk group. OS and RFS between two risk groups were analyzed and compared by Kaplan-Meier analysis and logRank test. TCGA, The Cancer Genome Atlas; GSE, Genomic Spatial Event; OS, overall survival; RFS, recurrence-free survival.

Figure 7. Kaplan-Meier curves for OS time (left) and RFS time (right) of patients in (A) TCGA and (B) GSE62254 sets. Patients of each set are divided by risk score into a high-risk group and a low-risk group. OS and RFS between two risk groups were analyzed and compared by Kaplan-Meier analysis and logRank test. TCGA, The Cancer Genome Atlas; GSE, Genomic Spatial Event; OS, overall survival; RFS, recurrence-free survival.

Preserved modules and the list of consensus DERs (Fig. 6A). Among these overlapped RNAs, 50 were lncRNAs, of which 32 were included in the blue module, 14 in the brown module, 3 in the turquoise module and 1 in the yellow module (Fig. 6B).

**Development and validation of an lncRNAs-based risk scoring system.** Based on the expression of these overlapped lncRNAs in the TCGA set, the LASSO-based Cox-PH model identified an 11-lncRNA signature that was significantly asso-
ZHANG et al: A PROGNOSTIC lncRNA SIGNATURE FOR GASTRIC CANCER

5586

Risk score=0.012437 x Exp_{ARHGAP5-AS1} + (-0.11914) x Exp_{FLVCR1-AS1} + 0.917082 x Exp_{H19} + (-0.49726) x Exp_{HOTAIR} + 1.179896 x Exp_{LINC00221} + (-0.70093) x Exp_{MCF2L-AS1} + (-0.09017) x Exp_{MUC2} + 0.257189 x Exp_{PRSS30P} + (-0.86146) x Exp_{SCARNA9} + 0.149341 x Exp_{TP53TG1} + (-0.92352) x Exp_{XIST}

Risk score was calculated for each patient. All patients in the TCGA set were split into a high-risk group and a low-risk group with the median risk score as the cutoff. Patients in the high-risk group (n=156) demonstrated significantly shorter OS time (15.56±13.15 months vs. 21.23±19.99, logRank P=7.44x10^{-5}) and RFS time (15.76±11.51 months vs. 21.72±21.03, logRank P=0.0117) compared with the patients in the low-risk group (n=155, Fig. 7A). Prognostic performance of this 11-lncRNA signature-based risk scoring system was tested in an independent set (GSE62254). All 300 patients in GSE62254 were divided into a high-risk group (n=150) and a low-risk group (n=150) by risk score. Similarly, OS time (54.79±31.83 months vs. 46.40±31.83, logRank P=0.0311) and RFS time (37.45±31.08 months vs. 29.99±28.11, logRank P=0.0282) were markedly elongated in the low-risk group relative to the high-risk group (Fig. 7B).

Function analysis of the 11-lncRNA signature. Among the 11 signature lncRNAs, 9 lncRNAs (FLVCR1-AS1, H19, LINC00221, MUC2, RSS30P, SCARNA9, TP53TG1, XIST, ARHGAP5-AS1, HOTAIR and MCF2L-AS1) were involved in the blue module, whereas another 2 lncRNAs (HOTAIR and MCF2L-AS1) were present in the brown module. Correlations between the 9 lncRNAs in the blue module and mRNAs revealed by the lncRNA signature were calculated using the following formula:

Risk score=0.012437 x Exp_{ARHGAP5-AS1} + (-0.11914) x Exp_{FLVCR1-AS1} + 0.917082 x Exp_{H19} + (-0.49726) x Exp_{HOTAIR} + 1.179896 x Exp_{LINC00221} + (-0.70093) x Exp_{MCF2L-AS1} + (-0.09017) x Exp_{MUC2} + 0.257189 x Exp_{PRSS30P} + (-0.86146) x Exp_{SCARNA9} + 0.149341 x Exp_{TP53TG1} + (-0.92352) x Exp_{XIST}

Function analysis of the 11-lncRNA signature. Among the 11 signature lncRNAs, 9 lncRNAs (FLVCR1-AS1, H19, LINC00221, MUC2, RSS30P, SCARNA9, TP53TG1, XIST and ARHGAP5-AS1) were involved in the blue module, whereas another 2 lncRNAs (HOTAIR and MCF2L-AS1) were present in the brown module. Correlations between the 9 lncRNAs in the blue module and mRNAs revealed by the
| GO category          | Term                          | Count | Genes                                                                 |
|----------------------|-------------------------------|-------|-----------------------------------------------------------------------|
| Biology process      | Immune response               |       | FDR                                                                  |
|                      | Regulation of cell activation | 30    | MICB, CD8A, LY86, HLA-DMA, CD94, CD47, CD80, FAS, FASLG, ITGB2, ITGAM,  |
|                      | Regulation of lymphocyte activation | 31    | ITGAL, MICB, CD8A, IL21R, KLRK1, PTPN22, IL7R, IL12R7, PTPN22, ZAP70, |
|                      | Positive regulation of immune system process | 33    | MICB, CD8A, IL21R, KLRK1, PTPN22, IL7R, IL12R7, PTPN22, ZAP70, |
|                      | Leukocyte activation          | 34    | ITGAL, MICB, CD8A, IL21R, KLRK1, PTPN22, IL7R, IL12R7, PTPN22, ZAP70, |
|                      | Regulation of T cell activation | 25    | PTPRC, SIT1, IL2RA, IKZF1, TNFRSF13B, CD40, PDCD1, CD28, IL2RA, |
|                      | Regulate of leukocyte activation | 28    | PTPRC, SIT1, IL2RA, IKZF1, TNFRSF13B, CD40, PDCD1, CD28, IL2RA, |
|                      | Cell activation               | 23    | PTPRC, SIT1, IL2RA, IKZF1, TNFRSF13B, CD40, PDCD1, CD28, IL2RA, |
|                      | Positive regulation of cell activation | 23    | PTPRC, SIT1, IL2RA, IKZF1, TNFRSF13B, CD40, PDCD1, CD28, IL2RA, |
| GO category                        | Term Count | Genes                                                                 | FDR   |
|-----------------------------------|------------|----------------------------------------------------------------------|-------|
| Hemopoietic or lymphoid organ development | 24         | PTPRC, CD3D, PLEK, IKZF1, CD8A, HCLS1, PTPN22, ITGA4, DOCK2, IRF8, ZAP70, CD4, CD2, CD9, CD79A, IRF4, SPN, RHOH | 2.60x10^-08 |
| Inflammatory response             | 26         | ITGAL, C3AR1, AIF1, LY86, CCR1, CXCL9, ITGB2, CCL5, C1QC, CCL4, AOAH, CIITA, IL7R, HLA-DMA, DOCK2, CXCL13, CXCR5, CXCL12, PARVG, SIGLEC10, ITGA4, SLAMF7, EMILIN2, CD84, SIGLEC1, CORO1A, CD3D, GPR171, CD40, GZMB, HLA-DRB5, HLA-DPA1, HLA-DPB1, HLA-DOA, HLA-DOB1, HLA-DOC1, PDCD1, PDCD1LG2, CD8, CD274, CD90, CD22, CD4, CD6, CD180, CD163, APOL3, PRKCQ, C1QB, SIGLEC1, CYBB, CCR5, CCR4, CXCL13, PLA2G7 | 8.34x10^-04 |
| KEGG Pathway                      | 26         | SELPG, PARVG, PTPRC, PLEK, SIGLEC10, ITGA4, SLAMF7, EMILIN2, CD84, SIGLEC1, CORO1A, CD3D, GPR171, CD40, GZMB, HLA-DRB5, HLA-DPA1, HLA-DPB1, HLA-DOA, HLA-DOB1, HLA-DOC1, PDCD1, PDCD1LG2, CD8, CD274, CD90, CD22, CD4, CD6, CD180, CD163, APOL3, PRKCQ, C1QB, SIGLEC1, CYBB, CCR5, CCR4, CXCL13, PLA2G7 | 8.58x10^-04 |
| Allograft rejection                | 12         | SELPG, PARVG, PTPRC, PLEK, SIGLEC10, ITGA4, SLAMF7, EMILIN2, CD84, SIGLEC1, CORO1A, CD3D, GPR171, CD40, GZMB, HLA-DRB5, HLA-DPA1, HLA-DPB1, HLA-DOA, HLA-DOB1, HLA-DOC1, PDCD1, PDCD1LG2, CD8, CD274, CD90, CD22, CD4, CD6, CD180, CD163, APOL3, PRKCQ, C1QB, SIGLEC1, CYBB, CCR5, CCR4, CXCL13, PLA2G7 | 6.74x10^-15 |
WGCNA were used to construct an IncRNA-mRNA network (Fig. 8A). Similarly, another IncRNA-mRNA network was built for the 2 IncRNAs (HOTAIR and MCF2L-AS1), in the brown module (Fig. 8B). The genes in the IncRNA-mRNA network that correlated with the 9 prognostic IncRNAs in the blue module were significantly associated with 23 GO biological process terms (including immune response, regulation of cell activation and regulation of lymphocyte activation) and 8 KEGG pathways (including cell adhesion molecules, allograft rejection and cytokine-cytokine receptor interaction; Table VI). The genes in the IncRNA-mRNA network that correlated with HOTAIR and MCF2L-AS1 were mainly associated with the cell cycle phase, cell cycle and mitotic cell cycle. In addition, 4 KEGG pathways were enriched for the genes in this IncRNA-mRNA network including cell cycle, DNA replication, progesterone-mediated oocyte maturation and steroid biosynthesis pathways (Table VII).

Discussion

A growing number of studies have demonstrated that aberrantly expressed IncRNAs are implicated in GC tumorigenesis and progression (30,31). Nonetheless, the prognostic significance of IncRNAs in GC remains to be elucidated. Based on the common RNAs data and corresponding clinical information of GC patients and normal controls which were obtained through data mining in NCBI GEO, EBI ArrayExpress and TCGA, a 11-lncRNA prognostic signature was identified by a series of bioinformatics analyses featuring WGCNA, the MetaDE method and a LASSO-based Cox-PH model. Furthermore, it was identified that patients could be classified into a high-risk group and a low-risk group by the risk score based on the 11-lncRNA signature in the training set, with noticeable separations being observed in the Kaplan-Meier curves between the 2 groups. The high-risk group exhibited significantly longer OS and PFS time compared with the low-risk group. The predictive ability of risk score was confirmed in an independent set. Therefore, the present study demonstrated that the 11-lncRNA signature has the potential for assessing survival rate of GC patients.

The 11-lncRNA signature determined in the study was comprised of FLVCR1-AS1, H19, LINC00221, MUC2, PRSS30P, SCARNA9, TP53TG1, XIST, ARHGAP5-AS1, HOTAIR and MCF2L-AS1. Among these lncRNAs, H19 is identified to be upregulated in plasma of GC patients and is proposed as a diagnostic biomarker (8). Increasing evidence also demonstrates that H19 upregulation promotes GC proliferation, migration and invasion (9,10). It has been established that MUC2 is associated with outcome of GC patients (32). lncRNA X inactive specific transcript (XIST) encoded by XIST gene acts as a regulator of X inactivation in mammals (33). Chen et al (34) observed upregulated XIST in GC tissue and identified that this lncRNA serves a regulatory role in GC progression via microRNA (miR)-101 and its direct target polycomb group protein enhancer of zeste homolog 2. HOTAIR transcribed from the HOXC locus is identified to be overexpressed in GC, which is a characteristic molecular alteration of GC (35). Furthermore, there is evidence that HOTAIR functions as a GC oncogene through regulating the expression of human epithelial growth factor receptor 2 by competing with miR-331-3p (12).
Table VII. Significant GO terms and KEGG pathways for the genes in the constructed lncRNA-mRNA network of two prognostic lncRNAs in the brown module.

| GO category               | Term                                  | Count | Genes                                                                 | FDR       |
|---------------------------|---------------------------------------|-------|----------------------------------------------------------------------|-----------|
| Biology                   | Cell cycle phase                      | 40    | E2F1, KIF23, PRC1, NEK3, NEK2, DBF4, TTK, PKMYT1, ANLN, AURKA, PTTG1, CEP55, AURKB, CCNE1, CDC2, CDC2A, CDC5, CDC6, CDC3, MKI67, MSH5, TPX2, SKP2, NUFB, CENPF, CDC20, BIRC5, CENPE, NDC80, ESPL1, PKB, CENK3, UBE2C, TACC3, CDC25B, CCNB1, MAD2L1, PLK1, POLDI, DSCC1 | 2.14x10^{−22} |
|                           | Cell cycle                            | 50    | E2F1, KIF23, CEPT7, PRC1, DBF4, E2F1, TTK, PKMYT1, AURKA, PTTG1, AURKB, CDC2, CDC6, CDC3, MKI67, MSH5, TPX2, SKP2, NUFB, CENPF, CDC20, BIRC5, CENPE, NDC80, ESPL1, PKB, CENK3, UBE2C, TACC3, CDC25B, CCNB1, MAD2L1, PLK1, POLDI, DSCC1 | 2.83x10^{−21} |
|                           | Mitotic cell cycle                    | 37    | KIF23, E2F1, PRC1, NEK3, NEK2, DBF4, TTK, PKMYT1, ANLN, AURKA, PTTG1, CEP55, AURKB, CCNE1, CENPA, CDC2, CDC2A, CDC6, CDC3, MKI67, MSH5, TPX2, SKP2, NUFB, CENPF, CDC20, BIRC5, CENPE, NDC80, ESPL1, PKB, CENK3, UBE2C, TACC3, CDC25B, CCNB1, MAD2L1, PLK1, POLDI, DSCC1 | 6.89x10^{−21} |
|                           | Cell cycle process                    | 42    | E2F1, KIF23, CEPT7, PRC1, NEK3, NEK2, DBF4, TTK, PKMYT1, ANLN, AURKA, PTTG1, AURKB, CDC2, CDC2A, CDC6, CDC3, MKI67, MSH5, TPX2, SKP2, NUFB, CENPF, CDC20, BIRC5, CENPE, NDC80, ESPL1, PKB, CENK3, UBE2C, CDC25B, CCNB1, MAD2L1, PLK1, POLDI, DSCC1 | 2.26x10^{−19} |
|                           | Mitosis                               | 28    | KIF23, NEK3, NEK2, PKMYT1, AURKA, ANLN, CEP55, AURKB, PTTG1, CDC2A, CDC6, CDC3, MKI67, MSH5, TPX2, SKP2, NUFB, CENPF, CDC20, BIRC5, CENPE, NDC80, ESPL1, PKB, UBE2C, TACC3, CDC25B, CCNB1, MAD2L1, PLK1, DSCC1 | 2.68x10^{−19} |
|                           | Nuclear division                      | 28    | KIF23, NEK3, NEK2, PKMYT1, AURKA, ANLN, CEP55, AURKB, PTTG1, CDC2A, CDC6, CDC3, MKI67, MSH5, TPX2, SKP2, NUFB, CENPF, CDC20, BIRC5, CENPE, NDC80, ESPL1, PKB, UBE2C, CDC25B, CCNB1, MAD2L1, PLK1, DSCC1 | 1.39x10^{−17} |
|                           | M phase of mitotic cell cycle         | 28    | KIF23, NEK3, NEK2, PKMYT1, AURKA, ANLN, CEP55, AURKB, PTTG1, CDC2A, CDC6, CDC3, MKI67, MSH5, TPX2, SKP2, NUFB, CENPF, CDC20, BIRC5, CENPE, NDC80, ESPL1, PKB, UBE2C, CDC25B, CCNB1, MAD2L1, PLK1, DSCC1 | 2.25x10^{−17} |
|                           | Organelle fission                     | 28    | KIF23, NEK3, NEK2, PKMYT1, AURKA, ANLN, CEP55, AURKB, PTTG1, CDC2A, CDC6, CDC3, MKI67, MSH5, TPX2, SKP2, NUFB, CENPF, CDC20, BIRC5, CENPE, NDC80, ESPL1, PKB, UBE2C, CDC25B, CCNB1, MAD2L1, PLK1, DSCC1 | 4.04x10^{−17} |
|                           | Cell division                         | 26    | KIF23, PRC1, NEK3, NEK2, ANLN, CEP55, PTTG1, AURKB, CCNE2, CCNE1, CDC2, CDC2A, CDC6, MKI67, MSH5, TPX2, SKP2, NUFB, CENPF, CDC20, BIRC5, CENPE, NDC80, ESPL1, UBE2C, CDC25B, CCNB1, MAD2L1, PLK1 | 3.42x10^{−12} |
|                           | Regulation of cell cycle              | 19    | E2F1, CDC6, HOXA13, NEK2, SKP2, CENPF, TTK, PKMYT1, ESPL1, CENPE, ANLN, BIRC5, TACC3, UBE2C, CENK3, CDT1, CCNE2, CCNB1, MAD2L1 | 4.79x10^{−05} |
|                           | Microtubule-based process             | 16    | KIFC2, KIF23, CEPT2, PRC1, NEK2, PSRC1, TTK, ESPL1, AURKA, NDC80, CENPE, TACC3, UBE2C, HOOK1, CENPA, KIF20A | 2.40x10^{−04} |
Investigation of lncRNA profiles in human cancer remains to be performed. Apart from H19, MUC2, XIST and HOTAIR, other prognostic lncRNAs have not been identified in GC. FLVCR1-AS1 has been reported in lung adenocarcinoma by a study based on an miR-lncRNA-mRNA network (36). TP53TG1 is a critical lncRNA responsible for correct response of p53 to DNA damage and acts as a tumor suppressor (37). There is evidence that TP53TG1 expression is elevated in human glioma tissue and TP53TG1 under glucose deprivation may promote cell proliferation and migration by influencing the expression of glucose metabolism associated genes in glioma (38). LINC00221 has been reported to be aberrantly expressed in bladder cancer (39). Li et al. (40) noted that PRSS30P is upregulated in lung adenocarcinoma. SCARNA9 is observed to be overexpressed in breast cancer cells on exposure to cadmium (41). However, ARHGAP5-AS1 and MCF2L-AS1 are rarely studied in cancer. In future studies, the expression levels of ARHGAP5-AS1 and MCF2L-AS1 will be investigated in clinical samples of GC patients since the prognostic value of these lncRNAs was observed for GC.

Correlations between the critical lncRNAs and mRNAs revealed by the WGCNA were used to construct lncRNA-mRNA networks. In order to investigate the molecular mechanisms of the 11 prognostic lncRNAs in GC, GO function and KEGG pathway enrichment analysis were performed for the genes in the construct lncRNA-mRNA networks. The results demonstrated that the genes correlated with the 9 lncRNAs in the blue module (FLVCR1-AS1, H19, LINC00221, MUC2, PRSS30P, SCARNA9, TP53TG1, XIST and ARHGAP5-AS1) were associated with the immune response, regulation of cell activation, regulation of lymphocyte activation and cytokine-cytokine receptor interaction. These results suggested that these 9 lncRNAs may serve important roles in the pathogenesis of GC by regulating their associated genes to affect the immune and inflammatory responses. The genes associated with the 2 lncRNAs (HOTAIR and MCF2L-AS1) in the brown module were revealed to be implicated in cell cycle regulation. This indicated that ARHGAP5-AS1 and MCF2L-AS1 may also be critical in the pathogenesis of GC by regulating their associated genes to influence the cell cycle. A growing body of evidence demonstrates the important roles of inflammation, immune and dysregulated cell cycle control in tumor growth and progression (42-44). Therefore, it can be concluded that the 11 critical lncRNAs may participate in the development and progression of GC by regulating their correlated genes to influence the immune response, inflammatory response and cell cycle.

Based on bioinformatics analysis of existing gene expression data from NCBI GEO, EBI ArrayExpress and TCGA, the present study identified an 11-lncRNA signature that could be used for predicting survival rate of GC patients. These 11 critical lncRNAs may participate in the pathogenesis of GC by regulating their associated genes that are associated with the immune response, inflammatory response and cell cycle. It is hoped that the present study may contribute to an improved understanding of the pathogenesis involved with lncRNAs in GC development and progression. Validation of this 11-lncRNA signature in large cohorts of GC patients and clinical trials are also essential in further investigation.
Acknowledgements
Not applicable.

Funding
No funding was received.

Availability of data and materials
The datasets analyzed during the present study are available from the corresponding author on reasonable request.

Authors' contributions
YZ and HL performed data analyses and wrote the manuscript. WZ and YC contributed significantly to the data analyses and critical revision of the manuscript. GH and WB conceived and designed the study. All authors read and approved the final manuscript.

Ethics approval and consent to participate
Not applicable.

Patient consent for publication
Not applicable.

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

References
1. IARC: World Cancer Report 2014. Stewart BW and Wild CP (eds). World Health Organization, Geneva, 2015.
2. Orditura M, Galizia G, Sforza V, Gambardella V, Fabozzi A, Laterza MM, Andreozzi F, Ventriglia J, Savastano B, Mabilia A, et al: Treatment of gastric cancer. World J Gastroenterol 20: 1635-1649, 2014.
3. Chen W, Zheng R, Baade PD, Zhang S, Zeng H, Bray F, Jemal A, et al: Cancer statistics in China, 2014. CA Cancer J Clin 64: 9-29, 2014.
4. Li H, Yu B, Li J, Su L, Yan M, Zhu Z and Liu B: Overexpression of IncRNA H19 enhances carcinogenesis and metastasis of gastric cancer. Oncotarget 5: 2318-2329, 2014.
5. Pan W, Liu L, Wei J, Ge Y, Zhang J, Chen H, Zhou L, Yuan Q, Zhou C and Yang M: A functional IncRNA HOTAIR genetic variant contributes to gastric cancer susceptibility. Mol Carcinog 55: 90-96, 2016.
6. Liu XH, Sun M, Nie FQ, Ge YB, Zhang EB, Yin DD, Kong R, Xia R, Lu KH, Li JH, et al: Lnc RNA HOTAIR functions as a competing endogenous RNA to regulate HER2 expression by sponging miR-331-3p in gastric cancer. Mol Cancer 13: 92, 2014.
7. Zhang EB, Kong R, Yin DD, You LH, Sun M, Han L, Xu TP, Xia R, Yang JS, De W and Chen JC: Long noncoding RNA ANRIL indicates a poor prognosis of gastric cancer and promotes tumor growth by epigenetically silencing of miR-99a/miR-449a. Oncotarget 5: 2276-2292, 2014.
8. Miao Y, Su J, Xu SY, Liang GY, Pu YP and Yin LH: Comprehensive analysis of a novel 4-lncRNA signature as a prognostic biomarker for human gastric cancer. Oncotarget 8: 75007-75024, 2017.
9. Danford T, Rolfe A and Gifford D: GSE: A comprehensive database system for the representation, retrieval, and analysis of microarray data. Pac Symp Biocomput: 539-550, 2008.
10. Ritchie ME, Shipson B, Wu D, Hu Y, Law CW, Shi W and Smyth GK: Limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res 43: e47, 2015.
11. Zhou M, Guo M, He D, Wang X, Cui Y, Yang H, Hao D and Sun J: A potential signature of eight long non-coding RNAs predicts survival in patients with non-small cell lung cancer. J Transl Med 13: 231, 2015.
12. Zhou M, Xu W, Yue X, Zhao H, Wang Z, Shi H, Cheng L and Sun J: Relapse-related long non-coding RNA signature to improve prognosis prediction of lung adenocarcinoma. Oncotarget 7: 29720-29738, 2016.
13. Larkin MA, Blackshields G, Brown NP, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, et al: Clustal W and Clustal X version 2.0. Bioinformatics 23: 2947-2948, 2007.
14. Zhao X, Xue Q, Liu Q, Guo Y and Chen Z: Colon cancer recurrence-associated genes revealed by WGCNA co-expression network analysis. Mol Med Rep 16: 6499-6505, 2017.
15. Langfelder P and Horvath S: WGCNA: An R package for weighted correlation network analysis. BMC Bioinformatics 9: 559, 2008.
16. Qi C, Hong L, Cheng Z and Yin Q: Identification of metastasis-associated genes in colorectal cancer using metaDE and survival analysis. Oncol Lett 11: 568-574, 2016.
17. Wang X, Kang DD, Shen K, Song C, Lu S, Chang LC, Liao SG, Hsu Z, Tang S, Ding Y, et al: An R package suite for microarray meta-analysis in quality control, differentially expressed gene analysis and pathway enrichment detection. Bioinformatics 28: 2534-2536, 2012.
18. Goeman JJ: L1 penalized estimation in the Cox proportional hazards model. Biostat J 52: 70-84, 2010.
19. Tibshirani R: The lasso method for variable selection in the Cox model. Stat Med 16: 385-395, 1997.
20. Cristescu R, Lee J, Nebozhy M, Kim KM, Ting JC, Wong SS, Liu J, Yue YG, Wang J, Yu K, et al: Molecular analysis of gastric cancer identifies subtypes associated with distinct clinical outcomes. Nat Med 21: 449-456, 2015.
21. Parrish RS and Spencer HJ III: Effect of normalization on significance testing for oligonucleotide microarrays. J Biopharm Stat 14: 575-589, 2004.
22. Huang da W, Sherman BT and Lempicki RA: Bioinformatics enrichment tools: Paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res 37: 1-13, 2009.
23. Huang da W, Sherman BT and Lempicki RA: Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc 4: 44-57, 2009.
24. Mikkelsen TS, Ku M, Jaffe DB, Issac B, Lieberman E, Giannoukos G, Alvarez P, Brockman W, Kim TK, Koche RP, Issac B, Lieberman E, et al: Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature 448: 553-560, 2007.
25. Liu J and Esteller M: IncRNAs and microRNAs with a role in cancer development. Biochim Biophys Acta 1859: 1440-1452, 2015.
26. Li H, Yu B, Li J, Su L, Yan M, Zhu Z and Liu B: Overexpression of IncRNA H19 enhances carcinogenesis and metastasis of gastric cancer. Oncotarget 5: 2318-2329, 2014.
27. Liu XH, Sun M, Nie FQ, Ge YB, Zhang EB, Yin DD, Kong R, Xia R, Lu KH, Li JH, et al: Lnc RNA HOTAIR functions as a competing endogenous RNA to regulate HER2 expression by sponging miR-331-3p in gastric cancer. Mol Cancer 13: 92, 2014.
28. Wang X, Kang DD, Shen K, Song C, Lu S, Chang LC, Liao SG, Hsu Z, Tang S, Ding Y, et al: An R package suite for microarray meta-analysis in quality control, differentially expressed gene analysis and pathway enrichment detection. Bioinformatics 28: 2534-2536, 2012.
29. Goeman JJ: L1 penalized estimation in the Cox proportional hazards model. Biostat J 52: 70-84, 2010.
30. Parrish RS and Spencer HJ III: Effect of normalization on significance testing for oligonucleotide microarrays. J Biopharm Stat 14: 575-589, 2004.
31. Huang da W, Sherman BT and Lempicki RA: Bioinformatics enrichment tools: Paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res 37: 1-13, 2009.
32. Huang da W, Sherman BT and Lempicki RA: Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc 4: 44-57, 2009.
33. Mikkelsen TS, Ku M, Jaffe DB, Issac B, Lieberman E, Giannoukos G, Alvarez P, Brockman W, Kim TK, Koche RP, Issac B, Lieberman E, et al: Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature 448: 553-560, 2007.
34. Liu J and Esteller M: IncRNAs and microRNAs with a role in cancer development. Biochim Biophys Acta 1859: 1440-1452, 2015.
35. Li H, Yu B, Li J, Su L, Yan M, Zhu Z and Liu B: Overexpression of IncRNA H19 enhances carcinogenesis and metastasis of gastric cancer. Oncotarget 5: 2318-2329, 2014.
36. Liu XH, Sun M, Nie FQ, Ge YB, Zhang EB, Yin DD, Kong R, Xia R, Lu KH, Li JH, et al: Lnc RNA HOTAIR functions as a competing endogenous RNA to regulate HER2 expression by sponging miR-331-3p in gastric cancer. Mol Cancer 13: 92, 2014.
37. Wang X, Kang DD, Shen K, Song C, Lu S, Chang LC, Liao SG, Hsu Z, Tang S, Ding Y, et al: An R package suite for microarray meta-analysis in quality control, differentially expressed gene analysis and pathway enrichment detection. Bioinformatics 28: 2534-2536, 2012.
38. Goeman JJ: L1 penalized estimation in the Cox proportional hazards model. Biostat J 52: 70-84, 2010.
39. Parrish RS and Spencer HJ III: Effect of normalization on significance testing for oligonucleotide microarrays. J Biopharm Stat 14: 575-589, 2004.
40. Huang da W, Sherman BT and Lempicki RA: Bioinformatics enrichment tools: Paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res 37: 1-13, 2009.
34. Chen DL, Ju HQ, Lu YX, Chen LZ, Zeng ZL, Zhang DS, Luo HY, Wang F, Qiu MZ, Wang DS, et al: Long non-coding RNA XIST regulates gastric cancer progression by acting as a molecular sponge of miR-101 to modulate EZH2 expression. J Exp Clin Cancer Res 35: 142, 2016.

35. Endo H, Shiroki T, Nakagawa T, Yokoyama M, Tamai K, Yamanami H, Fujiya T, Sato I, Yamaguchi K, Tanaka N, et al: Enhanced expression of long non-coding RNA HOTAIR is associated with the development of gastric cancer. PLoS One 8: e77070, 2013.

36. Li DS, Ainiwaer JL, Sheyhiding I, Zhang Z and Zhang LW: Identification of key long non-coding RNAs as competing endogenous RNAs for miRNA-mRNA in lung adenocarcinoma. Eur Rev Med Pharmacol Sci 20: 2285-2295, 2016.

37. Diaz-Lagares A, Crujeiras AB, Lopez-Serra P, Soler M, Setien F, Goyal A, Sandoval J, Hashimoto Y, Martinez-Cardús A, Gomez A, et al: Epigenetic inactivation of the p53-induced long noncoding RNA TP53 target 1 in human cancer. Proc Natl Acad Sci USA 113: E7535-E7544, 2016.

38. Chen X, Gao Y, Li D, Cao Y and Hao B: LncRNA-TP53TG1 participated in the stress response under glucose deprivation in glioma. J Cell Biochem 118: 4897-4904, 2017.

39. Wang H, Niu L, Jiang S, Zhai J, Wang P, Kong F and Jin X: Comprehensive analysis of aberrantly expressed profiles of lncRNAs and miRNAs with associated ceRNA network in muscle-invasive bladder cancer. Oncotarget 7: 86174-86185, 2016.

40. Li J, Li P, Zhao W, Yang R, Chen S, Bai Y, Dun S, Chen X, Du Y, Wang Y, et al: Expression of long non-coding RNA DLX6-ASI in lung adenocarcinoma. Cancer Cell Int 15: 48, 2015.

41. Lubovac-Pilav Z, Borràs DM, Ponce E and Louie MC: Using expression profiling to understand the effects of chronic cadmium exposure on MCF-7 breast cancer cells. PLoS One 8: e84646, 2013.

42. Candido J and Hagemann T: Cancer-related inflammation. J Clin Immunol 33 (Suppl 1): S79-S84, 2013.

43. Elinav E, Nowarski R, Thaiss CA, Hu B, Jin C and Flavell RA: Inflammation-induced cancer: Crosstalk between tumours, immune cells and microorganisms. Nat Rev Cancer 13: 759-771, 2013.

44. Evan GI and Vousden KH: Proliferation, cell cycle and apoptosis in cancer. Nature 411: 342-348, 2001.