OSgc: A Web Portal to Assess the Performance of Prognostic Biomarkers in Gastric Cancer

Longxiang Xie†, Qiang Wang†, Zhongyi Yan†, Yali Han†, Xiaoyu Ma†, Huimin Li†, Lu Zhang†, Xianzhe Li* and Xiangqian Guo*

1 Department of Preventive Medicine, Institute of Biomedical Informatics, Cell Signal Transduction Laboratory, Bioinformatics Center, Henan Provincial Engineering Center for Tumor Molecular Medicine, School of Software, School of Basic Medical Sciences, Henan University, Kaifeng, China, 2 Department of Thoracic Surgery, The Affiliated Nanshi Hospital of Henan University, Nanyang, China

Evaluating the prognostic value of genes of interest in different populations of gastric cancer (GC) is difficult and time-consuming for basic and translational researchers even though many datasets are available in public dataset depositories. In the current study, we developed a robust web-based portal called OSgc (Online consensus Survival analysis of gastric cancer) that enables easy and swift verification of known and novel biomarker candidates in GC. OSgc is composed of gene expression profiling data and clinical follow-up information of 1,824 clinical GC cases, which are collected from 7 public independent datasets derived from Gene Expression Omnibus (GEO) and The Cancer Genome Atlas (TCGA). By OSgc, users input the official gene symbol and will promptly retrieve the Kaplan–Meier survival plot with hazard ratio (HR) and log rank p value on the output webpage, by which users could assess the prognostic value of interesting genes for GC patients. Five survival end points containing overall survival, progression-free survival, progression-free interval, relapse-free survival, and disease-free survival could be measured in OSgc. OSgc can greatly help cancer biologists and clinicians to explore the effect of gene expression on patient survival. OSgc is freely available without restrictions at http://bioinfo.henu.edu.cn/GC/GCList.jsp.

Keywords: gastric cancer, gene expression, GEO, OSgc, prognostic biomarker, survival analysis, TCGA

INTRODUCTION

Gastric cancer (GC) is the fourth leading factor of cancer mortality in the world. In 2020, GC occurred in 1,089,103 people and resulted in 768,793 deaths. Although there are many advances in treatment of GC, patients have poor prognosis and the 5-year survival rate is just 5%–20%. Prognostic biomarkers can assist clinicians in assessing the risk of clinical outcomes including cancer recurrence or disease progression in the future (1, 2). Molecular characteristics such as gene expression and somatic mutations have been reported to represent the primary source of prognostic biomarker (3, 4). A recent study showed that high SETD2 (SET domain-containing protein 2, also known as HYPB) expression

Abbreviations: GC, gastric cancer; GEO, Gene Expression Omnibus; TCGA, The Cancer Genome Atlas; SETD2, SET domain-containing protein 2; BC, breast cancer; OC, ovarian cancer; HCC, hepatocellular carcinoma; LUCA, lung carcinoma.
was correlated with better prognosis for GC patients, and its overexpression in GC cell lines significantly inhibits cell proliferation, migration, and invasion (5). Moreover, lower MTBP (MDM2-binding protein) expression and HOXA5 (homeobox A5) expression were significantly associated with longer overall survival time in GC (6). However, these present retrospective cohort studies were limited to relatively small case series and further validation is required when these findings are going to be translated.

The expression profiling of gastric cancer has been performed using high-throughput technology such as microarrays and RNA sequencing. These data have been used to measure the association of mRNAs to clinical outcomes in GC patients, while a key step for biomarker development is the biomarker validation in multiple independent cohorts. Even though massive public datasets are available, multistep specialized analyses such as exploring repositories and acquiring and processing data make it difficult for most researchers. Previous studies have reported several good online prognosis tools including PROGgene (7), PRECOG (8), OncoLnc (9), GEPIA (10), KM plotter (11, 12), and ITTACA (13), which are available to help explore changes of individual genes and their association with GC patients’ survival. However, these above tools are restricted regarding low number of clinical cases, lacking most updated data, or login/registering, or limited survival terms.

Hence, to aid and facilitate the evaluation and verification of prognostic biomarkers in independent cohorts, we developed OSgc, a free and easy-to-use web portal to perform the survival analysis in GC. OSgc is composed of 7 public datasets with available follow-up data for 1,824 GC cases from TCGA and GEO databases, and by which survival analysis can be completed in minutes.

MATERIAL AND METHODS

Data Collection
Data were collected by searching for keywords related to gastric cancer, clinical outcome, mRNA profiles, and ≥20 samples. The searches were performed in GEO (http://www.ncbi.nlm.nih.gov/geo) and TCGA (https://tcga-data.nci.nih.gov/tcga). TCGA data are level 3 RNA-Seq data.

Web Implementation
The OSgc web portal was set up as we previously developed with minor modification (14, 15). In brief, OSgc contains two main components: storage and data analysis (Figure 1). A Java implementation was used to realize OSgc. OSgc used the SQL Server database to provide the storage and management of the gene expression profiles and clinical data for GC and used the Browser/Server architecture network management system to manage the database. R packages including “survminer” and “ggplot2” were used to plot the Kaplan–Meier curves.

Searching Previously Reported Prognostic Biomarkers
A PubMed search was performed to find gastric cancer’s prognostic biomarkers using keywords “gastric cancer,” “stomach cancer,” “adenocarcinoma,” “GC,” “survival,” and “gene expression.” Only studies published in English were encompassed. Eligibility criteria also contained the investigation of the biomarker in >30 patients while biomarkers described only in experimental models or meta-analysis or bioinformatics were excluded.

Statistical Analysis
The relationship between clinic-pathological factors and clinical survival outcomes was analyzed by GraphPad Prism 8. R package “survival” was used to perform Cox regression analysis to calculate hazard ratios (HR) and 95% confidence intervals (95% CIs). HR with their corresponding 95% CIs was assessed to explore the prognostic significance of gene of interest in gastric cancer. If a HR is >1 and the 95% CI did not cross 1, the typed gene will show a worse prognosis in the high gene expression group. If a HR is <1 and the 95% CI did not cross 1, it suggests a better prognosis of typed gene in the high gene expression group. Prognostic value was assessed by KM plot analysis and log-rank test. p value < 0.05 is regarded as statistically significant.

RESULTS

Clinical Information of GC Datasets Used in OSgc
To our knowledge, OSgc provides the largest compilation of expression profiling datasets related to clinical outcomes,
comprising 7 datasets and 1,824 clinical GC cases. The median age of these patients is 67 years, and the ratio of male to female is 2.2:1. 1,392 patients have OS, 776 patients have DFS, 420 patients have PFS, and 337 patients have RFS. A summary of above GC cohorts is shown in Table 1.

To explore the relationship between clinical characteristics and outcomes, we performed Kaplan–Meier plots for GC patients stratified by TNM stage and gender in datasets used for OSgc (Figure 2). In these patients, TNM stage was significantly related with OS \((p < 0.0001)\) and DFS \((p < 0.0001)\), as we knew (22). Nevertheless, gender showed no significant association with OS \((p = 0.4939)\) and DFS \((p = 0.7764)\).

**Web Interface**

As shown in Figure 3, OSgc could be easily used by typing only the gene symbol and selecting the individual/combined cohort, stage, gender, grade, and age on the input webpage. After then, “Kaplan–Meier plot” is clicked, and the survival outcome plots will be shown up on the output webpage in less than a minute (might vary if advanced plots are selected). Furthermore, users can easily obtain KM plots for multiple genes (with maximum 5 genes).

**Validation and Application**

To assess the capabilities of OSgc in the evaluation of prognostic biomarkers, we collected 20 previously published GC prognostic biomarkers through PubMed search (shown in Table 2). These published prognostic biomarkers include 16 unfavorable and 4 favorable prognostic biomarkers which have been verified by tissue-based immunohistochemistry (IHC), Western blot, and RT-PCR in literatures. To test these published prognostic biomarkers in OSgc, OS or DFS was selected as survival term in the combined datasets in OSgc. Combined datasets mean that each

| TABLE 1 | Clinical characteristics of individual dataset in OSgc. |
|---|---|---|---|---|---|---|---|
| ID | GPL | Sample size | Gender | Number of deaths | Stage (I/II/III/IV) | Median age (years) | Survival terms | Reference |
| GSE22377 | GPL570 | 43 | Female/male | 30 | – | 64 | OS, DFS | (16) |
| GSE26253 | GPL8432 | 432 | – | – | 68/167/130/67 | 59 | DFS | (17) |
| GSE29272 | GPL96 | 126 | 27/99 | 31 | 1/44/81 | 5/4/108/8 | OS | (18) |
| GSE57303 | GPL570 | 70 | 14/56 | 7 | 0/4/34/32 | 68 | OS | (19) |
| GSE62254 | GPL570 | 300 | 101/199 | 152 | 30/96/95/77 | 64 | OS, DFS | (20) |
| GSE84437 | GPL6947 | 433 | 137/296 | 209 | 11/38/92/292 | 62 | OS | (21) |
| TCGA-GC RNAseq | 420 | 134/286 | 169 | 10/150/251 | 55/128/202/40 | 67 | OS, RFS, PFI, PFS | (20) |
| Total | 1824 | 441/951 | 598 | 11/194/332 | 171/593/677/519 | 9 samples were GX that grade cannot be defined. |

\[ \text{FIGURE 2} \mid \text{The association of GC patient survival with gender (A, B) and stage (C, D) in OSgc.} \]
cohort was divided separately into strata by selecting the proper cutoff value, which are then put together for survival analysis. The hazard ratio and \( p \) value in original studies and OSgc are given in Table 2. The test results showed that the prognostic values of 17 genes line up with previous reports, while 3 genes did not reach significance in OSgc.

**Comprehensive Analysis of the Prognosis Significance of E2Fs by OSgc**

Increasing evidence has indicated that E2Fs, a family of critical transcription factors that regulate cell cycle progression and other cellular processes, are aberrantly expressed and involved in the tumor progression in various malignant tumors (41, 42). Recently, four research groups had systematically studied the expression patterns and prognostic values of eight E2Fs in patients with breast cancer (BC) (43), ovarian cancer (OC) (44), hepatocellular carcinoma (HCC) (45), and lung carcinoma (46) by investigating a series of databases. For example, Huang et al. had shown that high expression of individual E2Fs was related with poor prognosis in HCC patients (45). However, the expression and prognostic significance of each E2F in gastric cancer have not yet been elucidated. Herein, we can easily explore the prognostic significance for all E2F members in GC by using web tool OSgc. The results showed that the higher transcriptional levels of both E2F2 and E2F8 were associated with better overall survival in gastric cancer patients (Figure 4). However, other E2Fs members were not significantly correlated with OS of GC patients (data not shown).

**DISCUSSION**

Prognostic biomarkers are an important supplement to traditional clinical and histopathological features, for example, tumor size and lymph node metastasis, which cannot completely predict the prognosis of patients with cancer. The development of biomarkers by genomic, transcriptomic, and proteomic methods holds the promise of “individualized medicine,” bringing a new ground-breaking point to disease diagnosis, classification, and prognosis. The gene expression profiling datasets in TCGA and GEO are of great value in deepening our understanding of the underlying molecular mechanisms involved in GC, as well as in the identification of novel diagnostic and prognostic biomarkers (47, 48). For the maximum utilization of GEO and TCGA data resources, it is very necessary to provide a web-based portal that
allows clinicians and cancer biologists (regardless of having bioinformatics background or not) to easily access, analyze, and visualize the data. To identify genes related with patient survival is one possible way to prioritize genes with oncogenic or tumor-suppressor properties for further study. The user-friendly web portal of OSgc promotes the gene identification for survival associations in GC. Compared with published survival web tools such as OncoLnc (9), GEPIA (10), and KM plotter (11, 12), GEPIA and KM plotters are good in performing an extensive survey of prognosis in general cancer types; however, it has limited cases of GC (GEPIA contains 375 patients from TCGA; KM plotter includes 1,440 cases from TCGA and GEO). OncoLnc only containing

### TABLE 2 | Test the performance of previously reported prognostic biomarker candidates using OSgc.

| Gene symbol | Literature data | OSgc data |
|-------------|-----------------|-----------|
| **HOXA5**   | 30 OS Unfavorable qRT-PCR and Western blot (23) | Combined 1.3724 (1.114–1.5818) 0.0015 ✓ |
| **CAP2**    | 436 OS Unfavorable RT-PCR (24) | Combined 1.5207 (1.2645–1.7858) <0.0001 ✓ |
| **LAMA4**   | 388 OS Unfavorable qRT-PCR (25) | Combined 1.4904 (1.2305–1.8053) <0.0001 ✓ |
| **MTBP**    | 352 OS Unfavorable qRT-PCR and Western blot (26) | GSE22377 2.7219 (1.2506–6.9242) 0.0116 ✓ |
| **RAI14**   | 68 OS Unfavorable IHC (27) | Combined 0.6117 (1.3304–1.9524) <0.0001 ✓ |
| **SETD2**   | 153 OS Favorable qRT-PCR and IHC (28) | Combined 0.6946 (1.3304–1.0078) 0.0603 x |
| **NDRG4**   | 286 OS Unfavorable IHC (29) | Combined 1.4149 (1.1904–1.6818) 1e-04 ✓ |
| **SPARC**   | 227 OS Unfavorable qRT-PCR (30) | Combined 1.4217 (1.1612–1.7407) 7e-04 ✓ |
| **HOXB9**   | 190 OS Favorable qRT-PCR (31) | Combined 1.8461 (0.7008–1.0215) 0.821 x |
| **DDC**     | 39 OS Favorable qRT-PCR and IHC (32) | GSE22377 2.805 (0.097–8.108) 0.0198 ✓ |
| **ERCC1**   | 106 OS Unfavorable qRT-PCR and IHC (33) | GSE62254 1.6054 (1.1377–2.2662) 0.007 ✓ |
| **STAT3**   | 50 OS Unfavorable RT-PCR, Western blot and IHC (34) | TCGA 1.4306 (1.0138–2.0188) 0.0416 ✓ |
| **IGFBP7**  | 247 OS Unfavorable qRT-PCR and IHC (35) | Combined 1.4577 (1.2253–1.7342) <0.0001 ✓ |
| **TIMP3**   | 17 OS Unfavorable qRT-PCR and IHC (36) | Combined 1.4779 (1.2283–1.7783) <0.0001 ✓ |
| **KLK6**    | 66 OS Unfavorable qRT-PCR and IHC (37) | GSE62254 1.4878 (1.0523–2.1037) 0.0246 ✓ |
| **NNMT**    | 641 OS Unfavorable qRT-PCR (38) | Combined 1.3734 (1.1529–1.6382) 4e-04 ✓ |
| **ATAD2**   | 166 OS Unfavorable qRT-PCR (39) | GSE22377 3.4084 (1.5908–7.3027) 0.0016 ✓ |
| **CXC3**    | 96 OS Favorable RT-PCR and qRT-PCR (40) | Combined 0.7203 (0.5923–0.8759) 0.0010 ✓ |
| **MYD3**    | 166 OS Unfavorable IHC (41) | Combined 1.1857 (0.9798–1.435) 0.0801 x |
| **S100A4**  | 434 OS Unfavorable qRT-PCR and Western blot (42) | Combined 1.2081 (1.0103–1.4447) 0.0882 ✓ |

Xie et al. Survival Analysis Tool for GC

![FIGURE 4](image-url) | The prognostic value of E2F2 and E2F8 in GC patients (OS in OSgc). (A) KM survival plot for E2F2 suggests that its high expression (red) indicates favorable prognosis (GC patients were separated by the quarter of gene expression level); (B) KM survival plot for E2F8 suggests that its high expression (red) indicates favorable prognosis (GC patients were separated by the quarter of gene expression level); E2F2, E2F transcription factor 2; E2F8, E2F transcription factor 8; GC, gastric cancer; OS, overall survival.
TCGA data was developed to assess the prognostic significance of non-coding genes. More importantly, OSgc has integrated seven GC cohorts and incorporated the clinical covariates to provide more informative survival plots to the researchers.

We have demonstrated here the usefulness of OSgc using 20 previously published prognostic biomarkers as examples for rapid survival analysis. 85% (17 of 20) of these reported prognostic biomarkers were confirmed to be prognostically significant in OSgc. The three genes without statistical significance in OSgc could be due to several factors including different detecting methods (the analysis method for the HOXB8 and SMYD3 prognosis study was IHC and starting material (the material for SETD2 prognosis study was paraffin-embedded FFPE tissues). Moreover, we used OSgc to quickly systematically analyze the potential prognostic values of E2F family members and found that both E2F2 and E2F8 are favorable prognostic biomarkers for GC patients. Previous studies indicated that E2F2 exhibited as a tumor suppressor in epithelial tissues (49) or Myc-induced T cell lymphomagenesis (50), and overexpression of E2F2 inhibited the progression of these tumors. E2F8 is also a critical tumor suppressor for postnatal liver development (51). Compared with other web tools, OSgc is a free tool with the largest number of gastric cancers to perform the survival analysis based on gene expression. In the future, we will check whether the new standard datasets (including gene expression profile and clinical survival information) from different databases (such as TCGA, GEO and ArrayExpress) come out every 3 months, then we will incorporate these new gene expression profiles as well as additional utilities which are suggested by the users.

All in all, OSgc is a free and easy-to-use web portal to assess the performance of potential prognostic biomarkers for gastric cancer.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material. Further inquiries can be directed to the corresponding authors.

AUTHOR CONTRIBUTIONS

XG, XL: study concept and design. LX, QW, ZY: acquisition of data. QW, LX, YH, HL, LZ: analysis and interpretation of data. LX, QW, ZY, XG, XL: draft of the manuscript. XG, LX, QW, ZY, XL: critical revision of the manuscript for intellectual content. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the National Natural Science Foundation of China (No. U2004136), supporting program for Central Plain Young Top Talents (ZYQQR201912176), Kaifeng Science and technology development Plan (2103005) program from the Academy for Advanced Interdisciplinary Studies of Henan University (Y21008L).

REFERENCES

1. Goossens N, Nakagawa S, Sun X, Hoshida Y. Cancer Biomarker Discovery and Validation. *Transl Cancer Res* (2015) 4:256–69. doi: 10.3978/j.issn.2218-676X.2015.06.04
2. Deng Y, Guo Y, Ma X, Chao X, Wang F, Cai L, et al. Systemic Analysis of the Expression and Prognostic Significance of PAKs in Breast Cancer. *Genomics* (2020) 112(3):2433–44. doi: 10.1016/j.ygeno.2020.01.016
3. Xie L, Deng Y, Guo J, Sun X, Xie T, Zhang L, et al. High KRT8 Expression Independently Predicts Poor Prognosis for Lung Adenocarcinoma Patients. *Genes* (2019) 10:36. doi: 10.3390/genes10010036
4. Zhao G, Liu C, Wen X, Luan G, Xie L, Guo X. The Translational Values of TRIM Family in Pan-Cancers: From Functions and Mechanisms to Clinics. *Pharmacol Ther* (2021) 227:107881. doi: 10.1016/j.pharmthera.2021.107881
5. Chen Z, Raghoonundun C, Chen W, Zhang Y, Tang W, Fan X, et al. SETD2 Indicates Favourable Prognosis in Gastric Cancer and Suppresses Cancer Cell Proliferation, Migration, and Invasion. *Biochem Biophys Res Commun* (2018) 498(3):57–85. S0006291X18304960. doi: 10.1016/j.bbrc.2018.03.022
6. Wang W, Chen Z, Jin J, Long Z, Liu X, Cai H, et al. MDM2 Binding Protein as a Predictor of Metastasis and a Novel Prognostic Biomarker in Patients With Gastric Cancer. *Oncol Lett* (2017) 14:6409–16. doi: 10.3892/ol.2017.7031
7. Goswami CP, Nakshatri H. *PROGeneV2: Enhancements on the Existing Database.* * BMC Cancer* (2014) 14:970. doi: 10.1186/1471-2407-14-970
8. Gentles AJ, Newman AM, Liu CL, Bratman SV, Feng W, Kim D, et al. The Prognostic Landscape of Genes and Infiltrating Immune Cells Across Human Cancers. *Nat Med* (2015) 21:938–45. doi: 10.1038/nm.3909
9. Anaya J. *OncoLnc: Linking TCGA Survival Data to miRNAs, lncRNAs, and IncRNAs.* *PeerJ Comput Sci* (2016) 2:e67. doi: 10.7717/peerj-cs.67
10. Tang Z, Li K, Kang B, Gao G, Li C, Zhang Z. GEPIA: A Web Server for Cancer and Normal Gene Expression Profiling and Interactive Analyses. *Nucleic Acids Res* (2017) 45(W1):W98–W102. doi: 10.1093/nar/gkx247

11. Szaś AM, Łanczyk A, Nagy Á, Förster S, Hark K, Green JE, et al. Cross-Validation of Survival Associated Biomarkers in Gastric Cancer Using Transcriptomic Data of 1,065 Patients. *OncoTarget* (2016) 7:49322–33. doi: 10.18632/oncotarget.10337
12. Gyorffy B, Surowiak P, Budczies J, Łanczyk A. Online Survival Analysis Software to Assess the Prognostic Value of Biomarkers Using Transcriptomic Data in Non-Small-Cell Lung Cancer. *PloS One* (2013) 8:e82241. doi: 10.1371/journal.pone.0082241
13. Elffil A, Lair S, Verbeke C, Rosa PL, Radvanyi F, Barillot E. ITTACA: A New Database for Integrated Tumor Transcriptome Array and Clinical Data Analysis. *Nucleic Acids Res* (2006) 34:D613–6. doi: 10.1093/nar/gkj022
14. Xie L, Wang Q, Deng Y, Ge L, Sun X, Li N, et al. OSkcar: A Web Tool for Identifying Prognostic Biomarkers in Kidney Renal Clear Cell Carcinoma. *Future Oncol (London England)* (2019) 15:3103–10. doi: 10.2217/fon-2019-0296
15. Xie L, Wang Q, Nan F, Ge L, Deng Y, Sun X, et al. OSacc: Gene Expression-Based Survival Analysis Web Tool For Adrenocortical Carcinoma. *Cancer Manag* (2019) 11:9145–52. doi: 10.2147/CMAR.2515586
16. Förster S, Gretschel S, Jons T, Yashiro M, Kummer W. THBS4, a Novel Stromal Molecule of Diffuse-Type Gastric Adenocarcinomas, Identified by Transcriptome-Wide Expression Profiling. *Mod Pathol* (2011) 24:1390. doi: 10.1038/modpathol.2011.99
17. Lee J, Sohn I, Do I-G, Kim K-M, Park SH, Park JO, et al. Nanostring-Based Multigene Assay to Predict Recurrence for Gastric Cancer Patients After Surgery. *PloS One* (2014) 9:e90133. doi: 10.1371/journal.pone.0090133
18. Soldner F, Laganière J, Cheng AW, Hockemeyer D, Gao Q, Alagappan R, et al. Generation of Isogenic Pluripotent Stem Cells Differing Exclusively at Two Early Onset Parkinson Point Mutations. *Cell* (2011) 146:318–31. doi: 10.1016/j.cell.2011.06.019
19. Qian Z, Zhu G, Tang L, Wang M, Zhang L, Fu J, et al. Whole Genome Gene Copy Number Profiling of Gastric Cancer Identifies PAK1 and KRAS Gene
