Original Article

# Phosphoglucose isomerase gene expression as a prognostic biomarker of gastric cancer

**Han-Chen Huang[1,2*], Xian-Zi Wen[3*], Hua Xue[1,2], Run-Sheng Chen[1,4], Jia-Fu Ji[3], Lei Xu[5,6]**

[1]Key Laboratory of RNA Biology, Institute of Biophysics, Chinese Academy of Sciences, Beijing 100101, China; [2]University of Chinese Academy of Sciences, Beijing 100049, China; [3]Key Laboratory of Carcinogenesis and Translational Research (Ministry of Education/Beijing), Division of Gastrointestinal Cancer Translational Research Laboratory, Peking University Cancer Hospital & Institute, Beijing 100142, China; [4]Guangdong Geneway Decoding Bio-Tech Co.Ltd, Foshan 528316, China; [5]Centre for Cognitive Machines and Computational Health (CMaCH), School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China; [6]Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong 999077, China

*These authors contributed equally to this work.

*Correspondence to*: Lei Xu. Centre for Cognitive Machines and Computational Health (CMaCH), School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China. Email: lxu@cs.sjtu.edu.cn; Jia-Fu Ji. Key Laboratory of Carcinogenesis and Translational Research (Ministry of Education/Beijing), Division of Gastrointestinal Cancer Translational Research Laboratory, Peking University Cancer Hospital & Institute, Beijing 100142, China. Email: jijiafu@bjmu.edu.cn; Run-Sheng Chen. Key Laboratory of RNA Biology, Institute of Biophysics, Chinese Academy of Sciences, Beijing 100101, China. Email: crs@sun5.ibp.ac.cn.

## Abstract

**Objective:** Tumor heterogeneity renders identification of suitable biomarkers of gastric cancer (GC) challenging. Here, we aimed to identify prognostic genes of GC using computational analysis.

**Methods:** We first used microarray technology to profile gene expression of GC and paired nontumor tissues from 198 patients. Based on these profiles and patients' clinical information, we next identified prognostic genes using novel computational approaches. Phosphoglucose isomerase, also known as glucose-6-phosphate isomerase (*GPI*), which ranked first among 27 candidate genes, was further investigated by a new analytical tool namely enviro-geno-pheno-state (E-GPS) analysis. Suitability of *GPI* as a prognostic marker, and its relationship with physiological processes such as metabolism, epithelial-mesenchymal transition (EMT), as well as drug sensitivity were evaluated using both our own and independent public datasets.

**Results:** We found that higher expression of *GPI* in GC correlated with prolonged survival of patients. Particularly, a combination of *CDH2* and *GPI* expression effectively stratified the outcomes of patients with TNM stage II/III. Down-regulation of *GPI* in tumor tissues correlated well with depressed glucose metabolism and fatty acid synthesis, as well as enhanced fatty acid oxidation and creatine metabolism, indicating that *GPI* represents a suitable marker for increased probability of EMT in GC cells.

**Conclusions:** Our findings strongly suggest that *GPI* acts as a novel biomarker candidate for GC prognosis, allowing greatly enhanced clinical management of GC patients. The potential metabolic rewiring correlated with *GPI* also provides new insights into studying the relationship between cancer metabolism and patient survival.

**Keywords:** Gastric cancer; gene expression profile; prognostic biomarker; phosphoglucose isomerase; tumor metabolism

Submitted Dec 28, 2018. Accepted for publication May 16, 2019.
doi: 10.21147/j.issn.1000-9604.2019.05.07
**View this article at:** https://doi.org/10.21147/j.issn.1000-9604.2019.05.07

## Introduction

Among all cases of cancer worldwide, gastric cancer (GC) ranks fifth in incidence, and third in cancer-related mortality (GLOBOCAN, 2012) (1). Despite recent improvements in treatment, patients with stage III and IV GC exhibit 5-year overall survival (OS) rates of 9.2%−19.8% and 4.0%, respectively (National Cancer Institute, 2014). With advances in high throughput technology, gene expression profiling has been widely adopted to identify molecular factors associated with progression, recurrence and metastasis of GC (2,3). However, biomarkers currently used in clinical management of GC patients are limited in prognostics and therapeutics, due to their insufficient sensitivity and specificity. Thus, exploring novel effective biomarkers is critical for improved clinical management of GC patients (4).

Previous studies using gene expression microarray to identify GC-related biomarkers, mostly focused on analyzing tumor tissues, rather than use large-scale data from tumors and paired adjacent nontumor tissues (2,3). Recent analysis on pan-cancer data demonstrated that nontumor tissues are complementary for improved cancer prognosis. These comparative studies also showed that cancer microenvironments play pivotal roles in cancer patient survival by influencing the adjacent cell metabolism or the *in situ* immunization (5). Reprogrammed metabolism is no longer considered a mere consequence of oncogenic transformation, but a critical hallmark of cancer (6,7). By summarizing the primary tumor-related metabolic processes, several systematic studies provide proof that metabolic genes are highly suitable markers for both clinical prognosis and therapy (8,9).

Glucose-6-phosphate isomerase (GPI) is a housekeeping cytosolic enzyme that catalyzes the interconversion between glucose-6-phosphate (G6P) and fructose-6-phosphate (F6P), a process that plays a pivotal role in glycolytic and gluconeogenic pathways. In contrast to normal cells which metabolize glucose mainly via oxidative phosphorylation (OXPHOS) under aerobic conditions, cancer cells favor glycolytic pathway (10,11). *GPI* gene expression is induced by transcription factors c-Myc and HIF-1 (12,13), and has been shown to be overexpressed in many types of cancer (14). GPI has been proposed to be the autocrine motility factor (AMF), a secretory protein, which may act as a cytokine (15). Though complete *GPI* knockdown (*GPI*-KO) failed to prevent tumor growth *in vivo*, as *GPI*-KO cells can still grow by reprogramming their bioenergetics metabolism to OXPHOS, cell growth was shown to slow down and turn extremely sensitive to inhibitors of the respiratory chain complex (16). This dramatic metabolic plasticity was also observed in other cancer cell lines (14).

Here, we performed gene expression profiling of GC tumor and paired adjacent nontumor tissues from 198 GC patients. A genome-scale screening of GC prognostic genes was implemented via Fisher's discriminant analysis (FDA) based phenotype-targeted test (FDA-based PT-test) and integrative hypothesis testing (IHT) analysis, resulting in 27 potential prognostic genes, which were highly correlated with cancer metabolism. *GPI*, as the top ranked gene, was specifically evaluated for its ability as a prognostic marker. Its relationships with metabolism, EMT, and drug sensitivity were revealed. Particularly, results from Enviro-geno-pheno-state (E-GPS) analysis (17) indicated that *GPI* expression effectively stratified the outcomes of *CDH2*-negative patients with tumor node metastasis (TNM) stage II/III. Collectively, our study suggests *GPI* as a promising biomarker for GC prognosis, and the analytical frameworks used in this study can provide a useful tool for cancer studies.

## Materials and methods

### *Gene expression profile and clinical data from Peking University Cancer Hospital*

A total of 198 patients with GC included in this study were surgically treated at Peking University Cancer Hospital between 2007 and 2010, and were followed up to March 2016. This investigation was performed after approval by the Ethics Committee of Peking University Cancer Hospital. General informed consent was obtained from each patient. After radical gastrectomy, resected specimens were processed routinely for microscopic pathological assessment, and tissues were sampled and snap-frozen in liquid nitrogen. Fresh human tissues were stored at −80 °C. To ensure the quality of tissues, routine histological evaluation was performed for each sample. The gene expression profile of these tumors and paired noncancerous tissues were performed using the Agilent human mRNA & lncRNA Array V4.0 platform. All the 198 microarrays passed the quality control and were thus processed with quantile normalization and log−2 transformation. We further performed the prognostic biomarker study based on these normalized expression values of the 20,205 mRNAs.

GC stage was classified according to the 2010 TNM classification recommended by the American Joint Committee on Cancer (AJCC 7th edition). T and N classification were assessed based on the final pathological results and M classification was determined by surgical findings. Early GC (EGC) was defined as a tumor that was confined to the mucosa or submucosa regardless of lymph node (LN) involvement. Advanced GC (AGC) was defined as a tumor that invaded the muscle proper or beyond. OS was calculated from the date of the initial surgery to the time of decease caused by the tumor or the date of the last follow-up. Progression-free survival (PFS) was calculated from the date of the initial surgery to the time of GC progression. None of the patients received chemotherapy or radiation therapy prior to surgery. A summary of clinical information is shown in *Supplementary Table S1*. The gene expression data together with clinical information of the 198 GC patients constituted our own dataset, which was named after the Peking University Health Science Center (PUHSC).

### *Screening of prognostic genes from tumor and paired nontumor expression profiles*

Most existing methods for tumor-nontumor paired data analysis treat the gene expression of adjacent nontumor tissues as normal backgrounds that vary between individuals. Thus, traditional analytical frameworks for biomarker screening use fold-changes of the tumor over nontumor to measure the relative expression of targeted genes (18). Consequently, only those genes with significant difference in expression values between tumor and nontumor tissues, i.e. differentially expressed genes, are screened out for further prognostic analysis. This has been limited in several respects. First, as the resection margin, adjacent nontumor tissues differ from the pure normal tissues, hence the relative difference in gene expression between tumor and nontumor may not truly reflect the real "somatic" aberrations. Second, the expression pattern of transcripts in tumor tissues is dynamic. That is to say, genes are selectively expressed in specific time and space. Therefore, only considering the differentially expressed genes may have failed to capture important information associated with carcinogens *in vivo*. Here, we performed PT-test to discover genes that differentiate good or poor survival outcomes using FDA. Furthermore, based on the results of FDA-based PT-test for each gene, an IHT was performed to select the prognostic biomarkers of GC.

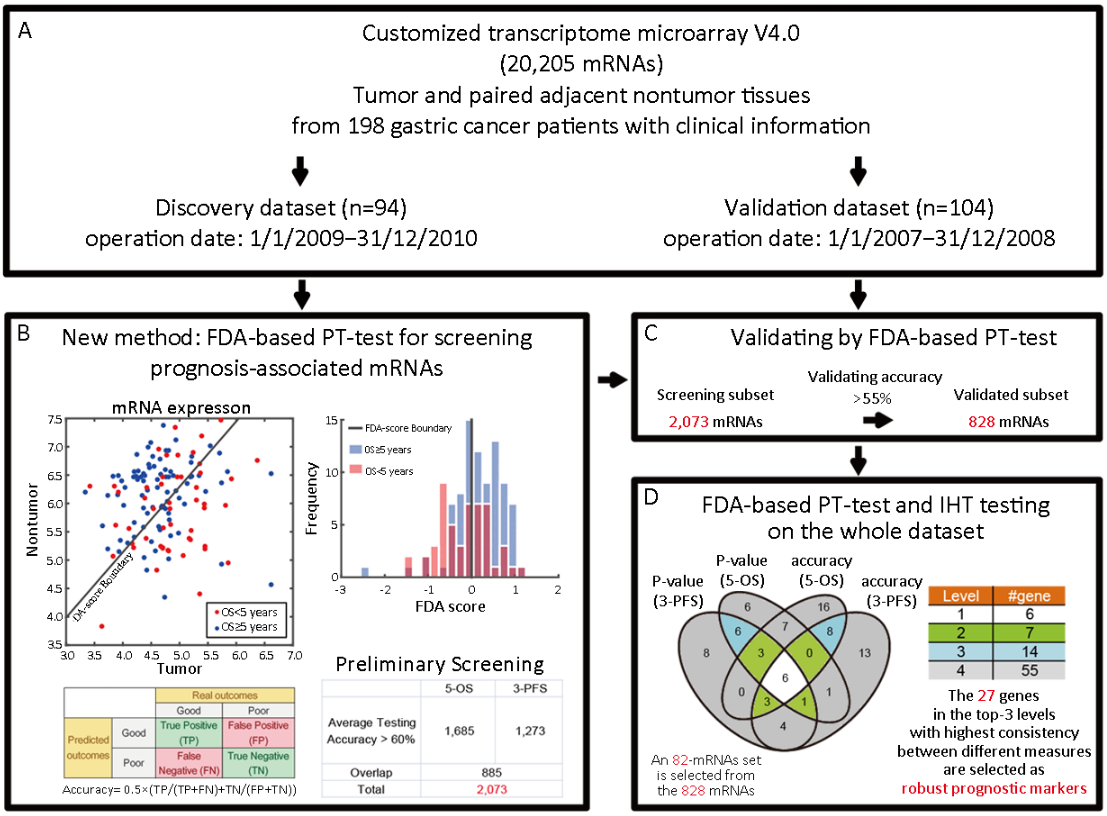
The preliminary screening of prognostic genes was implemented based on tumor-nontumor paired GC profiles of 198 Chinese patients. Details on patients and samples, RNA extraction and microarray processing were expanded in *Supplementary materials*.

### *FDA-based PT-test*

For each gene, the expression value in tumor and nontumor tissue of each patient was treated as a 2D sample. Each sample point was labeled as "good" or "poor" according to the outcome of patients. Based on Fisher's linear discriminant, all samples were projected onto a unit-length normal vector of the classification boundary, resulting in a set of 1D values. These values were adjusted to FDA-scores by subtracting the average of the mean centers of two classes. Based on the FDA-score, a weighted combination of the gene expression in tumor and paired nontumor, we obtained the P-value of two-sample *t*-test, and the predictive classification accuracy of outcomes. To obtain stable classification results, we calculated the average testing accuracy of FDA-based PT-test for each mRNA by 100 Monte-Carlo cross-validations, with 70% samples for training and 30% for testing. The average testing accuracy was calculated by the mean value of all the 100 testing accuracies. For the classification accuracy, balanced accuracy (definition refers to the contingency table in *Figure 1B*) was selected to avoid the bias of imbalanced datasets.

### *Integrative hypothesis test*

For a model-based test (e.g., Student's *t*-test), the P-value measures the difference between the models (e.g., means) of two populations, while the classification rate or accuracy measures how well the discriminant boundary separates the two populations. Since these two measures are complementary to each other, it is important to coordinate them with an integrated approach. Moreover, though it is known that the 5-year OS is a gold standard as an endpoint in clinical trials (19), 3-year PFS, as a potential surrogate of 5-year OS, has also received much attention recently (20). Hence we adopted the IHT (21) that incorporates both 5-year OS and 3-year PFS to evaluate each gene based on both P-value and classification accuracy. We used the FDA-based PT-test to select candidate mRNAs into a twin-set with one comprising of top-50 mRNAs with smallest Student's P-values and the other comprising of



| | | Real outcomes | |
|---|---|---|---|
| | | Good | Poor |
| Predicted outcomes | Good | True Positive (TP) | False Positive (FP) |
| | Poor | False Negative (FN) | True Negative (TN) |

Accuracy= 0.5×(TP/(TP+FN)+TN/(FP+TN))

| | 5-OS | 3-PFS |
|---|---|---|
| Average Testing Accuracy > 60% | 1,685 | 1,273 |
| Overlap | 885 | |
| Total | 2,073 | |

| Level | #gene |
|---|---|
| 1 | 6 |
| 2 | 7 |
| 3 | 14 |
| 4 | 55 |

**Figure 1** A flowchart of prognostic markers screening. (A) Introduction of the dataset based on tumor and paired adjacent nontumor tissues from 198 patients; (B) Preliminary screening by Fisher's discriminant analysis (FDA) based phenotype-targeted test (FDA-based PT-test) for each transcript on the discovery dataset. The scatter plot (left panel) represents the expression of gene X in tumor tissues (X-axis) and paired nontumor tissues (Y-axis). The FDA-scores in the histogram (right panel) are generated by projecting 2D expression into 1D value via FDA; (C) Validation of the genes from preliminary screening; (D) A subset of 828 validated genes is further analyzed by FDA-based PT-test and integrative hypothesis testing (IHT). Top-82 markers were located in different levels of Venn diagram, in which six genes were in the first level (white), seven in the second level (green) and fourteen in the third level (blue). The table on the right shows the corresponding numbers of mRNAs in each level of the hierarchy. OS, overall survival; PFS, progression-free survival.

top-50 mRNAs with highest classification accuracies. We obtained one such twin-set of mRNAs for 3-year PFS and another one for 5-year OS. Finally, the consensus of the four subsets, which was obtained by Venn Diagram as illustrated in *Figure 1D*, produced a hierarchy of candidate genes as biomarkers. Genes on the higher level received bigger consensus than those within lower levels and thus were more robust.

### *E-GPS approach*

We used a recently proposed E-GPS approach (17) to identify the role of a biomarker in a certain condition or environment. The E-GPS method is performed in the joint domain $D_{eg\varphi}$ of enviro-measure *e* (e.g., clinicopathological features such as TNM staging and Lauren classification, or well-known biomarkers proposed by previous studies), geno-measure *g* (e.g., gene features such as mutation and expression of targeted biomarkers), and pheno-measure $\varphi$ (e.g., occurrence of disease-related events such as metastasis and death of patients). In such a joint domain, each sample is represented by a triple-measured element ($\boldsymbol{e}$, $\boldsymbol{g}$, $\boldsymbol{\varphi}$). A collection of adjacent samples may share a common system status namely "state" *s*. In prognostic studies, three types of states are defined and can be learned from given sample sets: 1) g-state, short for "good" state, wherein an element has a high enough probability of good outcome; 2) p-state, short for "poor" state, wherein an element has a high enough probability of poor outcome; and 3) c-state, short for "confusing" state, wherein patients with good outcomes are considerably mixed with those with poor outcomes.

In the current study, in order to investigate the prognostic feature of *GPI*, we incorporated this newly discovered biomarker with previously well-studied biomarker, *CDH2*, and assigned the former as geno-measure and the latter as enviro-measure. Therefore, the E-GPS analysis was performed based on the joint expression of *GPI* and *CDH2*. Without loss of generality, we considered the support vector machine (SVM)-based model to classify patients into different states, as illustrated in *Supplementary materials*.

## Results

### *Identification of candidate prognostic biomarkers of GC*

In order to identify prognostic genes of GC, we performed FDA-based PT-test and IHT. As shown in *Figure 1*,*2*, we discovered 27 genes with top performance in either P-value or classification accuracy for both 5-year OS and 3-year PFS. We classified these genes as robust prognostic genes (*Supplementary Table S2*). As shown in *Figure 2B*, the FDA-based PT-test efficiently distinguished outcomes by using information from both tumor and nontumor. Our screening results showed that mRNA expressions of six genes including *GPI*, *PLA2G2A*, *COASY*, *FXN*, *EIF3B* and *IGFBP2* were located at the highest level, and their four indices all ranked as top-50 on a whole-genome scale. The Kaplan-Meier (KM) plots (*Figure 2C* & *Supplementary Figure S1*) showed that the outcome of patients could be well distinguished by these six genes.

We found these candidate genes have close relationships with metabolism. For example, *GPI* is a glucose metabolic gene involved in an early step of glycolysis. *PLA2G2A* and *IGFBP2* are involved in lipid metabolism and have been already reported as prognostic markers for GC (22-24). In order to obtain an overview of the biological features of the 27 prognostic markers, we performed enrichment analysis using the gene set enrichment analysis (GSEA), Kyoto Encyclopedia of Genes and Genomes (KEGG) and Gene Ontology (GO) enrichment. As shown in *Supplementary Table S3*, all 27 genes were closely linked to processes such as cell cycle, AMPK, p53 signaling pathway, or metabolism. GO enrichments based on tumor and nontumor tissues showed that 18 of the 27 genes obtained more enrichments with metabolic gene sets in nontumor tissues compared to tumor tissues, which was consistent with results obtained previously (5).

Considering that these genes are related to metabolic processes, we next analyzed the enrichment patterns of different metabolic pathways. Based on the GSEA toward metabolic gene signatures collected by a previous study (9), we found that our targeted genes were significantly enriched in processes related to OXPHOS, fatty acid oxidation (FAO), creatine metabolism (CM) and so on, in both tumor and nontumor tissues (*Figure 2D*).

### *Stratifying role of GPI for prognosis of CDH2-negative patients*

In order to investigate the prognostic features of *GPI*, we next performed E-GPS analysis to observe how the prognostic classification of patients can be improved by using *GPI* and the well-known GC biomarker, *CDH2*, an important signature of the occurrence of EMT process (25).

As shown in *Figure 3A*, patients with higher expression of *CDH2* had poorer survival compared with that of the *CDH2*-low patients. The separation by the 90th percentile of *CDH2* expression yielded a prognostic differentiation between the two subgroups of patients (*Figure 3B*, log-rank P=7.4e−02). However, the outcome of patients with low *CDH2* expression was still confused and thus needed a better subdividing (*Figure 3A*). To this end, we next performed SVM-based E-GPS analysis based on mRNA expression levels of both *GPI* and *CDH2*. As a result, the patients were classified into three states (*Figure 3C*), allowing us to stratify the outcome of patients significantly (*Figure 3D*,*E*). Moreover, outcomes of patients with high-*CDH2* or low-*CDH2* expression were further stratified into different levels by E-GPS analysis, as shown in *Figure 3F*,*G*.

We further found that TNM stages provided complementary information to differentiate the prognosis of patients more precisely to certain E-GPS state. For example, TNM stages differentiated the mixed outcomes within $S_3$, generating more differentiable subgroups (*Figure 3H*), allowing us to construct a molecular-diagnostic hybrid classification tree for GC prognosis (*Figure 3I*).

As a counterpart of the stratification of $S_1$, $S_2$, and $S_3$ by TNM stage (*Supplementary Figure S2A−C*), we also investigated how well the E-GPS marker differentiated the outcome of patients with each TNM stage (*Supplementary Figure S2D−G*). Comparing TNM I and IV, the prognosis prediction of TNM II and III were significantly improved.

Next, we performed analyses focusing on the *CDH2*-low patients with stage II/III, due to two main reasons: first, the

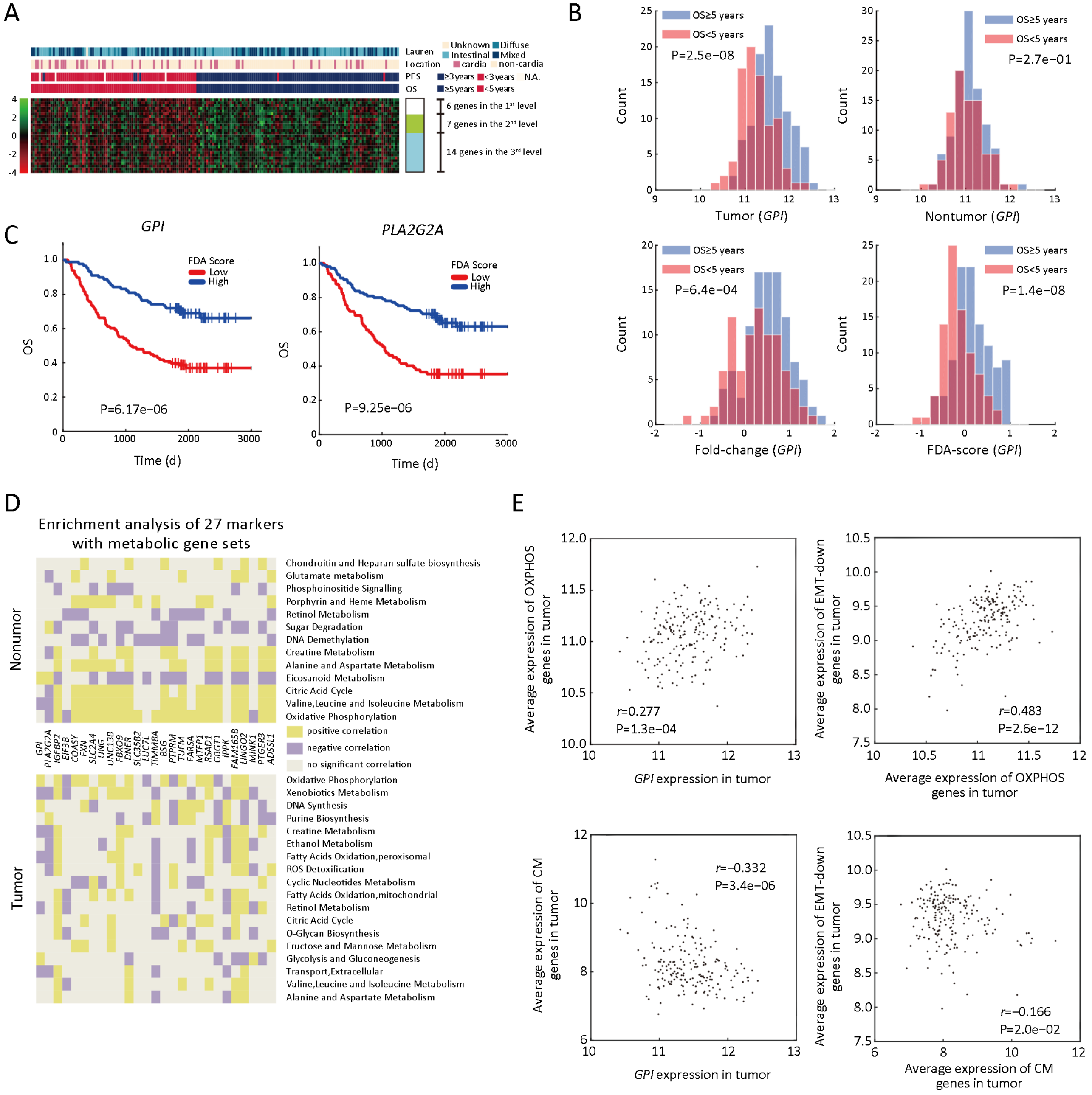


**Figure 2** Prognostic features and metabolic enrichment analyses of 27 gene markers. (A) Heat map of Fisher's discriminant analysis (FDA) scores of 27 genes; (B) Histograms of different 1D-representation of glucose-6-phosphate isomerase (*GPI*), including expression in tumor (top left), nontumor (top right), relative expression of tumor *vs*. nontumor (bottom left), and FDA-score (bottom right) of *GPI* expression in tumor and paired nontumor. P-values are generated by *t*-test between good- [overall survival (OS) ≥5 years.] and poor-surivival (OS <5 years.) groups (P); (C) Kaplan-Meier (KM) plots based on FDA-score of *GPI* (P=6.17e−06) and *PLA2G2A* (P=9.25e−06); (D) Summarizing of gene set enrichment analysis (GSEA) of 27 genes focusing on metabolism pathways in tumor (lower part) and nontumor (upper part); (E) 2D scatter plot of the expression composed by *GPI* & oxidative phosphorylation (OXPHOS) (top left) (*r*=0.277, P=1.3e−04), *GPI* & creatine metabolism (CM) (bottom left) (*r*=−0.332, P=3.4e−06), OXPHOS & "EMT-down genes" (top right) (*r*=0.483, P=2.6e−12), and CM & "EMT-down genes" (bottom right) (*r*=−0.166, P=2.0e−02). EMT-down genes, genes down-regulated in EMT process.

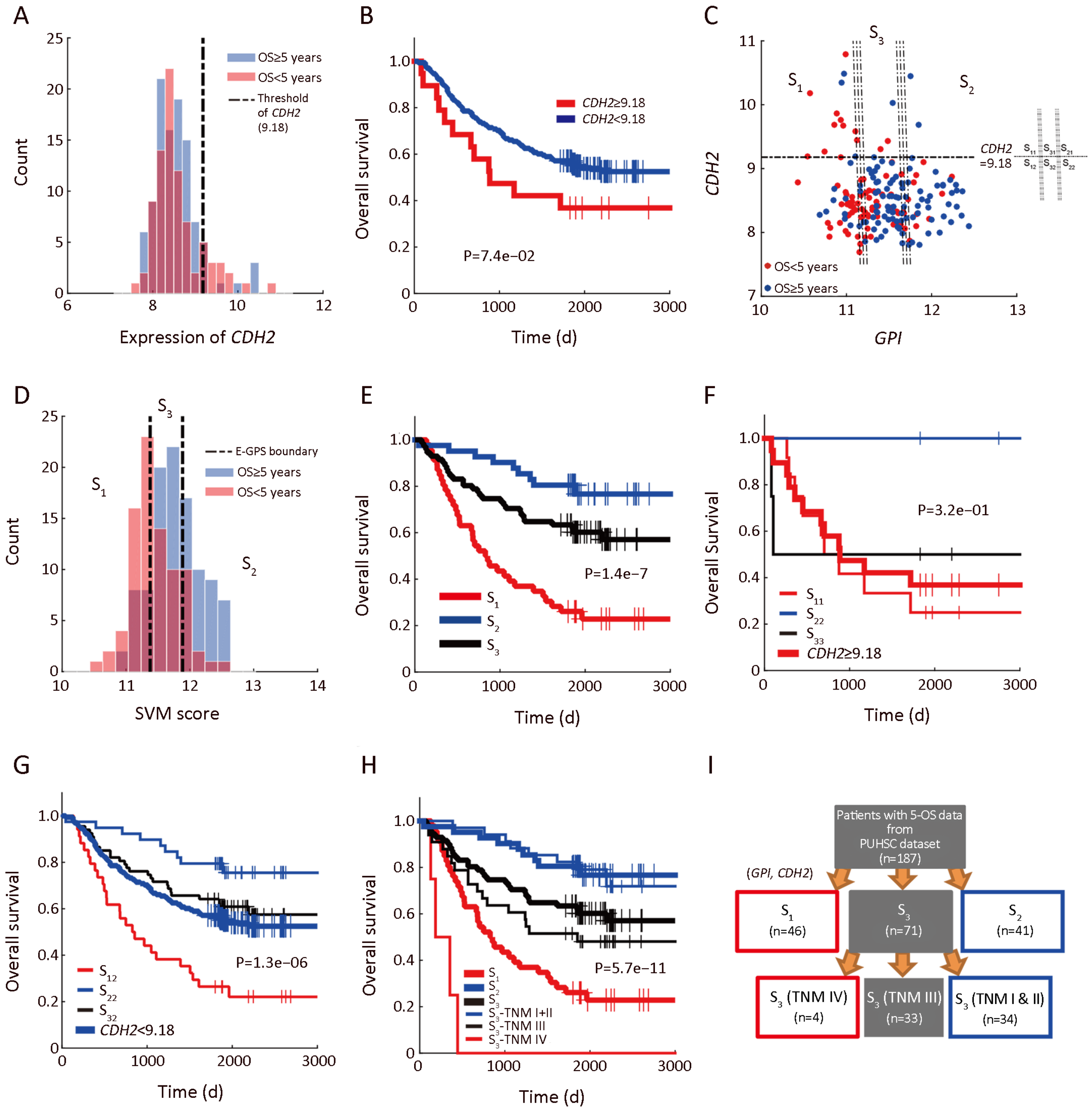


**Figure 3** Enviro-geno-pheno-state (E-GPS) analysis of *GPI*, *CDH2*. (A) Histogram of *CDH2* expression in tumor tissues; (B) Kaplan-Meier (KM) plots of *CDH2*-high (≥9.18) and *CDH2*-low (<9.18) group, showing that the expression of *CDH2* alone could not distinguish the prognosis of patients (P=7.4e−02); (C) E-GPS states defined by *GPI*, *CDH2*. Patients are stratified into three groups according to the joint expression of the two genes. Each group corresponds to one of the E-GPS states $S_1$, $S_2$, and $S_3$. Patients of *CDH2*-high group are separated into $S_{11}$, $S_{21}$, and $S_{31}$ by *GPI* expression level. Similarly, patients of *CDH2*-low group are separated into $S_{12}$, $S_{22}$ and $S_{32}$; (D) Histogram of SVM-scores and the corresponding states $S_1$, $S_2$ and $S_3$, generated by E-GPS analysis of *GPI*, *CDH2*; (E) KM plot of patients grouped by $S_1$, $S_2$ and $S_3$ states (P=1.4e−7); (F) KM plot of *CDH2*-high patients grouped by $S_{11}$, $S_{21}$ and $S_{31}$ substates (P=3.2e−01); (G) KM plot of *CDH2*-low patients grouped by $S_{12}$, $S_{22}$ and $S_{32}$ substates (P=1.3e−06); (H) KM plot of patients grouped by $S_1$, $S_2$ and $S_3$ states with the help of TNM staging (P=5.7e−11); (I) Molecular-diagnostic hybrid tree obtained by E-GPS analysis. Patients in the red boxes have relatively poor survival; patients in the blue boxes have good survival; patients in the gray blocks have confused survival. GPI, glucose-6-phosphate isomerase; OS, overall survival; SVM, support vector machine; PUHSC, Peking University Health Science Center.

sample size of this part of patients (144 of 198) is larger, ensuring a higher reliability of E-GPS analysis; second, the outcomes of them were more confused comparing with the rest of the patients. Using univariate E-GPS analysis based on *GPI* expression, *CDH2*-low patients with stage II/III were classified into three groups with significantly different levels of outcome (*Figure 4A*,*B*), leading to a reconstruction of molecular-diagnostic hybrid classification tree (*Figure 4C*). To validate this prognostic classification, we performed a combined analysis using independent data from public source (2, 26, 27). We extracted *CDH2*-lower patients with TNM II and III stages, and performed univariate E-GPS analyses using *GPI* based on Singaporean dataset (*Figure 4D*), Korean dataset (*Figure 4E*) and The Cancer Genome Atlas (TCGA) mRNA-sequencing data of GC (*Figure 4F*), respectively. The results from all the three validation datasets were consistent with that of our own dataset.

In addition, an immunohistochemistry (IHC) experiment for GPI was performed on 50 paired GC patients, which were randomly selected from the 198 patients. Among the 50 patients, 32 were *CDH2*-lower and with TNM II/III stages. The prognostic effectiveness of GPI was also observed in protein level (*Supplementary Figure S3*), however, further research based on larger sample size is required. Moreover, we checked GC cell line data (n=37) from Cancer Cell Line Encyclopedia (CCLE, https://portals.broadinstitute.org/ccle), and found marginally significant differential expression of *GPI* in the mRNA level (P=0.06) between the group of *CDH2*-high (mRNA expression of *CDH2*>0) and *CDH2*-low (mRNA expression of *CDH2*<0) cell lines, which is consistent with the result shown by the scatter plot of *Figure 3C*.

### *Decreased GPI expression in GC tissues is an indicator of rewiring of metabolism associated with poor outcome of patients*

In order to investigate the features of the mRNA expression of *GPI* as a metabolic indicator, we examined the relationship of *GPI* and several pathways of metabolism processes. GC patients exhibiting higher expression levels of OXPHOS in tumor tissues also displayed increased expression levels of glycolysis and *GPI*, which coordinated with the enhanced biosynthesis of glycogen (*r*=0.35, P=5.7e−06), fatty acid (*r*=0.32, P=6.2e−06) and DNA (*r*=0.48, P=4.4e−12). In addition, our results also showed that the mRNA expression level of *GPI* was negatively correlated with several catabolic pathways, such as mitochondrial FAO (*r*=−0.11, P=1.2e−01) and peroxisomal FAO (*r*=−0.19, P =1.1e−02).

Furthermore, we focused on the relationship between *GPI*, its related metabolic pathways, and EMT process. We selected a group of genes that are down-regulated in EMT (shortly DR in EMT) as epithelial markers, similar to those in a previous study (3). As shown in *Figure 2E*, OXPHOS were positively correlated with *GPI* as well as the epithelial markers. Similar correlation was also found between fatty acid synthesis (FAS) and *GPI* (*r*=0.32, P=6.2e−06) in tumor tissues. On the other hand, the positive correlation between glycogen degradation and the up-regulated genes in EMT (shortly UR in EMT) was detected (*r*=0.65, P=1.7e−23). These results suggested that tumors exhibiting up-regulated glycogen degradation as well as down-regulated glucose metabolism and lower FAS levels are compromised in cell growth; however, such tumor cells are potentially undergoing metastasis in a EMT-like manner. In addition, we found that the activation of CM was negatively correlated with the expression levels of DR genes (*Figure 2E*).

Furthermore, we found that mRNA expression levels of OXPHOS and FAS in tumor tissues were correlated with prolonged survival, whereas FAO and CM were associated with poor prognosis of GC patients. However, we further observed that the prognostic feature of the single pathways was heterogenic in different datasets, suggesting that we need an integrated signal to combine these pathways for prognosis. Hence, we first screened out the *GPI*-related metabolic pathways that were more stable for prognosis, and then combined these pathways to obtain an integrated signal. As a result, we chose OXPHOS, FAO, FAS, and CM according to their causal effects on prognosis. In most cases (≥75%), each of the four pathways maintained similar prognostic property across four different datasets (Singaporean, Korean, TCGA, and our dataset).

Next, we assigned OXPHOS and FAS as the positive signals for patients' survival, while FAO and CM as negative signals, based on their metabolic and prognostic features. The relative difference between the positive and negative signals (OXPHOS+FAS−FAO−CM) was used as an integrated signal, the expression of which was also positively correlated with *GPI* expression.

As shown in *Figure 5A*,*B*, we found that expression levels of the integrated signal correlated well with prognosis of GC patients. This result was further validated by the combined analysis on three independent public datasets

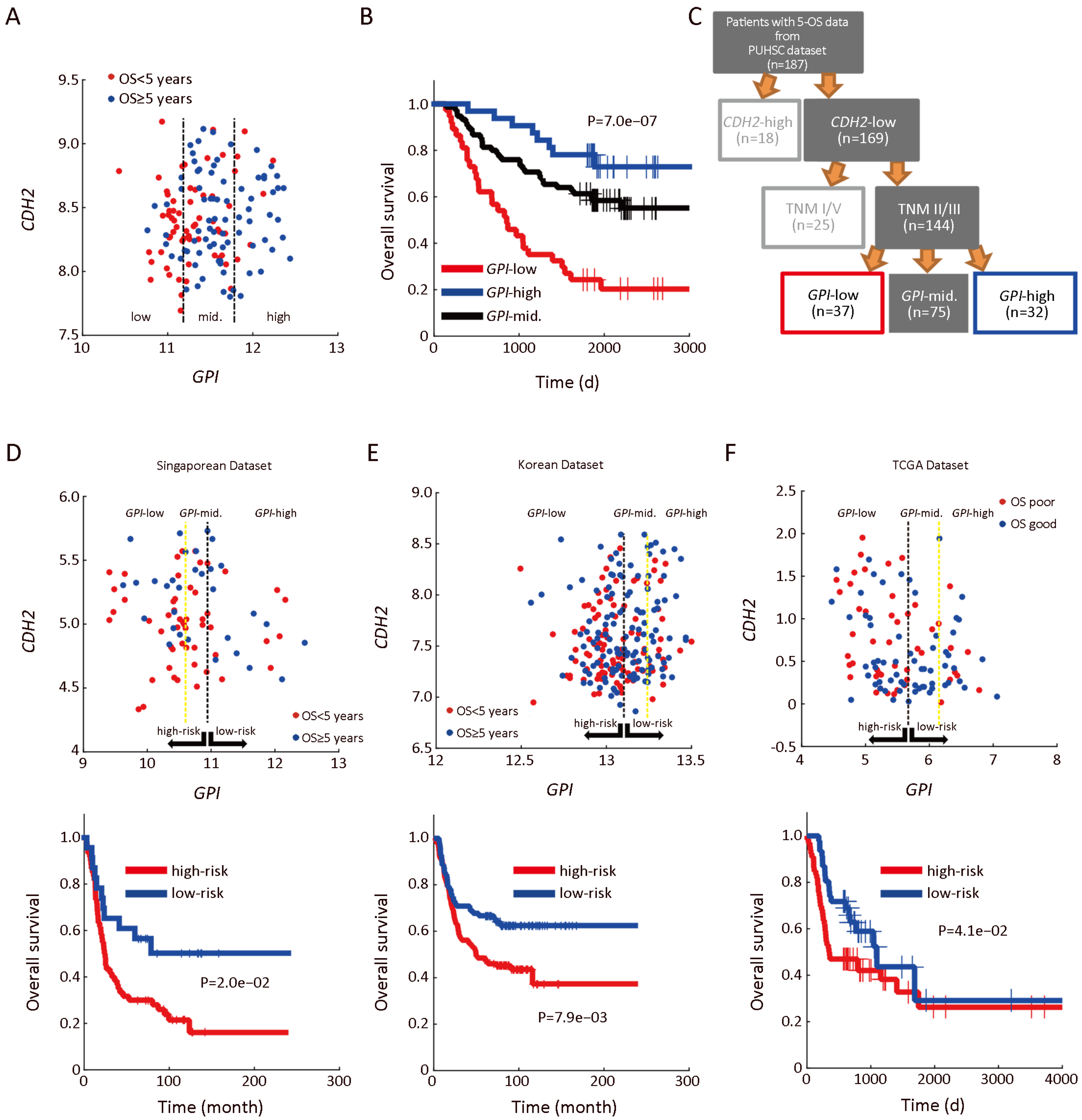


**Figure 4** Enviro-geno-pheno-state (E-GPS) analysis of *CDH2*-negative patients with TNM II/III stages based on *GPI*. (A) *GPI-CDH2* expression plot of *CDH2*-negative patients with TNM II/III stages. The two parallel lines in the scatter plots classify the *GPI* expression in tumor tissues into three levels, which are obtained from a univariate E-GPS analysis (*Supplementary materials*); (B) Kaplan-Meier (KM) plot of patients with each E-GPS state (P=7.0e−07); (C) An adjusted molecular-diagnostic hybrid classification tree. Patients in the red boxes have relatively poor survival; patients in the blue boxes have good survival; patients in the gray blocks have confused survival; (D−F) *GPI-CDH2* expression plot (upper) and KM plot (lower) of *CDH2*-negative patients with TNM II/III stages from the Singaporean dataset (P=2.0e−02), Korean dataset (P=7.9e−03) and TCGA dataset (P=4.1e−02), respectively. GPI, glucose-6-phosphate isomerase; PUHSC, Peking University Health Science Center.

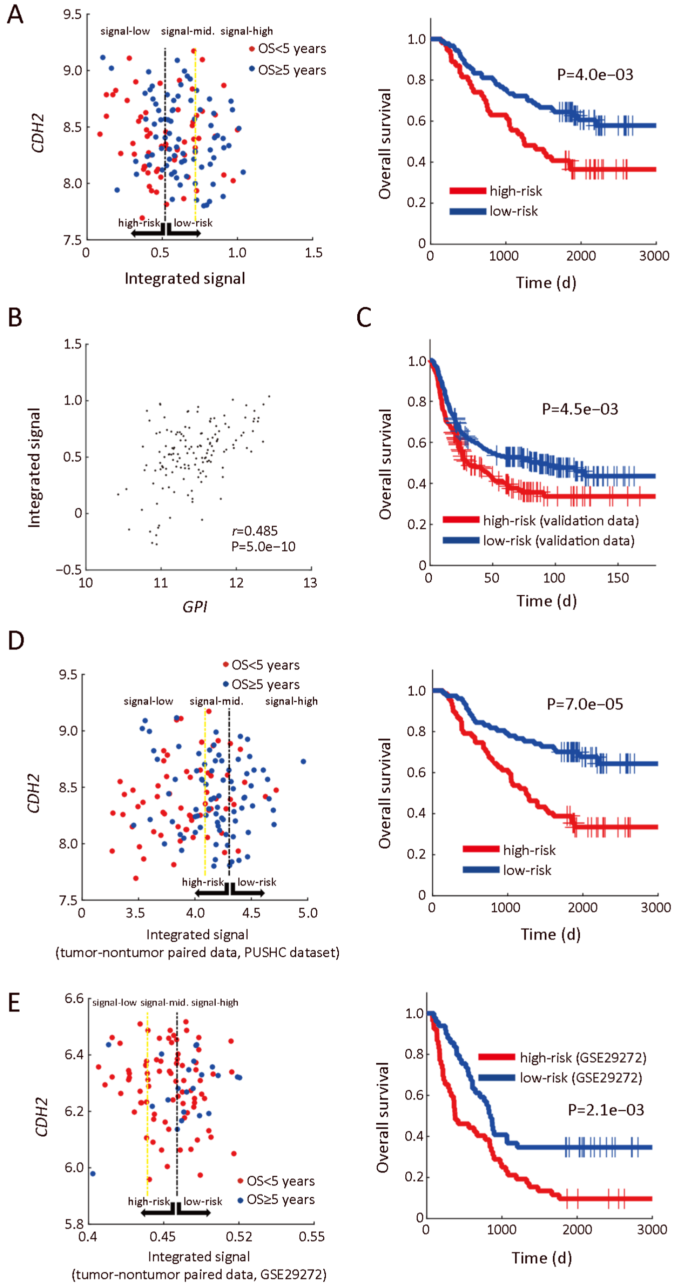


**Figure 5** Enviro-geno-pheno-state (E-GPS) analysis of *CDH2*-negative patients with TNM II/III based on integrated metabolic signals. (A) Joint expression plot (left) of the integrated signal and *CDH2*, and Kaplan-Meier (KM) plot (right) (P=4.0e−03) of *CDH2*-negative patients with TNM II/III stages. E-GPS states are defined by the integrated metabolic signal, which is represented by (OXPHOS+FAS−FAO−CM), based on the gene expression in tumor tissues from our own dataset; (B) Joint expression plot of *GPI* and the integrated signal (*r*=0.485, P=5.0e−10); (C) KM plot of *CDH2*-negative patients with TNM II/III stages, based on the combination of samples from three independent datasets (Korean, Singaporean and TCGA) (P=4.5e−03); (D) Joint expression plot (left) of the integrated signal (based on tumor and paired nontumor tissues) and *CDH2* (based on tumor tissues only), and KM plot (right) of *CDH2*-negative patients with TNM II/III stages (P=7.0e−05). E-GPS states are defined by the integrated metabolic signal, which is represented by [($OXPHOS_{tumor}$+ $OXPHOS_{nontumor}$)/2+($FAS_{tumor}$−$FAO_{tumor}$+$FAS_{nontumor}$−$FAO_{nontumor}$)/2−($CM_{tumor}$−$CM_{nontumor}$)], based on the gene expression in tumor and paired nontumor tissues from our own dataset; (E) Joint expression plot (left) and KM plot (right) based on the dataset GSE29272 (P=2.1e−03). GPI, glucose-6-phosphate isomerase; PUHSC, Peking University Health Science Center.

with E-GPS method (*Figure 5C*).

The relationship between metabolic pathways and prognosis was also analyzed in nontumor tissues. The higher level of OXPHOS, as well as the relative difference between FAS and FAO in nontumor tissue were related to prolonged survival. Higher levels of CM in nontumor also exhibited positive correlation with good survival outcomes, which is opposite to the case of CM in tumor tissues. Hence, we further integrated these pathways using the formula of $(OXPHOS_{tumor}+OXPHOS_{nontumor})/2 + (FAS_{tumor}-FAO_{tumor}+FAS_{nontumor}-FAO_{nontumor})/2 - (CM_{tumor}-CM_{nontumor})$, in which the molecular information generated by corresponding nontumor tissue was included. The integrated signal was analyzed by E-GPS analysis in both our own dataset and GSE29272 (28). As shown in *Figure 5D,E*, the integrated signal still could separate outcome of patient with GC effectively.

### *Pharmacogenomic analyses showed GPI indicating drug sensitivity*

We performed pharmacogenomic analysis based on the public data obtained from a recent large-scale study (29), which claimed that all 27 GC cell lines as drug-resistant against more than 900 drugs tested. Here, we compared the drug sensitivity of different cell lines using the half maximal inhibitory concentration ($IC_{50}$) (drug concentration that reduces viability by 50%) value. We examined 217 drugs that were previously tested on more than 70% of the cell lines, and found that the $IC_{50}$ values of 150 drugs were negatively correlated with mRNA expression levels of *GPI* in cell lines, suggesting that GC cells with higher expression of *GPI* might be more sensitive to certain drugs used during chemotherapy.

For example, 5-fluorouracil (5-FU), recommended as one of the first-line agents for the treatment toward advanced GC (30), was tested on 22 GC cell lines. As shown in *Figure 6A*, the linear regression revealed that *GPI* expression and $IC_{50}$ of 5-FU were negatively correlated. Moreover, we found that *GPI* mRNA expression significantly differed between $IC_{50}$-high (≥4) group and $IC_{50}$-low (<4) group, suggesting that cell lines with lower *GPI* expression exhibit higher drug resistance (*Figure 6B*). In addition, we checked the two first-line drugs commonly used for GC patients, Cisplatin and Docetaxel, and found little correlation between *GPI* and $IC_{50}$ of those drugs ($r$=0.08, P=7.1e−02 for Cisplatin and $r$=−0.11, P=6.1e−02 for Docetaxel).

## Discussion

In this study, we identified the mRNA expression of *GPI* as a promising indicator for GC prognosis. The close relationships between *GPI* and EMT as well as metabolic processes were revealed by computational analysis. We also proposed an integrative signal to combine the levels of activeness of four metabolic pathways including OXPHOS, FAS, FAO and CM, which were significantly correlated with *GPI*.

We found that lower level of OXPHOS in tumor tissues is positively correlated with poorer survival outcome, which is not only consistent with the previous study (13), but also validated by independent cohorts from public database. It is now widely accepted that the downregulation of OXPHOS in cancer cells can be explained by the hypothesis of mitochondrial dysfunction under metabolic stress,

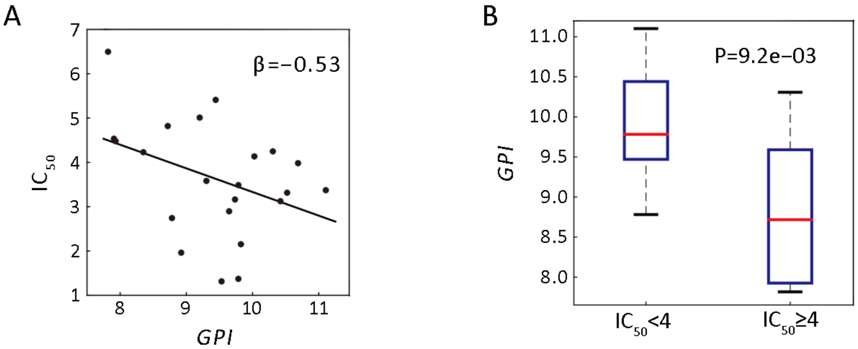


**Figure 6** Correlation analysis of *GPI* expression and 5-fluorouracil (5-FU) drug sensitivity. (A) Scatter plot shows the negative relationship between half maximal inhibitory concentration ($IC_{50}$) and *GPI* expression of 22 GC cell lines. The β-value represents the coefficient of linear regression; (B) Box-plot of *GPI* expression of gastric cancer (GC) cell lines grouped by higher sensitivity ($IC_{50}$ <4, n=13) and lower sensitivity ($IC_{50}$ ≥4, n=9) (P=9.2e−03). The P-value is generated by Student's *t*-test. GPI, glucose-6-phosphate isomerase.

corresponding to the significantly enhanced glycolysis even under aerobic condition in cancer cells, also known as Warburg effect (31): the growth of tumor is initiated by irreversible damage to respiration and persists due to increased anaerobic metabolism. However, we found that the prognostic effect of glycolytic signal in GC is contradictory.

Notably, our study revealed that the expression of *GPI* is positively correlated not only with glycolysis, but also with OXPHOS expression, indicating that the competition between OXPHOS and glycolysis is not a contradictory phenomenon in cancer cells, but an adaptation to provide sufficient ATP for tumor cell survival (10). Furthermore, we found that the expression of both OXPHOS and glycolysis in tumor tissues are negatively correlated to EMT. A plausible explanation for our finding is provided by a previous study, where the loss of glucose uptake was observed in mammary epithelial cells that were detached from extracellular matrix (32).

We also found that in addition to glucose-related pathways, various types of metabolic pathways were also correlated to EMT. For instance, CM is positively correlated with EMT process. While we found a negative correlation with CM and *GPI* expression, implying a possible metabolic rewiring.

Recently, a study on extracellular metabolic energetic of colon cancer found that upregulation of CKB, a key factor involved in CM progression, facilitates energy storage in colon cancer cells, while also promotes their survival during intrahepatic hypoxia after liver metastasis (33), resulting in poor prognosis for colon cancer patients.

Fatty acid metabolism plays significant roles in proliferation and survival of cancer cells (34). The *de novo* FAS contributes to membrane synthesis for cell growth and proliferation. In contrast, FAO (also known as β-oxidation) provides extra ATPs for cancer cells, subsequently promotes cell survival on loss of attachment (LOA), thus acting as ATP rescuer when glucose uptake and catabolism are inhibited by LOA (35). Our study showed that patients with higher relative activation of FAS to FAO have prolonged survival. This phenomenon coincided with a previous study reporting that consumption of NADPH decreased together with the downregulation of FAS under the stress of energy (36). In contrast, the upregulation of FAO facilitated the generation of NADPH. As a consequence, cell death is inhibited, suggesting the switch from FAS to FAO in tumor tissues may promote the survival of cancer cell (36).

## Conclusions

Using new computational approaches, our study identified *GPI* as a promising biomarker for reliable prognosis of GC. We also anticipate that IHT, E-GPS analysis, and prognostic analysis by causal inference will serve as promising tools to discover suitable biomarkers.

## Acknowledgements

This study was supported by grants from the Ministry of Science and Technology of the People's Republic of China (No. SS2014AA020603), Beijing Municipal Administration of Hospitals Clinical Medicine Development of Special Funding Support (No. ZYLX201701), Beijing Municipal Science and Technology Commission (No. D1311 00005313010), the National Natural Science Foundation of China (No. 31520103905), the National High Technology Research and Development Program ("863" Program) of China (No. 2015AA020108), and the Zhi-Yuan chair professorship start-up grant WF220103010 from Shanghai Jiao Tong University.

## Footnote

*Conflicts of Interest*: The authors have no conflicts of interest to declare.

## References

1. Torre LA, Bray F, Siegel RL, et al. Global cancer statistics, 2012. CA Cancer J Clin 2015;65:87-108.
2. Lee J, Sohn I, Do IG, et al. Nanostring-based multigene assay to predict recurrence for gastric cancer patients after surgery. PLoS One 2014;9: e90133.
3. Cristescu R, Lee J, Nebozhyn M, et al. Molecular analysis of gastric cancer identifies subtypes associated with distinct clinical outcomes. Nat Med 2015; 21:449-56.
4. Lordick F, Janjigian YY. Clinical impact of tumour biology in the management of gastroesophageal cancer. Nat Rev Clin Oncol 2016;13:348-60.
5. Huang X, Stern DF, Zhao H. Transcriptional profiles from paired normal samples offer complementary information on cancer patient survival — evidence from TCGA pan-cancer data. Sci Rep 2016;6:20567.
6. Hanahan D, Weinberg RA. Hallmarks of cancer: the

next generation. Cell 2011;144:646-74.

7. Sun L, Suo C, Li ST, et al. Metabolic reprogramming for cancer cells and their microenvironment: Beyond the Warburg Effect. Biochim Biophys Acta Rev Cancer 2018;1870:51-66.

8. Martinez-Outschoorn UE, Peiris-Pagés M, Pestell RG, et al. Cancer metabolism: a therapeutic perspective. Nat Rev Clin Oncol 2017;14:11-31.

9. Gaude E, Frezza C. Tissue-specific and convergent metabolic transformation of cancer correlates with metastatic potential and patient survival. Nat Commun 2016;7:13041.

10. Ždralević M, Marchiq I, de Padua MMC, et al. Metabolic Plasiticy in Cancers — Distinct Role of Glycolytic enzymes GPi, LDHs or Membrane Transporters MCTs. Front Oncol 2017;7:313.

11. Porporato PE, Filigheddu N, Pedro JMB, et al. Mitochondrial metabolism and cancer. Cell Res 2018;28:265-80.

12. Kim JW, Zeller KI, Wang Y, et al. Evaluation of myc E-box phylogenetic footprints in glycolytic genes by chromatin immunoprecipitation assays. Mol Cell Biol 2004;24:5923-36.

13. Funasaka T, Yanagawa T, Hogan V, et al. Regulation of phosphoglucose isomerase/autocrine motility factor expression by hypoxia. FASEB J 2005;19:1422-30.

14. Pusapati RV, Daemen A, Wilson C, et al. mTORC1-dependent metabolic reprogramming underlies escape from glycolysis addiction in cancer cells. Cancer Cell 2016;29:548-62.

15. Liotta LA, Mandler R, Murano G, et al. Tumor cell autocrine motility factor. Proc Natl Acad Sci U S A 1986;83:3302-6.

16. de Padua MC, Delodi G, Vučetić M, et al. Disrupting glucose-6-phosphate isomerase fully suppresses the "Warburg effect" and activates OXPHOS with minimal impact on tumor growth except in hypoxia. Oncotarget 2017;8:87623-837.

17. Xu L. Enviro-geno-pheno state approach and state based biomarkers for differentiation, prognosis, subtypes, and staging. Applied Informatics 2016;3:4.

18. Li J, Chen Z, Tian L, et al. LncRNA profile study reveals a three-lncRNA signature associated with the survival of patients with oesophageal squamous cell carcinoma. Gut 2014;63:1700-10.

19. Driscoll JJ, Rixe O. Overall survival: still the gold standard: why overall survival remains the definitive end point in cancer clinical trials. Cancer J 2009; 15:401-5.

20. Burzykowski T, Buyse M, Piccart-Gebhart MJ, et al. Evaluation of tumor response, disease control, progression-free survival, and time to progression as potential surrogate end points in metastatic breast cancer. J Clin Oncol 2008;26:1987-92.

21. Xu L. A new multivariate test formulation: theory, implementation, and applications to genome-scale sequencing and expression. Applied Informatics 2016;3:1.

22. Leung SY, Chen X, Chu KM, et al. Phospholipase A2 group ⅡA expression in gastric adenocarcinoma is associated with prolonged survival and less frequent metastasis. Proc Natl Acad Sci U S A 2002;99: 16203-8.

23. Zhang C, Yu H, Xu H, et al. Expression of secreted phospholipase A2-Group ⅡA correlates with prognosis of gastric adenocarcinoma. Oncol Lett 2015;10:3050-8.

24. Hur H, Yu EJ, Ham IH, et al. Preoperative serum levels of insulin-like growth factor-binding protein 2 predict prognosis of gastric cancer patients. Oncotarget 2017;8:10994-1003.

25. Cisło M, Filip AA, Arnold Offerhaus GJ, et al. Distinct molecular subtypes of gastric cancer: from Laurén to molecular pathology. Oncotarget 2018; 9:19427-42.

26. Lei Z, Tan IB, Das K, et al. Identification of molecular subtypes of gastric cancer with different responses to PI3-kinase inhibitors and 5-fluorouracil. Gastroenterology 2013;145:554-65.

27. Cancer Genome Atlas Research Network. Comprehensive molecular characterization of gastric adenocarcinoma. Nature 2014;513:202-9.

28. Wang G, Hu N, Yang HH, et al. Comparison of global gene expression of gastric cardia and noncardia cancers from a high-risk population in China. PloS One 2013;8:e63826.

29. Iorio F, Knijnenburg TA, Vis DJ, et al. A landscape of pharmacogenomic interactions in cancer. Cell 2016;166:740-54.

30. Van Cutsem E, Moiseyenko VM, Tjulandin S, et al. Phase Ⅲ study of docetaxel and cisplatin plus fluorouracil compared with cisplatin and fluorouracil

as first-line therapy for advanced gastric cancer: a report of the V325 Study Group. J Clin Oncol 2006; 24:4991-7.

31. Zheng J. Energy metabolism of cancer: Glycolysis versus oxidative phosphorylation (Review). Oncol Lett 2012;4:1151-7.
32. Schafer ZT, Grassian AR, Song L, et al. Antioxidant and oncogene rescue of metabolic defects caused by loss of matrix attachment. Nature 2009;461:109-13.
33. Loo JM, Scherl A, Nguyen A, et al. Extracellular metabolic energetics can promote cancer progression. Cell 2015;160:393-406.
34. Kuo CY, Ann DK. When fats commit crimes: fatty acid metabolism, cancer stemness and therapeutic resistance. Cancer Commun (Lond) 2018;38:47.
35. Carracedo A, Cantley LC, Pandolfi PP. Cancer metabolism: fatty acid oxidation in the limelight. Nat Rev Cancer 2013;13:227-32.
36. Jeon SM, Chandel NS, Hay N. AMPK regulates NADPH homeostasis to promote tumour cell survival during energy stress. Nature 2012;485:661-5.

**Cite this article as:** Huang H, Wen X, Xue H, Chen R, Ji J, Xu L. Phosphoglucose isomerase gene expression as a prognostic biomarker of gastric cancer. Chin J Cancer Res 2019;31(5):771-784. doi: 10.21147/j.issn.1000-9604.2019.05.07

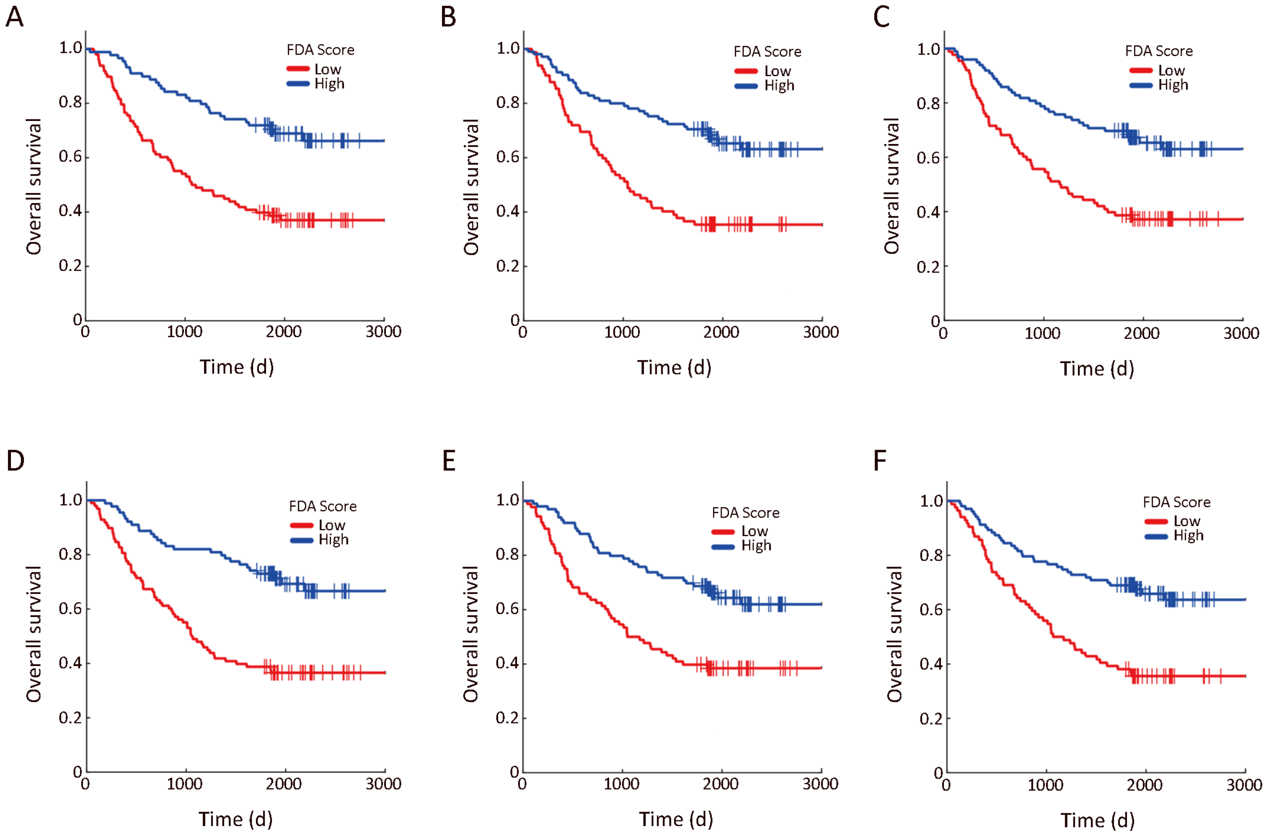


**Figure S1** Kaplan-Meier (KM) plots of six mRNAs in the first level. KM plots for overall survival of patients divided by Fisher's discriminant analysis (FDA) scores of mNRAs in the first level (see *Figure 1* in the main text). Blue curves represent the low-risk group with FDA scores higher than the median value; red curves represent the high-risk group with FDA scores lower than the median value. (A) GPI (P=6.17e−06); (B) PLA2G2A (P=9.25e−06); (C) IGFBP2 (P=4.41e−05); (D) EIF3B (P=1.66e−06); (E) COASY (P=6.76e−05); (F) FXN (P=2.15e−05).

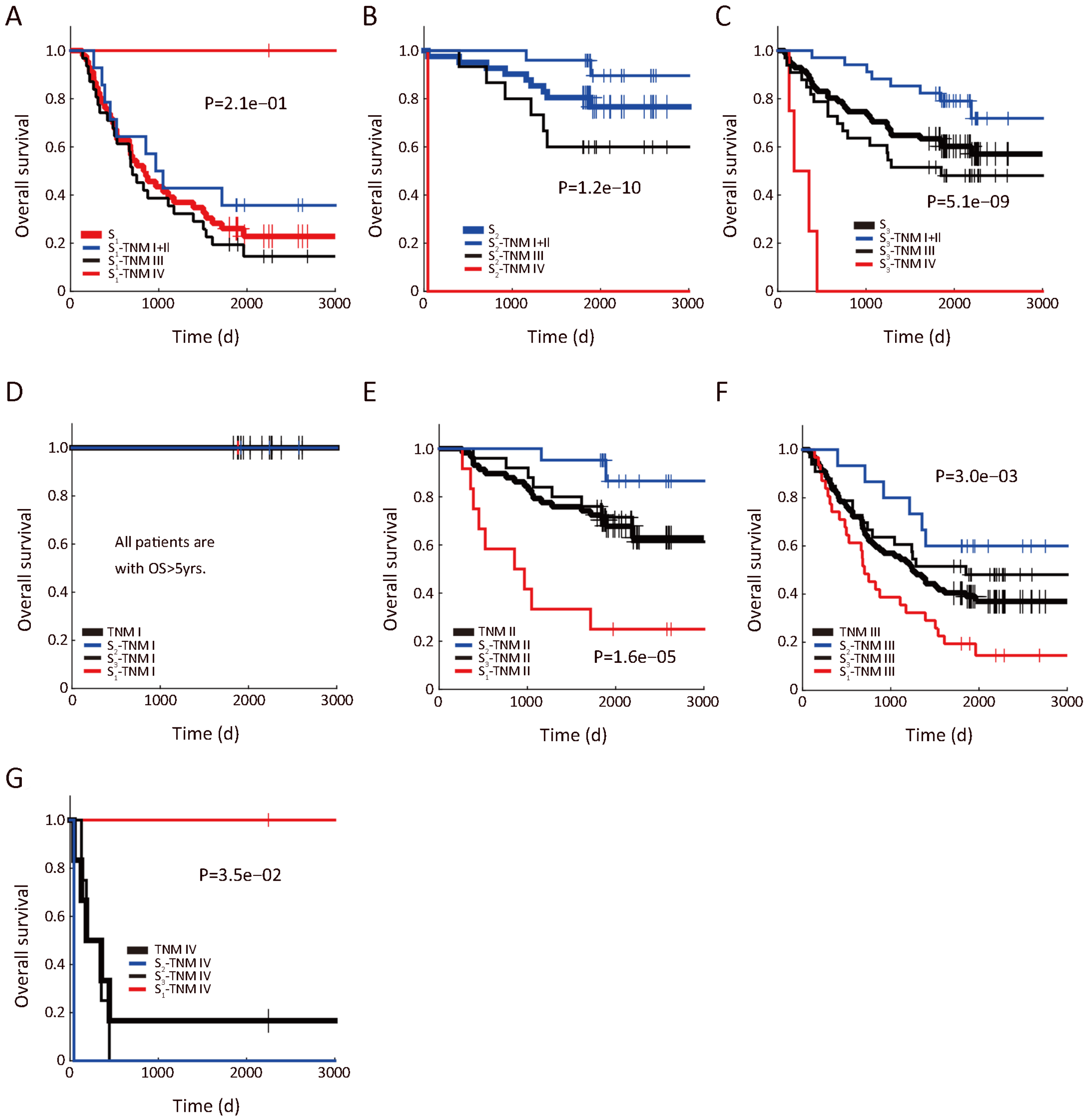


**Figure S2** Prognostic stratification coupling (*GPI*, *CDH2*) and TNM stage. (A−C) TNM stratifies patients dividing by *GPI*, *CDH2* into $S_1$ (A), $S_2$ (B) and $S_3$ (C), respectively; (D−G) $S_1$−$S_3$ stratify patients with TNM stage I (D), II (E), III (F) and IV (G), respectively. GPI, glucose-6-phosphate isomerase.

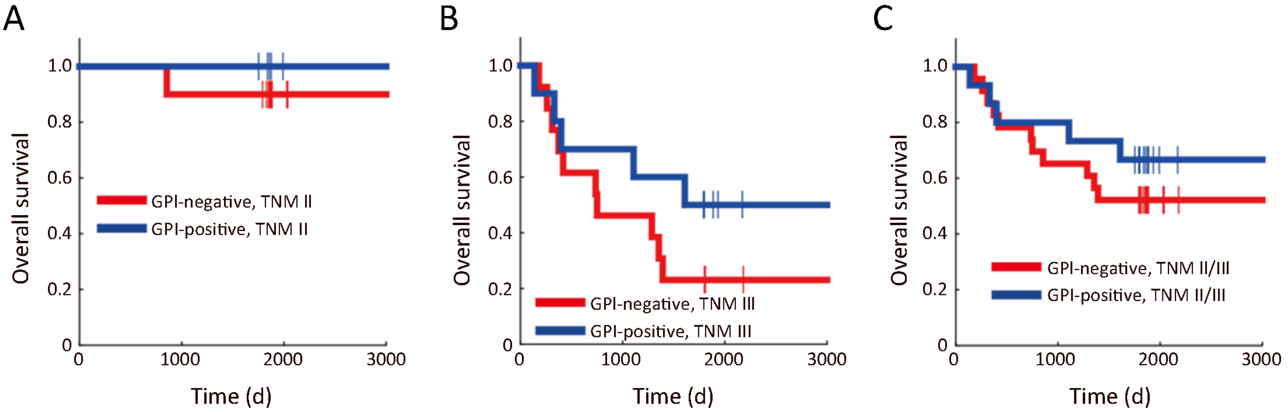
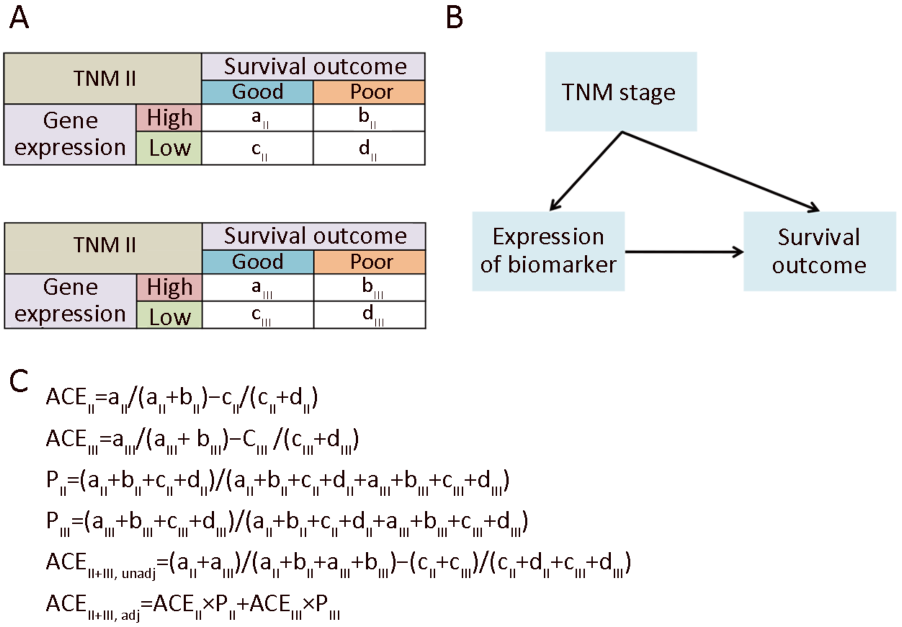


**Figure S3** Survival analysis of *CDH2*-lower patients with stage II/III based on immunohistochemistry (IHC) of phosphoglucose isomerase (GPI). Kaplan-Meier (KM) plots of *CDH2*-lower patients in TNM II (P=4.8e−01) (A), TNM III (P=2.2e−01) (B), and TNM II/III (P=4.0e−01) (C), separated by GPI expression in protein level.



| TNM II | | Survival outcome | |
|---|---|---|---|
| | | Good | Poor |
| Gene expression | High | $a_{II}$ | $b_{II}$ |
| | Low | $c_{II}$ | $d_{II}$ |

| TNM II | | Survival outcome | |
|---|---|---|---|
| | | Good | Poor |
| Gene expression | High | $a_{III}$ | $b_{III}$ |
| | Low | $c_{III}$ | $d_{III}$ |

$$ACE_{II}=a_{II}/(a_{II}+b_{II})-c_{II}/(c_{II}+d_{II})$$

$$ACE_{III}=a_{III}/(a_{III}+b_{III})-C_{III}/(c_{III}+d_{III})$$

$$P_{II}=(a_{II}+b_{II}+c_{II}+d_{II})/(a_{II}+b_{II}+c_{II}+d_{II}+a_{III}+b_{III}+c_{III}+d_{III})$$

$$P_{III}=(a_{III}+b_{III}+c_{III}+d_{III})/(a_{II}+b_{II}+c_{II}+d_{II}+a_{III}+b_{III}+c_{III}+d_{III})$$

$$ACE_{II+III,\ unadj}=(a_{II}+a_{III})/(a_{II}+b_{II}+a_{III}+b_{III})-(c_{II}+c_{III})/(c_{II}+d_{II}+c_{III}+d_{III})$$

$$ACE_{II+III,\ adj}=ACE_{II}\times P_{II}+ACE_{III}\times P_{III}$$

**Figure S4** Average causal effect (ACE) adjustment. (A) Contingency tables for TNM II and III; (B) Causal graph; (C) Computation of ACEs.

**Table S1** General information of patients

| Clinical items | n/n |
|---|---|
| Gender | |
| Female/Male | 52/146 |
| Age (year) | |
| ≤60/>60 | 80/118 |
| Tumor location | |
| Cardia/Noncardia | 46/152 |
| Lauren Classification | |
| Intestinal/Diffuse/Mix/unknown | 86/26/74/12 |
| TNM stage | |
| I/II/III/IV | 21/75/95/7 |
| Vessel carcinoma embolus | |
| Found/Not found/Unknown | 114/82/2 |
| Lymph node metastasis | |
| Metastasis/No metastasis/Unknown | 143/54/1 |
| 3-year PFS | |
| Good/Poor/Unknown* | 109/78/11 |
| 5-year OS | |
| Good/Poor/Unknown** | 103/84/11 |

PFS, progression-free survival; OS, overall survival; *, good: PFS ≥3 years, poor: PFS <3 years; **, good: OS ≥5 years, poor: OS <5 years.

**Table S2** List of top-27 gene markers

| Probe ID | Gene symbol | P_paired* | Fold-change** | 3-year PFS*** | | 5-year OS*** | | Rank | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | P | | Accuracy rate | |
| | | | | P | Accuracy rate | P | Accuracy rate | 3-year PFS | 5-year OS | 3-year PFS | 5-year OS |
| 143 | *GPI* | 2.11e–10 | 1.34e+00 | 8.27e–08 | 6.74e–01 | 1.37e–08 | 6.75e–01 | 1 | 1 | 1 | 1 |
| 247 | *PLA2G2A* | 8.65e–15 | 8.75e+00 | 4.97e–06 | 6.36e–01 | 3.05e–06 | 6.55e–01 | 12 | 6 | 35 | 11 |
| 489 | *IGFBP2* | 1.78e–17 | 3.51e-01 | 7.81e–06 | 6.40e–01 | 1.03e–05 | 6.51e–01 | 15 | 39 | 26 | 17 |
| 2266 | *EIF3B* | 2.57e–19 | 1.56e+00 | 3.77e–06 | 6.61e–01 | 6.05e–06 | 6.54e–01 | 10 | 25 | 3 | 13 |
| 14925 | *COASY* | 3.74e–01 | 1.03e+00 | 4.55e–07 | 6.47e–01 | 1.69e–06 | 6.41e–01 | 2 | 4 | 10 | 40 |
| 19021 | *FXN* | 5.78e–01 | 1.01e+00 | 1.13e–06 | 6.48e–01 | 2.87e–07 | 6.47e–01 | 4 | 2 | 8 | 24 |
| 2516 | *SLC2A4* | 1.23e–18 | 3.90e–01 | 1.97e–05 | 6.52e–01 | 6.13e–05 | 6.56e–01 | 29 | 130 | 6 | 10 |
| 6546 | *UNG* | 4.40e–01 | 1.04e+00 | 3.48e–05 | 6.23e–01 | 1.57e–05 | 6.51e–01 | 43 | 49 | 99 | 18 |
| 8952 | *UNC13B* | 2.31e–04 | 8.43e–01 | 1.61e–05 | 6.40e–01 | 9.38e–06 | 6.36e–01 | 23 | 32 | 25 | 59 |
| 16146 | *FBXO9* | 4.81e–15 | 7.35e–01 | 2.59e–05 | 6.23e–01 | 5.39e–06 | 6.40e–01 | 33 | 23 | 104 | 45 |
| 16998 | *DNER* | 4.75e–19 | 1.92e–01 | 3.82e–05 | 6.60e–01 | 1.81e–05 | 6.63e–01 | 47 | 62 | 5 | 5 |
| 18841 | *SLC35B2* | 1.62e–07 | 1.20e+00 | 1.72e–05 | 6.04e–01 | 6.35e–06 | 6.39e–01 | 24 | 26 | 428 | 49 |
| 19837 | *LUC7L* | 1.72e–10 | 8.44e–01 | 9.30e–06 | 6.61e–01 | 9.72e–05 | 6.66e–01 | 17 | 180 | 4 | 2 |
| 3836 | *TIMM8A* | 1.52e–10 | 1.31e+00 | 1.83e–04 | 6.45e–01 | 5.50e–05 | 6.65e–01 | 152 | 116 | 14 | 3 |
| 5211 | *BSG* | 1.51e–02 | 9.00e–01 | 1.04e–05 | 6.07e–01 | 4.42e–06 | 6.09e–01 | 20 | 17 | 365 | 473 |
| 6133 | *PTPRM* | 1.97e–06 | 7.46e–01 | 4.04e–05 | 6.14e–01 | 7.75e–06 | 6.32e–01 | 49 | 30 | 223 | 87 |
| 6515 | *TUFM* | 1.59e–04 | 1.15e+00 | 2.53e–05 | 6.22e–01 | 5.57e–06 | 6.37e–01 | 32 | 24 | 114 | 55 |
| 7402 | *FARSA* | 1.65e–16 | 1.42e+00 | 3.64e–06 | 6.07e–01 | 1.36e–05 | 6.25e–01 | 9 | 46 | 370 | 164 |
| 11851 | *MTFP1* | 5.38e–01 | 9.78e–01 | 8.92e–05 | 6.39e–01 | 1.11e–04 | 6.39e–01 | 86 | 196 | 29 | 47 |
| 12660 | *RSAD1* | 5.20e–07 | 8.29e–01 | 6.63e–04 | 6.35e–01 | 1.86e–04 | 6.43e–01 | 430 | 274 | 40 | 34 |
| 13918 | *GBGT1* | 9.42e–22 | 3.87e–01 | 4.74e–04 | 6.46e–01 | 2.16e–04 | 6.41e–01 | 334 | 300 | 13 | 44 |
| 14162 | *IPPK* | 4.51e–18 | 1.24e+00 | 4.83e–04 | 6.67e–01 | 3.54e–03 | 6.50e–01 | 343 | 1,711 | 2 | 22 |
| 16406 | *FAM165B* | 8.37e–20 | 3.76e–01 | 1.16e–04 | 6.47e–01 | 2.12e–05 | 6.60e–01 | 103 | 67 | 12 | 7 |
| 17780 | *LINGO2* | 3.05e–19 | 3.51e–01 | 3.09e–04 | 6.34e–01 | 4.48e–05 | 6.51e–01 | 251 | 103 | 42 | 19 |
| 18194 | *MINK1* | 9.33e–05 | 1.13e+00 | 2.67e–05 | 6.05e–01 | 1.51e–05 | 6.25e–01 | 34 | 48 | 410 | 166 |
| 19676 | *PTGER3* | 2.25e–19 | 2.45e–01 | 6.31e–04 | 6.40e–01 | 2.93e–04 | 6.47e–01 | 415 | 365 | 27 | 25 |
| 19747 | *ADSSL1* | 5.30e–07 | 7.46e–01 | 4.00e–05 | 5.84e–01 | 1.23e–05 | 6.01e–01 | 48 | 42 | 1,310 | 740 |

The color of each cell in the second column represents the level of the corresponding gene locating in *Figure 1D*. *, the paired *t*-test P value in NT-test; **, the corresponding fold-change; ***, P value and accuracy rate using Fisher's discriminant analysis (FDA), respectively; PFS, progression-free survival; OS, overall survival.

**Table S3** KEGG/GSEA/GO enrichment analyses on the top-27 genes

| Pathways/Gene sets | 1 | | | | | | 2 | | | | | | | 3 | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| in enrichment analysis | *GPI* | *PLA2G2A* | *IGFBP2* | *EIF3B* | *COASY* | *FXN* | *SLC2A4* | *UNG* | *UNC13B* | *FBXO9* | *DNER* | *SLC35B2* | *LUC7L* | *TIMM8A* | *BSG* | *PTPRM* | *TUFM* | *FARSA* | *MTFP1* | *RSAD1* | *GBGT1* | *IPPK* | *FAM165B* | *LINGO2* | *MINK1* | *PTGER3* | *ADSSL1* |
| KEGG pathways | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| hsa04110 Cell cycle | √ | | | √ | | | √ | √ | | √ | | | √ | √ | | √ | | √ | | | √ | √ | √ | | | √ | |
| hsa04152 AMPK signaling pathway | | √ | √ | | | | √ | | | √ | √ | | | | √ | √ | | | | √ | √ | | √ | √ | | √ | |
| hsa04115 p53 signaling pathway | √ | | | | | | √ | | | √ | √ | | | √ | | √ | | √ | | | √ | | √ | | | √ | |
| hsa04210 Apoptosis | √ | | | √ | | √ | | | | √ | | | √ | | √ | √ | | | | | | | | | √ | | |
| hsa04510 Focal Adhesion | | | √ | | √ | | √ | | | √ | √ | | √ | | | | | | √ | | | | √ | | | | |
| GSEA sets | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| VECCHI_GASTRIC_ CANCER | √ | √ | √ | √ | √ | √ | √ | √ | | √ | √ | | | √ | | √ | | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ |
| CHANG_CYCLING_ GENES | √ | √ | √ | √ | √ | √ | √ | √ | | √ | √ | | | √ | √ | √ | | √ | | | √ | √ | √ | √ | √ | √ | |
| BENPORATH_ PROLIFERATION | | √ | √ | | | | | | | √ | √ | | √ | √ | | √ | | | | √ | √ | | √ | √ | | | √ |
| KOBAYASHI_EGFR_ SIGNALING_24HR_DN | √ | √ | √ | √ | √ | √ | √ | √ | | √ | √ | | √ | √ | | √ | | √ | √ | | √ | √ | √ | √ | √ | √ | |
| GO pathways | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| Cell cycle, replication and division associated | √ | | | √ | √ | | √ | √ | | √ | | √ | | √ | | √ | √ | √ | √ | | √ | √ | √ | | | √ | |
| DNA repair | √ | | | √ | √ | | | √ | | | | √ | | √ | | √ | √ | √ | √ | | | √ | | | | | |
| Digestion | | √ | √ | | | | | | | | √ | | | | | √ | | | | √ | √ | | √ | √ | | √ | √ |
| Metabolic process | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | | √ | √ |

KEGG, Kyoto Encyclopedia of Genes and Genomes; GSEA, gene set enrichment analysis; GO, Gene Ontology.

## Supplementary materials

### *Extended information for prognostic marker selection*

#### Venn diagram consensus (VDC)

As briefly introduced in *Figure 1*, a Venn diagram was generated based on the overlap of the four subsets of mRNAs, which contained top-ranked ones measured by P value and testing accuracy of both 3-year progression-free survival (PFS) and 5-year overall survival (OS). As shown in the Venn diagram (*Figure 1D*), 82 mRNAs were involved in the four circles. The 27 mRNAs lying in the white, green and blue regions performed well in 3-year PFS and 5-year OS consistently. Considering the markers showing prognostic value in both PFS and OS are more likely to be practical in clinical use, we assigned the scheme of hierarchy for candidate mRNAs selection as shown in *Figure 1D*.

### *Enrichment and correlation analyses of candidate genes, metabolic pathways and epithelial-mesenchymal transition (EMT)*

#### Enrichment analysis of Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways

In total, 60 key KEGG pathways that play vital roles in tumor formation and evasion were selected from KEGG database and GSEA 'c2' gene sets ( http://software.broadinstitute.org/gsea/msigdb/genesets.jsp?collection=C2). Using Fisher's exact test, we generated P values for each pathway to evaluate its enrichment effect with the highly correlated genes of each candidate gene.

#### Gene set enrichment analysis (GSEA) enrichment

For each candidate gene marker, we performed GSEA for 'c2' gene sets, based on the gene expression in both tumor and nontumor tissues. In metabolic analysis, we simply adopted the metabolic gene sets collected by Gaude *et al* (1).

#### Gene Ontology (GO) enrichment

We performed GO enrichment analysis for 16 gene sets in GO enrichment system associated to cancer development and metastasis such as cell cycle, apoptosis, microtubule-based movement, cell adhesion, and so on. Based on the P values and enrichment scores calculated online by DAVID (https://david.ncifcrf.gov/), the significance of GO enrichment is evaluated.

#### Correlation analysis for metabolism pathways and EMT

In correlation analysis, the mean value of expressions of genes in the same metabolic pathway is generated to represent the expression level of the pathway. The EMT-signal is generated by the average expression of genes that up-/down-regulated in EMT process as collected by the previous study from Asian Cancer Research Group (2).

### *Extended information for Enviro-geno-pheno-state (E-GPS) analysis*

#### Implementations of E-GPS analysis

In the current E-GPS analysis, we aimed to find the optimal separating boundary to classify samples by their survival outcomes. We combined the tumor expression values of two different genes as 2D samples. We first applied support vector machine (SVM) (3) to project the 2D data into 1D values (namely SVM-scores). Next, we applied univariate E-GPS analysis based on the SVM-scores as follow procedures.

(1) All SVM-scores were sorted by increasing order;

(2) Every possibility of the two boundaries to separate these 1D data is enumerated, resulting in three groups of samples, corresponding to three E-GPS states. For each state *s*, a dedication degree (D-degree), $r_s$ is calculated by Eq. (2) in Ref. (4);

(3) The cost function *J* is generated by combining the D-degree of three states, i.e. $J=\sum_{s\in S \text{with} n_s\geq m \text{and} r_s>min(r_s)} \log(1-r_s)$, where $n_s$ is the total number of samples in state *s* from *S* (the complete set of the three states), and *m* is a threshold of minimum sample size for each state that is defined as 20% of the all samples, i.e., $m$ = 37 in our study;

(4) The best boundaries were selected when *J* reached the minimum. To enhance the robustness of the boundaries, we extended the line-form boundary to produce a band-form boundary. All points located within the band area (the region of line boundary ± 0.1 standard deviation of SVM-scores) were excluded from further calculation and optimization of the cost function (4). These samples were also excluded in the stratification for survival outcomes. As a result, the total number of samples in the three states is smaller than the total number of the whole sample set.

### Univariate E-GPS analysis on *CDH2*-negative patients in advanced stages by *GPI* and integrated metabolic signal

In some cases, we also performed univariate E-GPS analysis for single genes or pathways based on their expression values or integrated signals. In such analyses, independent datasets for validation were involved. However, these datasets usually contain few samples, if we still adopt band-form boundaries, the further loss of samples may lead to unreliable results. Hence, we used line-form boundaries for validation sets. Due to the same reason, the three E-GPS states were degenerated as two E-GPS states by merging two adjacent states that were with more similar outcomes.

## *Causal inference for prognosis with TNM staging and metabolic pathways*

We observed that 53%–60% of the metabolic signals indicated the prognosis in the same direction for patients with TNM II & III. However, the remaining metabolic pathways display different properties between the two stages. For example, glycogen degeneration acted as an indicator for good outcome in TNM II ($ACE_{II} >0$), meanwhile it was a risk factor for survival in TNM III ($ACE_{III} <0$). After merging patients of TNM II & III together, its divergent effects between TNM II & III counteracted, but still resulted in a total tendency of indicating slightly better survival (unadjusted $ACE_{II+III} >0$). Even more, we found for some metabolic signals, it could perform as a factor of poor outcome in TNM II & III, respectively, but the total effect of which unexpectedly turned to be positively correlated with good prognosis, which reminded us of the well-known Yule-Simpson's paradox. Such divergent even paradoxical phenomena uncovered that, in survival analysis, the Yule-Simpson's paradox also existed and might has been neglected for a long time. It means that, besides the traditional separate analyses for patients in different stages, the total effect of the gene expression marker should also be re-evaluated by solving the Yule-Simpson's paradox.

According to Pearl's and Rubin's causal theory (5-7), the ACE of the biomarker's expression toward the outcome could be adjusted by taking the TNM as a causal factor into consideration. We first built a causal graph model of the stage (TNM II or III), status of expression (high or low), and survival outcome (good or poor) as shown in *Supplementary Figure S4*. The ACE value of each metabolism pathway was calculated, based on which we evaluated the consistence of the indicating direction of metabolic signals for prognosis among patients with TNM II, III, and all patients of both stages. We further evaluate the consistence of the metabolic signals for prognosis in independent datasets including Singaporean, Korean, and TCGA data. Among the results of the 96 metabolism pathways, twenty-five of which (26%) were all consistent with the ones in our dataset.

## *Information of data sources and molecular experiments*

### Gene expression profile data sources

(1) Data from Peking University Cancer Hospital

A total of 198 patients with GC included in this study received their diagnosis and were surgically treated at Peking University Cancer Hospital between 2007 and 2010, and were followed up to March 2016. This investigation was performed after approval by the Ethics Committee of Peking University Cancer Hospital. General informed consent was obtained from each patient.

After radical gastrectomy, resected specimens were processed routinely for macroscopic pathological assessment, and tissues were sampled and snap-frozen in liquid nitrogen. Fresh human tissues were stored at −80 °C and fixed with 10% formalin in phosphate-buffered saline. To ensure the quality of the tissues, routine histological evaluation was performed for each sample.

GC stage was classified according to the 2010 tumor-node metastasis (TNM) classification recommended by the American Joint Committee on Cancer (AJCC 7th edition). T and N classification were assessed based on the final pathological result and M classification was determined by surgical findings. Early GC (EGC) was defined as a tumor that was confined to the mucosa or submucosa regardless of lymph node (LN) involvement. Advanced GC (AGC) was defined as a tumor that invaded the muscle proper or beyond. OS was calculated from the date of the initial surgery to the time of death caused by the tumor or the date of the last follow-up. PFS was calculated from the date of the initial surgery to the time of GC progression. None of the patients received chemotherapy or radiation therapy prior to surgery. A summary of clinical information is shown in *Supplementary Table S1*.

The mRNA expression profile of these tumors and paired noncancerous tissues were performed using the Agilent human mRNA & lncRNA Array V4.0 platform. All the 198 microarrays passed the quality control and were thus processed with quantile normalization and log-2 transformation. We further performed the prognostic biomarker study based on these normalized expression values of the 20,205 mRNAs.

(2) Independent validation datasets

For the Singaporean dataset, we combined the raw data of GSE34942 and GSE15459 together, using RMA algorithm from "Affy" package in R language to generate the normalized data. For the Korean dataset (GSE26253) and GSE29272 dataset from China, we adopted the same normalizing method with RMA separately. For TCGA, we downloaded the pre-processed data from UCSC cancer genome browser. We found that there are no enough patients with OS>5 years in TCGA, hence we grouped the 'good' and 'poor' survival into different strategies as we utilized for our own dataset and the other validated datasets. For TCGA, we followed the outcome classification defined by Gaude *et al* (2).

(3) Pharmacogenomic data

We downloaded the raw data of gene expression profiles of 965 cell lines from EBI database (Dataset ID: E-MTAB-3610). RMA normalization was performed. The drug sensitivity information valued by half maximal inhibitory concentration ($IC_{50}$) was directly adopted from the supplementary data of its originated study.

### Molecular experiments

RNA extraction. Total RNA was extracted using the Trizol reagent (Invitrogen) according to the manufacturer's protocol. The purity and concentration of RNA were determined by OD260/280 using spectrophotometer (NanoDrop 34 ND-1000). RNA integrity was determined by 1% formaldehyde denaturing gel electrophoresis.

Immunohistochemistry experiments of formalin-fixed paraffin-embedded (FFPE) sample. Four-micrometer sections from FFPE tissues were deparaffinized in xylene and rehydrated through graded alcohol. Antigen retrieval was performed by autoclaving in 0.01 mol/L citrate buffer (pH 6.0) for 3 min, followed by immersion in 3% hydrogen peroxide methanol for 10 min to block endogenous peroxidase activity. The section were then blocked with normal sheep serum (DAKO, Hamburg, Germany) for 90 min at room temperature and then incubated with glucose-6-phosphate isomerase (GPI) polyclonal antibody (Bethyl Laboratories, Inc., USA) diluted at 1:6,000 overnight at 4 °C. Diaminobenzidine was used as a chromogen, followed by counterstaining with hematoxylin. We regarded as GPI expression-positive when 10% or more cancer cells exhibited GPI in the cytoplasm. The expression of GPI was assessed independently by two experienced pathologists who were blinded to the patients' clinical outcomes. There was a high level of consistency among the two pathologists, and in the few discrepant cases (<5%) a consensus was reached after joint review.

## References

1. Gaude E, Frezza C. Tissue-specific and convergent metabolic transformation of cancer correlates with metastatic potential and patient survival. Nat Commun 2016;7:13041.
2. Cristescu R, Lee J, Nebozhyn M, et al. Molecular analysis of gastric cancer identifies subtypes associated with distinct clinical outcomes. Nat Med 2015;21:449-56.
3. Suykens JAK, Van Gestel T, De Brabanter J, et al. Least squares support vector machines. Singapore: World Scientific

Publishing, 2002.

4. Xu L. Enviro-geno-pheno state approach and state based biomarkers for differentiation, prognosis, subtypes, and staging. Applied Informatics 2016;3:4.
5. Pearl J. Simpson's paradox: An anatomy. Department of Statistics, UCLA, 2011.
6. Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. J Educ Psychol 1974;66:688-701.
7. Greenland S, Pearl J, Robins JM. Confounding and collapsibility in causal inference. Statist Sci 1999;14:29-46.