doi:10.4149/209_150531N303

Discovery of signature genes in gastric cancer associated with prognosis

X. ZHAO[1], H. CAI[2], X. WANG[3], L. MA[1,*]

[1]Department of General Surgery, Changhai Hospital, Second Military Medical University, Shanghai, China; [2]Department of Epidemiology, Second Military Medical University, Shanghai, China; [3]Department of Gynecology and Obstetrics, the 306 Hospital of PLA, Beijing ,China

**Correspondence: maliyye77@163.com*

Received May 31, 2015 / Accepted November 10, 2015

Gene expression profiles of gastric cancer (GC) were analyzed with bioinformatics tools to identify signature genes associated with prognosis. Four gene expression data sets (accession number: GSE2685, GSE30727, GSE38932 and GSE26253) were downloaded from Gene Expression Omnibus. Differentially expressed genes (DEGs) were screened out using significance analysis of microarrays (SAM) algorithm. P-value < 0.05 and |fold change| >1 were set as the threshold. A co-expression network was constructed for the GC-related genes with package WGCNA of *R*. Modules were disclosed with WGCNA algorithm. Survival-related signature genes were screened out via COX single-variable regression. A total of 3210 GC-related genes were identified from the 3 data sets. Significantly enriched GO biological process terms included cell death, cell proliferation, apoptosis, response to hormone and phosphorylation. Pathways like viral carcinogenesis, metabolism, EBV viral infection, and PI3K-AKT signaling pathway were significantly over-represented in the DEGs. A gene co-expression network including 2414 genes was constructed, from which 7 modules were revealed. A total of 17 genes were identified as signature genes, such as DAB2, ALDH2, CD58, CITED2, BNIP3L, SLC43A2, FAU and COL5A1. Many signature genes associated with prognosis of GC were identified in present study, some of which have been implicated in the pathogenesis of GC. These findings could not only improve the knowledge about GC, but also provide clues for clinical treatments.

Key words: gastric cancer, differentially expressed genes, functional enrichment analysis, gene co-expression network, survival curve, prognosis, signature genes

Gastric cancer (GC) is the second most common cause of cancer-related death worldwide. Risk factors of GC include diet, tobacco[1], infection of *Helicobacter pylori* and familiar GC [2]. The incidence, diagnostic studies, and therapeutic options have undergone big changes in the last decades, but the prognosis for GC patients remains poor, especially in more advanced stages [3]. Recurrence following surgery is a major problem, and is often the ultimate cause of death.

Molecular pathology can be helpful not only to understand the disease pathogenesis, but also to give useful prognostic molecular markers. Overexpression of p53 has been reported in 17–91% of invasive GC [4]. Sumiyoshi et al. suggest that the elevatedp53 is a marker for an unfavorable prognosis in GC [5]. Yokobori et al. further point out that p53 alters FBXW7 expression and the disruption of both p53 and FBXW7 contributes to poor prognosis in GC [6]. Besides, many biological prognostic factors have been proposed, such as p21 [7, 8], vascular endothelial growth factor (VEGF) [9], cyclin D2 [10] and HER2 [11].

Many studies have adopted microarray technology to identify critical genes in GC [12-14]. Lee et al. report that CDH17 is a prognostic marker for early stage GC [14]. Nishigaki et al. find aberrant expression of R-RAS in GC using microarrays and prove that blocking of the R-RAS-signaling pathway has great potential for GC therapy [15]. Obviously, these gene expression data are not fully utilized due to limited tools or restricted purpose. Mining valuable information with currently available bioinformatics tools from the plenty of gene expression data is of great significance.

In present study, we obtained a great number of GC-related genes via differential analysis of 3 gene expression data sets. Using another data set, a gene co-expression network was constructed for the GC-related genes and modules were disclosed. Then signature genes associated with survival time were screened out via Cox regression analysis.

Materials and methods

Gene expression data. Four gene expression data sets were downloaded from Gene Expression Omnibus[16]. GSE2685 [17]included22 primary human advanced gastric cancer tissues and 8 noncancerous gastric tissues. Affymetrix Human Full Length HuGeneFL Array (Affymetrix Inc., Santa Clara, California, USA) was used. GSE30727 analyzed 30 pairs of normal-cancer stomach tissues using the Affymetrix Human Exon 1.0 ST platform (Affymetrix Inc., Santa Clara, California, USA). GSE38932 [18] contained 12 gastric tumors and 12 paired adjacent non-tumoral gastric tissues. The platform was HEEBO Human oligo array. GSE26253 [19] included 432 formalin fixed paraffin embedded gastric tumor tissues and Illumina HumanRef-8 WG-DASL v3.0 was used (Illumina, USA).

Pre-treatment of raw data. Probes were mapped to genes. For genes corresponding to more than one probe, gene expression levels were determined by the average probe values [20]. Log2 conversion and quantile normalization [21] were applied on data. Genes with more than 20% missing values were removed and others were filled with average expression level.

Differential analysis. Significance analysis of microarrays (SAM) algorithm [22] was adopted to screen out differentially expressed genes (DEGs). SAM can reduce the false-positive rate in multiple testing via controlling false discovery rate (FDR). Relative difference (statistic *d*) is calculated as follow:

$$d = \frac{X'_1 | X'_2}{S + s_0} \quad (1)$$

Statistic *d* measures the relative differences in gene expression levels and it is the corrected *t*. X_1' represents the average expression level of a gene under certain state, X_2' represents the average expression level of a gene under another state, and *s* represents the variance of a gene.

FDR< 0.05 and log2|fold change| >1 were set as the threshold to screen out DEGs.

Functional enrichment analysis. Gene Onotology (GO) enrichment analysis and pathway enrichment analysis were performed for the DEGs with toppGene [23], a test based upon hypergeometric distribution. FDR < 0.05 was set as the cut-off value.

Gene co-expression network analysis. A gene co-expression network was constructed for the GC-related genes with package WGCNA (weighted gene co-expression network analysis)[24] of *R*. The connection coefficient *a* was calculated for a pair of genes as follow:

$$a_{ij} = S_{ij}^{\beta}, \text{ where } S_{ij} = |\text{cor}(x_i, x_j)| \quad (2)$$

Where *Xi* and *Xj* represent expression vectors of gene *i* and gene *j*, *cor* represents Pearson correlation coefficient of the two vectors. Pearson correlation coefficient is converted into connection coefficients *aij* via exponential transformation. Exponential transform can strengthen strong correlation but weaken weak correlation, and thus improve the reliability of the network.

WGCNA algorithm takes topological properties into consideration to identify modules from the network. The algorithm considers not only the two directly connected genes, but also others genes linked with the two genes. It calculates weighting coefficient *Wij* from connection coefficient *aij* as follow:

$$W_{ij} = \frac{l_{ij} + a_{ij}}{\min\{k_i, k_j\} + 1 - a_{ij}} \quad (3)$$

$$\text{where } l_{ij} = \sum_u a_{iu} a_{uj}, \; k_i = \sum_u a_{iu}$$

Wij considers overlap in the neighboring genes of node *i* and node *j*. Modules were identified via hierarchical clustering of the weighting matrix *W*.

Survival analysis. After we selected several GC-related genes and then clustered them into different modules basing on the different expressions, the patients were classified as having or not having relapse based on module genes using Support Vector Machine (SVM)[25]. A tenfold cross-validation method was chosen to evaluate the classification result. Survival-related signature genes were screened out via single-variable Cox regression analysis.

Results

Differentially expressed genes. A set of 5,524, 8,988 and 20,860 genes are detected in the data sets of GSE2685, GSE38923 and GSE30727, respectively. The box plots are shown in Figure 1, which indicates a uniform expression distribution of the samples after normalization. According to the criteria (FDR< 0.05 and log2|fold change| > 1), the DEGs were screened out in three data sets which contained both tumor and normal tissues. Consequently, a cohort of1282 DEGs were obtained in GSE2685, 28 DEGs in GSE38932 and 2041 DEGs in GSE30727. As shown in Figure 2, only a few genes were common among the 3 sets of DEGs, suggesting high heterogeneity in GC samples. The 3 sets of DEGs were combined and a total of 3210 genes were acquired, which were regarded as GC-related genes.

Functional enrichment analysis result. Significantly enriched GO biological process (BP) terms for the DEGs included cell death, cell proliferation, apoptosis, response to hormone and phosphorylation (Supplementary Table 1A). Pathway enrichment analysis revealed viral carcinogenesis, metabolism, EBV viral infection, and PI3K-AKT signaling pathway (Supplementary Table 1B). These functions and pathways have been implicated in the pathogenesis of GC [26].

Gene co-expression network and signature genes. To further screen out signature genes from the 3210 GC-related

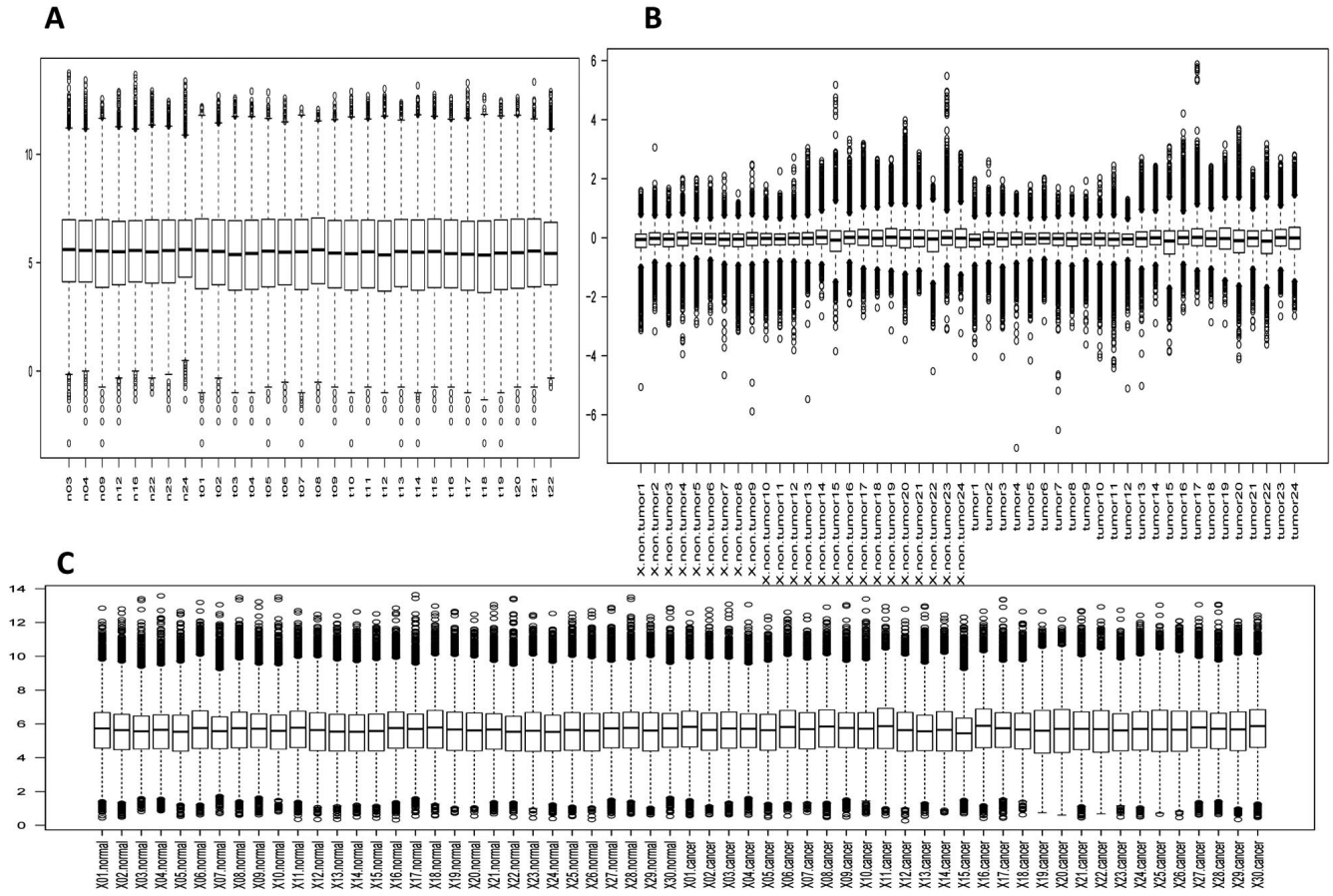

Figure 1. Box plots of the 3 gene expression data sets. A: GSE2685; B: GSE38932; C: GSE30727.

genes, another gene expression data set (GSE25263), which comprised only 432 GC patient samples with clinical information, was acquired from GEO. Same pre-treatment was applied on the raw data. Box plot for the 20 samples randomly selected out of the 432 samples also revealed a uniform expression among samples after normalization (Figure 3).

The 3210 GC-related genes showed various expression patterns in the 432 samples (Figure 4). A co-expression network including 2414 genes was constructed with WGCNA algorithm, from which 7 modules were identified (Supplementary Table 2).

Functional enrichment analysis was performed for the genes in each module. No significant term was enriched in module "green" and terms enriched in module "red" was not associated with GC. GC-related terms were identified in other 5 modules ("Yellow", "blue", "grey", "Brown" and "turquoise") and thus they were considered as GC-related modules.

We found that relapse could well classify patients with different survival time by Kaplan-Meier test ($P < 0.05$) (Figure 5). The genes of module "yellow" could separate patients with different prognosis and the error rate of cross-validation was 0.041. Cox regression analysis was conducted for the 65 genes in the "yellow" module to find out genes closely associated with survival time. With the adjusted p-value < 0.1 as the cut-off, a total of 17 genes were revealed (Supplementary Table 3), such as Dab homolog 2 (DAB2), aldehyde dehydrogenase 2 family (ALDH2), CD58 molecule (CD58), Cbp/p300-interacting transactivator, with Glu/Asp-rich carboxy-terminal domain, 2 (CITED2) and BCL2/adenovirus E1B 19kDa interacting protein 3-like (BNIP3L).

Discussion

In present study, we acquired a total of 3210 GC-related genes through differential analysis of 3 gene expression data sets. Functional enrichment analysis was applied on these genes. Cell proliferation and apoptosis were significantly enriched, which were closely related to cancer development. Pathway enrichment analysis showed that viral carcinogenesis, EBV viral infection, and PI3K-AKT signaling pathway were significantly over-represented in the DEGs. Epstein–Barr virus (EBV) has been linked to GC [27, 28]. The EBV is detected in the tissue of about 10% of gastric carcinoma cases throughout

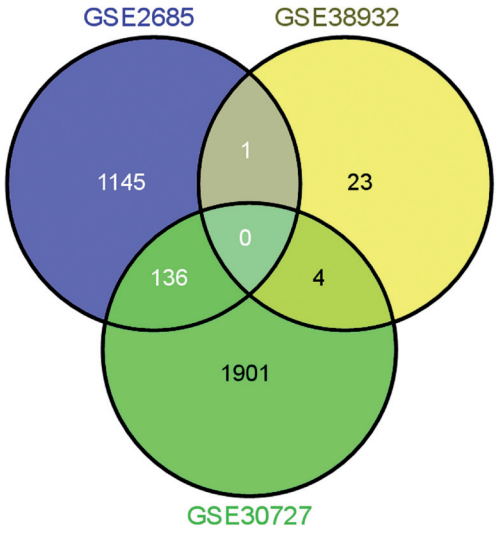

Figure 2. Venn diagram of the 3 sets of differentially expressed genes.

the world. Takada et al. report that EBV contributes to the maintenance of the malignant phenotype of EBV positive GC [28]. Phosphatidylinositol-3-kinase (PI3K) is a lipid kinase and produces phosphatidylinositol-3,4,5-trisphosphate (PI(3, 4, 5)P3), a second messenger essential for the translocation of Akt to the plasma membrane where it is phosphorylated and activated by phosphoinositide-dependent kinase (PDK) 1 and PDK2. Activation of Akt plays a pivotal role in fundamental cellular functions such as cell proliferation and survival by phosphorylating a variety of substrates. PI3K-AKT signaling pathway has been closely associated with various aspects of cancers [29, 30]. Osaki et al. report that inhibition of the PI3K-Akt signaling pathway enhances the sensitivity of Fas-mediated apoptosis in human GC cell line, MKN-45 [31]. These significantly enriched biological processes and pathways confirmed the reliability of the DEGs.

To screen out key genes associated with prognosis in GC patients, another gene expression data set was obtained, based upon which a weighted gene co-expression network was constructed with WGCNA method. A total of 7 modules were identified. Finally, a total of 17 genes were closely related to survival time and thus were considered as signature genes. Some of them have been implicated in GC according to previous studies, such as DAB2 [32], ALDH2 [33], CD58 [34], CITED2 [35] and BNIP3L [36].ALDH2 belongs to the aldehyde dehydrogenase family of proteins. Aldehyde dehydrogenase is the second enzyme of the major oxidative pathway of alcohol metabolism. Shin et al. find that ALDH2 polymorphisms modify the susceptibility to the development of GC associated with alcohol intake [37]. Wang et al. carry out a meta-analysis and suggest that ALDH2 and ADH1 genetic polymorphisms may play crucial roles in the pathogenesis of GC [33]. Mayer et al. find that expression of CD56 by more than 50% of the tumor cells correlates with tumor recurrence and decreased survival time and it may be involved in the development of distant metastases of GC [34].CITED2 is a gene that mediates sensitivity to chemotherapeutics. Regel et al. indicate that levels of CITED2 in gastric tumors correlate with patients' response to epirubicin[35], suggesting it might be a therapeutic target to modulate chemotherapy and thus benefit prognosis.

With regard to the other genes (or the protein products), SLC43A2 (Solute carrier group of membrane transport protein, member 2), a Na+-independent transporter that has significant role in the transport of macromolecule amino acid across membranes, is verified to be elevated in most metastatic GCs [38]. This cancer-related protein is reported associated

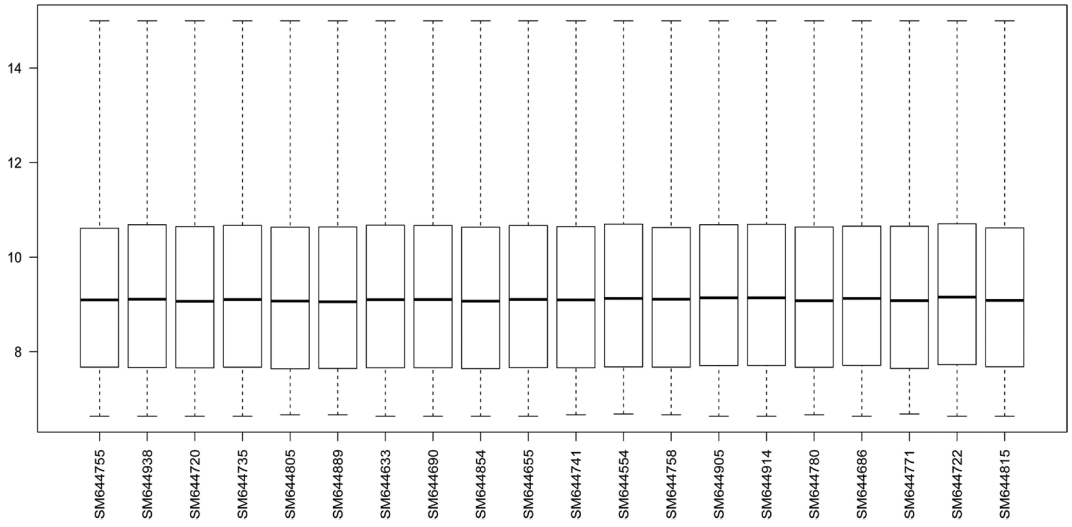

Figure 3. Box plot of 20 samples randomly selected from GSE25263.

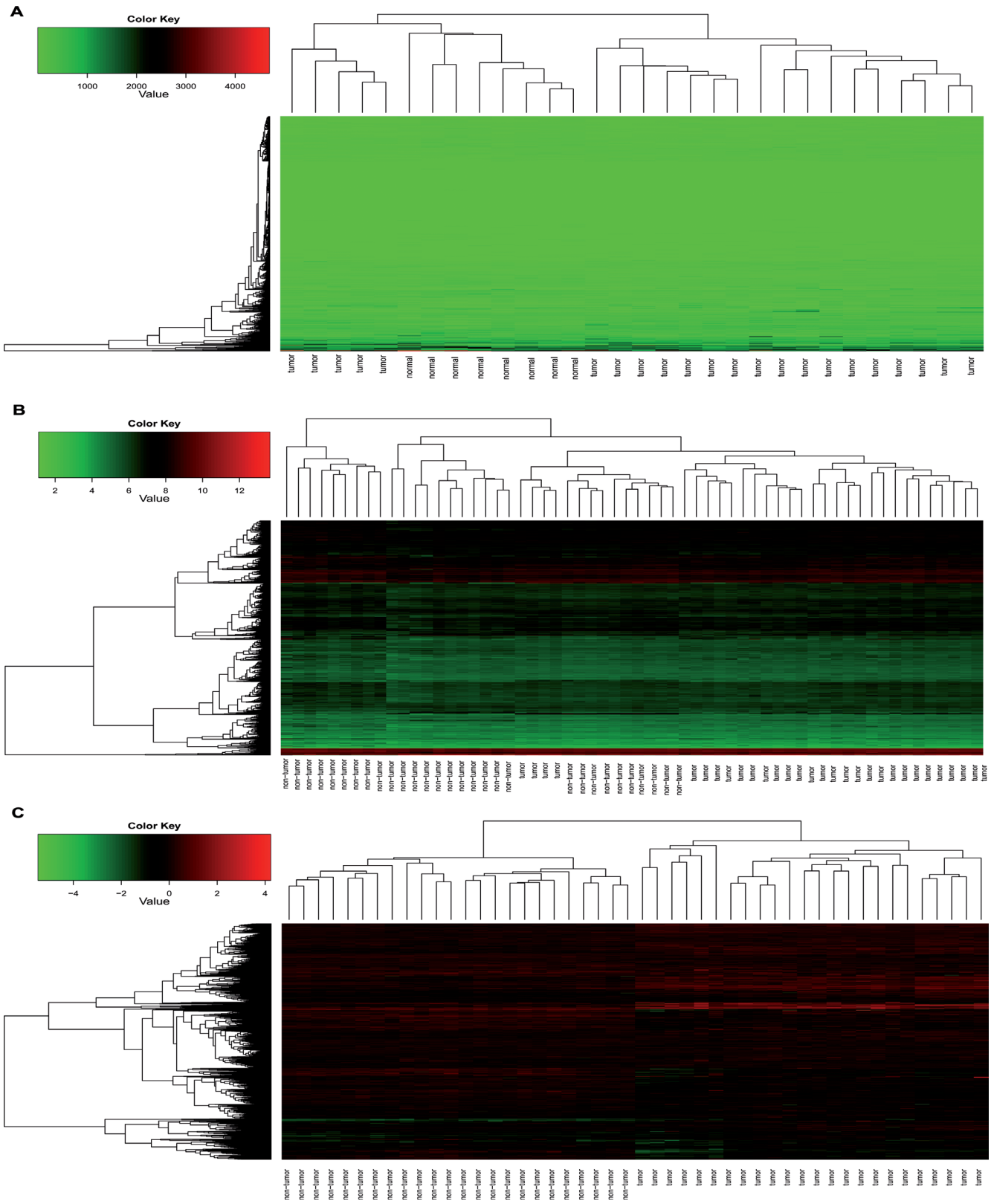

Figure 4. Heatmapof the differentially expressed genes (DEGs) in three data sets. A: DEGs in GSE2685; B: DEGs in GSE30727; C: DEGs in GSE38932. X-axis represents the tumor and the normal samples and Y-axis indicates the differentially expressed genes.

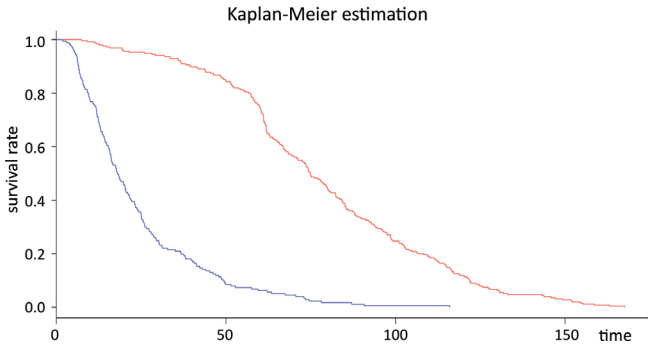

Figure 5. Survival curves of patients with (blue line) or without relapse (red line)

with the pancreatic cancer survival [39]. These provide potent evidence that SLC43A2 might also be implicated in GC survival time, as suggested in our results. In breast cancer, the homologous FAU gene (Finkel-Biskis-Reillymurine sarcoma virus) is found in the region of 11q13~q14, which other cell cycle-related genes are also located, and acts as a tumor suppressor in vitro [40]. However, at present, no data indicate the association between this gene and GC, suggesting that FAU might be a novel indicator for the prognosis of GC. Overexpression of the collagen COL5A1 is discovered in malignant lesion of stomach, suggesting the important role of COL5A1 in the GC progression [41]. These illustrations collectively suggest that the three genes might be novel biomarkers for the GC prognosis.

Overall, 17 genes significantly linked with survival time of GC patients were disclosed in present study. First, they could be used for prognosis. Second, some of them might be potential targets and thus could be exploited for GC therapy. Besides, the 3210 GC-related genes could also provide clues for future researches on GC.

Acknowledgements: This study was supported by Establishment and verification of molecular signature for the prediction of gastric cancer prognosis(2014M552576).

References

[1] SJODAHL K, LU Y, NILSEN T I, YE W, HVEEM K, et al. Smoking and alcohol drinking in relation to risk of gastric cancer: a population-based, prospective cohort study. Int J Cancer 2007;120: 128–132. http://dx.doi.org/10.1002/ijc.22157

[2] BARBER M, FITZGERALD R C, CALDAS C. Familial gastric cancer – aetiology and pathogenesis. Best Pract Res Clin Gastroenterol 2006;20: 721–734. http://dx.doi.org/10.1016/j.bpg.2006.03.014

[3] CATALANO V, LABIANCA R, BERETTA G D, GATTA G, DE BRAUD F, et al. Gastric cancer. Critical Reviews in Oncology / Hematology 71: 127–164. http://dx.doi.org/10.1016/j.critrevonc.2009.01.004

[4] FENOGLIO-PREISER C M, WANG J, STEMMERMANN G N, NOFFSINGER A. TP53 and gastric carcinoma: a review. Hum Mutat 2003;21: 258–270. http://dx.doi.org/10.1002/humu.10180

[5] SUMIYOSHI Y, KAKEJI Y, EGASHIRA A, MIZOKAMI K, ORITA H, et al. Overexpression of hypoxia-inducible factor 1alpha and p53 is a marker for an unfavorable prognosis in gastric cancer. Clin Cancer Res 2006;12: 5112–5117. http://dx.doi.org/10.1158/1078-0432.CCR-05-2382

[6] YOKOBORI T, MIMORI K, IWATSUKI M, ISHII H, ONOYAMA I, et al. p53-Altered FBXW7 expression determines poor prognosis in gastric cancer cases. Cancer Res 2009;69: 3788–3794. http://dx.doi.org/10.1158/0008-5472.CAN-08-2846

[7] XIANGMING C, HOKITA S, NATSUGOE S, TANABE G, BABA M, et al. p21 expression is a prognostic factor in patients with p53-negative gastric cancer. Cancer Lett 2000;148: 181–188. http://dx.doi.org/10.1016/S0304-3835(99)00335-3

[8] KOURAKLIS G, KATSOULIS I E, THEOCHARIS S, TSOUROUFLIS G, XIPOLITAS N, et al. Does the expression of cyclin E, pRb, and p21 correlate with prognosis in gastric adenocarcinoma? Dig Dis Sci 2009;54: 1015–1020. http://dx.doi.org/10.1007/s10620-008-0464-y

[9] LIETO E, FERRARACCIO F, ORDITURA M, CASTELLANO P, MURA A L, et al. Expression of vascular endothelial growth factor (VEGF) and epidermal growth factor receptor (EGFR) is an independent prognostic indicator of worse outcome in gastric cancer patients. Ann Surg Oncol 2008;15: 69–79. http://dx.doi.org/10.1245/s10434-007-9596-0

[10] TAKANO Y, KATO Y, VAN DIEST P J, MASUDA M, MITOMI H, et al. Cyclin D2 overexpression and lack of p27 correlate positively and cyclin E inversely with a poor prognosis in gastric cancer cases. Am J Pathol 2000;156: 585–594. http://dx.doi.org/10.1016/S0002-9440(10)64763-3

[11] GRAVALOS C, JIMENO A. HER2 in gastric cancer: a new prognostic factor and a novel therapeutic target. Ann Oncol 2008;19: 1523–1529. http://dx.doi.org/10.1093/annonc/mdn169

[12] KIM J M, SOHN H Y, YOON S Y, OH J H, YANG J O, et al. Identification of gastric cancer-related genes using a cDNA microarray containing novel expressed sequence tags expressed in gastric cancer cells. Clin Cancer Res 2005;11: 473–482.

[13] MYLLYKANGAS S, JUNNILA S, KOKKOLA A, AUTIO R, SCHEININ I, et al. Integrated gene copy number and expression microarray analysis of gastric cancer highlights potential target genes. Int J Cancer 2008;123: 817–825. http://dx.doi.org/10.1002/ijc.23574

[14] LEE H J, NAM K T, PARK H S, KIM M A, LAFLEUR B J, et al. Gene expression profiling of metaplastic lineages identifies CDH17 as a prognostic marker in early stage gastric cancer. Gastroenterology 2010;139: 213–225 e213.

[15] NISHIGAKI M, AOYAGI K, DANJOH I, FUKAYA M, YANAGIHARA K, et al. Discovery of aberrant expression of R-RAS by cancer-linked DNA hypomethylation in gastric cancer using microarrays. Cancer Res 2005;65: 2115–2124. http://dx.doi.org/10.1158/0008-5472.CAN-04-3340

[16] BARRETT T, SUZEK T O, TROUP D B, WILHITE S E, NGAU W C, et al. NCBI GEO: mining millions of expression profiles--database and tools. Nucleic Acids Res 2005;33: D562–566. http://dx.doi.org/10.1093/nar/gki022

[17] HIPPO Y, TANIGUCHI H, TSUTSUMI S, MACHIDA N, CHONG J M, et al. Global gene expression analysis of gastric cancer by oligonucleotide microarrays. Cancer Res 2002;62: 233–240.

[18] BIZAMA C, BENAVENTE F, SALVATIERRA E, GUTIERREZ-MORAGA A, ESPINOZA J A, et al. The low-abundance transcriptome reveals novel biomarkers, specific intracellular pathways and targetable genes associated with advanced gastric cancer. Int J Cancer 2014;134: 755–764. http://dx.doi.org/10.1002/ijc.28405

[19] LEE J, SOHN I, DO I G, KIM K M, PARK S H, et al. Nanostring-based multigene assay to predict recurrence for gastric cancer patients after surgery. PloS one 2014;9: e90133. http://dx.doi.org/10.1371/journal.pone.0090133

[20] MA H, SCHADT E E, KAPLAN L M, ZHAO H. COSINE: COndition-SpecIfic sub-NEtwork identification using a global optimization method. Bioinformatics 2011;27: 1290–1298. http://dx.doi.org/10.1093/bioinformatics/btr136

[21] FERRARI F, BORTOLUZZI S, COPPE A, SIROTA A, SAFRAN M, et al. Novel definition files for human GeneChips based on GeneAnnot. BMC Bioinformatics 2007;8: 446. http://dx.doi.org/10.1186/1471-2105-8-446

[22.] LARSSON O, WAHLESTEDT C, TIMMONS J A. Considerations when using the significance analysis of microarrays (SAM) algorithm. BMC Bioinformatics 2005;6: 129. http://dx.doi.org/10.1186/1471-2105-6-129

[23] CHEN J, BARDES E E, ARONOW B J, JEGGA A G. ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. Nucleic Acids Res 2009;37: W305–311. http://dx.doi.org/10.1093/nar/gkp427

[24] LANGFELDER P, HORVATH S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics 2008;9: 559. http://dx.doi.org/10.1186/1471-2105-9-559

[25] LIU J, LI S C, LUO X. Iterative reweighted noninteger norm regularizing SVM for gene expression data classification. Comput Math Methods Med 2013;2013: 768404. http://dx.doi.org/10.1155/2013/768404

[26] Comprehensive molecular characterization of gastric adenocarcinoma. Nature 2014; 513: 202–209. http://dx.doi.org/10.1038/nature13480

[27] FUKAYAMA M, HAYASHI Y, IWASAKI Y, CHONG J, OOBA T, et al. Epstein-Barr virus-associated gastric carcinoma and Epstein-Barr virus infection of the stomach. Lab Invest 1994;71: 73–81.

[28] TAKADA K. Epstein-Barr virus and gastric carcinoma. Mol Pathol 2000;53: 255–261. http://dx.doi.org/10.1136/mp.53.5.255

[29] HENNESSY B T, SMITH D L, RAM P T, LU Y, MILLS G B. Exploiting the PI3K/AKT pathway for cancer drug discovery. Nat Rev Drug Discov 2005;4: 988–1004. http://dx.doi.org/10.1038/nrd1902

[30] ALTOMARE D A, TESTA J R. Perturbations of the AKT signaling pathway in human cancer. Oncogene 2005;24: 7455–7464. http://dx.doi.org/10.1038/sj.onc.1209085

[31] OSAKI M, KASE S, ADACHI K, TAKEDA A, HASHIMOTO K, et al. Inhibition of the PI3K-Akt signaling pathway enhances the sensitivity of Fas-mediated apoptosis in human gastric carcinoma cell line, MKN-45. J Cancer Res Clin Oncol 2004;130: 8–14. http://dx.doi.org/10.1007/s00432-003-0505-z

[32] DOTE H, TOYOOKA S, TSUKUDA K, YANO M, OTA T, et al. Aberrant promoter methylation in human DAB2 interactive protein (hDAB2IP) gene in gastrointestinal tumour. British journal of cancer 2005;92: 1117–1125. http://dx.doi.org/10.1038/sj.bjc.6602458

[33] WANG H L, ZHOU P Y, LIU P, ZHANG Y. ALDH2 and ADH1 genetic polymorphisms may contribute to the risk of gastric cancer: a meta-analysis. PloS one 2014;9: e88779. http://dx.doi.org/10.1371/journal.pone.0088779

[34] MAYER B, LORENZ C, BABIC R, JAUCH K W, SCHILDBERG F W, et al. Expression of leukocyte cell adhesion molecules on gastric carcinomas: possible involvement of LFA-3 expression in the development of distant metastases. Int J Cancer 1995;64: 415–423. http://dx.doi.org/10.1002/ijc.2910640611

[35] REGEL I, MERKL L, FRIEDRICH T, BURGERMEISTER E, ZIMMERMANN W, et al. Pan-histone deacetylase inhibitor panobinostat sensitizes gastric cancer cells to anthracyclines via induction of CITED2. Gastroenterology 2012;143: 99–109 e110.

[36] ARAO T, YANAGIHARA K, TAKIGAHIRA M, TAKEDA M, KOIZUMI F, et al. ZD6474 inhibits tumor growth and intraperitoneal dissemination in a highly metastatic orthotopic gastric cancer model. Int J Cancer 2006;118: 483–489. http://dx.doi.org/10.1002/ijc.21340

[37] SHIN C M, KIM N, CHO S I, KIM J S, JUNG H C, et al. Association between alcohol intake and risk for gastric cancer with regard to ALDH2 genotype in the Korean population. Int J Epidemiol 2011;40: 1047–1055. http://dx.doi.org/10.1093/ije/dyr067

[38] XU L, WANG F, XU X-F, MO W-H, WAN R, et al. Data mining of microarray for differentially expressed genes in liver metastasis from gastric cancer. Frontiers of Medicine in China 2010;4: 247–253. http://dx.doi.org/10.1007/s11684-010-0027-4

[39] WU T T, GONG H, CLARKE E M. A transcriptome analysis by lasso penalized Cox regression for pancreatic cancer survival. Journal of bioinformatics and computational biology 2011;9: 63–73. http://dx.doi.org/10.1142/S0219720011005744

[40] VALLADARES A, SALAMANCA F, MADRIGAL-BUJAIDAR E, ARENAS D. Identification of chromosomal changes with comparative genomic hybridization in sporadic breast cancer in Mexican women. Cancer genetics and cytogenetics 2004;152: 163–166. http://dx.doi.org/10.1016/j.cancergencyto.2003.11.016

[41] ZHAO Y, ZHOU T, LI A, YAO H, HE F, et al. A potential role of collagens expression in distinguishing between premalignant and malignant lesions in stomach. The Anatomical Record 2009;292: 692–700. http://dx.doi.org/10.1002/ar.20874

Table 1A: Top 20 GO biological process terms significantly enriched in the differentially expressed genes

ID	Name	p-value	q-value Bonferroni
GO:0071310	cellular response to organic substance	1.16E-24	1.07E-20
GO:0010941	regulation of cell death	1.19E-20	1.10E-16
GO:0009719	response to endogenous stimulus	1.39E-20	1.28E-16
GO:1901700	response to oxygen-containing compound	3.84E-20	3.53E-16
GO:0008283	cell proliferation	4.63E-20	4.25E-16
GO:0043067	regulation of programmed cell death	1.91E-19	1.76E-15
GO:0042981	regulation of apoptotic process	1.97E-19	1.81E-15
GO:0009725	response to hormone	8.32E-19	7.65E-15
GO:0016310	phosphorylation	9.79E-19	9.00E-15
GO:0006915	apoptotic process	1.01E-17	9.32E-14
GO:0012501	programmed cell death	1.40E-17	1.29E-13
GO:0001775	cell activation	2.70E-17	2.48E-13
GO:0006468	protein phosphorylation	5.03E-17	4.63E-13
GO:0032268	regulation of cellular protein metabolic process	8.97E-17	8.25E-13
GO:0048584	positive regulation of response to stimulus	1.23E-16	1.13E-12
GO:0042325	regulation of phosphorylation	1.32E-16	1.21E-12
GO:0014070	response to organic cyclic compound	1.62E-16	1.49E-12

GO:0034097	response to cytokine	1.73E-16	1.59E-12
GO:0040011	locomotion	2.05E-16	1.88E-12
GO:0042127	regulation of cell proliferation	2.12E-16	1.95E-12

Table 1B: Top 20 pathways significantly enriched in the differentially expressed genes

Pathway Name	**Source**	**p-value**	**q-value BH-adjust**
Viral carcinogenesis	KEGG	1.51E-10	4.68E-07
Metabolism	REACTOME	9.31E-10	2.89E-06
PI3K-Akt signaling pathway	KEGG	3.14E-07	9.73E-04
Epstein-Barr virus infection	KEGG	6.99E-07	2.17E-03
Valine, Leucine and Isoleucine Degradation	SMPDB	1.55E-06	4.81E-03
Amoebiasis	KEGG	3.69E-06	1.15E-02
Viral myocarditis	KEGG	5.77E-06	1.79E-02
Metabolism of lipids and lipoproteins	REACTOME	7.33E-06	2.27E-02
HTLV-I infection	KEGG	1.13E-05	3.49E-02
Focal adhesion	KEGG	1.20E-05	3.73E-02
Focal Adhesion	WikiPathways	1.33E-05	4.14E-02
Asthma	KEGG	1.48E-05	4.60E-02
Valine, leucine and isoleucine degradation	KEGG	1.64E-05	5.07E-02
Allograft rejection	KEGG	1.89E-05	5.86E-02

MAP00280 Valine leucine and isoleucine degradation	GenMAPP	2.91E-05	9.01E-02
Developmental Biology	REACTOME	2.98E-05	9.26E-02
Pathways in cancer	KEGG	3.79E-05	1.18E-01
Interferon Signaling	REACTOME	3.93E-05	1.22E-01
Cytokine Signaling in Immune system	REACTOME	4.88E-05	1.51E-01
mitogen activated protein kinase signaling	Pathway Ontology	5.19E-05	1.61E-01

Table 2: Seven modules derived from the co-expression network

Moduleclasses	**Number of genes**	**Module genes (examples)**	**Biological functions**
Blue	351	IL10,TIMP1,FNTB, GATA6,FABP4…	Cell proliferation, metabolic, cell death
Brown	176	SP1,KLF11,NCOA6,KLF6, SRC,HLX,CASP8…	immune system development, viral process
Green	45	AXIN1,CAPN1,EWSR1	NA
grey	1308	MTG1,SPRY1,MDM2, ADAR,FABP1,MEN1…	cell death, apoptotic process, cell proliferation, metabolic
Red	33	OLFM1,HMGB2,TRADD, MYO1B,CAT,SPAST	catabolic process
Turquoise	435	RPS27A,PFN1,RPL23, UBA52,CCL5,USF1	immune response, viral process
Yellow	65	STK4,RSPO3,MITF, DAB2,CTNNB1,SMAD3	Wnt signaling pathway, cell death, apoptosis, immune system

Table 3: Genes significantly associated with survival time from module “yellow”

Modulegenes	**coef**	**exp(coef)**	**se(coef)**	**z**	**Pr(>|z|)**
SLC43A2	0.2446	1.2770	0.0617	3.9643	7.36E-05
FAU	0.5348	1.7072	0.1650	3.2416	0.0012
DAB2	-0.2460	0.7819	0.0783	-3.1410	0.0017
COL5A1	-0.2177	0.8043	0.0771	-2.8253	0.0047
ZCCHC2	-0.0891	0.9147	0.0323	-2.7572	0.0058
ISY1	0.1563	1.1692	0.0574	2.7245	0.0064
WDR1	-0.2144	0.8070	0.0845	-2.5382	0.0111
DOCK10	-0.1230	0.8843	0.0491	-2.5052	0.0122
C9orf142	0.2126	1.2369	0.0853	2.4931	0.0127
SH3BP4	-0.1579	0.8539	0.0648	-2.4363	0.0148
MRPS16	0.4509	1.5698	0.1882	2.3963	0.0166
ALDH2	-0.0784	0.9246	0.0367	-2.1359	0.0327
UBE2H	-0.2403	0.7864	0.1203	-1.9972	0.0458
MAEA	0.1293	1.1380	0.0692	1.8685	0.0617
CD58	-0.0660	0.9361	0.0358	-1.8440	0.0652
CITED2	-0.0691	0.9332	0.0402	-1.7214	0.0852
BNIP3L	-0.0776	0.9253	0.0462	-1.6804	0.0929

coef: coefficient

Table 1A: Top 20 GO biological process terms significantly enriched in the differentially expressed genes

ID	Name	p-value	q-value Bonferroni
GO:0071310	cellular response to organic substance	1.16E-24	1.07E-20
GO:0010941	regulation of cell death	1.19E-20	1.10E-16
GO:0009719	response to endogenous stimulus	1.39E-20	1.28E-16
GO:1901700	response to oxygen-containing compound	3.84E-20	3.53E-16
GO:0008283	cell proliferation	4.63E-20	4.25E-16
GO:0043067	regulation of programmed cell death	1.91E-19	1.76E-15
GO:0042981	regulation of apoptotic process	1.97E-19	1.81E-15
GO:0009725	response to hormone	8.32E-19	7.65E-15
GO:0016310	phosphorylation	9.79E-19	9.00E-15
GO:0006915	apoptotic process	1.01E-17	9.32E-14
GO:0012501	programmed cell death	1.40E-17	1.29E-13
GO:0001775	cell activation	2.70E-17	2.48E-13
GO:0006468	protein phosphorylation	5.03E-17	4.63E-13
GO:0032268	regulation of cellular protein metabolic process	8.97E-17	8.25E-13
GO:0048584	positive regulation of response to stimulus	1.23E-16	1.13E-12
GO:0042325	regulation of phosphorylation	1.32E-16	1.21E-12
GO:0014070	response to organic cyclic compound	1.62E-16	1.49E-12

GO:0034097	response to cytokine	1.73E-16	1.59E-12
GO:0040011	locomotion	2.05E-16	1.88E-12
GO:0042127	regulation of cell proliferation	2.12E-16	1.95E-12

Table 1B: Top 20 pathways significantly enriched in the differentially expressed genes

Pathway Name	**Source**	**p-value**	**q-value BH-adjust**
Viral carcinogenesis	KEGG	1.51E-10	4.68E-07
Metabolism	REACTOME	9.31E-10	2.89E-06
PI3K-Akt signaling pathway	KEGG	3.14E-07	9.73E-04
Epstein-Barr virus infection	KEGG	6.99E-07	2.17E-03
Valine, Leucine and Isoleucine Degradation	SMPDB	1.55E-06	4.81E-03
Amoebiasis	KEGG	3.69E-06	1.15E-02
Viral myocarditis	KEGG	5.77E-06	1.79E-02
Metabolism of lipids and lipoproteins	REACTOME	7.33E-06	2.27E-02
HTLV-I infection	KEGG	1.13E-05	3.49E-02
Focal adhesion	KEGG	1.20E-05	3.73E-02
Focal Adhesion	WikiPathways	1.33E-05	4.14E-02
Asthma	KEGG	1.48E-05	4.60E-02
Valine, leucine and isoleucine degradation	KEGG	1.64E-05	5.07E-02
Allograft rejection	KEGG	1.89E-05	5.86E-02

MAP00280 Valine leucine and isoleucine degradation	GenMAPP	2.91E-05	9.01E-02
Developmental Biology	REACTOME	2.98E-05	9.26E-02
Pathways in cancer	KEGG	3.79E-05	1.18E-01
Interferon Signaling	REACTOME	3.93E-05	1.22E-01
Cytokine Signaling in Immune system	REACTOME	4.88E-05	1.51E-01
mitogen activated protein kinase signaling	Pathway Ontology	5.19E-05	1.61E-01

Table 2: Seven modules derived from the co-expression network

Moduleclasses	**Number of genes**	**Module genes (examples)**	**Biological functions**
Blue	351	IL10,TIMP1,FNTB, GATA6,FABP4…	Cell proliferation, metabolic, cell death
Brown	176	SP1,KLF11,NCOA6,KLF6, SRC,HLX,CASP8…	immune system development, viral process
Green	45	AXIN1,CAPN1,EWSR1	NA
grey	1308	MTG1,SPRY1,MDM2, ADAR,FABP1,MEN1…	cell death, apoptotic process, cell proliferation, metabolic
Red	33	OLFM1,HMGB2,TRADD, MYO1B,CAT,SPAST	catabolic process
Turquoise	435	RPS27A,PFN1,RPL23, UBA52,CCL5,USF1	immune response, viral process
Yellow	65	STK4,RSPO3,MITF, DAB2,CTNNB1,SMAD3	Wnt signaling pathway, cell death, apoptosis, immune system

Table 3: Genes significantly associated with survival time from module “yellow”

Modulegenes	**coef**	**exp(coef)**	**se(coef)**	**z**	**Pr(>|z|)**
SLC43A2	0.2446	1.2770	0.0617	3.9643	7.36E-05
FAU	0.5348	1.7072	0.1650	3.2416	0.0012
DAB2	-0.2460	0.7819	0.0783	-3.1410	0.0017
COL5A1	-0.2177	0.8043	0.0771	-2.8253	0.0047
ZCCHC2	-0.0891	0.9147	0.0323	-2.7572	0.0058
ISY1	0.1563	1.1692	0.0574	2.7245	0.0064
WDR1	-0.2144	0.8070	0.0845	-2.5382	0.0111
DOCK10	-0.1230	0.8843	0.0491	-2.5052	0.0122
C9orf142	0.2126	1.2369	0.0853	2.4931	0.0127
SH3BP4	-0.1579	0.8539	0.0648	-2.4363	0.0148
MRPS16	0.4509	1.5698	0.1882	2.3963	0.0166
ALDH2	-0.0784	0.9246	0.0367	-2.1359	0.0327
UBE2H	-0.2403	0.7864	0.1203	-1.9972	0.0458
MAEA	0.1293	1.1380	0.0692	1.8685	0.0617
CD58	-0.0660	0.9361	0.0358	-1.8440	0.0652
CITED2	-0.0691	0.9332	0.0402	-1.7214	0.0852
BNIP3L	-0.0776	0.9253	0.0462	-1.6804	0.0929

coef: coefficient