Weighted Gene Co-expression Network Analysis Reveals ALDH1B1 As a Prognositc Biomarker Involved in Gastric Cancer Development

Yi Li
The First Hospital of Lanzhou University
Ke Pu
The First Hospital of Lanzhou University
Yuping Wang
The First Hospital of Lanzhou University
Yongning Zhou (zhouynlzu@outlook.com)
The First Hospital of Lanzhou University

Research Article

Keywords: WGCNA, Gastric cancer, Hub gene, Prognostic biomarkers

Posted Date: November 15th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-962825/v1

Abstract

Background

Gastric cancer (GC) is one of the leading cancers associated with high mortality and poor prognosis mainly due to its relatively late diagnosis and the limited therapeutic options. Consequently, screening for prognostic GC biomarkers and novel molecular therapeutic targets is necessary to promote patient outcomes.

Methods

Weighted gene co-expression network analysis (WGCNA), a systems biology approach, was applied to analyze the mRNA sequencing data and clinical information of GC patients obtained from The Cancer Genome Atlas (TCGA). Gene modules and clinical traits were constructed according to the Pearson correlation analysis, and the gene ontology (GO) and functional enrichment analysis of meaningful modules were carried out. Hub genes from meaningful modules were screened out by two approaches: the intra-modular and protein-protein interaction (PPI) analysis methods. Next, through upstream regulatory analysis, hub genes with high connectivity degree were further validated with differential expression analysis, Kaplan-Meier survival analysis, and the Cox regression model.

Results

We found that seven modules were associated with the following clinical traits: anatomical location of gastric adenocarcinoma, histological type, histological grade, and pathological stage. The hub gene *ALDH1B1* was found to have potential as a biomarker for gastric cancer cells, the relationship between this hub gene and gastric cancer drug treatment is also worthy of attention.

Conclusion

These findings may contribute to understanding the GC tumourigenic mechanisms, as well as provide new potential prognostic factors and molecular therapeutic targets for GC. The *ALDH1B1* hub gene also provides a new vantage point for further clinical experiments and large-scale cohort studies to validate its association with GC patient survival, and provide a new direction for the research of gastric cancer drug treatment.

Background

Despite advances in cancer research, gastric cancer remains the fourth most common cancer and the second leading causes of cancer-related mortalities worldwide. [1] Over 950,000 new cases of GC are diagnosed annually, and an estimated 720,000 GC-related deaths were reported in 2012.[2] GC is a highly heterogeneous disease in terms of tumour architecture and growth, cell differentiation, histogenesis, and molecular pathogenesis. This diversity is reflected by the array of GC histopathological classification schemes.[3] For example, the Lauren classification [4] classifies GC tumours into intestinal and diffuse types. A biopsy sample of intestinal-type GC demonstrates well-demarcated, polypoid-glandular formation, with non-infiltrative proliferative stromal and tubular epithelium, resembling colorectal adenocarcinomas. The diffuse-type exhibits solitary and/or poorly cohesive infiltrative neoplastic which spread diffusely along the gastric wall and invaginate into muscularis propria and underlying lymphatics, with no gland formation.

The majority of new cases of GC present in an advanced and unresectable stage because asymptomatic patients were inadequately screened or managed for non-specific early-stage GC. Despite receiving the standard of care chemotherapy, these patients with advanced GC have a median survival time less than one year. This may be explained by the limited therapeutic efficacy of chemotherapy for late-stage GC. [5] [6]

To address the limited therapeutic options, molecular-targeted therapies for advanced GC have been introduced clinically. Up to 20% of GC biopsies overexpress human epidermal growth factor receptor 2 (HER2) due to *HER2* gene amplification within malignant cells[7]. Furthermore, HER2 expression is correlated to the grading of stages of tumour differentiation. Trastuzumab, a humanized monoclonal antibody directed against HER2, has been demonstrated to improve overall survival in advanced GC patients and is now approved for first-line treatment of advanced HER2-positive gastric adenocarcinomas.

Anti-angiogenic agents have also been investigated as treatments for advanced GC, as vascular endothelial growth factor (VEGF) expression is correlated to tumour angiogenic microenvironment, tumour size, and metastatic dissemination. Patients receiving a combination of capecitabine, cisplatin, and bevacizumab, a monoclonal antibody directed against VEGF, did not demonstrate greater survival compared to the placebo, capecitabine-cisplatin, or bevacizumab-only groups in the AVAGAST trial[8]. Given the poor long-term therapeutic outcome associated with chemotherapy and scarce molecular-targeted therapy for advanced GC, developing and trialing other molecular-targeted agents is warranted to improve prognosis and increase overall survival rate. Discovering novel candidate genes involved in tumorigenesis may pave the way for novel molecular-targeted therapies.

While the weighted gene co-expression network analysis (WGCNA) approach is widely used in tumourigenesis research, it has not yet been widely applied to the GC-derived dataset documented in The Cancer Genome Atlas (TCGA). The TCGA's large-scale database, detailing gene sequencing data and clinicopathological information, enables systematic analysis of molecular mechanisms underlying the development and progression of various clinical features associated with cancers. [9]

WGCNA is a genome-wide analysis method used to construct gene co-expression networks based on inter-sample similarities in expression profiles. It is capable of ascribing genes with function and infer gene-disease associations. Highly co-expressed genes are interconnected in the network. Among these, similar genes can be sub-classified into different co-expression modules based on the different clinical traits. Each co-expression modules consists of functionally related genes that are also involved in solitary functions. Hub genes are identified through the most central and connected genes in the modules. These modules and their candidate genes may be associated with the biological processes of GC, such as tumorigenesis, pathogenesis, tumour progression,

neoplasm invasion, and metastasis. Thus, the significant module genes have the potential to have widespread clinical implications as either potential warning signatures or therapeutic targets.

In this study, WGCNA is applied to screen out key biomarkers associated with GC clinical features based on the correlation analysis between mRNA data of GC patient samples and clinical information acquired from the TCGA database. These information modules may lead to new prognostic markers or therapeutic targets.

Methods

Gene expression data and pre-processing

RNA sequencing data sets of GC and the corresponding clinical traits were obtained from the TCGA database (May 3, 2018) (http://portal.gdc.cancer.gov). The sequencing data of mRNA expression profiles was derived from the platform of Illumina HiSeq V2 RSEM genes. Gene expression level was calculated as measurement of fragments per kilobase of transcript per million mapped reads (FPKM). Related clinical information was extracted for WGCNA analysis, which included sex, age at initial pathological diagnosis, anatomical location, AJGG clinical TNM staging (clinical stage, Tumor, Lymph Node and Metastasis), pathological stage (pathological stage I, II, III and IV), histological grade (grade 1, 2, 3 and 4), and histological type.

According to the Lauren classification, GC is categorized into two main histological types, diffuse and intestinal, in addition to the mixed and indeterminate types. [4] In the WHO classification, GC is subdivided into tubular, papillary, mucinous, poorly cohesive, or rare variants, based on the predominant histological patterns of the carcinoma. [10] The tubular and papillary carcinomas corresponds closely to the Lauren classification intestinal-type. Similarly, the poorly cohesive carcinomas (encompassing cases constituted partly or totally by signet ring cells) corresponds closely to the Lauren classification diffuse-type.

Gene co-expression network construction analysis

Co-expression network were conducted based on the protocol of WGCNA package in R software[11, 12]. Firstly, the similarity matrix between the gene expression profile was constructed according to the pairwise Pearson's correlation coefficients, which reflect the degree of consistency between gene expression profiles. Next, we used the function power adjacency to transform the similarity matrix into an adjacency matrix. The adjacency matrix was used to measure the connectivity strengths between pairwise genes.

Scale-free gene co-expression networks were constructed using the function power. [11] To validate the results of network construction as being reliable, outliers with connectivity values of less than -2.5 were excluded[12].When a scale-free topology index R^2 calculated by function PickSoftThreshold is greater than 0.85, it is suggestive of the network topology approximating to scale-free status, and the corresponding power value is selected. Subsequently, the adjacency matrix was transformed into a Topological Overlap Matrix (TOM) and the corresponding dissimilarity TOM (dissTOM). The topological overlap is a biologically meaningful measure of gene similarity on basis of co-expression relationships between paired genes.[13]

To classify the genes with similar expression profiles into the same module, module classification was conducted with the dynamic branch-cut method to generate modules according to the dissTOM-based hierarchical clustering. During the process, a deep-split value of 2 to branch splitting and a minimum-size cut-off of 30 (minClusterSize=30) for the module size were selected to prevent abnormal modules from being produced.

Constructing module-trait relationships of Gastric Cancer

The module eigengene (ME) as the dominant component of the selected module was evaluated by function ME. It represents the gene expression profiles of a given module and captures the maximum variation in the module. If the MEs' correlation of modules were greater than 0.75, the modules would show the similar expression profiles and be integrated together. The correlation between MEs and clinical traits were assessed by Pearson's correlation tests, and $p < 0.05$ was deemed strongly correlational.

Finding meaningful modules and functional annotation

The correlation between modules and corresponding clinical traits was estimated by using Pearson's correlation test. Modules found to be closely correlated to clinical traits were identified as biologically meaningful modules. However, the mechanism of how module genes influenced the related clinical-traits remained unknown, so all genes in the meaningful modules were mapped onto the DAVID database for Gene Ontology (GO) analysis and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis.[14] A $P<0.01$ and false discovery rate (FDR) <0.01 were set as the cut-off criteria to screen the annotation information.

Hub genes identification and association analysis

Module hub genes are highly connected intra-modular genes, which have the highest module membership (MM) scores (Pearson's correlation >0.8) to every module. MM is a virtual gene which describes the relationship between the ME and the gene expression profile. As well, MM quantifies an associated degree to describe how close a gene is to the corresponding module. Gene significance (GS) parameter is introduced to weigh the correlation (Pearson's correlation >0.2) between a hub gene and clinical significance. Therefore, based on GS and MM, genes having a significance for a clinical feature and module membership can be identified for intra-modular analysis.[12] Representative network plots depicting the modules were constructed using Cytoscape software. Node centrality, assessed by node degree and number connections, identifies functionally hub genes. Moreover, protein-protein interaction (PPI) information is assessed by the online tool, search tool for the retrieval of interacting genes (STRING)[15]. The software, STRING, was used to explore any relationship of PPI networks among the meaningful modules. Meaningful modules genes were uploaded onto the database and a plausible PPI network was constructed with chosen confidence

>0.4 and was imported into Cytoscape software. In the PPI network, genes having a connectivity value of ≥10 were defined as hub genes. Hub genes being in co-expression network and in PPI network were regarded as common hub genes.

Hub gene with upstream regulatory network analysis

To explore the regulatory mechanisms at the transcriptional level, we investigated the hub genes as targets. First, we used the comprehensive experimentally validated miRNA-gene interaction data collected from TarBase and miRTarBase database to identifying upstream miRNAs of the hub genes. The ENCODE ChIP-seq database was then used to examine their upstream TFs. The networks were analyzed by NetworkAnalyst online tool and constructed by using Cytoscape to search for TFs and for miRNAs.

Hub genes validation

The gene expression profiling interactive analysis (GEPIA) online tool (http://gepia.cancer-pku.cn)[16] was used to perform the differential expression of hub gene. The GEPIA was a newly developed interactive web server that uses a standard processing pipeline to analyze RNA sequencing expression data referenced from TCGA and GTEx Projects. Furthermore, survival analyses were performed with another online database (http://kmplot.com) which provided published microarray datasets for four cancer types besides GC, and included clinical data and gene expression information of 1065 GC patients' datasets. A Kaplan Meier survival plot was generated and a hazard ratio with 95% confidence intervals and log-rank *p* values were calculated and plotted in R, by using Bioconductor packages[17].

Statistical analysis

The data were analyzed using stata14.0 software. Continuous variables were assessed by student's t-test, while group comparisons of count data were assessed by chi-square test. The univariate and multivariate Cox proportional hazards regression models were evaluated for estimating the independent prognostic values. A $P<0.05$ was deemed statistically significance.

Results

Gene co-expression network of GC

A flow chart of this study is shown in Figure 1. The RNA sequencing data of GC from 375 patients downloaded from TCGA database and the expression values of 56831 genes,were applied to construct the co-expression network. Cluster analysis was performed on these samples using the flashClust package. After outliers were removed, the soft threshold power value was sat to four (Figure 2), to define the adjacency matrix between all genes with power function, according to the standard scale-free network distribution.

Forty-six gene co-expression modules, clustered in size from 31 to 2474 genes, were partitioned by hierarchical clustering and by dynamic tree cutting (Figure 3A). The unique colour identifier for each module is listed in Additional Table 1. Among these modules, three were merged because of their similar MEs, and forty-three modules were identified in total. The gray module represented a gene set that was not assigned to any of the modules. In total, the interaction of the forty-two co-expression modules were analyzed (Figure 3B).

Table 1
The correlation of three genes (*FAM83D, ITGAL* and *ALDH1B1*) expression and clinicpathological features of GC patients

Parameters	*FAM83D*				*ITGAL*				*ADLH1B1*		
	Case	Low expression	High expression	*P* value	Case	Low expression	High expression	*P* value	Case	Low expression	High expression
Age (Mean± SD)	415	65.85±10.67	65.31±10.60	0.61	410	65.69±10.84	65.55±10.57	0.89	410	65.58±10.62	65.65±10.79
Sex											
Female	147	74	73	0.38	147	60	87	0.09	147	79	68
Male	268	147	121		263	130	133		263	149	114
Pathological T											
I-II	110	61	49	0.59	107	61	46	**0.01**	107	62	45
III-IV	305	160	145		303	129	174		303	166	137
Pathological N											
N0	123	64	59	0.72	121	60	61	0.39	121	58	63
N1-3	292	158	134		289	130	159		289	170	119
Pathological M											
M0	367	190	177	0.10	362	167	195	0.82	362	193	169
M1-3	48	31	17		48	23	25		48	35	13
Pathological stages											
I-II	186	103	83	0.44	184	94	90	0.08	184	96	88
III-IV	229	118	111		226	96	130		226	132	94
Histological type											
Intestinal	176	88	88	0.61	173	98	75	**<0.001**	174	101	73
Diffuse	239	133	106		236	91	145		236	126	110
Histological grade											
G1-G2	160	84	76	0.76	159	104	55	**<0.001**	159	92	67
G3-GX	255	137	118		251	86	165		251	136	115
Anatomical subdivision				**0.02**				0.48			
Cardia/GEJ	97	58	39		95	50	45		95	66	29
Fundus	143	82	61		147	62	85		147	80	67
Antrum	156	70	86		157	73	84		157	75	82
others	19	11	8		11	5	6		11	7	4
Living status											
Yes	252	147	105	**0.008**	251	121	130	0.34	251	140	111
No	163	74	89		159	69	90		159	88	71

Abbreviation: GEJ: gastroesophageal junction.

As shown in the network heatmap plot (Additional Figure. 1), each module represented individual validation to one another. To further explore the co-expression similarity of the whole modules, MEs were quantified and were clustered according to modules' correlation (Figure. 4A). These 42 modules were classified into four main cluster sets. Within these cluster sets, the big clustering branch included 26 modules that could be subdivided into 12 sub-clusters.

The remaining three clusters were subdivided into 16 other modules. The left and right branch of the clustering tree were both also divided into five sub-clusters. This result was consistent with the heatmap plot of the adjacencies (Figure. 4B).

Modules correspond to clinical significance

To investigate the corresponding clinical feature of the module, the Pearson correlations were analyzed between MEs and clinical traits, including: sex, age, clinical TNM stage, histological type, anatomical subdivision of neoplasm, histological grade and pathological stage (Figure 5). Eight modules were positively or negatively correlated with the defined clinical traits. The relative highest association with the module-feature relationship, was between the tan module and the neoplasm of anatomical location (at the gastroesophageal junction, GEJ) in histological region. The second significant association was between turquoise, violet modules, and the histopathological differential features of diffuse-type, histological grade G3, and pathological stage Ia.

A scatter plot of gene significance (GS) versus module membership (MM) were produced with select modules that demonstrated high module-feature associations. (Additional Figure 2). GS and MM were that demonstrated high correlation include the tan, turquoise, purple modules. The black modules had a relatively weaker GS-MM correlation and were consequently excluded from the functional enrichment analysis.

Functional enrichment analysis of genes in meaningful modules

Biological significance of selected modules was assessed using GO term function analysis, specifically: biological process, cellular component and molecular function, and KEGG pathway enrichment analysis. All genes in modules of interest were imported to DAVID software and string-db online analysis.

Light-green module genes were from the nuclear component and involved in RNA binding and poly (A) RNA binding, and were correlated to RNA splicing and RNA processing (Additional Table 2). Enrichment analysis showed that the tan module was involved mainly in epidermis tissue development and cell-cell junction during GC progression. Coding proteins from the tan module genes were also mostly correlated to the composition of extracellular exosome and zona adherent junction. Grey60 module genes were involved in cellular transcription, their coding proteins were the component of nucleus and nucleoplasm.

In modules of histological type, the turquoise, orange and salmon modules underwent enrichment analysis. The steel-blue module was excluded from this analysis. The turquoise module was associated with oxidation-reduction process and cell-cell adhesion, which were enriched in metabolic pathways. The orange module genes were correlated to angiogenesis. The salmon module genes were shown to play a role in exerting regulatory roles in inflammatory and immune responses. The coding proteins of these genes were the component of extracellular exosome and lysosomal lumen, the pathway was therefore involved with lysosome and phagosome.

The histological grade and pathological stage were also assessed in the modules. Here, the violet modules were excluded from enrichment analysis due to insufficient enrichment results. The purple module genes were from cytosolic transduction signal molecules, which were involved in the positive regulation of GTPase activity and protein binding, and were mainly enriched in immune response such as leucocyte migration. The pathway of this module was mainly regulated via NF-κB signaling pathway and human T-lymphotropic virus I (HTLV-I) infection.

Identification of hub genes from meaningful modules

The role of hub genes in modules suggest the biological significance of these modules. High MM value determines which genes were in network center and played essential biological function in whole network. To identify central nodes of these modules, intramodular analysis were conducted in seven modules including grey60, light-green, orange, purple, salmon, tan, and turquoise modules. From this, 293 genes with high intramodular connectivity value were identified as hub genes (Figure 6 and Additional Table 3). Another analytical approach, PPI network analysis, was introduced to analyze module genes. Hub gene criteria was set as the threshold value of confidence >0.4 and connectivity degree of ≥10. A total of 140 genes with high connectivity degree were obtained from six modules analyzed (the grey60, light-green, orange, purple, salmon and tan modules) (Additional Table 3). Turquoise modules were not assessed by String software, the number of genes exceeded the threshold setting of gene capacity. Finally, 47 intersectional hub genes from two analytical methods were collated, which were defined as common hub genes.

Common hub gene upstream regulator analysis

To investigate the upstream regulators of common hub genes, the miRNAs and TFs were predicted with the two databases of TarBase and miRTarBase[18]. High connectivity value ≥10 was set as a criterion to screen for the common hub genes. As shown in Additional Table 4, predicted miRNAs of four modular genes were constructed, while grey60 and tan modules were excluded due to low connectivity of modular genes. To further understand the regulatory mechanism between TFs and common hub genes, TF-gene networks of each module were constructed by an online tool, NetworkAnalyst[18], (Additional Table 5). Finally, 16 common hub genes with high connectivity degree with miRNAs and TFs from different modules, were selected as GC biomarkers.

Identification and validation of common hub genes

To validate the 16 common hub genes of modules further, differential expression and survival analyses were performed by GEPIA online tool and KM plotter database separately. Five common hub genes, *ITGAL, CD86, ALDH1B1, FAM83D* and *HSPB6* had significantly different expression (Figure 7A). Survival analysis showed *ITGAL, ALDH1B1* and *FAM83D* were correlated with better prognostic outcomes for GC patients (Figure 7B). The genes *CD86* and *HSPB6* were not included in the analysis because their abnormal regulation implied a longer survival time. The clinicopathological parameters of three common hub genes were analyzed in Table 1. Compared to *ITGAL* and *ALDH1B1*, high *FAM83D* expression was only associated with the living status of GC patients (p = 0.002).

Discussion

The progression and prognosis of GC in different patients remains variable. Further studies in biomarkers with improved prognostic or predictive value are needed to investigate GC molecular pathogenesis mechanisms as this can help lead to more statistically accurate conclusions and provide clinical information for best evidence-based practice. Here, we used WGCNA analysis to screen biomarkers associated with GC progression and prognosis. WGCNA has proven to have significant advantages over other methods, such as focusing on the association between co-expression modules and clinic traits, which can produce results with much higher reliability and biological significance. [19]

The characteristics of clinical traits-associated modules was changed variously followed by the different clinical traits. Seven modules that were positively correlated with GC clinical traits were selected with the Pearson correlation and *p* value. Related to the anatomical location of GC, functional enrichment analyses showed that three modules genes, light-green, tan, grey60, were involved in RNA splicing or RNA processing, cell-cell adhesion, epidermis tissue development, and cellular transcription. RNA alternative splicing is often implicated in tumourigenic progression, including cancer initiation and metastasis. Aberrant patterns of splicing have been shown to lead to transcriptome instability, growth-inhibiting signals resistance, uncontrolled metabolic processes, and dysregulated cellular phenotype, all of which contribute to cancerous growth[20–23]. Moreover, cell-cell adhesion plays a pivotal role in metastasis and invasion as the lost of adhesive molecules increases the infiltrative and metastatic potential of tumor cells. [24] Cellular transcription regulation, including coding RNA and non-coding RNAs, has also been closely correlated to cancer progression and prognosis. Although the above biological processes were associated with GC oncogenesis, no reported studies thus far have unveiled the relationship between the aberrant biological processes and GC anatomical location.

Turquoise, orange, and salmon modules, represented the three histological subtypes, diffuse, mucious and tubular type. These modules were shown to be independently in cell adhesion, angiogenesis and inflammatory response. The turquoise module was also found to be associated with patients' age. Similarly, the salmon module was also implicated in the size or expansive scope of tumors. Diffuse-type GC has a well-delineated pathogenesis wherein most diffuse-type GC has been shown to be driven primarily by defective intercellular adhesions caused by the loss of E-cadherin expression[25]. About 50% of the diffuse-type GC and lymph-node metastases is associated with somatic mutations within the E-cadherin gene; this is not the case in intestinal-type GC. [26] Consequently, abnormal E-cadherin variants is closely related to the diffuse-type GC. However, no direct proof exists to elucidate how cell-cell adhesion mediates tumourigenesis of GC in patients with increasing median age. Studies have described the role of angiogenesis, but there is no consensus on the molecular subtyping of GC, further investigation on whether the mucinous gastric adenocarcinoma is closely associated with the angiogenesis molecules will be necessary. Inflammatory responses facilitate GC progression through several mechanisms. First, the activation of the COX2/PGE-2 signaling appears to be among the major pathways.[27] The innate immune response also mediates carcinogenesis through the TLR/MyD88 adapter signaling[28, 29]. There also appears to be crosstalk between tumour cell-derived and macrophage-derived inflammatory factors during *H. pylori* infection which promote the inflammatory microenvironment disorder that leads to the acquisition of tumour stem cell properties and progression into gastric epithelium. While many studies have investigated how inflammation plays a role in GC, limited studies have examined the inflammatory changes that occur in tubular-type GC.

The histological grades were associated with the turquoise and purple modules. Cell-cell adhesive loss of turquoise module genes were involved in inhibiting apoptosis and invasion of GC cells, thereby providing opportunities for tumour cell expansion and metastasis[30–32]. Adhesive molecules played a crucial role in histological malignancy grading, but which histological grade the module genes being to could not be predicted definitely. Purple module involved components of signal transduction in NF-κB signaling. The NF-κB signaling pathway, has been shown to be upregulated in many steps of tumorigenesis. For example, *H. pylori* infection in gastric epithelial cells in GC induces activation and nuclear translocation of NF-κB [33]. Consequently, HuR, a direct transcript target of NF-κB, is activated in GC cell lines, exerting proliferative and anti-apoptotic effects on GC[34]. However, there is no significant link between NF-κB and pathological grade.

Among these meaningful GC progression-associated modules, high connectivity value of module genes from intramodular analysis and PPI analysis were selected out as common hub genes. Meanwhile, the upstream regulatory networks of common hub genes also were investigated. Next, differential analyses and survival analyses led to the selection of three common hub genes, *ITGAL*, *ALDH1B1*, *FAM83D*, from the turquoise module. Further analysis and literature review illustrated that the hub gene *ALDH1B1* have potential as a biomarker for gastric cancer cells.

ALDH1B1 is a member of the ALDH family and one of the isoenzymes of the aldehyde dehydrogenase 1(ALDH1), and the highly ALDH1 activity has been determined in a variety of cancer stem cells including gastric cancer[35–37]. On the basis of previous studies on ALDH1,many researchers have conducted research on the ALDH1 isoenzymes. Kai Li and Jia-Xin Shen et al both found that the high level of *ALDH1B1* was associated with better overall survival in GC patients in their[38, 39].And this is consistent with our study.

In addition,ALDH can also play a role as a functional regulators in cancer stem cells,which is also worthy of attention[40].First, ALDH is an essential component for the biosynthesis of retinoic acid (RA) and other molecular regulators of cellular function.In this way,ALDH can promote drug resistance through retinoid signaling. Second, Cancer cells with high expression level of ALDH can develop drug resistance of the ALDH-specific activity catabolyzes particular drugs through oxidation of the specific aldehyde of the drugs,such as cyclophosphamide and its active derivative hydroperoxycyclophosphamide (4-HC)[41, 42],doxorubicin[43], cisplatin [44], arabinofuranosyl cytidine (Ara-C)[45], and dacarbazine[46].As a member of the ALDH family, we believe that *ALDH1B1* may have similar functions and these studies will provide us with directions for further research on *ALDH1B1*.

Several limitations of our study should be taken into account. Although the most pivotal hub genes out of 16 common hub genes was filtered out by bioinformatics methods, inadequate persuasive experimental evidence exist to support our findings' impact on the prognostic value of biomarker in GC patients.

In summary, WGCNA, a systems biology approach, was adopted and used to analyze RNAseq data and clinical information of GC patients in TCGA database. *ALDH1B1*, which has been previously suggested to play a crucial molecule in GC development and progression, was as a possible new GC prognostic marker and therapeutic target. Nevertheless, this biomarker needs to be further validated with basic experiments and large-scale cohort studies.

Declarations

Ethics approval, guidelines and consent to participate:

Not applicable

Consent for publication

Not applicable

Availability of data and materials

The basic used in this study can be obtained from the TCGA(https://www.genome.gov/Funded-Programs-Projects/Cancer-Genome-Atlas) database, and other data have been included in the manuscript.And we declare that all methods were carried out in accordance with relevant guidelines and regulations.

Competing interests

The author reports no conflicts of interest in this work.

Funding

This research was funded by the National Natural Science Foundation of China(Grant No. 71964021), the National Key R&D Program of China(2016YFC1302201), the Fundamental Research Funds for the Central Universities(lzujbky-2021-kb35), the Open Foundation of Key Laboratory of Biotherapy and Regenerative Medicine of Gansu Province(zdsyskfkt-201705).

Authors' contributions

YL, KP: study concept and design, title, abstract, full-text screening, data abstraction, statistical analysis, data interpretation, drafting the article a

YPW: data interpretation, revision of the manuscript

YNZ: critical revision of the manuscript

Acknowledgements

I would like to thank all co-authors that were involved in this study. We are especially appreciative of the support from ProfessorsYongning Zhou.

References

1. Jemal A, Bray F, Center MM, Ferlay J, Ward E, Forman D: **Global Cancer Statistics**. *Ca-Cancer J Clin* 2011, **61**(2):69–90.
2. Ferlay J, Steliarova-Foucher E, Lortet-Tieulent J, Rosso S, Coebergh J-WW, Comber H, Forman D, Bray F: **Cancer incidence and mortality patterns in Europe: estimates for 40 countries in 2012**. *European journal of cancer* 2013, **49**(6):1374–1403.
3. Van Cutsem E, Sagaert X, Topal B, Haustermans K, Prenen H: **Gastric cancer**. *Lancet* 2016, **388**(10060):2654–2664.
4. Lauren P: **The two histological main types of gastric carcinoma: diffuse and so-called intestinal-type carcinoma. An attempt at a histo-clinical classification**. *Acta pathologica et microbiologica Scandinavica* 1965, **64**:31–49.
5. Cervantes A, Roda D, Tarazona N, Rosello S, Perez-Fidalgo JA: **Current questions for the treatment of advanced gastric cancer**. *Cancer Treat Rev* 2013, **39**(1):60–67.
6. Group G: **Role of chemotherapy for advanced/recurrent gastric cancer: an individual-patient-data meta-analysis**. *European journal of cancer* 2013, **49**(7):1565–1577.
7. Van Cutsem E, Bang YJ, Feng-Yi F, Xu JM, Lee KW, Jiao SC, Chong JL, Lopez-Sanchez RI, Price T, Gladkov O *et al*: **HER2 screening data from ToGA: targeting HER2 in gastric and gastroesophageal junction cancer**. *Gastric Cancer* 2015, **18**(3):476–484.
8. Ohtsu A, Shah MA, Van Cutsem E, Rha SY, Sawaki A, Park SR, Lim HY, Yamada Y, Wu J, Langer B *et al*: **Bevacizumab in combination with chemotherapy as first-line therapy in advanced gastric cancer: a randomized, double-blind, placebo-controlled phase III study**. *J Clin Oncol* 2011, **29**(30):3968–3976.
9. Tomczak K, Czerwinska P, Wiznerowicz M: **The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge**. *Contemp Oncol (Pozn)* 2015, **19**(1A):A68-77.
10. Bosman FT, Carneiro F, Hruban RH, Theise ND: **WHO classification of tumours of the digestive system**: World Health Organization; 2010.
11. Langfelder P, Horvath S: **WGCNA: an R package for weighted correlation network analysis**. *BMC Bioinformatics* 2008, **9**:559.
12. P M, S H, R M, M G, W SK: **Adult mesenchymal stem cells and cell surface characterization - a systematic review of the literature**. *Open Orthop J* 2011, **5**(Suppl 2):253–260.
13. Zhang B, Horvath S: **A general framework for weighted gene co-expression network analysis**. *Stat Appl Genet Mol Biol* 2005, **4**:Article17.

14. Huang DW, Sherman BT, Lempicki RA: **Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources**. *Nat Protoc* 2009, **4**(1):44–57.
15. Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, Simonovic M, Roth A, Santos A, Tsafou KP *et al*: **STRING v10: protein-protein interaction networks, integrated over the tree of life**. *Nucleic Acids Res* 2015, **43**(Database issue):D447-452.
16. Tang Z, Li C, Kang B, Gao G, Li C, Zhang Z: **GEPIA: a web server for cancer and normal gene expression profiling and interactive analyses**. *Nucleic Acids Res* 2017, **45**(W1):W98-W102.
17. Szász AM, Lánczky A, Nagy Á, Förster S, Hark K, Green JE, Boussioutas A, Busuttil R, Szabó A, Győrffy B: **Cross-validation of survival associated biomarkers in gastric cancer using transcriptomic data of 1,065 patients**. *Oncotarget* 2016, **7**(31):49322–49333.
18. Xia J, Gill EE, Hancock RE: **NetworkAnalyst for statistical, visual and network-based meta-analysis of gene expression data**. *Nat Protoc* 2015, **10**(6):823–844.
19. Chou WC, Cheng AL, Brotto M, Chuang CY: **Visual gene-network analysis reveals the cancer gene co-expression in human endometrial cancer**. *BMC Genomics* 2014, **15**:300.
20. David CJ, Manley JL: **Alternative pre-mRNA splicing regulation in cancer: pathways and programs unhinged**. *Genes Dev* 2010, **24**(21):2343–2364.
21. Lee SC, Abdel-Wahab O: **Therapeutic targeting of splicing in cancer**. *Nat Med* 2016, **22**(9):976–986.
22. Oltean S, Bates DO: **Hallmarks of alternative splicing in cancer**. *Oncogene* 2014, **33**(46):5311–5318.
23. Warzecha CC, Carstens RP: **Complex changes in alternative pre-mRNA splicing play a central role in the epithelial-to-mesenchymal transition (EMT)**. *Semin Cancer Biol* 2012, **22**(5-6):417–427.
24. Birchmeier W, Behrens J: **Cadherin expression in carcinomas: role in the formation of cell junctions and the prevention of invasiveness**. *Biochim Biophys Acta* 1994, **1198**(1):11–26.
25. Barber M, Murrell A, Ito Y, Maia AT, Hyland S, Oliveira C, Save V, Carneiro F, Paterson AL, Grehan N *et al*: **Mechanisms and sequelae of E-cadherin silencing in hereditary diffuse gastric cancer**. *J Pathol* 2008, **216**(3):295–306.
26. Becker KF, Atkinson MJ, Reich U, Becker I, Nekarda H, Siewert JR, Hofler H: **E-cadherin gene mutations provide clues to diffuse type gastric carcinomas**. *Cancer Res* 1994, **54**(14):3845–3852.
27. Grivennikov SI, Greten FR, Karin M: **Immunity, inflammation, and cancer**. *Cell* 2010, **140**(6):883–899.
28. Maeda Y, Echizen K, Oshima H, Yu L, Sakulsak N, Hirose O, Yamada Y, Taniguchi T, Jenkins BJ, Saya H *et al*: **Myeloid Differentiation Factor 88 Signaling in Bone Marrow-Derived Cells Promotes Gastric Tumorigenesis by Generation of Inflammatory Microenvironment**. *Cancer Prev Res (Phila)* 2016, **9**(3):253–263.
29. Pradere JP, Dapito DH, Schwabe RF: **The Yin and Yang of Toll-like receptors in cancer**. *Oncogene* 2014, **33**(27):3485–3495.
30. Metcalfe A, Streuli C: **Epithelial apoptosis**. *Bioessays* 1997, **19**(8):711–720.
31. Kantak SS, Kramer RH: **E-cadherin regulates anchorage-independent growth and survival in oral squamous cell carcinoma cells**. *J Biol Chem* 1998, **273**(27):16953–16961.
32. Day ML, Zhao X, Vallorosi CJ, Putzi M, Powell CT, Lin C, Day KC: **E-cadherin mediates aggregation-dependent survival of prostate and mammary epithelial cells through the retinoblastoma cell cycle control pathway**. *J Biol Chem* 1999, **274**(14):9656–9664.
33. Sharma SA, Tummuru MK, Blaser MJ, Kerr LD: **Activation of IL-8 gene expression by Helicobacter pylori is regulated by transcription factor nuclear factor-κB in gastric epithelial cells**. *The Journal of Immunology* 1998, **160**(5):2401–2407.
34. Kang MJ, Ryu BK, Lee MG, Han J, Lee JH, Ha TK, Byun DS, Chae KS, Lee BH, Chun HS *et al*: **NF-kappaB activates transcription of the RNA-binding factor HuR, via PI3K-AKT signaling, to promote gastric tumorigenesis**. *Gastroenterology* 2008, **135**(6):2030–2042, 2042 e2031-2033.
35. Ehlers CL: **Variations in ADH and ALDH in Southwest California Indians**. *Alcohol Res Health* 2007, **30**(1):14–17.
36. Douville J, Beaulieu R, Balicki D: **ALDH1 as a functional marker of cancer stem and progenitor cells**. *Stem Cells Dev* 2009, **18**(1):17–25.
37. Ma I, Allan AL: **The role of human aldehyde dehydrogenase in normal and cancer stem cells**. *Stem Cell Rev Rep* 2011, **7**(2):292–306.
38. Li K, Guo X, Wang Z, Li X, Bu Y, Bai X, Zheng L, Huang Y: **The prognostic roles of ALDH1 isoenzymes in gastric cancer**. *Onco Targets Ther* 2016, **9**:3405–3414.
39. Shen JX, Liu J, Li GW, Huang YT, Wu HT: **Mining distinct aldehyde dehydrogenase 1 (ALDH1) isoenzymes in gastric cancer**. *Oncotarget* 2016, **7**(18):25340–25349.
40. Vassalli G: **Aldehyde Dehydrogenases: Not Just Markers, but Functional Regulators of Stem Cells**. *Stem Cells Int* 2019, **2019**:3904645.
41. Hilton J: **Role of aldehyde dehydrogenase in cyclophosphamide-resistant L1210 leukemia**. *Cancer Res* 1984, **44**(11):5156–5160.
42. Sládek NE, Kollander R, Sreerama L, Kiang DT: **Cellular levels of aldehyde dehydrogenases (ALDH1A1 and ALDH3A1) as predictors of therapeutic responses to cyclophosphamide-based chemotherapy of breast cancer: a retrospective study. Rational individualization of oxazaphosphorine-based cancer chemotherapeutic regimens**. *Cancer Chemother Pharmacol* 2002, **49**(4):309–321.
43. Honoki K, Fujii H, Kubo A, Kido A, Mori T, Tanaka Y, Tsujiuchi T: **Possible involvement of stem-like populations with elevated ALDH1 in sarcomas for chemotherapeutic drug resistance**. *Oncol Rep* 2010, **24**(2):501–505.
44. Kawasoe M, Yamamoto Y, Okawa K, Funato T, Takeda M, Hara T, Tsurumi H, Moriwaki H, Arioka Y, Takemura M *et al*: **Acquired resistance of leukemic cells to AraC is associated with the upregulation of aldehyde dehydrogenase 1 family member A2**. *Exp Hematol* 2013, **41**(7):597-603.e592.
45. Magni M, Shammah S, Schiró R, Mellado W, Dalla-Favera R, Gianni AM: **Induction of cyclophosphamide-resistance by aldehyde-dehydrogenase gene transfer**. *Blood* 1996, **87**(3):1097–1103.

46. Prasmickaite L, Engesaeter B, Skrbo N, Hellenes T, Kristian A, Oliver NK, Suo Z, Maelandsmo GM: **Aldehyde dehydrogenase (ALDH) activity does not select for cells with enhanced aggressive properties in malignant melanoma**. *Plos One* 2010, **5**(5):e10731.

Figures

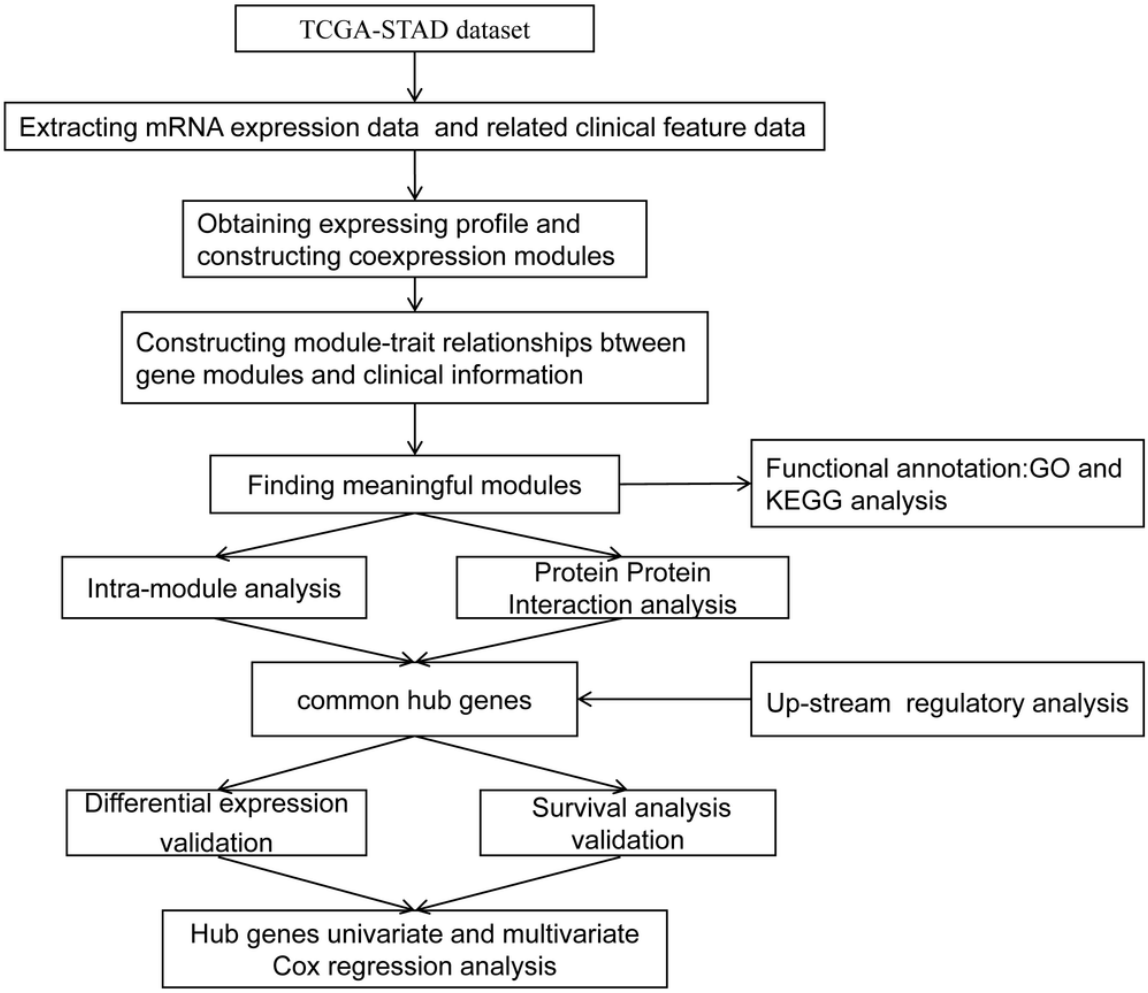

Figure 1

Flow chart of data preparation, processing, analysis and validation in this study

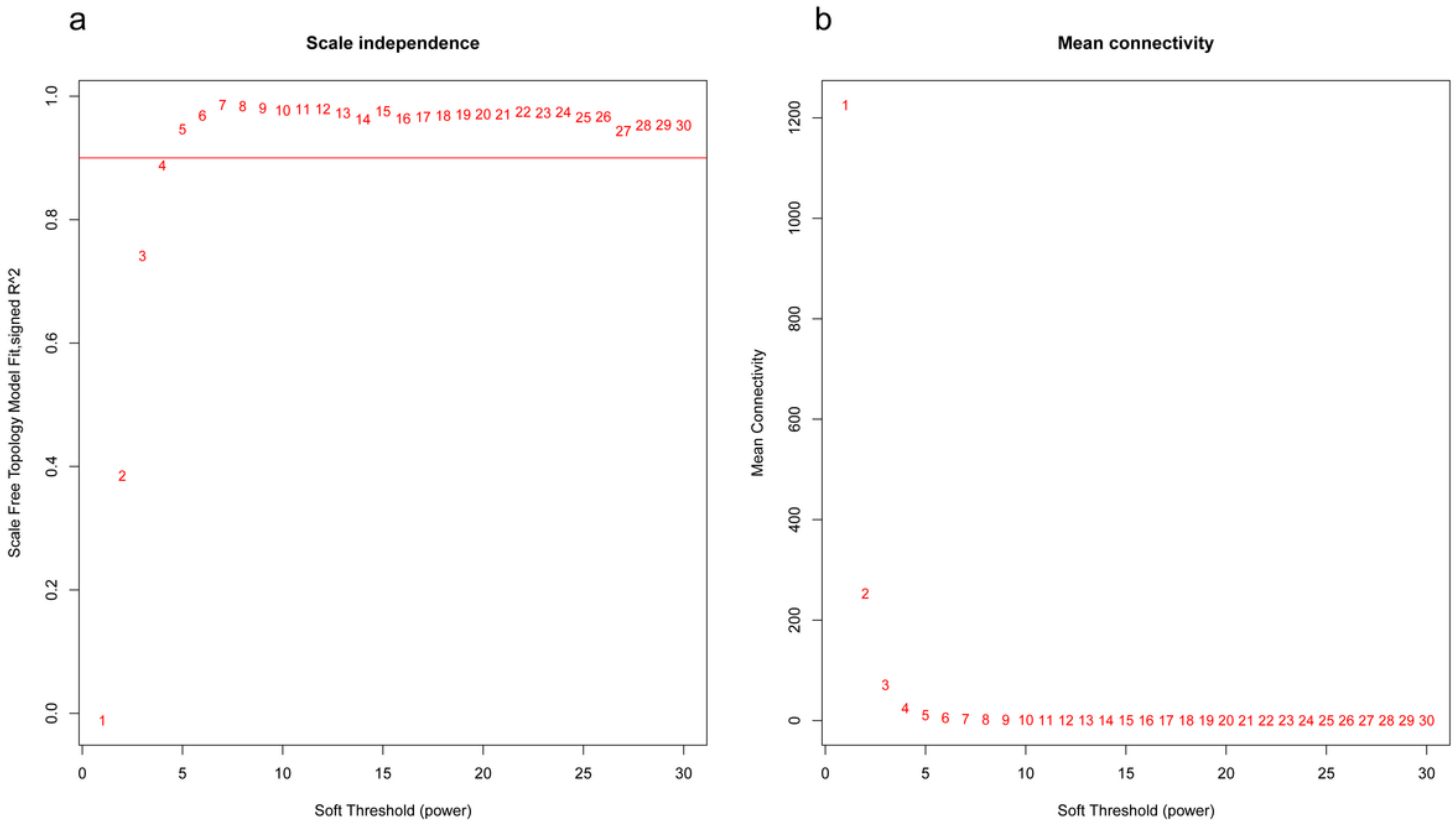

Figure 2

Analysis of network topology for various softthresholding powers. (A) Analysis of the scale-free fit index for various soft-thresholding powers (β). (B) Analysis of the mean connectivity for various soft-thresholding powers

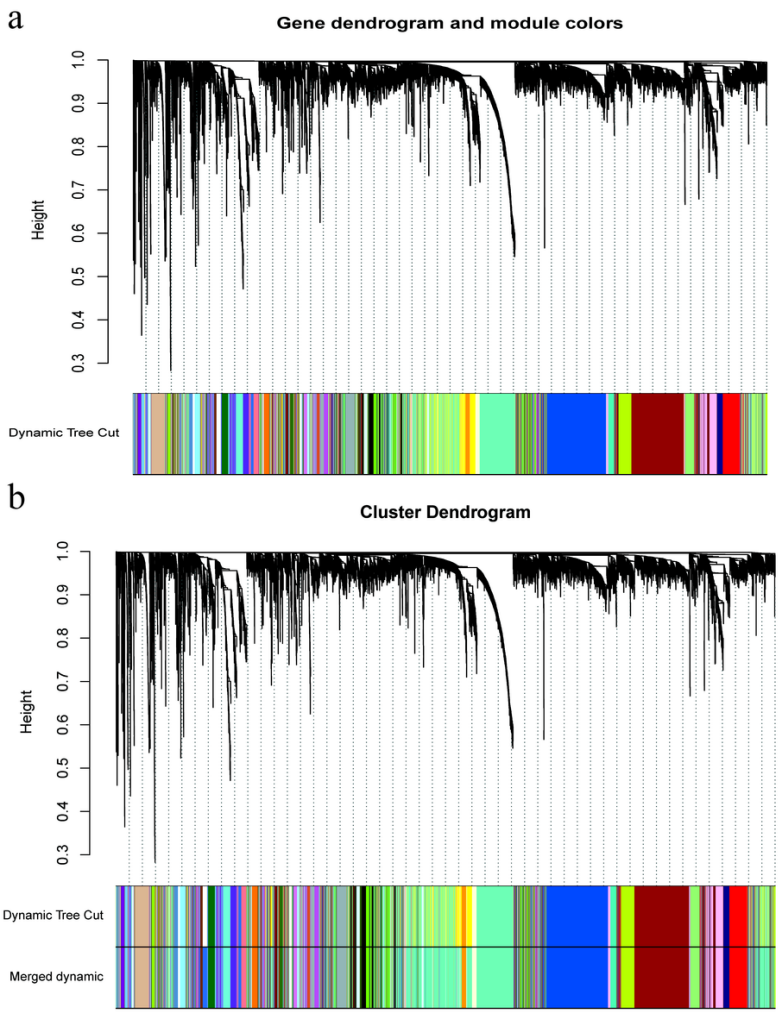

Figure 3

Cluster dendrogram was generated by hierarchical clustering based on dissimilarity measure (dis-TOM) of genes. The branches correspond to modules of highly interconnected groups of genes. Two colored bars below the dendrogram represent the original modules and merged modules. Forty-five modules were identified by Dynamic Tree Cutting method. Each module was assigned a color as an identifier. Forty-three modules were generated after merging according to the correlation of modules.

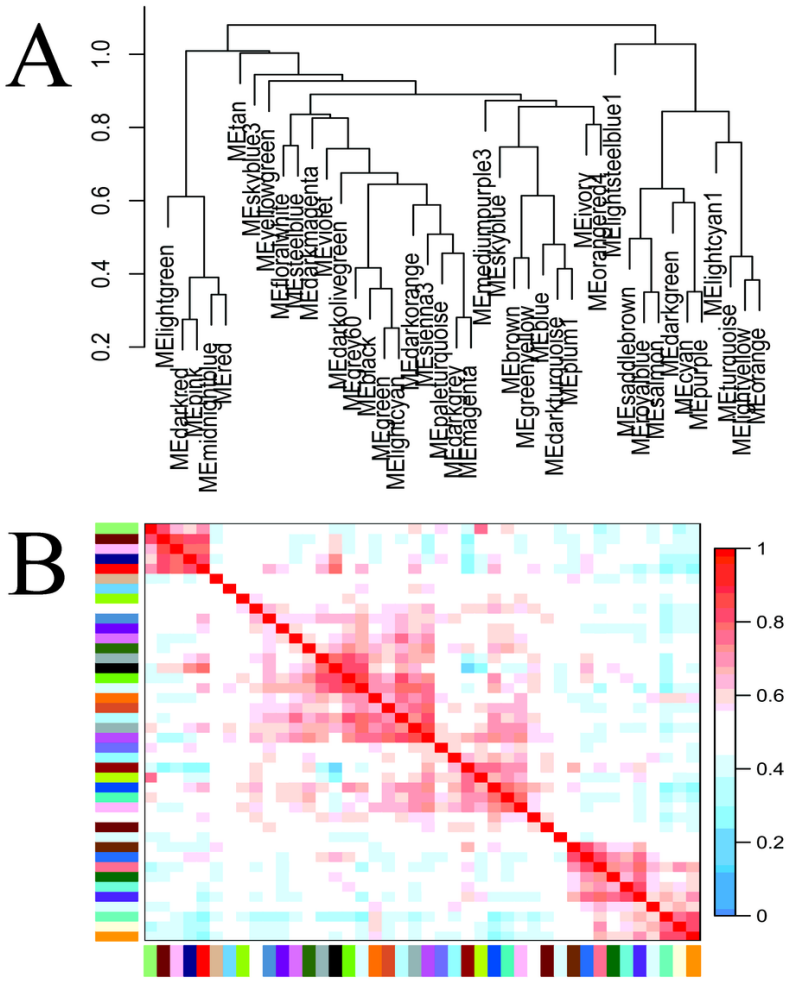

Figure 4

The eigengene network including dendrogram and heatmap shows the correlation among the module and clinical traits. (A) Hierarchical clustering of MEs indicated the branches of the dendrogram cluster together eigengenes which are positively correlated. (B) Heatmap plot of the adjacencies in the hub gene network, red represents positive correlation with high adjacency, while blue color represents negative correlation with low adjacency. Squares of red color along the diagonal are the meta-module. the meta-modules is positively correlated with clinical traits.

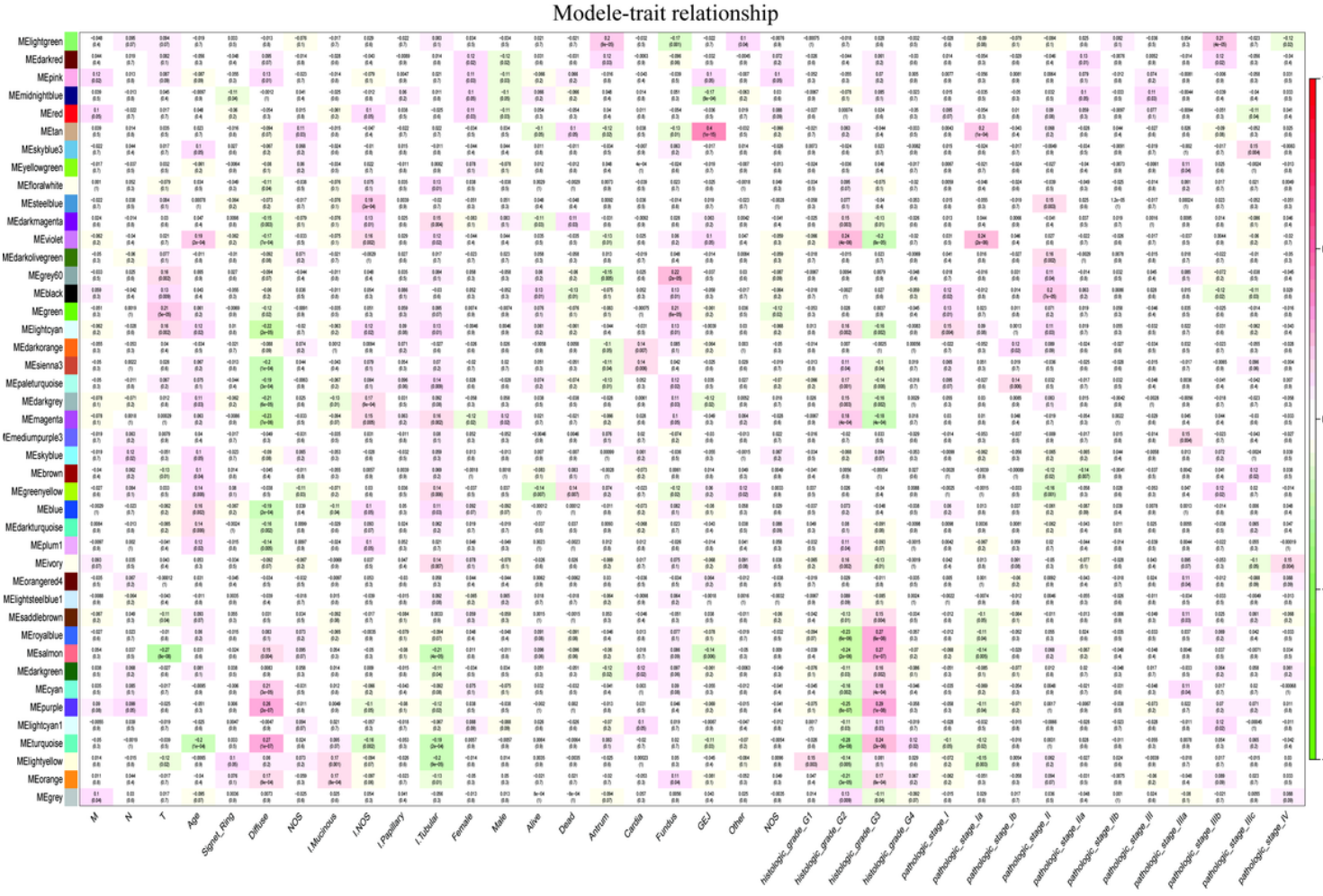

Figure 5

Module trait relationships and p-values for selected traits. Each row corresponds to a ME, and each column corresponds to a clinical trait. Each cell contains a corresponding correlation and p-value. The table is color-coded by correlation according to the color legend.

Figure 6

Co-expression network of most connected genes in the modules including (a) light-green, (b) orange, (c) purple, (d) salmon, (e) tan and (f) turquoise. Nodes and edges represent the correlation of genes. The nodes with saturated circle represented network hub genes and the edge width was proportional to the score of weight by WGCNA.

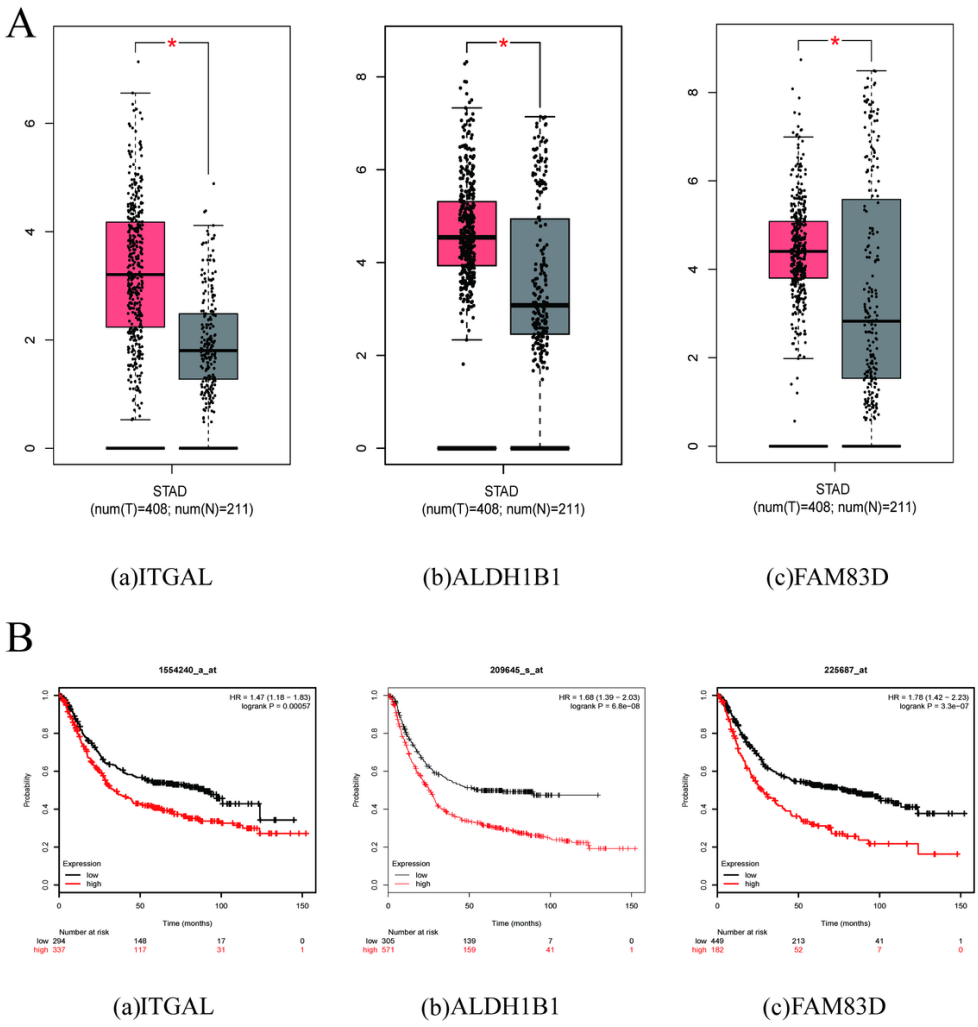

Figure 7

Hub genes validation. (A) Validation of the gene expression levels between GC sample and normal tissue. (a) ITGAL (b) ALDH1B1, (c) FAM83D. (B) Overall survival analysis of three hub genes in darkgreen, pink, and magenta modules individually (a) ITGAL (b) ALDH1B1, (c) FAM83D. Red line represented the samples with gene highly expressed and blue line was for the samples with gene lowly expressed. HR: hazard ratio.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementaryfiles.doc