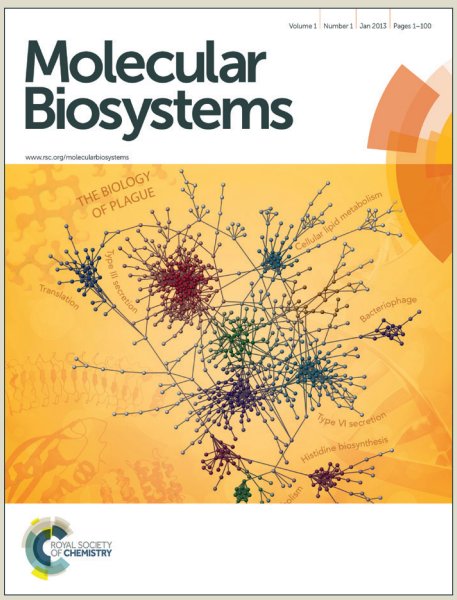
Molecular BioSystems

Accepted Manuscript



ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/molecularbiosystems

Journal Name

RSCPublishing

ARTICLE

A system level analysis of gastric cancer across tumor stages with RNA-seq data†

Cite this: DOI: 10.1039/xoxxooooox

Jun Wu,[a] Xiaodong Zhao,[b] Zongli Lin*[c] and Zhifeng Shao[d]

Received ooth January 2012,
Accepted ooth January 2012

DOI: 10.1039/xoxxooooox

www.rsc.org/

Gastric cancer is the third leading cause of cancer-related death in the world. Over the past decades, with the development of high-throughput technologies and the application of various statistical tools, cancer research has witnessed remarkable advancements. However, no system level analysis has taken into account the cancer stages, which are known to be extremely important in prognosis and therapy. In this study, we aimed to carry out a system level analysis over dynamics of the network structure across the normal phenotype and the four tumor stage phenotypes. We analyzed 276 samples of primary tumor tissues including normal and four tumor stage phenotypes to reveal the dynamics of the five phenotype-specific co-expression networks. Our analysis reveals that the structure of the normal network is dramatically different from that of a tumor network. The analysis of connectivity dynamics exhibits that hub genes present in the normal network but not in the tumor networks play important roles in tumorigenesis and hub genes unique to a tumor network are enriched in specific biological terms. Moreover, we found three interesting clusters of genes which possess specific dynamic features across the five phenotypes and are enriched in stage-specific biological terms. Integrating the results from the expression analysis and the connectivity analysis elucidates that the stages of tumor should be taken into consideration and a system level analysis serves as a complement to and a refinement of the traditional expression analysis.

Introduction

Gastric (stomach) cancer is the third leading cause of cancer-related death in both genders worldwide. According to GLOBOCAN 2012, almost one million new cases of gastric cancer (952,000 cases, 6.8% of the total cancer burden) were estimated to have occurred in 2012[1]. Of these cases, more than 70\% occurred in developing countries and half in Eastern Asia (mainly in China). Although the developments of techniques in diagnosis and treatment have improved the survival rate of gastric cancer, there are still many challenges[2].

Over the past decades, cancer research has experienced remarkable advancements with the development of high-throughput technologies and the application of various statistical tools[3-5]. In the analysis of the genetic pattern between tissues in different phenotypes, the most commonly used method is the differential expression analysis, such as DESeq2, baySeq, edgeR and DEpln[6-9]. These methods detect potential cancer associated genes based on the assumption that prognostic genes may express significantly differentially. Unfortunately, these methods treat genes as individuals and lose sight of the associations among them. Moreover, carcinogenesis is a complex process involving gradual accumulation and interaction of genetic mutations[10,11]. Biological networks, such as the protein-protein interaction network, the metabolic network, the gene regulatory network and the gene co-expression network, are very useful vehicles to a deep understanding of the cancer on the system level. Gene co-expression networks serve as a means to explore the functionality of genes on the systems level[12]. Compared with other types of biological networks, the gene co-expression network has several advantages, including its ability to build cancer-type-specific networks, nearly complete coverage of human genes and little bias due to the knowledge obtained from the published literature[13,14]. The weighted gene co-expression network analysis (WGCNA)[14-16] is a sophisticated method designed for constructing co-expression networks from gene expression data, and has been found to be one of the methods that performed best for constructing global co-expression networks[17]. Yang et al. utilized the WGCNA to statistically analyze the properties of the prognostic genes from the system perspective for glioblastoma multiforme, ovarian serous cystadenocarcinoma, breast invasive carcinoma and kidney renal clear cell carcinoma[13]. Anglani et al. combined the differential expression analysis with the gene co-expression network analysis to improve the classical enrichment pathway analysis[11].

However, on the system level, no analysis method has taken the cancer stages into consideration. Genes exhibit different behaviors (e.g., expression levels) in different stages of the

development and confounding of the tumor stages can introduce errors or biases in the analysis of cancer data[18,19]. In addition, cancer stages are also extremely important to prognosis and the confirmation of the cancer stage is a key factor in deciding the best way to treat the cancer.

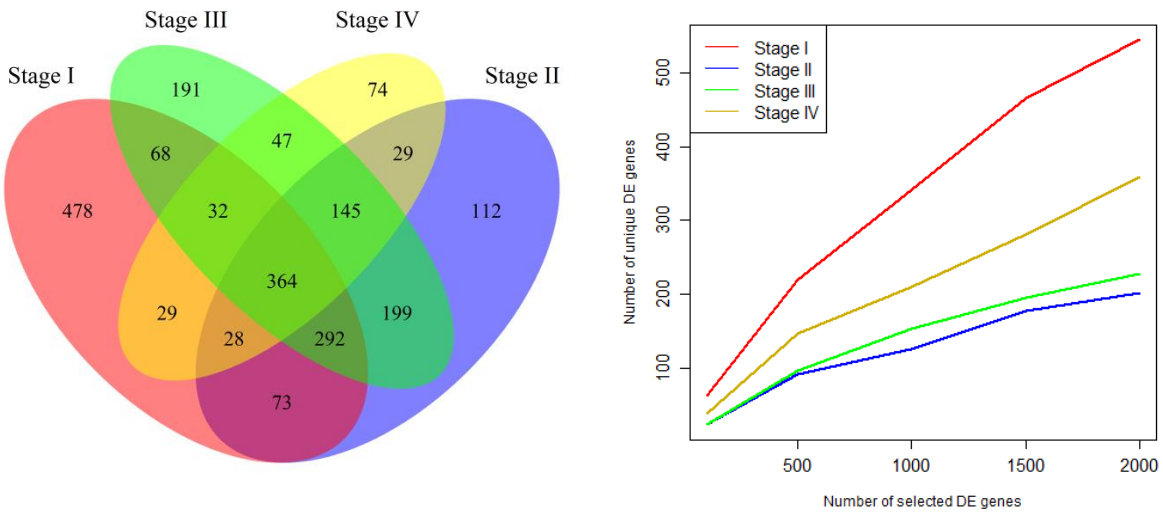
In this paper, we report on a comprehensive analysis of the genetic patterns of gastric tissues in the normal and different cancer stages. The differential expression analysis and the co-expression network analysis were applied to investigate the expression and the system level dynamic properties. DESeq2 methods was applied to identify the differentially expressed (DE) genes in the tissues in different tumor stages in comparison with the normal tissues and the relationship among the DE gene lists of the corresponding tumor stages were studied. Gene ontology (GO) terms were used to investigate the enrichment of the biological process (BP) and the KEGG pathway in these DE genes. The system level properties were studied by means of the gene co-expression network. We found that the structure of the normal network is more compact than those of the tumor networks in different cancer stages and the loss of connectivity in the tumor networks with respect to the normal network is a common trait among the different cancer stages. Genes with extremely large connectivity are more important than other genes in organizing the global network structure. We found that about 75% hub genes in the normal network are depleted in the tumor networks. These genes have previously been reported to play important roles in tumorigenesis. Integrating the results of the differential expression analysis and the connectivity analysis, we identified six genes, THBS2, COL4A1, COL12A1, NOTCH1, STK3 and PXDN, all of which have been reported to be closely associated with gastric cancer or other types of cancer. The results indicate that the stages of tumor should be taken into consideration and a system level analysis serves as a complement to and a refinement of the traditional methods.

Results

We analyzed expression levels of 276 human gastric cancer mRNAomes, including normal and tumor tissues. Both the raw count data and fragments per kilobase of exon per million fragments mapped (FPKM) count data are used. The raw count data was used in the differential expression analysis for the suggestion of DESeq2 method and the FPKM count data was used to construct the co-expression network for the unbiased measure of gene expression.

Differential expression analysis of gastric cancer samples with distinct stages

We compared the expression of genes in the normal samples with the genes in the samples from the four tumor stages, respectively. The approach to identifying the differentially expressed (DE) genes will be discussed later in the section of Materials and Methods. The genes satisfy the following three conditions are considered DE genes: 1) not background noise (see Material and Methods); 2) the adjusted p-value less than 0.001; 3) the fold change level larger than 2. Finally, we detected 1,364, 1,242, 1,338 and 748 DE genes in Stages I to IV, respectively. To investigate the overlaps among these sets of DE genes, we used the Venn diagram to show the results (see Fig. 1(a)). In the figure we can see that there are 364 common genes that differentially expressed in the four tumor stages. To further show the difference, we selected the top 100, 500, 1,000, 1,500 and 2,000 genes as the DE genes according to the adjusted p-value. The unique DE genes for each tumor stage were shown in Fig. 1(b). We can see that there actually exist considerable differences across the tumor stages, which should also be considered in related works, such as biological markers identification and cancer prognosis.



(a) (b)

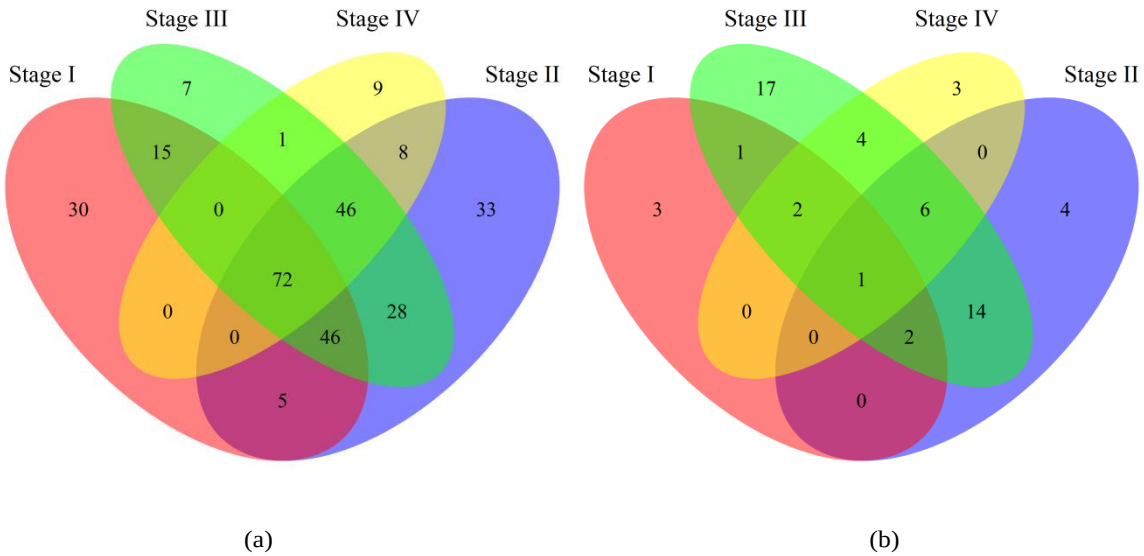
Fig. 1 The difference of DE genes identified in each tumor phenotypes compared to the normal phenotype. (a) Venn diagram of DE genes of different tumor stages. "Stage I", "Stage II", "Stage III" and "Stage IV" represented the corresponding DE genes. (b) The plot of unique DE genes versus different top number selected in each tumor phenotype.

To further investigate the difference of these DE genes, we applied the Gene Ontology (GO) enrichment analysis to search for the enriched biological function items. The GO enrichment analysis was performed with R package GOstats[20]. Hypergeometric test included in the GOstats package was used to test the enrichment of a GO term in each gene list. The p-values obtained were adjusted using the Benjamini and Hochberg procedure (BH-adjusted p-value) for multiple

comparisons. GO terms with a BH-adjusted p-value less than 0.001 were regarded as significantly enriched. We divided the DE genes of each stage into two parts, the up-regulated and the down-regulated genes. Then eight DE gene lists were obtained for further GO term enrichment analysis.

We first tested the enrichment of the biological process (BP) for each gene list, and obtained 168, 238, 215 and 136 enriched BPs in the up-regulated genes of Stages I to IV, respectively (see Table S1, ESI†). The Venn diagram to illustrate the overlap of the enriched BPs was shown in Fig. 2(a). There were 72 BPs commonly enriched in the up-regulated genes across the four tumor stages, including mitotic cell cycle, nuclear division, cell cycle and DNA replication. These commonly enriched BPs are mainly involved in the cell growth and development. We also identified the enriched BPs in the down-regulated genes across the four tumor stages (see Table S2, ESI†) and the Venn diagram was also shown in Fig. 2(b). The only commonly enriched BP was the oxidation-reduction process. It was reported that moderate oxidation can help with the immune system, while too much oxidation can damage the DNA to cause malignant cell or inhibit the mechanism that can clear cancer cell[21-23]. Reduction/oxidation (redox) imbalance may cause the cancer progression[24], and the down-regulation of genes involved in oxidation-reduction process may play an important role in tumorigenesis.

We also identified enriched KEGG pathways in the DE genes across tumor stages. The KEGG pathways with a BH-adjusted p-value less than 0.05 were regarded as significantly enriched. In the up-regulated genes, we finally identified 26, 23, 27 and 19 enriched KEGG pathways for each tumor stages (see Table S3, ESI†). We found that 11 KEGG pathways are commonly enriched across the four tumor stages, including DNA replication, TGF-beta signalling pathway, focal adhesion and ECM-receptor interaction. It has previously been reported that the aberrated expression of genes involved in DNA replication and TGF-beta signalling pathway contribute to the carcinogenesis and cancer progression[25,26]. ECM-receptor interaction and the focal adhesion pathway have been reported as associated with the progression of gastric cancer[27]. Moreover, the up-regulated genes in early gastric cancer tissues were intrinsically associated with ECM-receptor interactions and focal adhesion[28]. In the down-regulated genes, we identified 20, 27, 33 and 27 KEGG pathways enriched for each tumor stages (see Table S4, ESI†). Only 4 KEGG pathways were commonly enriched across all stages, including oxidative phosphorylation and gastric acid secretion which were usually associated with the gastric cancer[21,23,29-31].



(a) (b)

Fig. 2 Venn diagrams show the overlapped relation across the enriched BPs in DE genes. (a) Venn diagram of enriched BP lists in up-regulated genes of four tumor stages. (b) Venn diagram of enriched BP lists in down-regulated genes of four tumor stages.

Modules identification in the phenotype-specific networks

For all the five phenotypes, which include the normal phenotype and four tumor phenotypes, we constructed gene co-expression networks based on the RNA-seq data using WGCNA. The gene expression level was used to filter the background noise and the Pearson's correlation coefficient was introduced to measure the association between gene pairs (see Material and Methods). Finally, we obtained five phenotype-specific co-expression networks, each containing the same set of 11,077 genes. The weight of the edge connecting two genes denotes the strength of their interaction.

Module is an important property in a co-expression network. It is a highly connected subgraph in the gene co-expression network. The genes in a common module have similar functions or are involved in a common biological process which causes many interactions among them[32]. We defined the modules in each phenotype-specific network with the R package WGCNA and 177, 81, 71, 69 and 67 modules were detected in the normal and Stages I-IV networks, respectively (see Fig. 3(a)). In the figure, modules are designated by various colors and gray regions denote genes outside of the modules. The corresponding proportions of the genes belonging to a module in these phenotype-specific networks are respectively 69.11%, 33.13%, 27.44%, 28.39% and 22.02%. We can see that more than two-thirds of genes are assigned to a module in the normal networks while only less than one-third of genes are assigned to a module in the corresponding four tumor networks. This implies that the structure of the normal network is more compact than those of the tumor networks. There are 491 genes that are always in the modules of the normal and the four tumor networks (see Fig. 3(b)). Through the GO term enrichment

analysis, we found that these genes are enriched in the KEGG pathways that are involved in tumorigenesis. For example, the three top enriched KEGG pathways are Focal adhesion, ECM-receptor interaction and cell adhesion molecules.

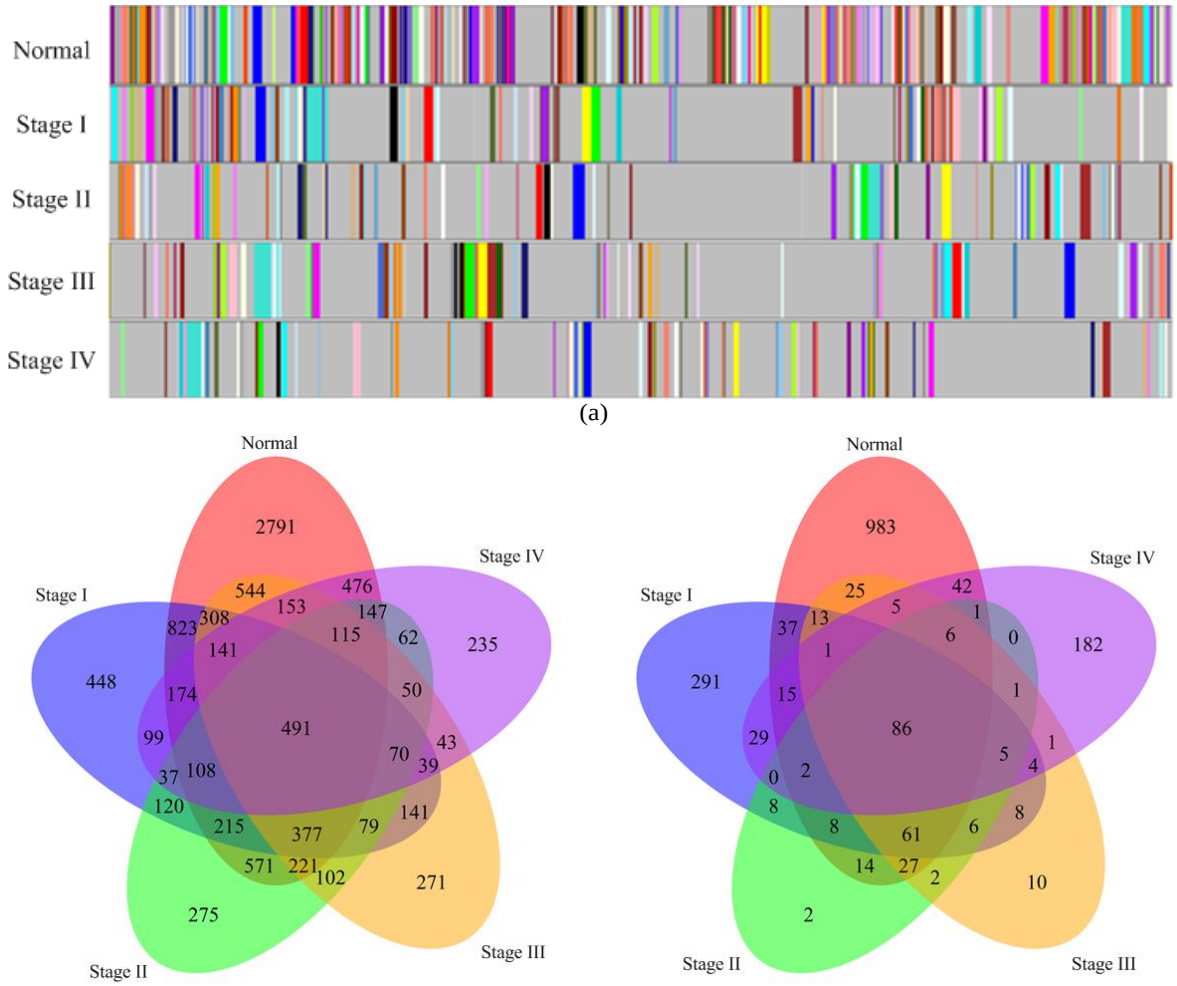


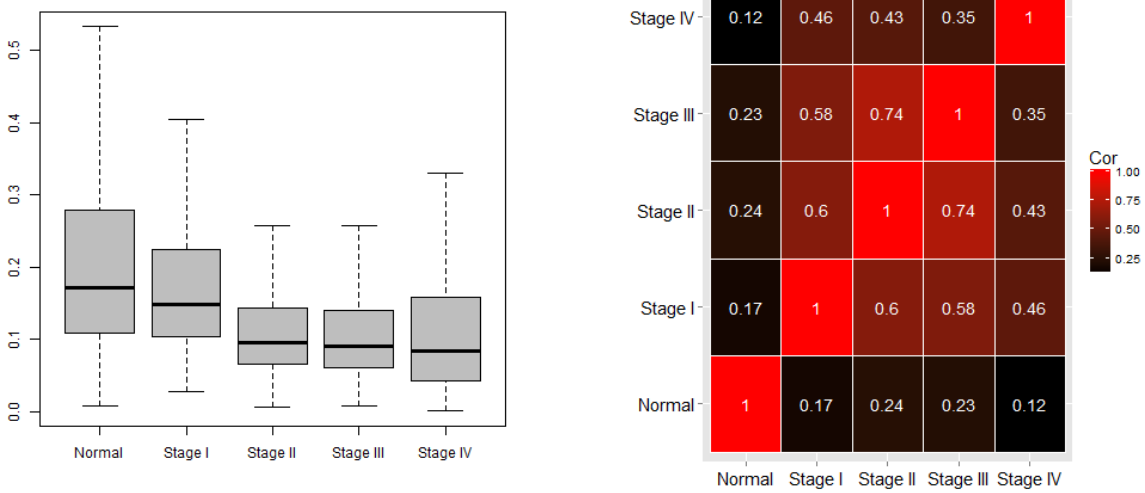
Fig. 3 Key properties in the weighted co-expression networks. (a) Modules defined with WCGNA. Color bands represent modules in the network and grey regions denote genes outside of modules. (b) The Venn diagram of module genes across the phenotype-specific networks. (c) The Venn diagram of hub-genes in the corresponding networks.

Connectivity dynamic analysis across the phenotype-specific networks

Gene connectivity is another key property to study gene co-expression networks. The gene connectivity reflects how frequently a gene connects with other genes and the rank of the connectivity, in some way, can indicate the importance of the gene. To facilitate the comparison of the connectivity measures among the networks, we normalized each gene connectivity value by dividing it with the maximum network connectivity, and obtained a connectivity profile for each gene. We first studied the general change of the gene connectivity across the five phenotype-specific networks and found that the gene connectivity decreases in the tumor networks (see Fig. 4(a)). A similar observation was also made in[11]. To measure the switching of the gene connectivity rank between any two networks, we computed the Spearman's correlation coefficient for each pair of networks and the results show that the gene connectivity rank in the normal network and in the tumor networks are largely uncorrelated, which indicates that the structure of the normal network and those of the tumor networks are very different (see Fig. 4(b)).

The dynamic feature of the gene connectivity across the five phenotype-specific networks was also analyzed. We utilized the K-means method to classify the genes into nine clusters according to their connectivity profiles (see Fig. 5(a)). A trial and error process indicates that the choice of $K = 9$ achieves a good tradeoff between the ratio of the between-cluster sum of squares and the total within-cluster sum of squares and the gene number in each cluster. We found that most genes (73.1%) maintain a low connectivity across the five phenotype-specific networks. Among the nine clusters, three clusters, Clusters 3, 4 and 9, display some interesting features (see Fig. 5(b)) and the corresponding genes in these three clusters are listed (see Table S5, ESI†). The connectivity values of genes in these three clusters are respectively higher in Stage IV, the normal and Stage I networks. To explore the biological terms of the genes in these clusters, R package GOstats was introduced to identify the significantly enriched KEGG pathways. The results revealed that the significantly enriched KEGG pathways are highly specific to clusters and the overlaps are small (see Fig. 5(c)). For example, the ECM-

receptor interaction pathway and the Wnt signaling pathway both are enriched in genes belonging to Cluster 4, in which the gene connectivity decreases in the tumor networks. The Wnt signaling pathway tightly regulates several important processes during the development, such as cell adhesion and growth. Deregulation of Wnt/beta-catenin signaling is frequently found in various human cancers[33]. ECM-receptor interaction is a significantly enriched pathway in gastric cancer and the deregulation of ECM-receptor interaction pathway emerge in most malignant cancer[34]. Furthermore, some enriched KEGG pathways in the Cluster 3 genes, whose connectivity is relatively high in the Stage IV networks, are involved in tumour cell invasion and metastasis, such as mRNA surveillance pathway. The mRNA surveillance pathway down-regulates aberrant E-cadherin transcript which is an adhesion molecule and acts as a tumor suppressor protein by inhibiting tumor cell invasion and metastasis, which are typical characters in stage IV tumor[35].



(a) (b)

Fig. 4 The connectivity difference between the normal and tumor networks. (a) The connectivity distributions of normal and four stage tumor networks. (b) The Spearman's correlation coefficient between each pair of connectivity of the five networks.

According to the connectivity, the genes can be classified into hub genes and non-hub genes. The hub genes possess extremely larger connectivity and are more important than other genes in organizing the global network structure and exchanging information. We mixed the gene connectivity values in all phenotype-specific networks and selected the top 5% largest value as the threshold that distinguishes the hub-genes from the non-hub genes. We obtained 1,326, 574, 229, 261 and 380 hub-genes in the normal and Stages I-IV networks, respectively (see Fig. 3(c)). We can see that about 75% of the hub genes in the normal network are depleted in the tumor networks. These genes may contribute to the tumorigenesis. For example, *ZMYND11* is a candidate tumor suppressor and is critical for the repression of a transcriptional program that is essential for tumor cell growth[36]. The expression of *ZMYND11* does not show a significant difference between the tumor phenotypes and the normal phenotype, and we can screen it out with a connectivity rank difference. We also observed that there are more hub genes unique to the Stage I and Stage IV networks than those unique to the other two tumor networks. Through a GO term enrichment analysis, we found that the hub-genes that uniquely exist in the Stage I network are mainly enriched in the processes of cell activation and immune system, while the hub-genes that uniquely exist in the Stage IV network are mainly enriched in the metabolic processes, such as RNA metabolic process and macromolecule metabolic process. Moreover, some of the hub genes that are unique to the Stage IV network, such as *TBK1*, play important roles in cell migration and invasion, which are typical characters in stage IV tumors. The loss of *TBK1* induces the epithelial-mesenchymal transition (EMT), which gives cells the ability to migrate and invade[37].

Integrating the results of the differential expression analysis and the gene connectivity analysis, we selected the genes that are commonly significantly differentially expressed in the four tumor phenotypes compared to the normal phenotype and the hub-genes that are uniquely to the normal network for an one-by-one analysis. Finally, we identified six genes, *THBS2*, *COL4A1*, *COL12A1*, *NOTCH1*, *STK3* and *PXDN*. *THBS2*, as a THBS family member, has been reported to regulate angiogenesis and its expression is aberrantly in gastric cancer, which indicates its critical role in cancer progression[38]. *COL4A1* and *COL12A1*, members of the collagen family, were reported to significantly up-regulate in gastric cancer[39]. Notch homolog 1 (*NOTCH1*) encodes a member of the Notch family and plays a role in a variety of developmental processes by controlling cell fate decision. The activated *NOTCH1* receptor promotes the progression of gastric cancer through regulating the expression levels of *STAT3* and *Twist*[40]. *STK3* encodes a serine/threonine protein kinase activated by proapoptotic molecules, indicating that the encoded protein functions as a growth suppressor. *STK3* is involved in the Hippo signaling pathway, which plays a pivotal role in organ size control and tumor supression by restricting proliferation and promoting apoptosis. Heme Oxygenase-1 (*HO-1*), expressed in many cancers, promotes growth and is implicated in tumor cell invasion and metastasis. The adhesion-promoting effects of *HO-1* are dependent on *PXDN* expression and the loss of *PXDN* leads to reduced cell attachment to Laminin and Fibronectin coated wells.

Conclusion

Gastric cancer is the third leading cause of cancer-related death in the world. From the public data of TCGA, we selected 276 gastric samples from primary tumor tissue and carried out a comprehensive analysis of gastric cancer across the normal and the tumor stages with RNA-seq data. The 276 samples were first classified into five phenotypes, the normal, and tumor Stages I-IV, according to the

clinical data. Our comprehensive analysis includes the differential expression analysis and the network structure analysis.

Gene differential expression analysis is a traditional approach to analyzing genes across different phenotypes. We compared the gene expression between the four tumor phenotypes and the normal phenotype, respectively, with the DESeq2 method. The results show that the gene expressions vary greatly from one tumor stage to another. To investigate the GO terms similarity, we carried out the GO enrichment analysis on the four DE gene lists and found that there are 72 commonly enriched BPs in the up-regulated genes and only one commonly enriched BPs in the down-regulated genes. The commonly enriched BPs in the up-regulated genes are mainly involved in cell growth and development, which are closely related with tumor cell proliferation, such as mitotic cell cycle, nuclear division, cell cycle and DNA replication. The only commonly enriched BP in the down-regulated genes is oxidation-reduction process. Moderate oxidation can help with the immune system, while excessive oxidation may inhibit the mechanism that can clear cancer cell. The KEGG pathway enrichment analysis identified 15 commonly enriched KEGG pathways, including DNA replication, TGF-beta signaling pathway, focal adhesion and ECM-receptor interaction.

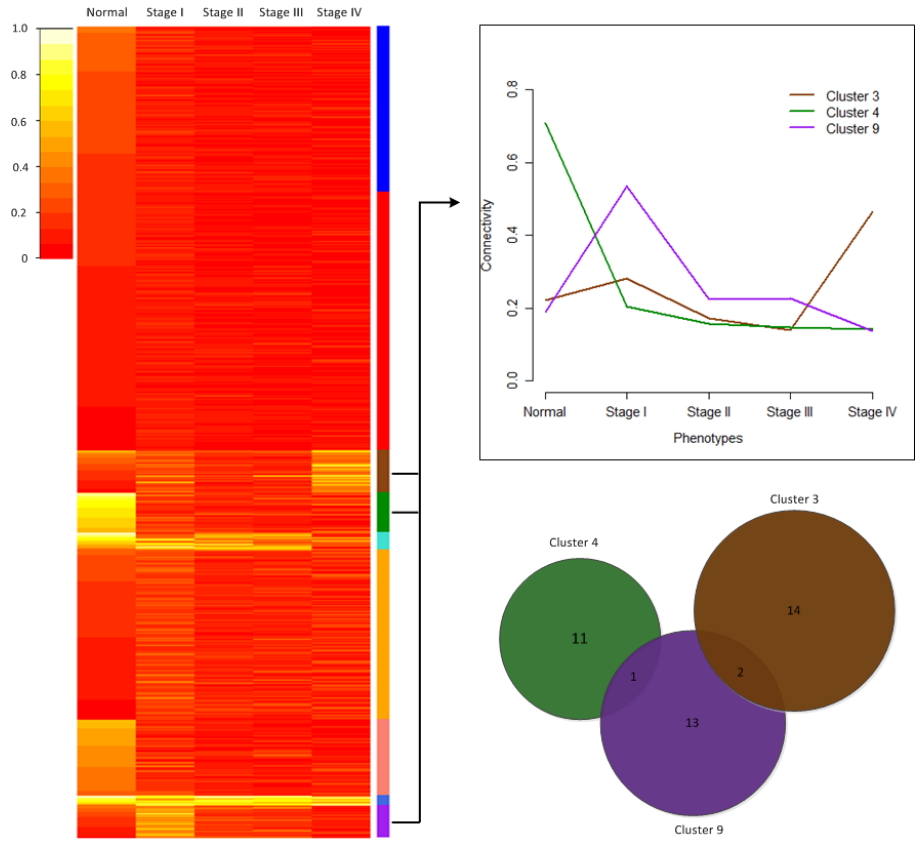


Fig. 5 K-means analysis based on gene connectivity across phenotypes. (a) Heatmap of K-means result. (b) The dynamic feature of connectivity mean of all genes belonging to three remarkable clusters across phenotypes. (c) The overlap relationships of significantly enriched KEGG pathways in the three clusters.

The WCGNA algorithm was applied to construct the gene co-expression networks and five phenotype-specific networks were obtained. We have analyzed two key properties of these networks in this study, modules and connectivity. Modules are highly connected subgraphs in gene co-expression networks. We found that about 70% genes are affiliated with modules in the normal network while only about 30% are affiliated with modules in the tumor networks. This indicates that the structure of the normal network is more compact than the structures of the tumor networks. Through connectivity analysis, we found that the loss of connectivity in the tumor networks with respect to the normal network is a common trait among different tumor stages and the connectivity ranks of genes are largely uncorrelated between the tumor networks and the normal networks. We profiled the connectivity dynamic features across the five phenotype-specific networks and divided them into 9 clusters based on K-means algorithm. We selected three interesting clusters for the GO enrichment analysis and found them all closely related to the pathways involved in tumorigenesis. The hub-genes which possess extremely high connectivity play more important roles than other genes in maintaining the global structure of the network. We selected the hub-genes in the five phenotype-specific networks, respectively. We found that about 75% hub genes in the normal network are depleted in the tumor networks and these genes may contribute to tumorigenesis. We also carried out the GO enrichment analysis of hub-genes that uniquely exist in the Stage I and Stage IV networks and found that the hub-genes unique to the Stage I network are mainly enriched in the processes of cell activation and immune system, while the hub-genes unique to the Stage IV network are mainly enriched in the metabolic processes, such as the RNA metabolic process and the macromolecule metabolic process. Integrating the results of the differential expression analysis and the connectivity analysis, we identified six genes, *THBS2*, *COL4A1*, *COL12A1*, *NOTCH1*, *STK3* and *PXDN*, all of which have been reported to be closely associated with the gastric cancer or other types of cancer. The results demonstrated that the stage separation and the combination of the expression analysis and the system level

analysis are necessary for a deep understanding of gastric cancer. This network analysis can also be used as a necessary method or a refinement to the traditional method in the works such as biomarker identification.

Materials and Methods

Data collection and differential expression analysis

We obtained the gene expression from the TCGA project webpage[41]. The details of the filter settings are shown in Table 1 and the corresponding clinical data was also downloaded. As described in[42], the gastric adenocarcinoma primary tumor tissues from patients not treated with prior chemotherapy or radiotherapy are selected. Finally, we obtained 276 samples, including 29 normal samples and 247 tumor samples. According to the clinical data, all these 276 samples were divided into 5 phenotypes, which are the normal phenotype and Stages I-IV phenotypes. The DESeq2 method, along with the raw count data, was used to implement the differential expression analysis. Trimmed mean of M-values (TMM) method [43] was employed for the normalization to reduce the bias across the samples. Differential expression analysis was applied to each tumor stage and the normal pair. The BH-adjusted p-value and the fold change level were used to screen the significantly differentially expressed genes. Moreover, genes with more than half of samples expressed less than 0.1 FPKM or a mean expression less than 2 FPKM were regarded as background noise. Genes with BH-adjusted p-value less than 0.001, fold change level larger than 2 and not background noise were regarded as significantly differentially expressed.

Table 1 TCGA filter settings

Select a disease	STAD – Stomach adenocarcinoma
Data Type	RNASeq
Data level	Level 3
Availability	Available
Other	default

Co-expression networks analysis

Given the RNA-seq data for the five phenotypes, including the normal phenotype and the four tumor stages, the WGCNA approach was utilized to construct a weighted gene co-expression network. To reduce the negative impact of background noise, we first filtered the background noise genes, which were defined above. Among the remaining genes, 11,484 are in the normal phenotype, 12,520 in Stage I, 12,680 in Stage II, 12,834 in Stage III and 12,649 in Stage IV. The common genes that remained in the five phenotypes were used to construct the phenotype-specific co-expression networks. The Pearson's correlation coefficient was used to measure the association of each gene pair. The key parameter β, which was used to maintain both the scale-free topology and a sufficiently high node connectivity, was optimized as recommended in the original manual[15], which is the lowest value for which the scale-free topology fit index reaches 0.9. The connection strength of any two genes was measured by the topology overlap matrix (TOM) provided in WGCNA. TOM takes both the co-expression pattern between two genes and the overlap of neighbouring genes into account. Moreover, TOM can be regarded as a filter that reduces the effect of weak connection to result in a more robust network.

Given a co-expression network, two key network properties were obtained, modules and connectivity. The modules of a network were identified by the dynamic hybrid tree cut algorithm provided in the WGCNA package and the parameters were set as the default values. The gene connectivity reflects how frequently a gene connects with other genes. The top 5% quintiles of the mixed gene connectivity values were set as the threshold to distinguish the hub-genes in each network. We used the Spearman's correlation coefficient of gene connectivity to measure the structure change of two networks.

Acknowledgements

This work was supported in part by the Longhua Medicical Project of the State Clinical Research Center of TCM at the Longhua Hospital (LYTD-21), in part by the State Key Development Program for Basic Research of China (2010CB529205 and 2013CB967402) and in part by the National Natural Science Foundation of China under grant No. 61221003, 91019004 and 91229123. The authors would like to thank the reviewers in advance for their comments.

Notes and references

[a] Department of Automation, Shanghai Jiao Tong University, and Key Laboratory of System Control and Information Processing of Ministry of Education, Shanghai, China; E-mail: junwu302@gmail.com.

[b] School of Biomedical Engineering, Shanghai Jiao Tong University, Shanghai, China; E-Mail: xiaodong122@yahoo.com.

[c] Charles L. Brown Department of Electrical and Computer Engineering, University of Virginia, Charlottesville, Virginia, United States of America; Email: zl5y@virginia.edu.

[d] School of Biomedical Engineering, Shanghai Jiao Tong University, Shanghai, China; E-Mail: zs9q@virginia.edu.

† Electronic Supplementary Information (ESI) available: See DOI: 10.1039/b000000x/

1 J. Ferly, I. Soerjomataram, M. Ervik, R. Dikshit, S. Eser, C. Mathers, M. Rebelo, D. M. Parkin, D. Forman, F. Bray, http://globocan.iarc.fr.
2 D.H. Roukos, *Cancer treatment reviews*, 2000, **26(4)**, 243-255.
3 H. Kitano, *Science*, 2002, **295(5560)**, 1662-1664.
4 H. Kitano, *Nature*, 2002, **420(6912)**, 206-210.
5 Y. Wang, X. S. Zhang, L. Chen, *BMC systems biology*, 2011, **5**, Suppl 1:S1.
6 T. J. Hardcastle, K. A. Kelly, BMC Bioinformatics, 2010, **11**, 422.
7 S. Anders, W. Huber, *Genome biology*, 2010, **11(10)**, R106.
8 M. D. Robinson, D. J. McCarthy, G. K. Smyth, *Bioinformatics*, 2010, **26(1)**, 139-140.
9 J. Wu, X. Zhao, Z. Lin, Z. Shao, *Journal of bioinformatics and computational biology*, 2015, **13(2)**, 1550001-1--1550001-16.
10 A. de la Fuente, Trends in genetics, 2010, **26(7)**, 326-333.
11 R. Anglani, T. M. Creanza, V. C. Liuzzi, A. Piepoli, A. Panza, A. Andriulli, N. Ancona, 2014, *PloS one*, **9(1)**, e87075.
12 J. Ruan, A. K. Dean, W. Zhang, BMC systems biology, 2010, **4**, 8.
13 Y. Yang, L. Han, Y. Yuan, J. Li, N. Hei, H. Liang, *Nat Commun*, 2014, **5**, 3231.
14 W. Zhao, P. Langfelder, T. Fuller, J. Dong, A. Li, S. Hovarth, *Journal of biopharmaceutical statistics*, 2010, **20(2)**, 281-300.
15 B. Zhang, S. Horvath, *Statistical applications in genetics and molecular biology*, 2005, **4**, Artical17.
16 P. Langfelder, S. Horvath, *BMC bioinformatics*, 2008, **9**, 559.

17 J. D. Allen, Y. Xie, M. Chen, L. Girard, G. Xiao, Comparing statistical methods for constructing large scale gene networks. *PloS one*, 2012, **7(1)**, e29348.
18 W. P. Kuo, T. K. Jenssen, P. J. Park, M. W. Lingen, R. Hasina, L. Ohno-Machado, *Proc AMIA Symp*, 2002, 415-419.
19 Z. Q. Fang, W. D. Zang, R. Chen, B. W. Ye, X. W. Wang, S. H. Yi, W. Chen, F. He, G. Ye, *Genetics and molecular research*, 2013, **12(2)**, 1479-1489.
20 S. Falcon, R. Gentleman, Bioinformatics, 2007, **23(2**), 257-258.
21 J. Ni, M. Mei, L. Sun, *Hepatogastroenterology*, 2012, **59(115)**, 671-675.
22 Y. Hiraku, *Fukuoka Igaku Zasshi*, 2014, **105(2)**, 33-41.
23 F. Farinati, R. Cardin, M. Bortolami, D. Nitti, D. Basso, M. de Bernard, M. Cassaro, A. Sergio, M. Rugge, *International journal of cancer*, 2008, **123(1)**, 51-55.
24 T. C. Jorgenson, W. X. Zhong, T. D. Oberley, *Cancer research*, 2013, **73(20)**, 6118-6123.
25 L. A. Loeb, C. F. Springgate, N. Battula, *Cancer research*, 1974, **34(9)**, 2311-2321.
26 R. Derynck, R. J. Akhurst, A. Balmain, *Nature genetics*, 2001, **29(2**), 117-129.
27 K. Hu, F. Chen, *Genetics and molecular biology*, 2012, **35(3)**, 701-708.
28 S. Nam, J. Lee, S. H. Goh, S. H. Hong, N. Song, S. G. Jang, I. J. Choi, Y. S. Lee, *International journal of oncology*, 2012, **41(5)**, 1675-1682.
29 M. H. Jung, S. C. Kim, G. A. Jeon, S. H. Kim, Y. Kim, K. S. Choi, S. I. Park, M. K. Joe, K. Kimm, Genomics, 2000, **69(3)**, 281-286.
30 P. Malfertheiner, *Digestive diseases*, 2011, **29(5)**, 459-464.
31 A.T.R. Axon, Adv Med Sci, 2007, **52**, 55-60.
32 S. Roy, D. K. Bhattacharyya, J. k. Kalita, *BMC bioinformatics*, 2014, **15**, Suppl 7:S10.
33 H. F. Zhang, Y. W. Xue, *Hepato-gastroenterology*, 2008, **55(84)**, 1126-1130.
34 D. Becker, I. Sfakianakis, M. Krupp, F. Staib, A. Gerhold-Ay, A. Victor, H. Binder, M. Blettner, T. Maass, S. Thorgeirsson, P. R. Galle, A. Teufel, *Molecular cancer*, 2012, **11**, 55, 2012.
35 R. Karam, J. Carvalho, I. Bruno, C. Graziadio, J. Senz, D. Huntsman, F. Carneiro, R. Seruca, M. F. Wilkinson, C. Oliveira, *Oncogene*, 2008, **27(30)**, 4255-4260.
36 H. Wen, Y. Li, Y. Xi, S. Jiang, S. Stratton, D. Peng, K. Tanaka, Y. Ren, Z. Xia, J. Wu, B. Li, M. C. Barton, W. Li, H. Li, X. Shi, *Nature*, 2014, **508(7495)**, 263-268.
37 K. M. Yang, Y. Jung, J. M. Lee, W. Kim, J. K. Cho, J. Jeong, S. J. Kim, *Cancer research*, 2013, **73(22)**, 6679-6689.
38 R. Sun, J. Wu, Y. Chen, M. Lu, S. Zhang, D. Lu, Y. Li, *Molecular cancer*, 2014, **13**, 255.
39 H. B. Jiang, T. J. Yang, P. Lu, Y. J. Ma, *European review for medical and pharmacological sciences*, 2014, **18(15)**, 2109-2115.
40 K. W. Hsu, R. H. Hsieh, K. H. Huang, A. L. Fen-Yau, C. W. Chi, T. Y. Wang, M. J. Tseng, K. J. Wu, T. S. Yeh, *Carcinogenesis*, 2012, **33(8)**, 1459-1467.
41 The TCGA Database, http://cancergenome.nih.gov/.
42 Cancer Genome Atlas Research Network, *Nature*, 2014, **513(7517)**, 202-209.
43 M. D. Robinson, A. Oshlack, *Genome biology*, 2010, **11(3)**, R25.