*J Genet Resour2020;6(2): 195-208* *Homepage: http://sc.journals.umz.ac.ir/*

**RESEARCH ARTICLE** *DOI: 10.22080/jgr.2020.19465.1198*

# Exploring Gene Signatures in Different Molecular Subtypes of Gastric Cancer (MSS/ TP53+, MSS/TP53-): A Network-based and Machine Learning Approach

Mehdi Sadeghi[1*†], Nafiseh Ghorbanpour[2†], Abolfazl Barzegar[2,3] and Iliya Rafiei[2]

[1] Department of Cell and Molecular Biology, Faculty of Science, Semnan University, Semnan, Iran
[2] Research Institute for Fundamental Sciences (RIFS), University of Tabriz, Tabriz, Iran
[3] Department of Medical Biotechnology, Faculty of Advanced Medical Sciences, Tabriz University of Medical Sciences, Tabriz, Iran

**ARTICLE INFO**

***Article history:***
Received 14 April 2020
Accepted 18 June 2020
Available online 2 July 2020

***Keywords:***
Gastric cancer
Molecular subtypes
Weighted gene co-expression network analysis
Decision tree
Network analysis

***Corresponding authors:**
✉ Mehdi Sadeghi
mehdisadeghi@semnan.ac.ir
† These authors have contributed equally to this work

p-ISSN 2423-4257
e-ISSN 2588-2589

**ABSTRACT**

Gastric cancer (GC) is one of the leading causes of cancer mortality, worldwide. Molecular understanding of GC's different subtypes is still dismal and it is necessary to develop new subtype-specific diagnostic and therapeutic approaches. Therefore, developing comprehensive research in this area is demanding to have a deeper insight into molecular processes, underlying these subtypes. In this study, a three-step methodology was developed to identify important genes and subnetworks in two subtypes of GC ($TP53^+$ and $TP53^-$). First, weighted gene co-expression network analysis was performed to explore co-expressed gene modules in both subtypes. Afterward, the relationship of each module with the tumor pathological stage (as a clinical trait indicating tumor progression) was studied by decision tree machine learning algorithm and the best predicting module was selected for further analysis (modules with 241 genes for $TP53^+$ and 1441 genes for $TP53^-$ were identified). Subsequently, a motif exploring and motif ranking analysis was implemented to explore three-member signature gene motifs in the selected modules' biological network. These motifs may have key regulatory roles in the studied GC subtypes. Motif members of $TP53^-$ mostly contain MAPK signaling pathway genes which show their key role in this subtype of GC. In the case of the $TP53^+$ subtype, our findings demonstrated that alternative splicing and *SNARE* proteins could prompt the initiation and advancement of the disease. These findings can be used to develop new diagnostic and therapeutic approaches based on the personalized medicine concept. This methodology could be implemented to unravel underlying mechanisms and pathways in other complex phenotypes and diseases.



***Please cite this paper as:*** Sadeghi M, Ghorbanpour N, Barzegar A, Rafiei I. Exploring gene signatures in different molecular subtypes of gastric cancer (MSS/ TP53+, MSS/TP53-): a network-based and machine learning approach. *J Genet Resour* 6(2): 195-208. doi: *10.22080/jgr.2020.19465.1198*.

## Introduction

Gastric cancer (GC) is one of the most common causes of cancer mortality, worldwide (Ferro *et al.*, 2014). Based on the global cancer statistics 2018, 1,000,000 newly diagnosed cases and an estimated 783,000 death (one in every 12 deaths, globally), making it the fifth most frequently diagnosed and the third leading cause of cancer death (Bray *et al.*, 2018). With 11644 new cases and 8965 death, gastric cancer is the second most frequently diagnosed and the first leading cause of cancer death in Iran (Bray *et al.*, 2018). There are serious challenges in GC treatment due to the poor prognosis of patients with advanced gastric cancer. Therefore, when the patients are diagnosed, they are most likely in the advanced stages of GC that leads to limited treatment options for patients and consequently a high mortality rate for GC patients. Since GCs molecular nature is not fully understood, further research on its molecular nature is required to

find novel biomarkers for better prognosis hence better treatment of GC (Shimizu *et al.*, 2017). The poor prognostics of GC could be the result of high clinical and pathological heterogeneity of the tumor (Cristescu *et al.*, 2015; Gullo *et al.*, 2018). Recent efforts to classify GC based on their molecular characterization into molecular subtypes has made it possible to avoid this heterogeneity and analyze each subtype separately. This effort is an essential step towards developing personalized medicinal treatment of gastric cancer (Lin *et al.*, 2015). A classification study based on gene expression data on gastric cancer was done by Asian Cancer Research Group in which four molecular subtypes were introduced; microsatellite instability (MSI), microsatellite stability (MSS)/epithelial to mesenchymal transition (EMT), MSS/TP53$^{+}$, MSS/ TP53$^{-}$. These subtypes have various molecular alterations, disease progression, and prognosis. Each subtype differs from the others based on their molecular characterization (Cristescu *et al.*, 2015).

MSS/ TP53$^{+}$ and MSS/ TP53$^{-}$ subtypes have been selected for further study, both subtypes are classified based on TP53 activity which is the most frequently mutated gene in GC. MSS/Tp53$^{+}$ group has intact TP53 activity and *MDM2* overexpression (Cristescu *et al.*, 2015). MSS/Tp53$^{-}$ group has genomic instability and TP53 mutation and recurrent amplification (Cristescu *et al.*, 2015). Despite this molecular subtyping our knowledge of different biological processes and pathways for each subtype is still dismal and therefore, systematic approaches including high-performance computational methods can give us a great perspective about the molecular mechanism behind the initiation and progression of different subtypes of TP53 activity related subtypes of gastric cancer. To gain such molecular understanding, the use of high throughput data such as microarray data and analyzing gene expression profiles has immense importance. Microarray data analysis gives out information regarding the complete transcription profiles of the cancer cells. Proper use of this data in a certain computational methodology could result in a better understanding of the biological processes and pathways underlying the progression of these sub-types. Also, signature genes and biomarkers in terms of cancer prognosis and progression can be explored (Kim *et al.*, 2004).

The weighted gene co-expression network analysis (WGCNA) is a network-based analysis that uses transcriptomics data to find highly co-expressed gene modules. Modules are groups of genes whose expression profiles are highly correlated across the samples (Zhang and Horvath, 2005) hence these genes are involved in certain biological processes and pathways. Moreover, a supervised decision tree algorithm allows building simple classifiers with gene expression data that can assign a label (such as target clinical traits). Supervised decision trees have been widely used for the classification of gene expression data (Dettling and Bühlmann, 2003). A developed decision tree model (based on a certain train set) can predict a target clinical trait with a particular prediction accuracy (based on a certain test set) in which demonstrates predictor variable capability to model a certain tree to predict a target variable. The tumor pathological stage can be used as an indicator of the advancement of the disease and may help the physician to predict how quickly cancer would spread, therefor this clinical trait is important in cancer prognosis (Edge and Compton, 2010). With the integration of external biological information, the implemented classifiers output has a necessary connection with the biological information and it can give an insight on the functional annotation of the data (Hira and Gillies, 2015) and directs the computational methodology towards certain biological information which determines cancer progression clinically.

We have developed a two-step workflow to explore significant gene signatures in two gastric cancer subtypes in this study (Fig. 1). First, we implemented a network-based approach (WGCNA) on the transcriptomics data of two subtypes of GC (MSS/ TP53+, MSS/ TP53-). WGCNA was used to configure important co-expressed networks (modules) for each subtype and then based on a machine learning approach (decision tree) we chose the most important module based on its ability to predict tumor pathological stage for each subtype. In the next step using the STRING database, we configured conserved experimentally validated interaction networks for the selected modules (based on the

accuracy of predicting the tumor pathological stage) (Szklarczyk *et al.*, 2015). Since both modules contained a large number of genes and hence interactions (MSS/ TP53+: 242 genes, MSS/ TP53-: 1449 genes) a motif detection and motif ranking analysis was performed to explore the most important gene signatures for each subtype.

The developed workflow was led to introduce five significant 3-node gene motifs for two molecular subtypes of gastric cancer, which can be considered as potential diagnostic and therapeutic markers. The developed methodology can also be used in the case of other complex biological phenotypes to unravel important signatures.

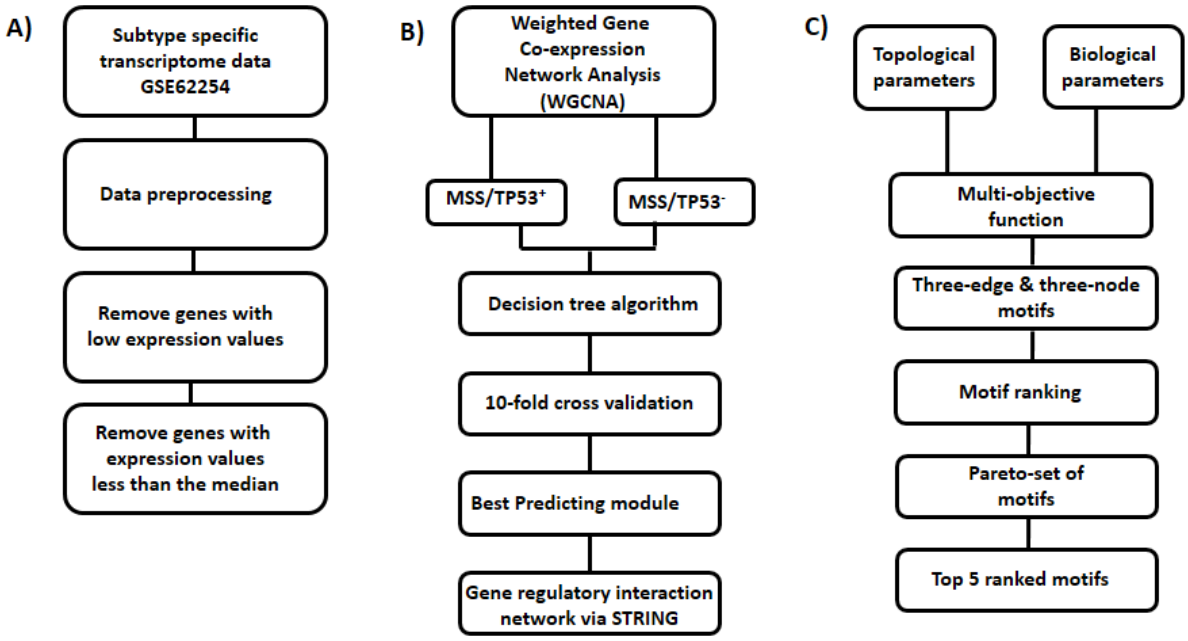


**Fig. 1.** Data analyzing workflow: Panel (A) shows the steps involved with data preprocessing including data normalization and feature filtering for GC subtype MSS/TP53$^+$ and MSS/TP53; Panel (B) shows the construction of weighted gene co-expression network and the steps of model tuning. The data is split into training and test sets. Models are tuned based on the training set and evaluated using the test set. Modules which best predict clinical trait using decision tree were selected and gene regulatory network analysis was performed, respectively; Panel (C) contains the motif exploring and motif ranking steps.

## Materials and Methods

### Datasets

The gene expression profile with the GSE62254 accession number from the Gene Expression Omnibus (GEO) database was used in this study. This dataset contains expression and clinical data for 300 primary gastric cancer (GC) patients. The Asian cancer research group defined 4 distinct GC molecular subtypes (MSI, MSS/EMT, MSS/TP53$^+$, TP53$^+$) from these samples which MSS/TP53$^+$ and MSS/TP53$^-$ subtypes (include 79 and 107 samples respectively) were analyzed in this study

### Data preprocessing

The R programming language (version 3.4.2) was used for statistical analysis in this study. preprocessing and normalization of raw data were performed using the “oligo” R package (Carvalho and Irizarry, 2010). The used microarray platform (Affymetrix human genome U133 plus 2.0 array) contains 54675 annotated genes. To decrease the number of input genes

and to avoid noisy results in WGCNA, some filtering methods should be done. First, we used hgu95av2.db R package and gapFilter from gene filter R package to select that one which has the most value of variance between Affymetrix IDs of one Entrez ID. After removing low-expressed genes the genes that had expression value more than mean in at least %25 of samples, selected. Using these filtering methods allows us to decrease the number of genes to 10377 and 10308 in subtypes TP53- and TP53+ respectively.

**Construction WGCNA**

The WGCNA for each subtype were constructed using the WGCNA-R package (Langfelder & Horvath, 2008). This procedure requires the following steps: Construction of a primary weighted co-expression network (similarity matrix) using Pearson's correlation coefficients for gene pairs.

$$S_{ij} = |\frac{1 + Cor(i,j)}{2}|$$

The $S_{ij}$ is the similarity and *cor(i,j)* is the Pearson correlation of *ith* and *jth* genes.
By raising the similarity measure to the power β, adjacency for a weighted co-expression network can be calculated by the $a_{ij} = |S_{ij}|^{\beta}$ formula.
Scale-free topology and power-law distribution of degree are key characteristic properties of gene co-expression networks. To construct the adjacency matrix, we selected the β power which complies with the following standards:
1. $R^2 > 0.8$
2. High mean connectivity
3. The slope of the regression line between log10(p(k)) and log10(k) should be near -1

**Network modules**

Clustering genes into highly co-expressed gene modules is the most used application of WGCNA. To grouping genes, we need an average linkage hierarchical clustering to accompany by a topologically based dissimilarity function. The topological overlap matrix (TOM) combines the connection strength between a pair of genes with their connections to other genes and the dissimilarity which calculated using TOM is considered as input in average linkage hierarchical clustering. Modules are branches of the resulting cluster tree and the cut-tree hybrid method was used to choose a cut-off height and minimum module size to identify modules. Dissimilarity measure based on TOM calculated as follow:

$$dissTOM_{ij} = 1 - TOM_{ij} = 1 - \frac{\sum_{u \neq i,j} a_{iu} a_{uj} + a_{ij}}{min\{k_i, k_j\} + 1 - a_{ij}}$$

Where $k_i = \sum_{u \neq i,j} a_{ui}$ indicates the network connectivity.

**Decision tree machine learning algorithm**

The decision tree method was used as a predictive machine learning technique. This method was implemented to select the most informative module in each subtype which can predict the tumor pathological stage more accurately. The R implementation of the CART algorithm (Decision tree is also known as Classification and Regression Trees (CART)) is called RPART (Recursive Partitioning and Regression Trees) that is available in "rpart" R package. A 10-fold cross-validation method was performed to evaluate model accuracy. In this analysis, the data set was split into 10 segments randomly that 9 of them were training data and the $10^{th}$ was test data. The CART modeling via rpart method was used to build a prediction model using training data for each module of subtypes. Then, models were evaluated by cross-validation analysis. The module with the highest prediction accuracy of each subtype was selected for subsequent analysis. caret (Kuhn, 2013) and rpart R packages were used to implement the decision tree algorithm in R.

**STRING database**

The STRING database is a precious global resource for the exploration and analysis of functional gene/protein interactions (Mering *et al.*, 2003). We use STRING to find conserved experimentally validated gene-gene interaction networks for the selected modules in the previous step. To create a STRING network the list of genes that present in the selected modules are entered into the STRING database to create conserved experimentally validated networks.

**Network motif identification**

Networks consist of smaller and repetitive structural units which are called a motif. Motifs

have an important role in biological networks and it is suggested that they accomplish overriding functions in biological networks. In this study, Cytoscape (Szklarczyk *et al.*, 2016) NetMatchStar plugin (Rinnone *et al.*, 2015) was used to find 3-node 3-edge network motifs in the selected gene regulatory networks for each subtype.

**Motif ranking**

To find the most important motifs in the respected networks, a previously developed motif ranking scheme by Khan *et al.* was implemented. The scheme is based on different topological and biological properties of involved genes in each motif. These properties contain (i) Topological parameters of Motif nodes including node degree and betweenness centrality, (ii) presence of motif genes in KEGG's "Pathways in cancer" pathway (KEGG: hsa05200), (iii) the gene prioritization score from Cytoscape GPEC plugin (Le and Pham, 2017); and (iv) gastric tumor subtype-specific gene expression log2 fold change in the transition from normal gastric tissue to tumor phenotype. To rank the explored motifs based on the mentioned parameters, the weighted multi-objective function was applied in the following formula.

$$GS_{ij} = \frac{w_{1j}}{2}.\frac{\langle nD\rangle_i}{\max(nD)} + \frac{w_{1j}}{2}.\frac{\langle nB\rangle_i}{\max(nB)} + w_{2j}.\frac{\langle PC\rangle_i}{\max(PC)} + w_{3j}.\frac{\langle GPS\rangle_i}{\max(GPS)} + w_{4j}.\frac{\langle |LFC|\rangle_i}{\max(LFC)}$$

$GS_{ij}$ is grade score for each motif (i= 1… n) in different weighting scheme (j= 1… 13) as said in supplementary Data1. Different weighting values including $w_{1j}$ to $w_{4j}$ are used to strike importance of used factors, $\langle nD\rangle_i$: average node degree for motif's node, $\langle nB\rangle_i$: average betweenness centrality of each node in a motif, $\langle PC\rangle_i$: number of genes in a motif involved in "pathways in cancer" KEGG pathway, $\langle GPS\rangle_i$: average gene prioritization score obtained from GPEC, $\langle |LFC|\rangle_i$: average absolute log2 fold change for the motif I (Khan *et al.*, 2017). Five different sets of weighting scenarios including 13 different weighting schemes were applied (Table 1) to remove biases between used parameters in motif prioritization. Each set pays more attention to specific parameters in Eq. 1. In the first set, only one parameter is more important for ranking. In sets 2-4, two, three, and four parameters are important respectively, and constantly have higher weights to the absolute LFC of the motif to explore tumor subtype-specific top-ranked motifs. In the fifth set, equal weights are allocated to all the parameters. These weighting schemes lead to a 13 ranking score for each motif. After removing duplicated motifs, we selected five top motifs from 13 ranking score output to further analysis.

**Table 1.** Weighting scenarios for motif ranking.

| Sets | $w_1$ | $w_2$ | $w_3$ | $w_4$ |
|---|---|---|---|---|
| | 1 | 0 | 0 | 0 |
| Set 1 | 0 | 0 | 1 | 0 |
| | 0 | 0 | 0 | 1 |
| | 1.4 | 0 | 0 | 3.4 |
| Set 2 | 0 | 1.4 | 0 | 3.4 |
| | 0 | 0 | 1.4 | 3.4 |
| | 1.8 | 1.8 | 0 | 3.4 |
| Set 3 | 1.8 | 0 | 1.8 | 3.4 |
| | 0 | 1.8 | 1.8 | 3.4 |
| | 1.16 | 1.16 | 1.8 | 3.4 |
| Set 4 | 1.16 | 1.8 | 1.16 | 3.4 |
| | 1.8 | 1.16 | 1.16 | 3.4 |
| Set 5 | 1.4 | 1.4 | 1.4 | 1.4 |

## Results

**WGCNA**

As mentioned in the methods section, to avoid the noisy result in WGCNA analysis, we applied a filtration step on expression data to reduce the number of genes. This step reduced the number of genes from 54675 to 10377 and 10308 in Tp53- and Tp53+ subtypes, respectively. The co-expressed gene modules were identified using the WGCNA R package. The most important step in WGCNA is choosing the power value, which has to satisfy scale-free topology and power-law distribution of degree in co-expressed modules. Eleven and 12 co-expression modules were identified using power 4 and 5 (Supplementary Table 1) in subtypes Tp53- and Tp53+, respectively. Modules are shown in different colors in cluster genes dendrogram for each subtype (Fig. 2).

**Decision tree**

The decision tree algorithm was implemented to find out the best predictor modules among identified modules to predict the tumor pathological stage. Decision trees for each module constructed by training data and the accuracy of each tree were calculated using test data (Table 2). As shown in table1, module

brown in subtype Tp53- and module green-yellow in subtype Tp53+ had the highest value of accuracy among modules and this means that these modules can predict the pathological stage of the tumor precisely.

## STRING database

After identification of the most significant modules for prediction of tumor pathological stage, the STRING database was used to explore conserved experimentally validated gene-gene interaction networks from the identified modules (brown and green-yellow modules for TP53- and TP53+ subtypes, respectively) in fig. 3.

**Table 2.** The table shows identified modules for each subtype. *

| Modules | TP53$^-$ | | Tp53$^+$ | |
|---|---|---|---|---|
| | **Freq** | **Accuracy*** | **Freq** | **Accuracy**** |
| Black | 497 | 31.25 | 504 | 40.74 |
| Blue | 1600 | 28.12 | 1906 | 29.63 |
| Brown | 1448 | 83.02 | 1617 | 25.93 |
| Green | 940 | 31.25 | 769 | 48.15 |
| Greenyellow | - | - | 241 | 55.56 |
| Grey | 2 | 46.88 | 110 | 48.15 |
| Magenta | 340 | 34.38 | 342 | 29.63 |
| Pink | 380 | 15.62 | 502 | 22.22 |
| Purple | 231 | 18.75 | 314 | 33.33 |
| Red | 600 | 21.88 | 603 | 25.93 |
| Turquoise | 2939 | 34.38 | 2617 | 40.74 |
| Yellow | 1400 | 34.38 | 783 | 40.74 |

*Freq column represents the number of genes for each module in a specific subtype and the second column shows obtained accuracy from the decision tree for each module; **Accuracy=%

## Network Analysis

After constructing a conserved gene regulatory network for each module, via STRING, a 3-edge 3-node motif exploring was performed (Supplementary data1). These motifs constitute basic regulation and protein organization patterns into modules. They represent the multiplexes of gene interactions that work together as a multi-component machine (Yeger-Lotem *et al.*, 2004). For the green-yellow module interaction network in TP53$^+$ we identified 17 motifs and for the brown module interaction network in TP53$^-$ we identified 1582 motifs, respectively (see supplementary data 1). From the identified motifs, we aimed to find the most important ones in each network. Towards this end, a motif ranking scheme using a multi-objective function for motif prioritization was performed for both interaction networks (supplementary data1). Motif ranking was based on both topological and biological parameters. Subsequently, we took the top 5 ranking motifs for each subtype (each containing three genes) and analyzed their enrichment and annotation regarding how they relate to the advancement of GC (tumor pathological stage). Some of the genes in these motifs were repeated constantly and constructed a sub-network in each subtype. These sub-networks may have important regulatory roles in both GC subtypes.

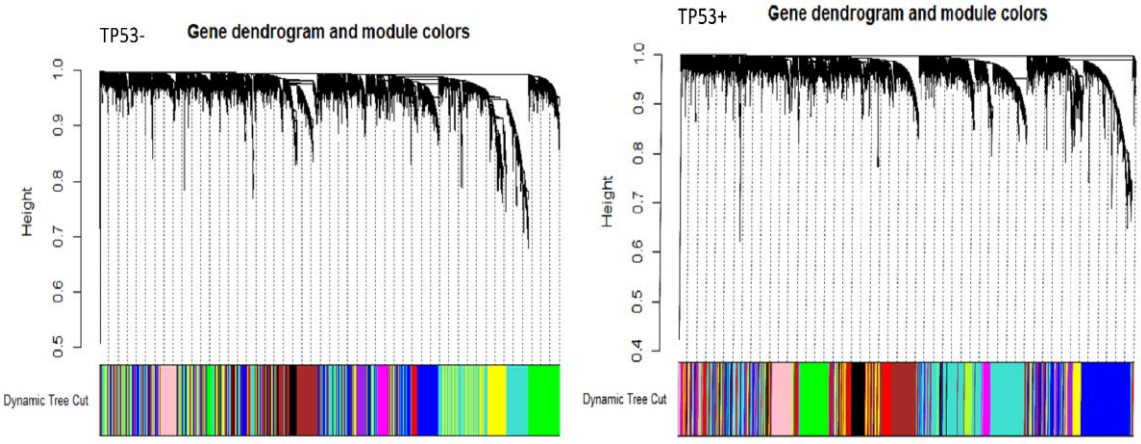


**Fig. 2.** Gene dendrogram of TP53$^-$ and TP53$^+$ gastric cancer subtypes: The hierarchical clustering tree was made using the dissimilarity measure based on Topology Overlap Matrix (dissTOM) of genes' expression values. The dynamic tree-cutting algorithm was used to create modules using the hierarchical clustering tree. This algorithm puts the genes with the lowest dissTOM in the same modules. 11 and 12 modules were identified for TP53$^-$ and TP53$^+$ respectively. Each color is assigned for each module as an identifier. The colored row below the dendrogram shows the merged modules.

## Discussion

Heterogeneity of gastric cancer has led to challenges in treating GC patients and it is one of the major reasons for poor clinical outputs of GC therapy. Therefore, in this study with the use of high-throughput data analyzing we aimed to pave the way to explore better prognostic biomarkers and gene signatures. In this study exploration of co-expression gene modules constructed from Transcriptomics data of 2 GC subtypes took place. Afterward, the module that was best able to predict the pathological stage of the tumor was identified using a machine learning approach (decision tree algorithm). The experimentally validated gene network of the selected module that was extracted from the STRING was analyzed to find the most significant functional motifs. These motifs are structural parts of the gene expression network which have important regulatory and phenotypic outcomes. All of the explored genes in the five top-ranked motifs for $TP53^-$ subtype (Table 3) have already been reported for GC in previous studies except *ATP6V0A1* and *ATP6AP1*. This shows the robustness of the computational approach which led to genes that are already known responsible for the progression of GC.

All 3 genes of Motif 1 are members of the Mitogen-activated protein kinases family and the main MAPK pathway which is *p38*. *MAPK14* is one of the four *p38* MAPK members that as shown in Fig. 3, is repeated in 3 top motifs. Thereby, this gene has a key role in the represented sub-network ($TP53^-$ brown module). In response to inflammatory cytokines or environmental stresses, MAP2K3 or MAP2K6 activate *MAPK14* by dual phosphorylation of the Thr-Gly-Tyr amino acid motif. A member of the *MAP3K* family such as *MAP3K7* activates the *MAP2K* tier, depending on the tissue and the stimuli type (Pritchard and Hayward 2013). Several studies have reported these genes as gastric cancer prognostic markers (Katoh and Katoh, 2009; Liu *et al.*, 2014; Parray *et al.*, 2014). Based on this, the explored motif which it's members are already known as responsible in GC development is capable of regulating and advancement of GC. However, their accurate course of action in GC progression and their ability to predict the tumor pathological stage can be studied more thoroughly in future studies. The repetition of these genes in our results is an indicator of the robustness of the analytical approach that took place in this study. All genes in motif 2 encode the V0 subunit of *V-ATPase* (Xu *et al.*, 2013) that they have not reported as GC related genes. However based on previous studies it is stated that they have a role in other types of cancer (Antonacopoulou *et al.*, 2008; Arif *et al.*, 2015; Hsin *et al.*, 2012). The effect of these genes and the underlying mechanism leading to GC progression is yet to be distinguished. However, they can be proper targets for future experimentations of GC related studies. The oncogenic behavior of the respected genes is proven in other previous studies but their course of action in GC progression is yet to be known.

**Table 3.** Five top-ranked motifs for subtypes $TP53^+$ and $TP53^-$.

| Top5 motifs | | GENE1 | GENE2 | GENE3 | Ranking Score |
|---|---|---|---|---|---|
| $TP53^+$ | 1 | *VAMP8* | *SEC22C* | *USO1* | 0.987451 |
| | 2 | *RBM22* | *SRRM2* | *SF3B1* | 0.980769 |
| | 3 | *SRRM2* | *PRPF6* | *SF3B1* | 0.932967 |
| | 4 | *HNRNPC* | *SRRM2* | *SF3B1* | 0.928079 |
| | 5 | *HNRNPC* | *RPL6* | *SRRM2* | 0.861993 |
| $TP53^-$ | 1 | *MAP2K6* | *MAPK14* | *MAP3K7* | 0.984842 |
| | 2 | *ATP6V0E2* | *ATP6V0A1* | *ATP6AP1* | 0.981421 |
| | 3 | *UBC* | *HSP90AB1* | *RPS6* | 0.979435 |
| | 4 | *PTPN11* | *MAP2K1* | *MAPK14* | 0.977016 |
| | 5 | *MAPK14* | *MAP3K7* | *HSP90AA1* | 0.961919 |

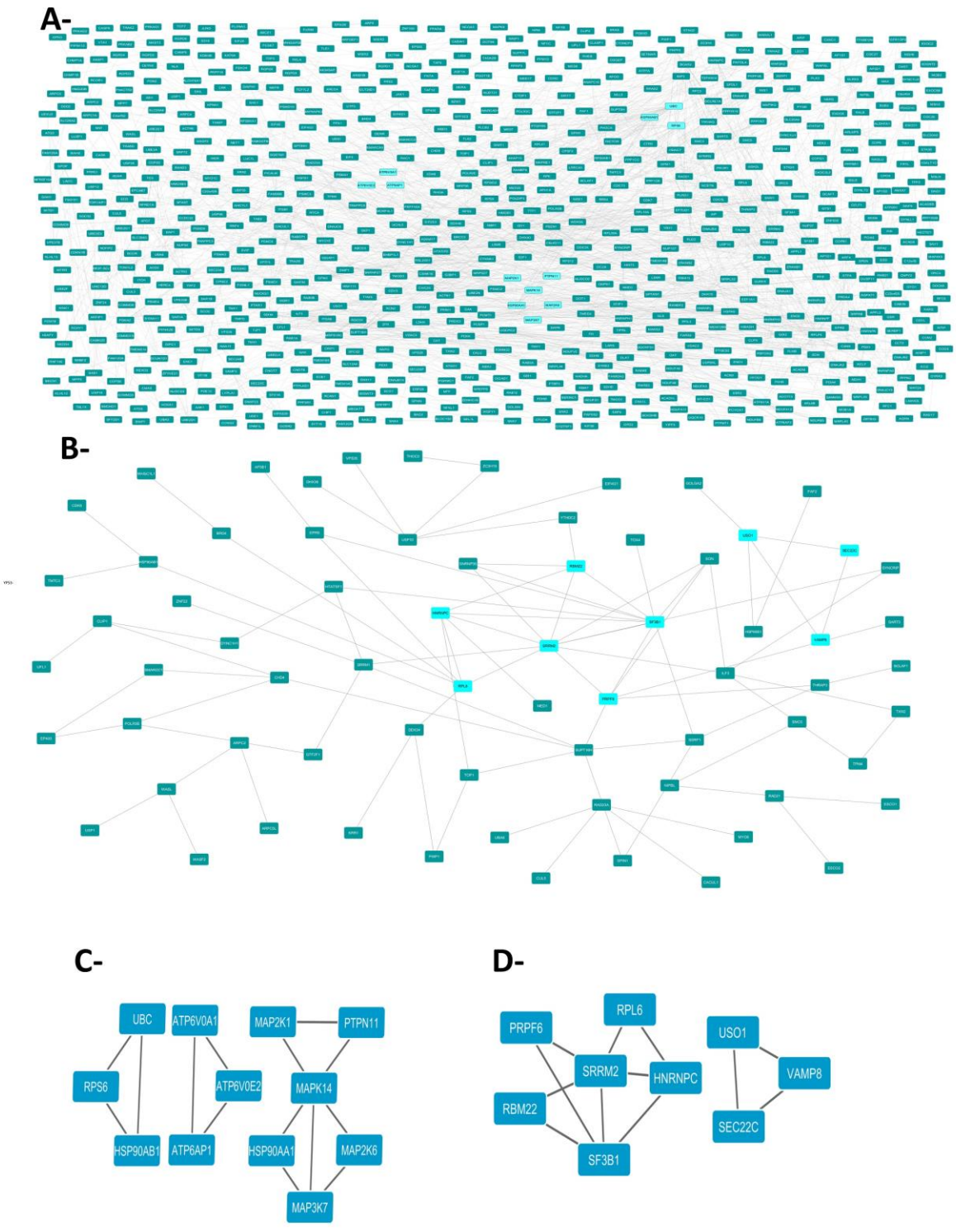


**Fig. 3**. Conserved experimentally validated STRING networks and the explored top-ranked motifs for TP53$^{-}$ and TP53$^{+}$ gastric cancer subtypes: A) STRING conserved experimentally validated network for the genes present in the TP53$^{-}$ subtype brown module; B) STRING conserved experimentally validated network for the genes present in the TP53$^{+}$subtype green-yellow module; Explored top-ranked motifs in the STRING networks highlighted with turquoise color nodes (A and B); The explored motifs from the STRING networks represented separately for TP53$^{-}$ and TP53$^{+}$ subtypes in panel C and D respectively.

*Polyubiquitin-C* which encoded by *UBC* exhibited an interaction with *calgranulin B*. Kim *et al.* reported the role of *calgranulin B* in gastrointestinal cancer cells and states that *calgranulin B* has a relationship with the tumor extracellular environment via *polyubiquitin-C* (Kim *et al.*, 2017).

Results of one study suggested that *HSP90* may play an important role in tumor invasion, metastasis, and prognosis, and could be a potential prognostic factor of gastric cancer (Wang *et al.*, 2013). Comparing the expression value of *HSP90AB1* between normal and cancerous gastric tissue, represented that expression of *HSP90* beta was increased in gastric cancer (Liu *et al.*, 1999). The respected expressional behavior is also witnessed in our results. Phosphorylation has a large influence on the function and regulation of *Hsp90* (Zuehlke *et al.*, 2015). *Hsp90* facilitates protein folding and leads proteins for degradation via the ubiquitin pathway. These opposite activities are done by *HSP90* when they bind to co-chaperones (*CHIP* and *HOP*) at C-termini (Muller *et al.*, 2013). In primary cancers, Phosphorylation of the c-terminal of *Hsp90α* is increased that leads to the elevation of *HOP* interaction with *HSP90α*, which leads to a high proliferation of cells, and faster tumor growth (Zuehlke *et al.*, 2015).

Analyses of common genes of the esophageal, gastric, and colon cancers showed that *HSP90AA1* is a gastrointestinal cancer-related gene which can be considered as a predictive biomarker for these cancers (Maghvan *et al.*, 2017). Another study showed that *HSP90AA1* up-regulated in metastatic GC compared with primary GC at transcriptional and translational levels, and also up-regulated at the translational level in primary GC compared with normal mucosa (Chang *et al.*, 2009). Thereby, since this gene has a known role in cancer progression its expression profile can be an important factor in determining the cancer stage. Based on its biological function and our findings differential expression of the respected gene is responsible for GC progression and based on our computational methodology its expression profile can have immense importance in predicting the tumor pathological stage. However, to determine the precise expression profile relationship with the tumor pathological stage more thorough experimental and computation experimentations are needed.

Studies showed that the Ribosomal protein family has a strong association with GC. Jiang *et al* (Jiang, Li, Jiang, & Shao, 2017), reported that reduction of phosphorylation of *RPS6* could alter *MEK* inhibition sensitivity of gastric cancer cells (Jiang *et al.*, 2017). Studies report the overexpression of *RPS6* in eight different colon cancers and adenomatous polyps. In other words, *RPS6* dysregulation may be a carcinogenic factor in gastric cancer (Guo *et al*., 2011) If so, *RPS6* can be a biomarker for GC progression (specifically the studied subtype) hence its pathological stage. *MAP2K1* in motif 4 also is one of the *MAPK* family that was reported as a gene which can predict the survival of GC patients (Xu *et al.*, 2010). This shows the importance of the respected gene in cancer progression and based on our computational approach and it is a clinical trait it can be used as a marker for determination of tumor pathological stage in the studied subtype.

*PTPN11* is a Protein Tyrosine Phosphatase that is upregulated in the $TP53^{-}$ subtype and according to previous reports it was overexpressed in tubular and intestinal types of gastric cancer tumor cells and may have a key role in gastric cancer pathogenesis thorough Helicobacter pylori infection (Kim *et al.*, 2010). However, its impact on the tumor pathological stage should be studied more thoroughly.

Regarding the $TP53^{+}$ subtype of GC, *VAMP8, SEC22C, USO1, RBM22, SRRM2, SF3B1, PRPF6, HNRNPC* and *RPL6* were genes in the five motifs with the highest-ranking score. Besides analyzing each gene and their relationship with GC separately, demonstrating each motif's relationship with GC in this research is of importance.

The first motif of $TP53^{+}$ contains three protein-coding genes as follows; *VAMP8, SEC22C, USO1*. *VAMP8* belongs to the synaptobrevin/vesicle-associated membrane protein subfamily of soluble N-ethylmaleimide-sensitive factor attachment protein receptors (*SNAREs*) (Wong *et al.*, 1998). Both *SEC22C* and *USO1* have *SNARE* binding functions, also recent studies have reported that SNARE proteins have an important role in tumorigenesis with several functions. These roles consist of

diverse functions including *SNARE-mediated trafficking* in tumor progression, cell migration, inflammatory response, autophagy, and cell survival in tumorigenesis (Meng and Wang, 2015). Differentiated expression of *SEC22C* and *USO1* indicates that regulation of *SNAREs* and *SNARE* binding proteins is compromised therefore *SNARE* proteins especially *VAMP8* could be a potential therapeutic target towards gastric cancer. Recent studies reported that *VAMP8* is associated with the development of tumors (Meng and Wang, 2015) and since the tumors, pathological stage represents cancer development and progression clinically this shows that *VAMP8* can be involved in GC development. Also, since the biological background stated that this gene has invasive carcinogenesis behaviors, it can be used as a biomarker for predicting the tumor pathological stage. This gene can be a target for future expression profiling studies in terms of exploring its precise ability to predict the tumor pathological stage as a biomarker.

All of the second, the third, fourth, and fifth motif of GC subtype TP 53+ genes; *SF3B1, SRRM2, RBM22, PRPF6, HNRNPC* (except *RPL6*), are a part of "mRNA splicing –major pathway" and "mRNA splicing via spliceosome" (GO:0000398), moreover a highly mutable gene such as *SF3B1* causes alternative splicing (AS) and miss splicing in which recent studies showed to have a direct relationship with gastric cancer advancement (David and Manley, 2010). Alternative splicing causes flexibility that cancer cells often use to their advantage to produce proteins that promote their growth and survival (David & Manley, 2010). recent studies proved that *PRPF6* motivates cancer proliferation by preferential splicing of genes that are responsible for the regulation of growth and the inhibition of such gene (selectively) abrogated the cancer growth which indicates the major importance of the latter gene in cancer occurrence and progression via splicing of distinct growth-related gene products (Adler et al., 2014).

These genes are not known to be directly gastric cancer related (except *HNRNPC*), although the relationship of alternative splicing and aberrant splicing with multiple cancer types and GC is inevitable. Also, this shows the importance of AS to be potentially involved in the progression and development of $TP53^+$ GC. Therefore, based on the computational methodology performed in this study we can reach two conclusions: 1- the importance of AS in GC progression and 2- the ability of the latter's genes expression profiles (combined) in determining the tumors pathological stage. However, all of these claims must be studied more thoroughly in the future and this hypothesis is based on the known biological behavior of these genes and their role as carcinogenesis and our computational approach. *SRRM2* and *SF3B1* are seen repeatedly in our motif list which may indicate the importance of *SRRM2* and *SF3B1* in the occurrence and development of gastric cancer, studies have also detected relevance of the mutation of both genes with various cancer types. For example *SF3B1* mutation (splicing factor gene) in chronic lymphocytic leukemia (Quesada *et al.*, 2012) and *SRRM2* (splicing factor gene) germline mutation in papillary thyroid carcinoma (Tomsic *et al.*, 2015) .however they are not known to be gastric cancer ($TP53^+$) related. Heterogeneous Nuclear RibonucleoproteinC (*HNRNPC*) has been studied and proved to be gastric cancer-related. Overexpression of *HNRNPC* promotes chemoresistance (up-regulation of this gene in our samples was also significant with the LFC of 1.951). Recent studies demonstrated this gene as a potential prognostic and therapeutic marker for GC (Huang *et al.*, 2016). Based on our computational method this gene is highly co-expressed with *SRRM2* and *SF3B1*, the importance and relation of these two genes are demonstrated towards advancement and chemoresistance of GC ($TP53^+$), respectively. This data suggests the usefulness of *SRRM2* and *SF3B1* alongside *HNRNPC* as a potential biological marker for GC ($TP53^+$).

Regarding the fifth motif of $TP53^+$, recent studies revealed that human ribosomal protein L6 (*RPL6*) has a role in protecting gastric cancer cells from drug-induced apoptosis and it can be used as a novel approach towards GC ($TP53^+$) therapy (Wu *et al.*,2011).

*RPL6* and *HNRNPC* both cause multidrug resistance. Inhibition of these genes both abrogated the growth of GC cells. This reveals the importance of these drug resistance genes in therapy (gene therapy) via RNA interference for

GC TP53$^{+}$. Moreover, the inner relationship of TP53$^{+}$ genes present *SRRM2* (splicing factor gene) as a hub gene (see Fig. 3) which demonstrates that AS (especially *SRRM2*) is utterly important in cancerous functions that are mentioned earlier (such as drug resistance). To clarify, this could indicate that the flexibility and diversity that cancerous cells develop in terms of becoming resistant to apoptotic drugs may be due to behaviors caused by AS and related genes (*SRRM2*, *SF3B1*). Based on this hypothesis their expression profile can be a great target for GC patients in terms of pathological prediction of patients.

## Conclusion

Tumor heterogeneity is one of the major challenges in analyzing multi-omics data (genomic, transcriptomic, proteomic, and metabolomics) from the sample pool. This heterogeneity is the basic concept for precision medicine. Inter-tumor and intra-tumor heterogeneity are one of the major factors in the diversity of prostate tumors that arise from genetic variations between tumor cells. The reasons for these heterogeneities are not well understood but they are critical in effective diagnosis and treatment of patients which are considered in this study. In this study, a two-step workflow has been used to explore significant gene signatures in two gastric cancer subtypes. First, a network-based approach (WGCNA) was implemented on the transcriptomics data of two subtypes of GC (MSS/ TP53$^{+}$, MSS/ TP53$^{-}$) in order of exploring important co-expressed networks (modules) for each subtype and then based on a machine learning approach (decision tree) we selected the best predicting module based on its ability to predict tumor pathological stage for each subtype based on the implemented ML algorithm and its prediction accuracy in each module. In this step, the hypothesis is that each module's test set capability to predict each patient's pathological stage determines the quality of the learned machine based on the train set (accuracy of the machine determines the characteristics of the gene set in each module). To clarify, the best predicting accuracy states that the respected module had better characteristics for pathological stage prediction. This means that the genes constructing the respected module can be responsible for cancer progression and are best capable of predicting the tumor pathological stage as biomarkers. In the next step using the STRING database, we configured conserved an experimentally validated interaction network for the selected modules. Since both modules contained a large number of genes and hence interactions (MSS/ TP53$^{+}$: 242 genes, MSS/ TP53$^{-}$: 1449 genes) a motif ranking approach was performed to explore the most important gene signatures for each subtype. Motif exploration and motif studies in this type of large network have led to a better understanding of the key and core regulatory features. In this study, we introduced five significant three-node gene motifs for two molecular subtypes of gastric cancer which can be considered as potential diagnostic and therapeutic markers. The developed methodology can also be used in the case of other complex biological phenotypes to unravel important signatures. Since all of the explored results had somehow a biological relationship with tumor progression, the robustness of this computational method on analyzing high throughput data and large expression networks with the use of patient clinical data is shown. The most important aim of this study was to pave the way for analyzing such data for future similar studies.

## Acknowledgments

This research was supported by the University of Tabriz and Semnan University. We would like to thank the Research Institute for Fundamental Sciences (RIFS)-University of Tabriz for financial supports.

## CRediT authorship contribution statement

We confirmed that, all authors were involved in writing this article. M. Sadeghi and N. Ghorbanpour have contributed equally to this article in the Conceptualization, Methodology, statistical analysis, programming and writing; A. Barzegar and I. Rafiei contributed in Conceptualization and Reviewing.

## References

Adler AS, McCleland M L, Yee S, Yaylaoglu M, Hussain S, Cosino E, Chopra VS. 2014. An integrative analysis of colon cancer identifies

an essential function for PRPF6 in tumor growth. *Genes Dev* 28:1068-1084.

Antonacopoulou AG, Grivas PD, Skarlas L, Kalofonos M, Scopa CD, Kalofonos HP. 2008. POLR2F, ATP6V0A1 and PRNP expression in colorectal cancer: new molecules with prognostic significance? *Anticancer Res* 28:1221-1227.

Arif S, Qudsia S, Urooj S, Chaudry N, Arshad A, Andleeb S. 2015. Blueprint of quartz crystal microbalance biosensor for early detection of breast cancer through salivary autoantibodies against ATP6AP1. *Biosens Bioelectron* 65:62-70.

Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. 2018. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin* 68:394-424.

Carvalho BS, Irizarry RA. 2010. A framework for oligonucleotide microarray preprocessing. *Bioinformatics* 26:2363-2367.

Chang W, Ma L, Lin L, Gu L, Liu X, Cai H, Zhang M. 2009. Identification of novel hub genes associated with liver metastasis of gastric cancer. *Int J Cancer* 125:2844-53.

Cristescu R, Lee J, Nebozhyn M, Kim KM, Ting JC, Wong SS, Xiang SY. 2015. Molecular analysis of gastric cancer identifies subtypes associated with distinct clinical outcomes. *Nat Med* 21:449-56.

David CJ, Manley JL. 2010. Alternative pre-mRNA splicing regulation in cancer: pathways and programs unhinged. *Genes Dev* 24:2343-64.

Dettling M, Bühlmann P. 2003. Boosting for tumor classification with gene expression data. *Bioinformatics* 19:1061-1069.

Edge SB, Compton CC. 2010. The American joint committee on cancer: the 7th edition of the AJCC cancer staging manual and the future of TNM. *Ann Surg Oncol* (6):1471-1474.

Ferro A, Peleteiro B, Malvezzi M, Bosetti C, Bertuccio P, Levi F, Lunet N. 2014. Worldwide trends in stomach cancer mortality and incidence (1980-2011) and predictions to 2015. *Eur J Cancer* 50: 1330-1344.

Gullo I, Carneiro F, Oliveira C, Almeida GM. 2018. Heterogeneity in gastric cancer: from pure morphology to molecular classifications. *Pathobiology* 85:50-63.

Guo X, Shi Y, Gou Y, Li J, Han S, Zhang Y, Huo J, Sun S. 2011. Human ribosomal protein S13 promotes gastric cancer growth through down-regulating p27Kip1. *J Cell Mol Med* 15:296-306.

Hira ZM, Gillies DF. 2015. A review of feature selection and feature extraction methods applied on microarray data. *Adv Bioinform* 2015.

Hsin IL, Sheu GT, Jan MS, Sun HL, Wu TC, Chiu LY, Ko JL. 2012. Inhibition of lysosome degradation on autophagosome formation and responses to GMI, an immunomodulatory protein from Ganoderma microsporum. *Br J Pharmacol* 167:1287-300.

Huang H, Han Y, Zhang C, Wu J, Feng J, Qu L, Shou C. 2016. HNRNPC as a candidate biomarker for chemoresistance in gastric cancer. *Tumor Biol* 37:3527-34.

Jiang B, Li S, Jiang Z, Shao P. 2017. Gastric cancer associated genes identified by an integrative analysis of gene expression data. *Biomed Res Int* 2017 :7259097.

Katoh M, Katoh M. 2009. Transcriptional mechanisms of WNT5A based on NF-κB, Hedgehog, TGFβ, and Notch signaling cascades. *Int J Mol Med* 23:763-769.

Khan FM, Marquardt S, Gupta SK, Knoll S, Schmitz U, Spitschak A, Engelmann D. 2017. Unraveling a tumor type-specific regulatory core underlying E2F1-mediated epithelial-mesenchymal transition to predict receptor protein signatures. *Nat Commun* 8:1-5.

Kim IJ, Kang HC, Park JG. 2004. Microarray applications in cancer research. Cancer research and treatment: *J Korean Cancer Assoc* 36:207-213.

Kim JS, Shin OR, Kim HK, Cho YS, An CH, Lim KW, Kim SS. 2010. Overexpression of protein phosphatase non-receptor type 11 (PTPN11) in gastric carcinomas. *Dig Dis Sci* 55:1565-1569.

Kim KH, Yeo SG, Yoo BC, Myung JK. 2017. Identification of calgranulin B interacting proteins and network analysis in

gastrointestinal cancer cells. *PLoS One* 12:e0171232.

Kuhn M, Johnson K. 2013. A Short Tour of the Predictive Modeling Process. InApplied predictive modeling (pp. 19-26). Springer, New York, NY.

Langfelder P, Horvath S. 2008. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinform* 9:559. doi:10.1186/1471-2105-9-559.

Le DH, Pham VH. 2017. HGPEC: a Cytoscape app for prediction of novel disease-gene and disease-disease associations and evidence collection based on a random walk on heterogeneous network. *BMC Syst Biol* 11:61. doi:10.1186/s12918-017-0437-x.

Lin X, Zhao Y, Song WM, Zhang B. 2015. Molecular classification and prediction in gastric cancer. *Comput Struct Biotechnol J* 13:448-458.

Liu G, Li DZ, Jiang CS, Wang W. 2014. Transduction motif analysis of gastric cancer based on a human signaling network. *Braz J Med Biol Res* 47:369-375.

Liu X, Ye L, Wang J, Fan D. 1999. Expression of heat shock protein 90 beta in human gastric cancer tissue and SGC7901/VCR of MDR-type gastric cancer cell line. *Chin Med J* 112:1133-1137.

Maghvan PV, Rezaei-Tavirani M, Zali H, Nikzamir A, Abdi S, Khodadoostan M, Asadzadeh-Aghdaei H. 2017. Network analysis of common genes related to esophageal, gastric, and colon cancers. *Gastroenterol Hepatol Bed Bench* 10:295-302.

Meng J, Wang J. 2015. Role of SNARE proteins in tumourigenesis and their potential as targets for novel anti-cancer therapeutics. *Biochim Biophys Acta Rev Cancer* 1856:1-2.

Mering CV, Huynen M, Jaeggi D, Schmidt S, Bork P, Snel B. 2003. STRING: a database of predicted functional associations between proteins. *Nucleic Acids Res* 31:258-261.

Muller P, Ruckova E, Halada P, Coates PJ, Hrstka R, Lane DP, Vojtesek B. 2013. C-terminal phosphorylation of Hsp70 and Hsp90 regulates alternate binding to co-chaperones CHIP and HOP to determine cellular protein folding/degradation balances. *Oncogene* 32:3101-3110.

Parray AA, Baba RA, Bhat HF, Wani L, Mokhdomi TA, Mushtaq U, Khanday FA. 2014. MKK6 is upregulated in human esophageal, stomach, and colon cancers. *Cancer Invest* 32:416-422.

Pritchard AL, Hayward NK. 2013. Molecular pathways: mitogen-activated protein kinase pathway mutations and drug resistance. *Clin Cancer Res* 19:2301-2309.

Quesada V, Conde L, Villamor N, Ordóñez GR, Jares P, Bassaganyas L, López-Guerra M. 2012. Exome sequencing identifies recurrent mutations of the splicing factor SF3B1 gene in chronic lymphocytic leukemia. *Nat Genet* 44:47-52.

Rinnone F, Micale G, Bonnici V, Bader GD, Shasha D, Ferro A, Giugno R. 2015. NetMatchStar: an enhanced Cytoscape network querying app. *F1000Research.* 4: 479. doi: 10.12688/f1000research.6656.2.

Shimizu D, Kanda M, Kodera Y. 2017. Review of recent molecular landscape knowledge of gastric cancer. *Histol Histopathol* 33:11-26.

Siegel RL, Miller KD, Fedewa SA, Ahnen DJ, Meester RG, Barzi A, Jemal A. 2017. Colorectal cancer statistics, 2017. *CA Cancer J Clin* 67:177-193.

Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, Kuhn M. 2015. STRING v10: protein–protein interaction networks, integrated over the tree of life. *Nucleic Acids Res* 43: 447-452.

Szklarczyk D, Morris JH, Cook H, Kuhn M, Wyder S, Simonovic M, Jensen LJ. 2016. The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible. *Nucleic Acids Res* 45: 362–368.

Tomsic J, He H, Akagi K, Liyanarachchi S, Pan Q, Bertani B, De La Chapelle A. 2015. A germline mutation in SRRM2, a splicing factor gene, is implicated in papillary thyroid carcinoma predisposition. *Sci Rep* 5:1-3.

Wang J, Cui S, Zhang X, Wu Y, Tang H. 2013. High expression of heat shock protein 90 is associated with tumor aggressiveness and poor prognosis in patients with advanced gastric cancer. *PLoS One* 8:e62876.

Wong SH, Zhang T, Xu Y, Subramaniam VN, Griffiths G, Hong W. 1998. Endobrevin, a novel synaptobrevin/VAMP-like protein

preferentially associated with the early endosome. *Mol Biol Cell* 9:1549-1563.

Wu Q, Gou Y, Wang Q, Jin H, Cui L, Zhang Y, Fan D. 2011. Downregulation of RPL6 by siRNA inhibits proliferation and cell cycle progression of human gastric cancer cell lines. *PLoS One* 6:e26401.

Xu K, Mao X, Mehta M, Cui J, Zhang C, Mao F, Xu Y. 2013. Elucidation of how cancer cells avoid acidosis through comparative transcriptomic data analysis. *PLoS One* 8:e71177.

Xu ZY, Chen JS, Shu YQ. 2010. Gene expression profile towards the prediction of patient survival of gastric cancer. *Biomed Pharmacother* 64:133-139.

Yeger-Lotem E, Sattath S, Kashtan N, Itzkovitz S, Milo R, Pinter RY, Margalit H. 2004. Network motifs in integrated cellular networks of transcription-regulation and protein–protein interaction. *Proc Natl Acad Sci* 101:5934-5939.

Zhang B, Horvath S. 2005. A general framework for weighted gene co-expression network analysis. *Stat Appl Genet Mol Biol* 4 (1). doi: 10.2202/1544-6115.1128.

Zuehlke AD, Beebe K, Neckers L, Prince T. 2015. Regulation and function of the human HSP90AA1 gene. *Gene* 570:8-16.