Identification of key genes and pathways involved in microsatellite instability in colorectal cancer

CHAORAN YU[1,2*], HIJU HONG[1,2*], SEN ZHANG[1,2], YAPING ZONG[1,2], JUNJUN MA[1,2],
AIGUO LU[1,2], JING SUN[1,2] and MINHUA ZHENG[1,2]

[1]Department of General Surgery; [2]Shanghai Minimally Invasive Surgery Center, Ruijin Hospital,
Shanghai Jiao Tong University School of Medicine, Shanghai 200025, P.R. China

Received April 17, 2018; Accepted November 14, 2018

DOI: 10.3892/mmr.2019.9849

Abstract. Microsatellite instability (MSI) has emerged as one of the key biological features of colorectal cancer (CRC). However, controversies remain regarding the association between the MSI status and clinicopathological characteristics of CRC. Therefore, it is crucial to identify potential key genes and pathways associated with MSI in CRC. In the present study, the GSE25071 gene expression profile was retrieved, with thirty-eight cases of microsatellite stable (MSS), five of MSI-High (MSI-H) and three of MSI-Low (MSI-L) CRC patients. The differentially expressed genes (DEGs) were analyzed by Gene Ontology (GO), Kyoto Encyclopedia of Genes and Genomes pathway enrichment, gene set enrichment analysis (GSEA) and gene co-expression network analysis. Weighted gene correlation network analysis (WGCNA) was used for the gene modules and correlation of clinical traits. A total of forty-nine DEGs were identified between MSI-H and MSS, including six upregulated and forty-three downregulated DEGs. Only the DEGs of MSI-H and MSS were subjected to subsequent analysis (limited number of DEGs of MSI-L and MSS, MSI-H and MSI-L). RNA metabolic process, endoplasmic reticulum and chemokine receptor binding were the top ranked terms in GO enrichment. The hub genes of co-expression network of DEGs included zinc finger protein (ZNF) 813, ZNF426, ZNF611, ZNF320 and ZNF573. The GSEA of MSI-H and MSS indicated that the mammalian target of rapamycin complex 1 signaling was significantly enriched with a nominal P-value of 0.038 and normalized enrichment score of 0.446. The WGCNA results showed that the pink module was the top in correlation with MSI status (R^2=0.5, P=0.0004). The genes in the pink module were significantly enriched in proteins targeting to endoplasmic reticulum, cytosolic part, structural constituent of ribosome and ribosome pathway. The hub genes identified in the pink module were ribosomal protein L12 (RPL12), RPS3A, RPS9, RPL27A, RPL7, RPL28, RPL14, RPS17, mitochondrial ribosomal protein L16, and G elongation factor, mitochondrial 2. The present study identified key genes and pathways associated with MSI, providing insightful mechanisms.

Introduction

Colorectal cancer (CRC) is among the most common malignancies worldwide, with a high incidence in the United States and Western Europe (1). In 2017, >130,000 newly cases were diagnosed and >50,000 individuals succumbed to mortality in the United States (1). The incidence rates and mortality rates have declined in patients with CRC aged >50 years, however, the incidence rates have increased by 22% in patients aged <50 years and the mortality rates have increased by 13% in the last decade (1). In China, the incidence and mortality rates of CRC have markedly increased (2). Despite substantial efforts in the establishment of early detection systems and chemotherapy reagents, the prognosis of CRC remains far from satisfactory for the majority of patients (3).

The therapeutic responses and survival outcomes of CRC are constrained by the clinical heterogeneity (4,5). Therefore, molecular markers emerge as efficient classifiers for CRC (6,7).

Correspondence to: Dr Minhua Zheng or Dr Jing Sun, Department of General Surgery, Ruijin Hospital, Shanghai Jiao Tong University School of Medicine, 197 Ruijin Er Road, Shanghai 200025, P.R. China
E-mail: zmhtiger@yeah.net
E-mail: sj11788@rjh.com.cn

[*]Contributed equally

Abbreviations: CI, confidence intervals; GEO, Gene Expression Omnibus; HR, hazard ratio; CRC, colorectal cancer; GO, Gene Ontology; KEGG, Kyoto Encyclopedia of Genes and Genomes; PPI, protein-protein interaction; STRING, Search Tool for the Retrieval of Interacting Genes; MCODE, Molecular Complex Detection; BP, biological processes; CC, cellular components; MF, molecular functions; WGCNA, weighted gene correlation network analysis; MSS, microsatellite stable; MSI-H, microsatellite instability high; MSI-L, microsatellite instability low; MMR, mismatch repair; FU, fluorouracil; HNPCC, hereditary nonpolyposis colorectal cancer; UICC, International Union Against Cancer; ZNF, zinc finger protein; RP, ribosomal protein; CIN, chromosomal instability

Key words: microsatellite instability, differentially expressed genes, Gene Ontology, Kyoto Encyclopedia of Genes and Genomes, gene set enrichment analysis, weighted gene correlation network analysis, colorectal cancer, protein-protein interaction

This is exemplified by the mutations identified in genes such as KRAS/BRAF/phosphatidylinositol-4,5-bisphosphate 3-kinase catalytic subunit α (PIK3CA), which have been widely accepted as clinical indicators for therapeutic decisions (6). However, the clinical outcomes of CRC remain largely diverse.

Microsatellite instability (MSI) is one of the most intensively investigated molecular markers in CRC (3). MSI indicates inactivation of the mismatch repair (MMR) genes and is frequently associated with the CpG island methylator phenotype, whereas microsatellite stable (MSS) is associated with chromosomal instability (CIN) (3). MSI CRCs are divided into MSI-High (MSI-H) or MSI-Low (MSI-L) subsets, based on the extent of the instability (8). The essence of MSI has been intensively evaluated in the National Comprehensive Cancer Network guideline (8). Noteworthy, 15% of patients with CRC show MSI whereas the remainder are characterized by MSS (9,10). CRCs with MSI-H often exhibit numerous distinct features, including a more proximal tumor position (11). Furthermore, >80% of patients with CRC with Lynch syndrome, a top-ranked inherited CRC-associated disease, exhibit MSI (12,13). Of note, ~10-20% of patients with CRC with Lynch syndrome manifest MSS, with diverse immunohistochemistry results (14).

MSI is one of the most promising markers investigated to date with prognostic and therapeutic values. Previously, patients with MSI were associated with a favorable prognosis compared with those with MSS (3,12,15). However, the prognostic role of MSS and MSI in CRC remains controversial. When patients with MSI-H and MSS/MSI-L received fluorouracil (FU), the significantly different prognostic values between the two became indistinguishable (16). In a MSI subset, those treated with FU had a poorer prognosis than those without FU (16). A recent study highlighted the predictive role of MMR status in immune checkpoint inhibition with pembrolizumab (7).

However, the mechanism underlying the association between MSI status and the clinicopathological characteristics of CRC remains to be fully elucidated. To gain better insight into the key genes and pathways involved in MSI of CRC, bioinformatics analysis of the GSE25071 gene expression profile, including 38 MSS, five MSI-H and three MSI-L samples, was conducted to identify potential key genes and pathways associated with MSI.

Materials and methods

Gene expression profile from the Gene Expression Omnibus (GEO) database. The gene expression profile, GSE25071, which contained 38 MSS colorectal cancer cases, five MSI-H cases and three MSI-L cases (17), was downloaded from the GEO database (http://www.ncbi.nlm.nih.gov/geo/) (18). The GSE25071 profile was further annotated by the profile GPL2986, ABI Human Genome Survey Microarray Version 2 (Applied Biosystems; Thermo Fisher Scientific, Inc., Waltham, MA, USA) (17). Briefly, familial adenomatous polyposis (FAP) syndromes, hereditary nonpolyposis colorectal cancer (HNPCC) and other types of cancer were excluded from the included cases. Clinical information including gender, age, tumor localization, tumor stage according to The International Union Against Cancer (UICC)/American Joint Committee on Cancer and MSI status were recorded (17). The total RNA of the samples was processed using the AllPrep DNA/RNA Mini kit (Qiagen, Inc., Valencia, CA, USA) for tumors and the Ambion RiboPure kit (Applied Biosystems; Thermo Fisher Scientific, Inc.) for normal colonic mucosa. The Chemilunimescent Detection kit from Applied Biosystems; Thermo Fisher Scientific, Inc. was used following the labeling process (NanoAmp RT-IVT Labeling kit; Applied Biosystems; Thermo Fisher Scientific, Inc., Waltham, MA, USA; DIG-UTP, Roche Diagnostics, Basel, Switzerland). Subsequently, the microarrays were scanned using the AB1700 Chemilunimescent microarray analyzer and further processed by the accompanying software (Applied Biosystems; Thermo Fisher Scientific, Inc.; version 1.1.1) (17). GSE18088, GSE13067 and GSE78220 were included for external validation of the differentially expressed genes (DEGs) determined in GSE25071. GSE18088 contained 34 cases of MSS and 19 cases of MSI with primary stage II colon cancer based on the UICC. The profile was based on the Affymetrix Human Genome U133 Plus 2.0 Array (Affymetrix; Thermo Fisher Scientific, Inc.) (19). GSE13067 contained 63 MSS and 11 MSI-H fresh-frozen primary CRC samples for the Affymetrix Human Genome U133 Plus 2.0 Array (Affymetrix; Thermo Fisher Scientific, Inc.) (20). GSE78220 contained 28 melanoma samples for transcriptomic analysis of anti-PD-1 therapy (pembrolizumab), with the Illumina HiSeq 2000 platform (GPL11154; Illumina Inc., San Diego, CA, USA) (21).

Identification of the DEGs. The DEGs listed in three groups of MSS vs. MSI, MSS vs. MSI-L and MSS vs. MSI-H were identified based on the GEO2R web-based tool (www.ncbi.nlm.nih.gov/geo/) (22). The Benjamini-Hochberg procedure (false discovery rate) was applied. The predefined cut-off values included adj. $P<0.05$ and $|\log$ fold change (logFC)$|>2$. Given the limited DEGs in MSS vs. MSI-L and MSS vs. MSI (H+L), only the expression data of DEGs identified in MSS vs. MSI-H group were processed in FunRich software (version 2.1.2; www.funrich.org) for a bidirectional hierarchical clustering plot (23).

Gene ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analyses of DEGs in the MSS vs. MSI-H group. For further functional enrichment analysis of the DEGs in MSS vs. MSI-H, GO enrichment analysis, including biological process (BP), molecular function (MF) and cellular component (CC), and KEGG analysis were performed using the Database for Annotation, Visualization, and Integrated Discovery (http://david.abcc.ncifcrf.gov/) web-based platform (24-26).

Gene co-expression network analysis. Only the DEGs of MSS vs. MSI-H were included for the subsequent analytic processes. All DEGs were imported into GeneMANIA, a web-based gene-gene interactions identification database (www.genemania.org) (27). The interaction list of the DEGs and additional genes (node degree ≥2) was output for visualization in the Cytoscape program (version 3.6.0; www.cytoscape.org/) (28). The hub genes with highest connected edges within the co-expression networks were determined. The highest ranked three modules were identified using the Molecular Complex Detection (MCODE) program (29). The mRNA expression of the hub genes were further externally validated in GSE18088, GSE13067 and GSE78220.

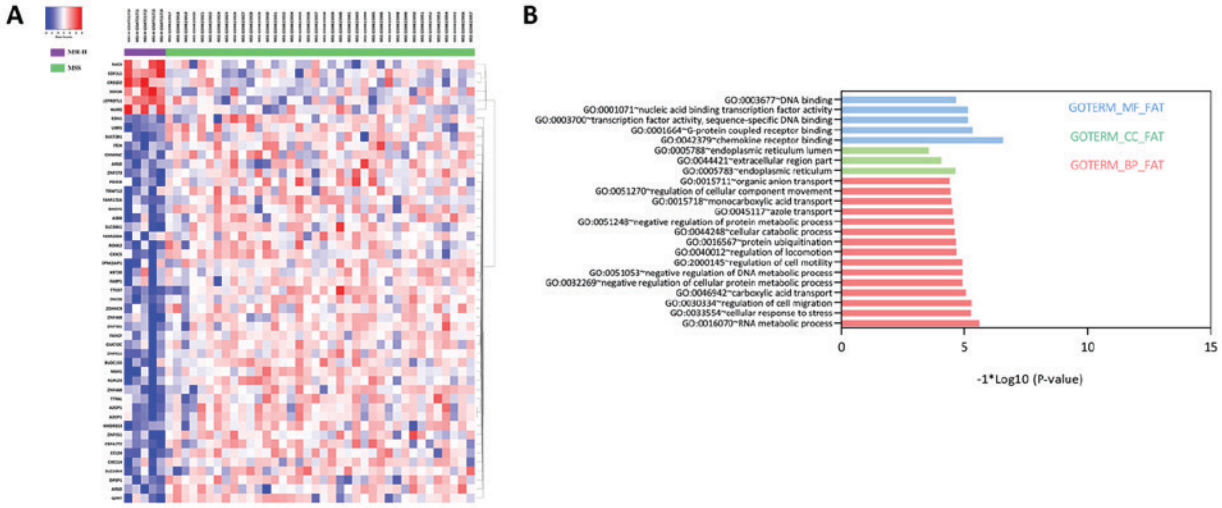

Figure 1. Bidirectional clustering heat map of the DEGs between MSI-H and MSS samples and GO enrichment of DEGs. (A) A total of 49 DEGs were hierarchically clustered and illustrated, with six upregulated (red) and 43 downregulated (blue). (B) 15 BP terms, three CC terms and five MF terms were significantly enriched. DEGs, differentially expressed genes; MSS, microsatellite stable; MSI, microsatellite instability high; GO, Gene Ontology; BP, biological process; CC, cellular component; MF, molecular function. Red indicates upregulation and blue indicates downregulation.

Protein-protein interaction (PPI) networks. All DEGs were further input to the Search Tool for the Retrieval of Interacting Genes/Proteins database (STRING version 10.0; www.string-db.org/) for interaction network at the protein level (30). The results were further visualized using Cytoscape with predefined cut-off values: Node degree, ≥2; k-score(value=2); node score, 0.2; max. depth, 100 (28).

Gene set enrichment analysis (GSEA). GSEA, released by the Broad Institute (software.broadinstitute.org/gsea/index.jsp), was used to cluster significant gene sets associated with given annotation terms (31). In the present study, the MSS and MSI-H samples were analyzed by GSEA with the annotation of 'hallmark gene sets'. The significant cut-off value was defined as P<0.05.

Weighted gene correlation network analysis (WGCNA). WGCNA was used for co-expression network construction based on the correlations among genes and identification of top-ranked gene modules and hub genes. The 'WGCNA' R package was implemented for all the genes in 46 samples (38 MSS, five MSI-H and three MSI-L, normal samples were excluded). Initially, Pearson's correlation of each pair genes was calculated and an adjacency matrix was constructed based on the Pearson's results and a predefined soft-thresholding parameter (β). Subsequently, the topological overlap matrix of the included genes and adjacency matrix was produced. Genes with similar expression trends were classified as the same modules eigengenes for further clinical traits correlation (32,33). The genes of the most correlated module were extracted for GO and KEGG analyses and PPI network construction. Hub genes were defined with the highest degrees.

Prognostic values of hub genes in WGCNA. The overall survival of hub genes determined by WGCNA was further examined in the PrognoScan database, a comprehensive platform for prognostic annotation (dna00.bio.kyutech.ac.jp/PrognoScan/) (34).

Results

Identification of upregulated and downregulated DEGs. A total of 49 DEGs, including six upregulated and 43 downregulated DEGs, were identified between MSI-H and MSS groups, illustrated by the bidirectional clustering heat map (Fig. 1A). However, no significant DEGs were identified between MSI-L and MSS, and only one DEG was identified between MSI-H and MSI-L. Therefore, subsequent investigations focused on the DEGs between MSI-H and MSS.

GO enrichment and KEGG pathway analyses of DEGs. A total of 15 BP, three CC and five MF terms were significantly enriched. Specifically, RNA metabolic process, endoplasmic reticulum and chemokine receptor binding were the top-ranked in each term, respectively (Fig. 1B). No significant KEGG pathway was enriched in the DEGs.

Co-expressed genes network analysis. To delineate the biological functions of the DEGs, a co-expression network of the DEGs with correlated additional genes was established based on the GeneMANIA program. A total of 66 nodes and 678 edges were determined (Fig. 2A). The top five hub genes with the highest degree were identified, including zinc finger protein (ZNF) 813, ZNF426, ZNF611, ZNF320 and ZNF573.

The top three modules with the highest scores were identified using the MCODE plugin (Fig. 2B-D). Among them, no particular KEGG pathway was significantly enriched. However, in the GO enrichment, the regulation of transcription, DNA-templated term was the highest ranked BP for module 1, and negative regulation of hydrolase activity for module 2. Subsequently, the hub genes were externally validated in GSE18088, GSE13067 and GSE78220. The expression levels of ZNF426, ZNF320 and ZNF573 were significantly reduced in MSI-H compared with MSS in GSE18088. The expression levels of ZNF813, ZNF426 and ZNF573 were significantly reduced in the MSI-H group compared with the MSS group

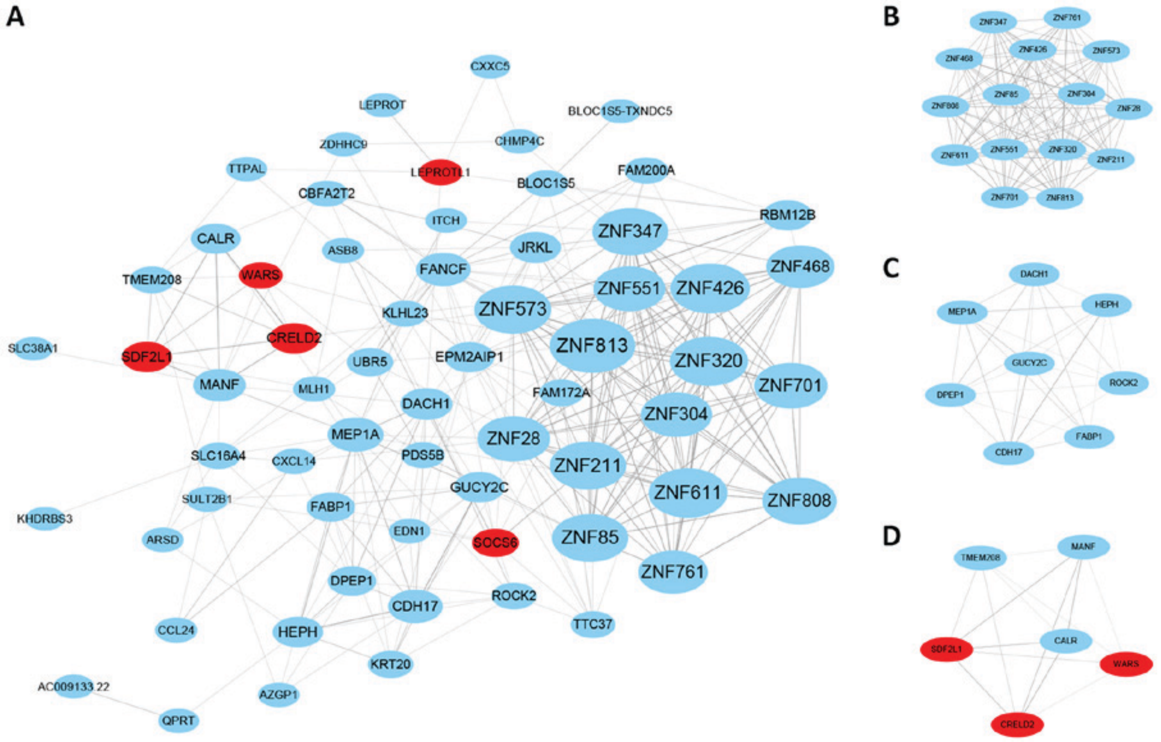

Figure 2. Co-expression network constructed by the differentially expressed genes and additional correlated genes from the GeneMANIA program. (A) Co-expression network; (B) top modules with highest scores; (C) second highest module; (D) third highest module. Red indicates upregulation; blue indicates downregulation; lines between nodes indicates interactions between genes.

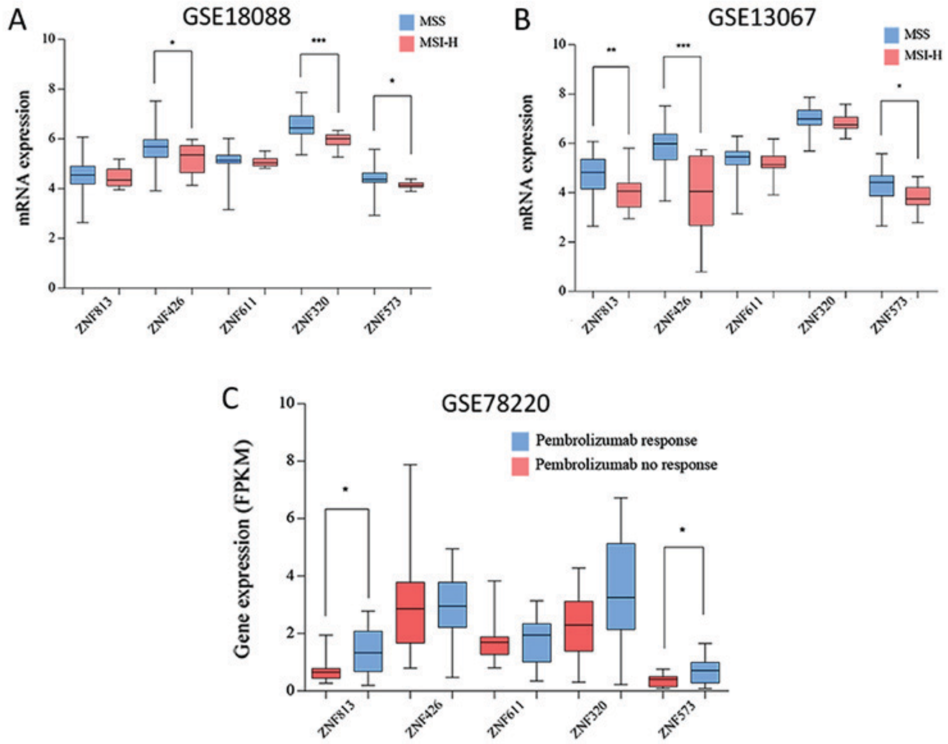

Figure 3. Hub gene expression in two independent genomic profiles. (A) Gene expression in MSS and MSI-H in GSE18088. (B) Gene expression in MSS and MSI-H in GSE13067. (C) Gene expression in GSE78220. MSS, microsatellite stable; MSI-H, microsatellite instability high. *P<0.05, **P<0.01, ***P<0.001. Data is presented as the mean ± standard deviation.

in GSE13067. The expression levels of ZNF813 and ZNF573 were significantly reduced in the pembrolizumab no-response group compared with the response group (Fig. 3A-C).

PPI network analysis. The minimum required interaction score of STRING was medium confidence (0.4) and the cut-off degree for the included nodes in Cytoscape was ≥1. The PPI

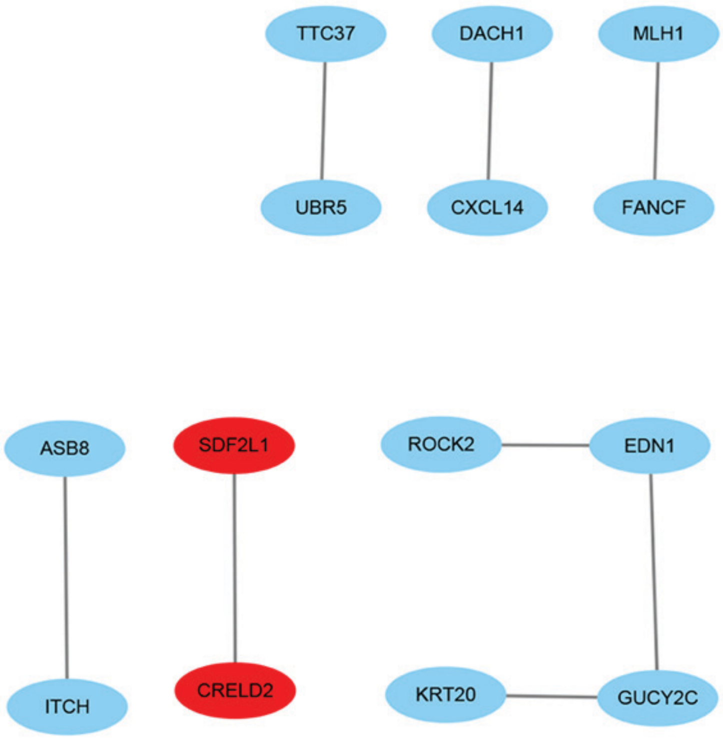

Figure 4. Protein-protein interaction networks of the differentially expressed genes. Red indicates upregulation, blue indicates downregulation.

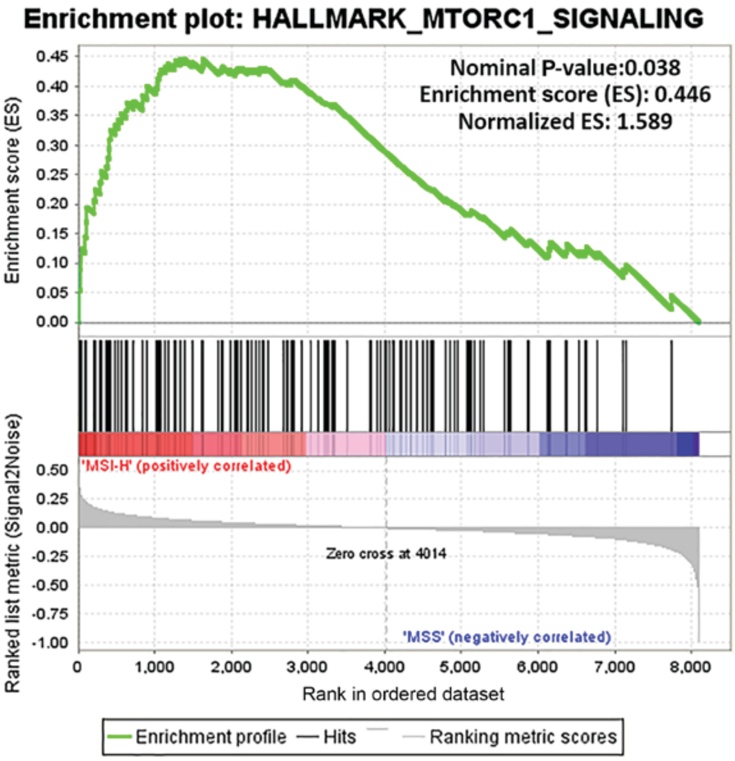

Figure 5. MTORC1 signaling is significantly enriched by gene set enrichment analysis between MSI-H and MSS groups. MTORC1, mammalian target of rapamycin complex 1; MSS, microsatellite stable; MSI-H, microsatellite instability high.

networks included 14 nodes and eight edges (Fig. 4), being distinct from the co-expression networks (Fig. 2A).

GSEA results. GSEA was used to determine the functions of gene sets between the MSI-H and MSS groups. Only one gene set, the mammalian target of rapamycin complex 1 (mTORC1) signaling (nominal P-value, 0.038; normalized enrichment score, 0.446) was significantly enriched in MSI-H, with none significantly enriched in MSS (Fig. 5). Furthermore, the 50 top ranked genes correlated with each

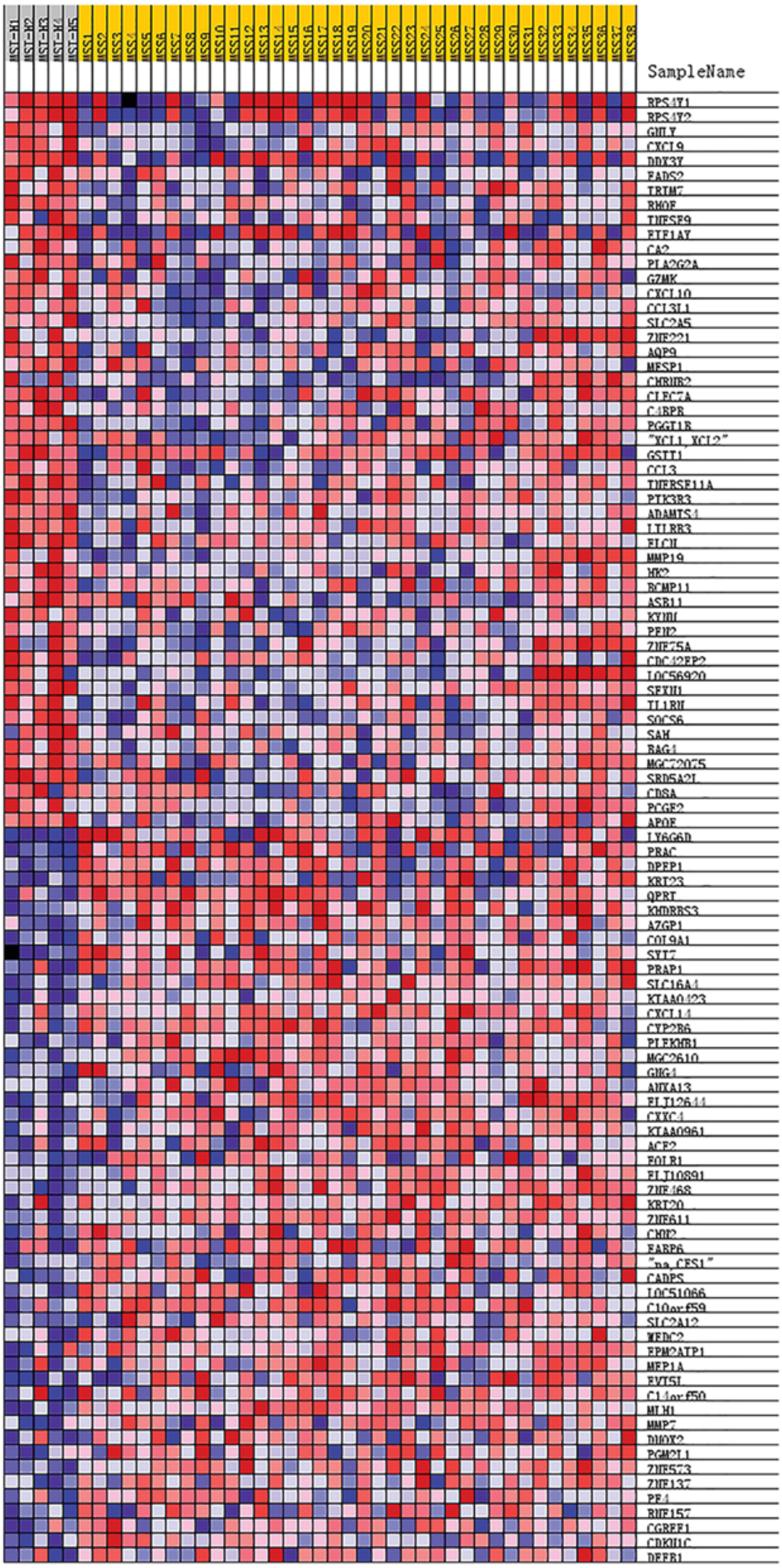

Figure 6. Heat map of the 50 top-ranked genes correlated with MSI-H and MSS, respectively. Red indicates high expression, blue indicates low expression, pink indicates moderate expression. MSS, microsatellite stable; MSI-H, microsatellite instability high.

phenotype (MSI-H and MSS) are illustrated with a heat map (Fig. 6).

WGCNA of the gene expression profile in all tumor samples. To further investigate the potential gene modules associated with MSI status, the WGCNA was conducted with R package. A total of 46 cases, including 38 MSS, eight MSI (five MSI-H and three MSI-L) were clustered (Fig. 7A). The power of β=9 was defined as the soft-thresholding value (scale free R^2=0.95; slope=-1.67; Fig. 7B-D). A total of 25 modules were identified (Fig. 8A). Noteworthy, the pink module was the most significantly correlated with MSI status (R^2=0.5, P=0.0004). Of note,

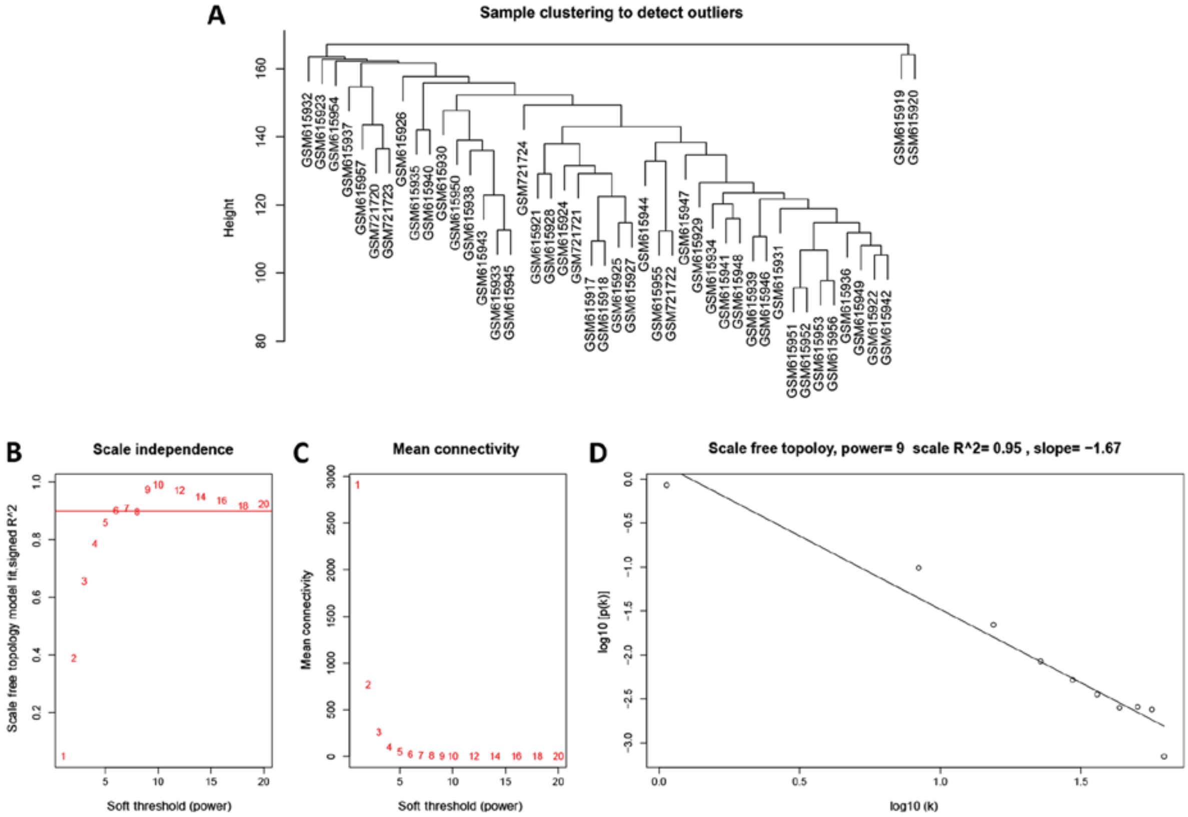

Figure 7. Weighted gene correlation network analysis of MSS vs. MSI. (A) Clustering dendrogram of MSS and MSI samples. (B) Scale-free fit index of included soft-threshold powers (β). (C) Mean connectivity of various powers (β). (D) Checking for scale-free topology with β=9. MSS, microsatellite stable; MSI-H, microsatellite instability high.

the pink module also featured a high correlation (R^2=0.53, P=0.0002) with tumor stage (Fig. 8B).

Identification of hub genes in the pink module associated with MSI status. The genes in the pink module were significantly enriched by protein targeting to endoplasmic reticulum in BP terms, cytosolic part in CC terms, structural constituent of ribosome in MF terms and ribosome in KEGG pathways. Subsequently, a PPI network (degree ≥1), with 55 nodes and 96 edges, was constructed based on the genes in the pink module. The top 10 hub genes with the highest degrees were determined, including ribosomal protein L12 (RPL12), ribosomal protein S3A (RPS3A), ribosomal protein S9 (RPS9), ribosomal protein L27a (RPL27A), ribosomal protein L7 (RPL7), ribosomal protein L28 (RPL28), ribosomal protein L14 (RPL14), ribosomal protein S17 (RPS17), mitochondrial ribosomal protein L16 (MRPL16) and G elongation factor, mitochondrial 2 (GFM2), as shown in Fig. 9.

Prognostic values of the hub genes in the pink module. The prognostic values of the hub genes were assessed in various independent datasets with different probe IDs and array types. In all, none of the genes exhibited significant prognostic values (Table I).

Discussion

The present study is the first, to the best of our knowledge, to use multiple bioinformatics analysis approaches to demonstrate the potential key genes and pathways associated with MSI status in patients with CRC. Among the significant GO terms, RNA metabolic process, endoplasmic reticulum and chemokine receptor binding were the top-ranked terms. No significant KEGG pathway was identified, however, using GSEA, the MTORC1 signaling pathway was significantly enriched. MTOR signaling pathways, consisting of at least two complexes, mTORC1 and mTORC2, receive a plethora of input factors and modulate a broad spectrum of downstream molecules (35,36). Noteworthy, the inhibition of mTOR1 only leads to mild protein synthesis reduction and potential influence upon the cell cycle process (35,37).

Previously, Choi *et al* investigated somatic mutational, intratumoral heterogeneity and expressional alterations of mTOR pathway-related genes in cancer with MSI, including PIK3CB, insulin receptor substrate 1/2 (IRS1), RPS6, eukaryotic translation initiation factor 4B (EIF4B), RPS6KA5 and PRKAA2 (38). Of the patients with MSI-H CRC, 8.9% harbored IRS1 frameshift mutations, whereas 10.1% harbored mutations in EIF4B and 3.8% in RPS6KA5. Noteworthy, no mutations was identified in MSS or MSI-L (38). The study by Choi *et al* and the present study demonstrated the potential roles of the ribosomal protein family associated with MSI status. In addition, differing from previous search strategies in published results (38), the present study illustrated how WGCNA can be implemented to predict new genes in the regulation of MSI in CRC.

Lin *et al* analyzed the mutations of 113 MSS and 29 MSI-H cases of CRC (39). Mutations of PIK3CA, phosphatase and

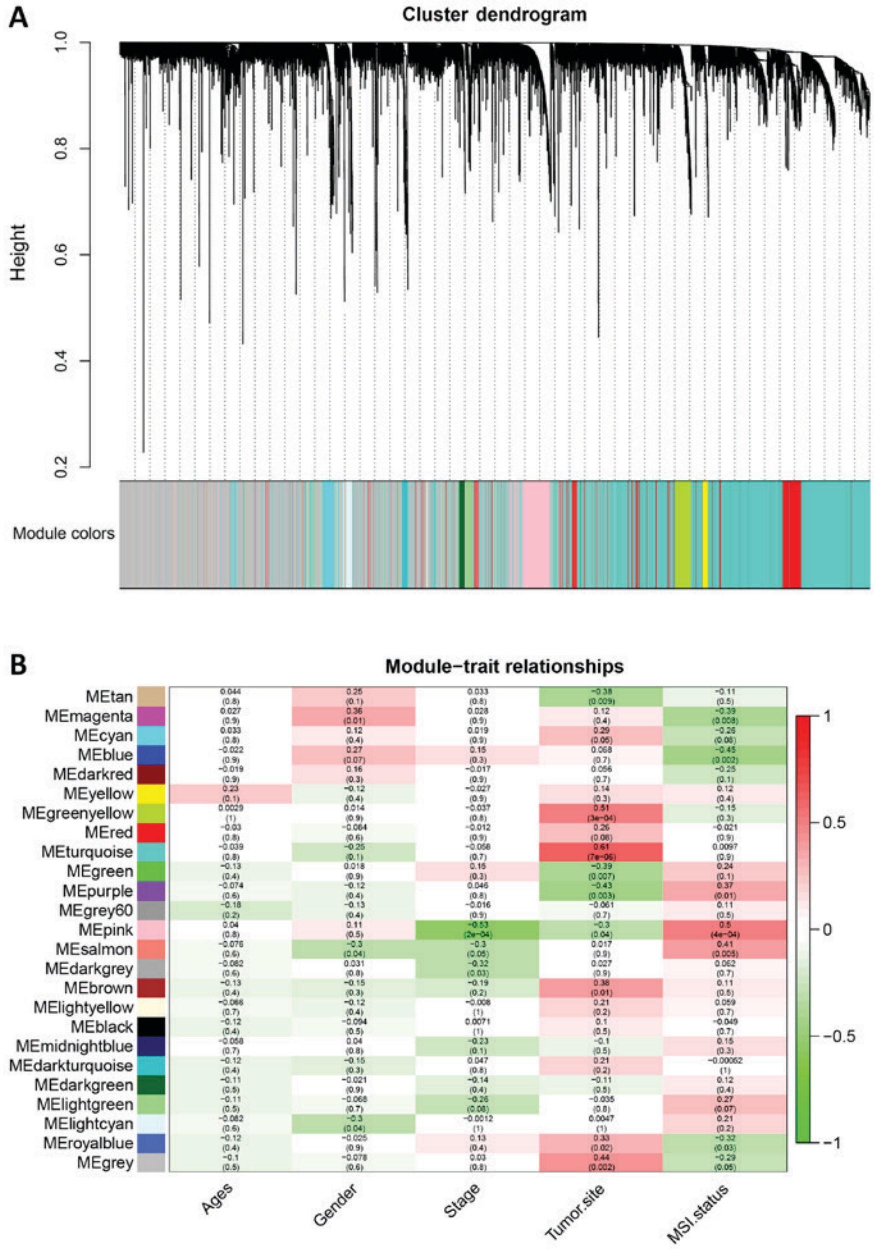

Figure 8. Identification of 25 modules in weighted gene correlation network analysis with a heat map of the correlation between clinical traits and colored modules. (A) Clustering dendrogram of all genes between MSI and MSS. (B) Heat map of the correlation between modules eigengenes and clinical traits. Numbers in the cells indicate correlation (P-value in brackets). MSS, microsatellite stable; MSI, microsatellite instability. Red represents positive correlation and green represents negative correlation.

tensin homolog and/or AKT1 in the mTOR pathway were found in 59% of the MSI-H patients compared with 19% of the patients with MSS (39). Methodologically, a 50-gene AmpliSeq Cancer Hotspot Panel was introduced by Lin *et al*, whereas the expression profile generated using the ABI Human Genome Survey Microarray was analyzed in the present study for GSEA and WGCNA. Collectively, the present study provided an insightful target and further complemented the results reported by Choi *et al* and Lin *et al* with regards to the multiple bioinformatics strategies.

In the co-expressed gene network, ZNF813, ZNF426, ZNF611, ZNF320 and ZNF573 were the top-ranked hub genes closely associated with MSI-H. Reduced expression levels of ZNF813 and ZNF573 were found in the MSI-H group and pembrolizumab no-response group. Of note, the check-point inhibitor pembrolizumab significantly prolonged the prognostic outcomes of patients with MSI-H CRC compared with those in the MSS group (7). This indicated that potential mechanisms exist between MSI status and the outcomes of check-point inhibitor treatment, further highlighting the predictive role of ZNFs.

ZNFs are one of the most common proteins in the eukaryotic system, with a broad range of biological functions, including DNA recognition, RNA transcription, apoptosis and protein structure (40). The five hub genes are mainly located in the nucleus and are involved in DNA-binding and transcription regulation. Novel topologies of numerous ZNF domains have provided evidence for structure/function relationships (40). To the best of our knowledge, the present study is the first highlighting the potential association between ZNFs and MSI in CRC.

Table I. Prognostic values of hub genes from PrognoScan.

ID_NAME	Dataset	Array type	Probe ID	N	COX P-value	HR (95% CI lower-CI upper)
GFM2	GSE17536	HG-U133_Plus_2	231917_at	177	0.52	1.20 (0.69-2.10)
GFM2	GSE17536	HG-U133_Plus_2	231918_s_at	177	0.43	1.19 (0.78-1.81)
GFM2	GSE17536	HG-U133_Plus_2	225392_at	177	0.35	1.33 (0.73-2.42)
GFM2	GSE17537	HG-U133_Plus_2	231917_at	55	0.64	0.75 (0.22-2.54)
GFM2	GSE17537	HG-U133_Plus_2	231918_s_at	55	0.54	1.24 (0.63-2.44)
GFM2	GSE17537	HG-U133_Plus_2	225392_at	55	0.47	1.37 (0.59-3.21)
MRPL16	GSE12945	HG-U133A	217980_s_at	62	0.82	1.12 (0.41-3.08)
MRPL16	GSE17536	HG-U133_Plus_2	217980_s_at	177	0.86	1.06 (0.59-1.89)
MRPL16	GSE17537	HG-U133_Plus_2	217980_s_at	55	0.623951	1.23 (0.53-2.86)
RPL12	GSE12945	HG-U133A	214271_x_at	62	0.76	2.21 (0.01-326.11)
RPL12	GSE12945	HG-U133A	200809_x_at	62	0.39	107.59 (0.00-4600568.31)
RPL12	GSE17536	HG-U133_Plus_2	214271_x_at	177	0.29	2.03 (0.55-7.52)
RPL12	GSE17536	HG-U133_Plus_2	200809_x_at	177	0.34	1.91 (0.50-7.28)
RPL12	GSE17537	HG-U133_Plus_2	214271_x_at	55	0.47	1.76 (0.38-8.13)
RPL12	GSE17537	HG-U133_Plus_2	200809_x_at	55	0.23	2.31 (0.59-8.99)
RPL14	GSE12945	HG-U133A	213588_x_at	62	0.45	7654467.33 (0.00-6592665664338867322355712.00)
RPL14	GSE12945	HG-U133A	219138_at	62	0.76	0.78 (0.17-3.69)
RPL14	GSE12945	HG-U133A	200074_s_at	62	0.90	1.07 (0.40-2.83)
RPL14	GSE17536	HG-U133_Plus_2	213588_x_at	177	0.78	0.84 (0.24-2.93)
RPL14	GSE17536	HG-U133_Plus_2	200074_s_at	177	0.57	0.75 (0.28-2.00)
RPL14	GSE17536	HG-U133_Plus_2	219138_at	177	0.52	0.72 (0.26-1.97)
RPL14	GSE17537	HG-U133_Plus_2	200074_s_at	55	0.95	0.96 (0.28-3.26)
RPL14	GSE17537	HG-U133_Plus_2	219138_at	55	0.89	0.95 (0.46-1.97)
RPL14	GSE17537	HG-U133_Plus_2	213588_x_at	55	0.86	0.88 (0.23-3.43)
RPL27A	GSE12945	HG-U133A	212044_s_at	62	0.88	1.06 (0.48-2.33)
RPL27A	GSE12945	HG-U133A	203034_s_at	62	0.72	0.80 (0.23-2.71)
RPL27A	GSE17536	HG-U133_Plus_2	223707_at	177	0.43	0.77 (0.40-1.49)
RPL27A	GSE17536	HG-U133_Plus_2	212044_s_at	177	0.69	0.87 (0.42-1.78)
RPL27A	GSE17536	HG-U133_Plus_2	203034_s_at	177	0.48	2.43 (0.21-27.88)
RPL27A	GSE17537	HG-U133_Plus_2	212044_s_at	55	0.23	0.75 (0.47-1.20)
RPL27A	GSE17537	HG-U133_Plus_2	203034_s_at	55	0.83	1.46 (0.05-41.30)
RPL27A	GSE17537	HG-U133_Plus_2	223707_at	55	0.28	0.50 (0.14-1.77)
RPL28	GSE12945	HG-U133A	213223_at	62	0.24	2.03 (0.62-6.63)
RPL28	GSE12945	HG-U133A	200003_s_at	62	0.58	1.34 (0.48-3.73)
RPL28	GSE17536	HG-U133_Plus_2	213223_at	177	0.95	0.98 (0.53-1.81)
RPL28	GSE17536	HG-U133_Plus_2	200003_s_at	177	0.81	0.83 (0.18-3.74)
RPL28	GSE17537	HG-U133_Plus_2	200003_s_at	55	0.19	2.72 (0.61-12.16)
RPL28	GSE17537	HG-U133_Plus_2	213223_at	55	0.99	1.01 (0.26-3.88)
RPL7	GSE12945	HG-U133A	212042_x_at	62	0.59	1.98 (0.17-23.44)
RPL7	GSE12945	HG-U133A	200717_x_at	62	0.42	3650.83 (0.00-1740006561282.69)
RPL7	GSE17536	HG-U133_Plus_2	239493_at	177	0.071	1.69 (0.96-3.00)
RPL7	GSE17536	HG-U133_Plus_2	212042_x_at	177	0.16	3.10 (0.63-15.23)
RPL7	GSE17536	HG-U133_Plus_2	200717_x_at	177	0.14	4.24 (0.63-28.50)
RPL7	GSE17537	HG-U133_Plus_2	239493_at	55	0.0050	0.09 (0.02-0.48)
RPL7	GSE17537	HG-U133_Plus_2	212042_x_at	55	0.45	0.38 (0.03-4.76)
RPL7	GSE17537	HG-U133_Plus_2	200717_x_at	55	0.53	0.47 (0.04-4.97)
RPS17	GSE12945	HG-U133A	201665_x_at	62	0.86	0.82 (0.08-8.15)
RPS17	GSE12945	HG-U133A	212578_x_at	62	0.66	1.91 (0.11-33.65)
RPS17	GSE17536	HG-U133_Plus_2	212578_x_at	177	0.082	4.17 (0.84-20.77)

Table I. Continued.

ID_NAME	Dataset	Array type	Probe ID	N	COX P-value	HR (95% CI lower-CI upper)
RPS17	GSE17536	HG-U133_Plus_2	201665_x_at	177	0.17	2.96 (0.64-13.76)
RPS17	GSE17537	HG-U133_Plus_2	212578_x_at	55	0.53	2.16 (0.20-23.51)
RPS17	GSE17537	HG-U133_Plus_2	201665_x_at	55	0.37	2.99 (0.27-32.95)
RPS3A	GSE12945	HG-U133A	201257_x_at	62	0.80	25.91 (0.00-2227310678806.87)
RPS3A	GSE17536	HG-U133_Plus_2	201257_x_at	177	0.97	1.04 (0.11-9.64)
RPS3A	GSE17537	HG-U133_Plus_2	201257_x_at	55	0.65	0.48 (0.02-11.00)
RPS9	GSE12945	HG-U133A	217747_s_at	62	0.85	0.90 (0.31-2.63)
RPS9	GSE12945	HG-U133A	214317_x_at	62	0.77	5.03 (0.00-240092.23)
RPS9	GSE17536	HG-U133_Plus_2	1557981_at	177	0.0018	0.06 (0.01-0.35)
RPS9	GSE17536	HG-U133_Plus_2	214317_x_at	177	0.81	1.22 (0.24-6.07)
RPS9	GSE17536	HG-U133_Plus_2	217747_s_at	177	0.66	1.30 (0.40-4.17)
RPS9	GSE17537	HG-U133_Plus_2	217747_s_at	55	0.99	1.01 (0.36-2.84)
RPS9	GSE17537	HG-U133_Plus_2	1557981_at	55	0.25	0.08 (0.00-5.87)
RPS9	GSE17537	HG-U133_Plus_2	214317_x_at	55	0.70	1.50 (0.19-11.58)

HR, hazard ratio; CI, confidence intervals; GFM, G elongation factor, mitochondrial 2; MRPL16, mitochondrial ribosomal protein L16; RP, ribosomal protein.

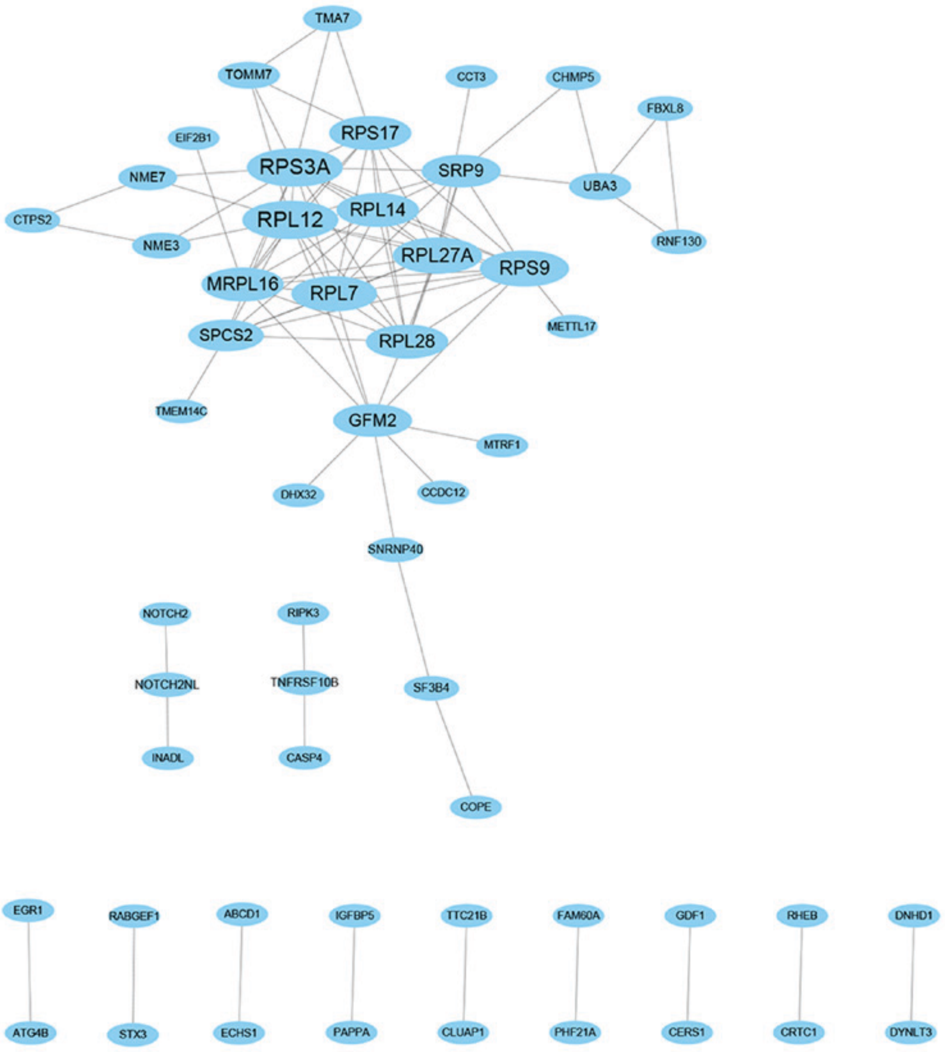

Figure 9. Protein-protein interaction networks of the extracted genes from the pink module.

The genes in the pink module of the WGCNA demonstrated the highest correlation with MSI. By performing further PPI network analysis, the top 10 hub genes were identified, including RPL12, RPS3A, RPS9, RPL27A, RPL7, RPL28, RPL14, RPS17, MRPL16 and GFM2. Given the high proportion of RPs in the hub gene list, RPs are of interest for further discussion and may be of significance to the mechanism underlying the MSI in CRC. The synthesis of RPs is the basis for the biological processes in each cell (41). The newly decoded crystal structures of ribosomes provide multiple traits associated with PPIs, RNA-protein and protein-drugs interactions (41). Noteworthy, the mutation of RPS20, part of the small ribosome subunit, can render individuals with MSS predisposition (42), highlighting the association between ribosome and MSI/MSS status. In addition, RPs are associated with biological RNA synthesis, one of the predominant features of cancer cells exposed to 5-FU treatment (43,44). Therefore, this clarified the role of MSI in FU-non-responders, at least in part. Of note, the prognostic evaluation of the hub genes indicated that their potential roles may not be directly associated with survival status.

Previously, Timmermann *et al* fully investigated the whole exome next generation sequencing of 454 patients with CRC and identify the significant 359 mutations in MSI and 45 mutations in MSS (45). In addition to the MSI and CIN subtypes, a third subtype associated with sessile-serrated adenomas was proposed (46). This newly added third subtype may partially clarify the limited DEGs identified in MSI-L vs. MSS and MSI vs. MSS in the present study.

The limitations of the present study include the comparably small sample size in MSI-H and lack of experimental validation for hub genes. Larger CRC samples with MSI/MSS and molecular biological experiments are required to specifically confirm the functions and mechanisms of the hub genes underlying MSI in CRC. However, to reduce the potential confounding factors produced by a single bioinformatics approach, the present study employed multiple bioinformatics patterns, including DEG analysis, GSEA and WGCNA. In addition, due to the limited number of patients and lack of survival data in original files, the prognostic values of hub genes identified by WGCNA were examined using the PrognoScan database.

In conclusion, the bioinformatics analysis performed in the present study identified key genes and pathways associated with MSI, and further elucidated insightful traits for potential mechanisms.

Acknowledgements

The authors would like to thank the Shanghai Institute of Digestive Surgery, Ruijin Hospital, Shanghai Jiao Tong University School of Medicine (Shanghai, China) for their academic support. The authors would also like to thank Dr Ernest Johann Helwig (Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China) for his helpful contributions with discussions and language editing.

Funding

This study was financially supported by the National Natural Science Foundation of China (grant nos. 81402423 and 81572818) and the Shanghai Municipal Commission of Health and Family Planning (grant no. 2017YQ062).

Availability of data and materials

The datasets supporting the conclusion of this article are included within the article.

Authors' contributions

CY, HH, SZ and JS performed experiments and data analysis; CY, HH, YZ, JM and MZ drafted the manuscript; CY, AL, JS, MZ, YZ and JM participated in the study design, data collection and revision process. All authors read and approved the final manuscript.

Ethics approval and consent to participate

Not applicable.

Patient consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

References

1. Siegel RL, Miller KD, Fedewa SA, Ahnen DJ, Meester RGS, Barzi A and Jemal A: Colorectal cancer statistics, 2017. CA Cancer J Clin 67: 177-193, 2017.
2. Chen W, Zheng R, Baade PD, Zhang S, Zeng H, Bray F, Jemal A, Yu XQ and He J: Cancer statistics in China, 2015. CA Cancer J Clin 66: 115-132, 2016.
3. Popat S, Hubner R and Houlston RS: Systematic review of microsatellite instability and colorectal cancer prognosis. J Clin Oncol 23: 609-618, 2005.
4. Markowitz SD and Bertagnolli MM: Molecular origins of cancer: Molecular basis of colorectal cancer. N Engl J Med 361: 2449-2460, 2009.
5. Budinska E, Popovici V, Tejpar S, D'Ario G, Lapique N, Sikora KO, Di Narzo AF, Yan P, Hodgson JG, Weinrich S, *et al*: Gene expression patterns unveil a new level of molecular heterogeneity in colorectal cancer. J Pathol 231: 63-76, 2013.
6. De Roock W, De Vriendt V, Normanno N, Ciardiello F and Tejpar S: KRAS, BRAF, PIK3CA, and PTEN mutations: Implications for targeted therapies in metastatic colorectal cancer. Lancet Oncol 12: 594-603, 2011.
7. Diaz LA Jr and Le DT: PD-1 blockade in tumors with mismatch-repair deficiency. N Engl J Med 373: 1979, 2015.
8. Benson AB Jr, Venook AP, Cederquist L, Chan E, Chen YJ, Cooper HS, Deming D, Engstrom PF, Enzinger PC, Fichera A, *et al*: Colon cancer, version 1.2017, NCCN clinical practice guidelines in oncology. J Natl Compr Canc Netw 15: 370-398, 2017.
9. Peltomaki P: Role of DNA mismatch repair defects in the pathogenesis of human cancer. J Clin Oncol 21: 1174-1179, 2003.
10. Cottrell S, Bodmer WF, Bicknell D and Kaklamanis L: Molecular analysis of APC mutations in familial adenomatous polyposis and sporadic colon carcinomas. Lancet 340: 626-630, 1992.
11. Cancer Genome Atlas Network: Comprehensive molecular characterization of human colon and rectal cancer. Nature 487: 330-337, 2012.
12. de la Chapelle A and Hampel H: Clinical relevance of microsatellite instability in colorectal cancer. J Clin Oncol 28: 3380-3087, 2010.

13. Aaltonen LA, Salovaara R, Kristo P, Canzian F, Hemminki A, Peltomäki P, Chadwick RB, Kääriäinen H, Eskelinen M, Järvinen H, *et al*: Incidence of hereditary nonpolyposis colorectal cancer and the feasibility of molecular screening for the disease. N Engl J Med 338: 1481-1487, 1998.
14. Pinol V, Castells A, Andreu M, Castellví-Bel S, Alenda C, Llor X, Xicola RM, Rodríguez-Moranta F, Payá A, Jover R, *et al*: Accuracy of revised Bethesda guidelines, microsatellite instability, and immunohistochemistry for the identification of patients with hereditary nonpolyposis colorectal cancer. JAMA 293: 1986-1994, 2005.
15. Walther A, Houlston R and Tomlinson I: Association between chromosomal instability and prognosis in colorectal cancer: A meta-analysis. Gut 57: 941-950, 2008.
16. Ribic CM, Sargent DJ, Moore MJ, Thibodeau SN, French AJ, Goldberg RM, Hamilton SR, Laurent-Puig P, Gryfe R, Shepherd LE, *et al*: Tumor microsatellite-instability status as a predictor of benefit from fluorouracil-based adjuvant chemotherapy for colon cancer. N Engl J Med 349: 247-257, 2003.
17. Ågesen TH, Berg M, Clancy T, Thiis-Evensen E, Cekaite L, Lind GE, Nesland JM, Bakka A, Mala T, Hauss HJ, *et al*: CLC and IFNAR1 are differentially expressed and a global immunity score is distinct between early-and late-onset colorectal cancer. Genes Immun 12: 653-662, 2011.
18. Edgar R, Domrachev M and Lash AE: Gene expression omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res 30: 207-210, 2002.
19. Gröne J, Lenze D, Jurinovic V, Hummel M, Seidel H, Leder G, Beckmann G, Sommer A, Grützmann R, Pilarsky C, *et al*: Molecular profiles and clinical outcome of stage UICC II colon cancer patients. Int J Colorectal Dis 26: 847-858, 2011.
20. Jorissen RN, Lipton L, Gibbs P, Chapman M, Desai J, Jones IT, Yeatman TJ, East P, Tomlinson IP, Verspaget HW, *et al*: DNA copy-number alterations underlie gene expression differences between microsatellite stable and unstable colorectal cancers. Clin Cancer Res 14: 8061-8069, 2008.
21. Hugo W, Zaretsky JM, Sun L, Song C, Moreno BH, Hu-Lieskovan S, Berent-Maoz B, Pang J, Chmielowski B, Cherry G, *et al*: Genomic and transcriptomic features of response to anti-PD-1 therapy in metastatic melanoma. Cell 165: 35-44, 2016.
22. Davis S and Meltzer PS: GEOquery: A bridge between the gene expression omnibus (GEO) and BioConductor. Bioinformatics 23: 1846-1847, 2007.
23. Pathan M, Keerthikumar S, Ang CS, Gangoda L, Quek CY, Williamson NA, Mouradov D, Sieber OM, Simpson RJ, Salim A, *et al*: FunRich: An open access standalone functional enrichment and interaction network analysis tool. Proteomics 15: 2597-2601, 2015.
24. Huang da W, Sherman BT and Lempicki RA: Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc 4: 44-57, 2009.
25. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, *et al*: Gene ontology: Tool for the unification of biology. Nat Genet 25: 25-29, 2000.
26. Kanehisa M and Goto S: KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res 28: 27-30, 2000.
27. Warde-Farley D, Donaldson SL, Comes O, Zuberi K, Badrawi R, Chao P, Franz M, Grouios C, Kazi F, Lopes CT, *et al*: The GeneMANIA prediction server: Biological network integration for gene prioritization and predicting gene function. Nucleic Acids Res 38 (Suppl_2): W214-W220, 2010.
28. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B and Ideker T: Cytoscape: A software environment for integrated models of biomolecular interaction networks. Genome Res 13: 2498-2504, 2003.
29. Bader GD and Hogue CW: An automated method for finding molecular complexes in large protein interaction networks. BMC Bioinform 4: 2, 2003.
30. 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 43: D447-D452, 2014.
31. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, *et al*: Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci USA 102: 15545-15550, 2005.
32. Langfelder P and Horvath S: WGCNA: An R package for weighted correlation network analysis. BMC Bioinformatics 9: 559, 2008.
33. Zhang B and Horvath S: A general framework for weighted gene co-expression network analysis. Stat Appl Genet Mol Boil: Aug 12, 2005 (Epub ahead of print).
34. Mizuno H, Kitada K, Nakai K and Sarai A: PrognoScan: A new database for meta-analysis of the prognostic value of genes. BMC Med Genomics 2: 18, 2009.
35. Abraham RT and Gibbons JJ: The mammalian target of rapamycin signaling pathway: Twists and turns in the road to cancer therapy. Clin Cancer Res 13: 3109-3114, 2007.
36. Han JM, Jeong SJ, Park MC, Kim G, Kwon NH, Kim HK, Ha SH, Ryu SH and Kim S: Leucyl-tRNA synthetase is an intracellular leucine sensor for the mTORC1-signaling pathway. Cell 149: 410-424, 2012.
37. Fingar DC and Blenis J: Target of rapamycin (TOR): An integrator of nutrient and growth factor signals and coordinator of cell growth and cell cycle progression. Oncogene 23: 3151-3171, 2004.
38. Choi MR, Yoo NJ, An CH and Lee SH: Frameshift mutations in mammalian target of rapamycin pathway genes and their regional heterogeneity in sporadic colorectal cancers. Hum Pathol 46: 753-760, 2015.
39. Lin EI, Tseng LH, Gocke CD, Reil S, Le DT, Azad NS and Eshleman JR: Mutational profiling of colorectal cancers with microsatellite instability. Oncotarget 6: 42334-42344, 2015.
40. Laity JH, Lee BM and Wright PE: Zinc finger proteins: New insights into structural and functional diversity. Curr Opin Struct Boil 11: 39-46, 2001.
41. Kothe U: Recent progress on understanding ribosomal protein synthesis. In: Comprehensive Natural Products II Chemistry and Biology, Section Amino Acids, Peptides and Proteins. Oxford, Elsevier, 2010.
42. Nieminen TT, O'Donohue MF, Wu Y, Lohi H, Scherer SW, Paterson AD, Ellonen P, Abdel-Rahman WM, Valo S, Mecklin JP, *et al*: Germline mutation of RPS20, encoding a ribosomal protein, causes predisposition to hereditary nonpolyposis colorectal carcinoma without DNA mismatch repair deficiency. Gastroenterology 147: 595-598.e5, 2014.
43. Longley DB, Harkin DP and Johnston PG: 5-fluorouracil: Mechanisms of action and clinical strategies. Nat Rev Cancer 3: 330-338, 2003.
44. Kurland CG and Maaløe O: Regulation of ribosomal and transfer RNA synthesis. J Mol Boil 4: 193-210, 1962.
45. Timmermann B, Kerick M, Roehr C, Fischer A, Isau M, Boerno ST, Wunderlich A, Barmeyer C, Seemann P, Koenig J, *et al*: Somatic mutation profiles of MSI and MSS colorectal cancer identified by whole exome next generation sequencing and bioinformatics analysis. PLoS One 5: e15661, 2010.
46. De Sousa E Melo F, Wang X, Jansen M, Fessler E, Trinh A, de Rooij LP, de Jong JH, de Boer OJ, van Leersum R, Bijlsma MF, *et al*: Poor-prognosis colon cancer is defined by a molecularly distinct subtype and develops from serrated precursor lesions. Nat Med 19: 614-618, 2013.