doi:10.4149/neo_2018_170708N470

# Integrated analysis of differentially expressed genes in esophageal squamous cell carcinoma using bioinformatics

Z. DONG, H. ZHANG, T. ZHAN, S. XU*

*Department of Gastroenterology, Tongji Hospital, Tongji University School of Medicine, Shanghai, China*

**Correspondence: xschang@163.com*

**Received July 8, 2017 / Accepted November 3, 2017**

Esophageal squamous cell carcinoma (ESCC) is a deadly disease. To identify key genes in esophageal squamous cell carcinoma, we followed a strategy utilizing the laiurger microarray dataset (GSE38129) as the training set and another independent microarray dataset (GSE20347) as the validation set. Following quality control, differentially expressed genes (DEGs) were obtained using R software. Functional enrichment analysis was performed using DAVID database and the DEG co-expression network was established with Weighted Gene Co-Expression Network Analysis (WGCNA) and visualized by Cytoscape. The prognosis-related hub genes were then identified by Kaplan-Meier analysis based on the TCGA database. A total of 188 DEGs were obtained; 88 up-regulated genes and 100 down-regulated. The up-regulated DEGs were significantly associated with extracellular matrix organization and disassembly while down-regulated DEGs were significantly related to keratinocyte differentiation. Blue and turquoise co-expression modules were established and 18 hub genes were identified. The blue module was associated with mitotic nuclear division, cell division and mitotic cytokinesis and the turquoise module was associated with collagen catabolic process, extracellular matrix organization and keratinocyte differentiation. We established that the TPX2, CDK1 and CEP55 blue module hub genes were associated with relapse-free survival, and our overall results not only identify key genes but also provide potential novel biomarkers for ESCC diagnosis and treatment.

*Key words: esophageal squamous cell carcinoma, differential expression genes, functional enrichment analysis, WGCNA, Kaplan-Meier analysis*

Esophageal cancer is the 8th most common and 6th most fatal cancer worldwide [1]. It is a deadly disease with roughly 480,000 new patients every year [2]. The esophageal squamous cell carcinoma (ESCC) is histologically the most prevalent type of esophageal cancer with increasing morbidity [3]. Lacking effective early diagnosis, ESCC is usually detected at an advanced stage when the patients cannot swallow solid foods anymore with poor prognosis and clinical outcome. Thus, further research is needed to discover more effective diagnostic methods in early stage and the molecular mechanisms underlying ESCC in order to improve prevention and prognosis.

Recently, molecular markers are thought to be predictive and prognostic markers in ESCC [4]. As for immunohistochemical result, a systemic review demonstrated that the survival rate of ESCC patients with HER2-positive expression decreased because of radiation resistance possibly [5]. Additionally, ESCC patients with abnormal p53 expression show several times more rapid progression of the disease [6]. In terms of blood-based markers, Shimada et al found that the combination of four antibodies (SURF1, HOOK2, LOC146223 and AGENCOURT_7565913) were highly specific for ESCC [7]. Moreover, tubulin beta chain, filamin A alpha isoform 1 and cytochrome b-c1 complex subunit 1 were identified as the differentially expressed proteins and biomarkers in ESCC patients serum [8]. According to epigenetic markers, the hypermethylation of CDKN2A, MGMT, APC, DAB2, CDH1 and DACT2 were found to be more frequent in ESCC tissues [9]. However, these biomarkers are still not effective for early diagnosis and prognosis prediction. Accordingly, it is worth exploring additional biomarkers for more effective diagnosis.

Microarray technology is a widely used tool to explore genetic alteration during tumorigenesis [10]. In our study, we followed a strategy utilizing the larger microarray dataset as training set and another independent microarray dataset as the validation set. The quality control was performed on both datasets in order to increase credibility of result. Functional enrichment and co-expression network analysis were applied for DEGs while survival analysis was used to identify key genes in ESCC.

## Materials and methods

**Microarray data.** The GSE38129 [11] and GSE20347 [12] gene expression profiles were downloaded from the Gene Expression Omnibus [13] which is based on the Affymetrix Human Genome U133A 2.0 Array platform. The GSE38129 dataset included 30 ESCC tissues and 30 pairs of normal esophageal squamous epithelium tissue (N) and GSE20347 comprised 17 ESCC tissues and 17 pairs of N tissue. We utilized the larger dataset (GSE38129) as the training set and an independent dataset (GSE20347) as the validation set.

**Quality control and DEG analysis.** Quality control included relative logarithmic expression (RLE), RNA degradation and principal component analysis (PCA). Differentially expressed genes (DEGs) in ESCC tissue compared with N tissue were identified by Linear Models for Microarray Analysis(Limma) package in R software [14], and the Benjamini-Hochberg false discovery rate corrected P values. Adjusted p-value at <0.05 and fold change (FC)>=4 formed DEG analysis cut-off criteria. Finally, hierarchical clustering analysis and DEG heatmap in GSE38129 and GSE20347 was conducted by heatmap package in R software.

**Functional enrichment analysis of DEGs.** Gene Ontology [15] and Kyoto Encyclopedia of Genes and Genomes [16] are widely used in bioinformatics to identify the most correlative biological process(BP) and relevant pathway information. KEGG and GO BP analysis through the Database for Annotation Visualization and Integrated Discovery online tool (DAVID;david.ncifcrf.gov/) identified DEG biological significance [17]. P-value adjusted by Benjamini-Hochberg to <0.05 established the cut-off criteria.

**Co-expression module detection.** Weighted gene co-expression network analysis (WGCNA) [18] was restricted to training dataset DEGs by the R software WGCNA package. An unsupervised co-expression relationship was constructed by adjacency matrix of connection strengths with Pearson's correlation coefficients. The adjacency was defined by a soft threshold β selected to amplify strong connections between genes and penalize weak ones. Herein, the soft threshold was set at β=18 with scale-free topology criterion and the modules were identified as gene sets with a high topologic overlap. The top 10% of genes with the highest network connectivity were identified as module hub genes. Average linkage hierarchical clustering comprised topological overlap matrix based dissimilarity measure, and minimum gene module size of 30 cut the branches. Finally, each module's co-expression network was presented by Cytoscape 3.4.0 [19].

**Survival analysis of hub genes.** The mRNA transcript per million (TPM) of 96 ESCC tissue samples and corresponding patient follow-up information were downloaded from UCSC Xena (https://xenabrowser.net/datapages/?host=https://tcga.xenahubs.net). The survival R package [20] was subjected to survival analysis to explore the prognosis value of hub genes. Kaplan-Meier survival curves were plotted and relapse-free survival (RFS) provided survival endpoints. Patients with ESCC were divided into low and high expression groups according to the median of each hub gene expression, with logrank p-value <0.05 significant.

**Table 1. The significant enriched GO BP terms and KEGG pathways.**

| | Description | No. enriched genes | Adjusted p-value |
|---|---|---|---|
| **Upregulated** | | | |
| GO:0030574 | Collagen catabolic process | 12 | $1.20\times10^{-11}$ |
| GO:0030198 | Extracellular matrix organization | 14 | $6.08\times10^{-9}$ |
| GO:0001501 | Skeletal system development | 11 | $4.15\times10^{-7}$ |
| GO:0022617 | Extracellular matrix disassembly | 9 | $8.00\times10^{-7}$ |
| GO:0007067 | Mitotic nuclear division | 11 | $7.00\times10^{-5}$ |
| GO:0030199 | Collagen fibril organization | 6 | $1.87\times10^{-4}$ |
| GO:0051301 | Cell division | 10 | $7.12\times10^{-3}$ |
| GO:0035987 | Endodermal cell differentiation | 4 | $3.20\times10^{-2}$ |
| GO:0000281 | Mitotic cytokinesis | 4 | $3.52\times10^{-2}$ |
| GO:0008283 | Cell proliferation | 9 | $3.83\times10^{-2}$ |
| hsa04512 | ECM-receptor interaction | 8 | $2.86\times10^{-5}$ |
| hsa05146 | Amoebiasis | 8 | $5.52\times10^{-5}$ |
| hsa04510 | Focal adhesion | 9 | $3.25\times10^{-4}$ |
| hsa04974 | Protein digestion and absorption | 6 | $2.13\times10^{-3}$ |
| hsa04151 | PI3K-Akt signaling pathway | 9 | $7.34\times10^{-3}$ |
| hsa05202 | Transcriptional misregulation in cancer | 6 | $2.73\times10^{-3}$ |
| **Downregulated** | | | |
| GO:0030216 | Keratinocyte differentiation | 7 | $1.81\times10^{-3}$ |

GO BP, Gene Ontology Biologic Process. KEGG, Kyoto Encyclopedia of Genes and Genomes.

## Results

**Quality control and DEG analysis.** Quality control included RLE, RNA degradation curve and PCA. The RLE in Figures 1A and 2A for both results showed high normalization level in the samples. The constantly rising RNA degradation curve in both results demonstrated the samples' undegraded RNA (Figures 1B, 2B) and PCA determined no absolute separation between different samples in either dataset, and thus imperfect quality (Figures 1C, 2V). Following integrated analysis of all of quality control processes, only 27 pairs of samples in GSE38129 and 16 in GSE20347 were retained and the following were excluded; GSM509791 and GSM509808 in GSE20347 and GSM935156, GSM935157, GSM935174, GSM935175, GSM935186 and GSM935187 in GSE38129. The PCA result of residual samples in GSE38129 and GSE20347 are presented in Figures 1D and 2D. The total of 188 DEGs obtained by Limma package comprised 88 up-regulated DEG's and 100 down-regulated and results are detailed in Supplementary Table 1. The DEGs in training and validation set heatmaps are listed in Figure 3.

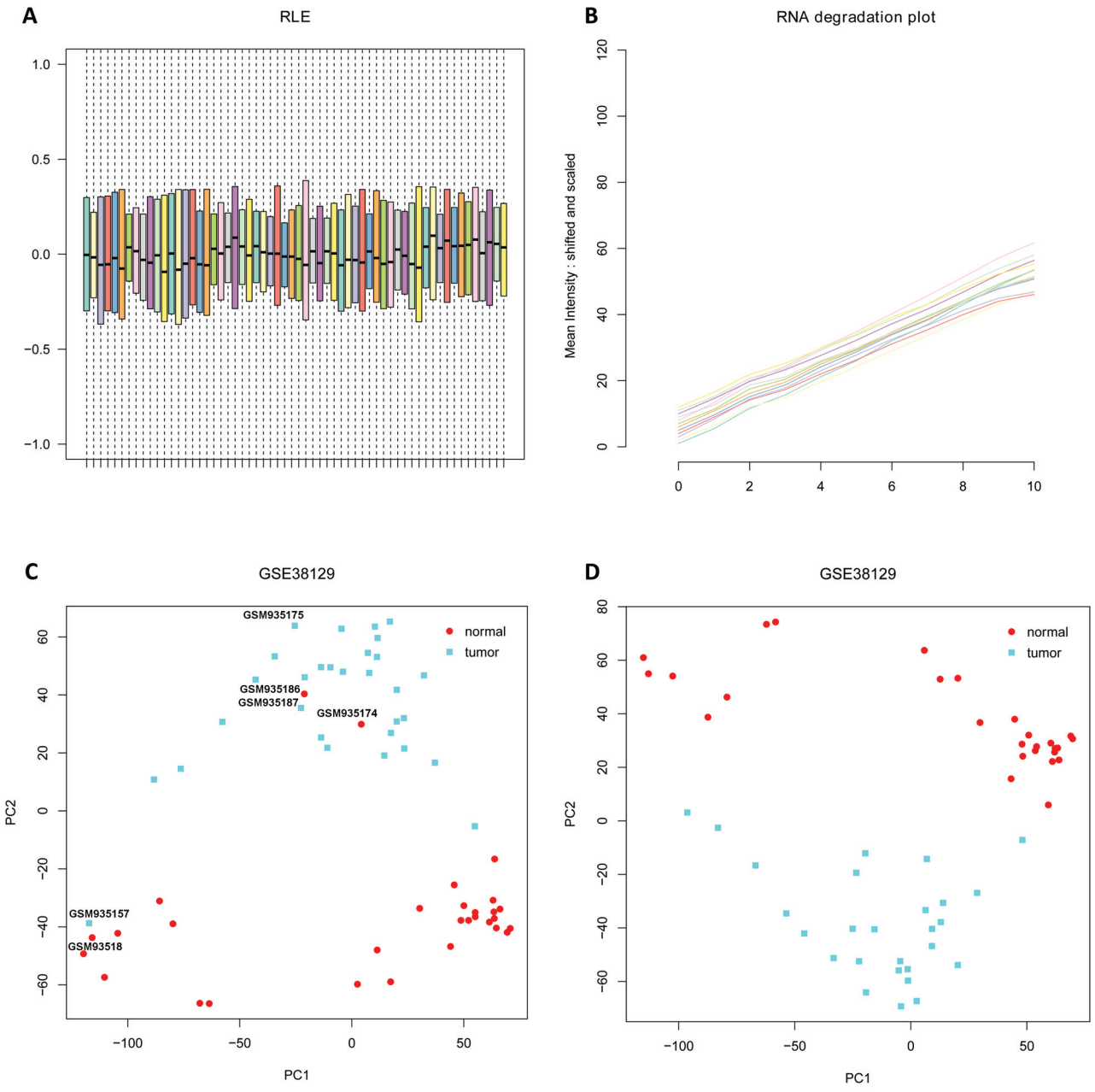


**Figure 1. Quality control of the GSE38129 dataset. RLE and RNA degradation results performed on (A) and (B). PCA result before and after excluding GSM935156, GSM935157, GSM935174, GSM935175, GSM935186 and GSM935187 performed on (C) and (D), respectively. The higher level of normalization in RLE and separation of different samples in PCA present higher quality micro-array. The continuously rising RNA degradation curve reveals undegraded RNA obtained from samples. RLE = relative logarithmic expression. PCA = principal component analysis.**

**Functional enrichment analysis.** GO BP analysis revealed that up-regulated DEGs were significantly associated with collagen catabolic process, extracellular matrix organization and extracellular matrix disassembly and down-regulated DEGs were significantly related to keratinocyte differentiation. In contrast, the KEGG pathway analysis demonstrated that the up-regulated genes were significantly enriched in ECM-receptor interaction, focal adhesion and the PI3K-Akt signaling pathway but it established no significantly enriched pathways for down-regulated genes. The basic results are highlighted in Table 1 and more detailed analysis is contained in Supplementary Table 2.

**Co-expression module detection and functional enrichment analysis.** WGCNA analyzed the 188 DEGs expression profiles in the construct of gene co-expression modules.

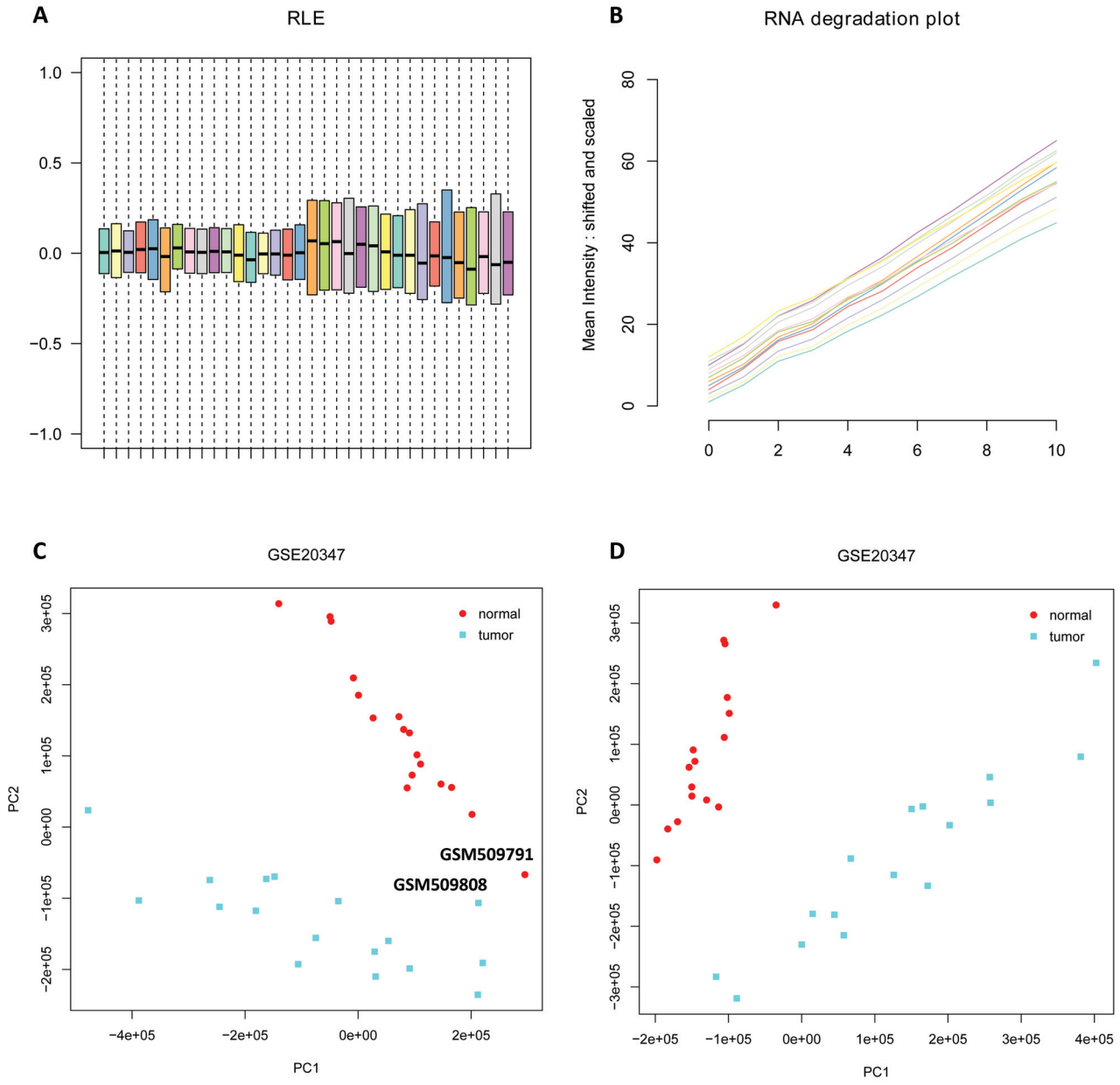


**Figure 2. Quality control of the GSE20347 dataset. RLE and RNA degradation results performed on (A) and (B). PCA result before and after excluding GSM509791 and GSM509808 performed on (C) and (D), respectively. The higher level of normalization in RLE and separation of different samples in PCA present higher quality micro-array. The constantly rising RNA degradation curve rising constantly reveals undegraded RNA obtained from samples. RLE = relative logarithmic expression. PCA = principal component analysis.**

'Blue' and 'Turquise' modules were identified, with 61 and 106 genes respectively, and 18 hub genes were identified. The co-expression networks of these modules and the hub genes of each module are pictured in Figures 4 and 5. Genes in blue modules are associated with mitotic nuclear division, cell division and mitotic cytokinesis in GO BP analysis and those in the turquoise modules denote collagen catabolic processes, extracellular matrix organization and keratinocyte differentiation in GO BP analysis, and protein digestion, amoebiasis and ECM-receptor interaction in KEGG analysis. Basic results are in Table 2 and greater detail is provided in Supplementary Table 3.

**Survival analysis of hub genes.** Follow-up information revealed that only 82 Relapse-Free Survival (RFS) patient notifications were obtained from the TCGA database. The prognostic value of the 18 hub genes was assessed by the R software survival package and Kaplan-Meier analysis established that the low expression of TPX2 ($p<0.001$), CEP55

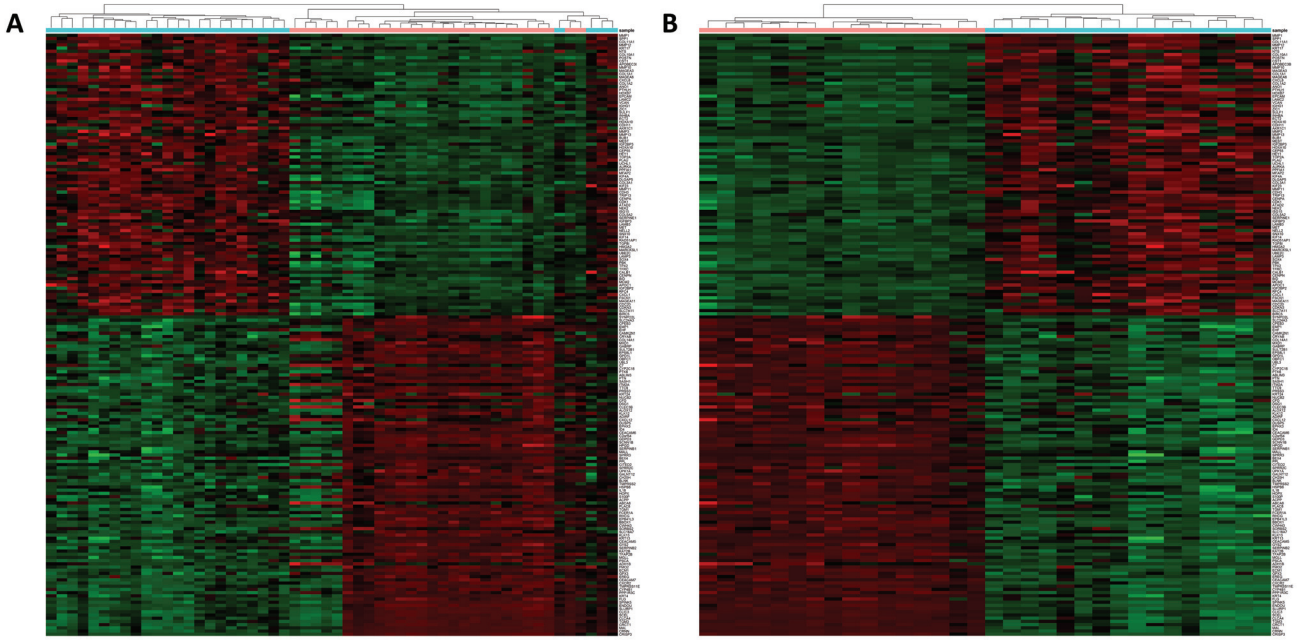


**Figure 3. Heatmap and hierarchical clustering analysis of DEGs in training set (A) and validation set (B). Color depth shows the expression of DEGs; with red illustrating over-expression and green low expression. The side color on the top shows sample classification; with turquoise illustrating ESCC samples and pink highlighting normal esophageal squamous epithelium. DEGs = differentially expressed genes. ESCC = esophageal squamous cell carcinoma.**

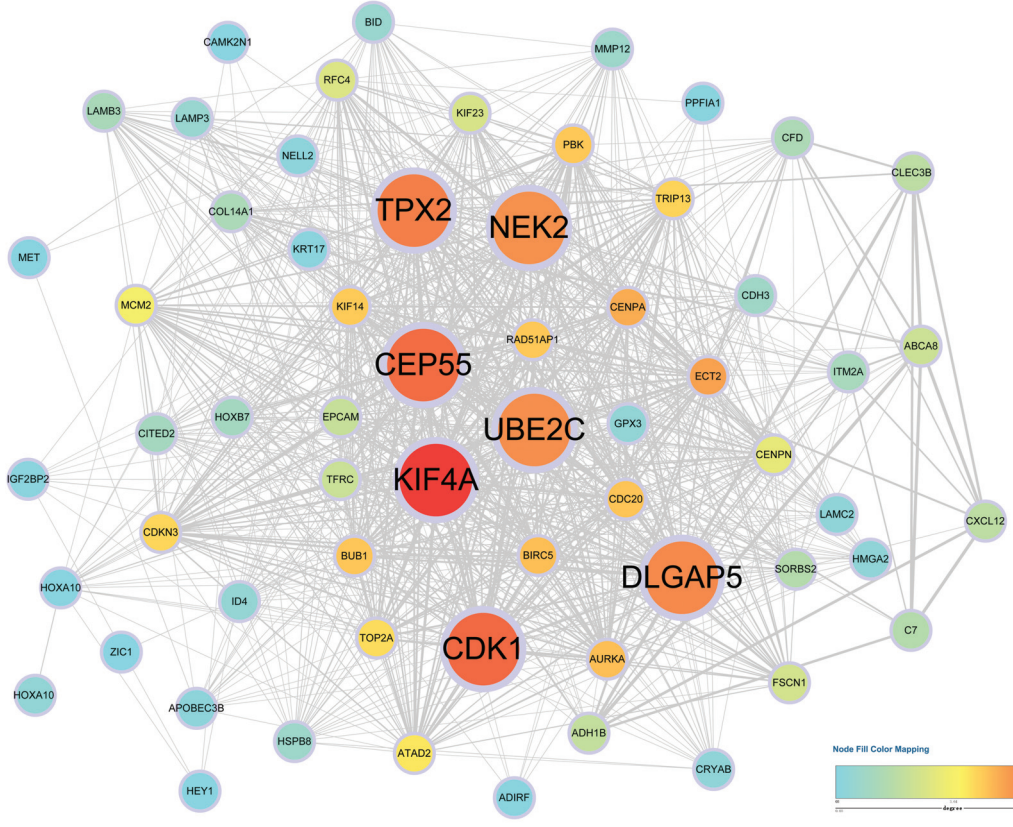


**Figure 4. The co-expression network of blue modules identified by WGCNA. Color depth presents node connectivity degree, with red illustrating higher degree and turquoise denoting lower degree. Larger size nodes identify hub genes with the top 10% highest network connectivity degree. WGCNA = Weighted Gene Co-Expression Network Analysis.**

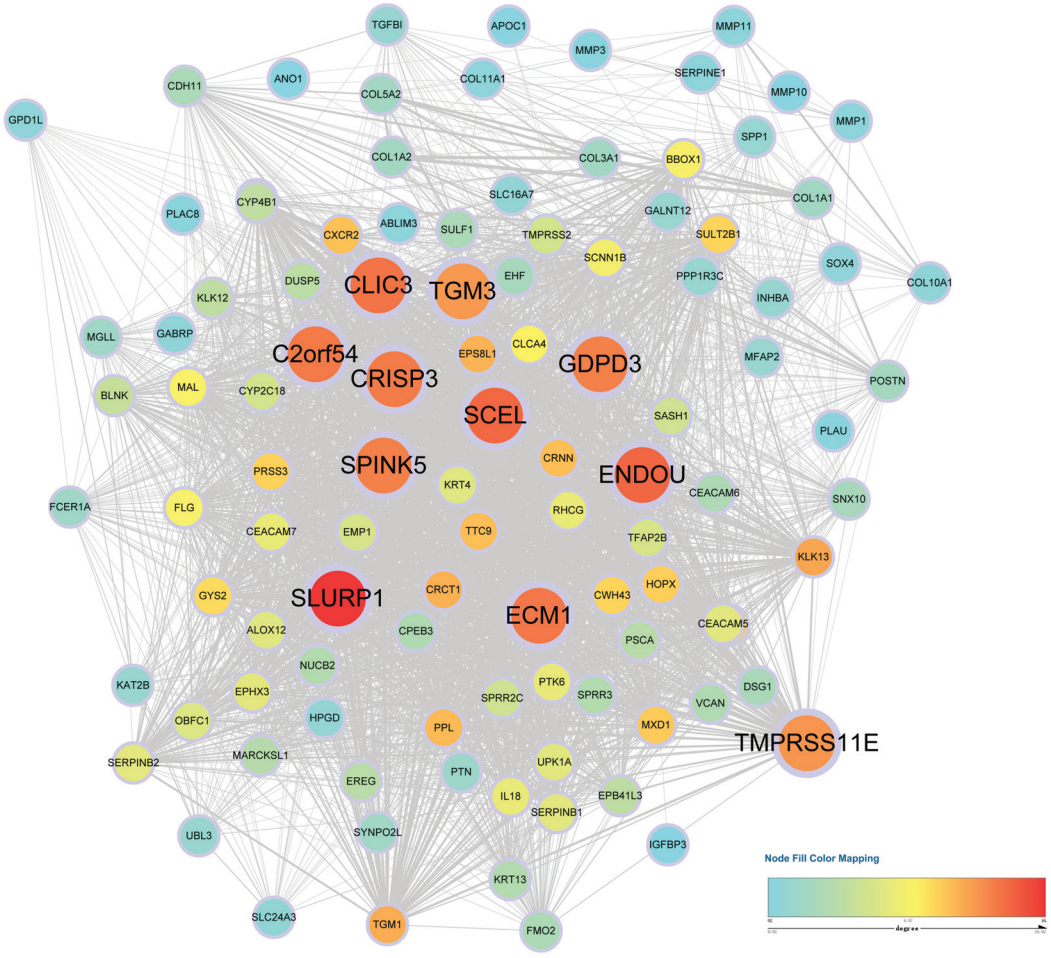


**Figure 5. The co-expression network of turquoise modules identified by WGCNA. Color depth identifies node connectivity degree; with red illustrating higher degree of connectivity and turquoise is for lower connectivity. The larger size nodes denote hub genes with the top 10% highest network connectivity degree. WGCNA = Weighted Gene Co-Expression Network Analysis.**

**Table 2. The significantly enriched GO BP and KEGG pathways for co-expression modules.**

| | Description | No. enriched genes | Adjusted p-value |
|---|---|---|---|
| **Module Blue** | | | |
| GO:0007067 | Mitotic nuclear division | 11 | $8.01\times10^{-6}$ |
| GO:0051301 | Cell division | 10 | $9.81\times10^{-4}$ |
| GO:0000281 | Mitotic cytokinesis | 4 | $2.71\times10^{-2}$ |
| **Module Turquoise** | | | |
| GO:0030574 | Collagen catabolic process | 10 | $7.46\times10^{-8}$ |
| GO:0030198 | Extracellular matrix organization | 13 | $4.91\times10^{-7}$ |
| GO:0001501 | Skeletal system development | 9 | $3.00\times10^{-4}$ |
| GO:0030199 | Collagen fibril organization | 6 | $5.13\times10^{-4}$ |
| GO:0030216 | Keratinocyte differentiation | 7 | $7.20\times10^{-4}$ |
| hsa04974 | Protein digestion and absorption | 7 | $1.26\times10^{-3}$ |
| hsa05146 | Amoebiasis | 7 | $1.82\times10^{-3}$ |
| hsa04512 | ECM-receptor interaction | 6 | $5.20\times10^{-3}$ |

GO BP, Gene Ontology Biologic Process. KEGG, Kyoto Encyclopedia of Genes and Genomes.

(p=0.017) and CDK1 (p=0.039) genes was associated with worse RFS in ESCC patients (Figure 6).

## Discussion

A total of 188 genes were identified herein. These comprised 88 up-regulated genes and 100 down-regulated genes and almost all DEGs obtained in the training set were verified in the validation set. The up-regulated genes were associated with extracellular matrix organization and disassembly while down-regulated genes were associated with keratinocyte differentiation. Further, WGCNA identified blue and turquoise co-expression modules consisting of 61 and 106 genes, respectively, and also 18 hub genes. The blue module was associated with mitosis and cell division and the turquoise module involved extracellular matrix organization and catabolic processes. Kaplan-Meier analysis of hub genes then revealed that low TPX2, CDK1 and CEP55 expression was significantly related to worse relapse-free

survival. Interestingly, these 3 prognosis-related hub genes were all part of the blue module and GO BP mitotic nuclear division. These are considered key genes and their related biologic processes are crucial to the mechanisms involved in ESCC progression. Moreover all entities in the enriched GO BP and KEGG pathways most likely participate in mechanisms underlying ESCC progression and these therefore require urgent attention.

The TPX2 target protein for Xenopus kinesin-like protein 2 is a microtubule-associated protein-coding gene located on chromosome 20q11.21. TPX2 is upregulated in multiple tumor types such as cervical and gastric cancer [21, 22] and its expression is associated with growth and metastasis in hepatocellular carcinoma [23]. Researchers report that TPX2 regulates ESCC cells proliferation and invasiveness [24] and Hsu et al demonstrated that high TPX2 expression was associated with worse overall survival and shorter disease-free survival [25]. While Hsu's work supports and confirms our determination of the crucial role of TPX2 in ESCC progression, we did not establish its same effect in prognosis.

The Cyclin-dependent kinase 1, protein coding gene on chromosome 10q21.2 has a crucial role in the G2/M cell cycle. CDK1 activation and the formation of the cyclin B1-CDK1 complex is controlled by inhibitory phosphorylation of tyrosine and threonine in the early cell cycle phases, and CDK1 is activated by CDC25C phosphatase in the late G2 phase which is an obligatory step in G2/M transition [26]. Moreover, over-expression of CDK1 has been reported in many tumors, including breast cancer and hepatocellular carcinoma [27, 28].

Although some research has demonstrated that aberrant CDK1 expression is associated with poor patient prognosis [29], to the best of our knowledge no previous research has investigated CDK1 in ESCC. Our experimental results strongly suggest that CDK1 is an esophageal squamous cell carcinoma G2/M pathway regulator and that the combination of CDK1 with other regulators such as CDC25 should enhance the prediction of patient prognosis.

Centrosomal protein 55 (CEP55) is the protein coding gene on chromosome 10q23.33, and research has elucidated that this protein has an important role in regulating the PI3K/AKT pathway and promoting tumorigenesis [30]. CEP55 has been reported to bind PIK3CA and regulate the PI3K/AKT pathway; thus promoting phosphorylation and stimulating activation of AKT which has diverse roles in the cell cycle, cell survival and protein synthesis. CEP55 has also been identified in prognostic signatures for multiple cancer cell lines [31] and its over-expression has been related to poor clinical parameters including tumor stage, margin status and plasma tumor marker levels. Moreover, Jiang et al illustrated that CEP55 over-expression is significantly associated with reduced overall patient survival after surgery and, importantly, that the 5-year survival rate of ESCC patients with CEP55 over-expression was lower

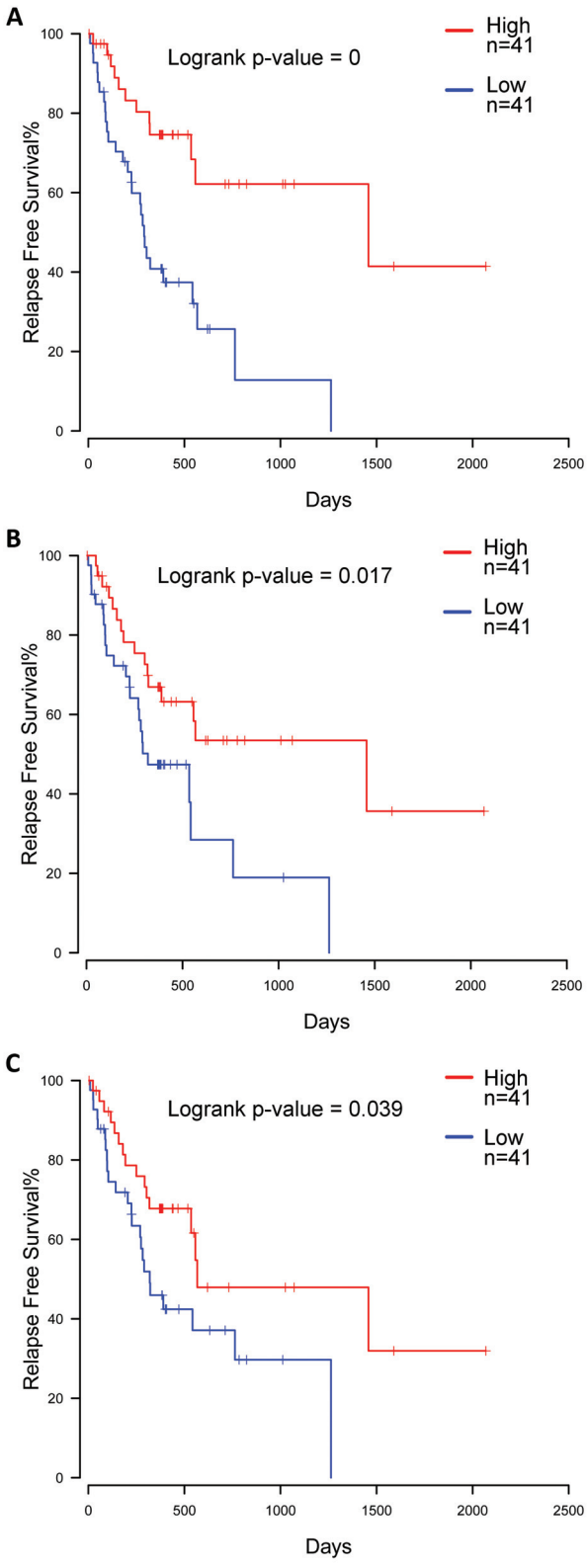


**Figure 6. Prognostic value of TPX2(A), CDK1(B) and CEP55(C) for relapse-free survival in ESCC patients. The patients were divided into high and low expression groups according to the median of each DEG expression. ESCC = esophageal squamous cell carcinoma.**

than those with under-expression [32]. However, we did not establish this finding, and we consider that our lack of support for Jiang's conclusion is due to the different populations, different expression measurements and different tumor tissue location in our individual research. Nevertheless, further assessment of CEP55 and ESCC relationships certainly appears warranted.

In conclusion, the results of our study proceed from screening 188 DEGs, and 2 co-expression modules and 18 hub genes were identified in these DEG's. These encode the enriched pathway, mitotic nuclear division, cell division, extracellular matrix organization and catabolic processes closely related to ESCC progression. Further, TPX2, CDK1 and CEP55 are considered key genes in esophageal squamous cell carcinoma prognosis. Since additional experiments and clinical research were not performed to verify resultant proteins ultimate significance, further study is advised. Focus on these key genes, their pathways, molecular mechanisms and clinical applications should enhance esophageal squamous cell carcinoma diagnosis and treatment.

**Supplementary information** is available in the online version of the paper.

Acknowledgments: This study was supported by the National Science Foundation of China (grant number 81370491), the National Science Foundation of China (grant number 8157040827) and National Science Foundation of China(grant number 81502041).

## References

[1] HONGO M, NAGASAKI Y, SHOJI T. Epidemiology of esophageal cancer: Orient to Occident. Effects of chronology, geography and ethnicity. J Gastroenterol Hepatol 2009; 24: 729–735. https://doi.org/10.1111/j.1440-1746.2009.05824.x

[2] JEMAL A, BRAY F, CENTER M M, FERLAY J, WARD E et al. Global cancer statistics. CA Cancer J Clin 2011, 61: 69–90. https://doi.org/10.3322/caac.20107

[3] HUANG F L, YU S J. Esophageal cancer: Risk factors, genetic association, and treatment. Asian Journal of Surgery, 2016. https://doi.org/10.1016/j.asjsur.2016.10.005

[4] MEHTA S, SHELLING A, MUTHUKARUPPAN A, LASHAM A, BLENKIRON C et al. Predictive and prognostic molecular markers for cancer medicine. Ther Adv Med Oncol 2010, 2: 125–148. https://doi.org/10.1177/1758834009360519

[5] CHAN DS, TWINE CP, LEWIS WG. Systematic Review and Meta-analysis of the Influence of HER2 Expression and Amplification in Operable Oesophageal Cancer J Gastrointest Surg 2012, 16: 1821–1829. https://doi.org/10.1007/s11605-012-1979-2

[6] HONG Y, MIAO X, ZHANG X, DING F, LUO A et al. The role of P53 and MDM2 polymorphisms in the risk of esophageal squamous cell carcinoma. Cancer Res 2005, 65: 9582–9587. https://doi.org/10.1158/0008-5472.CAN-05-1460

[7] SHIMADA H, NAKASHIMA K, OCHIAI T, NABEYA Y, TAKIGUCHI M et al. Serological identification of tumor antigens of esophageal squamous cell carcinoma. Int J Oncol 2005, 26: 77–86.

[8] FAN NJ, GAO CF, WANG XL. Tubulin beta chain, filamin A alpha isoform 1, and cytochrome b-c1 complex subunit 1 as serological diagnostic biomarkers of esophageal squamous cell carcinoma: a proteomics study. Omics 2013, 17: 215–223. https://doi.org/10.1089/omi.2012.0133

[9] AHRENS TD, WERNER M, LASSMANN S. Epigenetics in esophageal cancers. Cell Tissue Res 2014; 356: 643–655. https://doi.org/10.1007/s00441-014-1876-y

[10] HELLER MJ. DNA microarray technology: devices, systems, and applications. Annu Rev Biomed Eng 2002, 4: 129–153. https://doi.org/10.1146/annurev.bioeng.4.020702.153438

[11] HU N, WANG C, CLIFFORD RJ, YANG HH, SU H et al. Integrative genomics analysis of genes with biallelic loss and its relation to the expression of mRNA and micro-RNA in esophageal squamous cell carcinoma. BMC Genomics 2015, 16: 732. https://doi.org/10.1186/s12864-015-1919-0

[12] HU N, CLIFFORD RJ, YANG HH, WANG C, GOLDSTEIN AM et al. Genome wide analysis of DNA copy number neutral loss of heterozygosity (CNNLOH) and its relation to gene expression in esophageal squamous cell carcinoma. BMC Genomics 2010, 11: 576. https://doi.org/10.1186/1471-2164-11-576

[13] BARRETT T, WILHITE SE, LEDOUX P, EVANGELISTA C, KIM IF et al. NCBI GEO: archive for functional genomics data sets-update. Nucleic Acids Res 2013, 41: D991–995. https://doi.org/10.1093/nar/gks1193

[14] RITCHIE ME, PHIPSON B, WU D, HU Y, LAW CW et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Research 2015, 43: e47. https://doi.org/10.1093/nar/gkv007

[15] ASHBURNER M, BALL CA, BLAKE JA, BOTSTEIN D, BUTLER H et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 2000, 25: 25–29. https://doi.org/10.1038/75556

[16] DU J, YUAN Z, MA Z, SONG J, XIE X et al. KEGG-PATH: Kyoto encyclopedia of genes and genomes-based pathway analysis using a path analysis model. Mol Biosyst 2014; 10: 2441–2447. https://doi.org/10.1039/c4mb00287c

[17] DA WH, SHERMAN BT, TAN Q, COLLINS JR, ALVORD WG et al. The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists. Genome Biol 2007, 8: R183. https://doi.org/10.1186/gb-2007-8-9-r183

[18] LANGFELDER P, HORVATH S. Fast R Functions for Robust Correlations and Hierarchical Clustering. J Stat Softw 2012, 46: pii: i11.

[19] SHANNON P, MARKIEL A, OZIER O, BALIGA NS, WANG JT et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 2003; 13: 2498–2504. https://doi.org/10.1101/gr.1239303

[20] THERNEAU TM, GRAMBSCH PM (Eds.). Modeling Survival Data: Extending the Cox Model. Springer-Verlag New York, 2000, pp. 350. ISBN 978-0-387-98784-2

[21] JIANG P, SHEN K, WANG X, SONG H, YUE Y et al. TPX2 regulates tumor growth in human cervical carcinoma cells. Mol Med Rep 2014, 9: 2347–2351. https://doi.org/10.3892/mmr.2014.2106

[22] TOMII C, INOKUCHI M, TAKAGI Y, ISHIKAWA T, OTSUKI S et al. TPX2 expression is associated with poor survival in gastric cancer. World J Surg Oncol 2017, 15: 14. https://doi.org/10.1186/s12957-016-1095-y

[23] HUANG Y, GUO W, KAN H. TPX2 Is a Prognostic Marker and Contributes to Growth and Metastasis of Human Hepatocellular Carcinoma. Int J Mol Sci 2014, 15: 18148–18161. https://doi.org/10.3390/ijms151018148

[24] LIU HC, ZHANG GH, LIU YH, WANG P, MA JF et al. TPX2 siRNA regulates growth and invasion of esophageal cancer cells. Biomed Pharmacother 2014, 68: 833–839. https://doi.org/10.1016/j.biopha.2014.08.008

[25] HSU PK, CHEN HY, YEH YC, YEN CC, WU YC et al. TPX2 expression is associated with cell proliferation and patient outcome in esophageal squamous cell carcinoma. J Gastroenterol 2014, 49: 1231–1240. https://doi.org/10.1007/s00535-013-0870-6

[26] WANG Z, SLIPICEVIC A, FORSUND M, TROPE CG, NESLAND JM et al. Expression of CDK1Tyr15, pCDK-1Thr161, Cyclin B1 (Total) and pCyclin B1Ser126 in Vulvar Squamous Cell Carcinoma and Their Relations with Clinicopatological Features and Prognosis. PLoS One 2015, 10: e0121398. https://doi.org/10.1371/journal.pone.0121398

[27] CHAE SW, SOHN JH, KIM DH, CHOI YJ, PARK YL et al. Overexpressions of Cyclin B1, cdc2, p16 and p53 in human breast cancer: the clinicopathologic correlations and prognostic implications. Yonsei Med J 2011, 52: 445–453. https://doi.org/10.3349/ymj.2011.52.3.445

[28] ZHAO J, HAN SX, MA JL, YING X, LIU P et al. The role of CDK1 in apoptin-induced apoptosis in hepatocellular carcinoma cells. Oncol Rep 2013, 30: 253–259. https://doi.org/10.3892/or.2013.2426

[29] SUNG WW, LIN YM, WU PR, YEN HH, LAI HW et al. High nuclear/cytoplasmic ratio of Cdk1 expression predicts poor prognosis in colorectal cancer patients. BMC Cancer 2014, 14: 951. https://doi.org/10.1186/1471-2407-14-951

[30] JEFFERY J, NEYT C, MOORE W, PATERSON S, BOWER NI et al. Cep55 regulates embryonic growth and development by promoting Akt stability in zebrafish. FASEB J 2015, 29: 1999–2009. https://doi.org/10.1096/fj.14-265090

[31] MARTINEZ-GARAY I, RUSTOM A, GERDES HH, KUTSCHE K. The novel centrosomal associated protein CEP55 is present in the spindle midzone and the midbody. Genomics 2006, 87: 243–253. https://doi.org/10.1016/j.ygeno.2005.11.006

[32] JIANG W, WANG Z, JIA Y. CEP55 overexpression predicts poor prognosis in patients with locally advanced esophageal squamous cell carcinoma. Oncol Lett 2017, 13: 236–242. https://doi.org/10.3892/ol.2016.5414

# Integrated analysis of differentially expressed genes in esophageal squamous cell carcinoma using bioinformatics

Z. DONG, H. ZHANG, T. ZHAN, S. XU*

## Supplemental Material

**Supplementary Table 1. Differentially expressed genes of ESCC.**

| Genes | log FC | Adjusted p-value |
|---|---|---|
| **Up-regulated** | | |
| MMP1 | –5.248 | 1.36E-11 |
| SPP1 | –4.696 | 4.27E-12 |
| COL11A1 | –4.334 | 4.18E-13 |
| MMP12 | –3.772 | 5.07E-13 |
| KRT17 | –3.667 | 3.94E-08 |
| NTS | –3.459 | 1.94E-04 |
| COL10A1 | –3.398 | 4.51E-14 |
| POSTN | –3.35 | 2.53E-07 |
| CST1 | –3.092 | 2.86E-09 |
| APOBEC3B | –3.085 | 7.90E-11 |
| MMP10 | –3.065 | 6.94E-07 |
| MAGEA3 | –3.053 | 7.94E-05 |
| COL1A1 | –3.046 | 3.68E-08 |
| MAGEA6 | –2.984 | 1.57E-04 |
| CXCL8 | –2.929 | 9.00E-08 |
| COL1A2 | –2.901 | 1.76E-07 |
| ANO1 | –2.881 | 2.63E-06 |
| PTHLH | –2.88 | 1.83E-05 |
| HOXB7 | –2.774 | 2.54E-15 |
| EPCAM | –2.725 | 1.81E-10 |
| LAMC2 | –2.714 | 1.56E-09 |
| VCAN | –2.713 | 1.88E-08 |
| IGHG1 | –2.635 | 6.95E-05 |
| ZIC1 | –2.63 | 6.11E-08 |
| SULF1 | –2.627 | 3.49E-08 |
| INHBA | –2.574 | 9.49E-07 |
| ECT2 | –2.561 | 4.59E-16 |
| HOXA10 | –2.559 | 3.79E-06 |
| CDH11 | –2.548 | 7.13E-08 |
| AKR1C1 | –2.507 | 3.34E-04 |
| MMP3 | –2.507 | 2.56E-06 |
| MMP13 | –2.482 | 1.37E-07 |
| BUB1 | –2.48 | 2.70E-14 |
| MEST | –2.417 | 8.70E-10 |
| IGF2BP3 | –2.409 | 5.32E-06 |
| HOXA10 | –2.398 | 4.80E-12 |
| CEP55 | –2.397 | 3.53E-13 |
| HEY1 | –2.387 | 3.68E-10 |
| TOP2A | –2.378 | 1.87E-12 |
| PLAU | –2.377 | 2.22E-13 |
| UCHL1 | –2.358 | 2.35E-05 |
| AURKA | –2.357 | 6.41E-17 |
| PPFIA1 | –2.353 | 1.89E-06 |
| MFAP2 | –2.34 | 3.34E-11 |
| KIF4A | –2.339 | 4.93E-16 |
| DLGAP5 | –2.322 | 3.82E-12 |
| COL3A1 | –2.281 | 1.09E-05 |
| **Up-regulated** | | |
| KIF23 | –2.279 | 2.46E-12 |
| MMP11 | –2.267 | 1.44E-11 |
| CDH3 | –2.266 | 7.97E-08 |
| TRIP13 | –2.247 | 1.75E-14 |
| CENPA | –2.239 | 5.19E-13 |
| CDK1 | –2.229 | 1.18E-12 |
| ATAD2 | –2.229 | 1.75E-14 |
| NEK2 | –2.219 | 1.78E-14 |
| ISG15 | –2.216 | 3.34E-07 |
| COL5A2 | –2.216 | 7.80E-07 |
| SERPINE1 | –2.213 | 1.11E-08 |
| IGFBP3 | –2.207 | 3.43E-06 |
| LAMB3 | –2.2 | 4.66E-09 |
| MET | –2.184 | 1.41E-11 |
| NELL2 | –2.184 | 2.41E-09 |
| SNX10 | –2.181 | 8.44E-09 |
| KIF14 | –2.179 | 1.75E-14 |
| RAD51AP1 | –2.167 | 1.78E-13 |
| TGFBI | –2.153 | 2.30E-06 |
| HMGA2 | –2.152 | 3.96E-07 |
| MARCKSL1 | –2.151 | 2.80E-08 |
| UBE2C | –2.144 | 8.45E-13 |
| LAMP3 | –2.142 | 7.74E-09 |
| SOX4 | –2.134 | 8.33E-10 |
| PBK | –2.127 | 1.72E-11 |
| TPX2 | –2.123 | 6.68E-12 |
| TFRC | –2.114 | 1.70E-11 |
| CALB1 | –2.101 | 4.14E-03 |
| CENPN | –2.081 | 1.96E-10 |
| BID | –2.072 | 9.24E-14 |
| MCM2 | –2.071 | 8.07E-16 |
| APOC1 | –2.067 | 5.37E-10 |
| IGF2BP2 | –2.065 | 1.61E-09 |
| RFC4 | –2.06 | 2.70E-14 |
| CXCL1 | –2.052 | 2.53E-06 |
| FSCN1 | –2.051 | 1.78E-13 |
| MAGEA11 | –2.051 | 2.41E-04 |
| CDC20 | –2.025 | 1.90E-10 |
| CDKN3 | –2.018 | 1.68E-12 |
| SLC7A11 | –2.015 | 4.29E-04 |
| BIRC5 | –2.009 | 7.08E-11 |
| **Down-regulated** | | |
| SYNPO2L | 2 | 1.55E-05 |
| SLC24A3 | 2.006 | 3.66E-10 |
| CPEB3 | 2.022 | 5.23E-14 |
| EMP1 | 2.037 | 3.49E-06 |
| EHF | 2.037 | 4.10E-03 |
| **Down-regulated** | | |
| CAMK2N1 | 2.039 | 8.71E-09 |
| CRYAB | 2.048 | 5.63E-08 |
| COL14A1 | 2.05 | 2.70E-07 |
| MXD1 | 2.051 | 4.75E-07 |
| GABRP | 2.053 | 2.83E-03 |
| SULT2B1 | 2.056 | 1.29E-05 |
| EPS8L1 | 2.08 | 5.55E-09 |
| GPD1L | 2.083 | 4.49E-14 |
| OBFC1 | 2.096 | 4.56E-10 |
| UBL3 | 2.103 | 1.06E-18 |
| C7 | 2.103 | 5.15E-04 |
| CYP2C18 | 2.105 | 5.75E-06 |
| PTK6 | 2.112 | 1.58E-04 |
| ABLIM3 | 2.113 | 3.04E-13 |
| PTN | 2.126 | 1.84E-07 |
| SASH1 | 2.133 | 2.51E-11 |
| ITM2A | 2.136 | 1.85E-06 |
| TTC9 | 2.15 | 3.28E-07 |
| PRSS3 | 2.15 | 2.39E-06 |
| KRT24 | 2.164 | 1.40E-03 |
| NUCB2 | 2.171 | 3.65E-12 |
| CFD | 2.177 | 4.65E-06 |
| DSG1 | 2.196 | 4.64E-04 |
| CLEC3B | 2.197 | 2.00E-04 |
| ALOX12 | 2.201 | 3.12E-04 |
| KLK12 | 2.211 | 9.40E-05 |
| ADIRF | 2.214 | 1.19E-09 |
| CXCL12 | 2.233 | 1.31E-04 |
| DUSP5 | 2.248 | 7.88E-08 |
| EPHX3 | 2.252 | 9.47E-05 |
| ID4 | 2.254 | 2.84E-11 |
| CEACAM6 | 2.267 | 3.24E-03 |
| C2orf54 | 2.267 | 5.70E-08 |
| GDPD3 | 2.276 | 1.75E-09 |
| SCNN1B | 2.284 | 6.00E-08 |
| HPGD | 2.287 | 1.51E-09 |
| SERPINB1 | 2.306 | 4.21E-06 |
| MALL | 2.31 | 3.31E-05 |
| SPRR3 | 2.32 | 1.01E-02 |
| BEX4 | 2.327 | 6.38E-07 |
| PPL | 2.34 | 4.13E-05 |
| CITED2 | 2.346 | 1.78E-14 |
| SPRR2C | 2.36 | 7.29E-04 |
| UPK1A | 2.365 | 2.54E-06 |
| GALNT12 | 2.38 | 9.48E-11 |
| CH25H | 2.403 | 2.37E-08 |
| BLNK | 2.412 | 5.56E-07 |
| **Down-regulated** | | |
| TMPRSS2 | 2.413 | 2.41E-07 |
| HSPB8 | 2.439 | 8.11E-11 |
| IL18 | 2.466 | 3.74E-05 |
| HOPX | 2.473 | 2.40E-05 |
| S100P | 2.491 | 1.30E-04 |
| ACPP | 2.494 | 2.35E-05 |
| ABCA8 | 2.512 | 7.12E-06 |
| PLAC8 | 2.532 | 7.69E-05 |
| TGM1 | 2.559 | 2.90E-05 |
| FCER1A | 2.569 | 3.98E-14 |
| RHCG | 2.585 | 3.08E-03 |
| EPB41L3 | 2.614 | 8.80E-13 |
| BBOX1 | 2.645 | 5.20E-07 |
| CWH43 | 2.664 | 3.02E-08 |
| SORBS2 | 2.672 | 6.30E-10 |
| SLC16A7 | 2.713 | 2.38E-09 |
| KLK13 | 2.718 | 3.27E-08 |
| KRT13 | 2.722 | 4.66E-03 |
| CEACAM5 | 2.723 | 1.20E-04 |
| GYS2 | 2.74 | 5.42E-10 |
| SERPINB2 | 2.743 | 8.22E-04 |
| KAT2B | 2.772 | 1.65E-16 |
| TFAP2B | 2.796 | 3.76E-08 |
| MGLL | 2.828 | 5.91E-14 |
| PSCA | 2.841 | 5.33E-07 |
| ADH1B | 2.87 | 2.52E-06 |
| FMO2 | 2.966 | 1.70E-06 |
| ECM1 | 2.982 | 1.78E-07 |
| GPX3 | 3.014 | 9.95E-13 |
| EREG | 3.118 | 7.05E-08 |
| CEACAM7 | 3.264 | 4.36E-06 |
| CXCR2 | 3.317 | 1.17E-09 |
| TMPRSS11E | 3.322 | 5.09E-05 |
| CYP4B1 | 3.345 | 1.29E-08 |
| PPP1R3C | 3.476 | 1.33E-12 |
| KRT4 | 3.632 | 1.12E-04 |
| FLG | 3.667 | 1.79E-07 |
| SPINK5 | 3.684 | 1.84E-05 |
| ENDOU | 3.703 | 9.76E-10 |
| SLURP1 | 3.789 | 4.27E-07 |
| CLIC3 | 3.798 | 9.27E-07 |
| SCEL | 3.827 | 2.80E-05 |
| CLCA4 | 3.966 | 5.96E-05 |
| TGM3 | 4.206 | 3.04E-06 |
| CRCT1 | 4.228 | 5.01E-06 |
| MAL | 4.754 | 1.02E-08 |
| CRNN | 4.917 | 1.62E-06 |
| CRISP3 | 5.781 | 4.91E-09 |

ESCC, esophageal squamous cell carcinoma. FC, fold change.

**Supplementary Table 2. The significant enriched GO BP terms and KEGG pathways.**

| | Description | Enriched genes |
|---|---|---|
| **Upregulated** | | |
| GO:0030574 | Collagen catabolic process | MMP10, COL3A1, COL1A2, COL1A1, MMP3, COL11A1, MMP13, COL5A2, MMP12, MMP1, COL10A1, MMP11 |
| GO:0030198 | Extracellular matrix organization | COL3A1, POSTN, COL5A2, LAMB3, SERPINE1, TGFBI, COL1A2, LAMC2, MFAP2, VCAN, COL1A1, COL11A1, SPP1, COL10A1 |
| GO:0001501 | Skeletal system development | PTHLH, COL3A1, COL1A2, HOXA10, SOX4, POSTN, VCAN, COL1A1, COL5A2, COL10A1, CDH11 |
| GO:0022617 | Extracellular matrix disassembly | MMP10, LAMB3, LAMC2, MMP3, MMP13, MMP12, MMP1, SPP1, MMP11 |
| GO:0007067 | Mitotic nuclear division | CDK1, CENPN, NEK2, BUB1, TPX2, AURKA, BIRC5, CDC20, PBK, CEP55, HMGA2 |
| GO:0030199 | Collagen fibril organization | COL3A1, COL1A2, COL1A1, COL11A1, COL5A2, MMP11 |
| GO:0051301 | Cell division | KIF14, CDK1, NEK2, BUB1, TPX2, AURKA, BIRC5, CDC20, UBE2C, HMGA2 |
| GO:0035987 | Endodermal cell differentiation | INHBA, LAMB3, HMGA2, COL11A1 |
| GO:0000281 | Mitotic cytokinesis | KIF23, KIF4A, CENPA, CEP55 |
| GO:0008283 | Cell proliferation | CXCL1, CDK1, DLGAP5, FSCN1, TGFBI, UCHL1, MET, BUB1, TPX2 |
| hsa04512 | ECM-receptor interaction | LAMB3, COL3A1, COL1A2, LAMC2, COL1A1, COL11A1, COL5A2, SPP1 |
| hsa05146 | Amoebiasis | LAMB3, COL3A1, COL1A2, CXCL8, LAMC2, COL1A1, COL11A1, COL5A2 |
| hsa04510 | Focal adhesion | LAMB3, COL3A1, MET, COL1A2, LAMC2, COL1A1, COL11A1, COL5A2, SPP1 |
| hsa04974 | Protein digestion and absorption | COL3A1, COL1A2, COL1A1, COL11A1, COL5A2, COL10A1 |
| hsa04151 | PI3K-Akt signaling pathway | LAMB3, COL3A1, MET, COL1A2, LAMC2, COL1A1, COL11A1, COL5A2, SPP1 |
| hsa05202 | Transcriptional misregulation in cancer | MET, CXCL8, HMGA2, MMP3, IGFBP3, PLAU |
| **Downregulated** | | |
| GO:0030216 | Keratinocyte differentiation | CRCT1, EREG, FLG, TGM1, SPRR3, TGM3, SCEL |

GO BP, Gene Ontology Biologic Process. KEGG, Kyoto Encyclopedia of Genes and Genomes.

**Supplementary Table 3. The significantly enriched GO BP and KEGG pathways for co-expression modules.**

| | Description | Enriched genes |
|---|---|---|
| **Module Blue** | | |
| GO:0007067 | Mitotic nuclear division | CDK1, CENPN, NEK2, BUB1, TPX2, AURKA, BIRC5, CDC20, PBK, CEP55, HMGA2 |
| GO:0051301 | Cell division | KIF14, CDK1, NEK2, BUB1, TPX2, AURKA, BIRC5, CDC20, UBE2C, HMGA2 |
| GO:0000281 | Mitotic cytokinesis | KIF23, KIF4A, CENPA, CEP55 |
| **Module Turquoise** | | |
| GO:0030574 | Collagen catabolic process | MMP10, COL3A1, COL1A2, COL1A1, MMP3, COL11A1, COL5A2, MMP1, COL10A1, MMP11 |
| GO:0030198 | Extracellular matrix organization | COL3A1, POSTN, COL5A2, SPINK5, SERPINE1, TGFBI, COL1A2, MFAP2, VCAN, COL1A1, COL11A1, COL10A1, SPP1 |
| GO:0001501 | Skeletal system development | COL3A1, COL1A2, SOX4, POSTN, VCAN, COL1A1, COL5A2, COL10A1, CDH11 |
| GO:0030199 | Collagen fibril organization | COL3A1, COL1A2, COL1A1, COL11A1, COL5A2, MMP11 |
| GO:0030216 | Keratinocyte differentiation | CRCT1, EREG, FLG, TGM1, SPRR3, TGM3, SCEL |
| hsa04974 | Protein digestion and absorption | COL3A1, PRSS3, COL1A2, COL1A1, COL11A1, COL5A2, COL10A1 |
| hsa05146 | Amoebiasis | COL3A1, COL1A2, SERPINB2, SERPINB1, COL1A1, COL11A1, COL5A2 |
| hsa04512 | ECM-receptor interaction | COL3A1, COL1A2, COL1A1, COL11A1, COL5A2, SPP1 |

GO BP, Gene Ontology Biologic Process. KEGG, Kyoto Encyclopedia of Genes and Genomes.