Bioinformatics analysis of esophageal squamous cell carcinoma genome microarray based on GEO database

Kaihang Deng [a], Lei Wu [b]

College of Liberal Arts and Sciences, National University of Defense Technology, Changsha, 410073, China

[a]dengkaihang16@nudt.edu.cn, [b]wulei@nudt.edu.cn

Abstract

To screen the differentially expressed genes and elucidate the mechanism of esophageal squamous cell carcinoma by analyzing the microarray data of esophageal squamous cell carcinoma(ESCC). The datasets of the related gene expression profiles were downloaded from GEO database and screened differentially expressed genes(DEGs) by using the R software package. The gene ontology(GO) functional annotations and KEGG pathway enrichments were performed using the DAVID database. The protein-protein interaction(PPI) network was established by STRING online tools and visualized by Cytoscape software. A total of 186 DEGs were screened from the datasets, among which 72 were upregulated and 114 were downregulated. A number of pathways appear to be altered in ESCC, including PI3K-Akt signaling pathway, focal adhesion, ECM-receptor interaction and so on. This study has employed bioinformatics method to screen the DEGs and pathways in the development of ESCC, and provides theoretical basis for early diagnose and accurate treatment of ESCC.

Keywords

Esophageal squamous cell carcinoma; differentially expressed gene; bioinformatics.

1. Introduction

Esophageal squamous cell carcinoma(ESCC) is one of the maglinant esophageal cancer which ranks 6th cancer-leading mortality in the world, with more than 455,000 new cases and 400,000 deaths per year worldwide[1-3]. The 5-year survival rate of patients with ESCC is less than 25%, but the patients in early stage could survive longer since they receive curative therapy before it develops into advanced cancer[4, 5]. However, the superficial ESCC is often difficult to identify due to minimal macroscopic and color changes[6]. At present, the risk factors for ESCC are not well established: though many studies indicate that gastroesophageal reflux disease(GERD), cigarette smoking, and obesity are the main risk factors for ESCC, some research focus more on oral hygiene and human papillomavirus(HPV) infection[7-12]. Since the pathogenesis is not clear yet, it is vital to study the underlying mechanisms and molecular events in ESCC so that the development of the early stage diagnostic kits can be focused on the reliable biological markers and may help to evalute the disease phase, therapy methods and so on. Thus, the gene expression microarray datasets were downloaded from the NCBI Gene Expression Omnibus (GEO) database (https://www.ncbi.nlm.nih.gov/geo/) to identify the differentially expressed genes(DEGs) using the R software, followed with the gene ontology(GO) pathway enrichment analysis on Database for Annotation, Visualization and Integrated Discovery(DAVID) online analysis (https://david.ncifcrf.gov/). Then the protein protein interaction(PPI) networks were established to excavate the hub genes and illustrate the molecular interaction in ESCC. In conclusion, a comprehensive bioinformatics analysis of identified DEGs was performed to elucidate potential pathways and biomarkers involved in the carcinogenesis and development of ESCC, which provides reliable molecular markers for future diagnosis and therapy.

2. Materials and methods

2.1 Identification of DEGs in ESCC

The expression profiling data GSE17351[13], GSE26886[14], and GSE20347[15] were downloaded from the GEO database (https://www.ncbi.nlm.nih.gov/geo/) and all based on a GeneChip Human Genome U133 Plus 2.0 Array platform (Affymetrix; Thermo Fisher Scientific, Inc., Waltham, MA, USA). The dataset of GSE17351 contains 5 ESCC samples and 5 normal samples. GSE26886 includes 9 ESCC samples, 19 normal samples and 41 other lesion samples. GSE20347 includes 17 ESCC samples and 17 normal samples. The raw data, platform and series matrix files were downloaded from the GEO database. The pretreatment of the profile raw data consists of three steps: background adjustment, normalization, summarization. The limma package in R software was used to execute the gene differential expression analysis . Samples met the cut-off criteria of the adjusted P-value (adj.P) <0.01 and |logFC| >1.5 were considered as DEGs.

2.2 GO and KEGG pathway enrichment analysis

The DAVID database provides abundant annotation tools for researchers to identify the biological function of genes[16]. GO annotations and KEGG pathway enrichment analysis were performed using a DAVID online tool on the DEGs. In this study, P < 0.05 was considered statistically significant difference.

2.3 PPI network and module analysis

The Search Tool for the Retrieval of Interacting Genes/Proteins (STRING; http://string.embl.de/) is an online software used to reveal the PPI networks between the known and predicted proteins[17]. In this study, a PPI network of identified DEGs in all three datasets was identified using the STRING database and the combined score > 0.4 was set as statistically significance. In addition, Cytoscape (http://www.cytoscape.org/) software (version 3.4.0; The Cytoscape Consortium, San Diego, CA, USA) (11,12) was used to visualize the PPI network. The DEGs which node degree > 10 were identified as the hub genes.

3. Results

3.1 Identification of DEGs in ESCC

Using the R limma package and setting p < 0.05 and |logFC| > 1.5 as cut-off criteria, 1672, 4224 and 911 DEGs were screen from GSE17351, GSE26886 and GSE20347 separately. Adopting the interactive online tools VENNY 2.1, 186 DEGs were identified, including 72 up-regulated genes and 114 down-regulated genes (Figure 1).

3.2 3.2 GO enrichment and KEGG pathway analysis

The DEGs were uploaded to the DAVID to perform the GO enrichment analysis and KEGG pathways. The P-value < 0.05 was set as the cut-off criteria. The GO enrichment analysis consists of three parts: biological process, cellular component, and molecular function. In the biological process category, the DEGs were enriched in extracellular matrix organization, extracellular matrix disassembly, collagen catabolic process, and skeletal system development. In the cellular component part, the DEGs were mainly congregated in extracellular exosome, extracellular region, extracellular matrix, proteinaceous extracellular matrix, and collagen trimer. In the molecular function group, the DEGs were mainly concentrated in serine-type endopeptidase activity (Figure 2). The pathways of DEGs included ECM-receptor interaction, amoebiasis, protein digestion and absorption, focal adhesion, arachidonic acid metabolism, PI3K-Akt signaling pathway, transcriptional misregulation in cancer, platelet activation, signaling pathways regulating pluripotency of stem cells, p53 signaling pathway and complement and coagulation cascades (Figure 3).

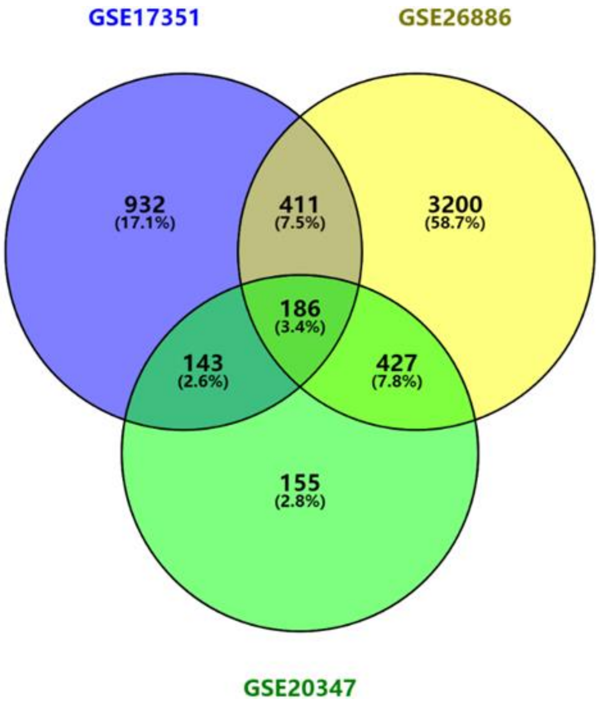

Figure 1. Identification of 186 DEGs from the three expression profiling data

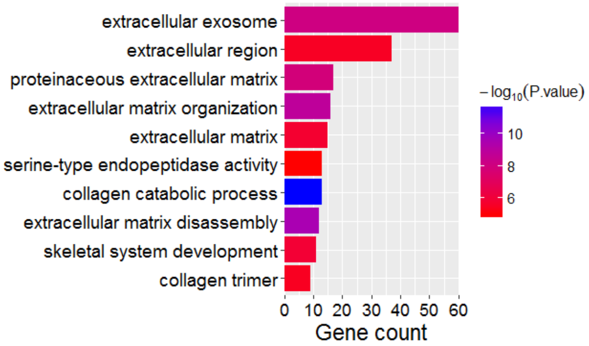

Figure 2. The GO enrichment analysis of ESCC.

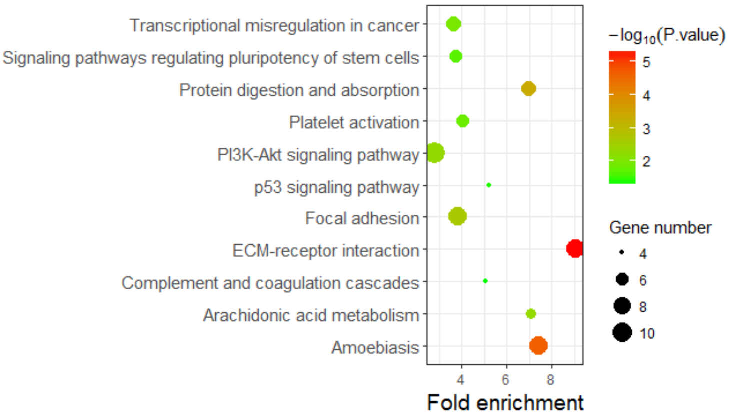

Figure 3. The KEGG pathway analysis of ESCC.

3.3 PPI network and module analysis

The DEGs were upload to the STRING database, then the interaction relationship table was imported into the Cytoscape software to visualize the PPI network. The network showed that the protein products encoded by 70 DEGs had protein protein interactions. The DEGs whose node degree > 10 were identified as the hub genes. The 7 hub genes, namely, MMP9, MMP3, TIMP1, COL1A2, COL3A1, MMP13 and SPP1 were all up-regulated genes .

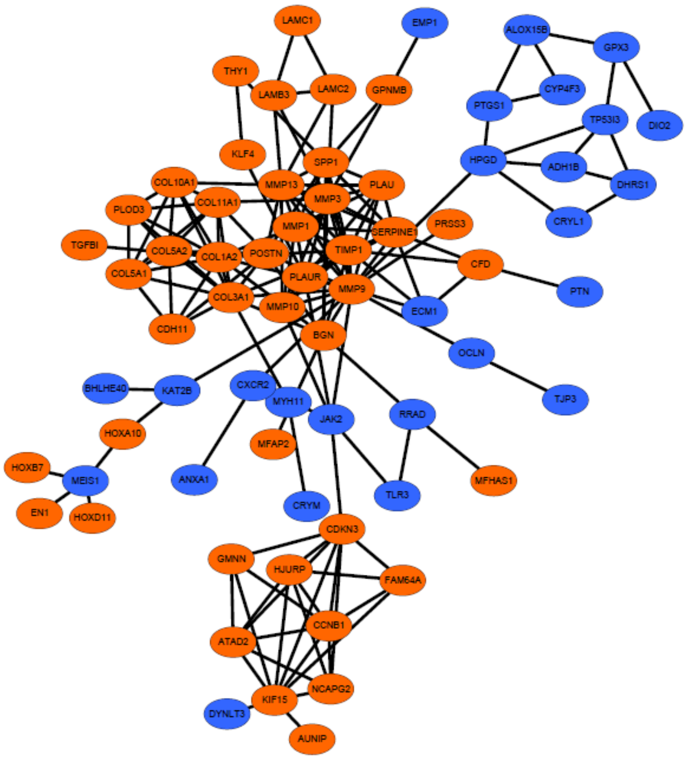

Figure 4. The PPI network of ESCC. Orange: up-regulated; Blue: down-regulated.

4. Discussion

ESCC is a malignant cancer with low 5-year survival rate in the world. Though there are some advances in the early detection, the therapies for ESCC have some limitations due to the lack of specificity of the target lesions. Therefore, an increasing number of interests focus on target therapy. Bioinformatics analyses of microarrays can screen the candidate therapeutic targets and elucidate the potential molecular mechanisms of ESCC. In the present study, GEO database was used to screen candidate expression profiling data. The expression profiling data were divided into cancer group and normal group for subsequent analysis. Based on the 3 profiles, GSE17351, GSE26886 and GSE20347, 186 DEGs were identified from the normal samples and tumor tissues, including 72 upregulated genes and 114 downregulated genes. Then function annotation and pathway analysis were obtained utilizing the GO enrichment and KEGG analysis tools. In the biological process category, the DEGs were enriched in extracellular matrix organization, extracellular matrix disassembly, collagen catabolic process, and skeletal system development. The components of extracellular matrix are structural substances and functional macromolecules. The synthesis and decomposition of extracellular matrix can affect the maintenance of cellular and tissue functions[18]. Collagen catabolic process will affect the structure and integrity of extracellular matrix scaffolds, which is very important for the development of some diseases[19]. The development of the skeletal system affects the contractility of cells and the activation of signals in the middle and lower states of the pathway[20]. The pathways of DEGs were mainly enriched in PI3K-Akt signaling pathway, ECM-receptor interaction, focal adhesion, and so on. These interactions can directly or indirectly regulate cell activity, such as adhesion, migration, differentiation, proliferation and apoptosis[21]. Using the STRING database and Cytoscape software, the PPI network was constructed and the hub genes were obtained, thus, MMP9, MMP3, TIMP1, COL1A2, COL3A1, MMP13 and SPP1. Among the hub genes, MMP9, MMP3 and MMP13 are all matrix metalloproteinases and important proteolytic enzymes, which can degrade extracellular matrix, basement membrane and stromal matrix, and play a key role in tumor invasion and metastasis[22]. TIMP1, a tissue inhibitor of metalloproteinase, can form a complex with MMP9, and inhibit the activity of MMP9 and play a negative role in tumor invasion and migration[23]. Both COL3A1 and COL1A2 are human type III collagen coding genes encoding fibrous collagen. Mutations in COL3A1 and COL1A2 are associated with a variety of tumors[24, 25]. SPP1 secrete phosphoprotein, namely osteopontin, which is classified as extracellular matrix, which promotes the adhesion and migration of cell[26].

This study has some limitations. First, the number of the expression profiles was small, which may weaken the reliability of the results. Besides, the results were not validated by qPCR or Western Blot.

5. Conclusion

Using mutiple ESCC expression profile datasets and integrated bioinformatic analysis, 186 DEGs were identified. The GO and KEGG analysis showed the significant enrichments and pathways of these DEGs. The hub genes were finally obtained as the key genes. The study improves the understanding of the mechanisms and underlying molecular events in ESCC, and the pathways and hub genes could be used as the therapeutic targets.

References

[1] RUSTGI A K, EL-SERAG H B. Esophageal carcinoma [J]. The New England journal of medicine, 2014, 371(26): 2499-509.

[2] PENNATHUR A, GIBSON M K, JOBE B A, et al. Oesophageal carcinoma [J]. Lancet (London, England), 2013, 381(9864): 400-12.

[3] FERLAY J, SOERJOMATARAM I, DIKSHIT R, et al. Cancer incidence and mortality worldwide Sources, methods and majo [J]. 2014,

[4] ENZINGER P C, MAYER R J. Esophageal cancer [J]. The New England journal of medicine, 2003, 349(23): 2241-52.

[5] PENNATHUR A, FARKAS A, KRASINSKAS A M, et al. Esophagectomy for T1 esophageal cancer: outcomes in 100 patients and implications for endoscopic therapy [J]. The Annals of thoracic surgery, 2009, 87(4): 1048-54; discussion 54-5.

[6] OHASHI S, MIYAMOTO S, KIKUCHI O, et al. Recent Advances From Basic and Clinical Studies of Esophageal Squamous Cell Carcinoma [J]. Gastroenterology, 2015, 149(7): 1700-15.

[7] SPECHLER S J. Barrett esophagus and risk of esophageal cancer: a clinical review [J]. Jama, 2013, 310(6): 627-36.

[8] LU C L, LANG H C, LUO J C, et al. Increasing trend of the incidence of esophageal squamous cell carcinoma, but not adenocarcinoma, in Taiwan [J]. Cancer Causes & Control, 2010, 21(2): 269-74.

[9] CASTRO C, BOSETTI C, MALVEZZI M, et al. Patterns and trends in esophageal cancer mortality and incidence in Europe (1980-2011) and predictions to 2015 [J]. Annals of Oncology Official Journal of the European Society for Medical Oncology, 2014, 25(1): 283-90.

[10] OTTERSTATTER M C, BRIERLEY J D, DE P, et al. Esophageal cancer in Canada: trends according to morphology and anatomical location [J]. Canadian journal of gastroenterology = Journal canadien de gastroenterologie, 2012, 26(10): 723-7.

[11] COOK M B, CHOW W, DEVESA S S. Oesophageal cancer incidence in the United States by race, sex, and histologic type, 1977‖[ndash]‖2005 [J]. 2009, 101(5): 855-9.

[12] ISLAMI F, KAMANGAR F. Helicobacter pylori and Esophageal Cancer Risk: A Meta-analysis [J]. Cancer Prevention Research, 2008, 1(5): 329.

[13] LEE J J, NATSUIZAKA M, OHASHI S, et al. Hypoxia activates the cyclooxygenase -2 -prostaglandin E synthase axis [J]. Carcinogenesis, 2010, 31(3): 427-34.

[14] WANG Q, MA C, KEMMNER W. Wdr66 is a novel marker for risk stratification and involved in epithelial-mesenchymal transition of esophageal squamous cell carcinoma [J]. BMC cancer, 2013, 13(1): 137.

[15] HU N, CLIFFORD R J, YANG H H, 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 [J]. Bmc Genomics, 2010, 11(1): 1-11.

[16] SHERMAN B T, DA W H, TAN Q, et al. DAVID Knowledgebase: a gene-centered database integrating heterogeneous gene annotation resources to facilitate high-throughput gene functional analysis [J]. Bmc Bioinformatics, 2007, 8(1): 426.

[17] VON M C, HUYNEN M, JAEGGI D, et al. STRING: a database of predicted functional associations between proteins [J]. Nucleic Acids Research, 2003, 31(1): 258.

[18] HAN B B, LI S, TONG M, et al. Fenretinide Perturbs Focal Adhesion Kinase in Premalignant and Malignant Human Oral Keratinocytes. Fenretinide's chemopreventive mechanisms include ECM interactions [J]. Cancer Prevention Research, 2015, 8(5): 419-30.

[19] BANERJEE T, MUKHERJEE S, GHOSH S, et al. Clinical significance of markers of collagen metabolism in rheumatic mitral valve disease [J]. Plos One, 2014, 9(3): e90527.

[20] ZEBDA N, DUBROVSKYI O, BIRUKOV K G. Focal adhesion kinase regulation of mechanotransduction and its impact on endothelial cell functions [J]. Microvascular Research, 2012, 83(1): 71.

[21] MARTINS F, DE SOUSA S C, DOS S E, et al. PI3K-AKT-mTOR pathway proteins are differently expressed in oral carcinogenesis [J]. Journal of Oral Pathology & Medicine, 2016, 45(10): 746-52.

[22] LI Y Y, ZHOU C X, GAO Y. Podoplanin promotes the invasion of oral squamous cell carcinoma in coordination with MT1-MMP and Rho GTPases [J]. American Journal of Cancer Research, 2014, 5(2): 514-29.

[23] IKEBE T, SHINOHARA M, TAKEUCHI H, et al. Gelatinolytic activity of matrix metalloproteinase in tumor tissues correlates with the invasiveness of oral cancer [J]. Clinical & Experimental Metastasis, 1999, 17(4): 315-22.

[24] MISAWA K, KANAZAWA T, MISAWA Y, et al. Hypermethylation of collagen α2 (I) gene (COL1A2) is an independent predictor of survival in head and neck cancer [J]. Cancer Biomarkers, 2011, 10(3-4): 135.

[25] LI J, DING Y, LI A. Identification of COL1A1 and COL1A2 as candidate prognostic factors in gastric cancer [J]. World Journal of Surgical Oncology, 2016, 14(1): 297.

[26] STANDAL T, BORSET M, SUNDAN A. Role of osteopontin in adhesion, migration, cell survival and bone remodeling [J]. Experimental Oncology, 2004, 26(3): 179.