Integrated TCGA and GEO analysis showed that SMAD7 is an independent prognostic factor for lung adenocarcinoma.

Zhou-Tong Dai
Wuhan University of Science and Technology
Jun Wang
Wuhan University of Science and Technology
Yuan Xiang
Wuhan University of Science and Technology
Jia Peng Li
Wuhan University of Science and Technology
Hui-Min Zhang
Wuhan University of Science and Technology
Tong Cun Zhang (zhangtongcun@wust.edu.cn)
Wuhan University of Science and Technology
Xianghua Liao (xinghualiao@hotmail.com)
Wuhan University of Science and Technology

Research article

Keywords: SMAD7, SMAD9, lung adenocarcinoma, prognostic value

Posted Date: April 1st, 2020

DOI: https://doi.org/10.21203/rs.3.rs-20228/v1

Version of Record: A version of this preprint was published at Medicine on October 30th, 2020. See the published version at https://doi.org/10.1097/MD.0000000000022861.

Abstract

The lack of effective markers leads to missed optimal treatment times, resulting in poorer prognosis In most cancers. SMAD family members are important cytokines in the transforming growth factor-beta (TGF-β) family. They jointly regulate the processes of cell growth, differentiation, and apoptosis. However, the expression of SMAD family genes in pan-cancers and their impact on prognosis have not been elucidated. Perl software and R software were used to perform expression analysis and survival curve analysis on the data collected by TCGA, GTEx and GEO, and the potential regulatory pathways were determined through GO enrichment and KEGG enrichment analysis. It was found that SMAD7 and SMAD9 expression decreased in Lung adenocarcinoma (LUAD), and their expression was positively correlated with survival time. Additionally, SMAD7 could be used as an independent prognostic factor for LUAD. In general, SMAD7 and SMAD9 can be used as prognostic markers of LUAD, further, SMAD7 is expected to become a therapeutic target for LUAD.

Background

Transforming growth factor-beta is distributed in various systems of the human body and is an extremely important type of growth factor that regulates the process of cell differentiation and maturation. It regulates a series of states such as cell migration, growth, differentiation, and apoptosis[1]. Early studies have shown that SMAD protein can be directly activated by TGFβ-induced cell membrane receptors to form transcription complexes, which further control the transcription of target genes in the nucleus of transcription. Therefore, SMAD protein not only becomes an important part of the TGFβ signaling pathways, but also regulates cell function together[2]. In recent years, it has been found to be related to the occurrence and development of various diseases, especially multiple malignant tumors[3].

Currently, 8 types of SMAD proteins are encoded in the human genome. SMAD1, SMAD5, SMAD9 (SMAD8) belong to the receptor substrates of AMH and BMPs in the TGFβ family[4], SMAD2 and SMAD3 are receptor substrates of activin, TGFβ, and Nodal pathways. SMAD4 assists all R-SMADs. SMAD6 and SMAD7 are inhibitory SMAD proteins[5]. Recent studies have shown that SMAD1 promotes colorectal cancer cell migration[6]. Meanwhile, in lung cancer, blocking SMAD2 and SMAD4 can block the function of TGFβ[7]. SMAD3 can participate in the EMT process of cervical cancer through Long noncoding RNA OIP5-AS1. The expression level of SMAD4 is positively correlated with the survival rate of colon cancer, and the lack of SMAD4 leads to a poor prognosis[8]. Down-regulation of SMAD5 can inhibit NPC cell proliferation, invasion and migration[9]. Up-regulation of SMAD6 can promote the appreciation of liver cancer cells[10]. In addition, according to different results, how SAMD7 regulates cancer cell proliferation and migration is still controversial[11]. The expression of SMAD9 is also closely related to risk of essential hypertension. However, the SMAD family's expression and function in some cancers are still unclear, and there are fewer reports about prognosis, especially in lung cancer. At present, the use of large databases and regulatory networks has been widely accepted in biology, such as the prognostic characteristics of melanoma by transcriptome analysis[12], and the development of new prognostic models in liver cancer[13]. In tihs study, a meta-analysis of the expression and prognostic value of SMAD

family members in cancer using multiple databases. We explore the expression, prognosis, clinical features, and possible regulatory pathways of SMAD family members in patients with LUAD, and provided a theoretical basis for future studies of SMAD family members in LUAD..

Materials And Methods

Expression and survival curve of SMAD family genes in cancer.

In order to analyze the expression of SMAD family genes in cancer, survival data and expressions of clinical from GTEx, TCGA and Oncomine databases were summarized to verify the expression of each SMAD family gene in cancer. ONCOMINE database is a gene chip-based database and integrated data extraction platform. In this database, you can set the conditions for filtering and extracting data according to your own needs (http://www.oncomine.org). In this study, we set the screening conditions as: " Analysis type: cancer vs normal analysis; P-value : 0.05; THRESHOLD (FOLD CHANGE): 2; THRESHOLD (GENE RANK): Top 10%." Meanwhile, The Gene Expression Profling Interactive Analysis (GEPIA) tool was used to analyze the clinical data of the GTEx and TCGA databases (http://gepia.cancer-pku.cn) and to compare the expression differences of the SMAD family in pan-cancers with their control of normal tissues and Its survival[14]. In addition, the Oncomine database was used to analyze the differential expression of SMAD families in each lung cancer subtype.

Analysis of clinical characteristics

The expression data of all genes in LUAD and their corresponding clinical information were extracted from the TCGA database. R software with limma was used to average the repeated data of each expression. in addition, Perl software was used to summarize the information into a matrix. The Wilcox Test was used on a single gene to examine the relationship with clinical characteristics.

Analysis of immunohistochemistry expression

The Human protein atlas (HPA) database is a large-scale protein research project, the main purpose of which is to map the positions of proteins encoded by expressed genes in human tissues and cells (https://www.proteinatlas.org/)[15]. In this study, SMAD7 and SMAD9 were expressed in LUAD as comparison conditions. The figures were obtained from the HPA database, and the results were shown as typical images.

Analysis of differential gene

Log (x + 1) processing was performed on data of the extracted TCGA clinical, and the samples were divided into high expression groups and low expression groups according to the median expression of the SMAD7 and SMAD9. R software with limma, pheatmap, and ggplot2 packages was used to filter the original data of the target gene in LUAD and normalize and screen out the differential genes between the two groups. DEGs were displayed using volcanic plot and heatmaps. Differential gene screening criteria | logFC | ≥2, P.adjust <0.05.

Analysis GO and KEGG

The cluster profiler package of R software was used to perform gene ontology (GO) functional analysis and kyoto encyclopedia of genes and genomes (KEGG) pathway analysis on the differential genes screened above, and the differences were screened using P.adjust <0.05 as the threshold[16-18]. P.adjust <0.05 is the main enrichment function and pathway of screening differential genes for conditional threshold.

Analysis of protein interaction network analysis

The selected differential genes were introduced into the Search Tool for the Retrieval of Interacting Genes (STRING). It is an online analysis website for protein-protein interaction (PPI) (https://string-db.org/)[19]. The results were imported into Cytoscape software[20], and key protein expression modules and key node genes were screened.

Analysis of GEO

The download data of GSE43767 from the database of GEO chips in NCBI (http://www.ncbi.nlm.nih.gov/geo/)[21, 22], which includes 113 samples and 29 samples of therapeutic or spontaneous abortion, 15 normal samples and 69 LUAD samples. R software with the limma and beeswarm packages were used to process the obtained data and draw difference expression heatmaps and volcano plot .

Analysis of gene set enrichment

Analysis of TCGA clinical data was used by GSEA with version of 4.0.3[23, 24]. According to the expression levels of SMAD7 and SMAD9, they were divided into two groups: high expression group and low expression group. The effect of their expression level on the gene set of various biological pathways was analyzed by GSEA. The gene set obtained from the MsigDB database of the GSEA website was used as the reference gene set, and the p value was calculated 1000 times per analysis cycle according to the weighted method.

Analysis of independent prognostic factor

R software with survival and survminer packages were used to analyze TCGA clinical clinical data. Both univariate analysis and multivariate analysis were COX proportional hazard regression models.

Result

Expression of SMAD protein family in pancreatic cancer

The expression of SMAD protein family members in human cancers at the mRNA level was analyzed by using the Oncomine online database. Analysis of expression differences between cancer and normal tissues according to the selected criteria, the results showed that there were 442, 458, 453, 459, 459, 448,

456, 389 independent studies in the database involving expressions from SMAD1 to SMAD9 (Figure 1). Interestingly, with the exception of a few SMAD genes that have increased expression in several specific cancers, SMAD protein family members have decreased expression in most cancers. In detail, SMAD1 expression increased in brain and cns cancer and lymphoma, the expression of SMAD5 increased in brain and cns cancer, colorectal cancer and kidney cancer, SMAD6 expression increased in esophageal cancer, and SMAD9 expression increased in brain and cns cancer., however, in the other cancer data, as shown in Table 1, except for testicular cancer, most members of the SMAD protein family have decreased expression, and there is no significant expression difference in other types of cancer.

In order to further determine the expression difference of SMAD protein family between cancer and normal tissues, the TCGA and GTEx database were used to jointly analyze the expression difference of SMAD protein family in 29 cancers, and a heatmap was drawn (Figure 2). The red box shows that the expression difference is statistically significant. And each gene is specifically expressed differently in cancer (Figures S1-S8). By combining the results of the Oncomine database, and using t test. Compared with normal tissues, it was found that the expressions of SMAD1, SMAD4, SMAD5, and SMAD7 were significantly different in Brain Lower Grade Glioma. In breast invasive carcinoma, SMAD9 expression was significantly different. There are significant differences in the expression of SMAD1 in Acute Myeloid LeukemiaL, significant differences in the expression of SAMD6, SMAD7 and SMAD9 in LUAD, and the expression of SMAD1 and SMAD7 in Lymphoid Neoplasm Diffuse Large B-cell Lymphoma There are significant differences in the expression of SMAD6 in Prostate adenocarcinoma, and significant differences in the expression of SMAD1 and SMAD7 in Testicular Germ Cell Tumors.

Prognostic analysis of SMAD protein family

To determine the prognostic values of the genes selected, Kaplan-Meier survival analysis was conducted on the genes selected above based on the clinical information in the TCGA database. In LUAD, SMAD6 (logrank p = 0.65, p (hr) = 0.66) cannot show an obvious correlation with Overall Survival (Figure 3A). Similarly, in all other cancers that have been analyzed, the differential genes for other SMAD protein family also showed the same negative results as SMAD6. However, in LUAD, both SMAD7 (logrank p = 0.0099, p (hr) = 0.01) and SMAD9 (logrank p = 0.0017, p (hr) = 0.0019) shown inFigure 3B.C showed positive results. The prognosis of SMAD7 and SMAD9 high expression groups were significantly better than that of low expression groups.

Clinical features of SMAD7 and SMAD9

In order to evaluate the clinical characteristics of SMAD7 and SMAD9, we extracted the expression data of SMAD7 and SMAD9 in TCGA in different types of lung cancer. they showed the same results as that from the combined analysis of TCGA and GTEx (Figure 4A). In LUAD, the expressions of SMAD7 (p = 1.76E-12) and SMAD9 (p = 1.64E-12) were reduced compared to normal tissues. However, as shown in Figure 4B.C.D, after analyzing their stage, gender, age and expression, it was found that the expression of SMAD7 has nothing to do with the stage, gender, and age. In SMAD9 (as shown in Figure 4E.F.G), although there are differences between Stage1 and Stage3 (p = 1.07E-2), there is no continuous

difference, thus, the expression of SMAD9 is independent of the stages. on the other hand, SMAD9 expression was slightly higher in women than that in men (p = 2.07E-2). In addition, the expression of SMAD9 is higher in young patients, but it is worth noting due to the insufficient sample size of young patients (n = 12).

Immunohistochemical image validation screening results

In order to verify the different expression of the SMAD7 and SMAD9 in LUAD, we extracted relevant IHC images from the Human Protein Atlas. The results showed that in normal tissues, the expression intensity of SMAD7 was mainly strong and that of SMAD9 was median (Figure 5A,C), while in LUAD, the expressions of SMAD7 and SMAD9 were both reduced (Figure 5B.D) .

Differentially expressed genes (DEGs) in TCGA

The data of LUAD in TCGA were divided into two groups of high expression and low expression according to the target gene median, and the DEG was used to screen the gene expression data between the two groups with limma in R software. According to the grouping result, a total of 12 DEGs of SMAD7 and 57 DEG24 of SMAD9 were identified from the TCGA database, (Figure 6.A.B).

Gene Ontology (GO)， Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analysis of differential genes

Cluster Profiler, org.Hs.eg.db, richplot and ggplot2 packages in R software were employed to analyze the functions of DEGs in LUAD. The results show that in GO enrichment, SMAD7 mainly participates in function of the regulation of blood vessel. SMAD9 is mainly involved in function of zymogen activation (Figure 7A). In the KEGG enrichment, SMAD7 is mainly involved in functions such as Protein digestion and absorption, and SMAD9 is mainly involved in functions such as the NOD-like receptor signaling pathway (Figure 7B).

Gene set enrichment analysis (GSEA) of the TCGA

Although DFGs have been used for GO and KEGG enrichment, it only screened for differential expressions and did not involve the degree and direction of differential gene expressions. Therefore, the function of SMAD7 and SMAD9 in LUAD was further analyzed by using GESA. In the GO and KEGG enrichment analysis results, SMAD7 positive mainly regulates processes such as CELLULAR RESPIRATION and RNA DEGRADATION, but SMAD7 negative mainly regulates processes such as REGULATION OF CELLULAR_RESPONSE and LEUKOCYTE TRANSENDOTHELIAL MIGRATION (Figure 8A.B). SMAD9 positive mainly regulates the processes such as CHRONIC INFLAMMATORY RESPONSE and GALACTOSE METABOLISM, while SMAD9 negative mainly regulates the processes such as LUNG ALVEOLUS DEVELOPMENT and GNRH SIGNALING PATHWAY (Figure 9A.B). These enrichment analysis results can better help us understand how SMAD7 and SMAD9 participate in the regulation of LUAD.

Network of DEGs protein-protein interactions (PPI)

The PPI network helps us to further explore the molecular mechanism of SMAD7 and SMAD9 in LUAD. the STRING network tool was used to analyze the identified DEGs. after hiding the disconnected nodes in the network for SMAD7 (Figure 10A), the PPI network of the DEGs consisted of 12 nodes and 15 edges. The top 5 of predicted functional partners are FGG, COL1A2, F2, SERPINC1 and LOX. Mainly through Platelet Aggregation, Common Pathway of Fibrin Clot Formation (FDR = 1.05E-10), Integrin cell surface interactions (FDR = 1.14E-9), Extra cellular matrix organization (FDR = 2.58E-09) and GRB2: SOS provides linkage to MAPK signaling for Integrins (FDR = 4.79E-6), which regulates the occurrence and development of LUAD. For SMAD9 (Figure 10B), the PPI network of the DEGs consisted of 54 nodes and 238 edges. The top 5 of predicted functional partners are FGB, HGF, F2, TRAF2 and OASL, which mainly involved in immune system (FDR = 1.05E-7), interferon alpha / beta signaling (FDR = 2.22E-7), cytokine signaling in immune system (FDR = 5.45E-07) and interferon signaling (FDR = 6.62E-6) , the finding shown that SMAD7 and SMAD9 is mainly regulate the occurrence and development of LUAD.

Identification of the expression of SMAD7 and SMAD9 in GSE43767

GSE43767 is a microarray study based on normal and LUAD patients. It includes data from 15 normal lung tissue samples and lung tissue from 69 LUAD patients. The samples were analyzed using limma package in R software. The results showed (Figure 11) that the expression of SMAD7 and SMAD9 was significantly reduced in cancer patients.

SMAD7 is an independent prognostic factor for LUAD

The clinical information of LUAD patients was extracted from the TCGA database, and some clinical information samples with some missing data were deleted, and the survival and survminer packages in R software were employed to analyze the data using the COX risk ratio model in with the expression of SMAD7 and SMAD9. It was shown (Table 2) that in SMAD7, both the results obtained based on univariate cox regression analysis (p = 1.42E-2) or multivariate cox regression analysis (p = 4.88E-4) are statistically significant (Figure 12A), indicating that SMAD7 can be used as an independent prognostic factor for LUAD. Unfortunately, SMAD9 cannot be used as an independent prognostic factor (p =0.086) (Figure 12B)

Discussion

At present, lung cancer is one of the first high-incidence malignancys among all tumors[25]. Histopathologically, it can be divided into non-small cell lung cancer and small cell lung cancer. Epidemiological statistics show that clinically, non-small cell lung cancer accounts for the vast majority. Non-small cell lung cancer can be divided into squamous cell carcinoma, adenocarcinoma and large cell lung cancer and other types[26]. LUAD is a primary epithelial tumor of the lung, which mostly originates from the bronchial mucosal epithelium or alveolar epithelium. In recent years, the incidence of LUAD has continued to rise, and has now become the most common type of lung cancer worldwide[27]. Unlike lung squamous cell carcinoma, which mostly manifests as central lung cancer, most LUADs occur in the peripheral parts of the lung, and there are no obvious clinical symptoms in the early stage, leading to a

poor prognosis. Timely detection and operation can effectively improve patients with LUAD Survival rate[28].

The SMAD signaling pathway is a key pathway for TGFβ transcription factor family to regulate cell proliferation, differentiation, metabolic migration, localization, and apoptosis[2]. As an important family of transcription factors, studying SMAD family genes can better allow us to understand the occurrence and metastasis of cancer, and thus develop new therapeutic approaches. Although many pathways have been reported, the pathways in some cancers have not been elucidated.

In this study, GTEx, TCGA, and GEO databases were used to analyze the expression of SMAD family in different cancers, and systematically to compare the mRNA expression differences of SMAD family genes in normal tissues and cancer tissues. At the same time, based on the screening results, the expression profile of SMAD family in LUAD was systematically revealed. These results show that the SMAD family plays an important role in the development of LUAD.

So far, there have been many studies on SMAD1 expression. In hepatocellular carcinoma, the expression of SMAD1 / SMAD5 / SMAD8 is reduced compared to normal tissues, but they are not potential biomarkers for hepatocellular carcinoma[29]. In prostate cancer, the expression of SMAD1 is increased[30]. However, there are no reports of SMAD1 on cancer prognosis. Similarly, SMAD2 is also involved in the regulation of many cancers. According to a study in 2018, overexpression of ATF4 can affect the survival rate of triple negative breast cancer patients through SMAD2[31]. Moreover, there is strong evidence that silencing SMAD2 can inhibit TGFβ function[32]. However, a previous study showed that the absence of SMAD2 can lead to reduced differentiation and increased EMT levels, leading to tumor metastasis[33]. At the same time, in patients with non-small cell lung cancer, high expression of p-SMAD2 is a predictive of poor clinical survival[34]. Like SMAD2, SMAD3 is considered a tumor suppressor. In lung cancer, the deletion of SMAD3 can inhibit the pathway of tumor growth through TGFβ, thereby promoting tumorigenicity[35]. And in lung cancer, the expression of SMAD3 increased. Although the number of samples studied in Oncomine was insufficient, it was consistent with our findings. It shows that our research has potential value[36]. Multiple reports have shown that SMAD4 is a tumor suppressor gene. For example, in pancreatic cancer, SMAD4 expression is reduced. The same is true for colorectal cancer[37, 38]. Moreover, special studies have shown that high expression of SMAD4 can prolong the survival time of patients. Although SMAD4 cannot be used as an independent prognostic factor, it can be used as an independent prognostic factor in combination with p-SMAD2[39]. For SMAD5, it has been reported that miR-145 can promote the migration and migration of esophageal cancer cells by inhibiting the expression of SMAD5[40]. Fstl1 can promote glioma growth through SMAD1 / SMAD5 / SMAD8[41]. At the same time, SMAD5 expression increases in prostate cancer, which is closely related to postoperative survival[40]. For SMAD6 gene, high expression of it can promote the development of glioma[42]. Similarly,, high expression of SMAD7 can lead to unsatisfactory prognosis for acute myeloid leukemia[43]. Unfortunately, the high expression of SMAD9 increases the EMT risk of oral squamous cell carcinoma[44]. These findings reveal that the SMAD family plays an important role in the development of cancer. In this study, SMAD7 and SMAD9 were significantly reduced in lung cancer. Moreover, the

expression of SMAD7 and SMAD9 is significantly correlated with the prognosis survival of patients. Through enrichment analysis, it was found that SMAD7 and SMAD9 mainly regulate the occurrence and development of lung cancer through the REGULATION OF CELLULAR_RESPONSE and GNRH SIGNALING PATHWAY pathways. At the same time, SMAD7 can be used as an independent prognostic factor to provide earlier detection and use of new treatments for lung cancer. Taken together, these results suggest that SMAD7 and SMAD9 may be markers and new therapeutic targets for LUAD. Correspondingly, these results need to be further verified by specific experiments.

Abbreviations

KEGG
kyoto encyclopedia of genes and genomes
GSEA
Gene Set Enrichment Analysis
GO
gene ontology
LUAD
lung adenocarcinoma

Declarations

The authors have no financial conflicts of interest.

Funding

This work was financially supported by National Natural Science Foundation of China (No. 31501149, 31770815, 31570764) and Hubei Natural Science Foundation (2017CFB537, 2019CFB529) and Educational Commission of Hubei (B2017009). Hubei Province Health and Family Planning Scientific Research Project (WJ2017M173, WJ2019M255) and the Applied Basic Research Project of Wuhan City (No. 2017060201010193) and the Science and Technology Young Training Program of the Wuhan University of Science and Technology (2016xz035, 2017xz027) and the Innovation and Entrepreneurship Fund for Graduate of Wuhan University of Science and Technology (JCX2016024, JCX2017032, JCX2017033).

Availability of supporting data

The data generated during this study are included in this article and its supplementary information files are available from the corresponding author on reasonable request.

Author contributions

X.H.L. and T.C.Z. designed research; X.H.L., Y.X., and Z.T.D. performed research., J.P.L., H.M.Z., Z.T.D., and J.W. analyzed data; and X.H.L., Z.T.D., and T.C.Z. wrote the paper. All authors read and approved the final manuscript.

Author's information

Xing-Hua Liao, Yuan Xiang, Jun Wang, Jia Peng Li, Hui-Min Zhang, Feng Huang, Han-Han Li, Zhou-Tong Dai, are from Institute of Biology and Medicine, Wuhan University of Science and Technology, P.R.China.

Tong-Cun Zhang is from Key Laboratory of Industrial Fermentation Microbiology, Ministry of Education and Tianjin, College of Biotechnology, Tianjin University of Science and Technology, P.R.China.

Consent for publication

All authors have read this manuscript and approved for the submission.

Competing interests

The authors declare that they have no competing interests.

References

1. Wan M, Li C, Zhen G, Jiao K, He W, Jia X, Wang W, Shi C, Xing Q, Chen YF *et al*: **Injury-activated transforming growth factor beta controls mobilization of mesenchymal stem cells for tissue remodeling**. *Stem Cells* 2012, **30**(11):2498-2511.
2. Schmierer B, Hill CS: **TGFbeta-SMAD signal transduction: molecular specificity and functional flexibility**. *Nat Rev Mol Cell Biol* 2007, **8**(12):970-982.
3. Pangas SA: **Bone morphogenetic protein signaling transcription factor (SMAD) function in granulosa cells**. *Mol Cell Endocrinol* 2012, **356**(1-2):40-47.
4. Wiegman EM, Blaese MA, Loeffler H, Coppes RP, Rodemann HP: **TGFbeta-1 dependent fast stimulation of ATM and p53 phosphorylation following exposure to ionizing radiation does not involve TGFbeta-receptor I signalling**. *Radiother Oncol* 2007, **83**(3):289-295.
5. Koller H, Hitzl W, Acosta F, Tauber M, Zenner J, Resch H, Yukawa Y, Meier O, Schmidt R, Mayer M: **In vitro study of accuracy of cervical pedicle screw insertion using an electronic conductivity device (ATPS part III)**. *Eur Spine J* 2009, **18**(9):1300-1313.
6. Yang D, Hou T, Li L, Chu Y, Zhou F, Xu Y, Hou X, Song H, Zhu K, Hou Z *et al*: **Smad1 promotes colorectal cancer cell migration through Ajuba transactivation**. *Oncotarget* 2017, **8**(66):110415-110425.
7. Chae DK, Ban E, Yoo YS, Kim EE, Baik JH, Song EJ: **MIR-27a regulates the TGF-beta signaling pathway by targeting SMAD2 and SMAD4 in lung cancer**. *Mol Carcinog* 2017, **56**(8):1992-1998.

8. Isaksson-Mettavainio M, Palmqvist R, Dahlin AM, Van Guelpen B, Rutegard J, Oberg A, Henriksson ML: **High SMAD4 levels appear in microsatellite instability and hypermethylated colon cancers, and indicate a better prognosis**. *Int J Cancer* 2012, **131**(4):779-788.
9. Li S, Zhao B, Zhao H, Shang C, Zhang M, Xiong X, Pu J, Kuang B, Deng G: **Silencing of Long Non-coding RNA SMAD5-AS1 Reverses Epithelial Mesenchymal Transition in Nasopharyngeal Carcinoma via microRNA-195-Dependent Inhibition of SMAD5**. *Front Oncol* 2019, **9**:1246.
10. Chen Z, Lu X, Jia D, Jing Y, Chen D, Wang Q, Zhao F, Li J, Yao M, Cong W *et al*: **Hepatic SMARCA4 predicts HCC recurrence and promotes tumour cell proliferation by regulating SMAD6 expression**. *Cell Death Dis* 2018, **9**(2):59.
11. Stolfi C, Marafini I, De Simone V, Pallone F, Monteleone G: **The dual role of Smad7 in the control of cancer growth and metastasis**. *Int J Mol Sci* 2013, **14**(12):23774-23790.
12. Thakur R, Laye JP, Lauss M, Diaz JMS, O'Shea SJ, Pozniak J, Filia A, Harland M, Gascoyne J, Randerson-Moor JA *et al*: **Transcriptomic Analysis Reveals Prognostic Molecular Signatures of Stage I Melanoma**. *Clin Cancer Res* 2019, **25**(24):7424-7435.
13. Li G, Xu W, Zhang L, Liu T, Jin G, Song J, Wu J, Wang Y, Chen W, Zhang C *et al*: **Development and validation of a CIMP-associated prognostic model for hepatocellular carcinoma**. *EBioMedicine* 2019, **47**:128-141.
14. Tang Z, Li C, Kang B, Gao G, Li C, Zhang Z: **GEPIA: a web server for cancer and normal gene expression profiling and interactive analyses**. *Nucleic Acids Res* 2017, **45**(W1):W98-W102.
15. Thul PJ, Akesson L, Wiking M, Mahdessian D, Geladaki A, Ait Blal H, Alm T, Asplund A, Bjork L, Breckels LM *et al*: **A subcellular map of the human proteome**. *Science* 2017, **356**(6340).
16. Yu G, He QY: **ReactomePA: an R/Bioconductor package for reactome pathway analysis and visualization**. *Mol Biosyst* 2016, **12**(2):477-479.
17. Yu G, Wang LG, Han Y, He QY: **clusterProfiler: an R package for comparing biological themes among gene clusters**. *OMICS* 2012, **16**(5):284-287.
18. Yu G, Wang LG, Yan GR, He QY: **DOSE: an R/Bioconductor package for disease ontology semantic and enrichment analysis**. *Bioinformatics* 2015, **31**(4):608-609.
19. 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* 2015, **43**(Database issue):D447-452.
20. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T: **Cytoscape: a software environment for integrated models of biomolecular interaction networks**. *Genome Res* 2003, **13**(11):2498-2504.
21. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Holko M *et al*: **NCBI GEO: archive for functional genomics data sets--update**. *Nucleic Acids Res* 2013, **41**(Database issue):D991-995.
22. Feng L, Wang J, Cao B, Zhang Y, Wu B, Di X, Jiang W, An N, Lu D, Gao S *et al*: **Gene expression profiling in human lung development: an abundant resource for lung adenocarcinoma prognosis**.

PLoS One 2014, **9**(8):e105639.

23. 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 U S A* 2005, **102**(43):15545-15550.
24. Mootha VK, Lindgren CM, Eriksson KF, Subramanian A, Sihag S, Lehar J, Puigserver P, Carlsson E, Ridderstrale M, Laurila E *et al*: **PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes**. *Nat Genet* 2003, **34**(3):267-273.
25. Torre LA, Siegel RL, Jemal A: **Lung Cancer Statistics**. *Adv Exp Med Biol* 2016, **893**:1-19.
26. Klugman M, Xue X, Hosgood HD, 3rd: **Race/ethnicity and lung cancer survival in the United States: a meta-analysis**. *Cancer Causes Control* 2019, **30**(11):1231-1241.
27. Siegel RL, Miller KD, Jemal A: **Cancer Statistics, 2017**. *CA Cancer J Clin* 2017, **67**(1):7-30.
28. Relli V, Trerotola M, Guerra E, Alberti S: **Distinct lung cancer subtypes associate to distinct drivers of tumor progression**. *Oncotarget* 2018, **9**(85):35528-35540.
29. Wang L, Ding Q, Zhao L, Pan Y, Song Z, Qin Y, Yan X: **Decreased BMP-7 and p-Smad1/5/8 expression, and increased levels of gremlin in hepatocellular carcinoma**. *Oncol Lett* 2018, **16**(2):2113-2118.
30. Qu F, Zheng J, Gan W, Lian H, He H, Li W, Yuan T, Yang Y, Li X, Ji C *et al*: **MiR-199a-3p suppresses proliferation and invasion of prostate cancer cells by targeting Smad1**. *Oncotarget* 2017, **8**(32):52465-52473.
31. Gonzalez-Gonzalez A, Munoz-Muela E, Marchal JA, Cara FE, Molina MP, Cruz-Lozano M, Jimenez G, Verma A, Ramirez A, Qian W *et al*: **Activating Transcription Factor 4 Modulates TGFbeta-Induced Aggressiveness in Triple-Negative Breast Cancer via SMAD2/3/4 and mTORC2 Signaling**. *Clin Cancer Res* 2018, **24**(22):5697-5709.
32. Yang J, Wahdan-Alaswad R, Danielpour D: **Critical role of Smad2 in tumor suppression and transforming growth factor-beta-induced apoptosis of prostate epithelial cells**. *Cancer Res* 2009, **69**(6):2185-2190.
33. Hoot KE, Lighthall J, Han G, Lu SL, Li A, Ju W, Kulesz-Martin M, Bottinger E, Wang XJ: **Keratinocyte-specific Smad2 ablation results in increased epithelial-mesenchymal transition during skin cancer formation and progression**. *J Clin Invest* 2008, **118**(8):2722-2732.
34. Chen Y, Xing P, Chen Y, Zou L, Zhang Y, Li F, Lu X: **High p-Smad2 expression in stromal fibroblasts predicts poor survival in patients with clinical stage I to IIIA non-small cell lung cancer**. *World J Surg Oncol* 2014, **12**:328.
35. Samanta D, Gonzalez AL, Nagathihalli N, Ye F, Carbone DP, Datta PK: **Smoking attenuates transforming growth factor-beta-mediated tumor suppression function through downregulation of Smad3 in lung cancer**. *Cancer Prev Res (Phila)* 2012, **5**(3):453-463.
36. Qian Z, Zhang Q, Hu Y, Zhang T, Li J, Liu Z, Zheng H, Gao Y, Jia W, Hu A *et al*: **Investigating the mechanism by which SMAD3 induces PAX6 transcription to promote the development of non-small cell lung cancer**. *Respir Res* 2018, **19**(1):262.

37. Yan P, Klingbiel D, Saridaki Z, Ceppa P, Curto M, McKee TA, Roth A, Tejpar S, Delorenzi M, Bosman FT *et al*: **Reduced Expression of SMAD4 Is Associated with Poor Survival in Colon Cancer**. *Clin Cancer Res* 2016, **22**(12):3037-3047.
38. Yamada S, Fujii T, Shimoyama Y, Kanda M, Nakayama G, Sugimoto H, Koike M, Nomoto S, Fujiwara M, Nakao A *et al*: **SMAD4 expression predicts local spread and treatment failure in resected pancreatic cancer**. *Pancreas* 2015, **44**(4):660-664.
39. Liu N, Qi D, Jiang J, Zhang J, Yu C: **Expression pattern of p-Smad2/Smad4 as a predictor of survival in invasive breast ductal carcinoma**. *Oncol Lett* 2020, **19**(3):1789-1798.
40. Zhang Q, Gan H, Song W, Chai D, Wu S: **MicroRNA-145 promotes esophageal cancer cells proliferation and metastasis by targeting SMAD5**. *Scand J Gastroenterol* 2018, **53**(7):769-776.
41. Jin X, Nie E, Zhou X, Zeng A, Yu T, Zhi T, Jiang K, Wang Y, Zhang J, You Y: **Fstl1 Promotes Glioma Growth Through the BMP4/Smad1/5/8 Signaling Pathway**. *Cell Physiol Biochem* 2017, **44**(4):1616-1628.
42. Jiao J, Zhang R, Li Z, Yin Y, Fang X, Ding X, Cai Y, Yang S, Mu H, Zong D *et al*: **Nuclear Smad6 promotes gliomagenesis by negatively regulating PIAS3-mediated STAT3 inhibition**. *Nat Commun* 2018, **9**(1):2504.
43. Zhang J, Zhang L, Cui H, Zhang X, Zhang G, Yang X, Yang S, Zhang Z, Wang J, Hu K *et al*: **High expression levels of SMAD3 and SMAD7 at diagnosis predict poor prognosis in acute myeloid leukemia patients undergoing chemotherapy**. *Cancer Gene Ther* 2019, **26**(5-6):119-127.
44. Chiba T, Ishisaki A, Kyakumoto S, Shibata T, Yamada H, Kamo M: **Transforming growth factor-beta1 suppresses bone morphogenetic protein-2-induced mesenchymal-epithelial transition in HSC-4 human oral squamous cell carcinoma cells via Smad1/5/9 pathway suppression**. *Oncol Rep* 2017, **37**(2):713-720.

Tables

Table1 Expression of SMAD famliy in other cancers in Oncomine

Gene	Cancer Type	Upregulation	Downregulation
SMAD1	Testicular Teratoma	0	3
	Yolk Sac Tumor	0	1
SMAD2	Parathyroid hyperplasia	1	0
	Non-Familial Multiple Gland Neoplasia	1	0
	Testicular Teratoma	0	3
SMAD3	Skin Carcinoma	1	0
	Adrenal Cortex Carcinoma	1	0
	Testicular Teratoma	1	0
	Vulvar Intraepithelial Neoplasia	0	1
SMAD4	Non-Familial Multiple Gland Neoplasia	1	0
	Parathyroid Gland Adenoma	1	0
	Pleural Malignant Mesothelioma	1	0
	Testicular Seminoma	0	2
	Uterine Corpus Leiomyoma	0	1
SMAD5	Teratoma, NOS	1	0
	Yolk Sac Tumor	1	0
SMAD6	Yolk Sac Tumor	1	0
	Skin Carcinoma	1	2
SMAD7	Teratoma, NOS	1	0
	Embryonal Carcinoma	1	0
	Mixed Germ Cell Tumor	1	0
	Yolk Sac Tumor	2	0
	Testicular Carcinoma	3	0
	Primitive Neuroectodermal Tumor	0	1
SMAD9	Malignant Fibrous Histiocytoma	0	1

Table2 Analysis of independent prognostic factor

	SMAD7							
Parameter	Univariate analysis				Multivariate analysis			
	HR	HR.95L	HR.95H	P	HR	HR.95L	HR.95H	pvalue
age	1.000845757	0.982401735	1.019636054	0.92901693	1.015716464	0.996187817	1.035627938	0.115404453
gender	1.00097432	0.698837323	1.433737948	0.995761563	0.832139732	0.57558904	1.203039815	0.328536239
stage	1.64465101	1.396688	1.93663649	2.42E-09	1.986915258	1.252324877	3.152402637	0.003552734
T	1.623091548	1.309819761	2.011289072	9.57E-06	1.243582527	0.972244126	1.590647307	0.082594154
M	1.681168333	0.923680619	3.059853055	0.08910352	0.387138789	0.121643708	1.232093665	0.10814002
N	1.792676516	1.464854278	2.193862653	1.47E-08	1.053602738	0.71290082	1.557129265	0.793329751
SMAD7	0.942661218	0.899203967	0.988218697	0.014201975	0.633340696	0.48995667	0.818685533	0.000487591

Supplemental Information Note

S1. Box plot of SMAD1 expression in cancer. The value is normalized by log2 (TPM + 1) for log-scale, * is the significant differences (p <0.05) after t test

S2. Box plot of SMAD2 expression in cancer. The value is normalized by log2 (TPM + 1) for log-scale, * is the significant differences ($p < 0.05$) after t test

S3. Box plot of SMAD3 expression in cancer. The value is normalized by log2 (TPM + 1) for log-scale, * is the significant differences ($p < 0.05$) after t test

S4. Box plot of SMAD4 expression in cancer. The value is normalized by log2 (TPM + 1) for log-scale, * is the significant differences ($p < 0.05$) after t test

S5. Box plot of SMAD5 expression in cancer. The value is normalized by log2 (TPM + 1) for log-scale, * is the significant differences ($p < 0.05$) after t test

S6. Box plot of SMAD6 expression in cancer. The value is normalized by log2 (TPM + 1) for log-scale, * is the significant differences ($p < 0.05$) after t test

S7. Box plot of SMAD7 expression in cancer. The value is normalized by log2 (TPM + 1) for log-scale, * is the significant differences ($p < 0.05$) after t test

S8. Box plot of SMAD8 expression in cancer. The value is normalized by log2 (TPM + 1) for log-scale, * is the significant differences ($p < 0.05$) after t test

Figures

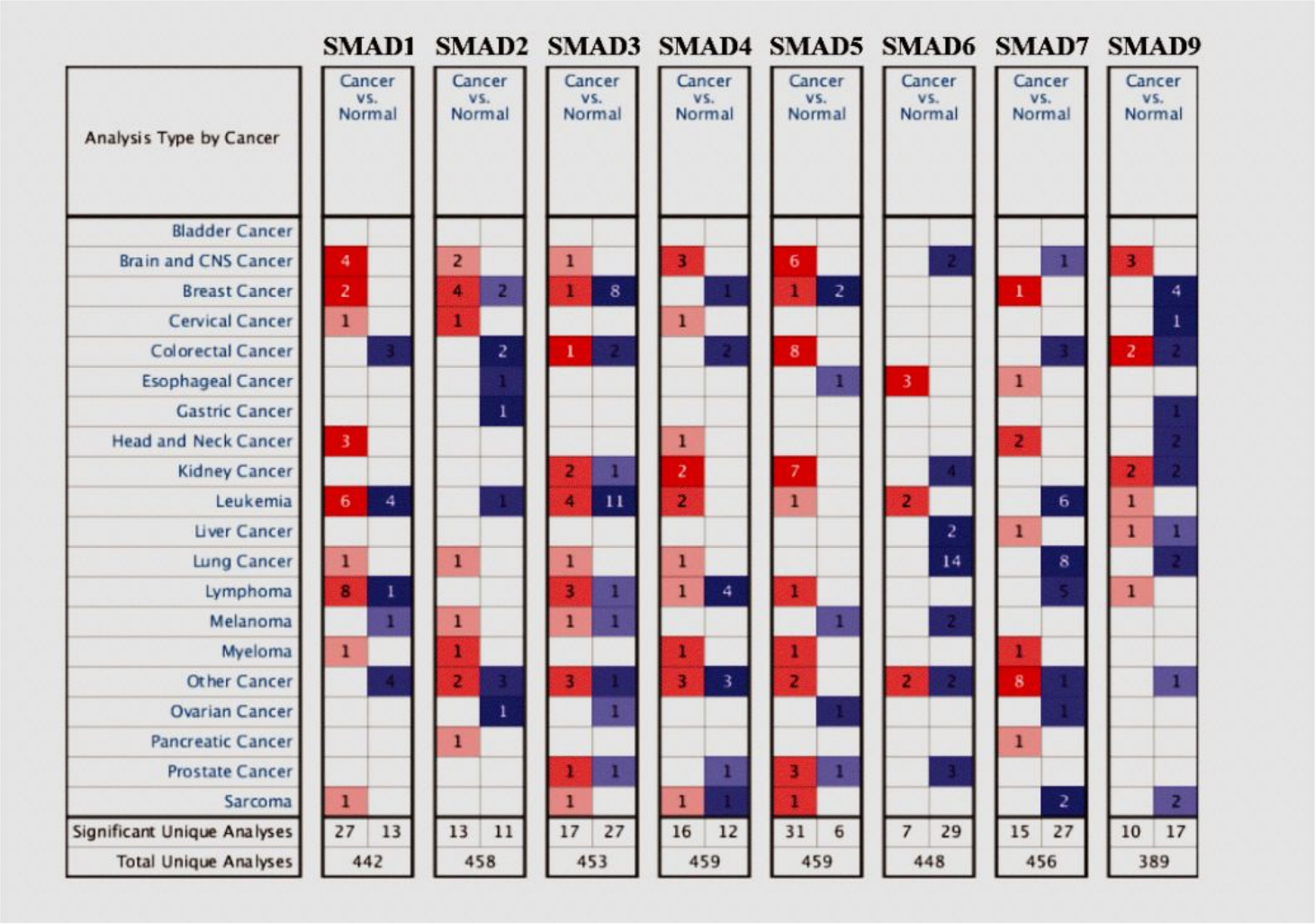

Figure 1

Analysis of the mRNA expression of SAMD family genes (cancer tissue vs normal tissue) from the Oncomine database. The Oncomine database was used to analyze. the number of statistically significant datasets from data filtered by the screening criteria. Red represents increased expression, blue represents decreased expression, and each group of underground numbers represents the total number of studies.

Figure 2

Analysis of SAMD family gene expression of GTEx and TCGA databases. Heatmaps show the expression of SMAD family members in tumors and normal tissues in all types of cancer. The value is normalized by log2 (TPM + 1) for log-scale. The red box represents significant differences after t-test, the results have. The abbreviations are the same as the tumor names in the TCGA database study.

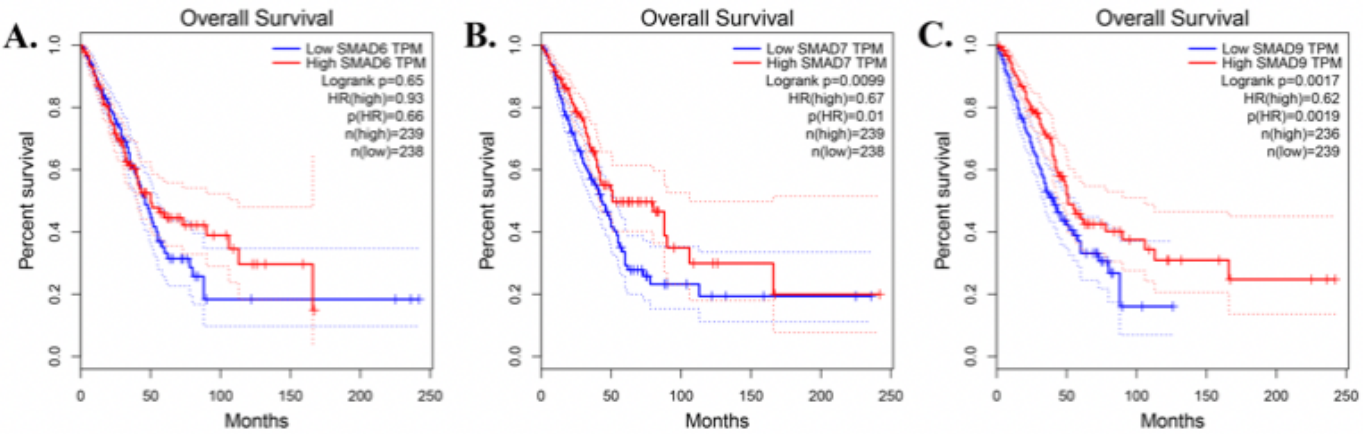

Figure 3

Prognostic value of SMAD family in LUAD. (A) Survival curve of SMAD6 in LUAD. (B) Survival curve of SMAD7 in LUAD. (C) SMAD9 survival curve in LUAD. Survival curves were drawn in median groups, and test survival curve is tested by COX and logrank.

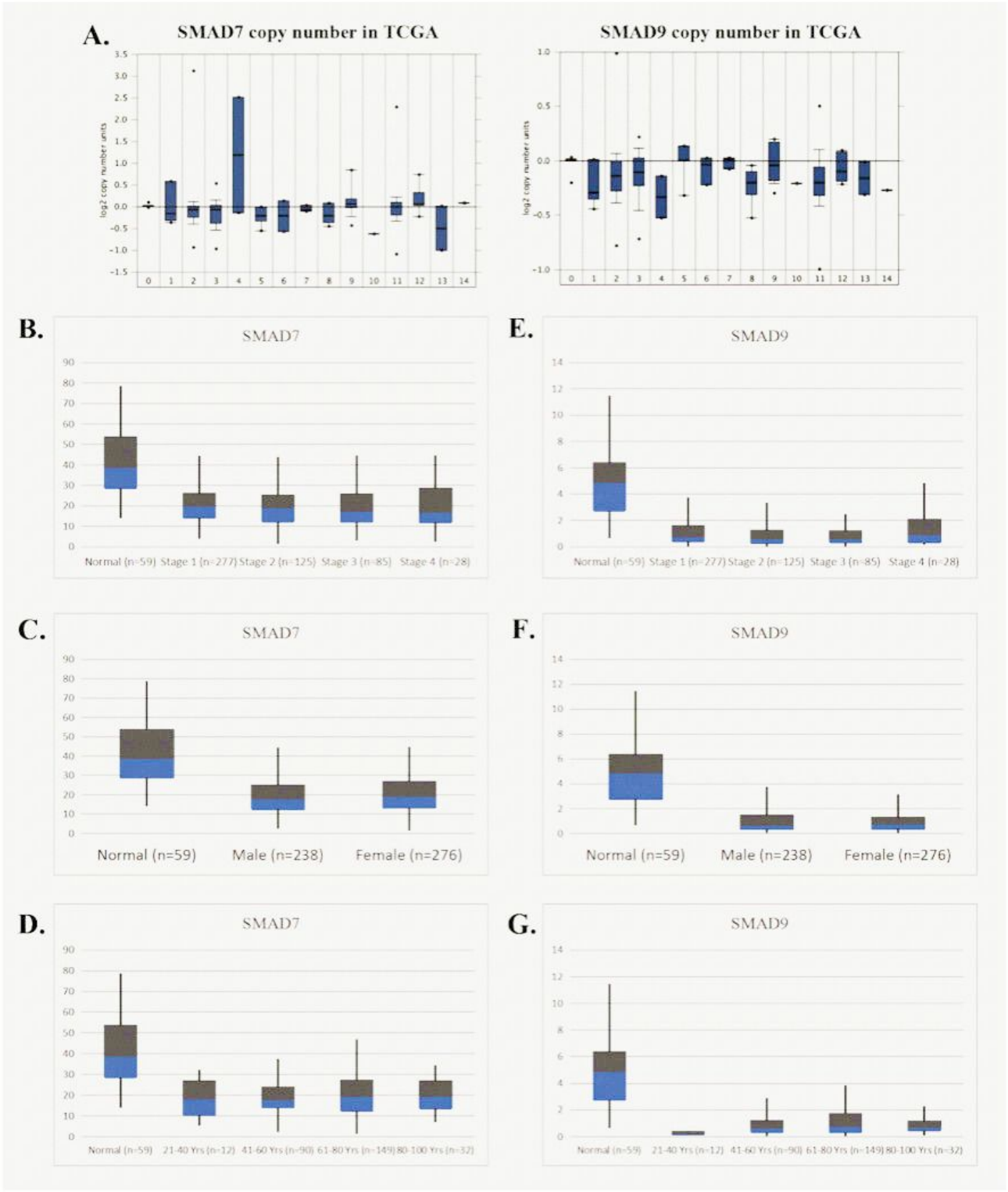

Figure 4

Clinical characteristics of SMAD7 and SMAD9.(A) Box plots indicate the mRNA expression levels of SMAD7 and SMAD9 in different types of lung cancer in the TCGA database. The values are normalized by log2. 0 = Normal (810); 1 : Acinar LUAD (6); 2: LUAD (261); 3 : LUAD, Mixed Subtype (67); 4 : Lung Clear Cell Adenocarcinoma (2); 5 : Lung Mucinous Adenocarcinoma (6); 6 : Micropapillary LUAD (3); 7: Mucinous Bronchioloalveolar Carcinoma (3); 8 :Non-Mucinous Bronchioloalveolar Carcinoma (9); 9:

Papillary LUAD (10); 10:Solid LUAD (1); 11 :Squamous Cell Lung Carcinoma (348); 12 :Squamous Cell Lung Carcinoma, Basaloid Variant (8); 13:Squamous Cell Lung Carcinoma, Papillary Variant (2); 14:Squamous Cell Lung Carcinoma, Small Cell Variant (1). (B) Expression of SMAD7 in LUAD at different stages. (C) Expression of SMAD7 in LUAD of different genders. (D) Expression of SMAD7 in LUAD at different ages. (E) Expression of SMAD9 in LUAD at different stages. (F) SMAD9 expression in LUAD of different genders. (G) SMAD9 expression in LUAD at different ages. The value is normalized by log2 (TPM + 1) for log-scale, * is the significant differences ($p < 0.05$) after t test

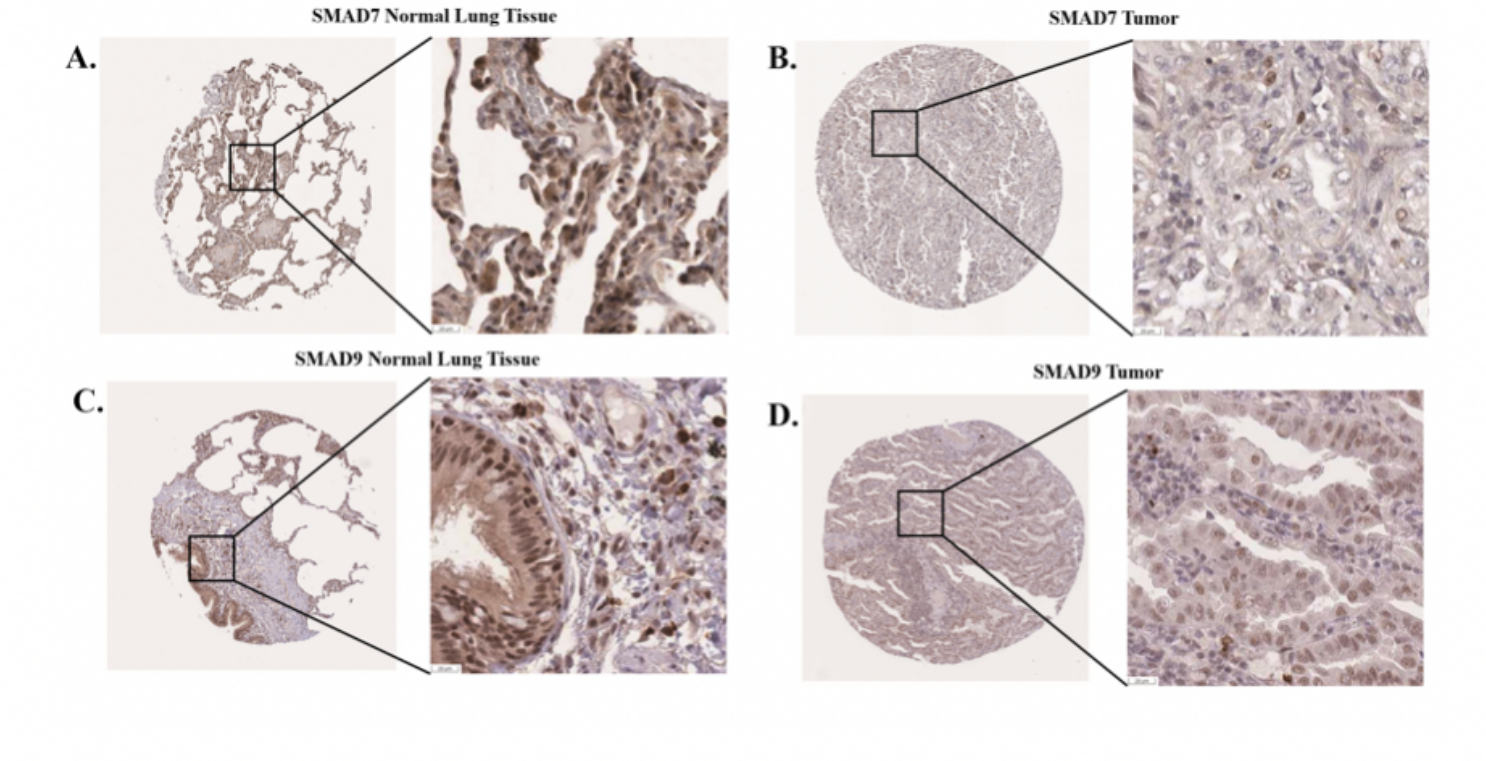

Figure 5

Typical IHC images of SMAD7 and SMAD9. (A) Expression of SMAD7 in normal lung tissue. HPA ID: HPA028897; Patient id: 3076. (B) Expression of SMAD7 in LUAD patients. HPA ID: HPA028897; Patient id: 3052. (C) Expression of SMAD9 in normal lung tissue. HPA ID: HPA031162; Patient id: 1678. (D) Expression of SMAD9 in LUAD patients. HPA ID: HPA031162; Patient id: 1847

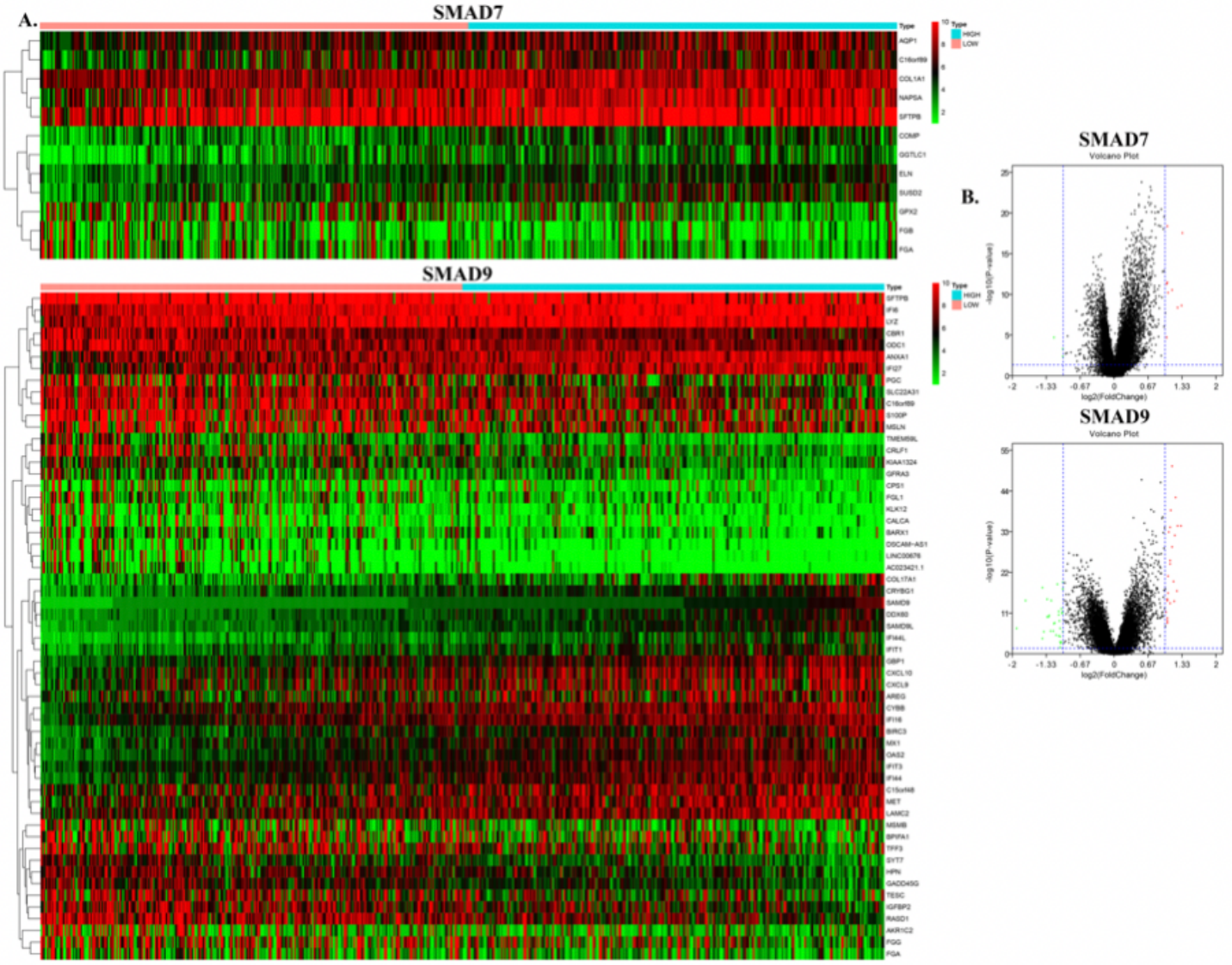

Figure 6

Heatmap and volcano plot of differential gene expression of SMAD7 and SMAD9. (A) Heatmap of DEGs screened based on clinical data from TCGA database. (B) Volcano plot of the DEGs. Red indicates genes that are up-regulated and green indicates genes that are down-regulated.

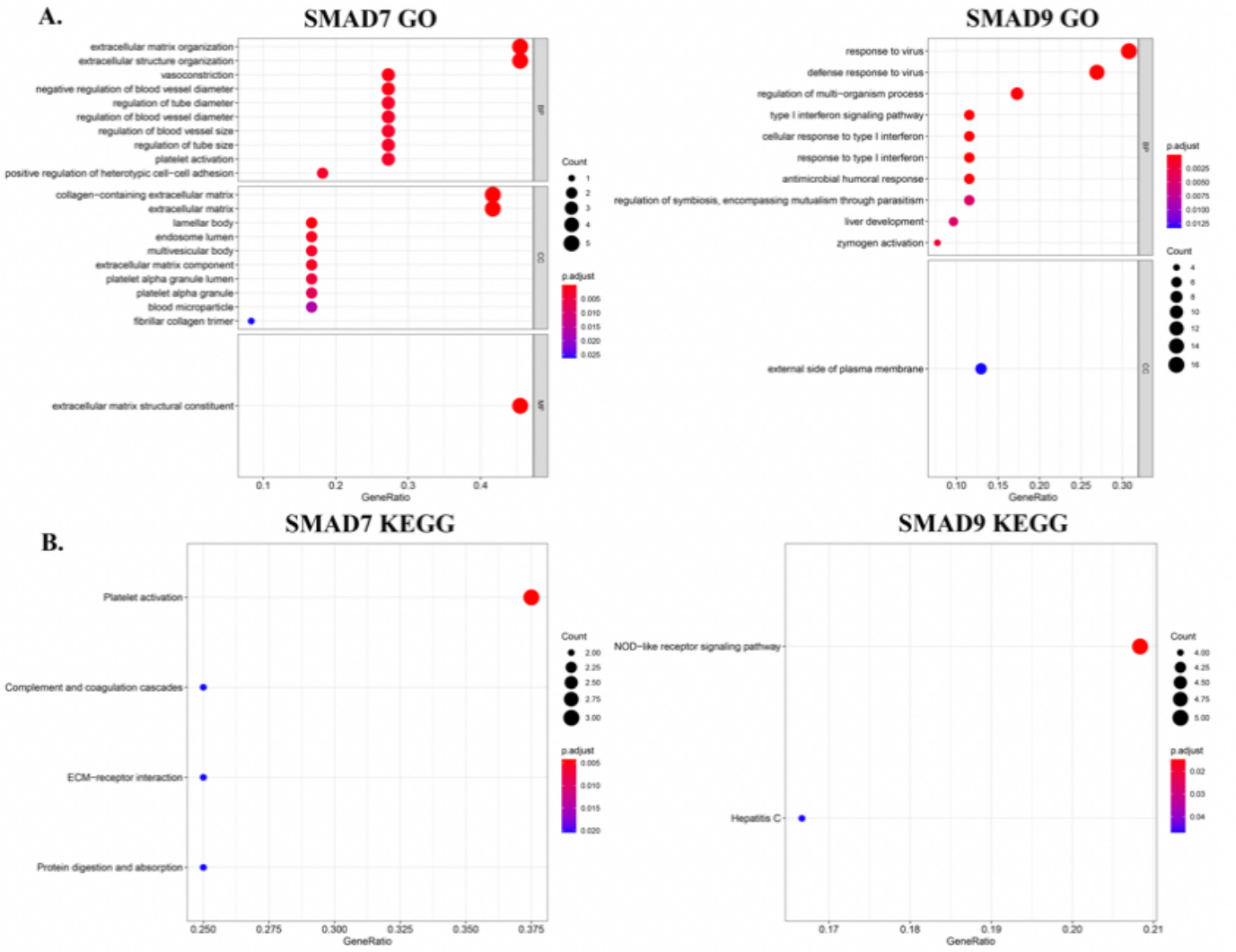

Figure 7

GO enrichment analysis and KEGG enrichment analysis of DEGs. (A) SMAD7. (B) SMAD9. The abscissa is the Rich factor. A larger value indicates a greater degree of enrichment. The ordinate is a pathway term with a higher degree of enrichment. The p value is the p value that has undergone multiple checks. The redder the color, the smaller the p value, indicating that the enrichment is more obvious. The size of the dot indicates the number of differential genes in the term, and the larger the dot, the more genes there are.

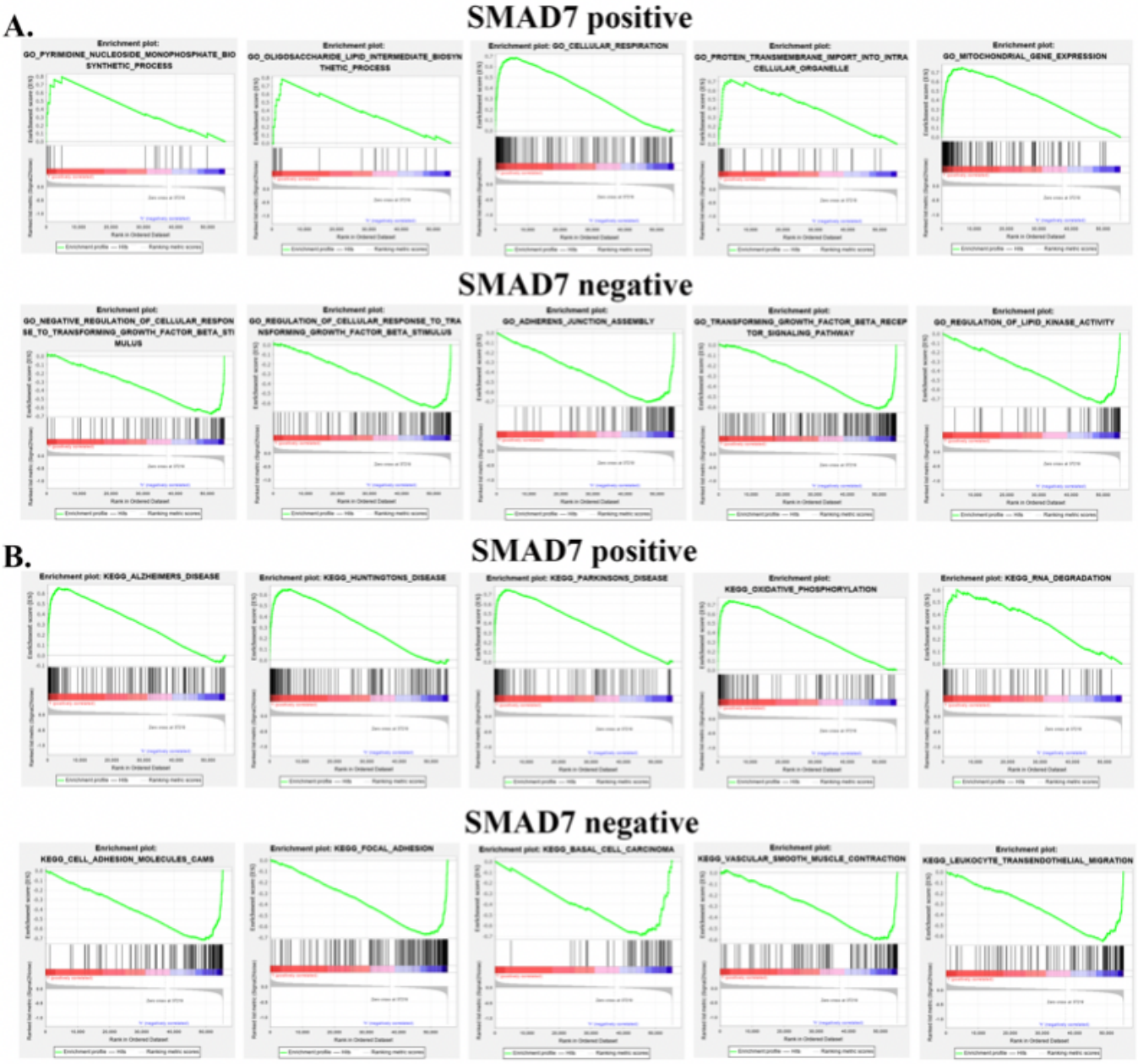

Figure 8

GO and KEGG enrichments analyses of SMAD7 by using GSEA. A. Top 5 positive enrichment scores (ES) . B. Top 5 negative ES.

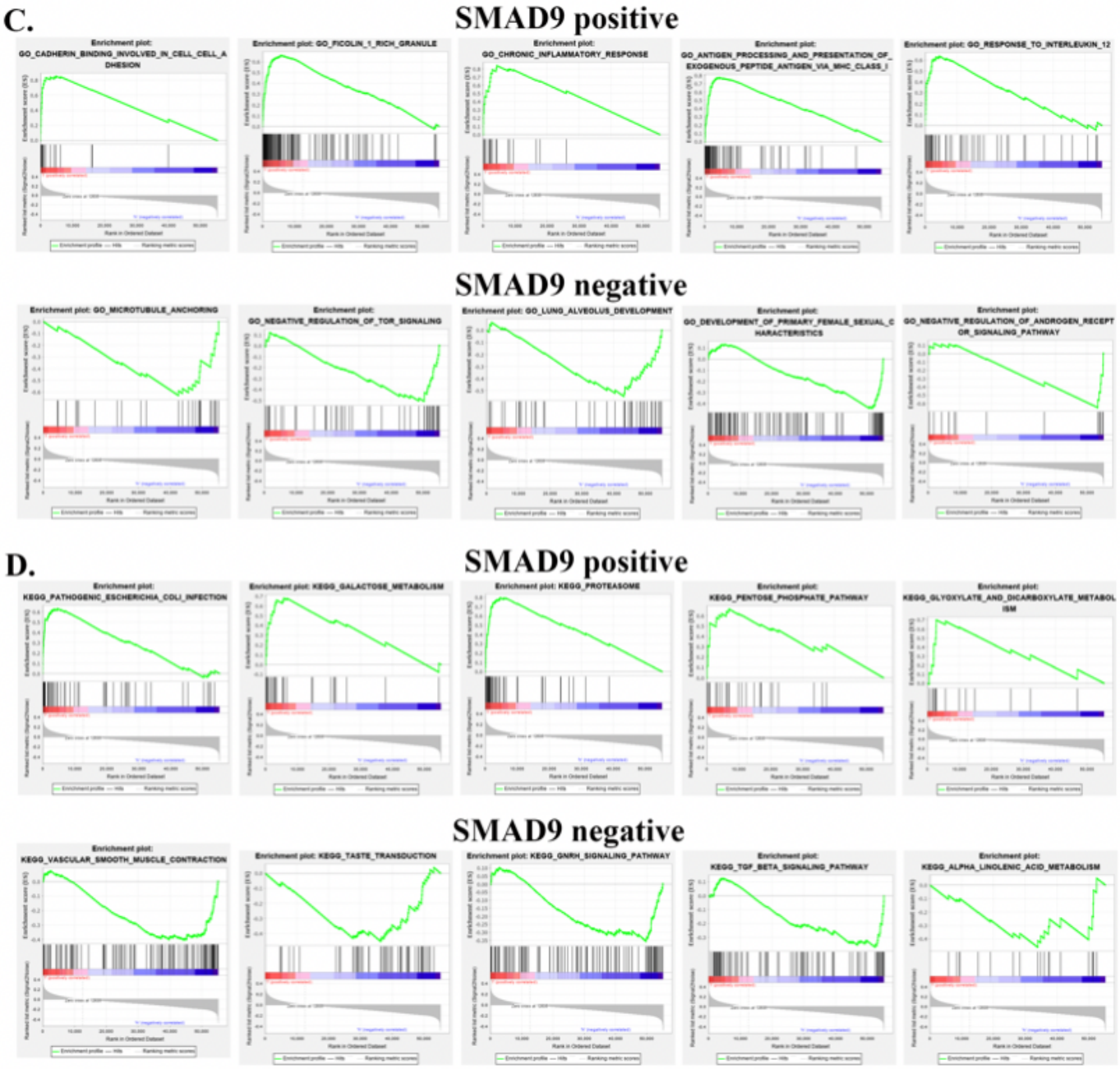

Figure 9

GO and KEGG enrichments of genes by SMAD9 analyzed using GSEA. A. Top 5 positive enrichment scores (ES) . B. Top 5 negative ES.

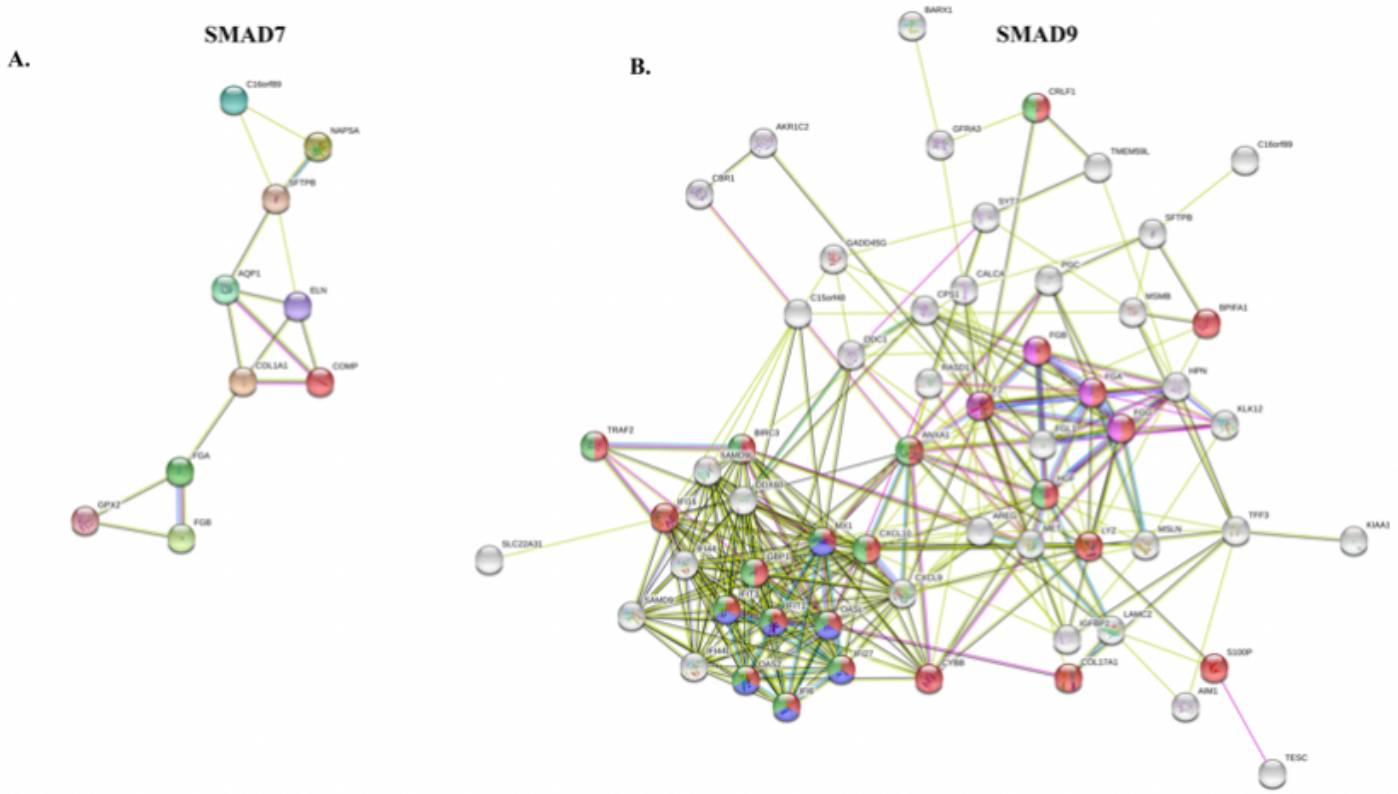

Figure 10

DEG's PPI network. A. PPI network between DEGs of SMAD7. B. PPI network between DEADs of SMAD9. All disconnected nodes are hidden.

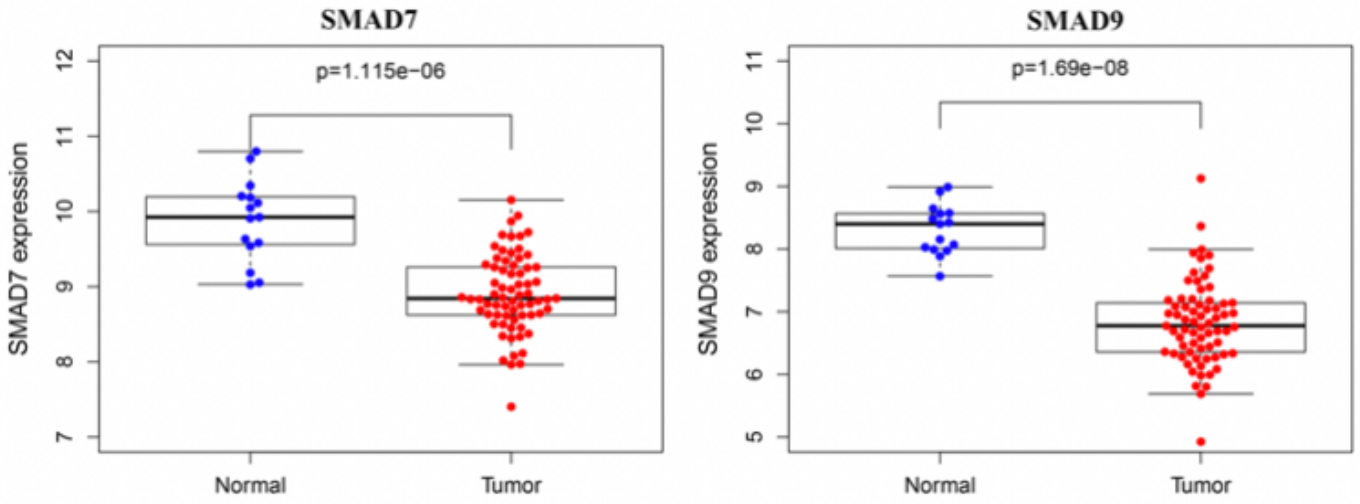

Figure 11

Expressions of SMAD7 and SMAD9 of the GSE43767 dataset in the GEO database. The ordinate is the expression value after taking log (x + 1) for the data result.

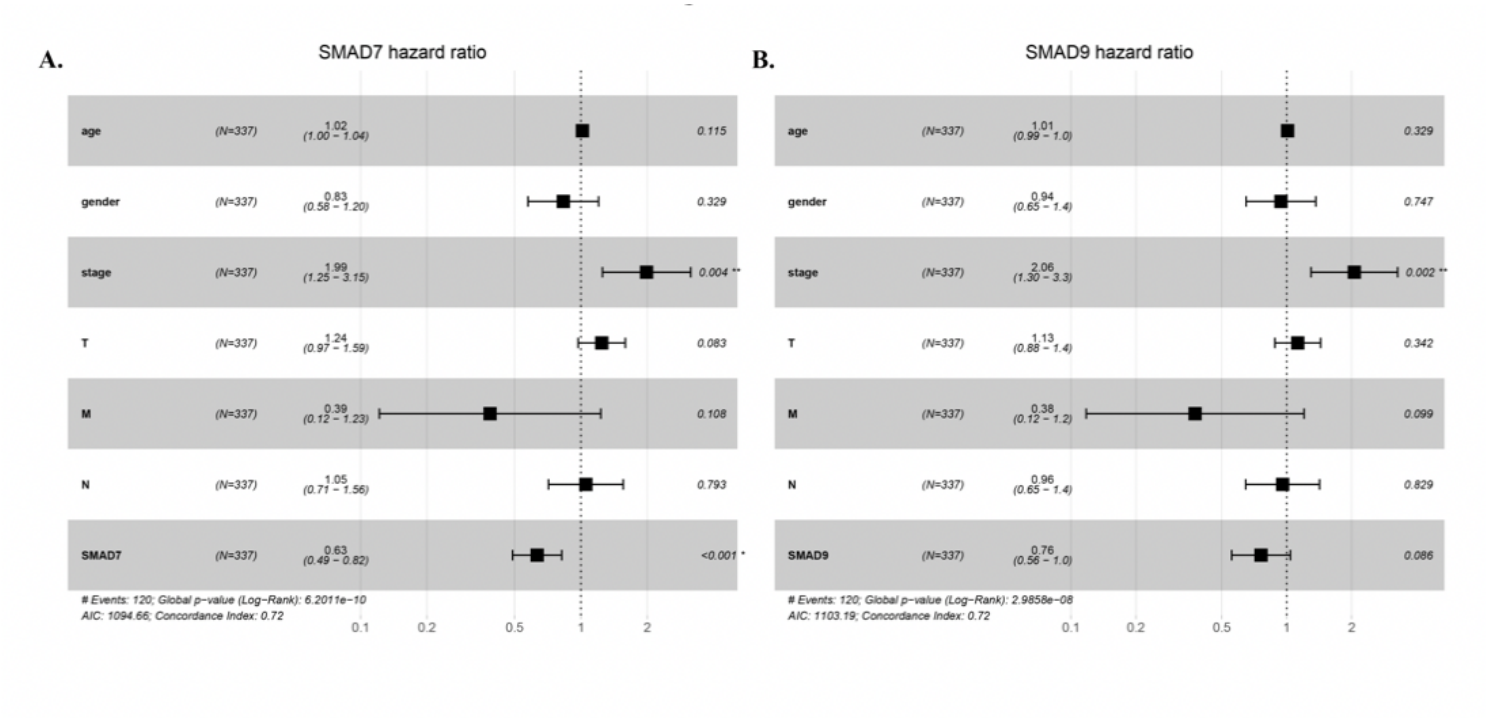

Figure 12

Multi-factor COX analysis forest map of SMAD7 and SMAD9.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementalFigures.docx