frontiers in Oncology

ORIGINAL RESEARCH
published: 22 May 2020
doi: 10.3389/fonc.2020.00681

Identification of Hub Genes Associated With Development of Head and Neck Squamous Cell Carcinoma by Integrated Bioinformatics Analysis

Chia Ying Li[1,2], *Jia-Hua Cai*[3†], *Jeffrey J. P. Tsai*[3] *and Charles C. N. Wang*[3*]

[1] *Department of Surgery, Show Chwan Memorial Hospital, Changhua, Taiwan,* [2] *Ph.D. Program in Tissue Engineering and Regenerative Medicine, National Chung Hsing University, Taichung, Taiwan,* [3] *Department of Bioinformatics and Medical Engineering, Asia University, Taichung, Taiwan*

OPEN ACCESS

Edited by:
Jorge A. R. Salvador,
University of Coimbra, Portugal

Reviewed by:
Lorenzo Lo Muzio,
University of Foggia, Italy
Sheng Li,
Wuhan University, China
Jun Liu,
Sichuan University, China

***Correspondence:**
Charles C. N. Wang
cnwang@asia.edu.tw

† *These authors have contributed equally to this work*

Specialty section:
This article was submitted to Head and Neck Cancer, a section of the journal Frontiers in Oncology

Received: *29 December 2019*
Accepted: *09 April 2020*
Published: *22 May 2020*

Citation:
Li CY, Cai J-H, Tsai JJP and Wang CCN (2020) Identification of Hub Genes Associated With Development of Head and Neck Squamous Cell Carcinoma by Integrated Bioinformatics Analysis. Front. Oncol. 10:681. doi: 10.3389/fonc.2020.00681

Improved insight into the molecular mechanisms of head and neck squamous cell carcinoma (HNSCC) is required to predict prognosis and develop a new therapeutic strategy for targeted genes. The aim of this study is to identify significant genes associated with HNSCC and to further analyze its prognostic significance. In our study, the cancer genome atlas (TCGA) HNSCC database and the gene expression profiles of GSE6631 from the Gene Expression Omnibus (GEO) were used to explore the differential co-expression genes in HNSCC compared with normal tissues. A total of 29 differential co-expression genes were screened out by Weighted Gene Co-expression Network Analysis (WGCNA) and differential gene expression analysis methods. As suggested in functional annotation analysis using the R clusterProfiler package, these genes were mainly enriched in epidermis development and differentiation (biological process), apical plasma membrane and cell-cell junction (cellular component), and enzyme inhibitor activity (molecular function). Furthermore, in a protein-protein interaction (PPI) network containing 21 nodes and 25 edges, the ten hub genes (S100A8, S100A9, IL1RN, CSTA, ANXA1, KRT4, TGM3, SCEL, PPL, and PSCA) were identified using the CytoHubba plugin of Cytoscape. The expression of the ten hub genes were all downregulated in HNSCC tissues compared with normal tissues. Based on survival analysis, the lower expression of CSTA was associated with worse overall survival (OS) in patients with HNSCC. Finally, the protein level of CSTA, which was validated by the Human Protein Atlas (HPA) database, was down-regulated consistently with mRNA levels in head and neck cancer samples. In summary, our study demonstrated that a survival-related gene is highly correlated with head and neck cancer development. Thus, CSTA may play important roles in the progression of head and neck cancer and serve as a potential biomarker for future diagnosis and treatment.

Keywords: head and neck squamous cell carcinoma, differential gene expression analysis, weighted gene co-expression network analysis, the differential co-expression genes, biomarkers

INTRODUCTION

Head and neck squamous cell carcinoma (HNSCC) is one of most common types of cancer in the world. HNSCC includes several malignancies that originate in the mouth, nasopharynx, oropharynx, hypopharynx, larynx, and neck (1). According to the published global cancer statistics report, there were more than an estimated 650,000 new cases and 330,000 deaths diagnosed in 2018 (2). Many lifestyle factors have been investigated, with tobacco use, alcohol consumption, human papillomavirus (HPV), and Epstein-Barr virus (EBV) infection being considered as the risk factors that are associated with the progression of HNSCC (3). However, HPV is currently the one most well-studied and frequently used biomarker in HNSCC (4–6). In the past several years, the treatments for managing head and neck cancer included the following: radiation therapy, surgery, and chemotherapy. Appropriate combinations of the three treatment modalities is selected according to the site of the cancer and the stage of the disease (1, 3). Although there are diverse treatments for HNSCC, patients have a limited survival advantage.

With the development of genomic technologies, bioinformatics has become increasingly popular for gene expression profiles analysis to study the molecular mechanisms of diseases and discover disease-specific biomarkers (7). One important method to understand the gene function and gene association from genome-wide expression is Weighted Gene Co-expression Network Analysis (WGCNA) (8). WGCNA can be used to detect co-expression modules of highly correlated genes and interested modules associated with clinical traits (9), providing great insight into predicting the functions of co-expression genes and finding genes that play key roles in human diseases (10–12). Furthermore, another powerful analysis within transcriptomics is differential gene expression analysis, which provides methods for studying molecular mechanisms underlying genome regulation and discovering quantitative changes in expression levels between experimental groups and control groups (13). Such gene expression differences can lead to the discovery of potential biomarkers for a particular disease. Therefore, using two approaches, the findings from WGCNA and differential gene expression analysis are combined to enhance the discriminating ability of highly related genes that are useful to serve as candidate biomarkers.

In this study, the mRNA expression data of HNSCC from the TCGA and GEO databases were analyzed by WGCNA and differential gene expression analysis to obtain differential co-expression genes. We further explored HNSCC development

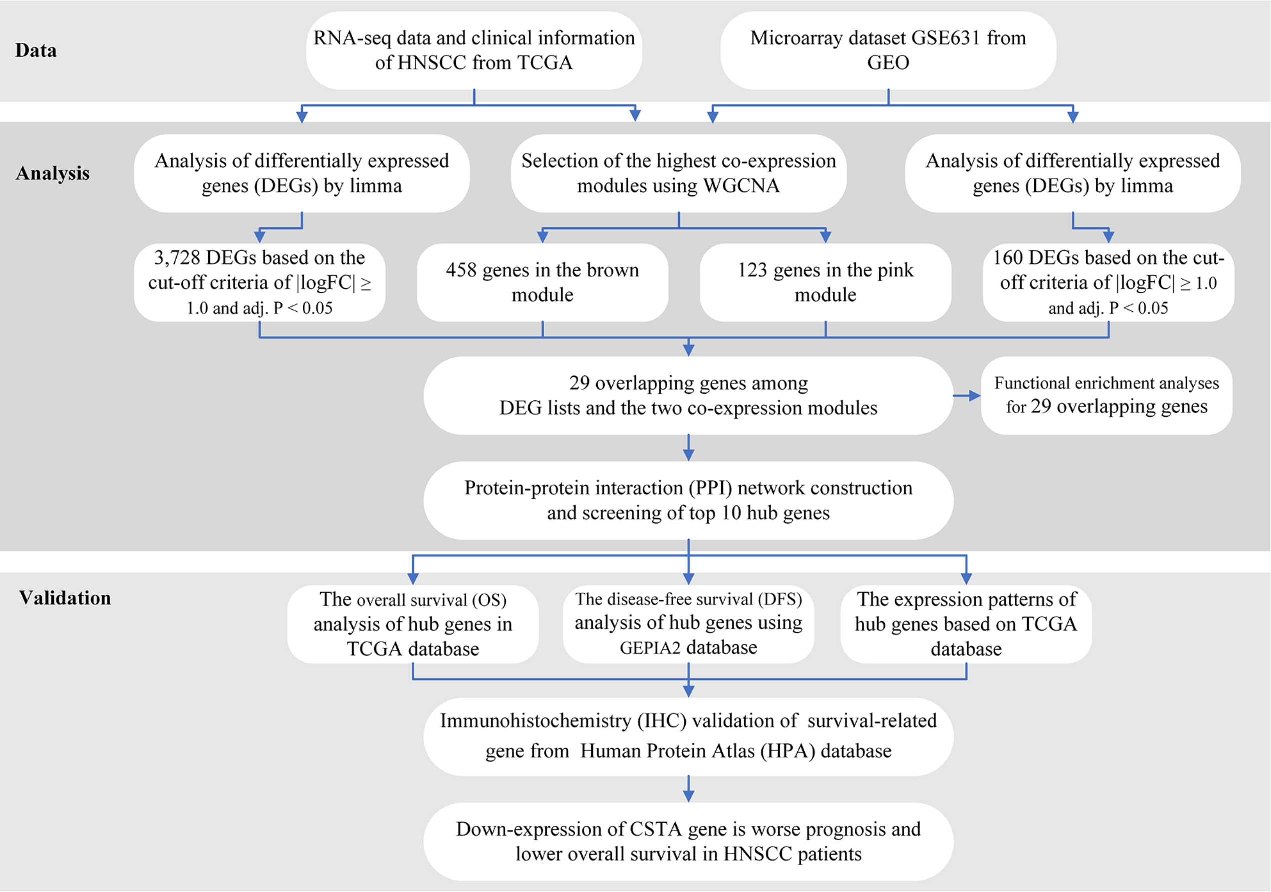
Data
RNA-seq data and clinical information of HNSCC from TCGA
Microarray dataset GSE631 from GEO
Analysis
Analysis of differentially expressed genes (DEGs) by limma
Selection of the highest co-expression modules using WGCNA
Analysis of differentially expressed genes (DEGs) by limma
3,728 DEGs based on the cut-off criteria of |logFC| ≥ 1.0 and adj. P < 0.05
458 genes in the brown module
123 genes in the pink module
160 DEGs based on the cut-off criteria of |logFC| ≥ 1.0 and adj. P < 0.05
29 overlapping genes among DEG lists and the two co-expression modules
Functional enrichment analyses for 29 overlapping genes
Protein-protein interaction (PPI) network construction and screening of top 10 hub genes
Validation
The overall survival (OS) analysis of hub genes in TCGA database
The disease-free survival (DFS) analysis of hub genes using GEPIA2 database
The expression patterns of hub genes based on TCGA database
Immunohistochemistry (IHC) validation of survival-related gene from Human Protein Atlas (HPA) database
Down-expression of CSTA gene is worse prognosis and lower overall survival in HNSCC patients

FIGURE 1 | Study design and workflow of this study.

through functional enrichment and protein-protein interaction (PPI) analysis combined with survival analysis. The study provides a potential basis to understand the cause and potential molecular events of HNSCC by analyzing differential co-expression genes for clinical diagnosis or treatment.

MATERIALS AND METHODS

The workflow of the analysis hub gene extraction curation pipeline is shown in **Figure 1**.

We elaborate on each step in the following sub-sections.

Datasets From TCGA and GEO Database

The gene expression profiles of HNSCC were downloaded from TCGA (https://portal.gdc.cancer.gov/) and GEO (https://www.ncbi.nlm.nih.gov/gds). In the TCGA database, all data on HNSCC and corresponding clinical information were freely downloaded by R package *TCGAbiolinks* (14). There were 544 NHSCC samples, including 500 head and neck cancers and 44 normal tissues, and RNAseq count data on 19,430 genes. A total of the data had been generated by using the Illumina HiSeq 2,000 platform, and were annotated to a reference transcript set of Human hg38 gene standard track. As suggested by the *edgeR* package tutorial (15), genes of low

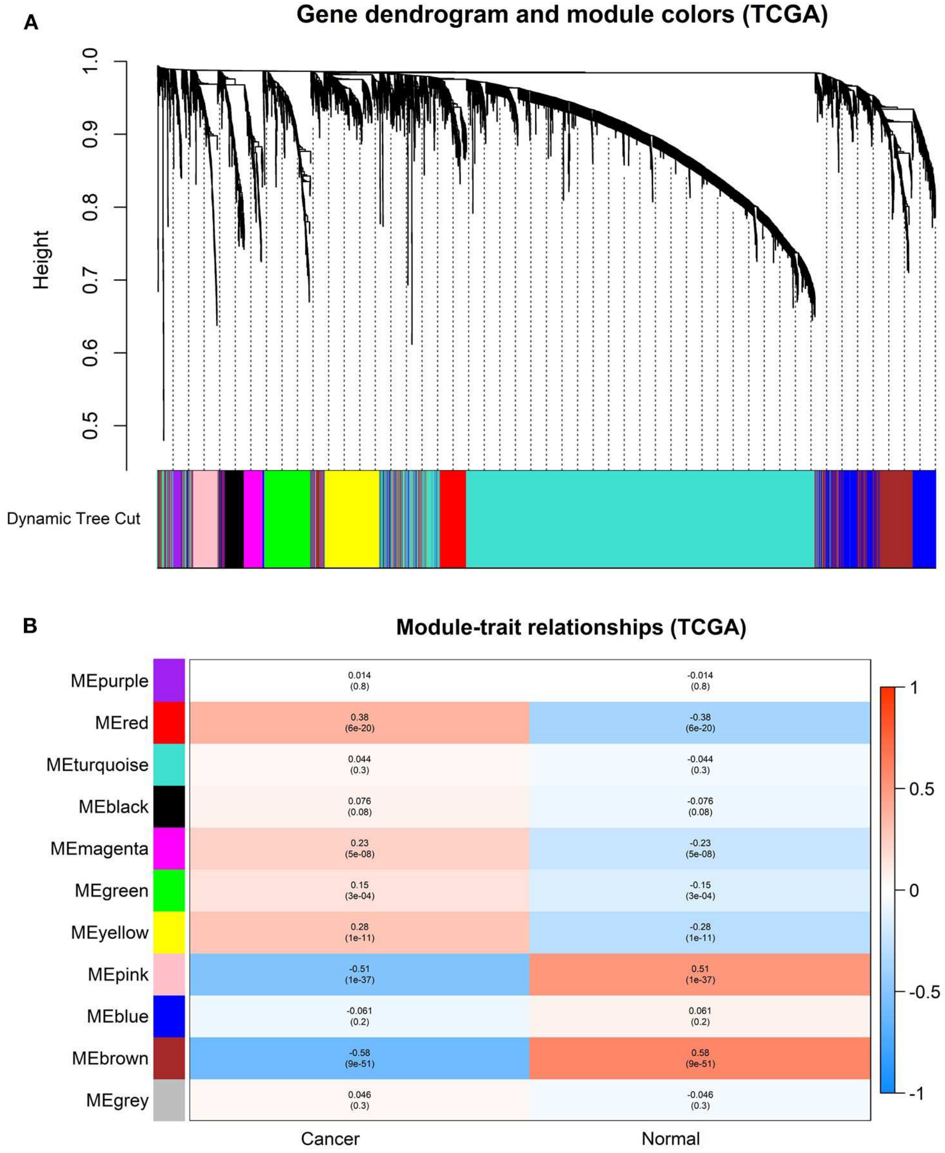

FIGURE 2 | Identification of modules associated with the clinical information in the TCGA-HNSCC dataset. **(A)** The Cluster dendrogram of co-expression network modules was ordered by a hierarchical clustering of genes based on the 1-TOM matrix. Each module was assigned different colors. **(B)** Module-trait relationships. Each row corresponds to a color module and column corresponds to a clinical trait (cancer and normal). Each cell contains the corresponding correlation and *P*-value.

read counts are usually not of interest for further analysis. So, we kept the genes with a cpm (count per million) ≥1 in this study. After filtering using function *rpkm* in *edgeR* package, which is calculated by dividing gene counts by gene length, a total of 15,367 genes with RPKM values were subject to our next analysis.

In addition, the normalized expression profiles of GSE6631, another gene expression profile of HNSCC from GEO, was obtained using R package *GEOquery* (16). GSE6631 consisted of 22 tumor samples and 22 paired normal tissues from patients with HNSCC, which were studied with the GPL8300 platform [HG_U95Av2] Affymetrix Human Genome U95 Version 2 Array. Probes were converted to the gene symbols based on a manufacturer-provided annotation file and duplicated probes for the same gene were removed by determining the median expression value of all its corresponding probes. As a result, a list of 9,203 genes were selected for the subsequent analysis.

Identification of Key Co-expression Modules Using WGCNA

Co-expression networks facilitate methods on network-based gene screening that can be used to identify candidate biomarkers and therapeutic targets. In our study, the gene expression data profiles of TCGA-HNSCC and GSE6631 were constructed to gene co-expression networks using the *WGCNA* package in R (8).

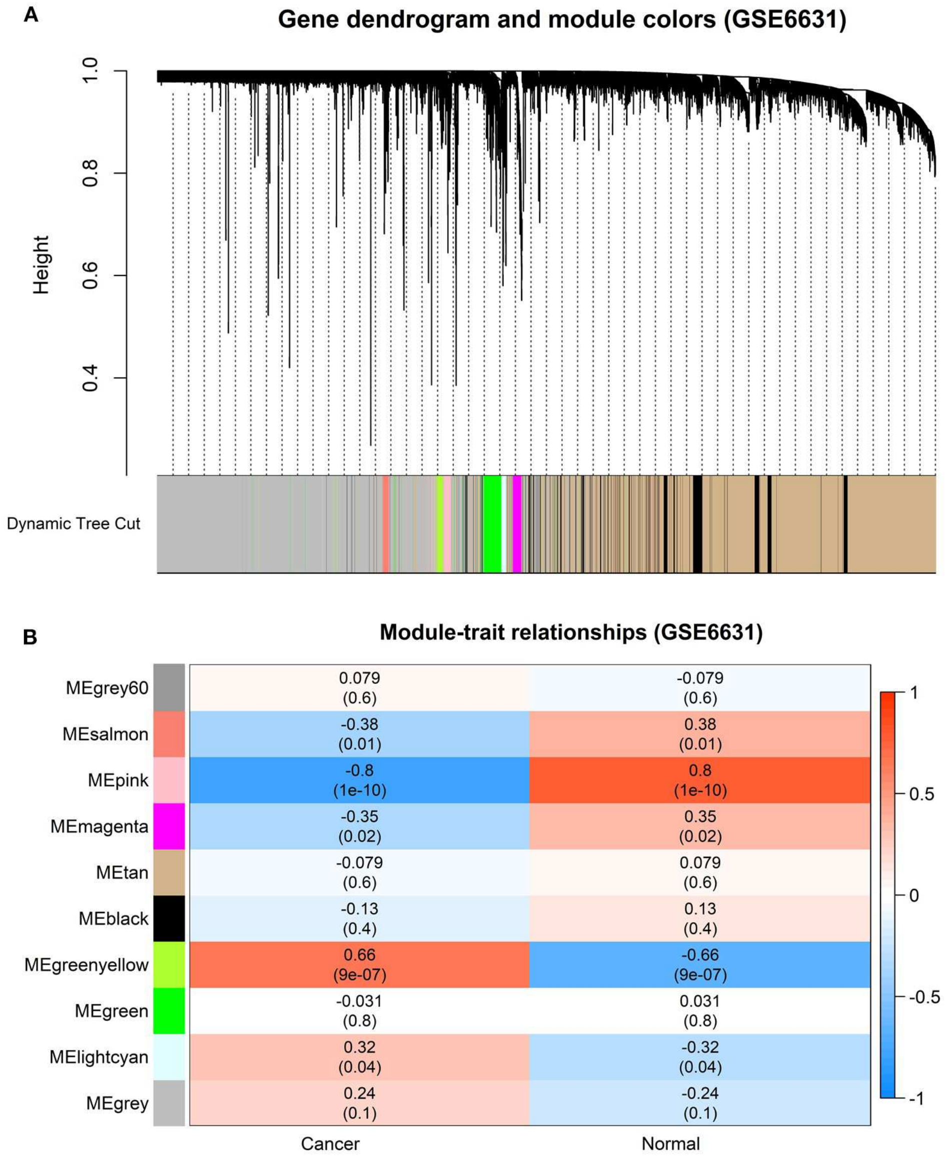

FIGURE 3 | Identification of modules associated with clinical information in the GSE6631 dataset. **(A)** The Cluster dendrogram of co-expression network modules was ordered by a hierarchical clustering of genes based on the 1-TOM matrix. Each module was assigned different colors. **(B)** Module-trait relationships. Each row corresponds to a color module and each column correlates to a clinical trait (cancer and normal). Each cell contains the corresponding correlation and *P*-value.

WGCNA was used to explore the modules of highly correlated genes among samples for relating modules to external sample traits. To build a scale-free network, soft powers β = 3 and 20 were selected using the function *pickSoftThreshold*. Next, the adjacency matrix was created by the following formula: $a_{ij} = |S_{ij}|^{\beta}$ (a_{ij}: adjacency matrix between gene i and gene j, S_{ij}: similarity matrix which is done by Pearson correlation of all gene pairs, β: softpower value), and was transformed into a topological overlap matrix (TOM) as well as the corresponding dissimilarity (1-TOM). Afterwards, a hierarchical clustering dendrogram of the 1-TOM matrix was constructed to classify the similar gene expressions into different gene co-expression modules. To further identify functional modules in a co-expression network, the module-trait associations between modules, and clinical trait information were calculated according to the previous study (17). Therefore, modules with high correlation coefficient were considered candidates relevant to clinical traits, and were selected for subsequent analysis. A more detailed description of the WGCNA method was reported in our previous study (17).

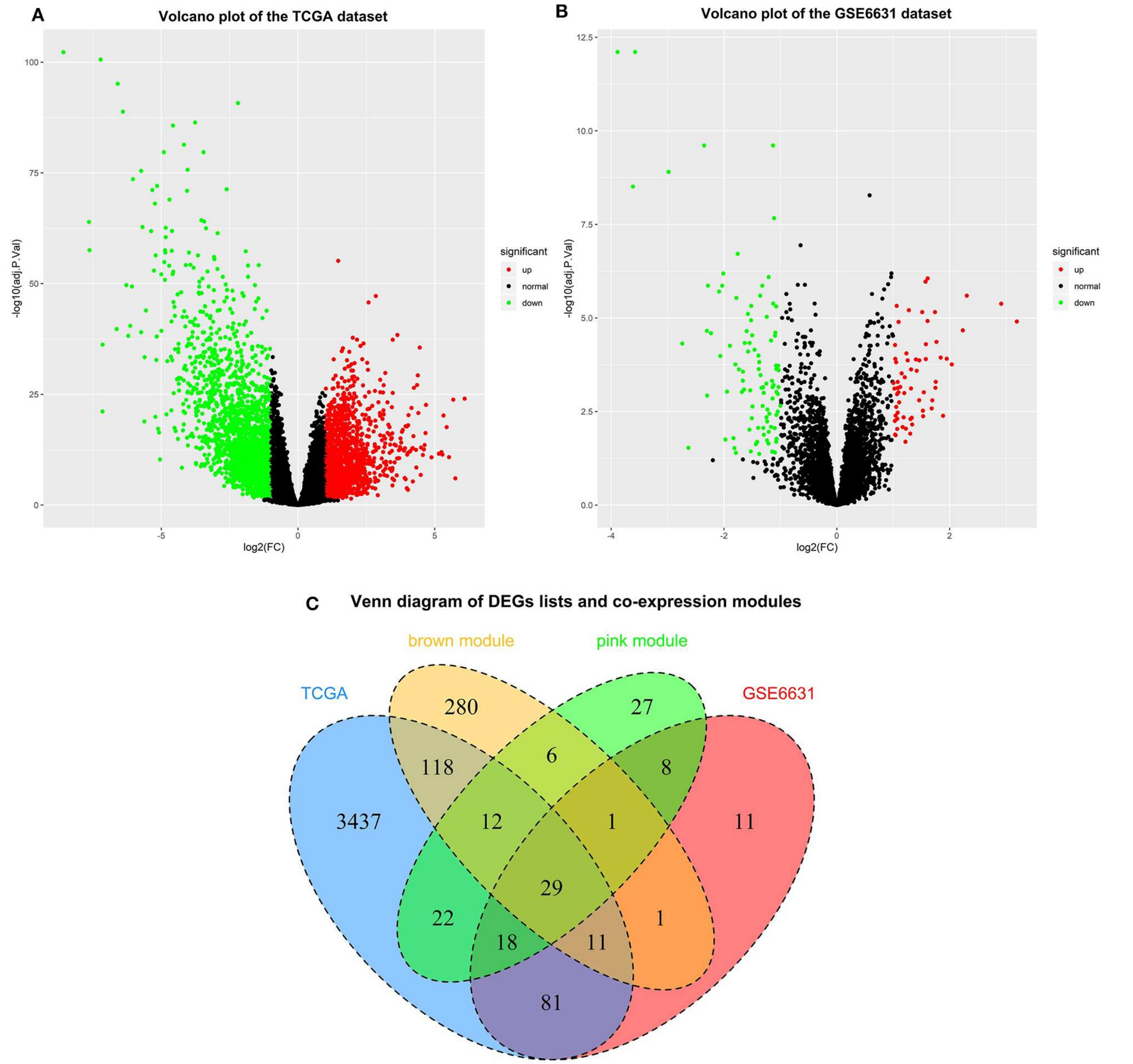

FIGURE 4 | Identification of differentially expressed genes (DEGs) among the TCGA and GSE6631 datasets of HNSCC with the cut-off criteria of |logFC| ≥ 1.0 and adj. $P < 0.05$. **(A)** Volcano plot of DEGs in the TCGA dataset. **(B)** Volcano plot of DEGs in the GSE6631 dataset. **(C)** The Venn diagram of genes among DEG lists and co-expression module. In total, 29 overlapping genes in the intersection of DEG lists and two co-expression modules.

Differential Expression Analysis and Interaction With the Modules of Interest

The R package *limma* (linear models for microarray data) provides an integrated solution for differential expression analyses on RNA-Sequencing and microarray data (18). In order to find the differentially expressed genes (DEGs) between HNSCC and normal tissues, *limma* was applied in the TCGA-HNSCC and GSE6631 dataset, respectively, to screen out DEGs. The *p*-value was adjusted by the Benjamini–Hochberg method to control for the false discovery Rate (FDR). Genes with the cut-off criteria of $|\text{logFC}| \geq 1.0$ and adj. $P < 0.05$ were regarded as DEGs. The DEGs of the TCGA-HNSCC and GSE6631 dataset were visualized as a volcano plot by using the R package *ggplot2* (19). Subsequently, the overlapping genes between DEGs and co-expression genes that were extracted from the co-expression network were used to identify potential prognostic genes, which were presented as a Venn diagram using the R package *VennDiagram* (20).

Functional Annotation for Genes of Interest

To explore Gene Ontology (GO) of selected genes, R package clusterProfiler package (21) was used to explore the functions among genes of interest, with a cut-off criterion of adjusted $p < 0.05$. GO annotation that contains the three sub-ontologies—biological process (BP), cellular component (CC), and molecular function (MF)—can identify the biological properties of genes and gene sets for all organisms (22).

Construction of PPI and Screening of Hub Genes

In our study, we used the STRING (Search Tool for the Retrieval of Interacting Genes) online tool, which is designed for predicting protein–protein interactions (PPI), to construct a PPI network of selected genes (23). Using the STRING database, genes with a score ≥ 0.4 were chosen to build a network model visualized by Cytoscape (v3.7.2) (24). In a co-expression network, Maximal Clique Centrality (MCC) algorithm was reported to be the most

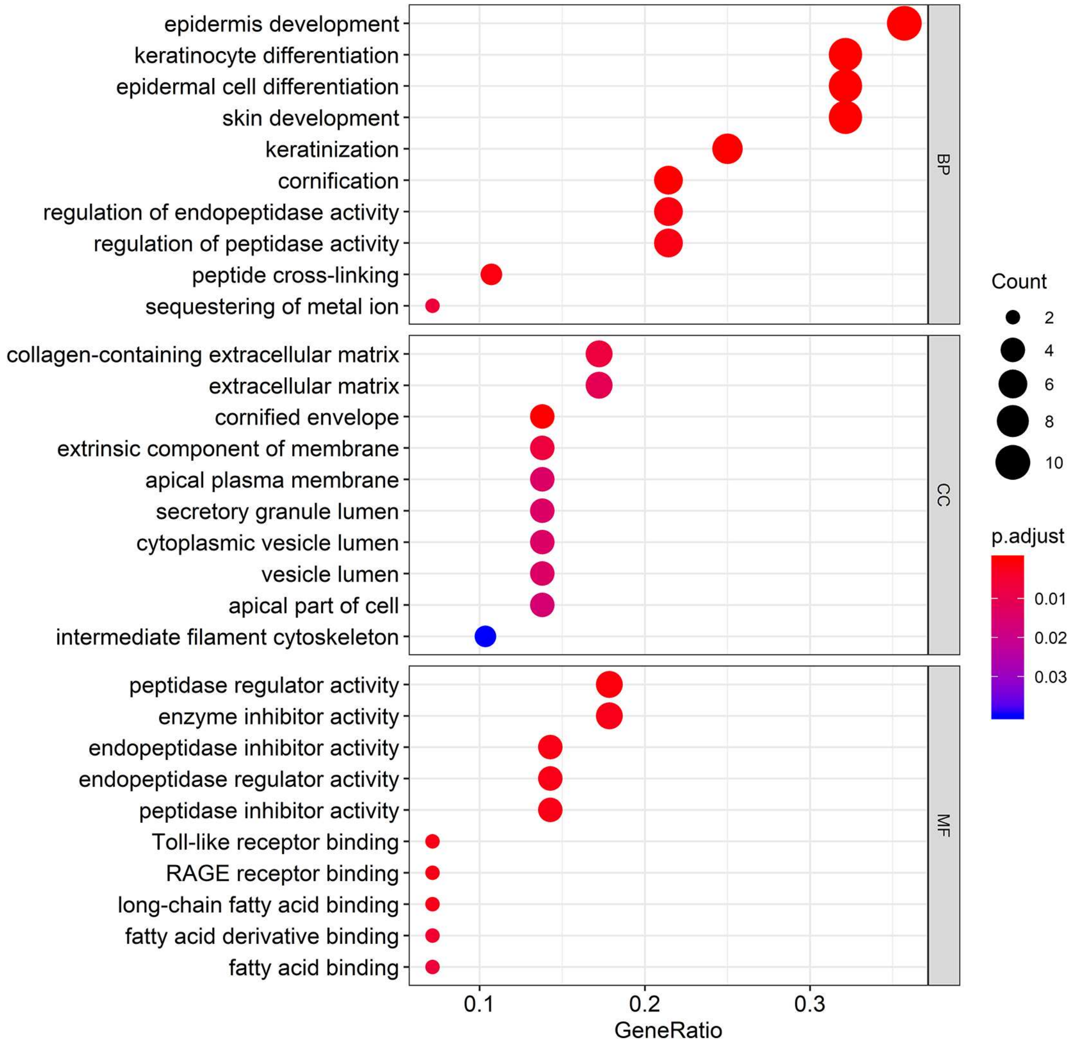

FIGURE 5 | Gene Ontology (GO) enrichment analysis for the genes in the brown module. The color represents the adjusted *p*-values (BH), and the size of the spots represents the gene number.

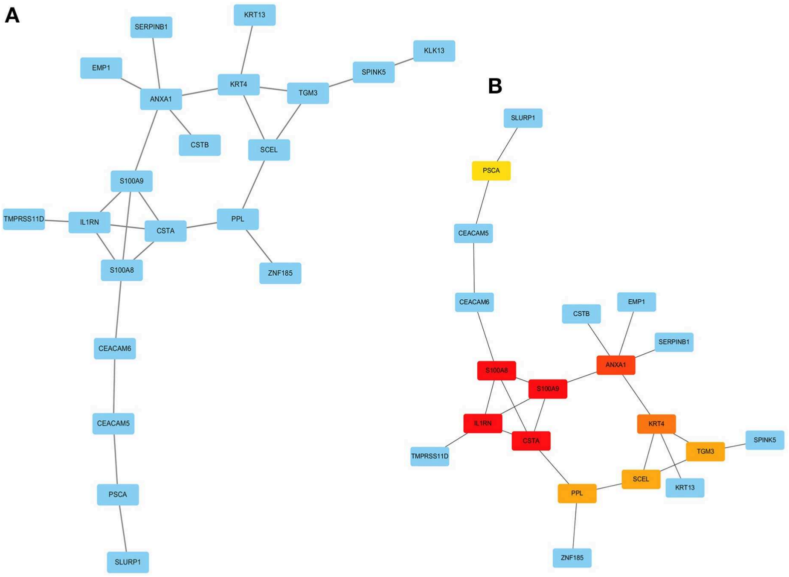

FIGURE 6 | Visualization of the protein-protein interaction (PPI) network and the candidate hub genes. **(A)** PPI network of the genes between DEG lists and two co-expression modules. The blue nodes represent the genes. Edges indicate interaction associations between nodes. **(B)** Identification of the hub genes from the PPI network using maximal clique centrality (MCC) algorithm. Edges represent the protein-protein associations. The red nodes represent genes with a high MCC sores, while the yellow node represent genes with a low MCC sore.

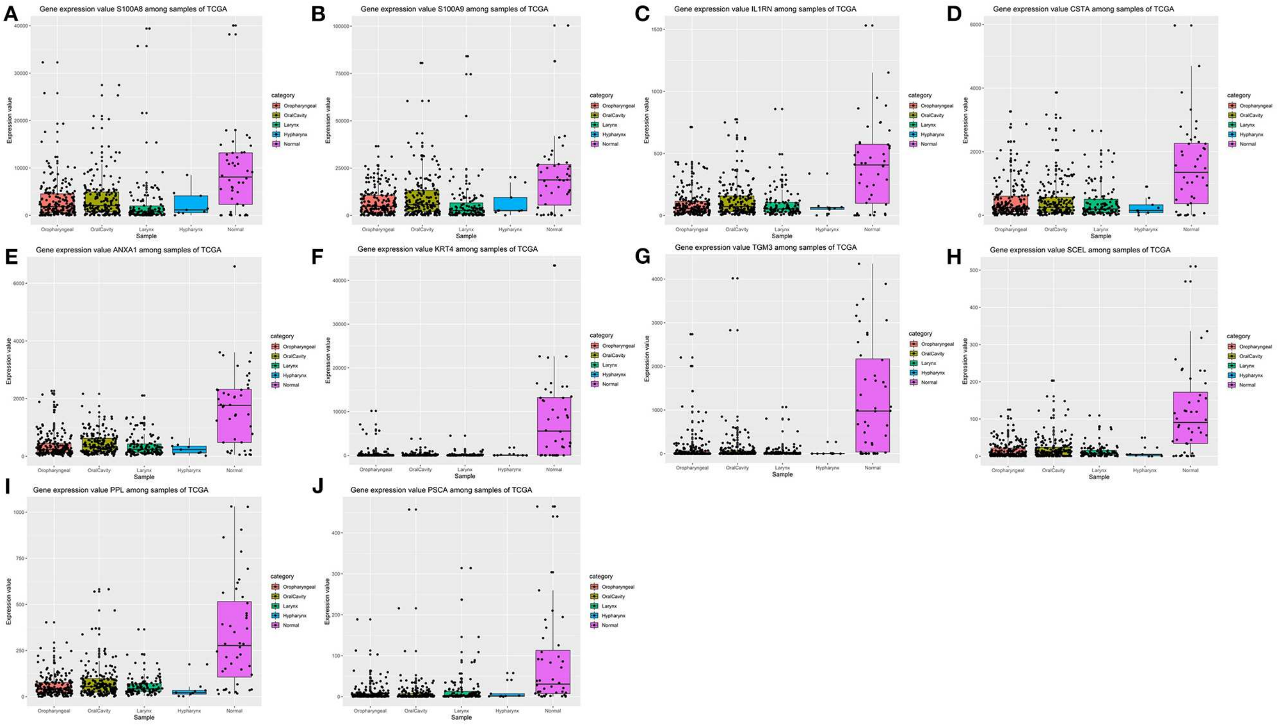

FIGURE 7 | Validation of expression levels of the ten hub genes among HNSCCs and normal tissues from the TCGA database. **(A)** Gene expression value S100A8 among samples of TCGA. **(B)** Gene expression value S100A9 among samples of TCGA. **(C)** Gene expression value IL1RN among samples of TCGA. **(D)** Gene expression value CSTA among samples of TCGA. **(E)** Gene expression value ANXA1 among samples of TCGA. **(F)** Gene expression value KRT4 among samples of TCGA. **(G)** Gene expression value TGM3 among samples of TCGA. **(H)** Gene expression value SCEL among samples of TCGA. **(I)** Gene expression value PPL among samples of TCGA. **(J)** Gene expression value PSCA among samples of TCGA.

effective method of finding hub nodes (25). The MCC of each node was calculated by CytoHubba, a plugin in Cytoscape (25). In this study, the genes with the top 10 MCC values were considered as hub genes.

Verification of the Expression Patterns and the Prognostic Values of Hub Genes

In order to confirm the reliability of the hub genes, we verified the expression patterns of the hub genes in different pathological tumors and normal tissues. The expression level of each hub gene between cancer and normal tissue was plotted as a box plot graph. Based on the data from the TCGA database, Kaplan–Meier univariate survival analysis was performed by using the *survival* package in R software to explore the relationship between overall survival (OS) and hub genes in patients. Moreover, the association between disease-free survival (DFS) and hub genes expressed in HNSCC patients was determined using the online tool GEPIA2 (26). In our study, only patients with completed follow-up times were selected for survival analysis and then divided into two separate groups based on the median expression value of hub genes. The survival-related hub genes with log-rank $p < 0.05$ were regarded as statistically significant.

Validation of Protein Expressions of Survival-Related Hub Genes by the HPA Database

The protein expression of the survival-related genes between HNSCC and normal tissues was determined using immunohistochemistry (IHC) from the Human Protein Atlas database (HPA, https://www.proteinatlas.org/). HPA is a valuable database that provides a large amount of transcriptomics and proteomics data in specific human tissues and cells for

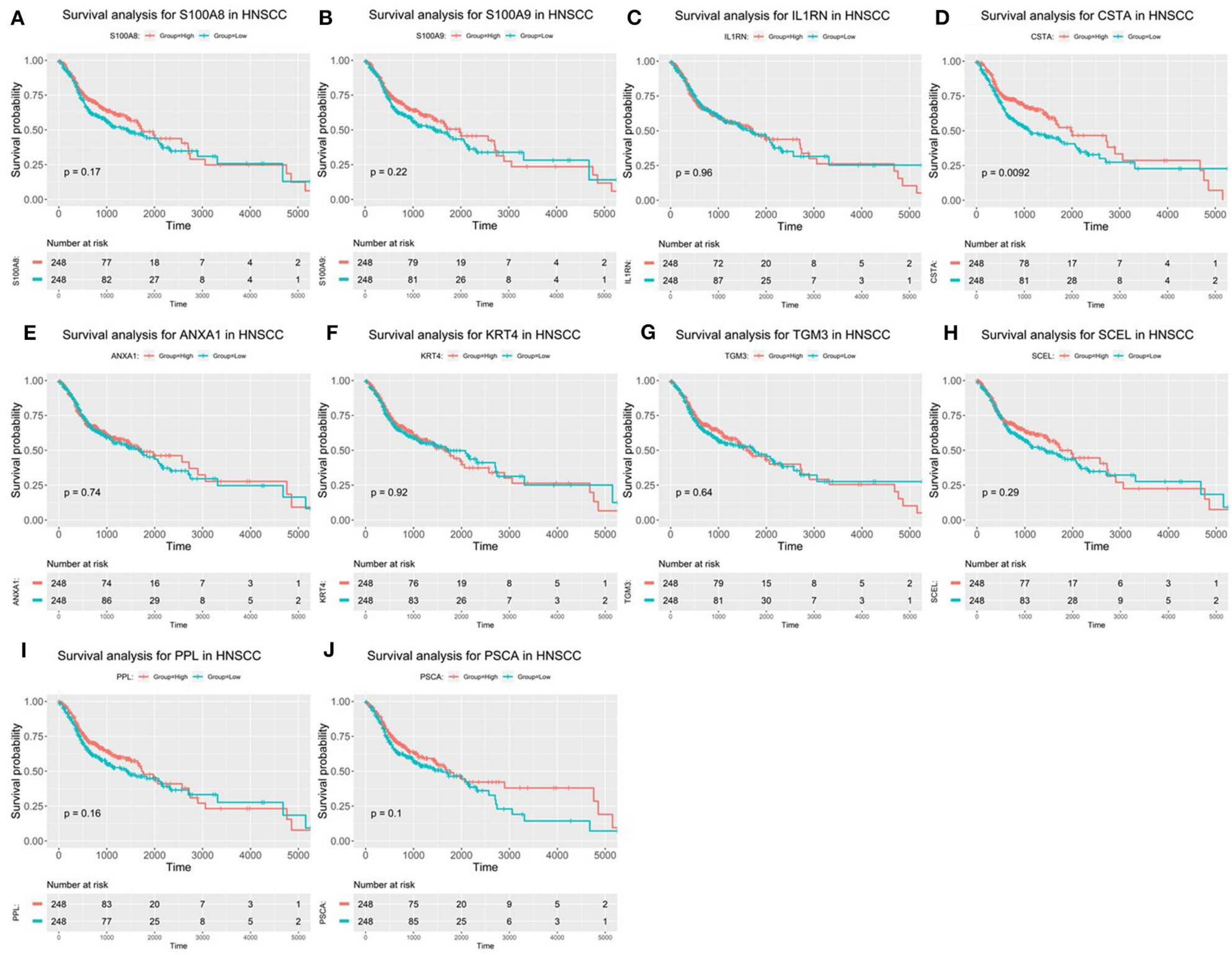

FIGURE 8 | Overall survival (OS) analysis of 10 hub genes in HNSCC patients from the GEPIA2 database. **(A)** Survival analysis for S100A8 in HNSCC. **(B)** Survival analysis for S100A9 in HNSCC. **(C)** Survival analysis for IL1RN in HNSCC. **(D)** Survival analysis for CSTA in HNSCC. **(E)** Survival analysis for ANXA1 in HNSCC. **(F)** Survival analysis for KRT4 in HNSCC. **(G)** Survival analysis for TGM3 in HNSCC. **(H)** Survival analysis for SCEL in HNSCC. **(I)** Survival analysis for PPL in HNSCC. **(J)** Survival analysis for PSCA in HNSCC. The patients were stratified into high-level group (red) and low-level group (green) according to median expression of the gene. Log-rank $P < 0.05$ was considered to be a statistically significant difference.

researchers (27). Moreover, the IHC-based protein expression pattern is the most common application of immunostaining to detect the relative location and abundance of proteins (28).

RESULTS

Construction of Weighted Gene Co-expression Modules

In order to find the functional clusters in HNSCC patients, the gene co-expression networks were constructed from the TCGA-HNSCC and GSE6631 datasets with the *WGCNA* package. With each module assigned a color, a total of 10 modules in the TCGA-HNSCC (**Figure 2A**) and nine modules in the GSE6631 (**Figure 3A**) were identified in the present study (excluding a gray module that was not assigned into any cluster). Then, we plotted the heatmap of module-trait relationships to evaluate the association between each module and two clinical traits (cancer and normal). The results of the module-trait relationships are presented in **Figure 2B**, **3B**, revealing that the brown module in the TCGA-HNSCC and pink module in the GSE6631 were found to have the highest association with normal tissues (brown module: $r = 0.58$, $p = 9e{-}51$; pink module: $r = 0.8$, $p = 1e{-}10$).

Identification of Genes Between the DEG Lists and Co-expression Modules

Based on the cut-off criteria of $|\text{logFC}| \geq 1.0$ and adj. $P < 0.05$, a total of 3,728 DEGs in the TCGA dataset (**Figure 4A**) and 160 DEGs in the GSE6631 dataset (**Figure 4B**) were found to be dysregulated in tumor tissues by the *limma* package. As shown in **Figure 4C**, 458 and 123 co-expression genes were found in the brown module of TCGA dataset and the pink module in GSE6631, respectively. In total, the 29 overlapping genes were extracted for validating the genes of co-expression modules (**Figure 4C**).

Functional Enrichment Analyses for the 29 Genes

To gain further insight into the potential functions of the 29 genes that overlapped with DEG lists and two co-expression modules, gene enrichment analysis was performed by the *clusterProfiler*

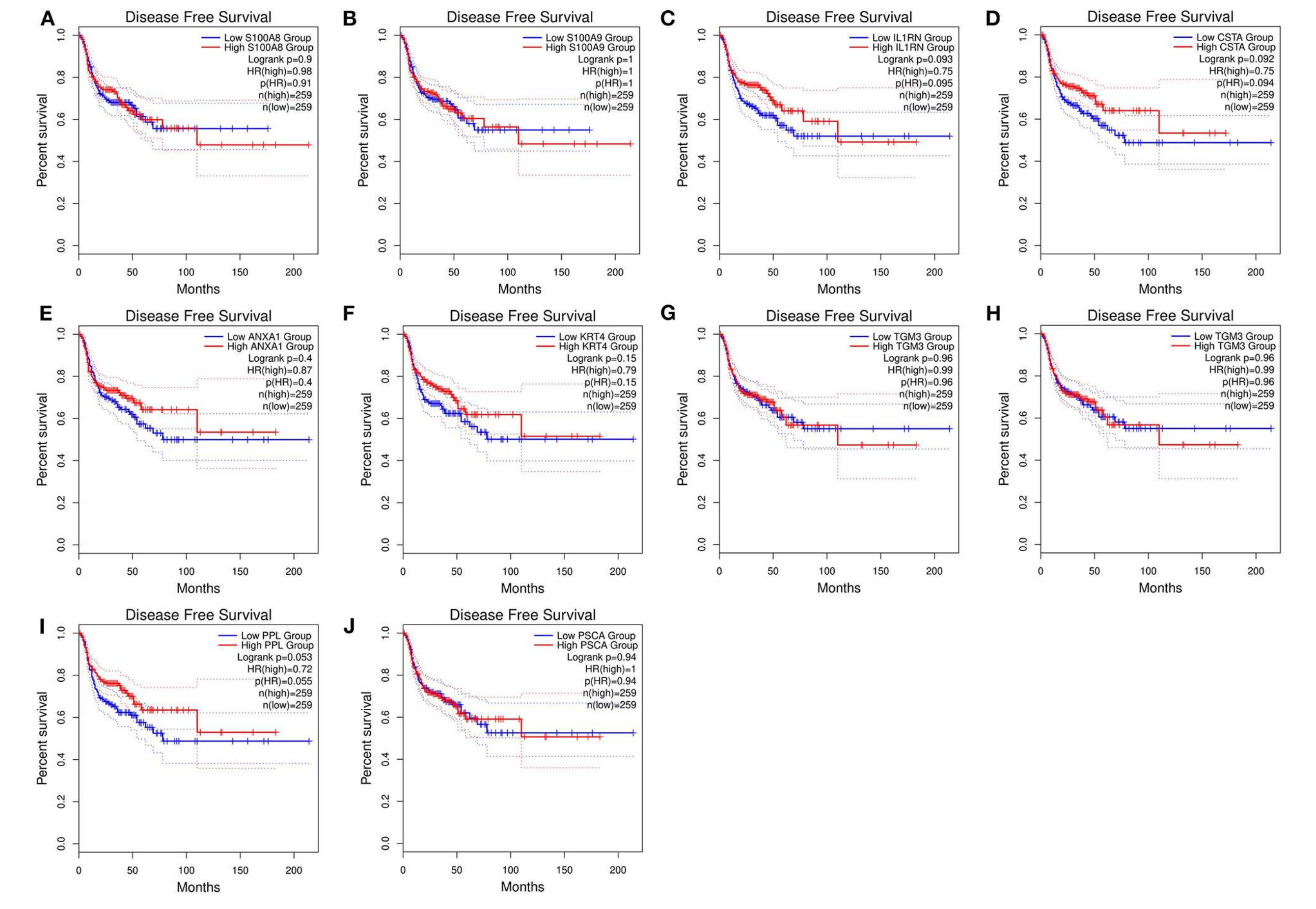

FIGURE 9 | Disease-free survival (DFS) analysis of 10 hub genes in HNSCC patients from the GEPIA2 database. **(A)** Survival analysis for S100A8 in HNSCC. **(B)** Survival analysis for S100A9 in HNSCC. **(C)** Survival analysis for IL1RN in HNSCC. **(D)** Survival analysis for CSTA in HNSCC. **(E)** Survival analysis for ANXA1 in HNSCC. **(F)** Survival analysis for KRT4 in HNSCC. **(G)** Survival analysis for TGM3 in HNSCC. **(H)** Survival analysis for SCEL in HNSCC. **(I)** Survival analysis for PPL in HNSCC. **(J)** Survival analysis for PSCA in HNSCC. The patients were stratified into high-level group (red) and low-level group (green) according to median expression of the gene. Log-rank $P < 0.05$ was considered to be a statistically significant difference.

package. After screening of GO enrichment analysis, we observed several enriched gene sets shown in **Figure 5**. The biological process (BP) of 29 genes are mainly enriched in epidermis development and epidermal cell differentiation. For the result of the cellular component (CC), it was revealed that these genes were mainly involved in apical plasma membrane, apical part of cell, and cell-cell junction. Moreover, in the molecular function (MF) analysis, peptidase regulator activity and enzyme inhibitor activity were suggested to be related to the 29 genes.

PPI Network Construction and Hub Genes Identification

The PPI network among the overlapped genes was established by using the STRING database, with 21 nodes and 25 edges (**Figure 6A**). The hub genes selected from the PPI network using the MCC algorithm of CytoHubba plugin were shown in **Figure 6B**. According to the MCC sores, the top ten highest-scored genes, including S100 calcium-binding protein A8 (S100A8), S100 calcium-binding protein A9 (S100A9), Interleukin-1 receptor antagonist (IL1RN), Cystatin A (CSTA), Annexin-A1 (ANXA1), Keratin 4 (KRT4), Transglutaminase 3 (TGM3), Sciellin (SCEL), Periplakin (PPL), and Prostate Stem Cell Antigen (PSCA), were selected as the hub genes.

Verification of the Expression Patterns, the Prognostic Values, and Protein Expression of Hub Genes

After the ten hub genes (S100A8, S100A9, IL1RN, CSTA, ANXA1, KRT4, TGM3, SCEL, PPL, and PSCA) were screened out by CytoHubba plugin, we verified the expression levels of the hub genes among the patients of the TCGA database. As shown in **Figure 7**, all of the ten hub genes were found to be significantly downregulated in HNSCC carcinoma compared with normal tissues. In addition, OS and DFS analyses of the ten hub genes were performed by Kaplan–Meier plotter using the R *survival* package (**Figure 8**) and the GEPIA2 database (**Figure 9**) for investigating the prognostic values of the hub gens in the HNSCC patients. Of the ten hub genes, the Kaplan–Meier analyses suggested that the lower expression level of CSTA was significantly associated with worse OS of the HNSCC patients ($P < 0.05$) (**Figure 8D**), while with DFS there was no significant difference observed in HNSCC patients with an expression level of CSTA ($P < 0.05$) (**Figure 9D**). Furthermore, the protein levels of the CSTA gene was significantly lower in tumor tissues compared with normal tissues based on the HPA database (**Figure 10**). All the above-mentioned observations confirmed down-expression of CSTA is associated with worse prognosis and lower overall survival in HNSCC patients.

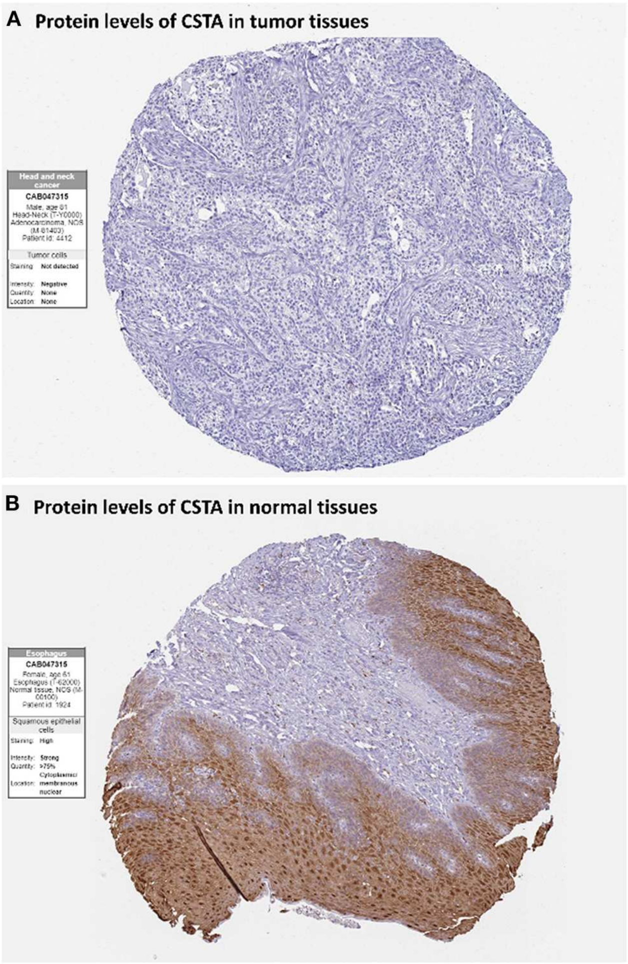

FIGURE 10 | Immunohistochemistry of the CSTA gene in HNSCC and normal tissues from the human protein atlas (HPA) database. **(A)** Protein levels of CSTA in HNSCC tissues (antibody CAB047315; staining: not detected; intensity: negative; quantity: none). **(B)** Protein levels of CSTA in normal oral mucosa tissues (antibody CAB047315; staining: high; intensity: strong; quantity: >75%).

DISCUSSION

Head and neck squamous cell carcinomas (HNSCC) are a group of cancers found in several regions, including the mouth, nose, throat, larynx, sinuses, or salivary glands. Although the treatment of head and neck cancer has improved, the prognosis of patients is generally poor due to the lack of precise molecular targets. Therefore, better biomarkers for specific prognosis and progression of HNSCC are demanded. In this study, a total of 29 significant genes with the same expression trends were identified in the TCGA and GSE6631 databases using integrated bioinformatic analysis. As suggested in functional annotation analysis by the *clusterProfiler* package, these genes were mainly enriched in epidermis development and differentiation, which are basic processes in cell proliferation. Furthermore, according to MCC scores from the CytoHubba plugin in Cytoscape, the top 10 HNSCC-related genes were screened out (namely S100A8, S100A9, IL1RN, CSTA, ANXA1, KRT4, TGM3, SCEL, PPL, and PSCA) and all their expression patterns were found be downregulated in HNSCC tissues compared with the normal controls. Among them, CSTA downexpression was significantly associated with poor overall survival in head and neck cancers. Finally, survival and immunohistochemical analysis for CSTA was carried out.

CSTA, also known as Cystatin A or stefin A, is a member of the cystatin superfamily. It is an intracellular inhibitor regulating the activities of cystatin proteinase and has an important role in desmosome-mediated cell-cell adhesion (29, 30). Furthermore, lower mRNA levels of CSTA have been reported in breast (31), prostate (32), skin (30), and esophagus tumors (33) as compared to adjacent control tissues (34, 35). In our study, CSTA was down-regulated in tumor tissues compared with normal tissues, showing a significant correlation with HNSCC. Previous studies demonstrated that higher levels of CSTA in tumor tissues have been shown to correlate with a favorable prognosis of patients with HNSCC, that was consistent with our finding of survival analysis (36–39).

As with all research, our study also had limitations about the classification of tumors to different subtypes. Although we provided a comprehensive bioinformatics analysis to identify potential diagnostic genes between cancer and normal tissues, it may not be very accurate for each patient with HNSCC subtypes. Moreover, the molecular mechanisms involved in the survival-related genes that affected the prognosis of HNSCC patients should be further validated through a series of experiments.

In summary, by integrating WGCNA with differential gene expression analysis, our study generated the significant survival-related gene CSTA that has potential for prognosis prediction in HNSCC.

DATA AVAILABILITY STATEMENT

All available data were analyzed in this study. Those can be found here: TCGA (https://portal.gdc.cancer.gov/), GEO (https://www.ncbi.nlm.nih.gov/gds), and HPA (https://www.proteinatlas.org/).

AUTHOR CONTRIBUTIONS

J-HC, CW, and CL: conceptualization and methodology. CW and J-HC: software and data curation. JT and CL: validation. J-HC and CL: writing—original draft preparation. CW and JT: writing—review, editing, and supervision.

FUNDING

This present study was funded by Asia University and Show Chwan Memorial Hospital (Grant no. RD108029).

REFERENCES

1. Marur S, Forastiere AA. Head and neck squamous cell carcinoma: update on epidemiology, diagnosis, and treatment. *Mayo Clin Proc.* (2016) 91:386–96. doi: 10.1016/j.mayocp.2015.12.017
2. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin.* (2018) 68:394–424. doi: 10.3322/caac.21492
3. Spitz MR. Epidemiology and risk factors for head and neck cancer. *Semin Oncol*. (1994) 21:281–8.
4. Marur S, D'Souza G, *Westra* WH, Forastiere AA. HPV-associated head and neck cancer: a virus-related cancer epidemic. *Lancet Oncol.* (2010) 11:781–9. doi: 10.1016/S1470-2045(10)70017-6
5. D'Souza G, Dempsey A. The role of HPV in head and neck cancer and review of the HPV vaccine. *Prev Med.* (2011) 53 (Suppl. 1):S5–11. doi: 10.1016/j.ypmed.2011.08.001
6. Ragin CC, Modugno F, Gollin SM. The epidemiology and risk factors of head and neck cancer: a focus on human papillomavirus. *J Dent Res.* (2007) 86:104–14. doi: 10.1177/154405910708600202
7. Can T. Introduction to bioinformatics. *Methods Mol Biol*. (2014) 1107:51–71. doi: 10.1007/978-1-62703-748-8_4
8. Langfelder PS. Horvath WGCNA: an R package for weighted correlation network analysis. *BMC Bioinform.* (2008) 9:559. doi: 10.1186/1471-2105-9-559
9. Zhang B, Horvath S. A general framework for weighted gene co-expression network analysis. *Stat Appl Genet Mol Biol.* (2005) 4:17. doi: 10.2202/1544-6115.1128
10. Li J, Zhou D, Qiu W, Shi Y, Yang JJ, Chen S, et al. Application of weighted gene co-expression network analysis for data from paired design. *Sci Rep*. (2018) 8:622. doi: 10.1038/s41598-017-18705-z
11. Saris CGJ, Horvath S, van Vught WJP, van Es MA, Blauw HM, Fuller TF, et al. Weighted gene co-expression network analysis of the peripheral blood from amyotrophic lateral sclerosis patients. *BMC Genomics.* (2009) 10:405. doi: 10.1186/1471-2164-10-405
12. Yang Y, Han L, Yuan Y, Li J, Hei N, Liang H. Gene co-expression network analysis reveals common system-level properties of prognostic genes across cancer types. *Nat Commun.* (2014) 5:3231. doi: 10.1038/ncomms4231
13. Segundo-Val IS, Sanz-Lozano CS. Introduction to the Gene Expression Analysis. *Methods Mol Biol.* (2016) 1434:29–43. doi: 10.1007/978-1-4939-3652-6_3
14. Colaprico A, Silva TC, Olsen C, Garofano L, Cava C, Garolini D, et al. TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Res*. (2016) 44:e71. doi: 10.1093/nar/gkv1507
15. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics.* (2010) 26:139–40. doi: 10.1093/bioinformatics/btp616
16. Sean D, Meltzer PS. GEOquery. *Bioinformatics.* (2007) 23:1846–7. doi: 10.1093/bioinformatics/btm254
17. Wang CCN, Li CY, Cai JH, Sheu PC, J.Tsai JP, Wu MY, et al. Identification of prognostic candidate genes in breast cancer by integrated bioinformatic analysis. *J Clin Med*. (2019) 8:1160. doi: 10.3390/jcm8081160
18. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res*. (2015) 43:e47. doi: 10.1093/nar/gkv007
19. Wickham H. *Ggplot2: Elegant Graphics for Data Analysis*. (2009) Dordrecht; New York, NY: Springer.
20. Chen H, and Boutros PC. VennDiagram: a package for the generation of highly-customizable Venn Euler diagrams in R. *BMC Bioinform*. (2011) 12:35. doi: 10.1186/1471-2105-12-35
21. Yu G, Wang LG, Han Y, He QY. clusterProfiler: an R package for comparing biological themes among gene clusters. *Omics.* (2012) 16:284–7. doi: 10.1089/omi.2011.0118
22. Gene Ontology Consortium. The Gene Ontology (GO) project in 2006. *Nucleic Acids Res*. (2006) 34:D322–6. doi: 10.1093/nar/gkj021
23. Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, et al. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res*. (2019) 47:D607–13. doi: 10.1093/nar/gky1131
24. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* (2003) 13:2498–504. doi: 10.1101/gr.1239303
25. Chin CH, Chen SH, Wu HH, Ho CW, Ko MT, Lin CY. cytoHubba: identifying hub objects and sub-networks from complex interactome. *BMC Syst Biol.* (2014) 8(Suppl. 4):S11. doi: 10.1186/1752-0509-8-S4-S11

26. Tang Z, Kang B, Li C, Chen T, Zhang Z. GEPIA2: an enhanced web server for large-scale expression profiling and interactive analysis. *Nucleic Acids Res.* (2019) 47:W556–60. doi: 10.1093/nar/gkz430
27. Thul PJ, Lindskog C. The human protein atlas: a spatial map of the human proteome. *Protein Sci.* (2018) 27:233–44. doi: 10.1002/pro.3307
28. Maity B, Sheff D, Fisher RA. Immunostaining: detection of signaling protein location in tissues, cells and subcellular compartments. *Methods Cell Biol.* (2013) 113:81–105. doi: 10.1016/B978-0-12-407239-8.00005-7
29. Blaydon DC, Nitoiu D, Eckl KM, Cabral RM, Bland P, Hausser I, et al. Mutations in CSTA, encoding Cystatin A, underlie exfoliative ichthyosis and reveal a role for this protease inhibitor in cell-cell adhesion. *Am J Hum Genet*. (2011) 89:564–71. doi: 10.1016/j.ajhg.2011.09.001
30. Gupta A, Nitoiu D, Brennan-Crispi D, Addya S, Riobo NA, Kelsell DP, et al. Cell cycle- and cancer-associated gene networks activated by Dsg2: evidence of cystatin a deregulation and a potential role in cell-cell adhesion. *PLoS ONE*. (2015) 10:e0120091. doi: 10.1371/journal.pone.0120091
31. Duivenvoorden HM, Rautela J, Edgington-Mitchell LE, Spurling A, Greening DW, Nowell CJ, et al. Myoepithelial cell-specific expression of stefin A as a suppressor of early breast cancer invasion. *J Pathol*. (2017) 243:496–509. doi: 10.1002/path.4990
32. Mirtti T, Alanen K, Kallajoki M, Rinne A, Soderstrom KO. Expression of cystatins, high molecular weight cytokeratin, and proliferation markers in prostatic adenocarcinoma and hyperplasia. *Prostate*. (2003) 54:290–8. doi: 10.1002/pros.10196
33. Luo A, Kong J, Hu G, Liew CC, Xiong M, Wang X, et al. Discovery of Ca2+-relevant and differentiation-associated genes downregulated in esophageal squamous cell carcinoma using cDNA microarray. *Oncogene*. (2004) 23:1291–9. doi: 10.1038/sj.onc.1207218
34. Kos J, Lah TT. Cysteine proteinases and their endogenous inhibitors: target proteins for prognosis, diagnosis and therapy in cancer (review). *Oncol Rep.* (1998) 5:1349–61. doi: 10.3892/or.5.6.1349
35. Kos J, Krasovec M, Cimerman N, Nielsen HJ, Christensen IJ, Brunner N. Cysteine proteinase inhibitors stefin A, stefin B, and cystatin C in sera from patients with colorectal cancer: relation to prognosis. *Clin Cancer Res.* (2000) 6:505–11.
36. Ma Y, Chen Y, Li Y, Grün K, Berndt A, Zhou Z, et al. Cystatin A suppresses tumor cell growth through inhibiting epithelial to mesenchymal transition in human lung cancer. *Oncotarget*. (2018) 9:14084–98. doi: 10.18632/oncotarget.23505
37. Strojan P, Budihna M, Smid L, Svetic B, Vrhovec I, Kos J, et al. Prognostic significance of cysteine proteinases cathepsins B and L and their endogenous inhibitors stefins A and B in patients with squamous cell carcinoma of the head and neck. *Clin Cancer Res*. (2000) 6:1052–62.
38. Anicin A, Gale N, Smid L, Kos J, Strojan P. Expression of stefin A is of prognostic significance in squamous cell carcinoma of the head and neck. *Eur Arch Otorhinolaryngol.* (2013) 270:3143–51. doi: 10.1007/s00405-013-2465-5
39. Ralhan R, Desouza LV, Matta A, Tripathi SC, Ghanny S, Datta Gupta S, et al. Discovery and verification of head-and-neck cancer biomarkers by differential protein expression analysis using iTRAQ labeling, multidimensional liquid chromatography, and tandem mass spectrometry. *Mol Cell Proteomics*. (2008) 7:1162–73. doi: 10.1074/mcp.M700500-MCP200

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.