A Novel Potential Diagnostic and Prognostic Biomarker for Esophageal Cancer

Fei Miao
Tianjin University of Finance and Economics
Fucun Xie
Peking Union Medical College Hospital
Xin Xu
Tianjin Medical University General Hospital
Zhisong Liu (zhisong.liu@stu.tjufe.edu.cn)
Tianjin University of Finance and Economics https://orcid.org/0000-0003-3213-4743

Primary research

Keywords: Esophageal cancer, Diagnosis, Prognosis, Biomarkers, Network analysis

Posted Date: May 5th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-410477/v1

Abstract

Background: Esophageal cancer (EC) is a lethal malignancy of the digestive tract with high morbidity and mortality rates, lacking molecular biomarkers for EC diagnosis and prognosis with high specificity and sensitivity.

Methods: Downloaded from the Cancer Genome Atlas (TCGA), the high-throughput transcriptome sequencing mRNA data was normalized, after which dysregulated genes were identified via limma package. Weight gene coexpression network analysis (WGCNA) was performed to select diagnostic modules and candidate genes. Later, Hub genes were screened through functional enrichment, LASSO Cox regression and survival analysis. Furthermore, we performed comparation with other biomarkers and investigated the association between our target gene and immune microenvironment.

Results: Esophageal adenocarcinoma subtype and esophageal squamous cell carcinoma subtype displayed similar expression patterns in terms of differentially expressed genes, which indicates the existence of consensus biomarkers. After a series of analyses, including weighted gene coexpression network analysis, Gene Ontology enrichment analysis and Cox regression analysis, *CMA1* was identified as an abnormally upregulated biomarker with highest diagnostic and prognostic potential. Moreover, *CMA1* might exert its effect by influencing the immune microenvironment.

Conclusion: CMA1 is a potential biomarker for EC in not only diagnosis but also prognosis. Future prospective clinic trials are expected to evaluate its performance in clinical practice.

1. Introduction

Esophageal cancer (EC) is a lethal malignancy of the digestive tract that occurs worldwide. It was estimated that there were over half a million new cases and related deaths in 2018.[1] Because of the lack of specific early symptoms and early screening methods, most patients with EC are diagnosed at advanced stages and miss the best treatment opportunity. Statistics show that more than 30% of patients with EC are diagnosed at an advanced stage, and the overall 5-year survival rate of EC is only 16%.[2] Traditional treatments, including surgery, radiotherapy and chemotherapy, are still the main strategies in EC. However, their outcomes are far from satisfactory.[3] Therefore, physicians and researchers are paying increasing attention to molecular biomarkers as both diagnostic and prognostic tools for EC, as they are efficient means for understanding the molecular basis of this disease and could also be exploited in new treatment strategies. Hence, an important task is to identify these biomarkers, which could thereby facilitate EC diagnosis and prognosis.

In the field of the transcriptome research, several studies have been published on the identification of biomarkers in EC; some biomarkers were identified for diagnosis or prognosis, and some were identified to distinguish between the esophageal squamous cell carcinoma (ESCC) or esophageal adenocarcinoma (EAC) subtypes of EC.[4–6] In general, the identified biomarkers are not broadly applicable. This situation

leads to the lower feasibility of the identified biomarkers because only small groups of patients may benefit.

Unlike these studies, we tried to identify important risk-related genes that have the ability not only to diagnose tumors but also to predict overall survival (OS) for EC within a set of differentially expressed protein-coding genes (PCGs). Unlike most of the studies that tested the reproducibility of the results in the last step, we tried to validate the results in most steps of our analyses.

2. Materials And Methods

2.1. Overall design

A schematic of our procedure is shown in Figure S1. An online high-throughput transcriptome sequencing mRNA data was extracted for the analyses. Raw data were processed and normalized, and dysregulated genes were identified between tumor samples and normal tissue samples. Hence, normalized data and differentially expressed gene (DEG) data were obtained. DEG data were divided into normal reference and tumor DEG data, and the latter set was further divided into ESCC and EAC subgroups. Network concepts were used to measure the similarity between ESCC and EAC. Tumor DEG data were further split into two sets, a tumor test set and a tumor reference set, according to the OS status of the patients. The OS of the patient corresponding to each sample was uncensored in the tumor test set and censored in the tumor reference set.

Duplicated tumor samples from a single case were removed to ensure that the corresponding clinical information had a unique match within the expression profile data. Moreover, outlier samples were removed from the tumor test set, tumor reference set and normal reference set. The tumor test set was used for the subsequent procedures, including weighted network construction, module detection, and module selection in the network. The tumor reference set was used to check the reproducibility of the network, examine the preservation of the modules, and match a diagnosis to the selected module. The normal reference set was also used for comparisons.

Hub genes were selected, and Gene Ontology (GO) enrichment analysis was performed within the selected module. Genes in the intersection of the hub gene set and the enriched gene set in the GO enrichment analysis were selected as candidates. Least absolute shrinkage and selection operator (LASSO) Cox regression was used to identify the biomarker among the candidates. Finally, the selected biomarker was widely explored and comprehensively compared with other biomarkers.

2.2. Data sources

High-throughput transcriptome sequencing data of EC samples were derived from the TCGA-ESCA project in The Cancer Genome Atlas (TCGA) platform, and mRNA expression profiles were extracted. Corresponding clinical information, including survival time, survival status and subtype of EC, was also obtained. Altogether, the mRNA profiles included data for 19645 PCGs among 171 samples, of which 160

were tumor samples and 11 were normal tissue samples. Of the tumor samples, 78 were from EAC cases, and 81 were from ESCC cases.

Another dataset was derived from the Genotype-Tissue Expression (GTEx) project for diagnostic validation purposes. The dataset included information for 1445 normal esophageal tissues.

2.3. Identification of DEGs

First, raw data were imported, organized, and filtered. The processed data were then transformed into log-counts per million (log-CPM) values, and the trimmed mean of M-values *(TMM)* was calculated as a scaling factor to normalize the data. Comparisons were made between tumor samples and normal tissue samples. Second, with the limma package, the mean-variance trend was investigated, and precision weight values were calculated to remove the heteroscedacity of the data. Finally, the *empirical Bayes* and *treat* algorithms were used to assess differential expression and perform gene set testing in linear models. DEGs were identified as significant if the fold change (FC) > 2 or < 0.5 and adjusted *p*-value < 0.05.

2.4. Similarity comparison for EAC and ESCC and assessment of the reproducibility of the network between the tumor test set and the tumor reference set

Outlier samples were detected and removed in advance. Standardized connectivity[7] was used as the statistic for detecting outliers, and the critical values were set at ± 5. The similarity between EAC and ESCC patients was explored by assessing network features, including the density, mean of the clustering coefficient, centralization, connectivity correlation, and adjacency correlation. A correlation-based adjacency matrix was used to explore the patterns. The same strategy was used to examine the reproducibility of the network between the tumor test set and the tumor reference set.

2.5. Weight gene coexpression network analysis (WGCNA)

2.5.1. Network construction, module exploration and preservation

A power function was used to construct the network, and a soft threshold that led the network to satisfy the feature of scale-free topology was determined. Distances were defined by dissimilarity of the topological overlap matrix (TOM), and then dynamic tree cut[8] techniques were used to detect modules. Module preservation was measured with composited statistics, *Zsummary* and *medianRank*.[9]

2.5.2. Module selection and diagnosis

The module of interest was selected based on three aspects. The first aspect is based on the concept of mean gene significance (GS). In a given module, the mean GS is defined by the mean absolute value of

the correlation between gene expression and uncensored survival time. The second aspect was the correlation between the module and survival. This can be measured by assessing the association between module eigengenes and uncensored survival time, where module eigengenes are defined as the first principal component of the modules. In addition, the association between intramodular connectivity and the absolute value of GS also was considered, as it is crucial to select hub genes in the downstream analysis. To ensure that the selected module was credible, further graphics were used to assign a diagnosis to the selected module of interest.

2.5.3. Hub genes selection and GO enrichment analysis

The relationship between intramodular connectivity and eigengene-based connectivity in the selected module was explored graphically; eigengene-based connectivity was represented as module membership (MM) and was defined as the absolute value of the correlation between the eigengenes and the expression values of each gene. Moreover, the association between MM and GS was investigated to decide the criterion for selecting the hubs.

The GO enrichment analysis results for the genes in the selected module were explored with the Database for Annotation, Visualization and Integrated Discovery (DAVID).[10, 11] Important GO terms were selected under the following criteria: *p*-value > 0.05, *fold enrichment* > 2, and *proportion of genes involved > 0.1*.

2.6. LASSO Cox regression and survival analysis

Genes related to important GO terms were identified, and those intersecting with hub genes were considered candidates. LASSO Cox regression analysis[12, 13] was used to select the final biomarker at the intersection. In brief, in one experiment, ten folds cross-validation was used to estimate the values of lambda. Then, *lambda.min*, the value of lambda that led to the minimum for the mean cross-validated error, was used to predict the coefficients of candidates. Inclusion or exclusion of the candidates was decided according to their corresponding nonzero or zero coefficients. The experiments were repeated 1000 times. The most frequent candidate was finally selected as the biomarker in our study. For the selected biomarker, OS features were investigated in the null model; Kaplan-Meier (KM) survival curves were generated, and the log-rank test was used to measure the difference in survival between patients with low expression levels and high expression levels of the marker.

2.7. Comprehensive comparisons with other biomarkers

A brief summary of recent studies that identified biomarkers in EC was made. These biomarkers include *MAGEA6*[4], *CXCL12*, and *CXCR4*.[5, 6] We broadly compared the performance of the biomarkers identified in our study and others. In addition, traditional biomarkers, including *PTGS2*, *EGFR*, *ERBB2*, and *TP53*, were also included in the comparisons.

In the diagnostic aspect, the diagnostic ability of the biomarkers was validated externally by comparing TCGA tumor samples and GTEx normal tissue samples. Boxplots were used to depict the differences, and the predictive ability of the biomarkers was measured via random forest models and corresponding

receiver operating characteristic (ROC) curves. In the prognostic aspect, bootstrap-validated C-index values were calculated to indicate the predictive ability of the biomarkers.

2.8. Target biomarker analysis

The Tumor IMmune Estimation Resource[14] (TIMER, https://cistrome.shinyapps.io/timer/) web server provides systemic analysis of immune infiltrates in various cancer types. Here, we mainly made use of the differential expression (Diff Exp) module, gene module, somatic copy number alteration (SCNA) module and correlation module. The Diff Exp module presents information on DEGs of intertest between tumor tissues and adjacent normal tissues across all TCGA tumor types, and significance is evaluated by Wilcoxon test. The gene module helped to estimate tumor purity[15] and the abundances of B cells, CD4 + T cells, CD8 + T cells, neutrophils, macrophages, and dendritic cells based on statistical deconvolution of immune infiltrates.[16] In the SCNA module, SCNAs are defined by GISTIC 2.0. The infiltration level for each SCNA category was compared between tumor samples and normal samples using a two-sided Wilcoxon rank-sum test. In addition, the correlation between the target biomarker and immune biomarkers was explored with adjustments for tumor purity.

2.9. Statistics

Statistical analyses were performed by R software (version 3.6.1). The main package used in our research was WGCNA. In addition, edgeR[17] was used to process raw data, and limma[18] was used to identify DEGs; survival[19], rms[20], pec[21] and glmnet[12] were used to analyze the data; ggplot2[22], survminer[23], GOplot[24], ggpubr[25] and RColorBrewer[26] were used to create graphs.

3. Results

3.1. Identification of DEGs

Figure 1 shows the procedures used to process, normalize data, and identify DEGs. Initially, none expressed genes in the profiles were filtered. In total, 969 out of 19645 genes were removed. Then, the sequencing depths of the mRNA profiles were explored and are shown in Fig. 1A. The maximum value was over 3 times the lowest value. Hence, normalization factors were calculated via the *TMM* method, and the corresponding results are shown in Fig. 1B. In addition, the data were transformed to a log-CPM scale for normalization.

The mean-variance trend was explored and the voom plot shows a decreasing trend between the means and variances in Fig. 1C, where the horizontal axis represents the normalized value and vertical axis represents the square-root of the standard deviations of the residual variances extracted from the fitting linear models in the voom function. Hence, normalized data were accommodated via precision weights; the result is shown in Fig. 1D, where the blue line represents the average $\log_2$ residual standard deviation. Finally, empirical Bayes moderation was carried out, and the *treat* method was used to identify DEGs between tumor samples and normal tissue samples. In brief, 542 DEGs were identified, of which 388 were

downregulated and 154 were upregulated. The results are shown visually as a volcano plot in Fig. 1E, where the green dots represent downregulated DEGs and the red dots represent upregulated DEGs.

3.2. Similarity comparison for EAC and ESCC and reproducibility checking of the network between the tumor test set and the tumor reference set

The similarity between EAC and ESCC subtypes was examined via network feature assessment. We also used the normal reference set to indirectly illustrate the similarity or dissimilarity between the sets. Outlier samples were removed before the analysis. One sample each was removed from the EAC and ESCC subsets. Figure 2A-B shows the corresponding cluster trees and heatmaps for the EAC group, ESCC group and normal reference group. The darker the color is, the higher the absolute correlation value. The EAC and ESCC groups showed similar patterns in the cluster tree and heatmap plot, whereas the normal reference group exhibited a quite different pattern.

We next looked at network features for each set. Mathematically, density values range from 0 to 1. The closer the value is to 1, the tighter the connection strength of the nodes in a network. The summarized table shows that the density values were 0.22, 0.21, and 0.51 for the EAC, ESCC, and normal reference sets, respectively. The mean clustering coefficient was used to quantify the internal extent of the network structure. The values of the two subtypes were 0.28 and 0.26. Centralization was used to measure star topological structure of a network. A value of 0 indicates that the connectivity of all nodes is equal, and a value of 1 indicates that a star topology exists. The results show that the values were relatively low in the three groups, which suggests that the network is far from having star topology. In addition, the correlation values for connectivity and adjacency were high in the EAC and ESCC subsets. In summary, the two subsets are quite similar. The similarity may be indirectly illustrated via the normal reference group, which exhibits a quite different pattern. In short, EAC and ESCC appear similar; thus, it may be feasible to identify a consensus biomarker for the EAC and ESCC groups.

The same strategy was used to examine the reproducibility of the network between the tumor test set and the tumor reference set. The corresponding results are shown in Fig. 2C-D. In brief, the two tumor sets are similar, which means that the network is stable and reproducible.

3.3. Network construction, module exploration and preservation

A power function was used to transform the correlation-based similarity matrix into an adjacency matrix to construct the weighted coexpression network, and the corresponding cluster dendrogram is shown in Fig. 3A. Under the TOM and dynamic tree cut algorithms, four modules were identified and labeled in different colors. Notably, the gray module did not hold statistical properties. Hence, we were not interested in this module, as hubs could not be identified via statistical procedures. The second largest module was

the turquoise module, which contained 100 DEGs, and the smallest module was the brown module, which contained 68 DEGs. To better illustrate the clustering results, a 3D plot is shown in Fig. 3B.

Module preservation was measured via *Zsummary* and *medianRank*. In brief, the module is preserved if *Zsummary* > 10, is not preserved if *Zsummary* < 2, and is weakly to moderately preserved if 2 < *Zsummary* < 10. In addition, a high-ranked module (according to the statistic of *medianRank*) is more preserved than a low-ranked module. Figure 3C-D show the preservation values for the tumor reference and normal reference sets. All modules were well preserved in the tumor reference set and less preserved in the normal reference set. In addition, the turquoise module was ranked as the highest module in terms of *medianRank*. In conclusion, the modules are preserved in the tumor reference set, which indicates that the identified modules are reproducible.

3.4. Module selection and diagnosis

The module of interest was selected based on Fig. 4A-C. GS values across modules were first explored. The analysis showed that the blue and turquoise modules had higher GS values than other modules (Fig. 4A). In addition, relations between module eigengenes and survival times were investigated via heatmaps (Fig. 4B). The red and blue colors represent positive and negative correlations. The darker the color is, the stronger the correlation. Both the blue module and the turquoise module were significantly correlated with survival time. Figure 4C shows the correlation between intramodular connectivity and GS. The blue module had the highest correlation among all the modules. Based on the above findings, the blue module was selected as the candidate module.

The blue module was then examined graphicly. The relationships between eigengenes and the expression level of each gene are shown in Fig. 4D. The plot shows that the highly expressed genes have corresponding positive eigengene values, and weakly expressed genes have negative eigengene values. Hence, eigengenes are representative of all the genes within the modules. In addition, Fig. 4E-F depicts the network plots of the blue module in the tumor test set and the tumor reference set. The line is colored red if the correlation is positive and blue if it is negative. The thickness of the line reflects the absolute correlation. The size of each black circle indicates the connectivity of the corresponding gene. Genes were ordered by connectivity in the tumor test set. The results indicated that all genes were positively correlated, as no blue line was observed. Subjectively, the two circle plots look similar, which indicates that the blue module is preserved and reproducible.

3.5. Hub genes selection and GO enrichment analysis

Figure 5A shows that MM and intramodular connectivity are high correlated. This indicates that the selection of hub genes via MM is practicable. Figure 5B shows the correlation between MM and GS. Conservatively, we set MM > 0.7 and GS > 0.2 as the criteria to select the hubs, as MM and GS are not highly correlated. As a result, 32 genes were identified as hubs.

In addition, GO enrichment analysis was performed for the blue module genes. Five important GO terms were identified, and the corresponding results are shown in Fig. 5C. Thirty-two genes were involved in

these GO terms, and the results are shown in a chord plot (Fig. 5D).

3.6. LASSO Cox regression and survival analysis

Twelve candidates were identified in the intersection between hub genes and highly enriched genes. The results of LASSO Cox regression for these 12 genes are shown graphically. Figure 6A shows the coefficients for candidates under different values of lambda, and Fig. 6B shows how many candidates were kept in the model under *lambda.min* in just one experiment. The results show that the log value for *lambda.min* is approximately − 2.9 in Fig. 6B, and only one coefficient was nonzero (Fig. 6A), which indicates that only one gene was retained in the model in the experiment. Figure 6C exhibits the results for 1000 experiments. *CMA1* was shown to be the most frequent candidate in LASSO Cox regression.

For *CMA1*, the null model for OS was explored in Fig. 6D. Five-year predictive ability was ignored in the following analyses because only 3 patients were at risk after the fifth year of survival. The predictive ability of *CMA1* for OS is shown in Fig. 6E, and the log-rank test showed that it is a significant predictor ($P = 0.009$).

3.7. Comprehensive comparisons with other biomarkers

Broad comparisons were made within PCG biomarkers for diagnostic and prognostic purposes. Important results are shown in Fig. 7. Boxplots for the expression values of the identified biomarkers between TCGA tumor samples and GTEx normal tissue samples are shown in Fig. 7A. It seems that the diagnostic ability of *MAGEA6*, *CXCL12*, and *CMA1* is acceptable, as the boxes overlapped less than other candidates between tumor tissues and normal tissues. The predictive abilities of the biomarkers were further examined via random forest models and corresponding ROC curves (Fig. 7B). *CMA1* ranked 3rd among all the biomarkers in terms of the area under the ROC curve (AUC) criterion.

The prognostic ability of the biomarkers was assessed according to the C-index, and its corresponding 95% bootstrap validated confidence interval (Fig. 7C). *CMA1* was the best biomarker, and its performance was slightly greater than that of *EGFR*. Furthermore, since only 15 patients assessed were at risk after 3 years of survival, we paid more attention to comparisons of biomarkers within 1 to 3 years of survival. The bootstrap cross-validated C-index was measured at a series of time points within one to three years, and the corresponding results are shown in Fig. 7D. Again, *CMA1* was the best biomarker at the selected time points. In summary, *CMA1* is the best choice for both diagnosis and prognosis prediction for all the identified biomarkers.

3.8. *CMA1* expression correlated with the immune infiltration level in EC

In many types of malignancies, *CMA1* was downregulated in tumor tissues compared to normal tissues (Fig. 8A). Tumor purity was defined as the proportion of cancer cells in the admixture [15]. Herein, the *CMA1* expression level was adversely associated with purity in ESCA, suggesting that *CMA1* is highly expressed in the microenvironment. Positive correlations were found between *CMA1* expression and CD4

+ T cell ($p < 0.01$), macrophage ($p < 0.01$), and dendritic cell ($p < 0.001$) infiltration levels (Fig. 8B). We further explored the association between the expression of *CMA1* and immune markers, including HLA-DPB1 for dendritic cells, FOXP3 for regulatory T cells (Tregs) and PD-1 (PDCD1) for exhausted T cells. All the markers in the scatter plots demonstrated a positive partial correlation ($p < 0.01$) (Fig. 8C). Figure 8D shows the distributions of each immune subset for each copy number status in EC. Only high amplification of B cells and arm-level deletion of CD4 + T cells and dendritic cells displayed significant differences ($p < 0.05$) between tumor tissues and normal tissues.

4. Discussion

EC is one of the most lethal malignancies of the digestive tract worldwide and has high morbidity and mortality rates. Traditional methods are unsatisfactory. Hence, clinicians and physicians are paying more attention to molecular biomarkers for EC diagnosis and prognosis. There are two main histological subtypes of EC: EAC and ESCC. Studies at the molecular level are usually based on one of these two subtypes of EC. In addition, diagnostic and prognostic studies are usually separated. In brief, the identified biomarkers are not broadly applicable. However, no evidence has shown that different molecular therapy methods should be developed that apply to different histological types of EC.

We aimed to identify the risk genes among PCGs that can be applied in both diagnosis and OS prediction. The feasibility of our aim was explored in advance via network analysis. After confirmation of the applicability of the design, a comprehensive study was generated to identify biomarkers via weighted coexpression network analysis, GO enrichment analysis, and LASSO Cox regression analysis. The reproducibility of the network and preservation of the modules were also validated in the analysis. These procedures are crucial, as they are prerequisites to ensure that the results are repeatable. Finally, *CMA1* was identified as a biomarker, and it was widely explored and comprehensively compared with other biomarkers identified in other studies of EC. The results show that *CMA1* is a promising candidate among these biomarkers due to its superior performance in both diagnostic and prognostic roles.

Downregulation of CMA1 expression was observed in various cancers, and the difference was relatively high between tumor and nontumor tissues in the EC dataset (Fig. 8A). Tumor infiltrating lymphocytes are prognostic biomarkers in various cancers.[27, 28] TIMER analysis revealed a correlation between the infiltration of many types of immune cells, especially dendritic cells, and the expression of *CMA1*, which indicated the possibility that *CMA1* influences tumor progression through immune infiltration.

Mast cells are recognized as a crucial part of the immune microenvironment in tumor tissues and modulate tumor progression by releasing protumorigenic and antitumorigenic molecules,[29] including transforming growth factor-beta, tumor necrosis factor-alpha, interleukin-8, FGF-2, and VEGF.[30] Increased activity of chymase and tryptase secreted by mast cells was observed at all stages of tumor progression and was accompanied by an increase in the number and size of blood vessels.[31]

In a breast cancer immunotyping model based on tumor-infiltrating immune cell subsets published in 2020,[32] patients with high B cell, NK cell, CD8 + cell, activated CD4 + memory T cell levels and low γδT

cell and activated mast cell levels had a better prognosis than those with other infiltration patterns, so activated mast cells are unfavorable factors in terms of breast cancer patient survival. Moreover, an *in vitro* animal experiment showed that coinjection of mast cells and HER2-positive breast cancer cells increased tumor engraftment and outgrowth in mice.[33] Additionally, a high level of mast cells in extratumoral benign prostate tissues was considered a risk factor for a poor prognosis in prostate cancer patients.[34] Mechanistically, it was observed that tumor-derived microvesicles from non-small-cell lung cancer (NSCLC) cells are internalized by mast cells, which enhances the migratory ability of mast cells and promotes the release of TNF-alpha and MCP-1.[35] In EC, however, there are some conflicting reports about the effect of mast cell infiltration on prognosis. Tinge *et al* found that the number of mast cells was not related to prognosis in EC patients,[36] but Fakhrjou *et al* pointed out that high mast cell density in the invasive edge of the tumor correlated with tumor progression and poorer survival.[37] Further studies are required to confirm the prognostic significance of mast cell infiltration in malignancies due to the complicated functions of mast cells.

Located at 14q11.2 in a cluster of genes encoding other proteases, *CMA1* encodes a chymotryptic serine proteinase belonging to the peptidase family S1.[38] It is secreted by mast cells and has been reported to play an essential role in the degradation of the extracellular matrix, the regulation of submucosal gland secretion, and the generation of vasoactive peptides. In the heart and blood vessels, the major convertor from angiotensin I to the vasoactive peptide angiotensin II is not angiotensin converting enzyme but CMA1, so *CMA1* aroused the interest of experts in heart[39] and hypertension research.[40]

In the field of oncology, however, *CMA1* is not well studied. There are only a few studies concerning *CMA1* in oncology. It was reported in 2015 that chymase has angiogenic effects and can induce tumor growth.[41] Interestingly, miR-9 induced an elevation of *CMA1* expression in both P815 (a mouse malignant mast cell line expressing activating KIT mutations) and mouse bone marrow-derived mast cells, supporting that miR-9 promotes metastasis by increasing the expression of proteases crucial for physical remodeling of the extracellular matrix.[42] Chymase-positive mast cells seem to be a biomarker of poor prognosis in breast cancer,[43] gastric cancer,[30] and lung cancer.[44]

In different malignancies, the mRNA levels of *CMA1* have various prognostic implications. Elevated *CMA1* mRNA expression levels were found in benign prostate hyperplasia tissue compared to prostate cancer tissue or normal prostate tissue.[45] Of note, the *CMA1* mRNA levels were lower in oral squamous cell carcinoma tissue than in normal tissue, and patients with high *CMA1* expression had apparently better prognoses ($p = 7*10^{-6}$).[46] The *CMA1* mRNA levels were elevated in the normal mucosa compared to the normal tissues and tumor tissues of colorectal carcinoma patients.[47] The upregulation of THBS2 and SPARC at the same time also implied that inflammatory activation of stromal fibroblasts is not only an integral part of colorectal carcinoma but can facilitate its invasion of the surroundings.[47] Increased *CMA1* mRNA was associated with an unfavorable prognosis in gastric cancer but a favorable prognosis in ovarian cancer.[48] Mechanistically, *CMA1* may influence the prognosis of cancer patients by altering the immune microenvironment. In our EC research, the *CMA1* expression level was positively correlated with the infiltration of dendritic cells, regulatory T cells, and exhausted T cells (Fig. 8B-C). The metastasis

promotion potential of dendritic cells is related to inducing Treg cell cytotoxicity and reducing the cytotoxicity of CD8 + T cells. In addition, overexpression of *CMA1* was correlated with poorer OS in gastric cancer patients at lymph node stages 1–3 but not in patients without lymph node metastasis. This discovery strongly suggests that the *CMA1* expression level is a potential predictive marker of metastasis. [48] However, if *CMA1* plays only a driving role in tumor development, it is difficult to explain its controversial prognostic implications across different tumors. Further studies are required to uncover the intricate regulatory network between *CMA1* and the immune microenvironment and develop it as a potential immune therapy target.

Conclusion

In summary, we revealed that *CMA1* is a potential diagnostic and prognostic biomarker of EC through a comprehensive analysis; *CMA1* might exert its effect by influencing the immune microenvironment. This finding may facilitate the individualized treatment of EC patients.

Declarations

Ethics approval and consent to participate

All the data were obtained from public databases according to their guidence, so ethics approval and consent to participate were not in requirement.

Consent for publication

All the authors have read the final edition of this manuscript and approve the publication.

Availability of data and material

RNA sequencing data and associated clinical features of ESCA cohort is available in the Cancer Genome Atlas (TCGA) database, and Genotype-Tissue Expression (GTEx) project is also accessible online for academy usages.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by the Tianjin Philosophy and Social Science research Project (TJTJ19-002), and the National Statistical Research Project (2018LZ21).

Authors' contributions

Fei Miao analyzed the data, wrote the article, and provided financial support for this work, Fucun Xie and Xin Xu analyzed the data and wrote the article, Zhisong Liu conceived the research and revised the article.

Acknowledgements

We thank the Cancer Genome Atlas (TCGA) and Genotype-Tissue Expression (GTEx) project for providing the data.

References

1. 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. Cancer J Clin. 2018;68(6):394–424.
2. Zhang Y. Epidemiology of esophageal cancer. World journal of gastroenterology: WJG. 2013;19(34):5598.
3. Thallinger CM, Kiesewetter B, Raderer M, Hejna M. Pre-and postoperative treatment modalities for esophageal squamous cell carcinoma. Anticancer research. 2012;32(11):4609–27.
4. Hao J, Li S, Li J, Jiang Z, Ghaffar M, Wang M, Jia R, Chen S, Wang Y, Zeng Y. Investigation into the expression levels of MAGEA6 in esophageal squamous cell carcinoma and esophageal adenocarcinoma tissues. Experimental therapeutic medicine. 2019;18(3):1816–22.
5. Łukaszewicz-Zając M, Mroczko B, Kozłowski M, Szmitkowski M: **The serum concentrations of chemokine CXCL12 and its specific receptor CXCR4 in patients with esophageal cancer**. *Disease markers* 2016, **2016**.
6. Sasaki K, Natsugoe S, Ishigami S, Matsumoto M, Okumura H, Setoyama T, Uchikado Y, Kita Y, Tamotsu K, Sakurai T. Expression of CXCL12 and its receptor CXCR4 correlates with lymph node metastasis in submucosal esophageal cancer. Journal of surgical oncology. 2008;97(5):433–8.
7. Oldham M, Langfelder P, Horvath S. Sample Networks for Enhancing Cluster Analysis of Genomic Data: Application to Huntington's Disease. BMC Syst Biol. 2011;6:63.
8. Langfelder P, Zhang B, Horvath S. Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. Bioinformatics. 2008;24(5):719–20.
9. Langfelder P, Luo R, Oldham MC, Horvath S. Is my network module preserved and reproducible? PLoS Comput Biol. 2011;7(1):e1001057.

10. Huang DW, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic acids research. 2009;37(1):1–13.
11. Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nature protocols. 2009;4(1):44.
12. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. Journal of statistical software. 2010;33(1):1.
13. Simon N, Friedman J, Hastie T, Tibshirani R. Regularization paths for Cox's proportional hazards model via coordinate descent. Journal of statistical software. 2011;39(5):1.
14. Li T, Fan J, Wang B, Traugh N, Chen Q, Liu JS, Li B, Liu XS. TIMER: A Web Server for Comprehensive Analysis of Tumor-Infiltrating Immune Cells. Cancer research. 2017;77(21):e108–10.
15. Aran D, Sirota M, Butte AJ. Systematic pan-cancer analysis of tumour purity. Nature communications. 2015;6:8971.
16. Li B, Severson E, Pignon JC, Zhao H, Li T, Novak J, Jiang P, Shen H, Aster JC, Rodig S, et al. Comprehensive analyses of tumor immunity: implications for cancer immunotherapy. Genome Biol. 2016;17(1):174.
17. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26(1):139–40.
18. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic acids research. 2015;43(7):e47–7.
19. Therneau T. **A Package for Survival Analysis in S. version 2.38**. In.; 2015.
20. Harrell FE Jr. **rms: Regression modeling strategies**. *R package version* 2016, 5(2).
21. Mogensen UB, Ishwaran H, Gerds TA. Evaluating random forests for survival analysis using prediction error curves. Journal of statistical software. 2012;50(11):1.
22. Wickham H: **ggplot2**: **elegant graphics for data analysis**: Springer; 2016.
23. Kassambara A, Kosinski M, Biecek P. **survminer: Drawing Survival Curves using'ggplot2'**. *R package version 03* 2017, 1.
24. Walter W, Sánchez-Cabo F, Ricote M. GOplot: an R package for visually combining expression data with functional analysis. Bioinformatics. 2015;31(17):2912–4.
25. Kassambara A. **ggpubr:"ggplot2" based publication ready plots**. *R package version 01* 2017, 6.
26. Neuwirth E, Brewer RC. **ColorBrewer palettes**. *R package version* 2014:1.1-2.
27. Gao G, Wang Z, Qu X, Zhang Z. Prognostic value of tumor-infiltrating lymphocytes in patients with triple-negative breast cancer: a systematic review and meta-analysis. BMC Cancer. 2020;20(1):179.
28. Idos GE, Kwok J, Bonthala N, Kysh L, Gruber SB, Qu C. The Prognostic Implications of Tumor Infiltrating Lymphocytes in Colorectal Cancer: A Systematic Review and Meta-Analysis. Scientific reports. 2020;10(1):3360.
29. Bao X, Shi R, Zhao T, Wang Y. **Mast cell-based molecular subtypes and signature associated with clinical outcome in early-stage lung adenocarcinoma**. *Mol Oncol* 2020.

30. Guidolin D, Ruggieri S, Annese T, Tortorella C, Marzullo A, Ribatti D. Spatial distribution of mast cells around vessels and glands in human gastric carcinoma. Clin Exp Med. 2017;17(4):531–9.
31. de Souza DA Jr, Toso VD, Campos MR, Lara VS, Oliver C, Jamur MC. Expression of mast cell proteases correlates with mast cell maturation and angiogenesis during tumor progression. PloS one. 2012;7(7):e40790.
32. Jiang J, Pan W, Xu Y, Ni C, Xue D, Chen Z, Chen W, Huang J. Tumour-Infiltrating Immune Cell-Based Subtyping and Signature Gene Analysis in Breast Cancer Based on Gene Expression Profiles. J Cancer. 2020;11(6):1568–83.
33. Majorini MT, Cancila V, Rigoni A, Botti L, Dugo M, Triulzi T, De Cecco L, Fontanella E, Jachetti E, Tagliabue E, et al: **Infiltrating mast cell-mediated stimulation of estrogen receptor activity in breast cancer cells promotes the luminal phenotype**. *Cancer Res* 2020.
34. Hempel Sullivan H, Heaphy CM, Kulac I, Cuka N, Lu J, Barber JR, De Marzo AM, Lotan TL, Joshu CE, Sfanos KS. High Extratumoral Mast Cell Counts Are Associated with a Higher Risk of Adverse Prostate Cancer Outcomes. Cancer Epidemiol Biomarkers Prev. 2020;29(3):668–75.
35. Salamon P, Mekori YA, Shefler I. Lung cancer-derived extracellular vesicles: a possible mediator of mast cell activation in the tumor microenvironment. Cancer immunology immunotherapy: CII. 2020;69(3):373–81.
36. Tinge B, Molin D, Bergqvist M, Ekman S, Bergstrom S. Mast cells in squamous cell esophageal carcinoma and clinical parameters. Cancer genomics proteomics. 2010;7(1):25–9.
37. Fakhrjou A, Niroumand-Oscoei SM, Somi MH, Ghojazadeh M, Naghashi S, Samankan S. Prognostic value of tumor-infiltrating mast cells in outcome of patients with esophagus squamous cell carcinoma. Journal of gastrointestinal cancer. 2014;45(1):48–53.
38. Caughey GH, Schaumberg TH, Zerweck EH, Butterfield JH, Hanson RD, Silverman GA, Ley TJ. The human mast cell chymase gene (CMA1): mapping to the cathepsin G/granzyme gene cluster and lineage-restricted expression. Genomics. 1993;15(3):614–20.
39. Stone G, Choi A, Meritxell O, Gorham J, Heydarpour M, Seidman CE, Seidman JG, Aranki SF, Body SC, Carey VJ, et al. Sex differences in gene expression in response to ischemia in the human left ventricular myocardium. Hum Mol Genet. 2019;28(10):1682–93.
40. Wu Y, Li Q, Yang K, Xiao C. [Association of CMA1 gene tag single nucleotide polymorphisms with essential hypertension in Yi population from Yunnan]. Zhonghua Yi Xue Yi Chuan Xue Za Zhi. 2014;31(4):449–54.
41. de Souza Junior DA, Santana AC, da Silva EZ, Oliver C, Jamur MC. The Role of Mast Cell Specific Chymases and Tryptases in Tumor Angiogenesis. Biomed Res Int. 2015;2015:142359.
42. Fenger JM, Bear MD, Volinia S, Lin TY, Harrington BK, London CA, Kisseberth WC. Overexpression of miR-9 in mast cells is associated with invasive behavior and spontaneous metastasis. BMC Cancer. 2014;14:84.
43. Glajcar A, Szpor J, Pacek A, Tyrak KE, Chan F, Streb J, Hodorowicz-Zaniewska D, Okon K. The relationship between breast cancer molecular subtypes and mast cell populations in tumor

microenvironment. Virchows Arch. 2017;470(5):505–15.

44. Ibaraki T, Muramatsu M, Takai S, Jin D, Maruyama H, Orino T, Katsumata T, Miyazaki M. The relationship of tryptase- and chymase-positive mast cells to angiogenesis in stage I non-small cell lung cancer. Eur J Cardiothorac Surg. 2005;28(4):617–21.
45. Neuhaus J, Schiffer E, Mannello F, Horn LC, Ganzer R, Stolzenburg JU. **Protease Expression Levels in Prostate Cancer Tissue Can Explain Prostate Cancer-Associated Seminal Biomarkers-An Explorative Concept Study**. *International journal of molecular sciences* 2017, 18(5).
46. Huang GZ, Wu QQ, Zheng ZN, Shao TR, Lv XZ. Identification of Candidate Biomarkers and Analysis of Prognostic Values in Oral Squamous Cell Carcinoma. Frontiers in oncology. 2019;9:1054.
47. Drev D, Bileck A, Erdem ZN, Mohr T, Timelthaler G, Beer A, Gerner C, Marian B. Proteomic profiling identifies markers for inflammation-related tumor-fibroblast interaction. Clin Proteomics. 2017;14:33.
48. Shi S, Ye S, Mao J, Ru Y, Lu Y, Wu X, Xu M, Zhu T, Wang Y, Chen Y, et al: **CMA1 is potent prognostic marker and associates with immune infiltration in gastric cancer**. *Autoimmunity* 2020:1–8.

Figures

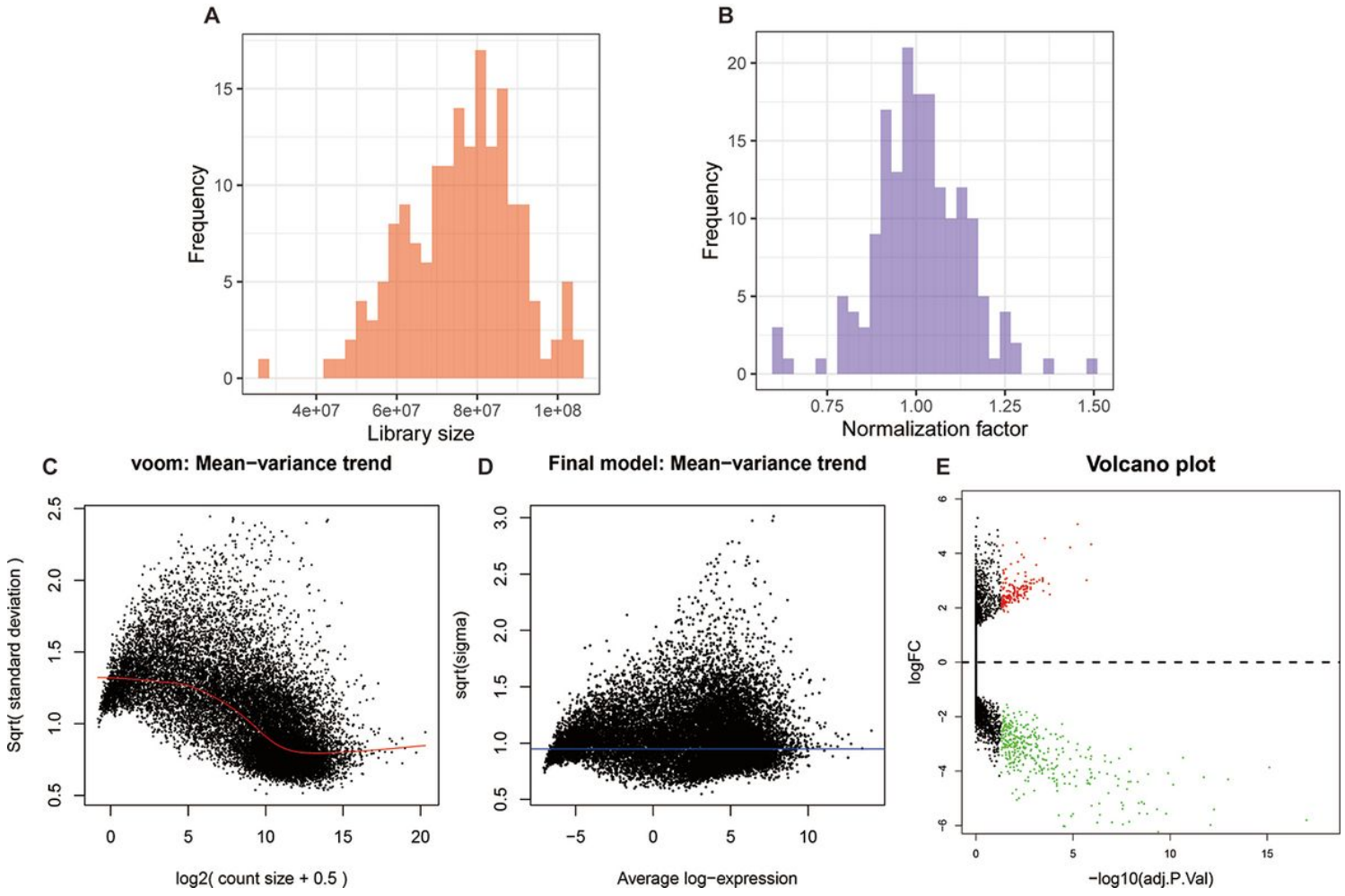
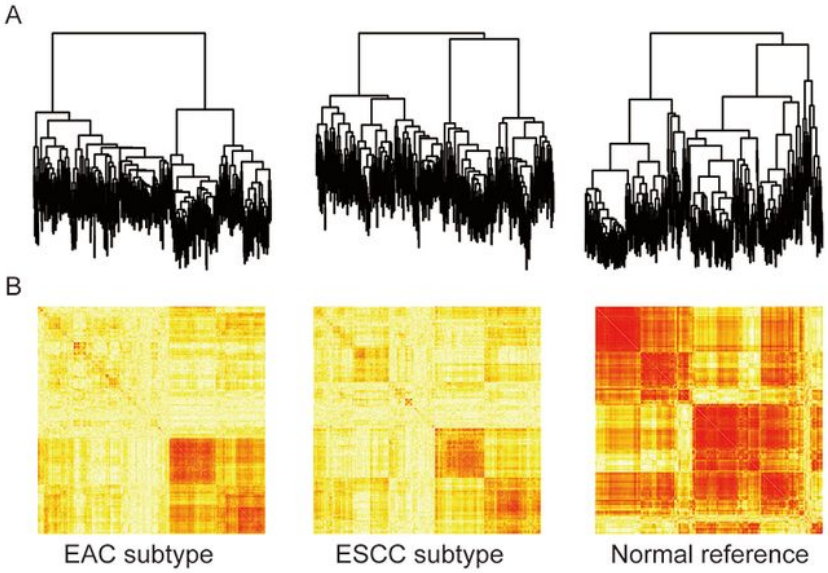
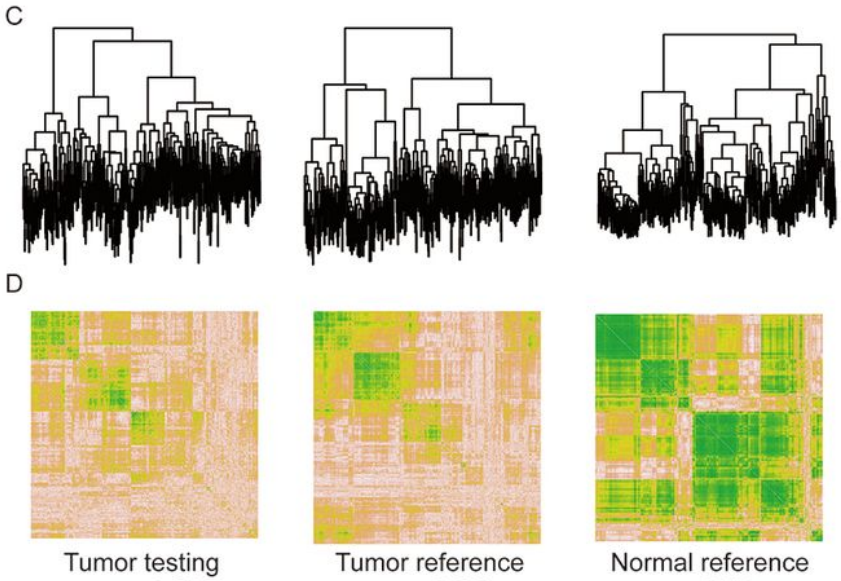


Figure 1

Main features in differential expression analysis. (A) Sequencing depths exploration. (B) Distribution of normalization factors. (C) Mean-variance trend. (D) Patterns after accommodated for the precision weights. (E) Volcano plot.



Density	0.217	0.209	0.511
Mean of clustering coefficient	0.279	0.257	0.575
Centralization	0.151	0.138	0.161
Connectivity correlation	NA	0.683	0.257
Adjacency correlation	NA	0.525	0.384



Density	0.211	0.233	0.511
Mean of clustering coefficient	0.252	0.289	0.575
Centralization	0.121	0.153	0.161
Connectivity correlation	NA	0.792	0.104
Adjacency correlation	NA	0.660	0.308

Figure 2

Primary assessment of the feasibility of the overall design. (A-B) Cluster tree and heatmap for EAC and ESCC comparison. (C-D) Cluster trees and heatmap for comparison of the two tumor sets.

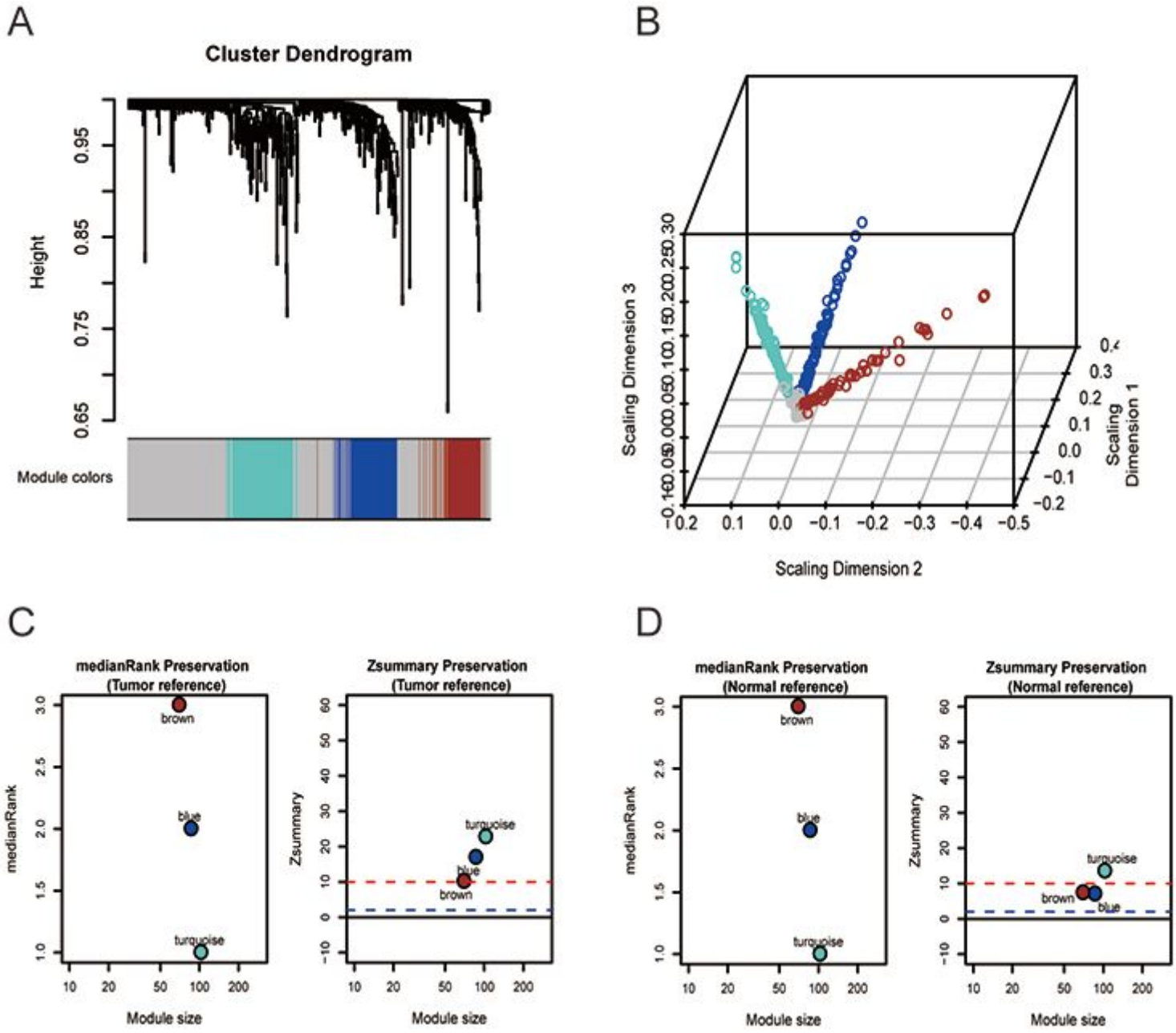


Figure 3

Module exploration and preservation assessment. (A) Cluster dendrogram and the corresponding modules. (B) Cluster results in a 3D plot. (C) Module preservation in the tumor reference set. (D) Module preservation in the normal reference set.

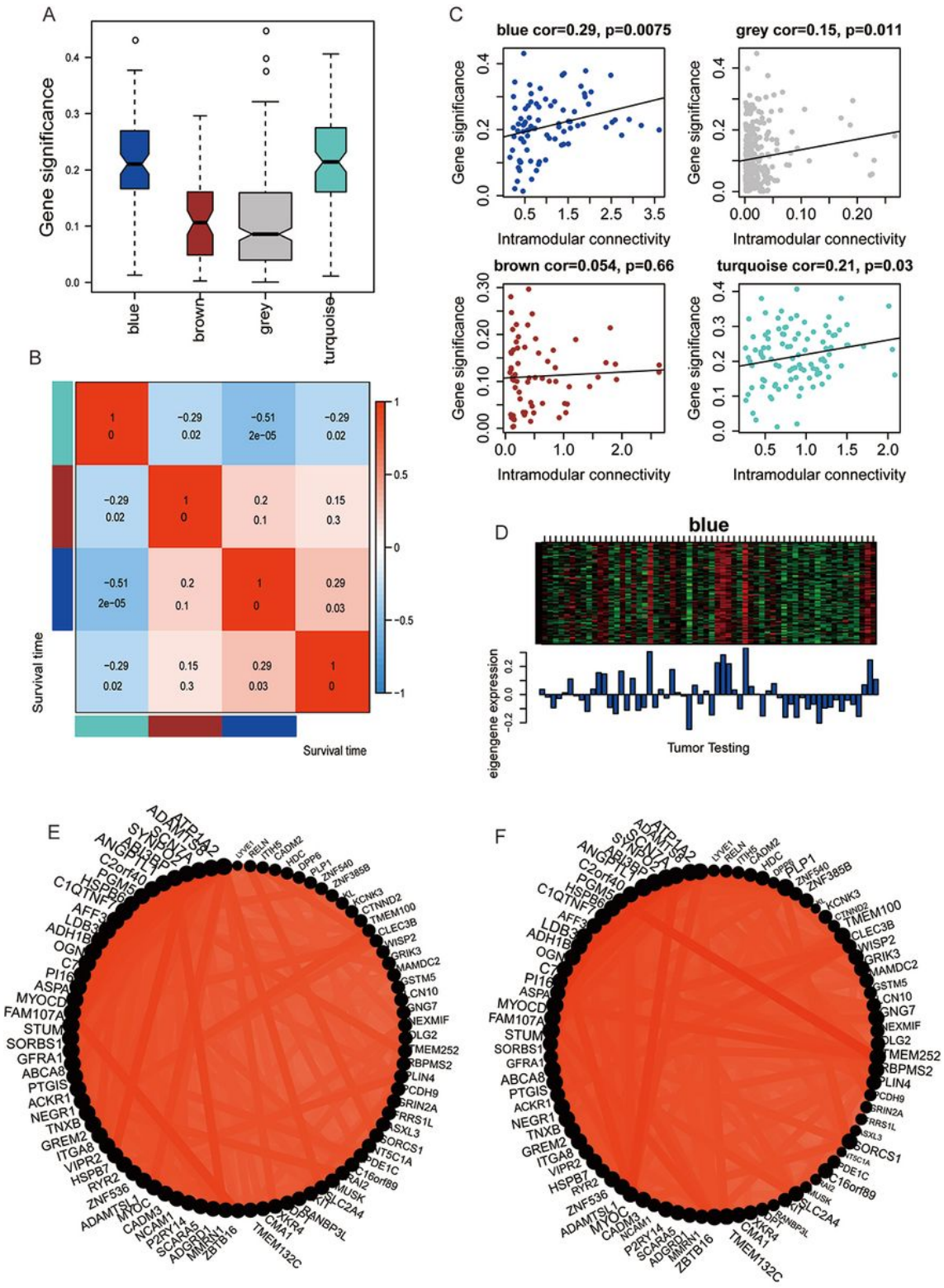


Figure 4

Module selection and visual diagnosis. (A) GS across the modules. (B) Association between module eigengenes and survival times. (C) Relation between GS and intramodular connectivity. (D) The relationships between eigengenes and the expression level of each gene. (E-F) Network plots of the blue module in the tumor test set and the tumor reference set.

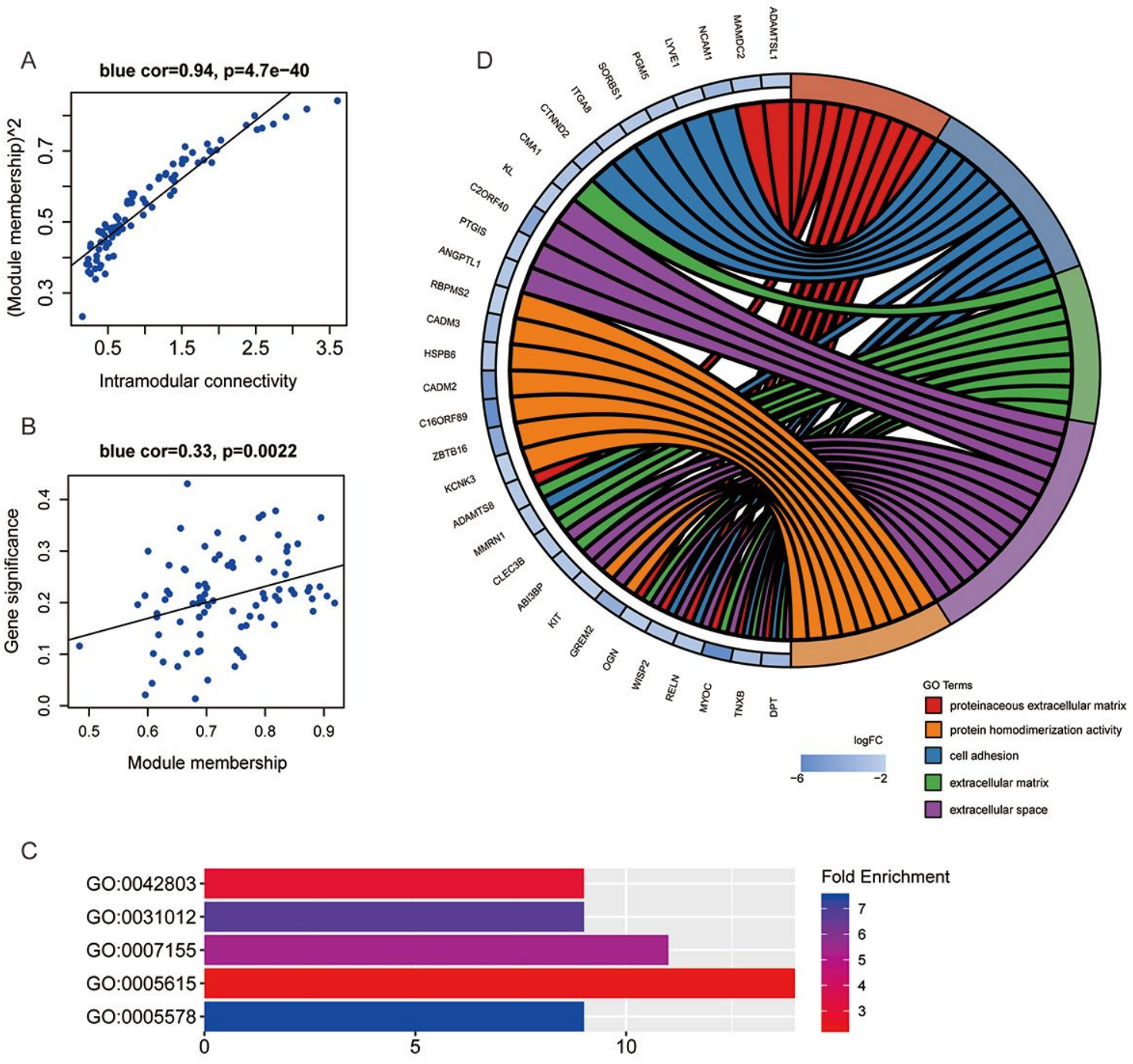


Figure 5

Hub genes selection and GO enrichment analysis. (A) Relationship between MM and intramodular connectivity. (B) Association between MM and GS. (C) Identification of important GO terms. (D) Genes related to the selected GO terms.

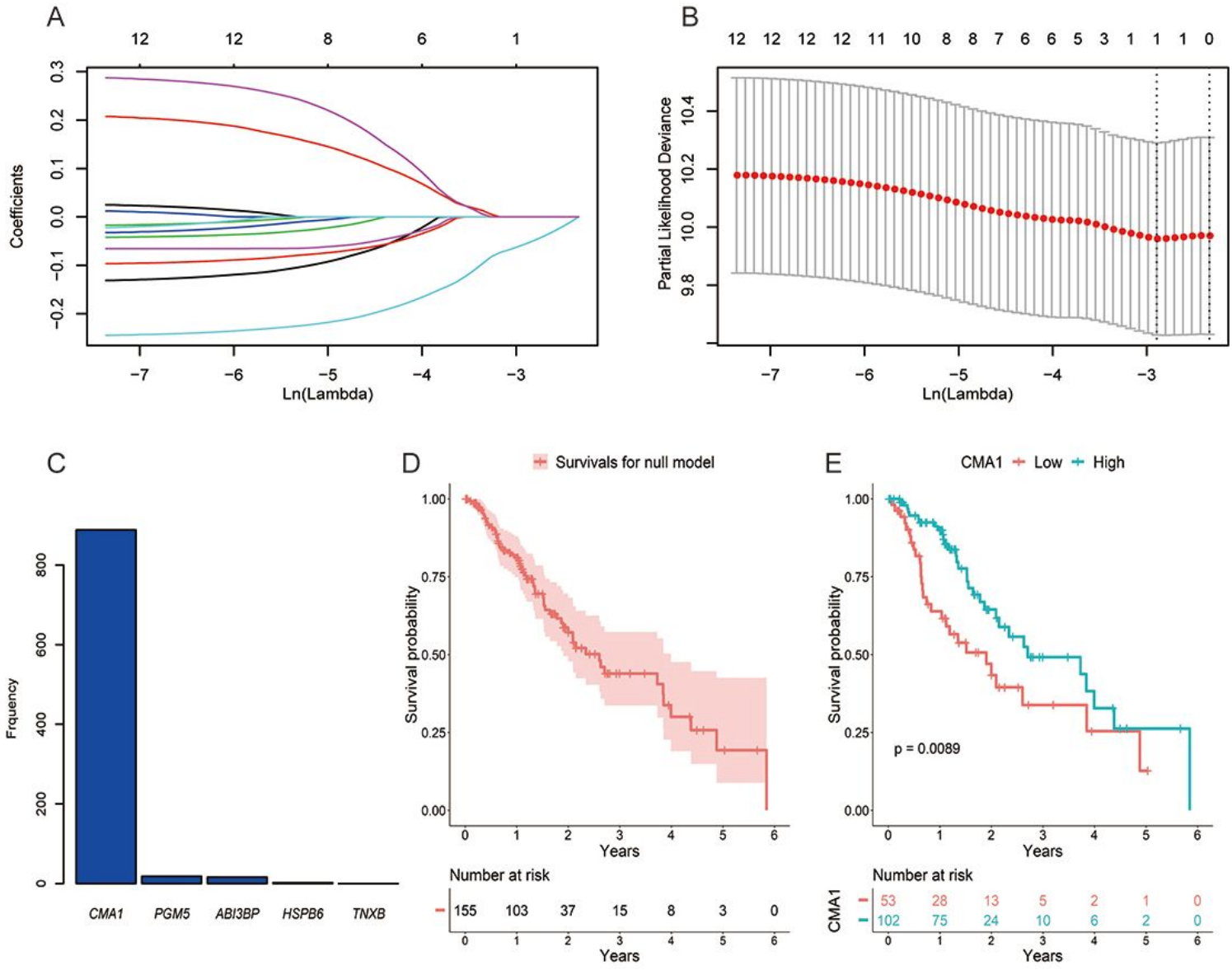


Figure 6

Biomarker selection and validation. (A) Coefficients for genes under different values of lambda in an experiment. (B) Number of genes that were kept in the model under the experiment. (C) Number of nonzero coefficients for the corresponding genes in 1000 experiments. (D) OS exploration. (E) Predictive ability of CMA1 in terms of OS.

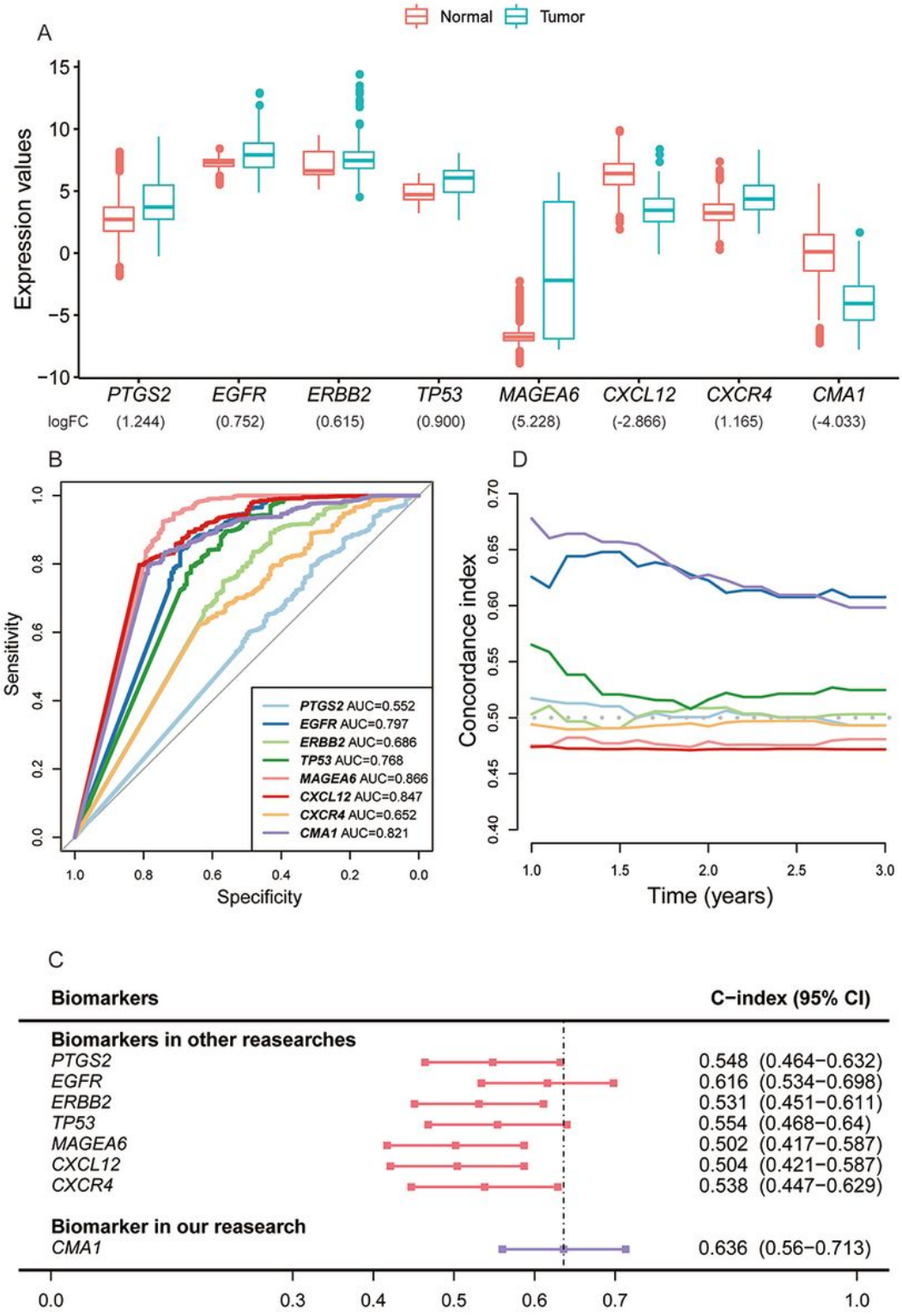


Figure 7

Comprehensive comparisons with biomarkers in other studies. (A) Diagnostic ability of the biomarkers, as shown in boxplots. (B) Diagnostic ability of biomarkers, as indicated by ROC curves derived from random forest models. (C) Prognostic ability of the biomarkers according to the C-index. (D) Prognostic ability of the biomarkers according to the C-index in the selected range of years.

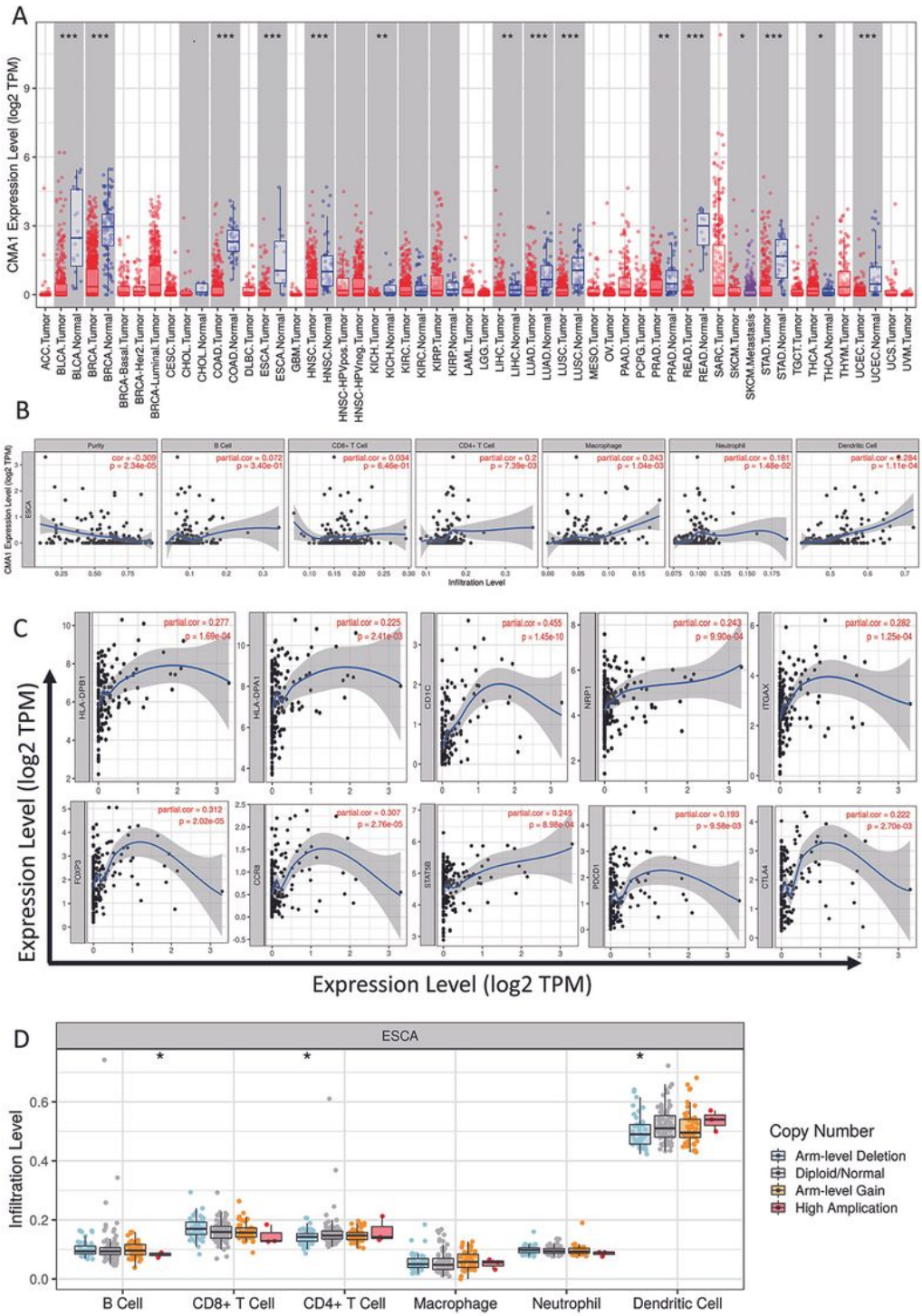


Figure 8

Association between CMA1 and immune system components. (A) Differential expression of CMA1 between tumor and adjacent normal tissues across all TCGA tumor types. The Wilcoxon test was employed (TPM, transcripts per million; *, p < 0.05; **, p < 0.01; ***, p < 0.001). (B) Correlation between CMA1 expression and tumor purity as well as immune cell infiltration levels. (C) Scatterplots showing the partial correlation between CMA1 expression and immune biomarkers, with adjustments for tumor purity

in EC, together with Spearman's rho value and estimated statistical significance. (D) Comparison of tumor infiltration levels in samples with different SCNAs for CMA1 in EC. Copy numbers were defined by GISTIC 2.0 and included deep deletion (-2), arm-level deletion (-1), diploid/normal (0), arm-level gain (1), and high amplification (2). Two-sided Wilcoxon rank-sum tests were employed to compare the infiltration level among samples with each copy number category and normal samples. (*, $p < 0.05$)

Supplementary Files

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

- Fig.1.jpg