The predicting roles of carcinoembryonic antigen and its underlying mechanism in the progression of coronavirus disease 2019

Runzhi Huang
Tongji University School of Medicine
Tong Meng
Tongji University School of Medicine
Qiongfang Zha
Shanghai Jiao Tong University School of Medicine Affiliated Renji Hospital
Kebin Cheng
Tongji University Affiliated Shanghai Pulmonary Hospital
Xin Zhou
Shanghai Jiaotong University First People's Hospital
Junhua Zheng
Shanghai Jiaotong University First People's Hospital
Dingyu Zhang
Wuhan Jinyintan Hospital
Ruilin Liu (18721881628@163.com)
Tongji University School of Medicine https://orcid.org/0000-0001-5498-9131

Research

Keywords: coronavirus disease 2019, carcinoembryonic antigen, carcinoembryonic antigen-related cell adhesion molecules, developing neutrophils, Type II pneumocyte

Posted Date: December 11th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-125433/v1

Abstract

Background

The coronavirus disease 2019 (COVID-19) has induced a worldwide pneumonia with a high infectivity and mortality. However, the predicting biomarkers and their potential mechanism in the progression of COVID-19 are not well known.

Objective

The aim of this study is to identify the candidate predictors of COVID-19 and investigate their underlying mechanism.

Methods

The retrospective study was conducted to identify the potential laboratory indicators with prognostic values of COVID-19 disease. Then, the prognostic nomogram was constructed to predict the overall survival of COVID-19 patients. Additionally, the scRNA-seq data of BALF and PBMCs from COVID-19 patients were downloaded to investigate the underlying mechanism of the most important prognostic indicators in lungs and peripherals, respectively.

Results

304 hospitalized adult COVID-19 patients in Wuhan Jinyintan Hospital were included in the retrospective study. CEA was the only laboratory indicator with significant difference in the univariate ($P < 0.001$) and multivariate analysis ($P = 0.021$). The scRNA-seq data of BALF and PBMCs from COVID-19 patients were downloaded to investigate the underlying mechanism of CEA in lungs and peripherals, respectively. The results revealed the potential roles of CEA were significantly distributed in Type II pneumocytes of BALF and developing neutrophils of PBMCs, participating in the progression of COVID-19 by regulating the cell-cell communication.

Conclusion

This study identifies the prognostic roles of CEA in COVID-19 patients and implies the potential roles of CEACAM8-CEACAM6 in the progression of COVID-19 by regulating the cell-cell communication of developing neutrophils and Type II pneumocyte.

Introduction

In December 2019, coronavirus disease 2019 (COVID-19) has been outbreaking in Wuhan China and rapidly spreads throughout the world inducing a worldwide panic[1]. The novel coronavirus was isolated from human airway epithelial cells and was named severe acute respiratory syndrome-related coronavirus 2 (SARS-CoV-2), which is highly infectious and induces a high fatality [2–5]. Nowadays, the underlying pathogenic mechanism of SARS-CoV-2 has been generally explored [6]. Similar to SARS-CoV, SARS-CoV-2 uses the receptors of angiotensin converting enzyme II (ACE2) for the entry and after receptor binding, while the spike (S) protein priming protease, such as cell surface transmembrane serine protease (TMPRSSs) and endosomal cathepsins, works in the membrane fusion [2, 7]. However, these proteases are not prognosis predictors of COVID-19 patients, which may be significantly associated with therapeutic decision-making.

Generally, patients' characteristics, nutritional status, clinical symptoms, comorbidities, inflammatory biomarkers and chest CT images are different in terms of patient outcome, however, whether these factors can serve as prognosis predictors for COVID-19 pneumonia is not clear. Regarding to chest CT images, consolidation, emphysema and residual healthy lung parenchyma are regarded as independent predictors in COVID-19 patients [8]. Additionally, high sensitivity C-reactive protein-prealbumin ratio (HsCPAR), high sensitivity C-reactive protein-albumin ratio (HsCAR) and low prognostic nutritional index (PNI) and the ratio of interleukin (IL)-6 to IL-10 were also reported to be related to the prognosis of COVID-19 patients [9, 10]. However, the potential mechanism of their predicting roles is unknown, neither is other candidate predictors.

In this study, in order to identify potential prognosis predictors for patients with COVID-19 pneumonia, we collected the demographic information, symptom, comorbidity, laboratory indexes, imaging, treatment, and prognosis of hospitalized adult COVID-19 patients. Based on the identified predictors, the prognostic nomogram was established to guide clinical decision-making. In addition, we also explore the underlying mechanism of candidate biomarkers based on single-cell transcriptomics of bronchoalveolar lavage fluid (BALF) from patients with or without COVID-19. This study will provide novel predictors and their potential mechanism in the infection and progression of COVID-19.

Materials And Methods

Patient selection and data extraction

This study was approved by the Ethics Committee of Jinyintan Hospital (KY-2020-58.01), followed the Standards for Reporting of Diagnostic Accuracy Studies Statement and Strengthening the Reporting of Observational Studies in Epidemiology (STROBE)[11, 12]. A total of 304 hospitalized adult COVID-19 patients diagnosed by Reverse Transcription-Polymerase Chain Reaction (RT-PCR) in Wuhan Jinyintan Hospital from January 1st 2020 to April 30th 2020 were included in the retrospective study. The exclusion criteria were: (1) Patients lower than 18 years old; (2) Non-hospitalized patients; (3) Patients with follow-up period less than 60 days; (4) Patients already at hospital admission due to other diseases; (5) Patients

whose survival time, endpoint (overall survival), demographic information, or treatment data were unknown; (6) Patients whose admission carcinoembryonic antigen (CEA) were unknown.

The clinical data in the study was retrieved from the electronic medical record system of Wuhan Jinyintan Hospital on initial admission, including variables of demographic information (age at diagnosis and gender), symptom (fever, cough, expectoration, shortness of breath and diarrhea), comorbidity (diabetes mellitus, hypertension, cardiovascular disease and cerebral infarction hypertension), and COVID-19 disease information (use of glucocorticoid, imaging score, nasal catheter, high flow oxygen intake, ventilation and disease stage). Additionally, laboratory indexes were also collected including CEA (ng/ml), albumin (g/l), hemoglobin (g/l), neutrophils ($\times 10^9$/l), lymphocytes ($\times 10^9$/l), C-reactive protein (CRP, mg/l), hypokalemia, hypocalcemia, hyponatremia, hyperkalemia and hypernatremia. As the endpoint, the survival time and overall survival status of each patient was retrieved.

Epidemiological statistical analysis

The retrospective study started with descriptive statistic: discontinuous variables were presented as percentages while continuous variables in normal distribution were described as mean ± standard deviation (SD) or else reported as median (range). Two statistic methods were applied to explore potential significant predictors. As initial parameter or non-parametric tests, Chi-Square test was used to compare the outcomes between discontinuous variables, and variance homogeneous and normal distributed continuous variables were compared by student t-test, otherwise, the Mann-Whitney U-test or Kruskal-Wallis H-test were used. Besides, Kaplan-Meier survival analysis was used to identify the prognostic value of each variable. Furthermore, predictors with statistical significance in both parameter or non-parametric tests and Kaplan-Meier survival analysis were selected to construct the multivariate Cox proportional hazards model. The nomogram was established based on the multivariate model to predict the prognosis of COVID-19 patients. Receiver operating characteristic (ROC) curve and calibration curve were drawn to evaluate the discrimination and calibration of the nomogram.

Processing of single-cell RNA-seq data

Single-cell RNA-sequence (scRNA-seq) data of COVID-19 patients' and healthy volunteer's bronchoalveolar lavage fluid (BALF, accession no. GSE145926)[13] and peripheral blood mononuclear cells (PBMCs, accession no. GSE150728)[14] were download from the Gene Expression Omnibus (GEO).

The initial data processing of single cell RNA-seq data started from the Cell Ranger Single Cell Software Suite 3.3.1 (http://10xgenomics.com/). The pair-ended reads fastq files were generated by the command named "cellranger mkfastq" and were trimmed to remove template switch oligo (TSO) sequence and poly-A tail sequence. Then, command of "cellranger count" was used to quantify the clean reads, aligned to the hg38 human transcriptome. The Seurat method was applied to integrated data analysis[15].

In terms of quality control (QC), genes with count greater than 1 and being expressed in at least 3 single cells were considered for further analysis. Cells with either fewer than 100,000 transcripts or fewer than 1,500 genes were filtered out.

In data processing, first, “vst” method was used to identify variable genes. Variable genes were the input as initial features for principal component analysis (PCA) [15]. Then, the principal components (PCs) with P values < 0.05 were filtered by the jackstraw analysis and were incorporated into further UMAP (Uniform Manifold Approximation and Projection) and t-distributed Stochastic Neighbor Embedding (t-SNE) to identify cell sub-clusters (resolution = 0.50)[16]. Only the genes with |log2 fold change (FC)| greater than 0.5 and false discovery rate (FDR) value < 0.05 were identified as differentially expressed genes (DEGs) among cell sub-clusters. The feature plots and violin plots were utilized to illustrate the distribution and expression of DEGs, respectively. Additionally, scMatch[17], singleR[18] and CellMarker[19] database were used as references to define each cluster. Cell trajectory and pseudo-time analysis was performed by monocle2[20]. Furthermore, 50 hallmark gene sets were retrieved from the Molecular Signatures Database (MSigDB) v7.1 (https://www.gsea-msigdb.org/gsea/msigdb/index.jsp) and Gene Set Variation Analysis (GSVA) was performed to absolutely quantify the activity of signaling pathway in each single cell [21, 22]. Furthermore, cellphoneDB was used to identify the cellular communication between pneumonocyte and immune cells[23].

Identification of the mechanism of abnormal CEA expression in COVID-19 patients

First of all, the distribution and expression of CEA-related genes (CRGs) including CEACAM1, CEACAM3, CEACAM4, CEACAM5, CEACAM6, CEACAM7, CEACAM8, CEACAM16, CEACAM18, CEACAM19, CEACAM20, CEACAM21, CEACAMP1, CEACAMP2, CEACAMP3, CEACAMP4, CEACAMP5, CEACAMP6, CEACAMP7, CEACAMP8, CEACAMP9, CEACAMP10, CEACAMP11 and CEACAM22P were visualized by feature plot and violin plot in BALF and PBMC scRNA-seq data. Then, co-expression (correlation) analysis were performed among CRGs, PRBPGs and 50 hallmark of gene sets to identify the potential downstream pathways. CellphoneDB algorithm was used to illuminate the cellular communication between cells with high CRGs expression and other cells. Besides, two data including scRNA-seq data of acute lung injury (ALI) mus musculus’s lung (GSE134383) and idiopathic pulmonary fibrosis (IPF) mus musculus’s lung (E-HCAD-14) were downloaded to evaluate the distribution and expression of CRGs, key receptor-ligand pair of cellular communication and potential downstream pathways[24-28].

Statistical analysis

Only p value of two-sided statistical testing lower than 0.05 was considered statistically significant. All statistical analysis was performed with R version 3.6.1 software (Institute for Statistics and Mathematics, Vienna, Austria; www.r-project.org).

Results

Patient characteristics and univariate analysis

A total of 304 hospitalized adult COVID-19 patients diagnosed by RT-PCR in Wuhan Jinyintan Hospital from January 16th 2020 to March 30th 2020 were included in this retrospective study. After fulfillment of all exclusion criteria, two patients with follow-up period less than 60 days and two patients whose admission CEA were unknown were excluded from the further analysis.

The baseline characteristics of 300 COVID-19 patients were described in Figure 1A. The cohort comprised 170 males and 130 females, with a median age of 63.0 (range, 21.0 - 90.0) years. After removing 17 of 28 laboratory indicators with missing values more than 20% of the sample size, the results of initial Kaplan-Meier survival analysis (Figure 1C-D) and parameter or non-parametric tests (Figure 1E) revealed that only five indicators (serum CEA, lymphocytes, neutrophils, CRP and albumin) were significantly associated with both the imaging score and prognosis of COVID-19 patients (Figure 1B).

Cox proportional hazards model and nomogram

CEA is the only laboratory indicator with significant difference in all the univariate and multivariate analysis. To identify the optimal cut off point of CEA, the cyclic log-rank test was performed. The results revealed that CEA = 7.3ng/ml was the optimal cut off point with the most significant P value in the log-rank test (Figure 2A-B). Then, 12 potential indicators showing prognostic values in Kaplan-Meier analysis were incorporated into the initial Cox proportional hazards models, along with two demographic information (age and gender). The final multivariable-models were constructed to confirm the effects of significant covariates in the initial models to the OS of COVID-19 patients (Figure 2C). Patients with lower CEA had better OS (HR, 0.57; 95% CI, 0.35 to 0.92; P = 0.021) in the multivariable model, suggesting CEA as prognostic indicator for COVID-19 patients independently.

The prognostic nomogram was constructed based on the multivariate Cox model including CEA to predict the 3-week and 5-week overall survival probability of COVID-19 patients (Figure 3A). The calibration curve and ROC curve (AUC = 0.776) suggested acceptable calibration and discrimination of the nomogram, respectively (Figure 3B; Figure S1A-B). Besides, the risk score (RS) was calculated by the formula generated by the multivariate Cox model. The scatter plot (Figure S1C) and risk curve (Figure S1D) of the model demonstrated the RS distribution based on risk score of each patient. Kaplan-Meier curve suggested the prognostic value of the RS (Figure 3C, P < 0.001). Besides, the residual distribution of the multivariate model was accessed by the residual plot (Figure S1E). Eventually, the RS was shown to be an independently prognostic indicator for COVID-19 patients in both univariate (HR = 34.215, 95%CI (17.827–65.687), P < 0.001, Figure 3D) and multivariate (HR = 1.281, 95%CI (1.214–1.353), P < 0.001, Figure 3E) Cox regression model corrected by demographics,.

Identification of the potential mechanism of CEA in COVID-19

The scRNA-seq data of bronchoalveolar lavage fluid (BALF) from three patients with moderate COVID-19 (C141, C142, C144), six patients with severe or critical infection (C143, C145, C146, C148, C149, C152), three healthy controls (C51, C52, C100) [13] were download from the GEO database. A UAMP analysis was performed in 63,010 cells in BALF and clearly identified 20 clusters and 11 cell types including B cell,

CD4+ T cell, CD8+ T cell, Dendritic cell, Macrophage, Monocyte, Natural killer cell, Neutrophil, T cell: gamma-delta, Type I pneumocyte, Type II pneumocyte (Figure 4A-B, Figure S2A-B). The expression levels and expression percentages of the marker genes in each cell type were displayed in Figure S2C and S2D, respectively. Except for macrophages and type I and type II pneumocytes, all other immune cells (B cell, CD4+ T cell, CD8+ T cell, Dendritic cell, Monocyte, Natural killer cell, Neutrophil and T cell: gamma-delta) were dominantly differentiated and chemotactic in the BALF of COVID-19 patients compared to healthy volunteer (Figure 4C). Furthermore, in terms of the expression and distribution of CRGs, CEACAM1, CEACAM3, CEACAM5, CEACAM6, CEACAM7, CEACAM8 and CEACAM21 were differentially expressed among moderate, severe/critical COVID-19 patients and healthy controls while CEACAM5, CEACAM6 were significantly localized in the type II pneumocytes of COVID-19 patients (Figure 4D-E). Especially, Figure 4F summarized the absolute quantification of 50 hallmark gene sets calculated the GSVA in type I and type II pneumocytes, suggesting that the interferon response and cell proliferation signaling pathways were significantly activated in type II pneumocytes highly expressing CRGs of COVID-19 patients. Besides, cell cycle analysis suggested that COVID-19 patients were more likely to have cells in the G2M and S stages (Figure S3A-B). Furthermore, cellphoneDB analysis illustrated that pneumocytes of COVID-19 patients communicated extensively with other immune cells through CRGs (Figure S3C).

Similarly, the scRNA-seq data of 94,448 PBMCs from six patients with moderate COVID-19 and six healthy volunteers were also download [14]. The UAMP analysis identified 18 clusters and 10 cell types including B cell, B cell Naïve, CD4+ T cell, CD8+ T cell, Macrophage-Monocyte, Myelocyte, Natural killer cell, Neutrophil, Plasma cell. Platelets (Figure 5A-B; Figure S4A-B). All types of immune cell were significantly differentiated and chemotactic in COVID-19 patients' PBMCs compared to healthy controls (Figure 5C). CEACAM1, CEACAM4, CEACAM6 and CEACAM8 were differentially expressed between PBMCs of COVID-19 patients and healthy controls, while CEACAM1, CEACAM6 and CEACAM8 were significantly localized in a novel cell subtype annotated as 'developing neutrophils', which was significantly differentiated and chemotactic only in COVID-19 patients with ARDS reported by Wilk, A.J., et al (Figure 5D-E) [14]. Additionally, dot plots summarized the results of Gene Ontology (GO) and the Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analysis. Based on GO analysis, the DEGs were associated with the neutrophil activation and degranulation (Figure 5F). According to KEGG analysis, the DEGs were related to protein processing in endoplasmic reticulum, phagosome, epstein-barr virus infection and tuberculosis (Figure 5F). Besides, cell cycle analysis suggested that the developing neutrophils in COVID-19 patients' PBMCs were all engaged in the G2M and S stages (Figure S4C-D).

The specific expressions of CRGs in COVID-19 patients

Due to the close correlation between CEA and ALI/IPF, we initially speculated that the poor prognosis of COVID-19 patients mediated by CEA might be related to ALI and IPF pathophysiologically. To validate this hypothesis, scRNA-seq data of ALI mouse lungs (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE134383) and IPF mouse lungs (https://www.ebi.ac.uk/gxa/sc/experiments/E-HCAD-14/results/tsne) were also downloaded to evaluate the distribution and expression of CRGs, key receptor-ligand pair of cellular communication and potential downstream pathways[24-28]. The UAMP analysis

identified 18 clusters and 6 cell types in ALI mouse lungs while there were no abnormal expressions of CRGs (Figure 6A-C). The interferon response and cell proliferation signaling pathways were not significantly activated in type II pneumocytes of ALI mouse lungs (Figure 6D). Similarly, abnormal expressions of CRGs were also not detected in 31 clusters and 10 cell types of IPF mouse lungs (Figure 6E-G). Besides, the heatmap of GSVA also showed that the interferon response and cell proliferation signaling pathways were not activated in type II pneumocytes of IPF mouse lungs (Figure 6H). Thus, the abnormal expressions of CRGs in COVID-19 patients were COVID-19-specific and not related to CEA involvement in ALI and IPF.

Protein-protein interaction (PPI) network of CRGs

String [29] database was used to construct the PPI network of CRGs, illustrating that several CRGs had direct PPIs with a variety of immune cell surface markers (Figure 7A-C). Besides, the protein expression levels of CRGs in normal lung samples of The Human Protein Atlas were also checked[30], showing that only CEACAM21 was stained moderately in pneumocytes while the proteins of CEACAM5, CEACAM6 and CEACAM8 were not detected in normal lung samples (Figure 7D). To sum up, we supposed that CEA can serve as a predictor for the poor prognosis of COVID-19 patients. In COVID-19, the developing neutrophils/neutrophil progenitors (highly expressed CEACAM8, ELANE and LYZ) can have the cross-talk with Type II pneumocyte (highly expressed CEACAM5 and CEACAM6) via CEACAM8-CEACAM6. This process may not only promote the differentiation of developing neutrophils and subsequently induce the ARDS, but also regulate the proliferation of Type II pneumocyte, which is the target cell of SARS-Cov-2.

Discussion

The COVID-19 has induced a worldwide pneumonia with a high infectivity and mortality[31]. The identification of predicting biomarkers may assist clinicians in decision-making, however, the candidate predictors are still unclear. In this study, we identified CEA as a potential biomarker for COVID-19 patients. To further explore the underlying mechanism, we used the single-cell transcriptomics of BALF from patients with or without COVID-19, along with the scRNA-seq data of ALI and IPF mouse lungs. We found that the developing neutrophils/neutrophil progenitors can have the cross-talk with Type II pneumocyte via CEACAM8-CEACAM6 in COVID-19 but not ALI and IPF. During this process, the differentiated developing neutrophils can promote ARDS and regulate the proliferation of Type II pneumocyte.

The predicting biomarkers are important for clinical decision-making; thus many efforts have been made to identify them in patients with COVID-19 pneumonia. Previously, the inflammatory biomarkers (IL-6, IL-8, IL-10 and ratio of IL-6 to IL-10), patients' characteristics (age) and chest CT images (consolidation, emphysema and residual healthy lung parenchyma) have been reported to predict the prognosis of COVID-19 patients[8, 32]. In addition, the novel method of machine learning is also used to precisely evaluate the COVID-19 pneumonia[33].

In this study, based on the clinical information of hospitalized adult COVID-19 patients, we identified CEA as prognostic indicator for COVID-19 patients independently. Additionally, the prognostic nomogram

including CEA was also constructed with a good applicability (AUC = 0.776). CEA, initially considered as an oncofetal protein, is an epithelial cell glycoprotein with a molecular mass of 180–200 kDa. At present, CEA is viewed as a normal epithelial molecule and its high expression is generally found in tumors[34].

In COVID-19, we also found that CEACAM8 is highly expressed in the developing neutrophils/neutrophil progenitors, while CEACAM5 and CEACAM6 are highly expressed in Type II pneumocyte. In humans, CEA and CEA subfamily members (CEACAMs) are cell surface heavily glycosylated proteins. In the bacterial or viral infection, CEA and CEACAM1 participate in the adherence of enteric bacteria to the apical membrane of colonic M cells in the human gut mucosa [35]. Besides, in the human respiratory tract, CEACAM1 and CEACAM5 increase the host susceptibility to bacterial infection upon viral challenge[36].

In COVID-19, the developing neutrophils were found to have cross-talk with Type II pneumocyte via CEACAM8-CEACAM6. Generally, CEACAM can be engaged in cell-cell contact and communication which may affect various signal transduction processes related to cell activation, differentiation, and apoptosis[37, 38]. In this process, CEACAM8-CEACAM6 regulation network may promote the differentiation of developing neutrophils, which are the newly annotated cells in patients with ARDS and represent neutrophils at various developmental stages [14]. The developing neutrophils may further lead to COVID-19 progression and induce the ARDS. Besides, it also regulates the proliferation of Type II pneumocyte, which highly expresses ACE2 and serves as the major infected cell type by SARS-CoV-2[39]. The increased amounts of Type II pneumocytes can be invaded by SARS-CoV-2, thereby leading to disease progression.

Conclusion

This study identifies the prognostic roles of CEA in COVID-19 patients and implies the potential roles of CEACAM8-CEACAM6 in the progression of COVID-19 by regulating the cell-cell communication of developing neutrophils and Type II pneumocyte. The abnormal expressions of CRGs in COVID-19 patients were COVID-19-specific and not related to CEA involvement in ALI and IPF.

Declarations

Ethical Approval and Consent to participate

This study was approved by the Ethics Committee of Jinyintan Hospital (KY-2020-58.01), followed the Standards for Reporting of Diagnostic Accuracy Studies Statement and Strengthening the Reporting of Observational Studies in Epidemiology (STROBE). The authors declare that there is no conflict of interests.

Consent for publication

Written informed consent for publication was obtained from all participants.

Availability of supporting data

The datasets generated and/or analysed during the current study are available in the Supplementary Material, Gene Expression Omnibus (GEO) (Accession no. GSE145926, GSE150728, GSE134383) and Single Cell Expression Atlas (Accession no. E-HCAD-14).

Competing interests

The authors declare that there is no conflict of interests.

Funding

This study was not supported by any funding.

Authors' contributions

Conception/design: Runzhi Huang, Tong Meng Qiongfang Zha, Kebin Cheng, Xin Zhou, Junhua Zheng, Dingyu Zhang and Ruilin Liu

Collection and/or assembly of data: Runzhi Huang, Tong Meng Qiongfang Zha, Kebin Cheng, Xin Zhou, Junhua Zheng, Dingyu Zhang and Ruilin Liu

Data analysis and interpretation: Runzhi Huang, Tong Meng Qiongfang Zha, Kebin Cheng, Xin Zhou, Junhua Zheng, Dingyu Zhang and Ruilin Liu

Manuscript writing: Runzhi Huang, Tong Meng Qiongfang Zha, Kebin Cheng, Xin Zhou, Junhua Zheng, Dingyu Zhang and Ruilin Liu

Final approval of manuscript: Runzhi Huang, Tong Meng Qiongfang Zha, Kebin Cheng, Xin Zhou, Junhua Zheng, Dingyu Zhang and Ruilin Liu

Acknowledgments

We thank the Gene Expression Omnibus (GEO) (Accession no. GSE145926, GSE150728, GSE134383) and Single Cell Expression Atlas (Accession no. E-HCAD-14) team for using their data.

References

1. Jin Y, Yang H, Ji W, Wu W, Chen S, Zhang W, Duan G. Virology, Epidemiology, Pathogenesis, and Control of COVID-19. *Viruses* 2020: 12(4).
2. Zhou P, Yang XL, Wang XG, Hu B, Zhang L, Zhang W, Si HR, Zhu Y, Li B, Huang CL, Chen HD, Chen J, Luo Y, Guo H, Jiang RD, Liu MQ, Chen Y, Shen XR, Wang X, Zheng XS, Zhao K, Chen QJ, Deng F, Liu LL, Yan B, Zhan FX, Wang YY, Xiao GF, Shi ZL. A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* 2020: 579(7798): 270-273.

3. Guan WJ, Ni ZY, Hu Y, Liang WH, Ou CQ, He JX, Liu L, Shan H, Lei CL, Hui DSC, Du B, Li LJ, Zeng G, Yuen KY, Chen RC, Tang CL, Wang T, Chen PY, Xiang J, Li SY, Wang JL, Liang ZJ, Peng YX, Wei L, Liu Y, Hu YH, Peng P, Wang JM, Liu JY, Chen Z, Li G, Zheng ZJ, Qiu SQ, Luo J, Ye CJ, Zhu SY, Zhong NS. Clinical Characteristics of Coronavirus Disease 2019 in China. *The New England journal of medicine* 2020.
4. Chen N, Zhou M, Dong X, Qu J, Gong F, Han Y, Qiu Y, Wang J, Liu Y, Wei Y, Xia J, Yu T, Zhang X, Zhang L. Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study. *Lancet* 2020: 395(10223): 507-513.
5. Huang C, Wang Y, Li X, Ren L, Zhao J, Hu Y, Zhang L, Fan G, Xu J, Gu X, Cheng Z, Yu T, Xia J, Wei Y, Wu W, Xie X, Yin W, Li H, Liu M, Xiao Y, Gao H, Guo L, Xie J, Wang G, Jiang R, Gao Z, Jin Q, Wang J, Cao B. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet* 2020: 395(10223): 497-506.
6. Zhang. H, Kang. Z, Gong H, Xu D, Wang J, Li Z, Li Z, Cui X, Xiao J, Zhan J, Meng T, Zhou W, Liu J, Xu H. Digestive system is a potential route of COVID-19: an analysis of single-cell coexpression pattern of key proteins in viral entry process. *Gut* 2020: 69(6): 1010-1018.
7. Wrapp D, Wang N, Corbett KS, Goldsmith JA, Hsieh CL, Abiona O, Graham BS, McLellan JS. Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. *Science (New York, NY)* 2020: 367(6483): 1260-1263.
8. Salvatore C, Roberta F, Angela L, Cesare P, Alfredo C, Giuliano G, Giulio L, Giuliana G, Maria RG, Paola BM, Fabrizio U, Roberta G, Beatrice F, Vittorio M. Clinical and laboratory data, radiological structured report findings and quantitative evaluation of lung involvement on baseline chest CT in COVID-19 patients to predict prognosis. *La Radiologia medica* 2020.
9. Xue G, Gan X, Wu Z, Xie D, Xiong Y, Hua L, Zhou B, Zhou N, Xiang J, Li J. Novel serological biomarkers for inflammation in predicting disease severity in patients with COVID-19. *International immunopharmacology* 2020: 89(Pt A): 107065.
10. McElvaney OJ, Hobbs BD, Qiao D, McElvaney OF, Moll M, McEvoy NL, Clarke J, O'Connor E, Walsh S, Cho MH, Curley GF, McElvaney NG. A linear prognostic score based on the ratio of interleukin-6 to interleukin-10 predicts outcomes in COVID-19. *EBioMedicine* 2020: 61: 103026.
11. Lachat C, Hawwash D, Ocké MC, Berg C, Forsum E, Hörnell A, Larsson C, Sonestedt E, Wirfält E, Åkesson A, Kolsteren P, Byrnes G, De Keyzer W, Van Camp J, Cade JE, Slimani N, Cevallos M, Egger M, Huybrechts I. Strengthening the Reporting of Observational Studies in Epidemiology-Nutritional Epidemiology (STROBE-nut): An Extension of the STROBE Statement. *PLoS medicine* 2016: 13(6): e1002036.
12. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, Lijmer JG, Moher D, Rennie D, de Vet HC. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. *Clinical chemistry and laboratory medicine* 2003: 41(1): 68-73.
13. Liao M, Liu Y, Yuan J, Wen Y, Xu G, Zhao J, Cheng L, Li J, Wang X, Wang F, Liu L, Amit I, Zhang S, Zhang Z. Single-cell landscape of bronchoalveolar immune cells in patients with COVID-19. *Nature*

medicine 2020: 26(6): 842-844.

14. Wilk AJ, Rustagi A, Zhao NQ, Roque J, Martínez-Colón GJ, McKechnie JL, Ivison GT, Ranganath T, Vergara R, Hollis T, Simpson LJ, Grant P, Subramanian A, Rogers AJ, Blish CA. A single-cell atlas of the peripheral immune response in patients with severe COVID-19. *Nature medicine* 2020.
15. Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature biotechnology* 2018: 36(5): 411-420.
16. Chung NC, Storey JD. Statistical significance of variables driving systematic variation in high-dimensional data. *Bioinformatics (Oxford, England)* 2015: 31(4): 545-554.
17. Hou R, Denisenko E, Forrest ARR. scMatch: a single-cell gene expression profile annotation tool using reference datasets. *Bioinformatics (Oxford, England)* 2019: 35(22): 4688-4695.
18. Aran D, Looney AP, Liu L, Wu E, Fong V, Hsu A, Chak S, Naikawadi RP, Wolters PJ, Abate AR, Butte AJ, Bhattacharya M. Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nature immunology* 2019: 20(2): 163-172.
19. Zhang X, Lan Y, Xu J, Quan F, Zhao E, Deng C, Luo T, Xu L, Liao G, Yan M, Ping Y, Li F, Shi A, Bai J, Zhao T, Li X, Xiao Y. CellMarker: a manually curated resource of cell markers in human and mouse. *Nucleic acids research* 2019: 47(D1): D721-d728.
20. Qiu X, Mao Q, Tang Y, Wang L, Chawla R, Pliner HA, Trapnell C. Reversed graph embedding resolves complex single-cell trajectories. *Nature methods* 2017: 14(10): 979-982.
21. Hänzelmann S, Castelo R, Guinney J. GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC bioinformatics* 2013: 14: 7.
22. Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, Tamayo P. The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell systems* 2015: 1(6): 417-425.
23. Efremova M, Vento-Tormo M, Teichmann SA, Vento-Tormo R. CellPhoneDB: inferring cell-cell communication from combined expression of multi-subunit ligand-receptor complexes. *Nature protocols* 2020: 15(4): 1484-1506.
24. Peyser R, MacDonnell S, Gao Y, Cheng L, Kim Y, Kaplan T, Ruan Q, Wei Y, Ni M, Adler C, Zhang W, Devalaraja-Narashimha K, Grindley J, Halasz G, Morton L. Defining the Activated Fibroblast Population in Lung Fibrosis Using Single-Cell Sequencing. *American journal of respiratory cell and molecular biology* 2019: 61(1): 74-85.
25. Lin SE, Barrette AM, Chapin C, Gonzales LW, Gonzalez RF, Dobbs LG, Ballard PL. Expression of human carcinoembryonic antigen-related cell adhesion molecule 6 and alveolar progenitor cells in normal and injured lungs of transgenic mice. *Physiological reports* 2015: 3(12).
26. Fahim A, Crooks MG, Wilmot R, Campbell AP, Morice AH, Hart SP. Serum carcinoembryonic antigen correlates with severity of idiopathic pulmonary fibrosis. *Respirology (Carlton, Vic)* 2012: 17(8): 1247-1252.
27. Ueno F, Kitaguchi Y, Shiina T, Asaka S, Miura K, Yasuo M, Wada Y, Yoshizawa A, Hanaoka M. The Preoperative Composite Physiologic Index May Predict Mortality in Lung Cancer Patients with

Combined Pulmonary Fibrosis and Emphysema. *Respiration; international review of thoracic diseases* 2017: 94(2): 198-206.

28. Yu WS, Lee JG, Paik HC, Kim SJ, Lee S, Kim SY, Park MS, Haam S. Carcinoembryonic antigen predicts waitlist mortality in lung transplant candidates with idiopathic pulmonary fibrosis. *European journal of cardio-thoracic surgery : official journal of the European Association for Cardio-thoracic Surgery* 2018: 54(5): 847-852.
29. Snel B, Lehmann G, Bork P, Huynen MA. STRING: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene. *Nucleic acids research* 2000: 28(18): 3442-3444.
30. Uhlén M, Fagerberg L, Hallström BM, Lindskog C, Oksvold P, Mardinoglu A, Sivertsson Å, Kampf C, Sjöstedt E, Asplund A, Olsson I, Edlund K, Lundberg E, Navani S, Szigyarto CA, Odeberg J, Djureinovic D, Takanen JO, Hober S, Alm T, Edqvist PH, Berling H, Tegel H, Mulder J, Rockberg J, Nilsson P, Schwenk JM, Hamsten M, von Feilitzen K, Forsberg M, Persson L, Johansson F, Zwahlen M, von Heijne G, Nielsen J, Pontén F. Proteomics. Tissue-based map of the human proteome. *Science (New York, NY)* 2015: 347(6220): 1260419.
31. Mafham MM, Spata E, Goldacre R, Gair D, Curnow P, Bray M, Hollings S, Roebuck C, Gale CP, Mamas MA, Deanfield JE, de Belder MA, Luescher TF, Denwood T, Landray MJ, Emberson JR, Collins R, Morris EJA, Casadei B, Baigent C. COVID-19 pandemic and admission rates for and management of acute coronary syndromes in England. *Lancet* 2020: 396(10248): 381-389.
32. Nagant C, Ponthieux F, Smet J, Dauby N, Doyen V, Besse-Hammer T, De Bels D, Maillart E, Corazza F. A score combining early detection of cytokines accurately predicts COVID-19 severity and intensive care unit transfer. *International journal of infectious diseases : IJID : official publication of the International Society for Infectious Diseases* 2020.
33. Pan P, Li Y, Xiao Y, Han B, Su M, Li Y, Zhang S, Jiang D, Chen X, Zhou F, Ma L, Bao P, Su L, Xie L. Prognostic Assessment of COVID-19 in ICU by Machine Learning Methods: A Retrospective Study. *Journal of medical Internet research* 2020.
34. Duffy MJ. Carcinoembryonic antigen as a marker for colorectal cancer: is it clinically useful? *Clinical chemistry* 2001: 47(4): 624-630.
35. Baranov V, Hammarström S. Carcinoembryonic antigen (CEA) and CEA-related cell adhesion molecule 1 (CEACAM1), apically expressed on human colonic M cells, are potential receptors for microbial adhesion. *Histochemistry and cell biology* 2004: 121(2): 83-89.
36. Klaile E, Klassert TE, Scheffrahn I, Müller MM, Heinrich A, Heyl KA, Dienemann H, Grünewald C, Bals R, Singer BB, Slevogt H. Carcinoembryonic antigen (CEA)-related cell adhesion molecules are co-expressed in the human lung and their expression can be modulated in bronchial epithelial cells by non-typable Haemophilus influenzae, Moraxella catarrhalis, TLR3, and type I and II interferons. *Respiratory research* 2013: 14(1): 85.
37. Gray-Owen SD, Blumberg RS. CEACAM1: contact-dependent control of immunity. *Nature reviews Immunology* 2006: 6(6): 433-446.

38. Khairnar V, Duhan V, Maney SK, Honke N, Shaabani N, Pandyra AA, Seifert M, Pozdeev V, Xu HC, Sharma P, Baldin F, Marquardsen F, Merches K, Lang E, Kirschning C, Westendorf AM, Häussinger D, Lang F, Dittmer U, Küppers R, Recher M, Hardt C, Scheffrahn I, Beauchemin N, Göthert JR, Singer BB, Lang PA, Lang KS. CEACAM1 induces B-cell survival and is essential for protective antiviral antibody production. *Nature communications* 2015: 6: 6217.
39. Ziegler CGK, Allon SJ, Nyquist SK, Mbano IM, Miao VN, Tzouanas CN, Cao Y, Yousif AS, Bals J, Hauser BM, Feldman J, Muus C, Wadsworth MH, 2nd, Kazer SW, Hughes TK, Doran B, Gatter GJ, Vukovic M, Taliaferro F, Mead BE, Guo Z, Wang JP, Gras D, Plaisant M, Ansari M, Angelidis I, Adler H, Sucre JMS, Taylor CJ, Lin B, Waghray A, Mitsialis V, Dwyer DF, Buchheit KM, Boyce JA, Barrett NA, Laidlaw TM, Carroll SL, Colonna L, Tkachev V, Peterson CW, Yu A, Zheng HB, Gideon HP, Winchell CG, Lin PL, Bingle CD, Snapper SB, Kropski JA, Theis FJ, Schiller HB, Zaragosi LE, Barbry P, Leslie A, Kiem HP, Flynn JL, Fortune SM, Berger B, Finberg RW, Kean LS, Garber M, Schmidt AG, Lingwood D, Shalek AK, Ordovas-Montanes J. SARS-CoV-2 Receptor ACE2 Is an Interferon-Stimulated Gene in Human Airway Epithelial Cells and Is Detected in Specific Cell Subsets across Tissues. *Cell* 2020: 181(5): 1016-1035.e1019.

Figures

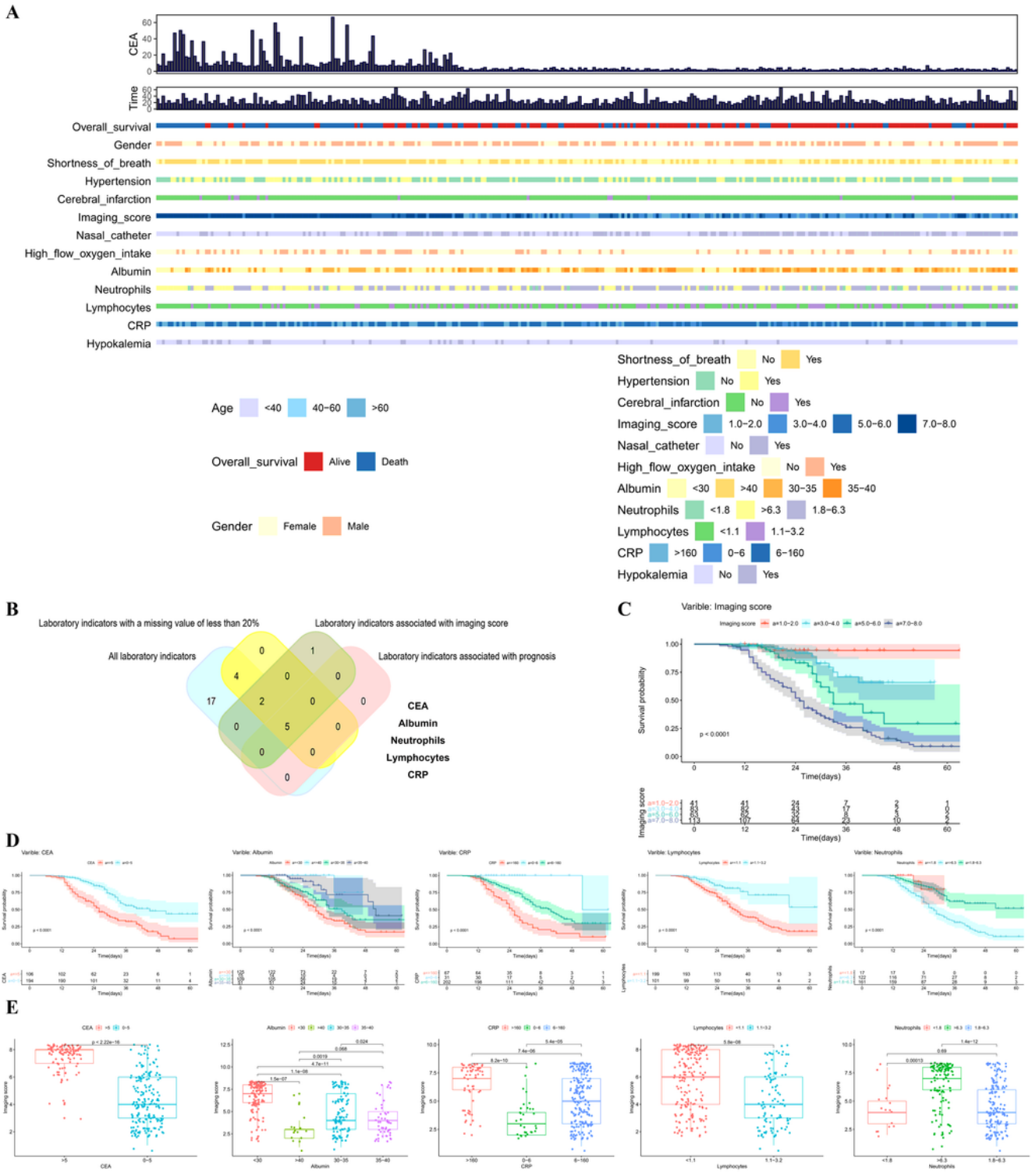

Figure 1

Patient characteristics and univariate analysis. The baseline characteristics of 300 COVID-19 patients were described in Figure 1A (A). The cohort comprised 170 males and 130 females, with a median age of 63.0 (range, 21.0 – 90.0) years. After removing 17 of the 28 laboratory indicators with missing values more than 20% of the sample size, the results of initial Kaplan-Meier survival analysis (C-D) and parameter or non-parametric tests (E) suggested that only five (serum CEA, lymphocytes, neutrophils, CRP

and albumin) indicators were significantly associated with both imaging score and prognosis COVID-19 patients (B).

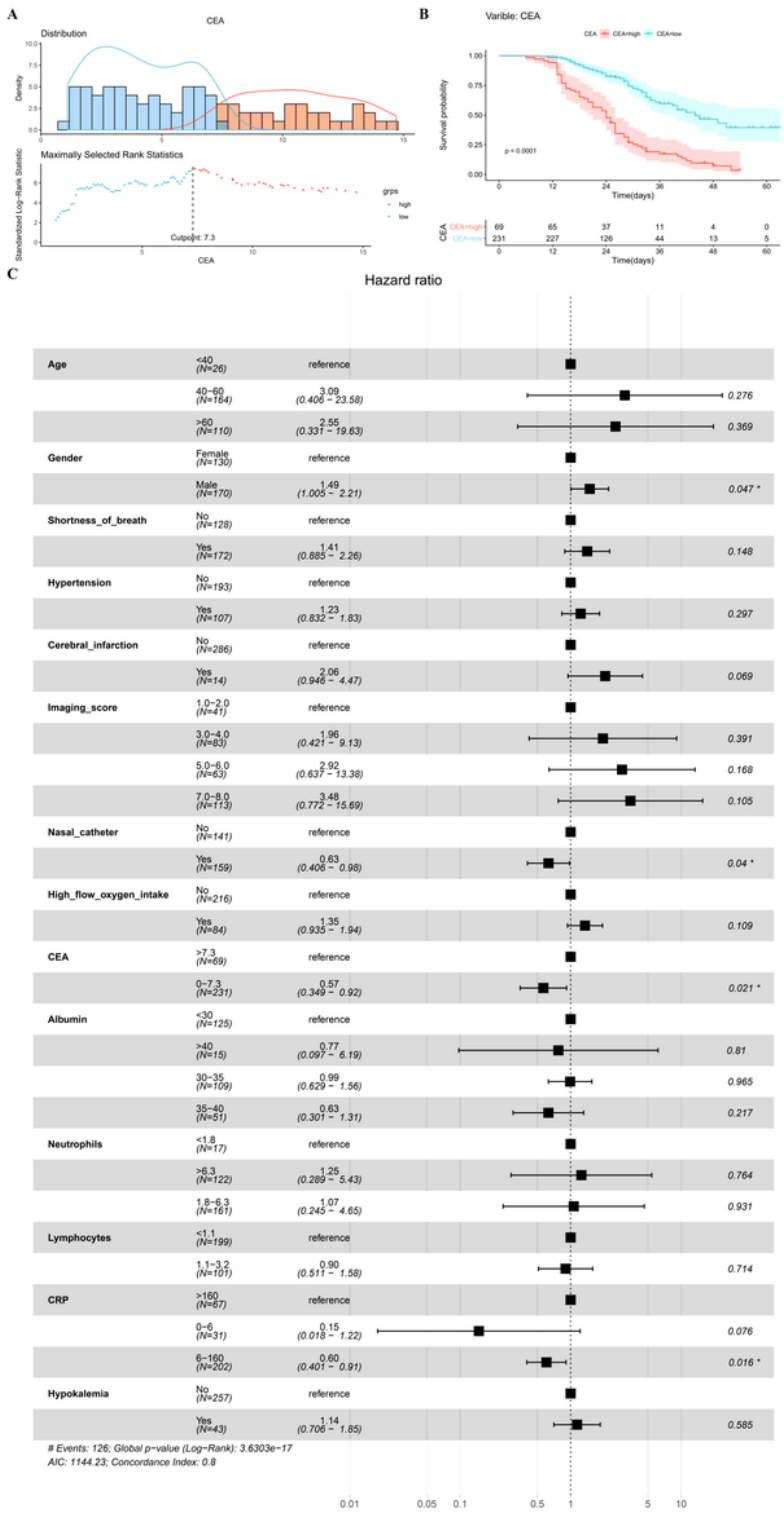

Figure 2

Cox proportional hazards model. CEA is the only laboratory indicator with significant results in all univariate and multivariate analyses. To identify the optimal cut off point of CEA, the cyclic log-rank test was performed. And the results showed that CEA = 7.3ng/ml was the optimal cut off point with the most

significant P value in log-rank test (A-B). Then, 12 potential significant indicators (showing prognostic values in Kaplan-Meier analysis) and two demographic information (age and gender) were incorporated into the initial Cox proportional hazards models, and the final multivariable-models were constructed to confirm the effects of significant covariates in the initial models to the OS of COVID-19 patients (C). Patients with lower CEA had better OS (HR, 0.57; 95% CI, 0.35 to 0.92; P = 0.021) in multivariable-model, which suggested that CEA independently prognostic indicator for COVID-19 patients.

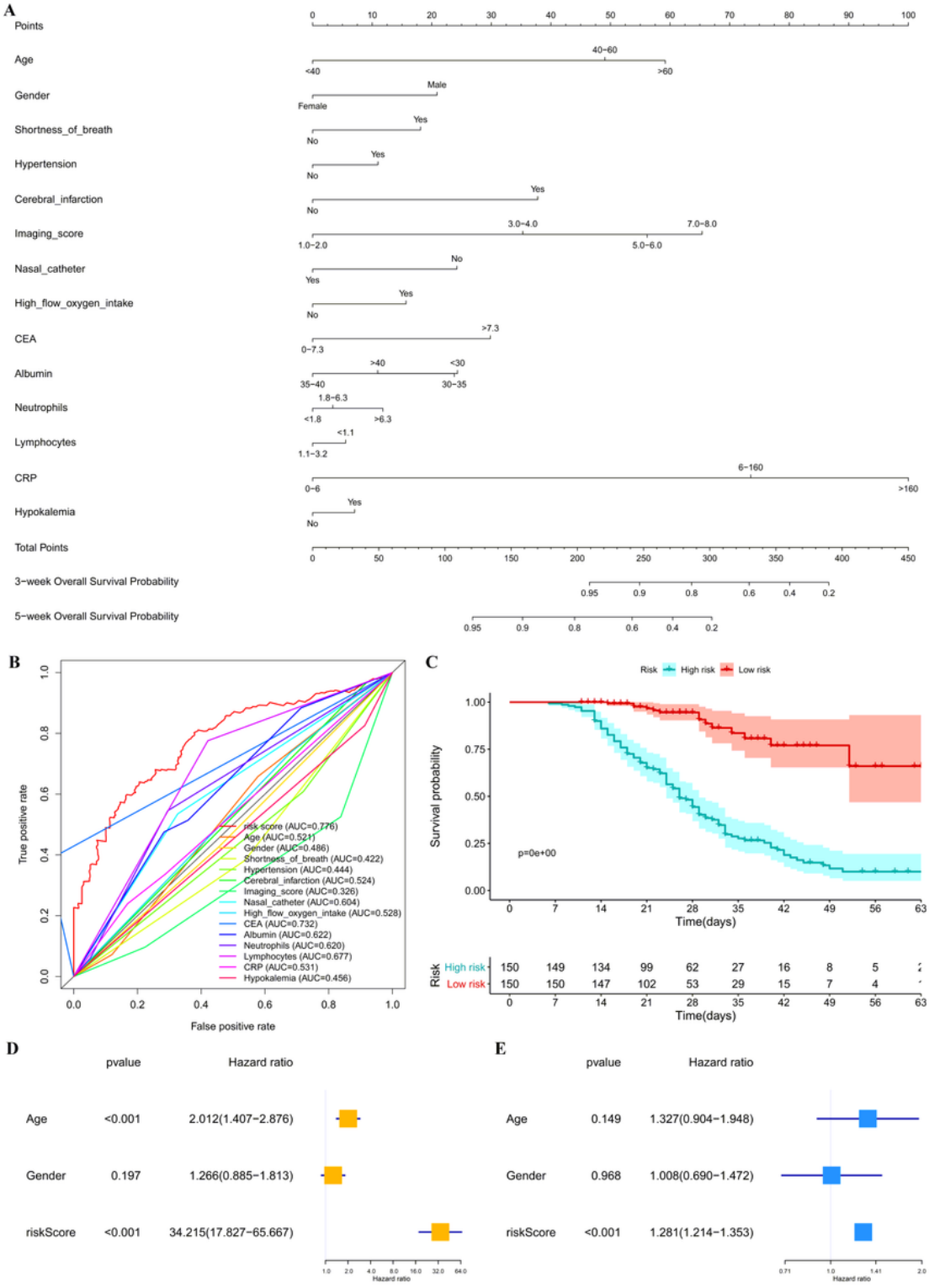

Figure 3

Construction and model diagnosis of prognostic nomogram. The prognostic nomogram was constructed based on the multivariate Cox model including CEA, which could predict the 3-week and 5-week overall survival probability of COVID-19 patients (A). The ROC curve (AUC = 0.776) suggested acceptable discrimination of the nomogram (B). Besides, the risk score (RS) was calculated by the formula generated by the multivariate Cox model. Kaplan-Meier curve suggested the prognostic value of the RS (C, $P < 0.001$). Eventually, in univariate (HR = 34.215, 95%CI (17.827–65.687), $P < 0.001$) (D) and multivariate (HR = 1.281, 95%CI (1.214–1.353), $P < 0.001$) (E) Cox regression model corrected by demographics, the RS was shown to be an independently prognostic indicator for COVID-19 patients.

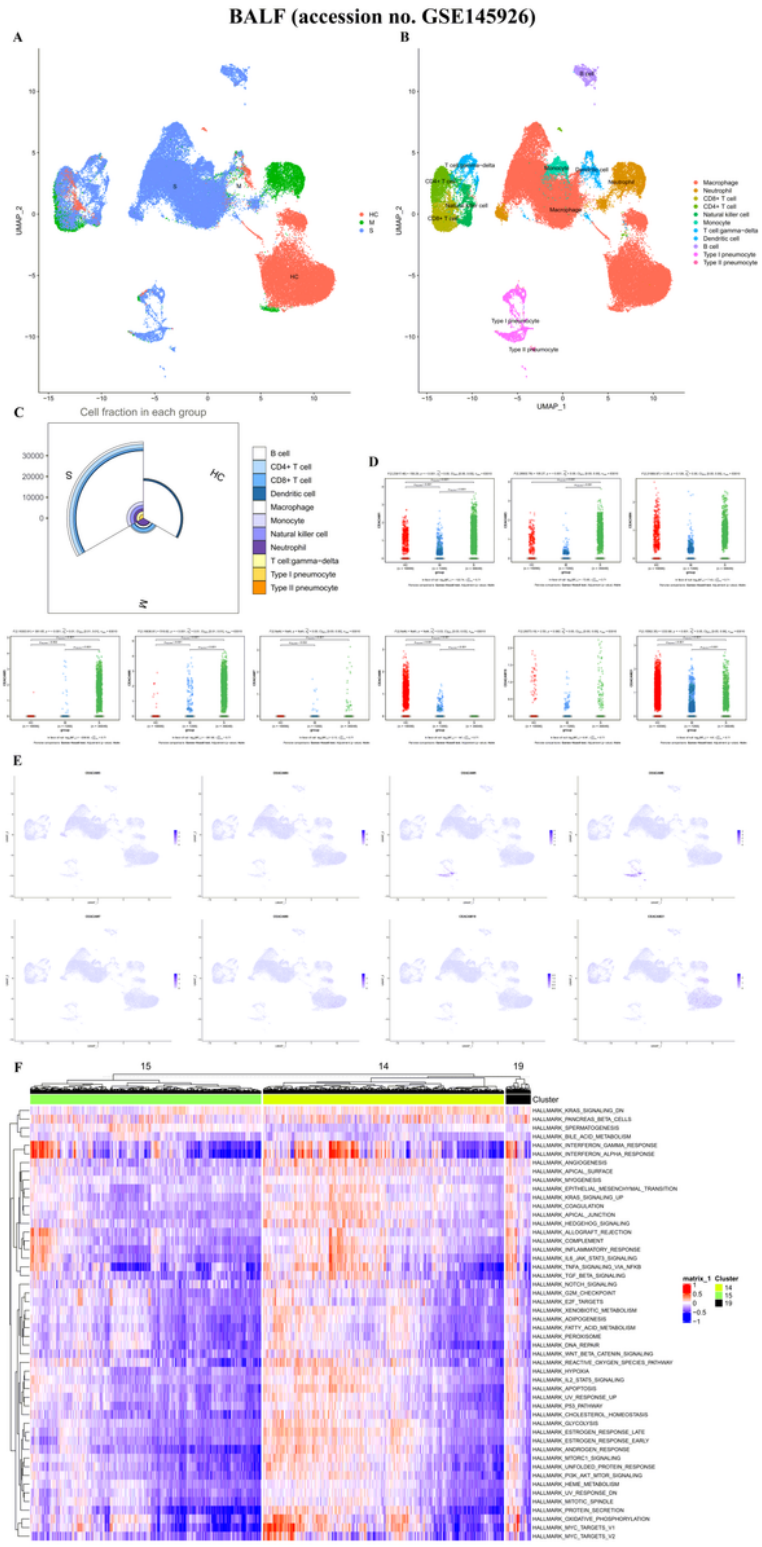

Figure 4

Identification of the mechanism of abnormal CEA expression in COVID-19 patients' and healthy volunteers' bronchoalveolar lavage fluid (BALF) scRNA-seq data of bronchoalveolar lavage fluid (BALF) from three patients with moderate COVID-19 (C141, C142, C144), six patients with severe or critical infection (C143, C145, C146, C148, C149, C152), three healthy controls (C51, C52, C100) (accession no. GSE145926) were download from the GEO database. A UAMP analysis was performed in 63,010 cells in

BALF and clearly identified 20 clusters and 11 cell types (B cell, CD4+ T cell, CD8+ T cell, Dendritic cell, Macrophage, Monocyte, Natural killer cell, Neutrophil, T cell: gamma–delta, Type I pneumocyte, Type II pneumocyte) (A-B). All other immune cells (B cell, CD4+ T cell, CD8+ T cell, Dendritic cell, Monocyte, Natural killer cell, Neutrophil and T cell: gamma–delta) except for macrophages and type I and type II pneumocytes were dominantly differentiated and chemotactic in COVID-19 patients' BALF compared to healthy volunteer's BALF (C). Furthermore, in terms of the expression and distribution of CRGs, CEACAM1, CEACAM3, CEACAM5, CEACAM6, CEACAM7, CEACAM8 and CEACAM21 were differentially expressing among moderate, severe/critical COVID-19 patients and healthy controls while CEACAM5, CEACAM6 were significantly localized in the type II pneumocytes of COVID-19 patients (D-E). Especially, figure 4F summarized the absolute quantification of 50 hallmark gene sets calculated the GSVA in type I and type II pneumocytes, suggesting that the interferon response and cell proliferation signaling pathways were significantly activated in type II pneumocytes highly expressing CRGs of COVID-19 patients (F).

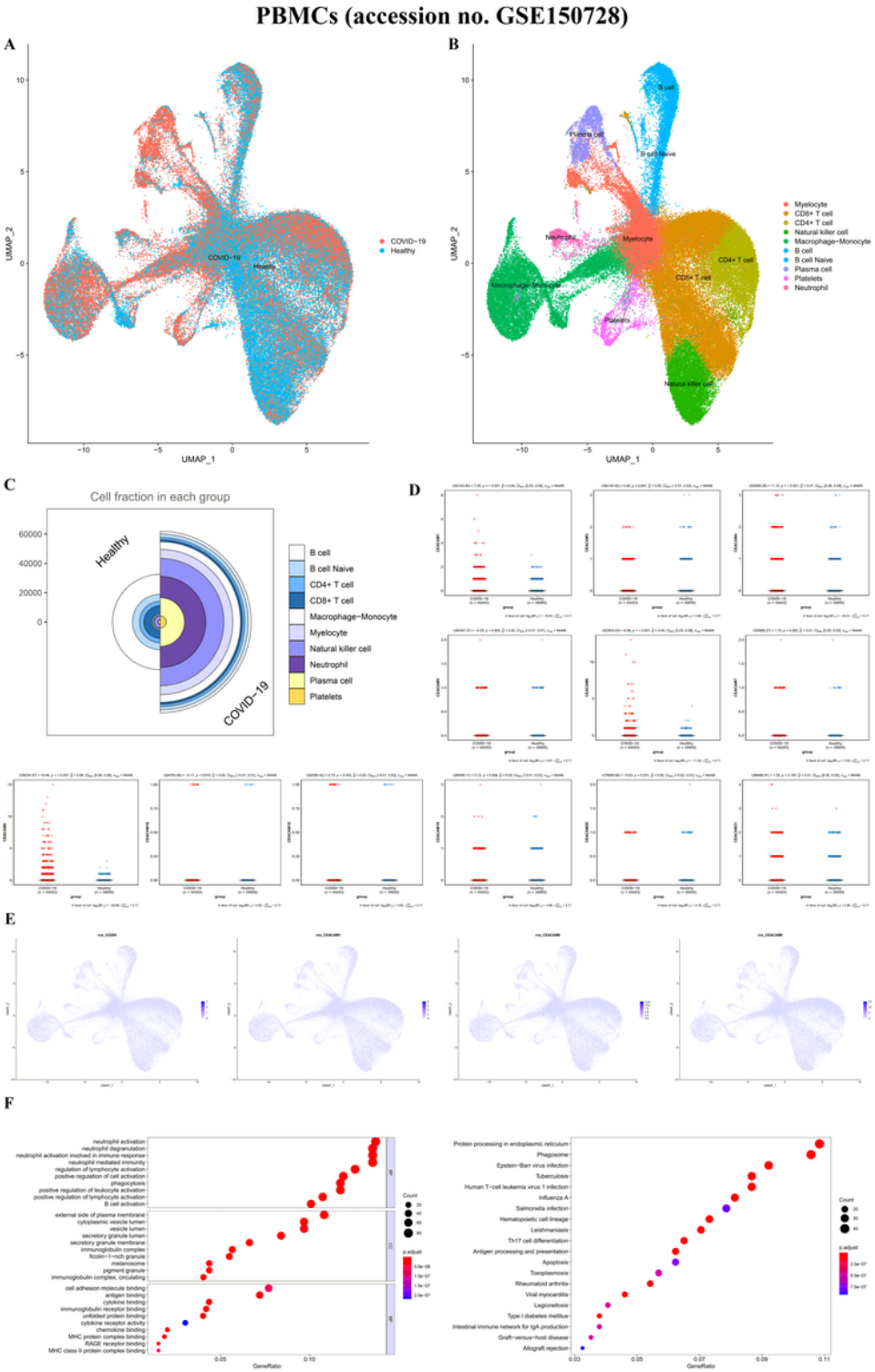

Figure 5

Identification of the mechanism of abnormal CEA expression in COVID-19 patients' and healthy volunteers' PBMCs scRNA-seq data of 94,448 PBMCs from six patients with moderate COVID-19 and six healthy volunteers were download from the GEO database (accession no. GSE150728). The UAMP analysis identified 18 clusters and 10 cell types (B cell, B cell Naïve, CD4+ T cell, CD8+ T cell, Macrophage−Monocyte, Myelocyte, Natural killer cell, Neutrophil, Plasma cell. Platelets) (A-B). All types of

immune cell were significantly differentiated and chemotactic in COVID-19 patients' PBMCs compared to healthy controls (C). What's more, CEACAM1, CEACAM4, CEACAM6 and CEACAM8 were differentially expressing between PBMCs of COVID-19 patients and healthy controls while CEACAM1, CEACAM6 and CEACAM8 were significantly localized in a novel cell subtype annotated as 'developing neutrophils', which was significantly differentiated and chemotactic only in COVID-19 patients with ARDS reported by Wilk, A.J., et al (D-E). Additionally, dot plots in figure 5F summarized the results of Gene Ontology (GO) and the Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analysis of the DEGs of the developing neutrophils (F).

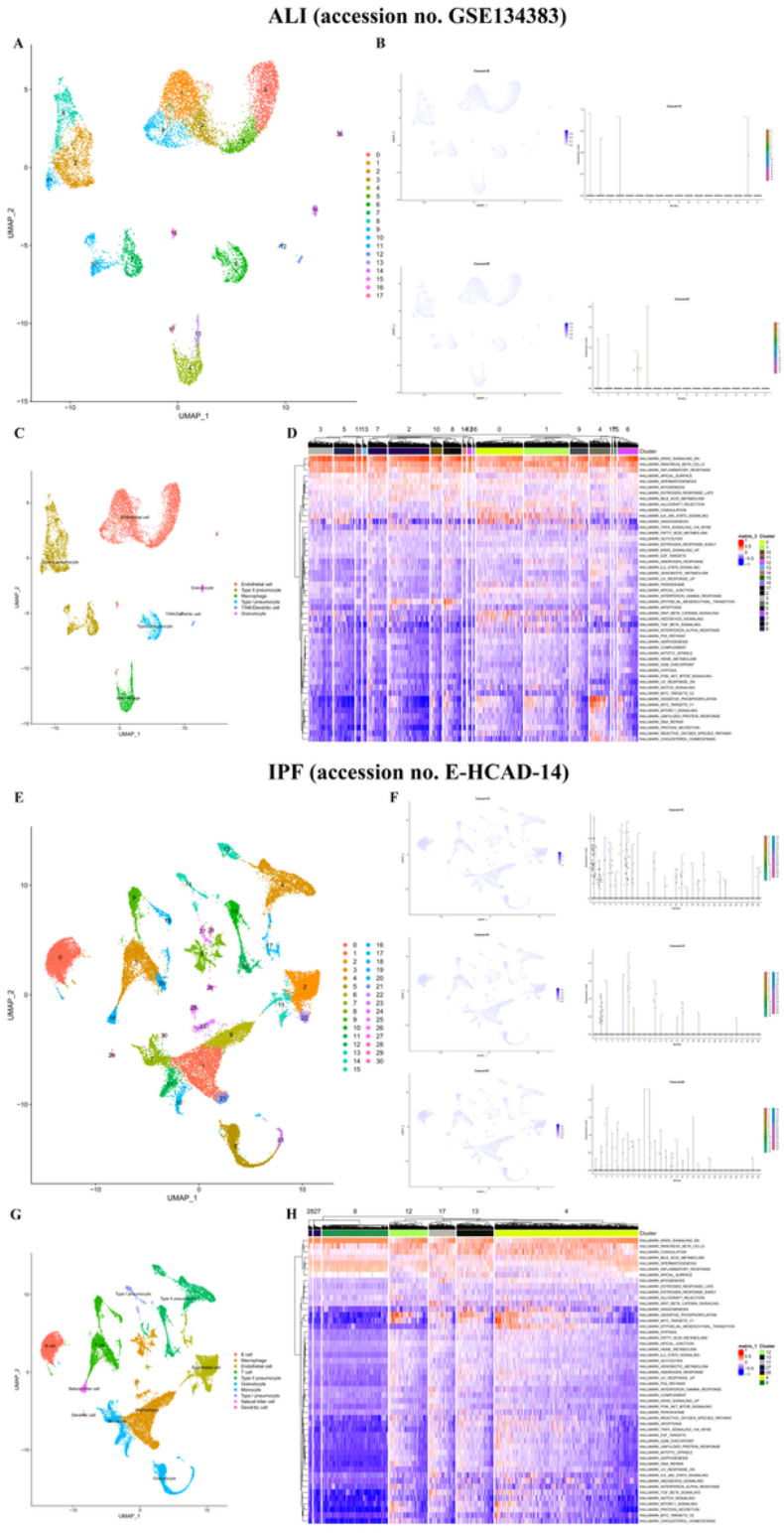

Figure 6

The abnormal expressions of CRGs in COVID-19 patients were COVID-19-specific and not related to CEA involvement in ALI and IPF. Due to the close correlation between CEA and ALI and IPF, we initially speculated that the poor prognosis of COVID-19 patients mediated by CEA might be related to ALI and IPF pathophysiologically. To validate this hypothesis, scRNA-seq data of ALI and IPF mouse lungs were also downloaded to evaluated the distribution and expression of CRGs, key receptor-ligand pair of cellular

communication and potential downstream pathways. The UAMP analysis identified 18 clusters and 6 cell types in ALI mouse lungs while there were no abnormal expressions of CRGs (A-C). And the interferon response and cell proliferation signaling pathways were not significantly activated in type II pneumocytes of ALI mouse lungs (D). Similarly, abnormal expressions of CRGs were also not detected in 31 clusters and 10 cell types of IPF mouse lungs (E-G). Besides, the heatmap of GSVA also showed that the interferon response and cell proliferation signaling pathways were not activated in type II pneumocytes of IPF mouse lungs (H).

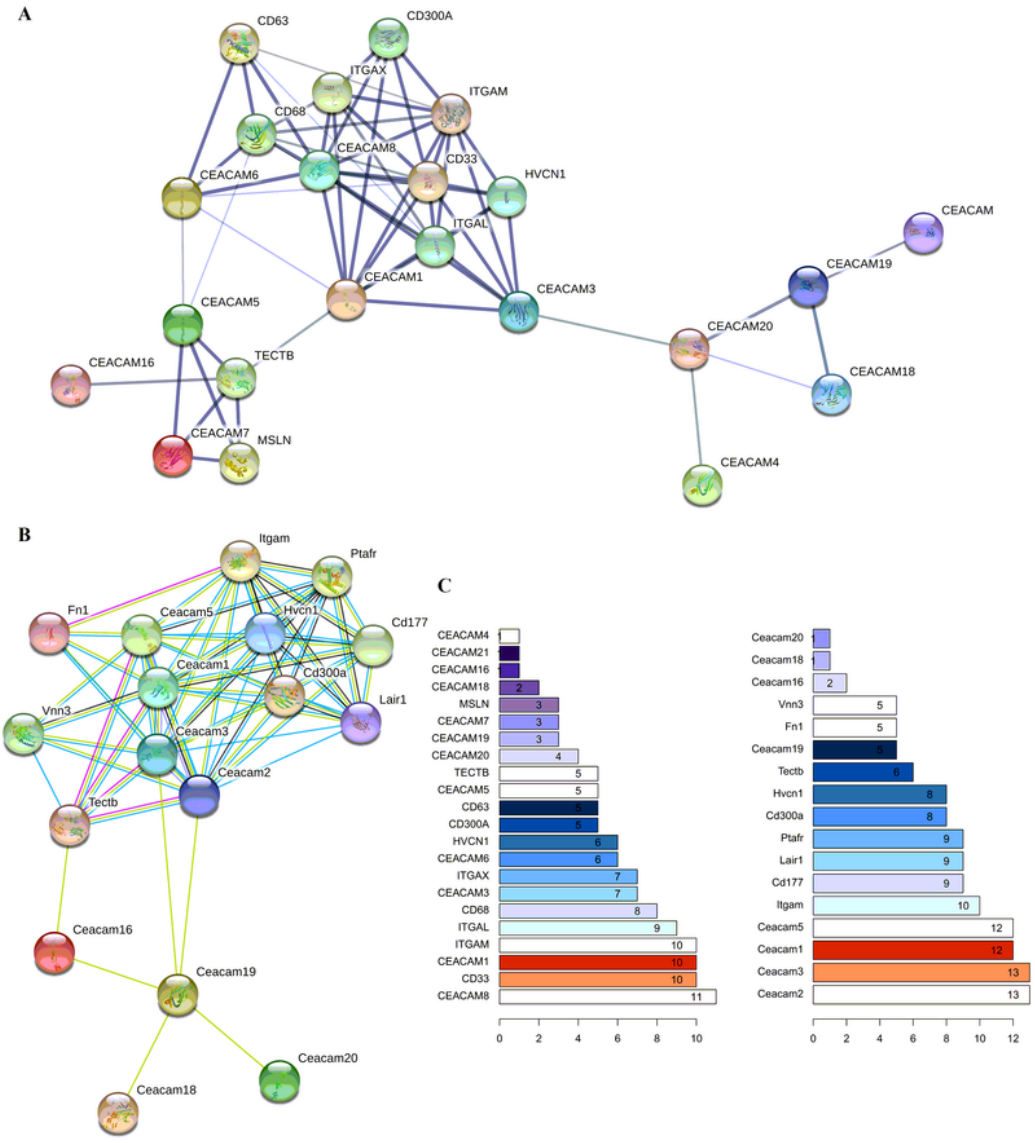

Figure 7

Protein-protein interaction (PPI) network of CRGs String database was used to construct the PPI network of CRGs, illustrating that several CRGs had direct protein-protein interactions with a variety of immune cell surface markers (A-C). Besides, the protein expression levels of CRGs in normal lung samples of The Human Protein Atlas were also checked, showing that only CEACAM21 were stained moderately in pneumocytes while the proteins of CEACAM5, CEACAM6 and CEACAM8 were not detected in normal lung samples (D).

Supplementary Files

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

- FigureS1.tif
- FigureS2.tif
- FigureS3.tif
- FigureS4.tif