# The transmembrane serine protease inhibitors are potential antiviral drugs for 2019-nCoV targeting the insertion sequence-induced viral infectivity

Tong Meng[1,2,10]†, Hao Cao[3,4]†, Hao Zhang[5,10]†, Zijian Kang[6,10], Da Xu[7,10], Haiyi Gong[5,10], Jing Wang[8], Zifu Li[8], Xingang Cui[7], Huji Xu[4,6], Haifeng Wei[5], Xiuwu Pan[7], Rongrong Zhu[9], Jianru Xiao[5]*, Wang Zhou[4,10]*, Liming Cheng[1]*, Jianmin Liu[8]*.

1 Division of Spine, Department of Orthopedics, Tongji Hospital affiliated to Tongji University School of Medicine, 200065 Shanghai, China

2 Tongji University Cancer Center, School of Medicine, Tongji University, 200092 Shanghai, China

3 School of Life Science and Biopharmaceutics, Shenyang Pharmaceutical University, 103 Wenhua Road, 110016 Shenyang, China

4 Peking-Tsinghua Center for Life Sciences, TsinghuaUniversity, 100084 Beijing, China

5 Department of Orthopaedic Oncology, Changzheng Hospital, Second Military Medical University, 200003 Shanghai, China

6 Department of Rheumatology and Immunology, Changzheng Hospital, Second Military Medical University, 200003 Shanghai, China

7 Depanrtment of Urology, The Third Affiliated Hospital of Second Military Medical University, 201805 Shanghai, China

8 Department of Neurosurgery, Changhai hospital, Second Military Medical University, 200003 Shanghai, China

9 Key Laboratory of Spine and Spinal Cord Injury Repair and Regeneration of Ministry of Education, Orthopaedic Department of Tongji Hospital, School of Life Science and Technology, Tongji University, 200092 Shanghai, China

10 Qiu-Jiang Bioinformatics Institute, 200003 Shanghai, China

†These authors contributed equally to this work, and all should be considered first author.

*Correspondence to: chstroke@163.com (Jianmin Liu)

limingcheng@tongji.edu.cn (Liming Cheng)
brilliant212@163.com (Wang Zhou)
jianruxiao83@163.com (Jianru Xiao)

**Abstract**

The 2019 novel coronavirus (2019-nCoV) induces an ongoing outbreak of pneumonia in China. The rapidly increasing infected cases suggest its effective transmission among humans, which largely depends on virus-binding host cell receptors and host cell proteases-cleaving virus spike protein. In order to identify the virus transmission of 2019-nCoV, the protease-induced spike protein cleavage ability between 2019-nCoV and SARS-CoV were compared. By protein sequence aligment and structure comparison, we found the key sequence of 675-QTQTNSPRRARSVAS-679 mediating 2019-nCoV spike protein cleavage. Its furin score (0.688) was higher than that of the corresponding sequence in SARS-CoV (0.139). In addition, the fragment of 680-SPRR-683 added two arginine hydrolysis sites (R682 and R683) on the surface and formed a loop for protease recognition. The molecular docking was based on the transmembrane serine protease (TMPRSS), the main proteases in coronavirus cleavage. Furthermore, as the cell receptor angiotensin converting enzyme II (ACE2) and cell proteases TMPRSSs are located in the same cell, the single-cell transcriptomes of normal human lung and gastroenteric system were used. The ACE2 and TMPRSSs were highly co-expressed in absorptive enterocytes, upper epithelial cells of esophagus and lung AT2 cells. In conclusion, this study provides the bioinformatics evidence for the increased viral infectivity of 2019-nCoV and indicates alveolus pulmonis, intestinal epithelium and esophagus epithelium as the potential target tissues. Due to the important roles of TMPRSSs in 2019-nCoV infection, transmembrane serine protease inhibitors may be the potential antiviral treatment options for 2019-nCoV infection.

Key words: 2019-nCoV, angiotensin converting enzyme II (ACE2), transmembrane serine protease (TMPRSSs), cleavage ability, viral infectivity

## Introduction

At the end of 2019, a rising number of pneumonia patients with unknown pathogen emerged from Wuhan to nearly the entire China[1]. A novel coronavirus was isolated and labeled as 2019 novel coronavirus (2019-nCoV) [2]. By complete genome sequence and pairwise protein sequence analysis, 2019-nCoV is suggested to be the species of severe acute respiratory syndrome (SARS) related coronaviruses (SARSr-CoV) [2]. Although 2019-nCoV is generally less pathogenic than SARS-CoV and Middle East respiratory syndrome coronavirus (MERS-CoV), it has a relatively high transmissibility [3].

With regard to SARSr-CoVs, the transmissibility and infectivity is largely controlled by the spike (S) surface envelope protein [4]. Its surface unit (S1) mediates the entry into host cells by binding to cell receptor and the transmembrane unit (S2) subunit regulates the fusion of viral and cellular membranes [5]. Prior to membrane fusion, the S protein should be cleaved and activated to allow for the fusion peptide releasing onto host cell membranes (Figure 1) [6]. 2019-nCoV uses the same cell receptor (angiotensin converting enzyme II, ACE2) as SARS-CoV, whereas it causes disease of different transmissibility and infectivity [7-9]. Thus, it is necessary to analyze the sequence of the S1 cleavage sites between 2019-nCoV and SARS-CoV and its proteolytic activation by host cell proteases.

The transmembrane serine proteases (TMPRSSs) were the main host cell proteases which cleave the S protein of human coronaviruses on the cell membrane. Nowadays, their hydrolytic effects have been widely reported in SARS-CoV and MERS-CoV[10]. In the alveolar cells, TMPRSSs, especially the TMPRSS2 and TMPRSS11D, are the main proteases which activate SARS-CoV S protein (SARS-S) for membrane fusion and cleave ACE2 for viral uptake[11,12]. Both 2019-nCoV and SARS-CoV belong to beta-coronavirus, with close genetic relationship and same host cell receptor, thus TMPRSSs may also be the candidate cell proteases in 2019-nCoV infection.

In this study, in order to identify the virus transmission of 2019-nCoV, the different hydrolysis abilities of TMPRSSs in the S1 cleavage sites between 2019-nCoV and SARS-CoV were compared by protein sequence aligment, furin score and structure comparison. As the cell colocalization of cell receptor ACE2 and cell proteases TMPRSSs, the single-cell transcriptomes of normal human lung and gastroenteric system were also used to identify the coexpressing cell composition and proportion. Our results may not only help to explain the high transmissibility of 2019-nCoV, but also identify the potential target tissues and explore the candidate therapeutic targets for 2019-nCoV infection.

## Materials and methods

### Structure modelling

The structures of 2019-nCoV S protein and TMPRSS2 were generated by SWISS-MODEL online server[13]. The structures were marked, superimposed and visualized by Chimera[14]. To further explore the possible catalytic mechanism of the 2019-nCoV S protein cleaved by TMPRSS2, ZDOCK program was used to predict their interaction [15]. A total of 5000 models were generated and were set to 50 clusters, then the best scoring models from the 5 largest clusters were selected for further analysis.

### Furin score

The fragmentation maps, scoring and residue coverage analysis were conducted using arginine and lysine propeptide cleavage sites prediction algorithms ProP 1.0 server[16].

### Single cell transcriptome data sources

Single cell transcriptome data were obtained from Single Cell Portal (https://singlecell.broadinstitute.org/single_cell), Human Cell Atlas Data Protal (https://data.humancellatlas.org) and Gene Expression Omnibus (GEO; https://www.ncbi.nlm.nih.gov/). Esophageal data were obtained from the research of E Madissoon *et al* containing 6 esophageal and 5 lung tissue samples[17]. Three lung datasets were obtained from GSE130148 [18], GSE122960[19] and GSE128169[20],

including four, five and eight lung tissues respectively. GSE134520 included 6 gastric mucosal samples from 3 non-atrophic gastritis and 2 chronic atrophic gastritis patients[21]. GSE134809 comprises 11 noninflammatory ileal samples from Crohn's disease patients[22]. The data from Christopher S *et al* consisted of 12 normal colon samples[23].

Quality control

Cells would be identified as poor-quality once (1) the number of expressed genes fewer than 200 or greater than 5000, or (2) more than 20% of UMIs being mapped to mitochondrial or ribosomal genes.

Data Integration, Dimension Reduction and Cell Clustering

Different methods were performed to process the downloaded data:

*1.* Esophagus dataset. Rdata were obtained and dimension reduction and clustering had already been implemented by the authors [17].

*2.* Lung, stomach and ileum datasets. We utilized functions in the Seurat package to normalize and scale the single-cell gene expression data[24]. Unique molecularidentifier (UMI) counts were normalized by the total number of UMIs per cell, multiplied by 10000 for normalization and log-transformed using the NormalizeData'' function. Then, multiple sample data within each dataset were merged using the "FindIntegrationAnchors" and "Integratedata" functions. After identifying highly variable genes (HVGs) using the "FindVariableGenes" function a principal component analysis (PCA) was performed on the single-cell expression matrix using the ''RunPCA'' function. The ''FindClusters'' function in the Seurat package was next utilized to conduct the cell clustering analysis into a graph structure in PCA space after constructing a K-nearest-neighbor graph based on the Euclidean distance in PCA space. Uniform Manifold Approximation and Projection (UMAP) visualization was performed for obtaining the clusters of cells.

*3.* Colon Dataset. The single cell data was processed with the R packages LIGER[25] and Seurat[24]. The gene matrix was first normalized to remove differences in sequencing depth and capture efficiency among cells. Variable genes in each dataset were identified using the "selectGenes" function. Then we used the "optimizeALS" function in LIGER to perform the integrative nonnegative matrix factorization and selecte a k of 15 and lambda of 5.0 to obtain a plot of expected alignment. The "quantileAlignSNF" function was then performed to builds a shared factor neighborhood graph to jointly cluster cells, then quantile normalizes corresponding clusters. Next nonlinear dimensionality reduction was calculated using the "RunUMAP" function and the results were visualized with UMAP.

Identification of cell types and Gene expression analysis

Clusters were annotated on the expression of known cell markers and the clustering information provided in the articles. Then, we utilized the "RunALRA" function to impute lost values in the gene expression matrix. The imputed gene expression was shown in Feature plots and violin plots. We used "Quantile normalization" in the R package preprocessCore (R package version 1.46.0. https://github.com/bmbolstad/preprocessCore) to remove unwanted technical variability across different datasets were further denoised to compare gene expression.

Endocytosis and exocytosis-associated genes signature

All pathways related to endocytosis or exocytosis were obtained from Harmonizome dataset [26]. To detect the expression levels of functional genesets, mean expression of an inflammation signature was calculated.

External validation

To minimize bias, external databases of Genotype-Tissue Expression (GTEx)[27], and The Human Protein Atlas[28] were used to detect gene and protein expression of ACE2 and TMPRSSs at the tissue levels including normal lung and digestive system, such as esophagus, stomach, small intestine and colon.

**Results**

The structure of the SARS-S and 2019-nCoV S protein homo-trimers

The structure of the SARS-S and 2019-nCoV S protein were compared. The insert aa675-690 to SARS-S aa661-672 with the structural missed residues are colored green (Figure 2A). Although the 2019-nCoV S protein has similar structure with SARS-S, there was a very obvious insertion sequence (QTQTNSPRRARSVASQS) in 2019-nCoV S protein. The insert aa675-690 of 2019-nCoV S protein that corresponds to the insert region of SARS-S is colored yellow (Figure 2B). The structural superimpose of SARS-S (wheat) and 2019-nCoV S protein (cyan) (Figure 2C).

We analyzed the amino acid sequence of the S1 cleavage sites among all the coronavirus family members and found that other coronaviruses, such as HCoV-OC43 and BatCoV-HKU5) also have the similar R-rich domain (Figure 3A). Then, the furin score was used to identify the cleavage efficiency of this sequence. The furin score of SARS was 0.139, whereas that of 2019-nCoV was 0.688, indicating that the insertion sequence may increase the ability of recognition and cleavage by TMPRSSs (Figure 3B). The obvious insertion sequence in 2019-nCoV S protein is at residue R685, especially at R682 and R683 (the yellow structure in Figure 3C). In addition, the insertion sequence of R682 and R683 was protruded from the molecular surface (Figure 3D).

Structure and catalytic mechanism of TMPRSSs

The catalytic triad comprised of H296, D345 and S441 are colored blue, green and cyan, respectively, the substrate binding residue D435 which located in the bottom of pocket is marked in red, the substrate binding pocket is deeper than most of serine proteinase (Figure 4A, B). Take TMPRSS2 as an example. Its bottom has a negatively charged aspartic acid residue which can facilitate the binding and stabilization of the long-chain positively charged amino acid residue of substrate. Polypeptide substrate analogue KQLR was presented in Figure 4C, with arginine, glutamine, leucine and lysine. The Figure 4D and 4E revealed the state of substrate analogue binding to the catalytic pocket (Figure 4D, E). We next simulated the conformation of the insertion

sequence in 2019-nCoV S protein and the TMPRSS2 by molecular docking. We found the insertion sequence formed a loop which was easily recognized by the catalytic pocket of TMPRSS2 and was subsequently cleaved (Figure 4F, G).

Annotation of cell types

In this study, 5 datasets with single-cell transcriptomes of esophagus, gastric, small intestine and colon were analyzed, along with lung. Based on Cell Ranger output, the gene expression count matrices were used to present sequential clustering of cells according to different organs or particular clusters.

Cell type-specific expression of ACE2, TMPRSS2, TMPRSS11D and ADAM17

After initial quality controls, 57,020 cells and 15 clusters were identified in the lung (Figure 5A). The detected cell types included ciliated, alveolar type 1 (AT1) and alveolar type 2 (AT2) cells, along with fibroblast, muscle, and endothelial cells. The identified immune cell types were T, B and NK cells, along with macrophages, monocytes and dendritic cells (DC). ACE2 was mainly expressed in AT2 cells along with AT1 and fibroblast cells, while TMPRSS2 was found in AT1 and AT2 cells; and TMPRSS11D was expressed in AT1 cells, fibroblast and macrophage (Figure 5B). A total of 17 TMPRSS genes were marked with green and ACE2 was marked with red. We found a obvious colocalization (yellow) between them in AT2 cell (Figure 5C). In addition, the total TMPRSS genes were highly expressed in AT1 cells, and less expressed in ciliated and AT2 cells (Figure 5D). The Violin plots showed the expression of ACE2 and each TMPRSS and found TMPRSS1, TMPRSS2 and TMPRSS3 had a obvious co-expression with ACE2 in AT1 and AT2 cells. TMPRSS11D was not found in any clusters, whereas ADAM17 was found in all clusters (Figure 5E). The Immunohistochemistry (IHC) images of ACE2, TMPRSS2 and TMPRSS11D in nomal lung showed a similar result (Figure 5F).

In the esophagus, 87,947 cells passed quality control and 14 cell types were identified. Over 90% cells belong to four major epithelial cell types: upper, stratified, suprabasal, and dividing cells of the suprabasal layer. ACE2 was highly expressed in upper and

stratified epithelial cells. The total TMPRSS genes were found in upper, stratified and dividing epithelial cells and glands. Most TMPRSS genes were expressed in upper epithelial cells. Thus, ACE2 and total TMPRSS genes were highly expressed in upper epithelial cells. (Figure 6A).

A total of 29,678 cells and 10 cell types were identified in the stomach after quality control with a high proportion of gastric epithelial cells, including antral basal gland mucous cells (GMCs), pit mucous cells (PMCs), chief cells and enteroendocrine cells. The expression of ACE2 are very low in all the clusters, and each TMPRSS gene was expressed in different cells with a relatively low expression(Figure 6B).

After quality controls, 11,218 cells and 5 cell types were identified in the ileum epithelia (Figure 6C). ACE2 were highly expressed in absorptive and progenitor enterocytes. Although the total TMPRSS genes could be found in all the cells, TMPRSS2, TMPRSS4, TMPRSS14 and TMPRSS15 have been found expressed in ACE-expressing cells (Figure 6C).

All the 47,442 cells from the colon were annotated after quality controls. Absorptive and secretory clusters were identified in epithelial cells. The absorptive epithelial cells included transit amplifying (TA) cells (TA 1, TA 2), immature enterocytes, and enterocytes. The secretory epithelial cells comprised progenitor cells (secretory TA, immature goblet) and for mature cells (goblet, and enteroendocrine). ACE2 was mainly found in enterocytes and less expressed in immature enterocytes. In the meantime, total TMPRSS genes were found in all the cells. TMPRSS2, TMPRSS3, TMPRSS4 and TMPRSS14 were found in ACE-expressing cells (Figure 6D).

The expressions of ACE2 and classic TTSPs (TMPRSS2, TMPRSS3, TMPRSS4, TMPRSS6, TMPRSS11D, TMPRSS14 were detected in lung and digestive tract clusters. An almost consistent expression and distribution was found between ACE2 and TMPRSS2 in all the 9 clusters, with high expression in absorptive enterocytes, upper epithelial cells of esophagus and lung AT2 cells (Figure 7A). The total TMPRSS genes were expressed in all the ACE2-expressing cells and they are highly

expressed in enterocytes, upper epithelial cells of esophagus and lung AT1 cells (Figure 7B). The endocytosis and exocytosis-associated genes which are related to the entry of virus into host cells were also detected in all the 9 clusters. The endocytosis signature were more expressed in colon and exocytosis signature were more expressed in upper epithelial cells of esophagus (Figure 7C). The RNA-seq data of lung, esophagus, stomach, small intestine, colon-transverse and colon-sigmoid were obtained from GTEx database. The expressions of ACE2 and TMPRSS2 also had a similar tendency and were highly expressed in small intestine and colon, while the TMPRSS11D was mainly found in the esophagus (Figure 7D).

**Discussion**

The coronaviruses is the common infection source of enteric, respiratory, and central nervous system in humans and other mammals[29]. At the beginning of the twenty-first century, two betacoronaviruses, SARS-CoV and MERS-CoV, result in persistent public panics and became the most significant public health events[30]. In December 2019, a novel identified coronavirus (2019-nCoV) induced an ongoing outbreak of pneumonia in Wuhan, Hubei, China [31]. The rapidly increasing number of 2019-nCoV-infected cases suggests that 2019-nCoV may be transmitted effectively among humans, implying a high pandemic potential of 2019-nCoV [2,31,32]. In this study, we found that the insertion sequence in 2019-nCoV S1 cleavage sites may enhance TMPRSS-induced cleavage ability and viral infectivity. The high co-expression of receptor ACE2 and protease TMPRSSs was found in absorptive enterocytes, upper epithelial cells of esophagus and lung AT2 cells. The important roles of TMPRSSs in 2019-nCoV infection indicate transmembrane serine protease inhibitors as the potential antiviral treatment options.

Previous studies identified that SARS-CoV mutated between 2002 and 2004 to better bind to its cell receptor, replicate in human cells and enhance the virulence [3]. Thus, it is important to explore whether 2019-nCoV behaves like SARS-CoV to adapt to the host cell and whether the specific structure of 2019-nCoV S protein seems better

suited to be activated by host cell proteases. Notably, SARS-CoV and 2019-nCoV share the same receptor protein ACE2[8,33]. Moreover, the receptor-binding domain (RBD) in S protein of 2019-nCoV may bind to ACE2 with the similar affinity as SARS-CoV RBD does, indicating the different virus transmission may be associated with the protease-induced spike protein cleavage ability[34]. As the common transmembrane proteases, TMPRSSs can cleave SARS-S near or at the cell surface and render host cell entry independent of the endosomal pathway using cathepsin B/L[35]. Different from cathepsin B/L, they can also promote viral spread in the host and cleave ACE2 to augment about 30-fold viral infectivity[11,36]. Structurally, TMPRSSs include extracellular domain, transmembrane domain and intracellular domain in which extracellular domain is the main catalytic domain. They show similar substrate-specificity and catalytic mechanism. All TMPRSSs possess a conserved aspartate residue at the bottom of the catalytic pocket which is relatively deep, facilitating the binding and stabilization of arginine or lysine residues in the P1 position[37,38].

In the SARS infection, the member of TMPRSS family, such as TMPRSS2, is shown to cleave SARS-S at residue R667 which is the S1/S2 cleavage site and residue R797 which is the S2' cleavage site [10]. Both the two hydrolysis sites of R667 and R797 are located on surface to facilitate the recognition and cleavage. Compared with SARS-S, the corresponding arginine hydrolysis sites of 2019-nCoV S protein has similar structure. However, there is an insertion sequence at residue R685 (corresponding to R667 in the SARS-S) which makes R682 and R683 protrude from the protein surface. Moreover, as for this insertion sequence, the furin score of 2019-nCoV was much higher than that of SARS-CoV. Due to the main hydrolysis site of TMPRSSs being R and the high furin score, we supposed that the additional arginine residues exposed from protein surface might enhance the recognition and cleavage activity of TMPRSSs and promote the viral infectivity of 2019-nCoV [37,38].

By the way, some researchers believed that there were four inserts in the S protein of 2019-nCoV from HIV sequence. However, the similar sequence of the reported fourth insertion site (680-SPRR-683) in 2019-nCoV was commonly found in many beta-coronavirus. Therefore, it is not scientific to consider the insertion sequence in 2019-nCoV S protein being artificial.

The entry of 2019-nCoV into host cells depends on the cell receptor recognition and cell proteases cleaving. Thus, the cell receptor ACE2 and cell proteases TMPRSSs should be colocalized in the same host cell. With the help of single cell sequencing, we found a strong co-expression between ACE2 and all TMPRSS family members in lung AT1 and AT2 cells, which was also the main damaged cell in 2019-nCoV pneumonia. In addition, they are also highly co-expressed in absorptive enterocytes and upper epithelial cells of esophagus, implying that intestinal epithelium and esophagus epithelium may also be the potential target tissues. This can explain the cases whose 2019-nCoV was detected in the esophageal erosion or stool specimen[31,39,40].

Due to the critical role of TMPRSSs in influenza virus and coronavirus infections, serine protease inhibitors have been used in the antiviral therapeutic strategy targeting TMPRSSs with high antiviral activities, such as camostat, nafamostat and leupeptin[36,41-43]. Nowadays, Remdesivir (GS-5734) has been used in the treatment of 2019-nCoV and the therapeutic effects are still unclear. Based on our results, we also supposed that TMPRSSs may also serve as candidate antiviral targets for 2019-nCoV infection and the clinical trails of transmembrane serine protease inhibitors may also be treatment options for 2019-nCoV infection.

## Conclusion

This study provides the bioinformatics evidence for the increased viral infectivity of 2019-nCoV and indicates alveolus pulmonis, intestinal epithelium and esophagus epithelium as the potential target tissues. In addition, due to the important roles of

TMPRSSs in 2019-nCoV infection, transmembrane serine protease inhibitors may be potential antiviral treatment options for 2019-nCoV infection

**Reference**


1. The L. Emerging understandings of 2019-nCoV. Lancet 2020.
2. Zhu N, Zhang D, Wang W, et al. A Novel Coronavirus from Patients with Pneumonia in China, 2019. The New England journal of medicine 2020.
3. Chen J. Pathogenicity and Transmissibility of 2019-nCoV-A Quick Overview and Comparison with Other Emerging Viruses. Microbes and infection 2020.
4. Walls AC, Xiong X, Park YJ, et al. Unexpected Receptor Functional Mimicry Elucidates Activation of Coronavirus Fusion. Cell 2019;176:1026-39.e15.
5. Hofmann H, Pohlmann S. Cellular entry of the SARS coronavirus. Trends in microbiology 2004;12:466-72.
6. Gallagher TM, Buchmeier MJ. Coronavirus spike proteins in viral entry and pathogenesis. Virology 2001;279:371-4.
7. Gui M, Song W, Zhou H, et al. Cryo-electron microscopy structures of the SARS-CoV spike glycoprotein reveal a prerequisite conformational state for receptor binding. Cell research 2017;27:119-29.
8. P Zhou XY, XG Wang, B Hu, L Zhang, W Zhang, HR Si, Y Zhu, B Li, CL Huang, HD Chen, J Chen, Y Luo, H Guo, RD Jiang, MQ Liu, Y Chen, XR Shen, X Wang, XS Zheng, K Zhao, QJ Chen, F Deng, LL Liu, B Yan, FX Zhan, YY

1 Wang, GF Xiao, ZL Shi. A pneumonia outbreak associated with a new
2 coronavirus of probable bat origin. Nature 2020.
3 9. Zhao S RJ, MUSA SS, Yang G, Lou Y, Gao D, Yang L, He D. . Preliminary
4 estimation of the basic reproduction number of novel coronavirus (2019-nCoV)
5 in China, from 2019 to 2020: A data-driven analysis in the early phase of the
6 outbreak. . bioRxiv 2020:916395.
7 10. Millet JK, Whittaker GR. Host cell proteases: Critical determinants of
8 coronavirus tropism and pathogenesis. Virus research 2015;202:120-34.
9 11. Heurich A, Hofmann-Winkler H, Gierer S, Liepold T, Jahn O, Pohlmann S.
10 TMPRSS2 and ADAM17 cleave ACE2 differentially and only proteolysis by
11 TMPRSS2 augments entry driven by the severe acute respiratory syndrome
12 coronavirus spike protein. Journal of virology 2014;88:1293-307.
13 12. Li F. Structure, Function, and Evolution of Coronavirus Spike Proteins.
14 Annual review of virology 2016;3:237-61.
15 13. Biasini M, Bienert S, Waterhouse A, et al. SWISS-MODEL: modelling
16 protein tertiary and quaternary structure using evolutionary information.
17 Nucleic acids research 2014;42:W252-8.
18 14. Pettersen EF, Goddard TD, Huang CC, et al. UCSF Chimera--a
19 visualization system for exploratory research and analysis. Journal of
20 computational chemistry 2004;25:1605-12.
21 15. Wiehe K, Pierce B, Mintseris J, et al. ZDOCK and RDOCK performance in

1 CAPRI rounds 3, 4, and 5. Proteins 2005;60:207-13.

2 16. Duckert P, Brunak S, Blom N. Prediction of proprotein convertase
3 cleavage sites. Protein engineering, design & selection : PEDS
4 2004;17:107-12.

5 17. Madissoon E, Wilbrey-Clark A, Miragaia RJ, et al. scRNA-seq assessment
6 of the human lung, spleen, and esophagus tissue stability after cold
7 preservation. Genome biology 2019;21:1.

8 18. Vieira Braga FA, Kar G, Berg M, et al. A cellular census of human lungs
9 identifies novel cell states in health and in asthma. Nature medicine
10 2019;25:1153-63.

11 19. Reyfman PA, Walter JM, Joshi N, et al. Single-Cell Transcriptomic
12 Analysis of Human Lung Provides Insights into the Pathobiology of Pulmonary
13 Fibrosis. American journal of respiratory and critical care medicine
14 2019;199:1517-36.

15 20. Valenzi E, Bulik M, Tabib T, et al. Single-cell analysis reveals fibroblast
16 heterogeneity and myofibroblasts in systemic sclerosis-associated interstitial
17 lung disease. Annals of the rheumatic diseases 2019;78:1379-87.

18 21. Zhang P, Yang M, Zhang Y, et al. Dissecting the Single-Cell
19 Transcriptome Network Underlying Gastric Premalignant Lesions and Early
20 Gastric Cancer. Cell reports 2019;27:1934-47.e5.

21 22. Martin JC, Chang C, Boschetti G, et al. Single-Cell Analysis of Crohn's

1 Disease Lesions Identifies a Pathogenic Cellular Module Associated with
2 Resistance to Anti-TNF Therapy. Cell 2019;178:1493-508.e20.
3 23. Smillie CS, Biton M, Ordovas-Montanes J, et al. Intra- and Inter-cellular
4 Rewiring of the Human Colon during Ulcerative Colitis. Cell
5 2019;178:714-30.e22.
6 24. Stuart T, Butler A, Hoffman P, et al. Comprehensive Integration of
7 Single-Cell Data. Cell 2019;177:1888-902.e21.
8 25. Welch JD, Kozareva V, Ferreira A, Vanderburg C, Martin C, Macosko EZ.
9 Single-Cell Multi-omic Integration Compares and Contrasts Features of Brain
10 Cell Identity. Cell 2019;177:1873-87.e17.
11 26. Rouillard AD, Gundersen GW, Fernandez NF, et al. The harmonizome: a
12 collection of processed datasets gathered to serve and mine knowledge about
13 genes and proteins. Database : the journal of biological databases and
14 curation 2016;2016.
15 27. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis:
16 multitissue gene regulation in humans. Science (New York, NY)
17 2015;348:648-60.
18 28. Uhlen M, Fagerberg L, Hallstrom BM, et al. Proteomics. Tissue-based map
19 of the human proteome. Science (New York, NY) 2015;347:1260419.
20 29. Perlman S, Netland J. Coronaviruses post-SARS: update on replication
21 and pathogenesis. Nature reviews Microbiology 2009;7:439-50.

1 30. de Wit E, van Doremalen N, Falzarano D, Munster VJ. SARS and MERS:
2 recent insights into emerging coronaviruses. Nature reviews Microbiology
3 2016;14:523-34.
4 31. Huang C, Wang Y, Li X, et al. Clinical features of patients infected with
5 2019 novel coronavirus in Wuhan, China. Lancet 2020.
6 32. Lee PI, Hsueh PR. Emerging threats from zoonotic coronaviruses-from
7 SARS and MERS to 2019-nCoV. Journal of microbiology, immunology, and
8 infection = Wei mian yu gan ran za zhi 2020.
9 33. Li W, Moore MJ, Vasilieva N, et al. Angiotensin-converting enzyme 2 is a
10 functional receptor for the SARS coronavirus. Nature 2003;426:450-4.
11 34. Tian XL LC, Huang A, Xia S, Lu SC, Shi ZL, Lu L, Jiang SB, Yang ZL, Wu
12 YL, Ying TL. Potent binding of 2019 novel coronavirus spike protein by a
13 SARS coronavirus-specific human monoclonal antibody. bioRxiv 2020.
14 35. Shirato K, Kawase M, Matsuyama S. Wild-type human coronaviruses
15 prefer cell-surface TMPRSS2 to endosomal cathepsins for cell entry. Virology
16 2018;517:9-15.
17 36. Zhou Y, Vedantham P, Lu K, et al. Protease inhibitors targeting
18 coronavirus and filovirus entry. Antiviral research 2015;116:76-84.
19 37. Herter S, Piper DE, Aaron W, et al. Hepatocyte growth factor is a preferred
20 in vitro substrate for human hepsin, a membrane-anchored serine protease
21 implicated in prostate and ovarian cancers. The Biochemical journal

2005;390:125-36.

38. Limburg H, Harbig A, Bestle D, et al. TMPRSS2 Is the Major Activating Protease of Influenza A Virus in Primary Human Airway Cells and Influenza B Virus in Human Type II Pneumocytes. Journal of virology 2019;93.

39. Holshue ML, DeBolt C, Lindquist S, et al. First Case of 2019 Novel Coronavirus in the United States. The New England journal of medicine 2020.

40. Guan WJ NZ, Hu Y, Liang WH, Ou CQ, He JX, Liu L, Shan H, Lei CL, Hui David S.C., Du B Clinical characteristics of 2019 novel coronavirus infection in China. medRxiv 2020.

41. Shen LW, Mao HJ, Wu YL, Tanaka Y, Zhang W. TMPRSS2: A potential target for treatment of influenza virus and coronavirus infections. Biochimie 2017;142:1-10.

42. Shin WJ, Seong BL. Type II transmembrane serine proteases as potential target for anti-influenza drug discovery. Expert opinion on drug discovery 2017;12:1139-52.

43. Yamamoto M, Matsuyama S, Li X, et al. Identification of Nafamostat as a Potent Inhibitor of Middle East Respiratory Syndrome Coronavirus S Protein-Mediated Membrane Fusion Using the Split-Protein-Based Cell-Cell Fusion Assay. Antimicrobial agents and chemotherapy 2016;60:6532-9.

1

2

3

4

5

6

7

8

9

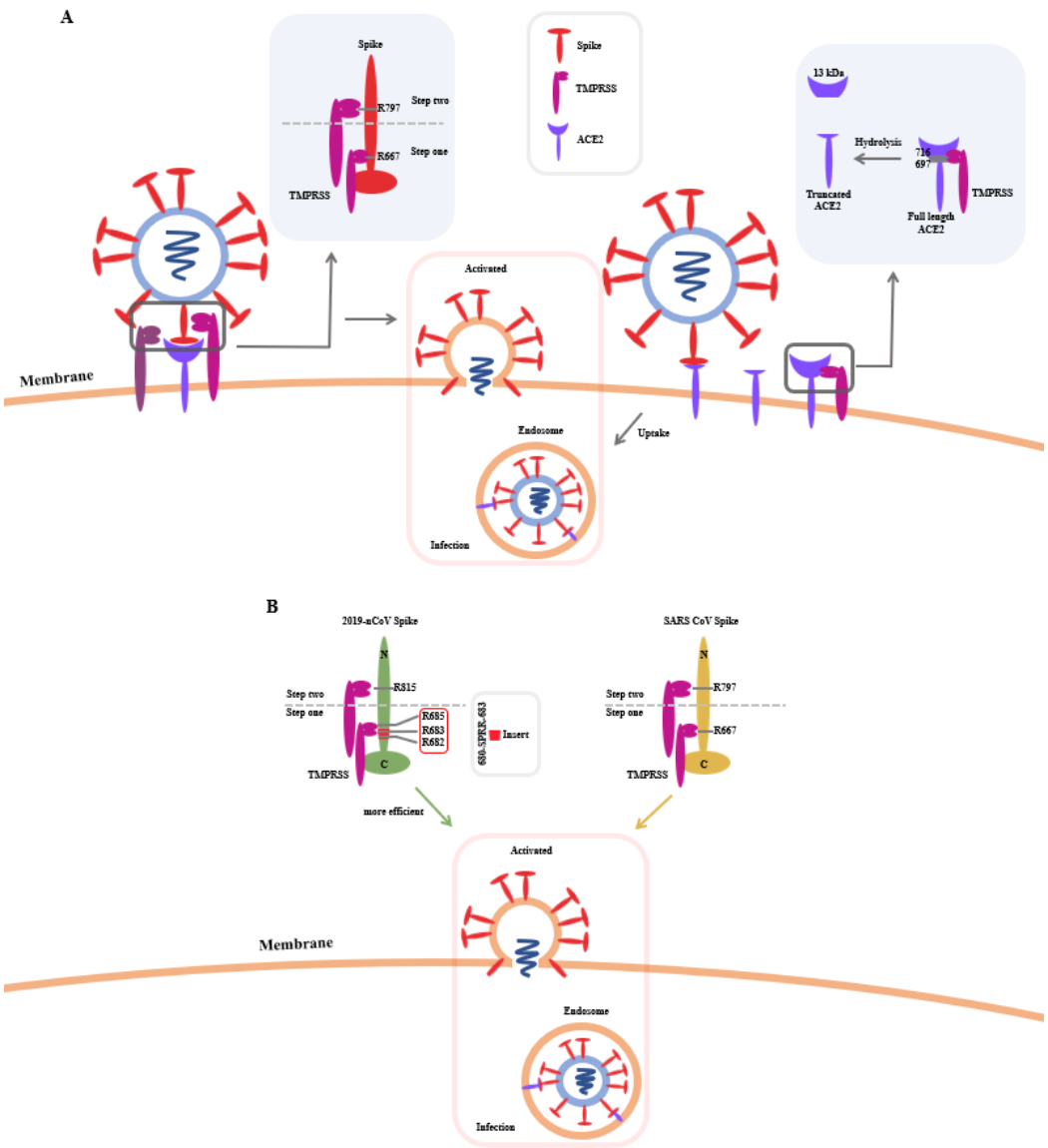


**Figure1. Role of TMPRSS2/HAT proteases in the cellular entry of 2019-nCoV.**

(A). Routes of coronavirus entering host cells. 2019-nCoV could enter host cells via two distinct routes, depending on the availability of cellular proteases required for activation of 2019-nCoV. The first route of activation can be achieved if the 2019-nCoV activating protease TMPRSS2 and ACE2 are co-expressed on the surface of target cells. The spiked protein binds to ACE2 through its S1 subunit and is treated by TMPRSS2 at the R667 or R797 in the S1/S2 site. This activates Spike protein and allows 2019-nCoV fusion at the cell surface. 2019-nCoV was encapsulated into

cellular vesicles before transport of virions into host cell endosomes. Uptake can be enhanced if TMPRSS2 also cleaves ACE2 amino acids 697 to 716, resulting in shedding of 13kD ACE2 fragment in culture supernatants. The second route can be pursued if there were no 2019-nCoV activating proteases expressed at the cell surface. While binding of virion-associated Spike protein to ACE2, the virions are taken up into endosomes, where 2019-nCoV could be cleaved and activated by the pH-dependent protease.

(B). The difference between 2019-nCoV and SARS-CoV in activating the Spike protein. The Spike protein of SARS involves two cleavage sites recognized by TMPRSS2, one at arginine 667 and one at arginine 797. (right). Compared with SARS-Cov, the S protein of 2019-nCoV (left)has an insertion sequence 680-SPRR-683 (grey box)at the TM cleavage site. We speculated that R682, R682 and R685 (red box) could be used as the most suitable substrates for TM, which can increase the cleavage efficiency of TM to S protein, promote the activation of 2019-nCoV, and make 2019-nCoV more infectious.

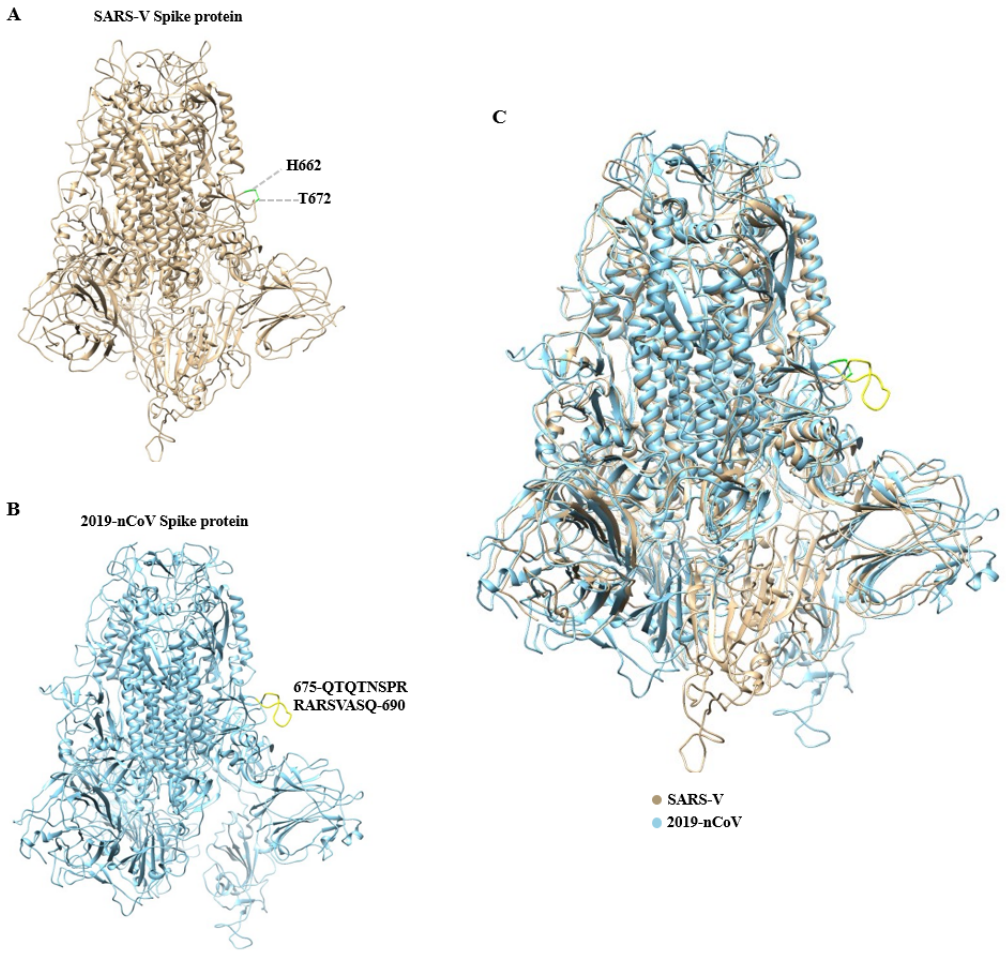


**Figure 2. Overall structure of the SARS-CoV Spike protein and 2019-nCoV Spike protein homo-trimers**

(A). Structure of the SARS-CoV Spike protein (from PDB: 5X5B). The insert aa675-690 to SARS-CoV Spike protein aa661-672 with the structural missed residues are colored green.

(B). Structure of the 2019-nCoV Spike protein (Modelled by SWISS-MODEL). The insert aa675-690 of 2019-nCoV Spike protein that corresponds to the insert region of SARS-V Spike protein is colored yellow

(C). The structural superimpose of SARS-CoV Spike protein (yellow) and 2019-nCoV Spike protein (blue)

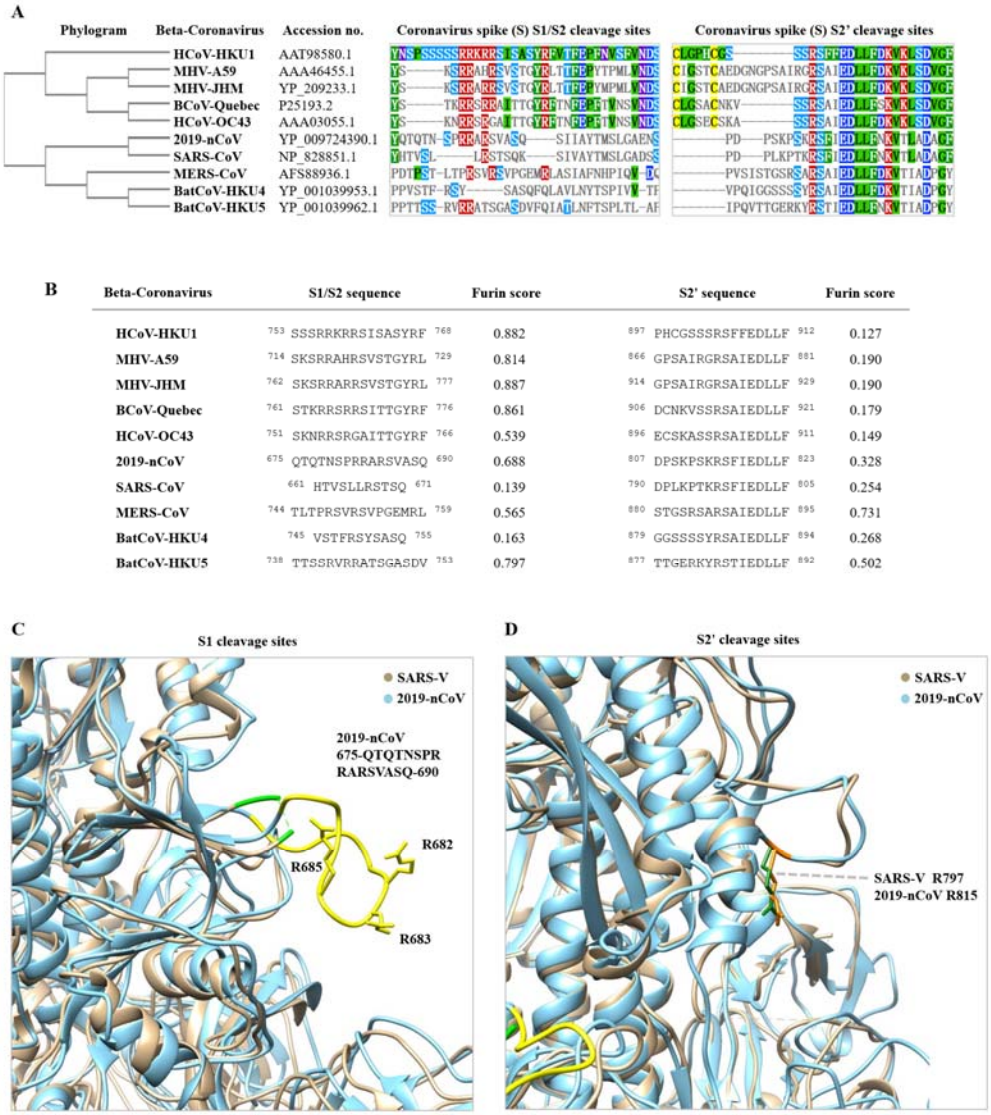


**Figure 3. The two potential cleavage sites of SARS-CoV Spike protein and 2019-nCoV Spike protein by TMPRSS2**.

(A). Phylogenetic tree based on the protein sequences of 2019-nCoV Spike protein with the Spike protein of SARS-CoV and other relative beta-coronaviruses, and the amino acid sequence alignment of two potential cleavage sites by TMPRSS2 of them. (B) The putative furin scores of the two potential cleavage sites of the coronaviruses. (C-D). Structure comparison in detail of the SARS-CoV Spike protein with 2019-nCoV Spike protein. Shown are the insert 675-690 of 2019-nCoV Spike protein (yellow) and the corresponding loci to SARS-CoV Spike protein 661-672 (green).

Three important residues, R682, R683, R685, are specially marked in (B). The similarly SARS-CoV R797 with 2019-nCoV R815 are colored forest green and orange, respectively (C).

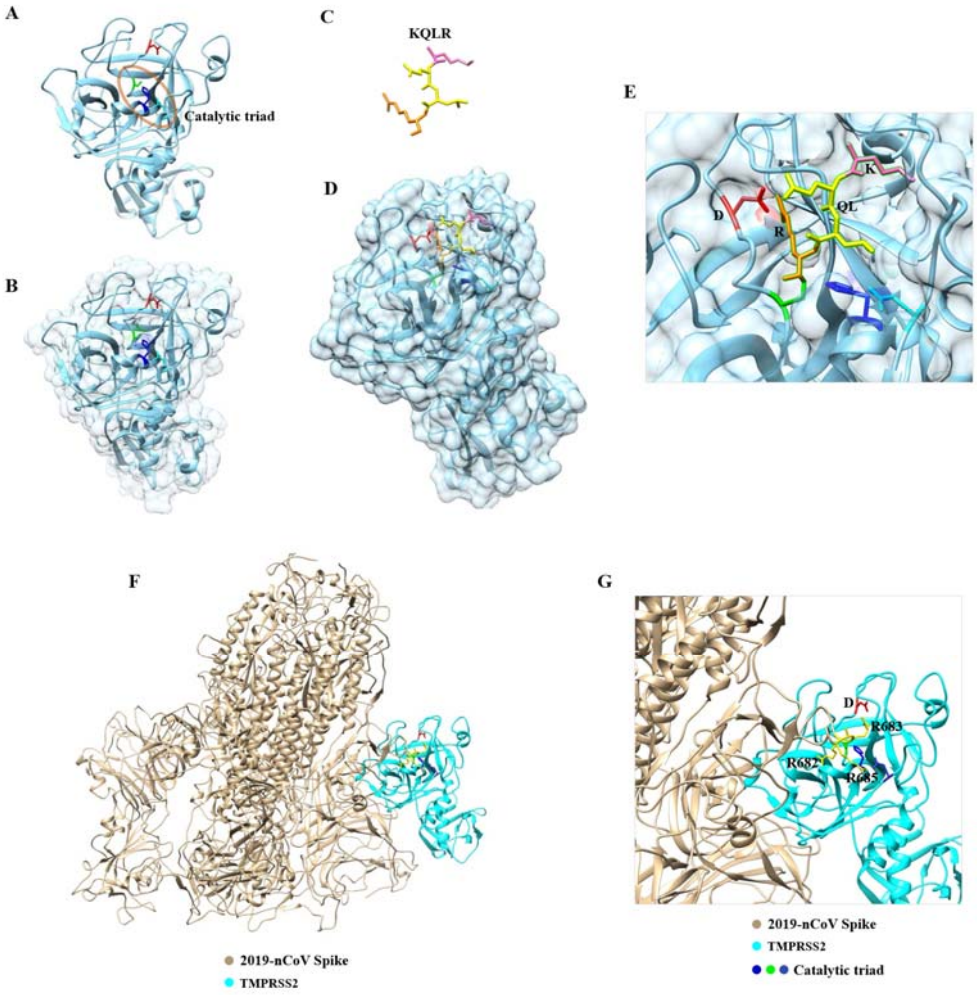


**Figure 4. Structure and catalytic mechanism of TMPRSS2**

**(A-B)**. Overall structure and surface of TMPRSS2 (Modelled by SWISS-MODEL). The TMPRSS2, catalytic triad comprised of H296, D345 and S441 are colored cyan, blue, cyan and green, respectively. The substrate binding residue D435 which located in the bottom of pocket is marked in red, the substrate binding pocket is deeper than most of serine proteinase.

**(C)**. Polypeptide substrate analogue KQLR. Cleaved site Arg is coloured orange. Gln, Leu are colored yellow, and Lys is colored pink.

**(D-E)**. The state of substrate analogue binding in the catalytic pocket, and the detail shown in (**E)**. The state of substrate analogue binding in the catalytic pocket, and the detail shown in E, Arg of substrate analogue is strongly interacted with D435, shown

in (**E)**, Arg of substrate analogue is strongly interacted with D435.

**(F-G)**. The predicted state of 2019-nCoV Spike protein binding to the catalytic pocket of TMPRSS2 and its detail. 2019-nCoV Spike protein, D345 of TMPRSS2 are coloured wheat and medium blue in here, respectively.

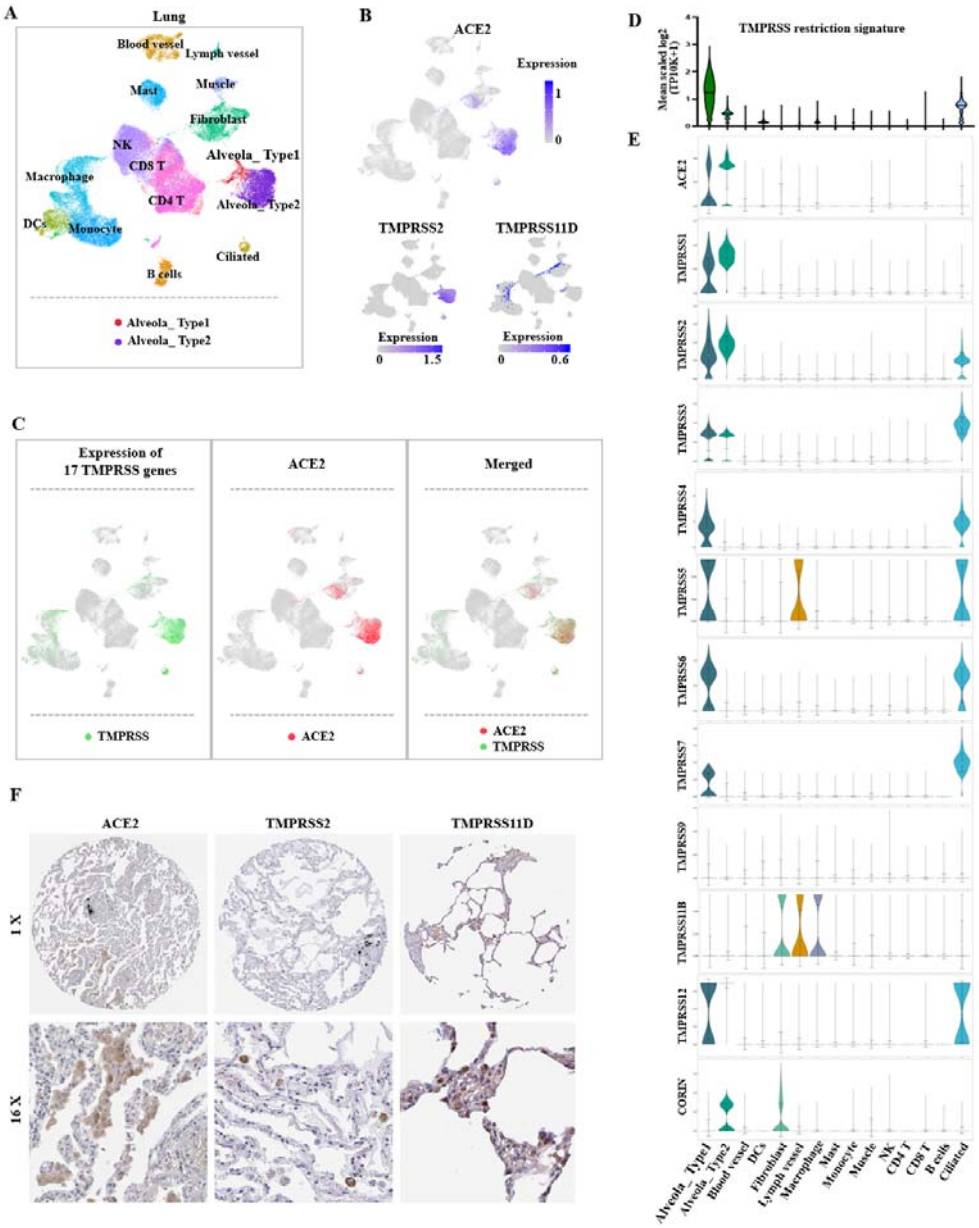


**Figure 5. Single-cell analysis and immunohistochemistry of normal lung tissue.**

(A). UMAP plots showing the landscape of cells from normal lung tissue. Fifteen clusters are colored, distinctively labeled.

(B). Feature plots demonstrating expression of ACE2, TMPRSS2 and TMPRSS11D across fifteen clusters.

(C) Feature plots demonstrating expression of ACE2 (red) and TMPSS genes(green). The plots were merged to show the co-expression of these genes (brown)

(D). Violin plots showing the mean expression of TMPRSS family genes across clusters. The expression is measured as the mean log2 (TP10K+1) value.

(E). Violin plots showing the mexpression of ACE2 and TMPRSS family genes, across clusters. The expression is measured as the log2 (TP10K+1) value.

(F). Immunohistochemical images showing the expression of ACE2, TMPRSS2 and TMPRSS11D in lung tissues.

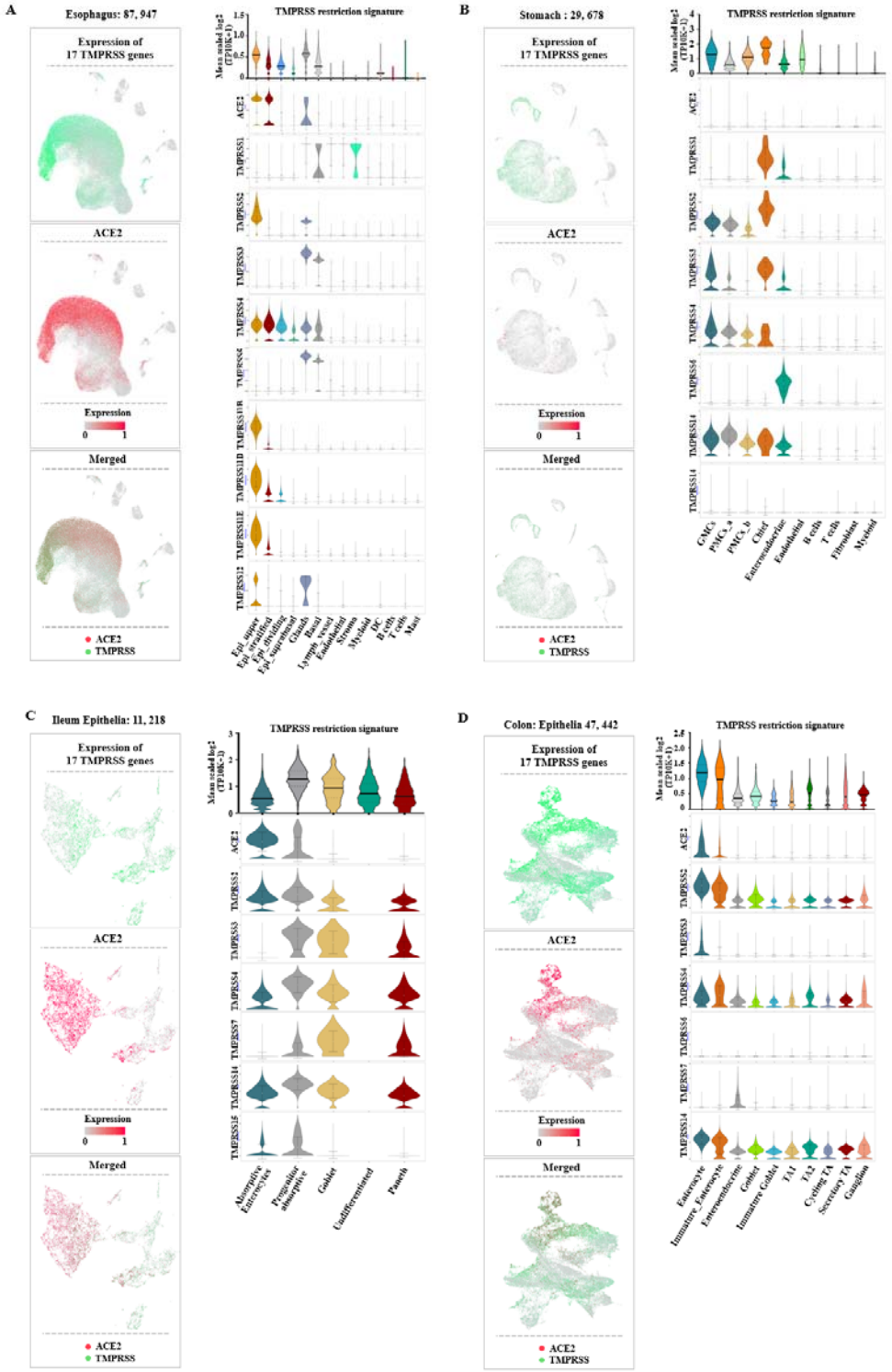


**Figure 6. Single-cell analysis of esophageal cells, gastric mucosal cells, ileal epithelial cells and colonic epithelial cells.**

(A). UMAP plots showing 87,947 esophageal cells. Fourteen clusters are colored, distinctively labeled. Feature plots demonstrate the expression of ACE2, TMPRSS2 and TMPRSS11D across esophageal clusters. Violin plots show the expression of ACE2 and TMPRSS family genes. The expression is measured as the log2 (TP10K+1) value.

(B). UMAP plots showing 29,678 gastric mucosal cells. Ten clusters are colored, distinctively labeled. Feature plots demonstrate expression of ACE2, TMPRSS2 and TMPRSS11D across gastric mucosal clusters. Violin plots show the expression of ACE2 and TMPRSS family genes. The expression is measured as in (A).

(C). UMAP plots showing 11,218 ileal epithelial cells. Five clusters are colored, distinctively labeled. Feature plots demonstrate expression of ACE2, TMPRSS2 and TMPRSS11D across ileal epithelial clusters. Violin plots show the expression of ACE2 and TMPRSS family genes. The expression is measured as in (A).

(D). UMAP plots showing 47,442 colonic epithelial cells. Ten clusters are colored, distinctively labeled. Feature plots demonstrate expression of ACE2, TMPRSS2 and TMPRSS11D across colonic epithelial clusters. Violin plots show the expression of ACE2 and TMPRSS family genes. The expression is measured as in (A).

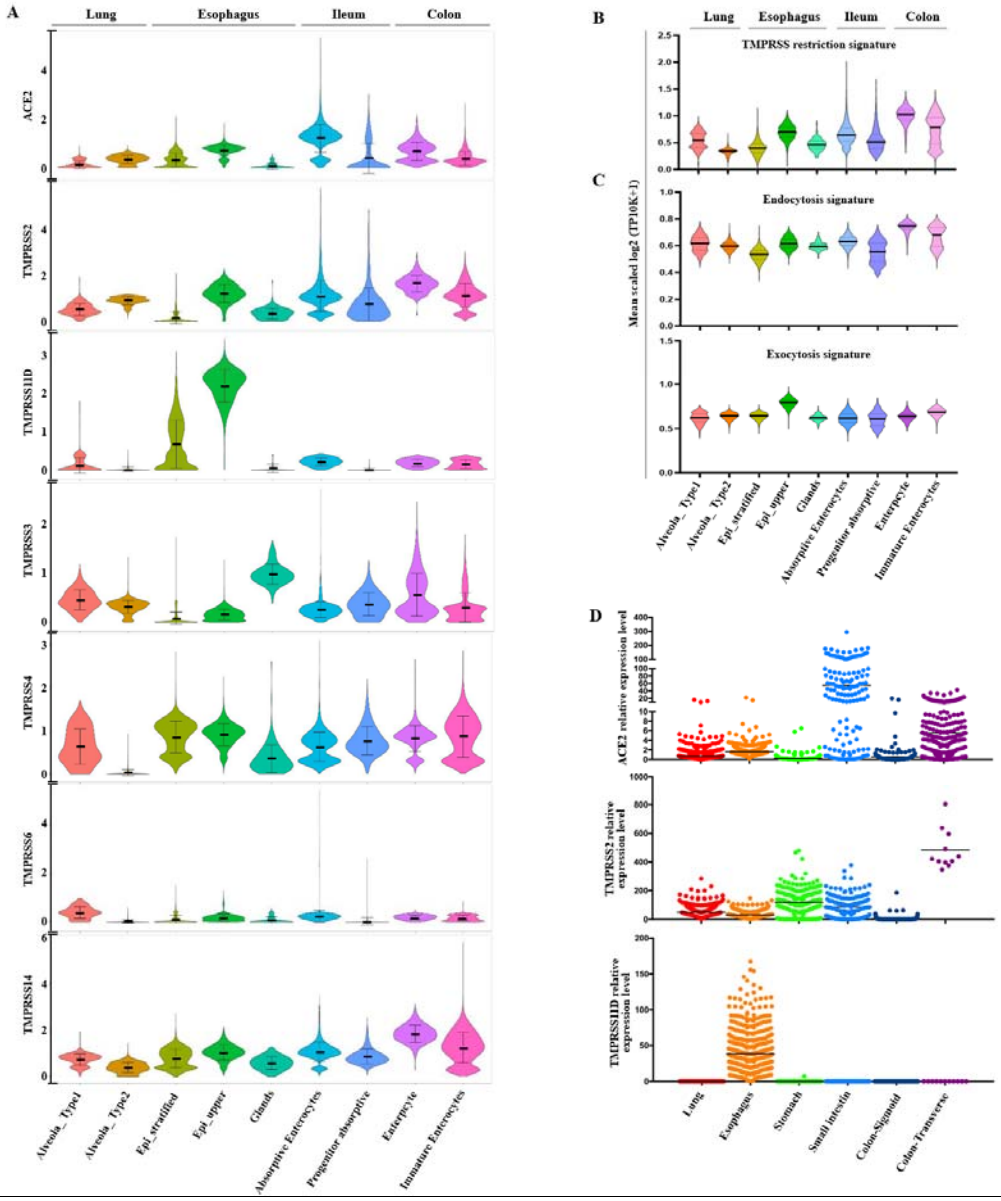


**Figure 7. Expression levels of ACE2, TMPRSSs, TMPRSS11D and functional gene sets in lung and digestive tracts.**

(A). Violin plots showing the expression levels of ACE2 and TMPRSS family genes in 2 lung clusters and 7 digestive tract clusters. The gene expression matrix was normalized and denoised to remove unwanted technical variability across the 4 datasets.

(B). Violin plots showing the mean expression level of TMPRSS family genes. The expression is measured as the mean log2 (TP10K+1) value.

(C). Violin plots showing the expression levels of endocytosis and exocytosis-associated genes. The expression is measured as the mean log2 (TP10K+1) value.

(D). Expression levels of ACE2, TMPRSS2 and TMPRSS11D at RNA level in different tissues. The expression is measured as the pTPM value in the RNA-seq data from the GTEx database.