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

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[3,10]*, 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, Shenyang, 110016, PR China

4 Peking-Tsinghua Center for Life Sciences, TsinghuaUniversity, Beijing, P.R. 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 (Jian-Min Liu)

limingcheng@tongji.edu.cn (Li-Ming Cheng)

brilliant212@163.com (Wang Zhou)

jianruxiao83@163.com (Jian-Ru Xiao)

Abstract

In December 2019, 2019 novel coronavirus (2019-nCoV) induced an ongoing outbreak of pneumonia in Wuhan, Hubei, China. It enters into host cell via cellular receptor recognization and membrane fusion. The former is based on angiotensin converting enzyme II (ACE2). In the latter process, type II transmembrane serine proteases (TTSPs) play important roles in spike protein cleavage and activation. In this study, we used the single-cell transcriptomes of normal human lung and gastroenteric system to identify the ACE2- and TTSP-coexpressing cell composition and proportion. The results revealed that TMPRSS2 was highly co-expressed with ACE2 in the absorptive enterocytes, upper epithelial cells of esophagus and lung AT2 cells, implying the important role of TMPRSS2 in 2019-nCoV infection. Additionally, sequence and structural alignment showed that 675-QTQTNSPRRARSVAS-679 was the key sequence mediating 2019-nCoV spike protein, and there was a inserted sequence (680-SPRR-683). We speculated that this insertion sequence especially the exposed structure at R682 and R683 may enhance the recognition and cleavage activity of TMPRSS2 and then increase its viral infectivity. In conclusion, this study provides the bioinformatics and structure evidence for the increased viral infectivity of 2019-nCoV and indicates transmembrane serine protease inhibitors as the antiviral treatment options for 2019-nCoV infection targeting TMPRSS2.

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 from the human airway epithelial cells and named 2019 novel coronavirus (2019-nCoV)[2]. By analyzing complete genome sequences, 2019-nCoV has an 86.9% nucleotide sequence identity to a severe acute respiratory syndrome (SARS)-like

coronavirus detected in bats (bat-SL-CoVZC45, MG772933.1) and is suggested to be the species of SARS related coronaviruses (SARSr-CoV) by pairwise protein sequence analysis[2].

As a human coronavirus (HCoV), its prerequisite of infection is entering host cells. During this process, the spike (S) glycoprotein of HCoV plays an important role[3]. The surface unit (S1) of S protein mediates the entry into host cells by binding to cellular receptor and the transmembrane unit (S2) subunit harbors the functional elements for the fusion of viral and cellular membranes [4]. It should be noted that the cleavage of S protein is priming to be activated which is essential for viral infectivity [5]. Thus, this process not only needs virus-binding components (cell receptors), but also requires virus protein-cleaving components (cell proteases)[6].

In 2019-nCoV infection, a metallopeptidase, angiotensin converting enzyme II (ACE2) is proved to be the cellular receptor, same as SARS-CoV infection[7, 8]. As the host cell protease of ACE2, the type II transmembrane serine proteases (TTSPs), such as TMPRSS2 and TMPRSS11D, can cleave and activate the SARS-CoV S protein (SARS-S) for membrane fusion and also cleave ACE2 to promote viral uptake [9, 10]. Thus, we suppose that TTSPs may also play a significant role in 2019-nCoV infection.

In order to identify the ACE2- and TTSP-coexpressing cell composition and proportion, we used the single-cell transcriptomes of normal human lung and gastroenteric system based on the public databases. TMPRSS2 was highly co-expressed with ACE2 in absorptive enterocytes, upper epithelial cells of esophagus and lung AT2 cells. In addition, we also explore the cleavage effects of TMPRSS2 in 2019-nCoV based on its structure. A striking finding was that the insertion sequence in 2019-nCoV S protein may increase the cleavage activity of TMPRSS2. Thus, as a previously reported antiviral target, transmembrane serine protease inhibitors may also be used for 2019-nCoV treatment targeting TMPRSS2.

Materials and methods

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 published by E Madissoon et al containing 6 esophageal and 5 lung tissue samples[11]. Three lung datasets were obtained from GSE130148 [12], GSE122960[13] and GSE128169[14], 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[15]. GSE134809 comprises and 11 noninflammatory ileal samples from Crohn's disease patients[16]. The data from Christopher S *et al* included 12 normal colon samples[17].

Quality Control

Cells would be identified as poor-quality ones if they met one of the following thresholds: (1) The number of expressed genes fewer than 200 or greater than 5000. (2) more than 20% of UMIs were mapped to mitochondrial or ribosomal genes.

Data Integration, Dimension Reduction and Cell Clustering

We perform Different methods to process data for according to the downloaded data. *Esophagus dataset*: Rdata were obtained from the research published by E. Madissoon *et al*[11]. Dimension Reduction and clustering had already beenimplemented by the authors.

Lung, stomach and ileum datasets: We utilized functions in the Seurat package to normalize and scale the single-cell gene expression data[18]. 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 perform ed on the single-cell expression matrix using the ‘‘RunPCA’’ function.

We next utilized the ‘‘FindClusters’’ function in the Seurat package 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.

Colon Dataset The single cell data was processed with the R packages LIGER[19] and Seurat[18]. 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 [20]. 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)[21], and The Human Protein Atlas[22] were used to detect gene and protein expression of ACE2 at the tissue levels including normal lung and digestive system, such as esophagus, stomach, small intestine and colon.

Modelling

The structures of 2019-nCoV Spike protein and TMPRSS2 were generated by using SWISS-MODEL online server[23]. The structures were marked, superimposed and visualized by using Chimera[24].

Results

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. The cell type identity in each cluster was annotated by the expression of the known cell type markers.

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 (Fig. 1A). 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 (Fig. 1B). The Violin plots showed that ACE2 and TMPRSS2 were co-expressed in AT1 and AT2

cells. TMPRSS11D was not found in any clusters, whereas ADAM17 was found in all clusters (Fig. 1C). The Immunohistochemistry (IHC) images of ACE2, TMPRSS2 and TMPRSS11D in nomal lung showed a similar result (Fig. 1D).

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, while TMPRSS2 and TMPRSS11D were found in upper epithelial cells. ADAM17 was found in almost all clusters (Fig. 2A).

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 and TMPRSS11D are relatively low in all the clusters, while TMPRSS2 was found in GMCs and PMCs. ADAM17 was also found in all clusters (Fig. 2B).

After quality controls, 11,218 cells and 5 cell types were identified in the ileum epithelia (Fig. 2C). ACE2 and TMPRSS2 were highly expressed in absorptive enterocytes. In the meantime, TMPRSS2 was also lowly expressed in goblet and Paneth cells. TMPRSS11D was not found in any cluster, whereas ADAM17 was found in all clusters except undifferentiated cells (Fig. 2C).

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, so was TMPRSS2. Other clusters also had a lower expression of TMPRSS2. TMPRSS11D was not found in any cluster and ADAM17 was expressed in enteroendocrine cells (Fig. 2D).

The expressions of ACE2, classic TTSPs (TMPRSS2, TMPRSS3, TMPRSS4, TMPRSS6, TMPRSS11D, TMPRSS14 and ADAM17 were detected in lung and digestive tract clusters. An almost consistent expression and distribution was only 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 (Fig. 3A). 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 (Fig. 3B). 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 (Fig. 3C).

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 (Fig. 4A). 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 (Fig. 4B). The structural superimpose of SARS-S (wheat) and 2019-nCoV S protein (cyan) (Fig. 4B).

The obvious insertion sequence in 2019-nCoV S protein is at residue R685, especially at R682 and R683 (the orange frame in Fig. 5A and the yellow structure in Fig. 5B). In addition, the insertion sequence of R682 and R683 was protruded from the molecular surface (Fig. 5B, C).

Structure and catalytic mechanism of TMPRSS2

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 (Fig. 6A, B). The bottom of TMPRSS2 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 Fig. 6C, with arginine, glutamine, leucine and lysine. The Fig. 6D and 6E revealed the state of substrate analogue binding to the catalytic pocket (Fig. 6D, E).

Discussion

The coronaviruses is the common infection source of enteric, respiratory, and central nervous system in humans and other mammals[25]. 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[26]. In December 2019, a novel identified coronavirus (2019-nCoV) induced an ongoing outbreak of pneumonia in Wuhan, Hubei, China [27]. Till now, the pathogenic mechanism of 2019-nCoV is still unclear. In this study, we found the high co-expression between ACE2 and TMPRSS2 in the absorptive enterocytes, upper epithelial cells of esophagus and lung AT2 cells. Additionally, the insertion sequence (680-SPRR-683) in 2019-nCoV S protein may increase the cleavage activity of TMPRSS2 and enhance the viral infectivity, indicating that TMPRSS2 may serve as a candidate antiviral target for 2019-nCoV infection.

Similar to SARS-CoV and MERS-CoV, 2019-nCoV utilizes ACE2 as a receptor for host cell entry [8]. In human HeLa cells, expressing ACE2 from human, civet, and Chinese horseshoe bat can help many kinds of SARSr-CoV enter into the cells including 2019-nCoV [8, 28-30]. During this process, the catalytic domain of ACE2 interacts with a defined receptor-binding domain (RBD) of CTD1 in S protein[7, 31]. The “up” and “down” transition of CTD1 allows ACE2 binding by regulating the relationship among CTD1, CTD2, S1-ACE2 complex and S2 subunit[32].

In this study, we found a strong co-expression between ACE2 and TMPRSS2. The latter can cleave SARS-S near or at the cell surface along with another TTSP human airway trypsin-like protease (HAT), also known as TMPRSS11D, and render host cell entry independent of the endosomal pathway using cathepsin B/L[33]. Different from cathepsin B/L, TMPRSS2 can also promote viral spread in the host [34]. In addition, TMPRSS2 and HAT can also cleave ACE2 to augment about 30-fold viral infectivity and during this process, the metalloprotease ADAM17 can compete with TMPRSS2 [9]. The detailed process is presented in Fig. 7. Due to the critical role of TMPRSS2 in influenza virus and coronavirus infections, serine protease inhibitors have been used in the antiviral therapeutic strategy targeting TMPRSS2 with high antiviral activities, such as camostat, nafamostat and leupeptin[34-37].

The TTSPs showed similar substrate-specificity and catalytic mechanism. As a TTSP, TMPRSS2, located on the cell surface, includes extracellular domain, transmembrane domain and intracellular domain structurally, in which extracellular domain is the main catalytic domain. TMPRSS2 can catalyze the hydrolysis of substrate protein with a strong substrate specificity. Moreover, based on previous study, as for TTSP family members, the P1 position of the cleaved substrates is arginine[38, 39]. This is mainly determined by the special structure of its catalytic pocket. The catalytic pocket of TMPRSS2 is relatively deep, and 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.

Generally, TMPRSS2 is shown to cleave SARS-S at residue R667 which is significantly associated with SARS-S activation for cell-cell fusion[40]. As the S1/S2 cleavage site, R667 is often embedded in different cleavage motifs and cleaved during S protein biogenesis[41]. Besides R667, residue R797 is aslo required for S protein activation by TMPRSS2. As the S2’ cleavage site, it is frequently cleaved during viral entry[41]. Both the hydrolysis sites of R667 and R797 are corresponding to the SARS-S sequence tagged site.

Similar to SARS-S, the 2019-nCoV S protein has three arginine hydrolysis sites located on its surface (R685, R682 and R683), with two from the insertion sequence in 2019-nCoV S protein (680-SPRR-683). As we above-mentioned, the main hydrolysis site of TMPRSS2 is the long-chain positively charged amino acid residue represented by R and more R often increase the viral infectivity[38, 39]. The insertion sequence of R682 and R683 was exposed from the molecular surface. Based on the catalytic mechanism of TMPRSS2, we supposed that the additional arginine residues linked structure exposed from protein surface may be more conducive to the recognition and cleavage activity of TMPRSS2, theraby enhancing the viral infectivity of 2019-nCoV.

Some researchers believed that there were four inserts in the S protein of 2019-nCoV, which may be the result of artificial modification. The SARS-S has been widely studied, including vaccine development and drug design[42, 43]. The key residues 663-VSLLRSTSQ-671 involved in SARS-S protein cleavage site was missed in all structures that have been resolved to date due to its flexibility. The reasonable modification of protein/enzyme function is to modify the key sites of protein/enzyme based on the existing structural studies. We can observe that in the key sequence of S protein splicing, 675-QTQTNSPRRARSVAS-679, only two bases of 685-RS-686 are same as SARS-S, which is most likely not designed for protein function.

Conclusion

ACE2 and TMPRSS2 were highly co-expressed to facilitate the entry of 2019-nCoV into host cells. In addition, the insertion sequence in 2019-nCoV S protein may increase the cleavage activity of TMPRSS2 and enhance the viral infectivity, indicating that transmembrane serine protease inhibitors may serve as the antiviral treatment options for 2019-nCoV infection targeting TMPRSS2.

Reference

1. The L. Emerging understandings of 2019-nCoV. Lancet. 2020.

2. Zhu N, Zhang D, Wang W, Li X, Yang B, Song J, Zhao X, Huang B, Shi W,

1 Lu R, Niu P, Zhan F, Ma X, et al. A Novel Coronavirus from Patients with
2 Pneumonia in China, 2019. N Engl J Med. 2020.
3 3. Walls AC, Xiong X, Park YJ, Tortorici MA, Snijder J, Quispe J, Cameroni E,
4 Gopal R, Dai M, Lanzavecchia A, Zambon M, Rey FA, Corti D, et al.
5 Unexpected Receptor Functional Mimicry Elucidates Activation of Coronavirus
6 Fusion. Cell. 2019; 176: 1026-39.e15.
7 4. Hofmann H, Pohlmann S. Cellular entry of the SARS coronavirus. Trends
8 Microbiol. 2004; 12: 466-72.
9 5. Gallagher TM, Buchmeier MJ. Coronavirus spike proteins in viral entry and
10 pathogenesis. Virology. 2001; 279: 371-4.
11 6. Shulla A, Heald-Sargent T, Subramanya G, Zhao J, Perlman S, Gallagher
12 T. A transmembrane serine protease is linked to the severe acute respiratory
13 syndrome coronavirus receptor and activates virus entry. J Virol. 2011; 85:
14 873-82.
15 7. Gui M, Song W, Zhou H, Xu J, Chen S, Xiang Y, Wang X. Cryo-electron
16 microscopy structures of the SARS-CoV spike glycoprotein reveal a
17 prerequisite conformational state for receptor binding. Cell Res. 2017; 27:
18 119-29.
19 8. P Zhou XY, XG Wang, B Hu, L Zhang, W Zhang, HR Si, Y Zhu, B Li, CL
20 Huang, HD Chen, J Chen, Y Luo, H Guo, RD Jiang, MQ Liu, Y Chen, XR Shen,
21 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. Heurich A, Hofmann-Winkler H, Gierer S, Liepold T, Jahn O, Pohlmann S.
4 TMPRSS2 and ADAM17 cleave ACE2 differentially and only proteolysis by
5 TMPRSS2 augments entry driven by the severe acute respiratory syndrome
6 coronavirus spike protein. J Virol. 2014; 88: 1293-307.
7 10. Li F. Structure, Function, and Evolution of Coronavirus Spike Proteins.
8 Annu Rev Virol. 2016; 3: 237-61.
9 11. Madissoon E, Wilbrey-Clark A, Miragaia RJ, Saeb-Parsy K, Mahbubani KT,
10 Georgakopoulos N, Harding P, Polanski K, Huang N, Nowicki-Osuch K,
11 Fitzgerald RC, Loudon KW, Ferdinand JR, et al. scRNA-seq assessment of
12 the human lung, spleen, and esophagus tissue stability after cold preservation.
13 Genome Biol. 2019; 21: 1.
14 12. Vieira Braga FA, Kar G, Berg M, Carpaij OA, Polanski K, Simon LM,
15 Brouwer S, Gomes T, Hesse L, Jiang J, Fasouli ES, Efremova M,
16 Vento-Tormo R, et al. A cellular census of human lungs identifies novel cell
17 states in health and in asthma. Nat Med. 2019; 25: 1153-63.
18 13. Reyfman PA, Walter JM, Joshi N, Anekalla KR, McQuattie-Pimentel AC,
19 Chiu S, Fernandez R, Akbarpour M, Chen CI, Ren Z, Verma R,
20 Abdala-Valencia H, Nam K, et al. Single-Cell Transcriptomic Analysis of
21 Human Lung Provides Insights into the Pathobiology of Pulmonary Fibrosis.

Am J Respir Crit Care Med. 2019; 199: 1517-36.

14. Valenzi E, Bulik M, Tabib T, Morse C, Sembrat J, Trejo Bittar H, Rojas M, Lafyatis R. Single-cell analysis reveals fibroblast heterogeneity and myofibroblasts in systemic sclerosis-associated interstitial lung disease. Ann Rheum Dis. 2019; 78: 1379-87.

15. Zhang P, Yang M, Zhang Y, Xiao S, Lai X, Tan A, Du S, Li S. Dissecting the Single-Cell Transcriptome Network Underlying Gastric Premalignant Lesions and Early Gastric Cancer. Cell Rep. 2019; 27: 1934-47.e5.

16. Martin JC, Chang C, Boschetti G, Ungaro R, Giri M, Grout JA, Gettler K, Chuang LS, Nayar S, Greenstein AJ, Dubinsky M, Walker L, Leader A, et al. Single-Cell Analysis of Crohn's Disease Lesions Identifies a Pathogenic Cellular Module Associated with Resistance to Anti-TNF Therapy. Cell. 2019; 178: 1493-508.e20.

17. Smillie CS, Biton M, Ordovas-Montanes J, Sullivan KM, Burgin G, Graham DB, Herbst RH, Rogel N, Slyper M, Waldman J, Sud M, Andrews E, Velonias G, et al. Intra- and Inter-cellular Rewiring of the Human Colon during Ulcerative Colitis. Cell. 2019; 178: 714-30.e22.

18. Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM, 3rd, Hao Y, Stoeckius M, Smibert P, Satija R. Comprehensive Integration of Single-Cell Data. Cell. 2019; 177: 1888-902.e21.

19. Welch JD, Kozareva V, Ferreira A, Vanderburg C, Martin C, Macosko EZ.

1 Single-Cell Multi-omic Integration Compares and Contrasts Features of Brain
2 Cell Identity. Cell. 2019; 177: 1873-87.e17.
3 20. Rouillard AD, Gundersen GW, Fernandez NF, Wang Z, Monteiro CD,
4 McDermott MG, Ma'ayan A. The harmonizome: a collection of processed
5 datasets gathered to serve and mine knowledge about genes and proteins.
6 Database (Oxford). 2016; 2016.
7 21. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis:
8 multitissue gene regulation in humans. Science. 2015; 348: 648-60.
9 22. Uhlen M, Fagerberg L, Hallstrom BM, Lindskog C, Oksvold P, Mardinoglu
10 A, Sivertsson A, Kampf C, Sjostedt E, Asplund A, Olsson I, Edlund K,
11 Lundberg E, et al. Proteomics. Tissue-based map of the human proteome.
12 Science. 2015; 347: 1260419.
13 23. Biasini M, Bienert S, Waterhouse A, Arnold K, Studer G, Schmidt T, Kiefer
14 F, Gallo Cassarino T, Bertoni M, Bordoli L, Schwede T. SWISS-MODEL:
15 modelling protein tertiary and quaternary structure using evolutionary
16 information. Nucleic Acids Res. 2014; 42: W252-8.
17 24. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng
18 EC, Ferrin TE. UCSF Chimera--a visualization system for exploratory research
19 and analysis. J Comput Chem. 2004; 25: 1605-12.
20 25. Perlman S, Netland J. Coronaviruses post-SARS: update on replication
21 and pathogenesis. Nat Rev Microbiol. 2009; 7: 439-50.

26. de Wit E, van Doremalen N, Falzarano D, Munster VJ. SARS and MERS: recent insights into emerging coronaviruses. Nat Rev Microbiol. 2016; 14: 523-34.

27. 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, et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. Lancet. 2020.

28. Yang XL, Hu B, Wang B, Wang MN, Zhang Q, Zhang W, Wu LJ, Ge XY, Zhang YZ, Daszak P, Wang LF, Shi ZL. Isolation and Characterization of a Novel Bat Coronavirus Closely Related to the Direct Progenitor of Severe Acute Respiratory Syndrome Coronavirus. J Virol. 2015; 90: 3253-6.

29. Ge XY, Li JL, Yang XL, Chmura AA, Zhu G, Epstein JH, Mazet JK, Hu B, Zhang W, Peng C, Zhang YJ, Luo CM, Tan B, et al. Isolation and characterization of a bat SARS-like coronavirus that uses the ACE2 receptor. Nature. 2013; 503: 535-8.

30. Hu B, Zeng LP, Yang XL, Ge XY, Zhang W, Li B, Xie JZ, Shen XR, Zhang YZ, Wang N, Luo DS, Zheng XS, Wang MN, et al. Discovery of a rich gene pool of bat SARS-related coronaviruses provides new insights into the origin of SARS coronavirus. PLoS Pathog. 2017; 13: e1006698.

31. Li F, Li W, Farzan M, Harrison SC. Structure of SARS coronavirus spike receptor-binding domain complexed with receptor. Science. 2005; 309: 1864-8.

32. Song W, Gui M, Wang X, Xiang Y. Cryo-EM structure of the SARS coronavirus spike glycoprotein in complex with its host cell receptor ACE2. PLoS Pathog. 2018; 14: e1007236.

33. Shirato K, Kawase M, Matsuyama S. Wild-type human coronaviruses prefer cell-surface TMPRSS2 to endosomal cathepsins for cell entry. Virology. 2018; 517: 9-15.

34. Zhou Y, Vedantham P, Lu K, Agudelo J, Carrion R, Jr., Nunneley JW, Barnard D, Pohlmann S, McKerrow JH, Renslo AR, Simmons G. Protease inhibitors targeting coronavirus and filovirus entry. Antiviral Res. 2015; 116: 76-84.

35. 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.

36. Shin WJ, Seong BL. Type II transmembrane serine proteases as potential target for anti-influenza drug discovery. Expert Opin Drug Discov. 2017; 12: 1139-52.

37. Yamamoto M, Matsuyama S, Li X, Takeda M, Kawaguchi Y, Inoue JI, Matsuda Z. 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. Antimicrob Agents Chemother. 2016; 60: 6532-9.

38. Herter S, Piper DE, Aaron W, Gabriele T, Cutler G, Cao P, Bhatt AS, Choe Y, Craik CS, Walker N, Meininger D, Hoey T, Austin RJ. Hepatocyte growth factor is a preferred in vitro substrate for human hepsin, a membrane-anchored serine protease implicated in prostate and ovarian cancers. Biochem J. 2005; 390: 125-36.

39. Limburg H, Harbig A, Bestle D, Stein DA, Moulton HM, Jaeger J, Janga H, Hardes K, Koepke J, Schulte L, Koczulla AR, Schmeck B, Klenk HD, 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. J Virol. 2019; 93.

40. Bertram S, Glowacka I, Muller MA, Lavender H, Gnirss K, Nehlmeier I, Niemeyer D, He Y, Simmons G, Drosten C, Soilleux EJ, Jahn O, Steffen I, et al. Cleavage and activation of the severe acute respiratory syndrome coronavirus spike protein by human airway trypsin-like protease. J Virol. 2011; 85: 13363-72.

41. Millet JK, Whittaker GR. Host cell proteases: Critical determinants of coronavirus tropism and pathogenesis. Virus Res. 2015; 202: 120-34.

42. McPherson C, Chubet R, Holtz K, Honda-Okubo Y, Barnard D, Cox M, Petrovsky N. Development of a SARS Coronavirus Vaccine from Recombinant Spike Protein Plus Delta Inulin Adjuvant. Methods Mol Biol. 2016; 1403: 269-84.

43. Du L, He Y, Zhou Y, Liu S, Zheng BJ, Jiang S. The spike protein of SARS-CoV--a target for vaccine and therapeutic development. Nat Rev Microbiol. 2009; 7: 226-36.

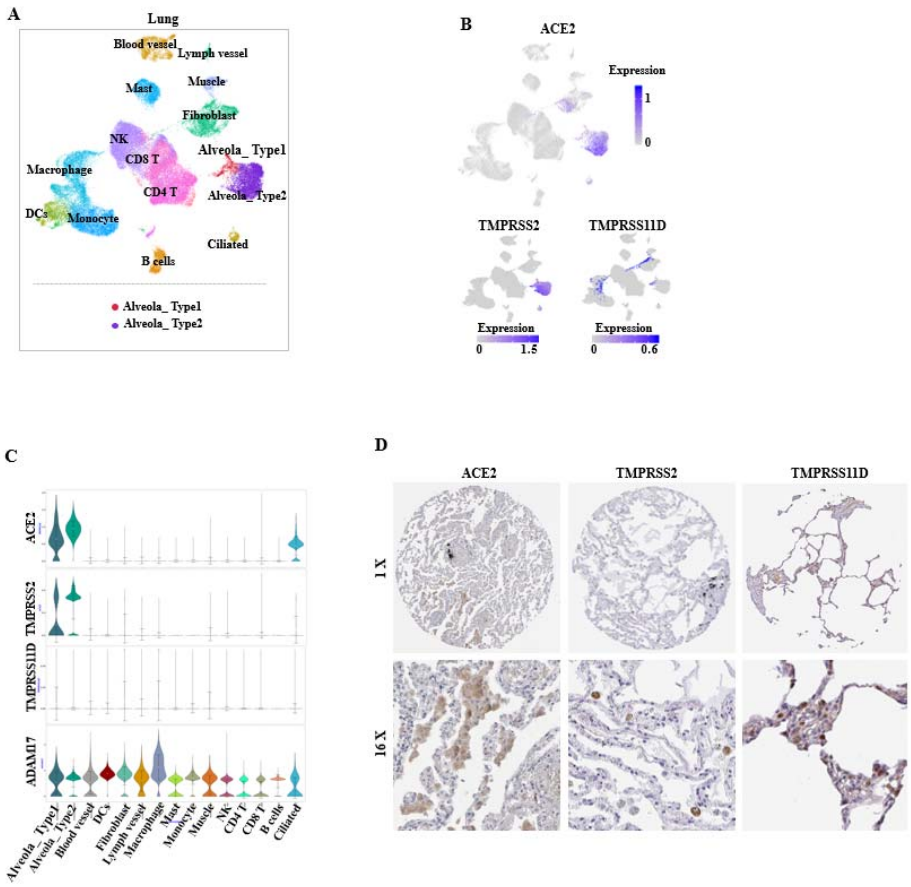

Figure 1. Single-cell analysis and immunohistochemistry of normal lung tissue.

(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). Violin plots showing the expression of ACE2, TMPRSS2, TMPRSS11D and ADAM17 across clusters. The expression is measured as the log2 (TP10K+1) value.

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

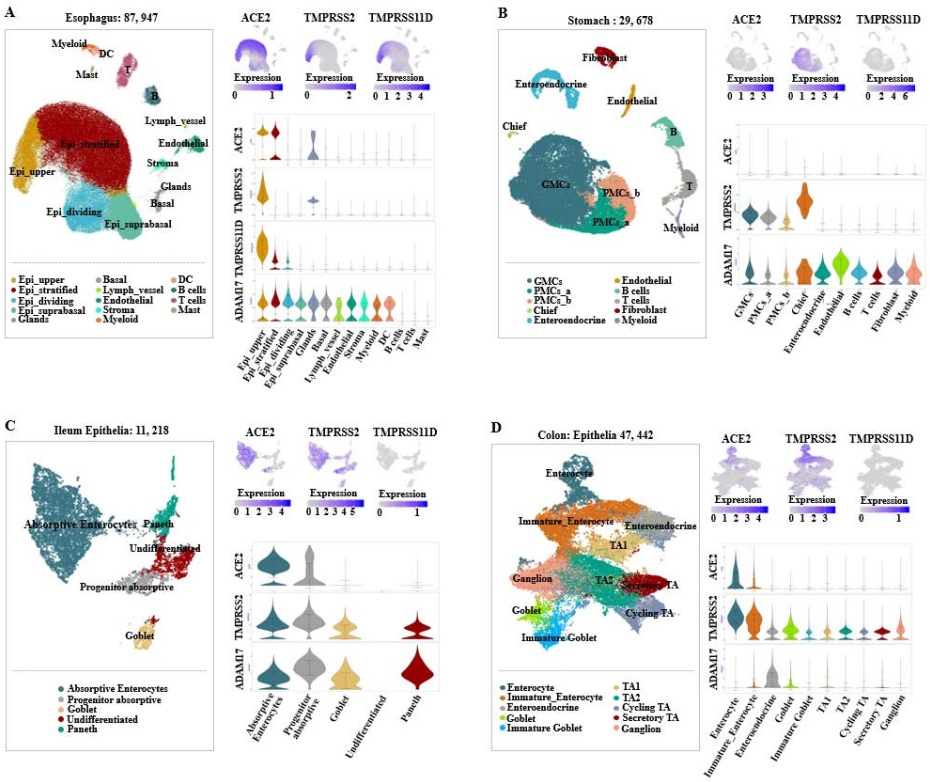

Figure 2. Single-cell analysis of esophageal cells, gastric mucosal cells, ileal epithelial cells and colonic epithelial cells.

(A). UMAP plots showing 87,947 esophageal cells. Fourteen clusters are colored, distinctively labeled. Feature plots demonstrating expression of ACE2, TMPRSS2 and TMPRSS11D across esophageal clusters. Violin plots showing the expression of ACE2, TMPRSS2, TMPRSS11D and ADAM17. 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 demonstrating expression of ACE2, TMPRSS2 and TMPRSS11D across gastric mucosal clusters. Violin plots showing the expression of ACE2, TMPRSS2, TMPRSS11D and ADAM17. The expression is measured as in (A).

(C). UMAP plots showing 11,218 ileal epithelial cells. Five clusters are colored, distinctively labeled. Feature plots demonstrating expression of ACE2, TMPRSS2 and TMPRSS11D across ileal epithelial clusters. Violin plots showing the expression of

ACE2, TMPRSS2, TMPRSS11D and ADAM17. The expression is measured as in (A).

(D). UMAP plots showing 47,442 colonic epithelial cells. Ten clusters are colored, distinctively labeled. Feature plots demonstrating expression of ACE2, TMPRSS2 and TMPRSS11D across colonic epithelial clusters. Violin plots showing the expression of ACE2, TMPRSS2, TMPRSS11D and ADAM17. The expression is measured as in (A).

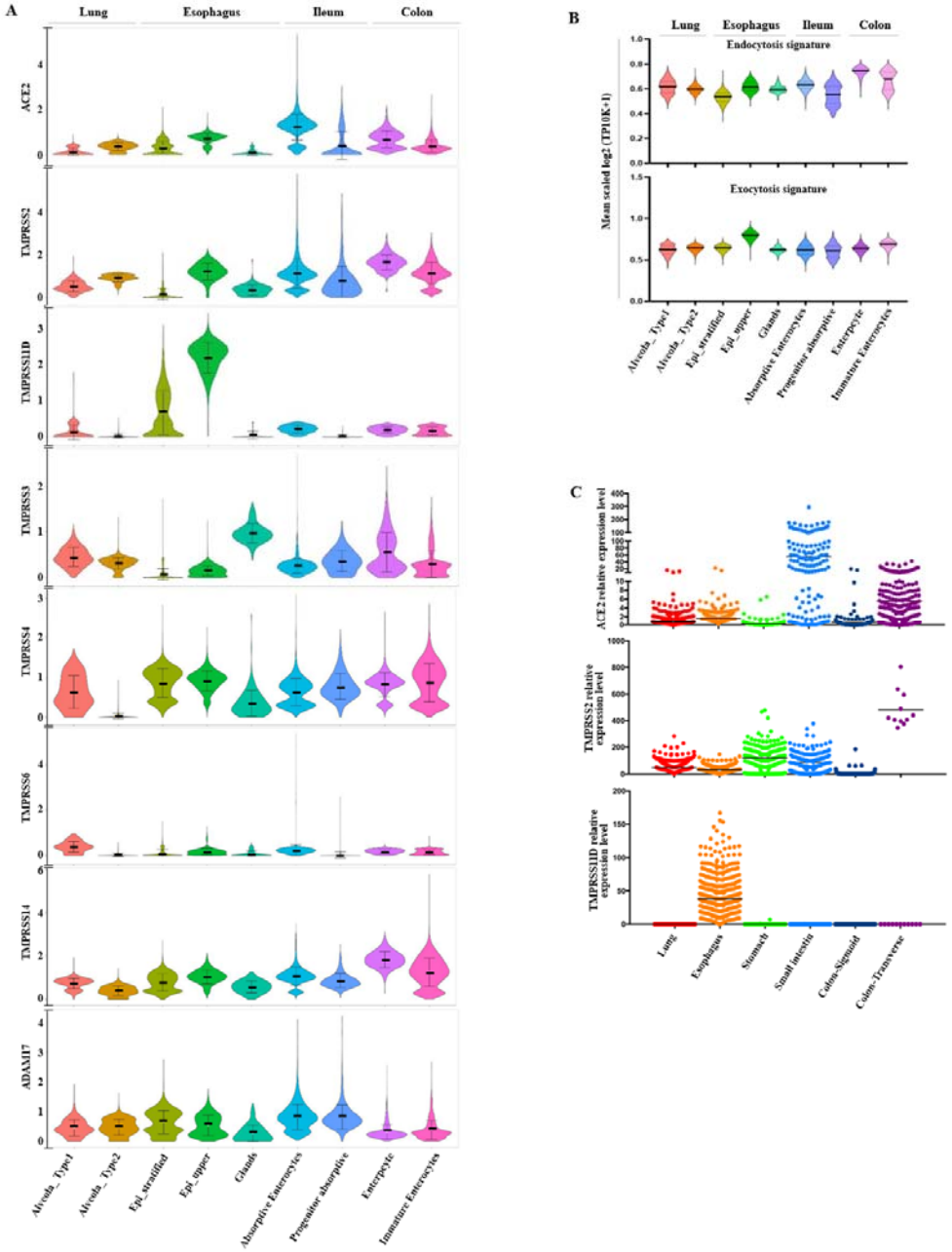

Figure 3. Expression levels of ACE2, TMPRSSs, ADAM17 and functional gene sets in lung and digestive tracts.

(A). Violin plots showing the expression levels of ACE2, TMPRSSs and ADAM17 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 expression levels of endocytosis and exocytosis-associated genes. The expression is measured as the mean log2 (TP10K+1) value.

(C). Expression levels of ACE2, TMPRSS2 and ADAM17 at RNA level in different tissues. The expression is measured as the pTPM value in the RNA-seq data from the GTEx database.

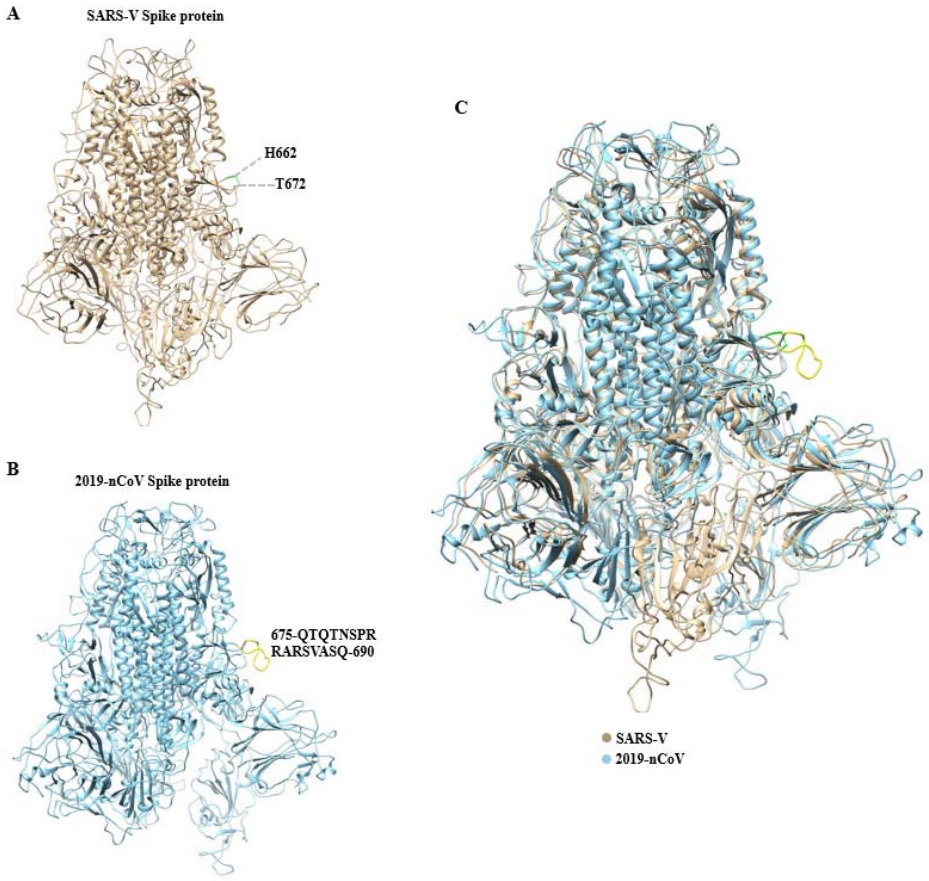

Fig 4. Overall structure of the SARS-CoV Spike protein and 2019-nCoV Spike protein homo-trimers

(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)

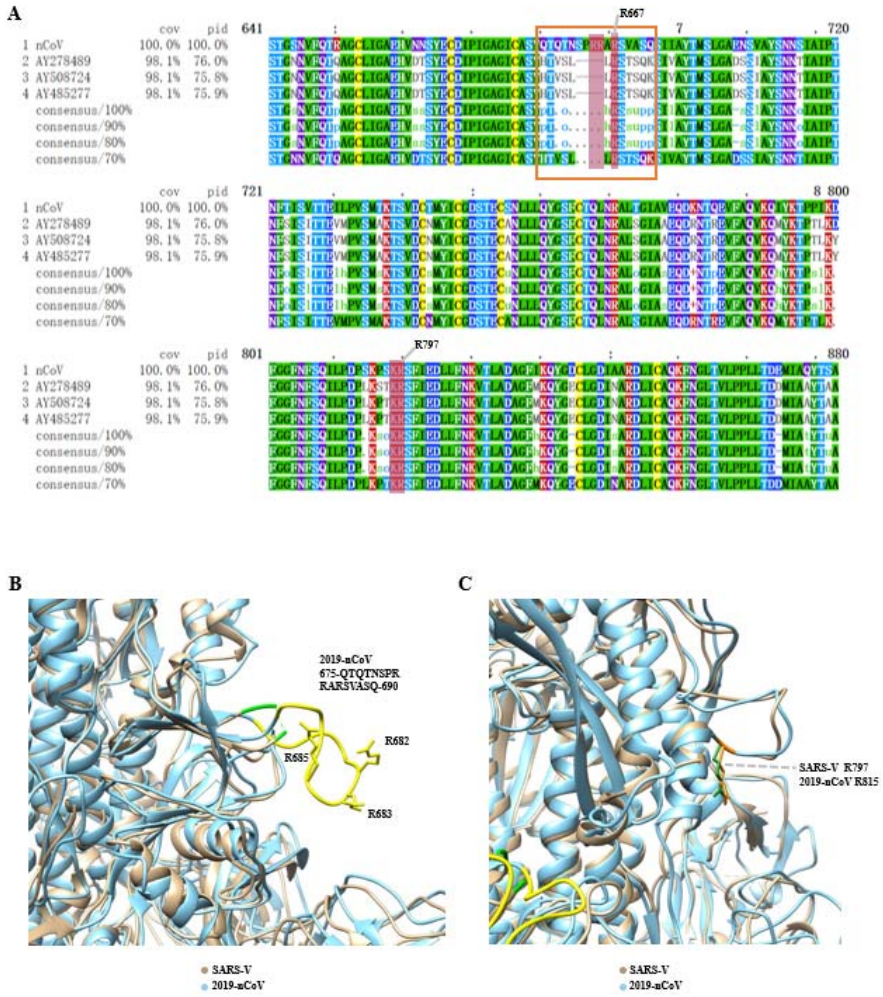

Fig 5. The two potential cleavage sites of SARS-CoV Spike protein and 2019-nCoV Spike protein by TMPRSS2.

(A). Amino acid sequence alignment of the SARS-CoV Spike protein with 2019-nCoV Spike protein. The insert 675-690 that corresponds to 2019-nCoV Spike protein is boxed in orange. Two potential cleavage sites by TMPRSS2, R667 and R797 are marked.

(B-C). Structural alignment 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**).

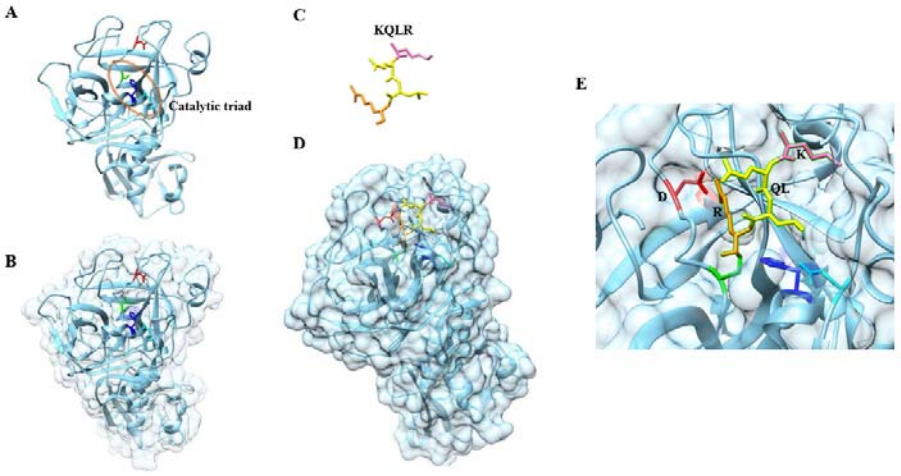

Fig 6. Structure and catalytic mechanism of TMPRSS2

(A-B). Overall structure and surface of TMPRSS2 (Modelled by SWISS-MODEL). 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.

(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.

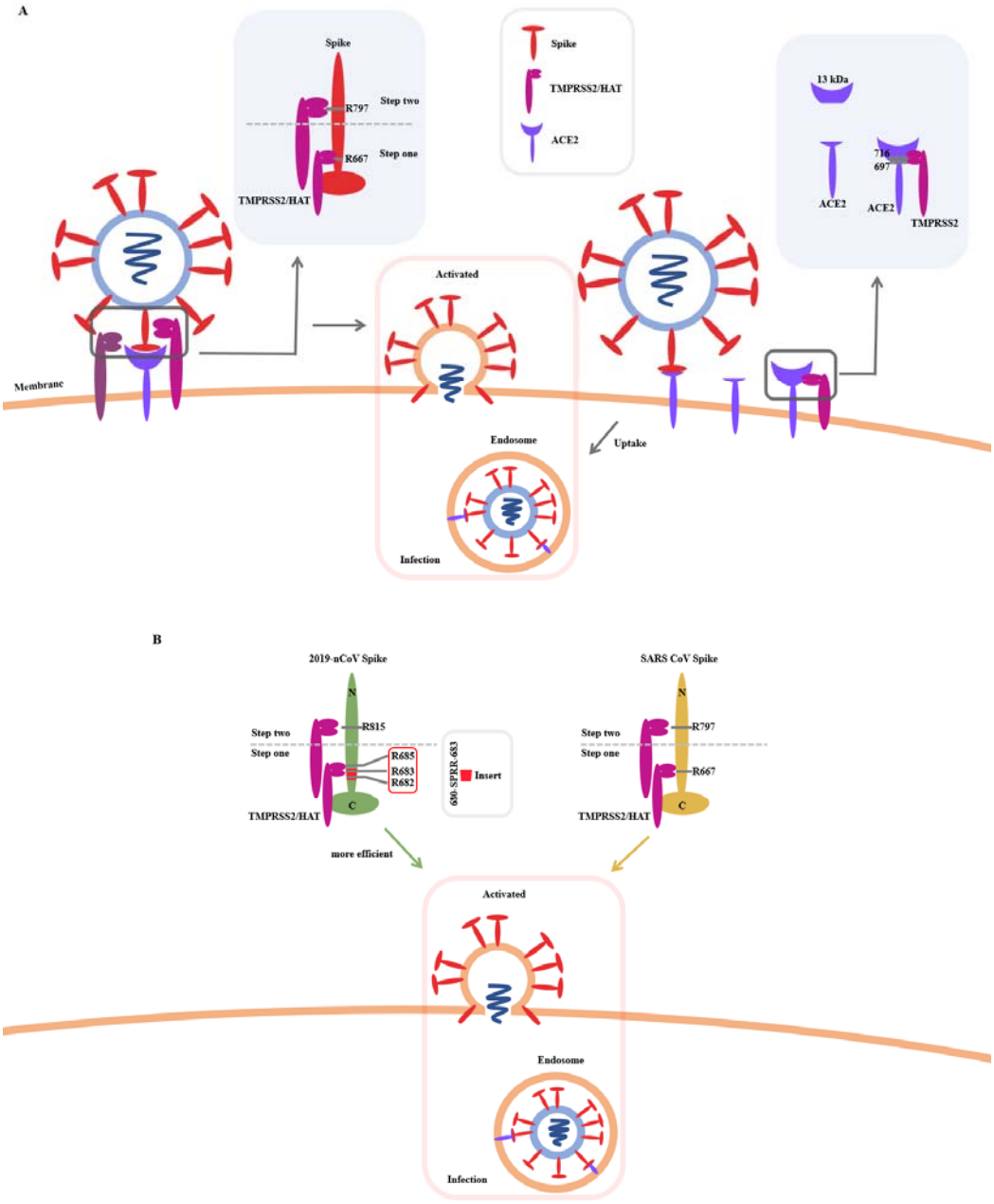

Fig 7：Role of TMPRSS2/HAT proteases in the cellular entry of 2019-nCoV.

(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.