Confronting the COVID-19 Pandemic with Systems Biology

Erica Teixeira Prates, Ph.D.[1‡], Michael R. Garvin, Ph.D.[1‡], Mirko Pavicic, Ph.D.[1], Piet Jones, M.S.[1,2], Manesh Shah, M.S.[1], Christiane Alvarez[1], David Kainer, Ph.D.[1], Omar Demerdash, M.D.[1], B Kirtley Amos, M.S.[5], Armin Geiger, M.S.[1,2], John Pestian[1,3,4], Kang Jin, B.S.[3,4], Alexis Mitelpunkt, M.D.[3,4], Eric Bardes, B.S.[3,4], Bruce Aronow, Ph.D.[3,4], Daniel Jacobson, Ph.D.[1,2,6*]

[*]Corresponding author: Daniel Jacobson, Ph.D.[1,2,6], jacobsonda@ornl.gov, 865 574 6134

[1] Oak Ridge National Laboratory, Biosciences Division, 1 Bethel Valley Road, MS 6420, Oak Ridge, TN 37830
[2] University of Tennessee Knoxville, The Bredesen Center for Interdisciplinary Research and Graduate Education, 444 Greve Hall, 821 Volunteer Blvd. Knoxville, TN 37996-3394
[3] Biomedical Informatics, Cincinnati Children's Hospital Research Foundation
[4] University of Cincinnati, S.10.200, 240 Albert Sabin Way, Cincinnati, OH 45229
[5] University of Kentucky, Department of Horticulture, N-318 Ag Sciences Center, Lexington, KY 40546-0091
[6] University of Tennessee Knoxville, Department of Psychology, Austin Peay Building, Knoxville TN 37996

[‡]Contributed equally

This manuscript has been authored by UT-Battelle, LLC under Contract No. DE-AC05-00OR22725 with the U.S. Department of Energy. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or

reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes. The Department of Energy will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan(http://energy.gov/downloads/doe-public-access-plan).

Summary

Background

The magnitude and severity of the COVID-19 pandemic cannot be overstated. Although the mortality rate is less than SARS and MERS, the global outbreak has already resulted in orders of magnitude more deaths. In order to tackle the complexities of this disease, a Systems Biology approach can provide insights into the biology of the virus and mechanisms of disease.

Methods

Using a Systems Biology approach, we have integrated genomic, transcriptomic, proteomic, and molecular evolution data layers to understand its impact on host cells. We overlay these analyses with high-resolution structural models and atomistic molecular dynamics simulations conducted on the Summit supercomputer at the Oak Ridge National Laboratory.

Findings

Transcriptomic and proteomic data indicate little to no expression of ACE2 in lung tissue. Molecular modeling simulations support ACE2 as the receptor for SARS-CoV-2, but ACE may also act as a receptor for the virus and may be important for entry of SARS-CoV-1. Gene expression data from bronchoalveolar lavage samples from COVID-19 patients identify upregulation of renin, angiotensin, and the angiotensin 1-7 receptor MAS as well as a cellular landscape consistent with large-scale dissolution of lung parenchyma tissues, likely comprised of all lung epithelial cell types as well as lymphatic endothelial cells, but an absence of cells, such as macrophages, normally essential for host defense.

Interpretation

Our analyses indicate that the commonly accepted view that SARS-CoV-2 enters host cells via ACE2 expressed in the lung is unlikely because ACE2 is undetectable there. Instead, given the greater target space of ACE2-positive nasal, oral, and gastrointestinal tissues, a more likely scenario suggests initial infection in those tissues is followed by a secondary infection via migration through the lymphatic system and bloodstream to the lung microvasculature. The elimination of macrophages and complete lack of activated cytokine signature in COVID-19 lung samples suggest that a major component of SARS-CoV-2's virulence is its net effect of causing a functional immune deficiency syndrome. Our structural analysis of the SARS-CoV-2 proteome suggests involvement of the highly conserved nsp5 protein as part of a major mechanism that suppresses the nuclear factor transcription factor kappa B (NF-κB) pathway, eliminating the host cell's interferon-based antiviral response.

Funding

Oak Ridge National Laboratory, Laboratory Directed Research & Development Fund, LOIS:10074

Office of Biological and Environmental Research, United States Department of Energy Office of Science, COVID-19 Testing Research & Development Priorities, ERKPA09.

National Institutes of Health, U24 HL148865: the LungMap Consortium

Research in context

Evidence before this study

We searched PubMed, medRxiv, bioRxiv, arXiv, NCBI GEO & SRA, PDB, Genotype-Tissue Expression Portal, Proteomics DataBase, ALlele FREquency Database, Global Initiative on Sharing All Influenza Data, Human Cell Landscape, Tabula Muris, Google and Google Scholar for peer-reviewed articles, preprints, datasets and research reports on all aspects of coronavirus and the genes, gene/protein expression and cell types mentioned in this manuscript up to April 3, 2020.

Added value of this study

By synthesizing information from a wide variety of data types, sources and scales (from atoms, alleles, transcripts, proteins, protein complexes, cells, tissues, organs and individuals to populations) with a Systems Biology approach we have provided mechanistic insight into the route of infection and spread of SARS-CoV-2 in the human body. In addition, this has provided insights into the body's responses to the virus, including a broader view of the role of the Renin-Angiotensin-System, and the likely targeted tissues and cell types that are responsible for COVID-19 symptoms. We have posited that ACE can be an alternate receptor for SARS-CoV-1 in some tissues and discussed how this may explain the behavior of SARS-CoV-2 in comparison to SARS-CoV-1. We have also used gene expression levels across organs and individuals to explain some of the possible causes for the different responses observed in individuals across the human population. Furthermore, we have elucidated the effects of COVID-19 on specific cell types in lung tissue. By predicting the structures of the entire SARS-CoV-2 proteome and doing a thorough structure-based mutational analysis of all SARS-CoV-2 proteins we have been able to provide functional insights as to the likely impacts of these mutations on transmission, virulence and pathogenicity which will serve as a significant resource to the scientific community for further investigation and validation.

Implications of all the available evidence

This study provides greater comprehension of the mechanistic and molecular underpinnings of COVID-19. In so doing it contributes to the understanding of the evolution of the virus and its varying effects on human hosts. This enhanced understanding has implications on disease spread and prevention, provides preliminary evidence for the design of new studies into the causes of susceptibility and severity of illness and provides information that will aid in the development of drugs (targeting viral and human proteins), vaccines, and diagnostic methodologies.

Introduction

The COVID-19 beta-coronavirus epidemic that originated in Wuhan, China in December of 2019 is now a global pandemic and is having devastating societal and economic impacts. The increasing frequency of the emergence of zoonotics such as Ebola, Severe Acute Respiratory Syndrome (SARS), and Middle East Respiratory Syndrome (MERS) (among others) are of grave concern because of their high mortality rate (15% - 90%). Fortunately, successful containment of those pathogens prevented global-scale deaths. In contrast, the current estimates of mortality for COVID-19 are much lower (~4%), but it is now clear that the virus has not been contained, which may be due to higher rates of asymptomatic transmission of those infected with SARS-CoV-2.[1,2] Given that the SARS-CoV-2 virus has been detected worldwide, rapid measures are needed to address the expanding epidemic.

Paradoxically, an opportunity that was unavailable with SARS, MERS or Ebola has arisen because of the intense, globally distributed focus of medical and scientific professionals on COVID-19 that is providing a wealth of highly diverse information and data types. A Systems Biology approach in combination with supercomputing can integrate these diverse sources of information to provide improved clarity and insights into the mechanisms of disease.

Using existing public multi-omics data layers and supercomputing to analyze available data on the COVID-19 pandemic, we find that viral entry into cells likely involves several components of the Renin Angiotensin System (RAS) besides angiotensin-converting enzyme 2 (ACE2). In addition, analysis of expression data from diverse sources and tissues suggests that the virus evolved to target the digestive system and that damage to the lung may be secondary. We propose that the non-respiratory route of infection is a major contributor to the wide geographic spread of SARS-CoV-2 via "asymptomatic" carriers.

By integrating extensive proteome structural analyses with the multi-omics data layers, the molecular differences between SARS-CoV-1 and SARS-CoV-2 were identified, and provide a basis for explaining the global spread of COVID-19. Structural models of all mature SARS-CoV-2 proteins were generated with an ensemble workflow of structure prediction methods to achieve the highest possible resolution. Along with the models, synopses of the likely relevant variations for each protein are made available for the use of the medical and scientific community. Additionally, we analyze the growing number of SARS-CoV-2 strains in the context of these structures and suggest functional changes that may result in alternative virulence or pathogenicity in different geographic locations.

Results

Several species of coronavirus (with highly variable host mortality rates) have evolved to use proteases in the Renin Angiotensin System (RAS) to gain entry to host cells[3] including the current SARS-CoV-2 virus, (figure 1A). Notably, in addition to ACE2, other RAS proteases that have not been widely discussed, such as those that produce angiotensin IV (ANPEP) and angiotensin 1-7 (ACE), may be important, given that they activate proinflammatory genes, which are components of the innate antiviral response and may be a key factor in mortality from coronaviruses.[4] The expression profile of RAS genes of cells from bronchoalveolar lavage samples (BAL) taken from individuals with COVID-19 indicate upregulation of renin, angiotensin, ACE2, and MAS (figure 1B). Induction of this pathway from viral infection also indirectly implicates ACE, because it is

needed for the production of the ligand for MAS: Ang 1-7. In addition, during the current pandemic, it may be informative to track several factors that are known to alter RAS, including drugs (R_X in figure 1A), outcomes of infection (*e.g.* dehydration from diarrhea), and genetic polymorphisms. For example, an *Alu* deletion in ACE is responsible for 40% of the variance in circulating levels of that enzyme[5] and the frequency of the allele varies considerably among populations (table S1).

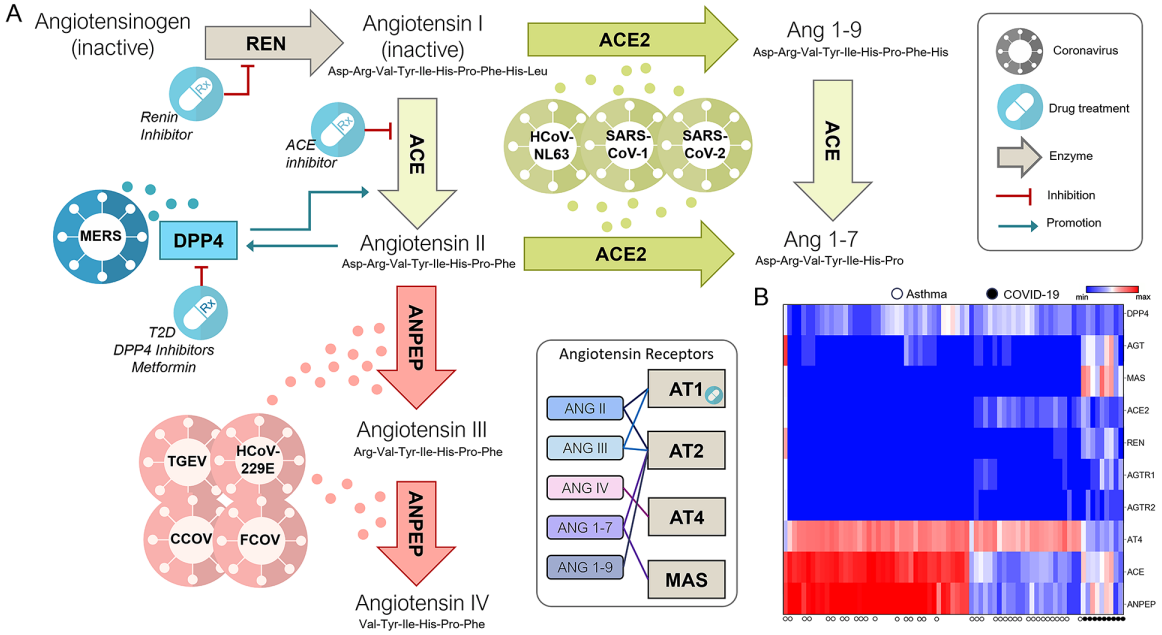


Figure 1. Brief summary of the Renin Angiotensin System (RAS). A) Angiotensinogen (AGT) is converted to angiotensin I by renin (REN), then to angiotensin II by Angiotensin Converting Enzyme (ACE), which is further cleaved to Angiotensin III and IV by Alanyl Amino Peptidase N (ANPEP). ACE2 converts angiotensin I and II to smaller peptides, angiotensin 1-9 or angiotensin 1-7. Several drugs used to treat hypertension and diabetes act on RAS, such as ACE inhibitors (e.g. captopril), RAS Inhibitors, and angiotensin receptor blockers. The various angiotensin peptides have different effects depending on which receptor is targeted. Four receptors are widely reported, namely, AT1, AT2, AT4, and MAS. Several coronaviruses have evolved to use RAS proteins to enter cells (color coded according to known entry point) but mortality rates vary widely (MERS, 50%; SARS, 15%; COVID-19, 4%; HCoVNL63, rare). B) Cells from bronchoalveolar lavage samples of COVID-19 infected individuals show increased expression of REN, AGT, and MAS1 and downregulation of ANPEP and AT4 (heatmap lower right). Dehydration due to diarrhea, vomiting and perspiration could further alter RAS and affect virus pathogenicity as could several classes of widely prescribed drugs. Common side effects of these drugs are diarrhea and dry cough, similar to COVID-19. Deletion of an *Alu* sequence in ACE is correlated with higher blood levels of the enzyme and therefore this allele may be an informative predictor for outcomes of COVID-19. Focus on the entire RAS system is warranted in the continuing pandemic. Abbreviations: MERS (Middle East Respiratory Syndrome), TGEV (Transmissible Gastroenteritis Virus), FCoV (Feline Coronavirus), CCoV (Canine Coronavirus), HCoV229E (Human Coronavirus 229E), HCoV-NL63 (Human Coronavirus NL63), SARS (Severe Acute Respiratory Syndrome), and SARS-CoV-2 (cause of current COVID-19 pandemic).

Is SARS-CoV-2 entering cells via ACE2 in the lungs?

A thorough analysis of the data from the Genotype-Tissue Expression (GTEx) Portal (https://www.gtexportal.org), the Proteomics DataBase (Proteomics DB) (https://www.proteomicsdb.org/) and the Human Cell Landscape[12] indicates ACE2 is either expressed at very low levels or is not detectable in tissues that are currently thought to be major entry points for SARS-CoV-1 and SARS-CoV-2 such as lung (figures 1, 2, 4, S1, Appendix 1.2). However, ACE2 is highly expressed in digestive tissue (gut, colon, rectum, ileum), reproductive tissue (testis, ovary), kidney, thyroid, adipose, breast and heart, possibly providing different viable routes for viral infection. Indeed, SARS-CoV-2 has been detected in feces in some individuals[6] and gastrointestinal disorders are now being reported as an early sign of COVID-19 infection.[7] Additionally, the intestine has been shown to be a viable route for infection for MERS-CoV,[8] suggesting that it survives the low pH of the stomach and is potentially replicating in digestive tissue.

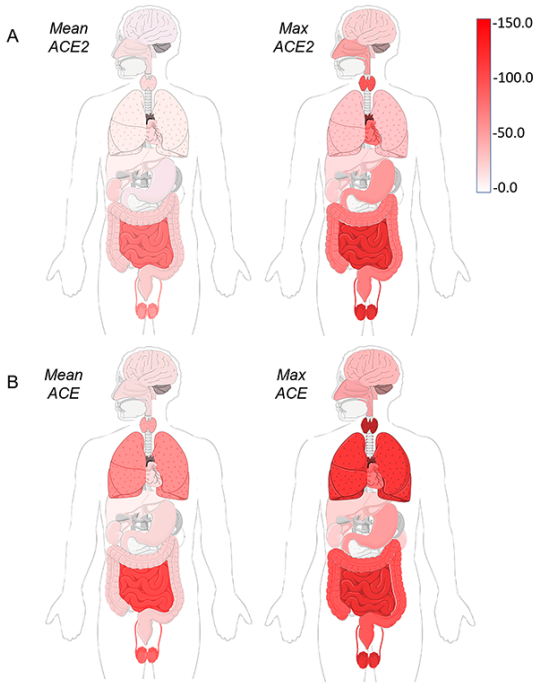


Figure 2. Mean and maximum expression levels (TPM) of ACE2 and ACE in different organs. Both ACE2 (A) and ACE (B) are highly expressed in digestive tissue and testis. Only ACE is highly expressed in the lung. Both ACE and ACE2 expression levels vary across individuals in the GTEx population.

Here, we revisit the hypothesis that infection occurs via the respiratory tract and posit that ACE may be the receptor for SARS-CoV-2 in the lungs, given that the mRNA levels of ACE are 39-fold higher than ACE2 which is often less than 1 TPM in lung tissue (figure 2, figure S1). Similarly, the ACE protein is highly expressed in lungs, tonsils, salivary glands, and oral epithelium, whereas ACE2 is not detected in these tissues and yet, most citations identify ACE2 as the only receptor for SARS-CoV-1, often stating that it is highly expressed in the lung, when GTEx expression data, our analysis of the Human Cell Landscape and a careful review of the

literature (Appendix 1.1) reveals that it is not. Furthermore, there is experimental evidence that ACE can mediate SARS-CoV-1 infection (Appendix 1.2*)*.

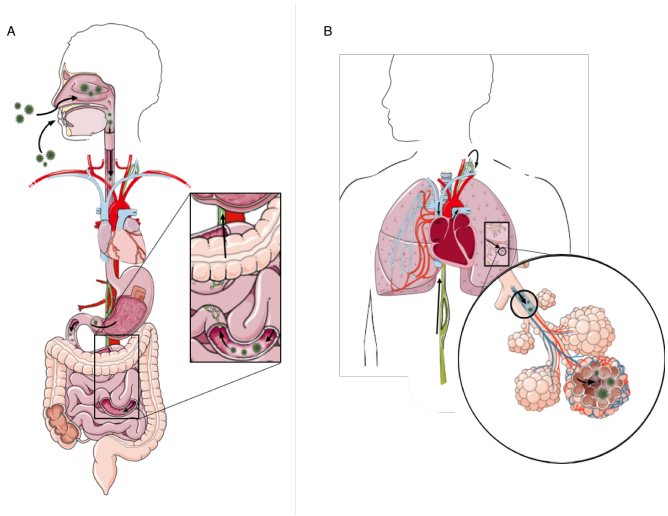


Figure 3. Possible oral/nasal route of infection. Intestinal tissue (high levels of ACE and ACE2) becomes a major reservoir of SARS-CoV-2 virus, that then migrates through the lymphatic system up to the thoracic duct, followed by entry into the venous system at the venous angle (Pirgoff's angle) between the left subclavian vein and left internal jugular vein (A). From there, it passes through the heart to reach the lungs, contacting the large surface area of the lung microvasculature, becoming a secondary site of infection. Infection in the lung may be facilitated by high levels of ACE and/or in individuals with unusually high levels of ACE2 expressed in the lung (B). From there the virus can spread through the circulatory systems to other tissues with high ACE/ACE2 levels that correspond to symptoms being reported in COVID-19, *e.g.* testes[9], thyroid (fatigue[10,11]), brain epithelial cells (headache[10]), etc. This path of infection to the lung is supported by both bulk tissue and single cell gene expression patterns.

What does cell-specific data tell us about ACE and ACE2 distribution?

We found considerable support of these differences in tissue expression of ACE vs. ACE2 in scRNA-Seq datasets (*e.g.* The Human Cell Landscape[12] and The Mouse Cell Atlas[13], figure S2). A single-cell atlas of human respiratory tissues available on the ToppCell server (toppcell.cchmc.org, Appendix 2.1) reveals that, in contrast to high levels of expression of ACE in lung endothelial cells, ACE2 is very sparsely expressed in several different lung cell types (figure 4). Single cell data from the mouse atlas revealed that ACE expression is high in lung endothelial and mast cells, trachea, pancreatic endothelial and stellate cells, and heart fibroblasts (figure S2). High ACE2 expression occurred in the large intestine epithelium, pancreatic endocrine cells, heart myofibroblasts and the tongue keratinocytes (figure 4). Taken together, these data indicate that ACE2 is more sparsely expressed than ACE and is not detected in the endothelial or fibroblast lineages that express ACE (figure 4). However, high levels of ACE2 were found in differentiating keratinocytes of both nasal passage and the tongue (figure 4), enterocytes of the small and large bowel, and several important endocrine cells of the pancreas. Finally, nasal epithelial brushings have been an important site for detection of SARS-CoV-2 and we therefore sought to determine if an ACE2-positive signal in nasal brushings could indicate the presence of virus-susceptible cell type(s). We analyzed expression patterns of 937 genes that are highly correlated/anti-correlated with ACE2 in lung RNAseq data from GTEx and from RNA measured

in nasal epithelium cells (from nasal brushings) in individuals of different ages. Although ACE2 is not highly expressed, there appear to be sub clusters of individuals in genes that have to do with development of the trigeminal nerve and microvilli in teens and adults. Infants do not express ACE2 to any appreciable extent (figure 5).

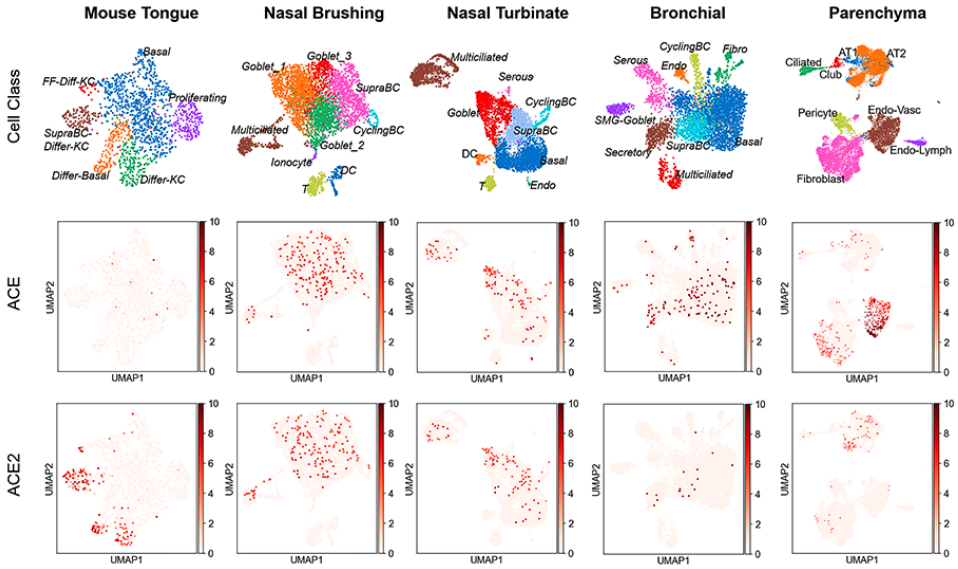


Figure 4. A variety of human single cell datasets from nose[14] to lower lung[15] were used to explore gene expression in previously identified cell lineages. The most consistent site of ACE2 expression was in differentiating keratinocytes of nasal brushing and nasal turbinates. Similarly, mouse ACE2 was most consistently expressed in differentiating keratinocytes of the tongue. Color scales are Log2(TPM+1). FF-Diff-KC, filiform differentiated keratinocytes; SupraBC-Differ-KC, suprabasal differentiating keratinocytes; Differ-Basal, differentiating basal cells; Differ-KC, differentiated keratinocytes; SupraBC, suprabasal cells; cyclingBC, cycling basal cells; T, T cells; DC, dendritic cells; Endo, endothelial cells; SMG-Goblet, submucosal gland Goblet cells; Fibro, Fibroblasts; AT1, Alveolar Type 1 cells; AT2, Alveolar Type 2 cells; Endo-Vasc, vascular endothelial cells; Endo_Lymph, lymph vessel endothelial cells.

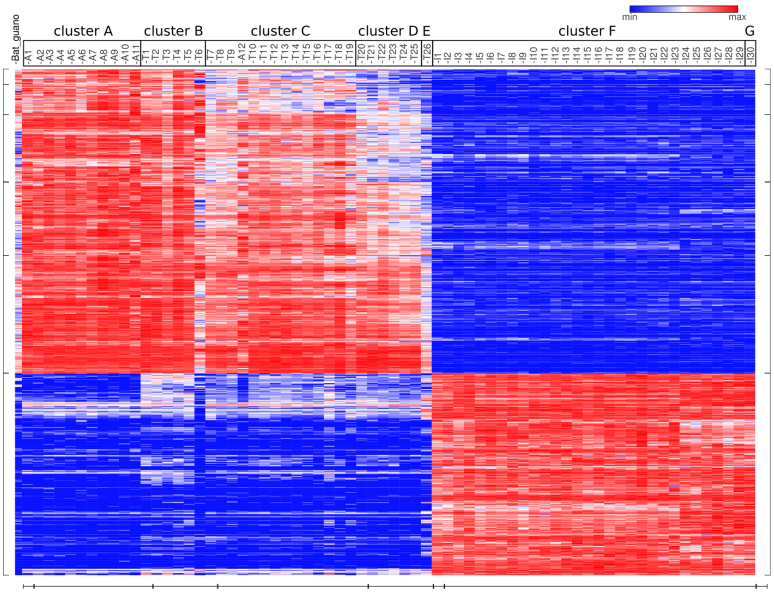


Figure 5. Clustering of 937 differentially (vertical axis) expressed genes that were also highly correlated with ACE2 in GTEx from lung (+/- 0.8) in nasal brushings in different age classes (horizontal axis). ACE2 was not detected in the majority of infant (I) nasal epithelium and was moderately expressed in adults (A) and teens (T) (range 1.0-5.2 TPM). Hierarchical clustering identified sub-patterns of expression within those expressing ACE2 (e.g. cluster C is specific to one adult and a small number of younger individuals). A GO analysis of the genes upregulated with ACE2 (top left quadrant) identified a 30-fold enrichment for trigeminal ganglion development and 26-fold enrichment for microvilli assembly Those that were down-regulated compared to ACE2 (*blue*, lower left quadrant) did not show significant GO enrichment but were mainly non-coding RNAs. A sample from a bat guano collector in Thailand infected with a HKU1 coronavirus appears on the far left.

ACE vs. ACE2 as the entry point for SARS-CoV-1 and SARS-CoV-2

Closer scrutiny of the initial report identifying ACE2 as the receptor for SARS-CoV-1 suggests it may be specific to kidney-derived cell lines and the potential for ACE as the receptor has not been adequately evaluated (Appendix 2.2). Applying computational methods to study the molecular biophysics of viral and host machinery has been an effective way of checking this hypothesis and identifying key molecular elements of interactions with these putative receptors that can then be further explored in depth.

Currently, high-resolution structures of the receptor binding domains (RBD) of the SARS-CoV-1 and SARS-CoV-2 spike glycoproteins in complex with the peptidase domain of ACE2 are available, and enable a detailed description of the interfacial interactions.[16] In order to explore the involvement of ACE and ACE2 in SARS-CoV-1 and -CoV-2 infection, we used these structures as a starting point for a comparative atomistic molecular dynamics (MD) study of the RBDs of the two viruses in complex with ACE2, here referred as RBD1-ACE2 and RBD2-ACE2 for SARS-CoV-1 and SARS-CoV-2, respectively. Additional simulations of the putative complexes with ACE were performed for the preliminary study of the stability of these complexes (RBD1-ACE and RBD2-ACE).

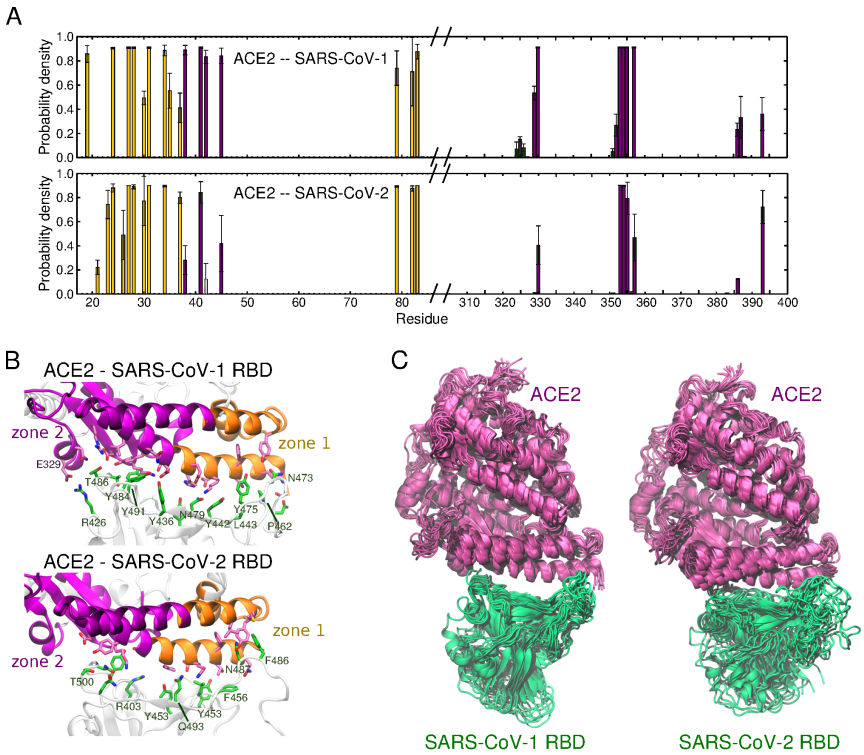


Figure 6. Analysis of simulations of SARS-CoV-1 and SARS-CoV-2 RBDs in complex with ACE2. A) Probability density of residues from ACE2 forming contacts with the RBDs. A maximum distance of 4 Å between any atom in a pair of residues was established. Bars with a standard deviation greater than 50% of the probability density are considered transient contacts in the simulations and not included in these plots. The colors of the bars correspond to zone 1 and zone 2 of ACE2, defined in (B), which shows the residues involved in contacts formed during more than 70% of the simulation time. RBD and ACE2 residues are represented as licorices, in *green* and *pink*, respectively. C) Superimposition of frames in a representative simulation of RBD1-ACE2 (*left*) and RBD2-ACE2 (*right*), using the initial position of ACE2 as reference for alignment. RBDs and ACE2 are represented in *green* and *pink*, respectively.

The amino acid sequences of SARS-CoV-1 and SARS-CoV-2 spike glycoproteins are 77% identical. The RBD is a region that harbors a high concentration of non-conservative substitutions, remarkably at regions that are known to directly bind to the host receptor (figure S3). These differences are expected to considerably influence the affinity of spike for ACE2. Tian *et al*. measured the binding of RBD2 to ACE2 with a biolayer interferometry binding assay and reported similar affinity of RBD1 and RBD2 to ACE2 (K_d=15·0 nM and 15·2 nM, respectively).[17] In contrast, Wrapp *et al*. reported a 10- to 20-fold higher affinity of RBD2 to ACE2, compared to RBD1.[18] Despite the disagreement on the relative affinity for ACE2, both reports establish that RBD2 forms a stable complex with ACE2, which was shown to be a host receptor for SARS-CoV-2 cell entry.[19]

Accordingly, RBD1-ACE2 and RBD2-ACE2 complexes are stable along all of the conducted MD simulations. The computed average number of contacts (residues with Cα less than 8 Å distant) is slightly smaller between RBD2 and ACE2 (22 ± 4) than in the complex with RBD1 (25 ± 3). This suggests that, if RBD2 has higher affinity to ACE2 than RBD1, as reported by Wrapp *et al*., this is resulting from stronger rather than additional interactions in RBD2-ACE2. As shown in figure 6A, the profile of ACE2 residues involved in persistent interactions with the RBDs is consistent in triplicate simulations. The contact profile, figure 6A-B, shows a slightly higher density of stable

contacts in zone 2 for RBD1-ACE2 compared to RBD2-ACE2, that is likely due to the additional salt bridge formed by RBD1 Arg^{426} and ACE2 Glu^{329}, which is lost with the substitution $Arg^{438}Asn$ in RBD2, as well as due to the presence of RBD1 Tyr^{484} (Gln^{498} in RBD2), packing with the hydrophobic tail of ACE2 Lys^{353}. The weaker interactions in zone 2 are at least partially compensated in RBD2-ACE2 in zone 1, where hydrophobic packing is enhanced by the bulky RBD2 Phe^{486} and Phe^{456} (Leu^{472} and Leu^{443} in RBD1, respectively).

The analysis of the conformational dynamics of the two complexes revealed an important structural difference that was not captured in previous analyses of static structures. The RBD from SARS-CoV-2 exhibits considerably higher flexibility compared to the RBD from SARS-CoV-1, as evident in the superimposition of the simulation frames (figure 6C). A close inspection of the structure strongly suggests that this is caused by the substitution of Lys^{447} by Asn^{460} in RBD2, resulting in the loss of a salt bridge with Asp^{407}, or Asp^{420} in RBD2 (figure S5). The weaker interaction with α3 helix "unlocks" loop β4-5, that mostly interacts with zone 1 of ACE2. The elongation of the loop with the additional glycine, Gly^{482}, further contributes to the higher conformational flexibility of the SARS-CoV-2 RBD.

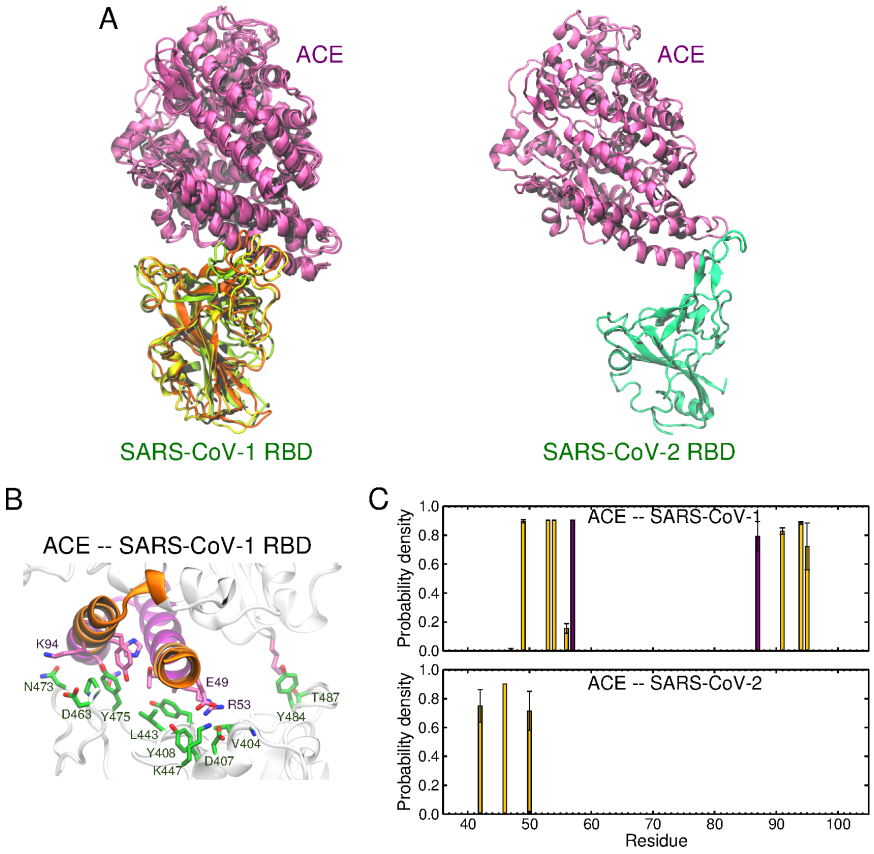


Figure 7. Analysis of simulations of SARS-CoV-1 and SARS-CoV-2 RBDs in complex with ACE. A) Superimposition of the last frames of the simulations of RBD1-ACE (*left*). For visual clarity, because the relative orientation of the proteins in RBD2-ACE2 is very flexible due to the small surface of contact, we only show the last frame of a representative simulation of RBD-ACE2 (*right*). RBDs and ACE2 are represented in *green* and *pink*, respectively. B) Residues involved in contacts formed during more than 70% of the simulation time. RBD and ACE residues are represented as licorices, in *green* and *pink*, respectively. C) Probability density of residues in ACE forming contacts with the RBDs. A maximum distance of 4 Å between any atom in a pair of residues was established. Bars

with standard deviation higher than 50% of the probability density are considered transient contacts in the simulations and not included in these plots. The colors of the bars correspond to zone 1 and zone 2 of ACE, shown in B.

In order to perform preliminary tests of the hypothesis that ACE is an alternative receptor for SARS-CoV-1 and SARS-CoV-2, we also conducted MD simulations of RBD1-ACE and RBD2-ACE. The peptidase domains of ACE and ACE2 are 40% identical and have a very similar fold (RMSD 6.6 Å) and therefore we assumed that the interaction with RBDs would occur in the same region in the fold. We built the initial structures by alignment and replacement of ACE2 by ACE in the complexes described above. On the putative complex interface, only 35% of the residues are similar or identical to residues identified as stable in the interaction of ACE2 with RBD1 or RBD2. Therefore, local structural rearrangements are expected to happen in the built RBD1/2-ACE complexes along the MD simulations. In order to allow structural adjustments to happen, we conducted long equilibration simulations involving multiple steps for gradual relaxation of the system (Appendix 2.3).

Our analysis verifies that, for both systems, the RBDs remain bound to ACE during the simulations. In all independent simulations of RBD1-ACE, a significant reorientation of the RBD1 is observed, so that the loop ß5-6 slides towards the center of the α1 helix of ACE. Figure 7A shows the superimposed last frames of the three simulations of RBD1-ACE. Persistent interactions are established involving the formation of three salt bridges, namely, Asp^{407}-Arg^{53}, Lys^{447}-Glu^{49}, Asp^{493}-Lys^{94}, from RBD1 and ACE, respectively (figure 7B).

The RBD2-ACE also converges to a common configuration in two of the three independent simulations of RBD2-ACE, with only few residues attaching the proteins together (figure 7A). In these simulations, the loop ß4-5 anchors the RBD2 to the N terminal of α1 helix and the nearby region of α2 helix of ACE, mostly involving only hydrophobic contacts between Phe^{456} and Tyr^{489} of RBD2 at the N-terminal of α1 (figure 7C).

Despite the fact that MD simulations of hundreds of nanoseconds cannot provide reliable quantitative estimates of binding affinity, they can be effectively used to explore the relative stability of the studied complexes. Taken together, our simulations demonstrate the convergence of stable and strong interactions between ACE and SARS-CoV-1 and weaker but still possible interactions between ACE and SARS-CoV-2, suggesting that ACE may allow for infection in tissues with low or undetectable levels of ACE2, but high ACE expression. Importantly, our results emphasize that future experiments should be designed to include the complete native spike protein, since intra-spike interactions of the RBD in the closed conformation are an important element competing with the stabilization of the open conformation of the spike via interaction with the host receptor.

The forgotten co-receptors of SARS-CoV-1 infection - other co-receptors that may be involved

In the study that originally identified ACE2 as the receptor for SARS-CoV-1, two additional proteins were found to bind the spike glycoprotein - major vault protein (MVP) and myosin 1B (MYO1b) - but they were not further explored at that time.[20] More recent data has indicated that these are, in fact, important elements for viral infection. MYO1b is directly linked to SARS because it is necessary for internalization of feline coronavirus (which also enters cells via the RAS) from the cell surface.[21] MVP is the major component of Vault ribonucleoprotein particles,

which are known to be involved in multidrug resistance in cancer cells[22,23] and are a central component of the rapid innate immune response in cells exposed to pathogens.[24]

As part of the immune response to lung infection by the bacterium *Pseudomonas aeruginosa,* MVP is recruited to lipid rafts that mediate internalization of the pathogen via endocytosis, and it is likely essential to initiate an apoptotic cascade.[25] These lipid rafts include 127 other proteins within them (table S2, henceforth referred to as "raft-127 genes"), including angiotensin II receptor-associated protein (AGTRAP), whose function is, in part, to internalize AT1-type receptors via endocytosis[26] and cycling to the endoplasmic reticulum and Golgi apparatus.[27] In addition, several antiviral response genes are present including MAVS, CD81, ITGA2, NECTIN1, TFRC, and LDRL. NECTIN1 is a receptor for many different viruses including herpes, measles, and pseudorabies virus, and LDLR is a known receptor for vesicular stomatitis virus[28] and facilitates production of hepatitis C virus.[29] Given the link between MVP and SARS-CoV-1 that was previously reported,[20] these findings indicate that other proteins are important for SARS-CoV-1 and SARS-CoV-2 entry into cells and their replication, likely by hijacking the lipid raft-mediated endocytosis that is part of a normal immune response.

Indeed, the importance of the raft-127 genes was revealed in a recent study that used affinity-purification mass spectrometry to identify host proteins that bind to those expressed by SARS-CoV-2; of the 332 proteins identified using a high-stringency filter, seven are found in the raft-127 set and four of those seven bind to nsp7, which is part of SARS-CoV-2 replication complex. Furthermore, these four proteins (RAB5C, RAB18, RHO and RALA) are all involved in the same process, endocytosis/vesicle trafficking. A relaxed filter identified another 33 raft-127 genes including Cathepsin D (CTSD), which is substantially downregulated in cells from bronchoalveolar lavage samples taken from COVID-19 patients, and here binds to the SARS-CoV-2 ORF8 protein. This protein harbors the mutation that defines the "S" and "L" lineages of the current pandemic [30] that have been associated with differing pathogenic behavior (figure S6). [31]

Gene expression patterns of SARS-CoV-2 infected bronchoalveolar lavage cells

As mentioned above we found that several components of RAS are upregulated in bronchoalveolar lavage (BAL) samples taken from COVID-19 patients (ACE2, AGT, MAS, and REN, figure 1). In order to gain a broader understanding of SARS-CoV-2 infection, we carried out a transcriptome-wide analysis of these samples and identified 2,057 significantly downregulated and 591 upregulated genes. The downregulated genes were highly enriched for innate and adaptive immune response and endosomal trafficking; the upregulated genes were highly enriched for non-coding RNA (https://compsysbio.ornl.gov/sars-cov-2-mechanisms-of-infection/). Notably, 57 of the 127-raft genes were significantly differentially expressed in the BAL compared to control (55 downregulated and two upregulated).

We then filtered these differentially expressed genes for those that were highly correlated or anti-correlated with ACE2 (figure 8, Appendix 2.1). COVID-19 BAL samples exhibited a striking gene expression signature of upregulated genes that are markers for lung cells including type I and type II alveolar epithelial cells, ciliated cells, club cells, and lymphatic endothelial cells. Noteworthy absences included any gene-expression based signature typical of macrophage, neutrophil, or lymphocytic cells nor were there inflammatory signatures. Furthermore, genes that were

significantly downregulated in the COVID-19 BAL samples are typical markers for myeloid and lymphoid lineages, particularly CSFR1+ macrophages. Given that these genes are those correlated with ACE2, which is not normally expressed in lung, these results are consistent with the release of a myriad of lung parenchymal cells into the fluid phase as a result of SARS-CoV-2 infection from a distant location (figure 2). The absence of macrophages and lack of an activated cytokine signature in COVID-19 lung samples suggest that a major component of SARS-CoV-2's virulence is its apparent ability to cause a functional immune deficiency syndrome in lung tissue.

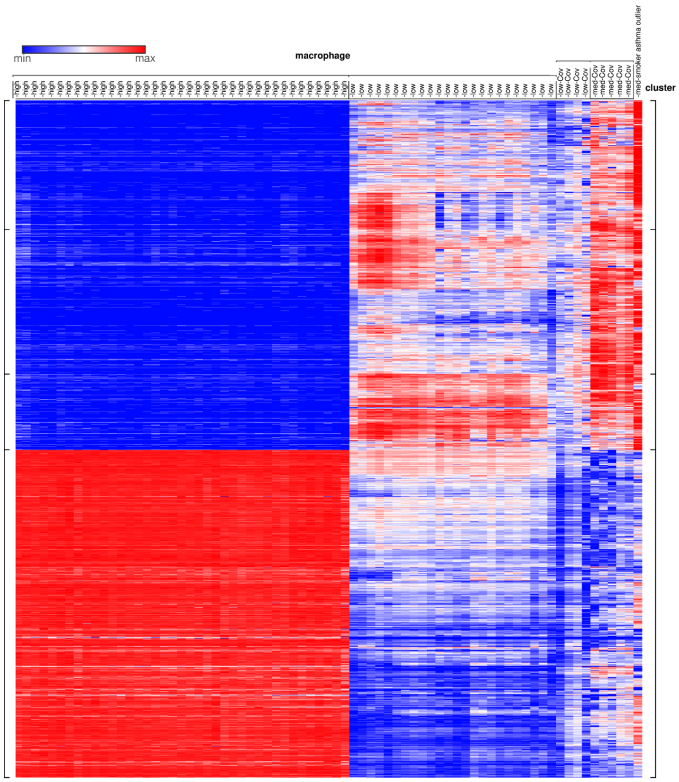


Figure 8. Overview of globally differentially expressed genes in COVID-19 BAL samples relative to controls. The COVID-19-overexpressed clusters are highly enriched for genes that define lung structural cell types including epithelial type I, type II, and lymphatic endothelial cells (upper right quadrant). There is also increased expression for genes related to intercellular or extracellular adhesion compared to controls, as well as additional genes in those categories not normally expressed in those cell types. The ACE2-correlated differentially expressed genes in the lower half of the heatmap define macrophage and dendritic cells, which are uniformly absent in the COVID-19 samples (lower right quadrant). Although 40% of the control BAL samples lacked this signature (low cluster), none of the COVID-19 samples matched the pattern in this subset of controls, indicating exclusion of these cells from BAL fluid in infected cells. The COVID-19-repressed signature (964 genes, https://compsysbio.ornl.gov/sars-cov-2-mechanisms-of-infection/) is a complete inventory of genes associated with macrophages and dendritic cells, suggesting that COVID-19 actively destroys these cells or reprograms them to the point of being unrecognizable.

Structural models are a valuable resource for Systems Biology approaches

Structural analyses and molecular dynamics studies of SARS-CoV-1 and SARS-CoV-2 spike proteins revealed differences that significantly affected the interaction with putative host receptors. Next, because SARS-CoV-1 is the closest coronavirus evolutionarily to SARS-CoV-2 that can infect humans and that has been intensely studied, we extended the structural analysis to all mature proteins of SARS-CoV-2 to identify other critical elements that may enhance the spread of COVID-19 compared to SARS. We used the currently available experimentally solved structures of SARS-CoV-2 proteins, and an ensemble workflow to generate structural models and profiles of all unsolved viral proteins (Appendix 2.5).

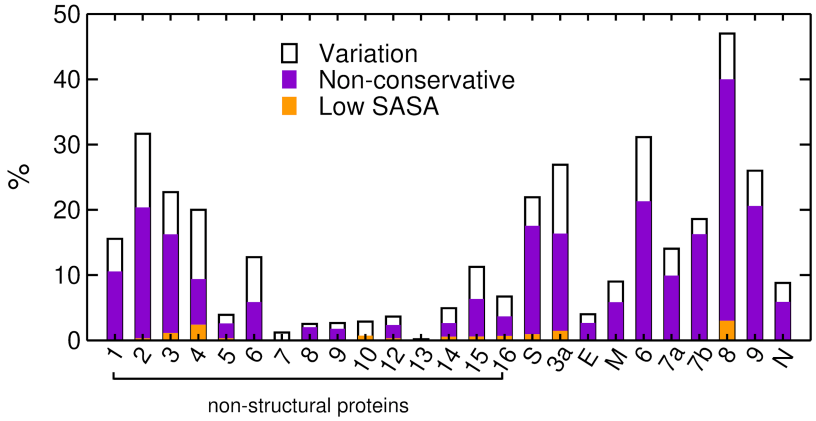


Figure 9. Distribution of sequence variation in SARS-CoV-2 proteins relative to SARS-CoV-1. Variations considered non-conservative, represented in *violet*, are defined in table S4. Variations occurring within protein cores (low solvent accessible surface area, SASA), are represented in *orange*. ORF10 is not included as it is not found in the SARS-CoV-1 proteome.

The proteome of SARS-CoV-2 includes four structural proteins, namely, spike (S), envelope (E), membrane protein (M), and nucleocapsid (N). It also produces 15 mature non-structural proteins (nsp1-nsp10 and nsp12-nsp16), and nine accessory proteins.[32] An in-depth comparative genome study reported that 380 amino acids that are fixed across thousands of SARS-like coronaviruses are changed, and specific to SARS-CoV-2,[33] suggesting that this minor portion may be essential for determining the pathogenic divergence of COVID-19. Here we verified that, in total, including the mutations that have occurred over the course of the pandemic to date, there are 1570 amino acid substitutions in SARS-CoV-2 relative specifically to SARS-CoV-1 among these proteins. As shown in figure 9, the majority of variations are non-conservative and distributed among the mature proteins, but several nsps are highly similar to their counterparts in SARS-CoV-1. The strong conservation of these nsps would tend to suggest strong purifying selection to maintain viral fitness. This relationship and their likely long term stability in the population makes them attractive drug targets. Except for ORF8, the most variable sequences diverge approximately 30% relative to SARS-CoV-1, which typically do not change global topologies and, consequently, the main protein function.

Additionally, visual inspection of non-conserved substitutions in the predicted structures combined with analyses of their structural profiles (i.e., predicted location of structured,

intrinsically disordered and transmembrane regions - Appendix 2.5), indicates that the great majority of them are located in superficial regions of the proteins or protein domains. That is, most of the substitutions do not significantly affect protein folding, but they potentially can affect post-translational modification patterns and protein function if located in key regions for interactions with other proteins, ligands and substrates. Along with the predicted models, we provide to the scientific community a synopsis for 27 mature viral proteins and nsp11 (short peptide), in which we mapped the variation relative to SARS-CoV-1, the mutations of SARS-CoV-2 occurring globally, and analyzed their potential relevance concerning the pathobiology of COVID-19. Here, as examples, we discuss the possible effects of non-conservative variations in the nonstructural proteins, nsp1 and nsp5.

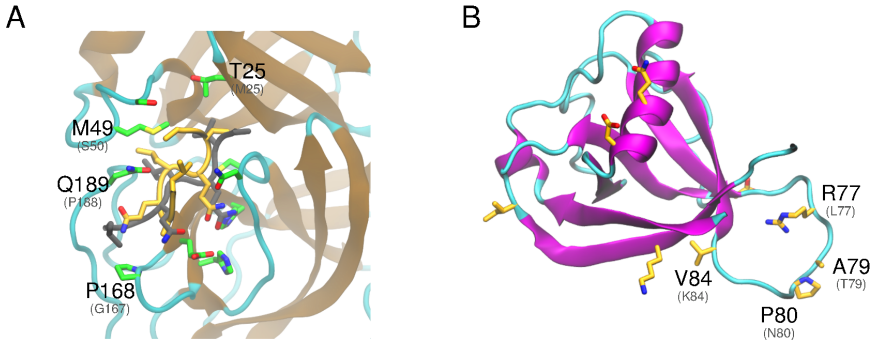


Figure 10. Analysis of the structural variation in SARS-CoV-2 nsp1 and nsp5 proteins. A) Close view of the catalytic site of nsp5. In *yellow*, NEMO is shown in the conformation predicted with docking. The conformation of NEMO in the PEDV 3CL^{pro} is also depicted, in *grey* (PDB id: 5zqg). B) Predicted fold of nsp1. Non-conservative substitutions relative to SARS-CoV-1 nsp1 are depicted in *yellow*. Substitutions discussed in the text are labeled, including the corresponding residue of the homologue (PEDV, in panel A and SARS-CoV-1, in B) in parentheses.

Pathogenic-relevant substitutions in nsp1 and nsp5

The nsp1 protein is associated with the degradation and suppression of the translation of host mRNA. 3CL^{pro} (nsp5), also commonly referred to as the main viral protease, cleaves the polyproteins translated from the invading RNA[34–36] to form the mature viral proteins (Appendix 5.5). We highlight molecular differences of these two proteins between SARS-CoV-1 and SARS-CoV-2 as they relate to host immune response and to pathogenicity divergence, being promising targets for drug development, drug repurposing, or vaccine production.

In addition to its role in processing the viral proteome, we propose that the highly conserved nsp5 protein may also be part of a major mechanism that suppresses the nuclear factor transcription factor kappa B (NF-κB) pathway, eliminating the host cell's interferon-based antiviral response. In SARS-CoV-1, several proteins have been reported to be interferon antagonists, including nsp1 and nsp3.[37,38] An additional mechanism of circumventing the interferon antiviral response is described for the porcine epidemic diarrhea virus (PEDV) as well as non-coronaviruses,[39] in which the 3CL^{pro} cleaves the NF-kB essential modulator (NEMO)[40] Given that the substrate binding site of SARS-CoV-2 3CL^{pro} is very similar to PEDV 3CL^{pro}, it is possible that SARS-CoV-2 3CL^{pro} is also active towards NEMO. Structural divergence is concentrated in the region corresponding to the S2 binding site of PEDV and in the peptide segment 45-51, in the catalytic entrance (figure 10A). As a preliminary test for this hypothesis, we conducted molecular docking of NEMO

targeting SARS-CoV-2 and PEDV $3CL^{pro}$ proteins. The best resulting substrate conformation has the Gln^{231} reaction center of NEMO positioned very similarly to the PEDV $3CL^{pro}$-NEMO crystal structure. The estimated binding affinity is -6·2 kcal/mol for SARS-CoV-2 $3CL^{pro}$-NEMO, and -7·4 kcal/mol for PEDV $3CL^{pro}$-NEMO. The binding site of SARS-CoV-2 $3CL^{pro}$ is conserved relative to SARS-CoV-1 $3CL^{pro}$, except by the substitution $Ala^{46}Ser$ in the entrance of the cleft, indicating that SARS-CoV-1 $3CL^{pro}$ may also be active towards NEMO. This result suggests that drug development targeting this mechanism may prove fruitful as it would allow for a normal host immune response to combat the pathogen.

Nsp1 is highly conserved between SARS-CoV-1 and SARS-CoV-2. Notably, in SARS-CoV-2, there are four substitutions in the less conserved β3-4 loop (figure 10B), namely, $Leu^{77}Arg$, $Thr^{79}Ala$, $Asn^{80}Pro$ and $Lys^{84}Val$. These substitutions may directly relate to pathogenicity as experimentally induced substitutions ($Arg^{73}Glu$, $Asp^{75}Arg$, $Leu^{77}Ala$, $Ser^{78}Glu$ and $Asn^{80}Gly$) in SARS-CoV-1 demonstrated increased inhibition of host gene expression, compared to the SARS-CoV-1 wild type.[41] Subsequent experiments in mice showed that the deletion of this loop in SARS-CoV-1 resulted in increased survival rate and less severe lung damage.[42] Given that this loop plays an essential role in the ability of nsp1 to impair host-translational activity, and the three substitutions in SARS-CoV-2 may be important elements of virulence divergence, this should be targeted in future studies that focus on disrupting infection.

Discussion

A Systems Biology perspective lends itself well to combating the current COVID-19 pandemic and for future zoonotic outbreaks as they will likely require multiple solutions from different biological and epidemiological perspectives. Here, our analyses advocate for a re-evaluation of the currently accepted view that the primary means of infection for the SARS-CoV-2 virus is via association with ACE2 in respiratory tissue. Gene expression and protein data suggest that a major route of entry may be digestive tissue, which is supported by recent reports and can be further verified with more in-depth analyses of existing and future medical data. This could have profound implications for early diagnosis and understanding of the frequency of less severe cases of COVID-19.

A non-respiratory path of infection has important implications for selecting the best methods to prevent the spread of SARS-CoV-2. If a major route of infection is nasal/oral that leads to strong intestinal colonization, and the poorer binding of SARS-CoV-2 Spike protein with ACE (compared to SARS-CoV-1) results in delays in the colonization (if any) of lung tissue, this may explain why many people are considered asymptomatic as they do not display the traditional symptoms (cough, fever, pneumonia). Such individuals could thus be shedding virus through feces and poor hygiene long before they display traditional symptoms (if ever), and even if they are not adversely affected. A recent study indicates that fecal-mediated viral aerosolization from toilets may be a source of infective viral particles found in the air of bathrooms. [44] It appears that there may be sufficient expression of ACE2 in nasal passages for SARS-CoV-2 to establish low level colonization that can lead to viral shedding via sneezing, nose blowing and lack of subsequent hand hygiene. Awareness of these routes of infection may offer ways to reduce the number and severity of COVID-19 cases.

Furthermore, as shown in figure 2 and figure S1, there is significant variation in ACE and ACE2 expression among organs which is consistent with some of the symptomology being reported for COVID-19, e.g. gastrointestinal,[7] testes,[9] thyroid (fatigue[10,11]), brain epithelial cells (headache[10]), etc. In addition, there is wide variation in tissue-specific expression across the GTEx population, which may partially explain differences in susceptibility and severity of illness being observed in COVID-19.

Given that we verified a weak binding of SARS-CoV-2 spike protein to ACE in our simulations, we urge that the hypothesis that ACE, which is highly expressed in lung tissue, is an alternative receptor for SARS-CoV-2 should be thoroughly tested experimentally. Our analyses suggest that ACE can be a secondary receptor for SARS-CoV-1 in the absence of ACE2, and it may have played a subtle but important role for the SARS outbreak. It could explain the lower incidence of respiratory distress in COVID-19 compared to SARS.[43] The simulations also indicate that the subtraction of a specific salt bridge due to the substitution $Lys^{460}Asn$ in SARS-CoV-2 spike glycoprotein (relative to SARS-CoV-1 S) releases the mobility of the long loop β4-5, which plays a major role in the interaction with ACE2, making the complex with ACE2 less rigid. The effectiveness of drugs targeting the S-ACE2 interface will likely be correlated with the differences in local mobility. We also identify more than 100 other host proteins that are likely important for viral infection and replication, which should be explored in greater detail for diagnostic and therapeutic potential.

The gene expression analysis has also given us fascinating insights into the overall role of the RAS system as well as into mechanisms of virus internalization. We identified candidates for co-receptors among the 127 lipid raft genes that are not currently being focused on by the COVID-19 research community. Notably, 57 of the 127-raft genes were significantly differentially expressed in COVID-19 patients. Finally, the expression analysis of the BAL samples has given us considerable data on the effect of the virus on cells found in the lung. It appears that COVID-19 likely induces increases in type I and type II alveolar epithelial cells, ciliated cells, club cells, and lymphatic endothelial cells in BAL samples. There also seemed to be a significant repression of myeloid and lymphoid lineages, particularly CSFR1+ macrophages. Macrophage, neutrophil, and lymphocytic cell inflammatory signatures were notably absent. We posit that COVID-19 actively destroys these cells or reprograms them to the point of being unrecognizable. This apparent elimination of functional macrophages and the complete lack of activated cytokine signature in COVID-19 lung BAL samples would seem to indicate that SARS-CoV-2's net effect of causing a functional immune deficiency syndrome is a considerable component of its virulence.

Along with the information of host proteins involved in infection, we provide an extensive structural analysis of the viral proteome, all of which is available in Appendix 5 and as a web resource (https://compsysbio.ornl.gov/covid-19/covid-19-structome/). The thorough structure-based mutational analysis of nsp1 and nsp5 that was discussed above as well as that done on the entire SARS-CoV-2 proteome has been performed in order to provide functional insights as to the likely impacts of these mutations on transmission, virulence and pathogenicity. The collective analysis also informs the identification of promising drug, vaccine and diagnostic targets for COVID-19. The workflow developed for this study can readily be implemented in future efforts against pathogen outbreaks.

Acknowledgements

This research used resources of the Oak Ridge Leadership Computing Facility, which is a DOE Office of Science User Facility supported under Contract DE-AC05-00OR22725. This research used resources of the Compute and Data Environment for Science (CADES) at the Oak Ridge National Laboratory. We thank Joao Gabriel F. M. Gazolla (Biosciences Division, Oak Ridge National Laboratory) for support in the design of figures.

Contributions

DJ conceived of the study, raised funding, supervised the study, performed data analysis and participated in writing and editing. ETP predicted protein structures, performed molecular dynamics simulations and analysis, analyzed protein structures and mutants, and participated in the writing and editing. MG collected and curated datasets, performed data analysis, analyzed protein structures and mutants, and participated in the writing and editing. PJ predicted protein structures, analyzed protein structures and mutants and participated in the writing. MS organized data, predicted protein structures, analyzed protein structures and mutants and participated in the writing. MP, OD, BKA, AG analyzed protein structures and mutants and participated in the writing. DK, CA performed data analysis and figure creation. JP participated in the early conceptual framework. BA, JK, AM, EB analyzed transcriptomic data, derived reference signatures, and developed infrastructure for comparative analysis of test signatures. BA, JK, AM participated in writing.

Role of the funding source

The funders of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report. The corresponding author had full access to all the data in the study and had final responsibility for the decision to submit for publication.

References

1 Cai J, Sun W, Huang J, Gamber M, Wu J, He G. Indirect Virus Transmission in Cluster of COVID-19 Cases, Wenzhou, China, 2020. *Emerg Infect Dis* 2020; **26**.

2 Department of Communicable Disease Surveillance and Response. Consensus document on the epidemiology of severe acute respiratory syndrome (SARS). World Health Organization, 2003.

3 Wevers BA, van der Hoek L. Renin–angiotensin system in human coronavirus pathogenesis. *Future Virology* 2010; **5**: 145–61.

4 Mehta P, McAuley DF, Brown M, *et al.* COVID-19: consider cytokine storm syndromes and immunosuppression. *Lancet* 2020; published online March 16.

5 Rigat B, Hubert C, Alhenc-Gelas F, Cambien F, Corvol P, Soubrier F. An insertion/deletion polymorphism in the angiotensin I-converting enzyme gene accounting for half the variance of serum enzyme levels. *J Clin Invest* 1990; **86**: 1343–6.

6 Wang W, Xu Y, Gao R, *et al.* Detection of SARS-CoV-2 in Different Types of Clinical Specimens.

JAMA 2020; published online March 11.

7 Hu Z, Chen Y, Zhang K, *et al.* Clinical characteristics and current treatment of critically ill patients with COVID-19 outside Wuhan, China: a multicenter, retrospective, observational study. 2020; Preprint at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3551426.

8 Zhou J, Li C, Zhao G, *et al.* Human intestinal tract serves as an alternative infection route for Middle East respiratory syndrome coronavirus. *Sci Adv* 2017; **3**: eaao4966.

9 Ma L, Xie W, Li D, *et al.* Effect of SARS-CoV-2 infection upon male gonadal function: A single center-based study. *medRxiv* 2020.

10 Yang X, Yu Y, Xu J, *et al.* Clinical course and outcomes of critically ill patients with SARS-CoV-2 pneumonia in Wuhan, China: a single-centered, retrospective, observational study. *The Lancet Respiratory Medicine* 2020.

11 Xu X-W, Wu X-X, Jiang X-G, *et al.* Clinical findings in a group of patients infected with the 2019 novel coronavirus (SARS-Cov-2) outside of Wuhan, China: retrospective case series. *BMJ* 2020; *368*: m606.

12 Han X, Zhou Z, Fei L, *et al.* Construction of a human cell landscape at single-cell level. *Nature* 2020; published online March 25.

13 Tabula Muris Consortium, Overall coordination, Logistical coordination, *et al.* Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature* 2018; **562**: 367–72.

14 Ruiz García S, Deprez M, Lebrigand K, *et al.* Novel dynamics of human mucociliary differentiation revealed by single-cell RNA sequencing of nasal epithelial cultures. *Development* 2019; **146**.

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

16 Yan R, Zhang Y, Li Y, Xia L, Guo Y, Zhou Q. Structural basis for the recognition of the SARS-CoV-2 by full-length human ACE2. *Science* 2020; **367**: 1444–8.

17 Tian X, Li C, Huang A, *et al.* Potent binding of 2019 novel coronavirus spike protein by a SARS coronavirus-specific human monoclonal antibody. *Emerg Microbes Infect* 2020; **9**: 382–5.

18 Wrapp D, Wang N, Corbett KS, *et al.* Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. *Science* 2020; **367**: 1260–3.

19 Hoffmann M, Kleine-Weber H, Krüger N, Müller M, Drosten C, Pöhlmann S. The novel coronavirus 2019 (2019-nCoV) uses the SARS-coronavirus receptor ACE2 and the cellular protease TMPRSS2 for entry into target cells. *bioRxiv* 2020.

20 Li W, Moore MJ, Vasilieva N, *et al.* Angiotensin-converting enzyme 2 is a functional receptor for the SARS coronavirus. *Nature* 2003; **426**: 450–4.

21 Dewerchin HL, Desmarets LM, Noppe Y, Nauwynck HJ. Myosins 1 and 6, myosin light chain kinase, actin and microtubules cooperate during antibody-mediated internalisation and trafficking of membrane-expressed viral antigens in feline infectious peritonitis virus infected monocytes. *Vet Res* 2014; **45**: 17.

22 Yu Z, Fotouhi-Ardakani N, Wu L, *et al.* PTEN associates with the vault particles in HeLa cells. *J Biol Chem* 2002; **277**: 40247–52.

23 Kickhoefer VA, Rajavel KS, Scheffer GL, Dalton WS, Scheper RJ, Rome LH. Vaults are up-regulated in multidrug-resistant cancer cell lines. *J Biol Chem* 1998; **273**: 8971–4.

24 Peng N, Liu S, Xia Z, *et al.* Inducible Major Vault Protein Plays a Pivotal Role in Double-Stranded RNA- or Virus-Induced Proinflammatory Response. *J Immunol* 2016; **196**: 2753–66.

25 Kowalski MP, Dubouix-Bourandy A, Bajmoczi M, *et al.* Host resistance to lung infection mediated by major vault protein in epithelial cells. *Science* 2007; **317**: 130–2.

26 Daviet L, Lehtonen JY, Tamura K, Griese DP, Horiuchi M, Dzau VJ. Cloning and characterization of ATRAP, a novel protein that interacts with the angiotensin II type 1 receptor. *J Biol Chem* 1999; **274**: 17058–62.

27 Lopez-Ilasaca M, Liu X, Tamura K, Dzau VJ. The Angiotensin II Type I Receptor-associated Protein, ATRAP, is a Transmembrane Protein and a Modulator of Angiotensin II Signaling. *Mol Biol Cell* 2003; **14**: 5038–50.

28 Makino S, Narayanan K. Faculty of 1000 evaluation for LDL receptor and its family members serve as the cellular receptors for vesicular stomatitis virus. F1000 - Post-publication peer review of the biomedical literature. 2013.

29 Syed GH, Tang H, Khan M, Hassanein T, Liu J, Siddiqui A. Hepatitis C virus stimulates low-density lipoprotein receptor expression to facilitate viral propagation. *J Virol* 2014; **88**: 2519–29.

30 Ceraolo C, Giorgi FM. Genomic variance of the 2019-nCoV coronavirus. *J Med Virol* 2020; **92**: 522–8.

31 Tang X, Wu C, Li X, *et al.* On the origin and continuing evolution of SARS-CoV-2. *National Science Review* 2020.

32 Gordon DE, Jang GM, Bouhaddou M, *et al.* A SARS-CoV-2-Human Protein-Protein Interaction Map Reveals Drug Targets and Potential Drug-Repurposing. *bioRxiv* 2020.

33 Wu A, Peng Y, Huang B, *et al.* Genome Composition and Divergence of the Novel Coronavirus (2019-nCoV) Originating in China. *Cell Host Microbe* 2020; **27**: 325–8.

34 Perlman S, Netland J. Coronaviruses post-SARS: update on replication and pathogenesis. *Nat Rev Microbiol* 2009; **7**: 439–50.

35 Ziebuhr J, Snijder EJ, Gorbalenya AE. Virus-encoded proteinases and proteolytic processing in the Nidovirales. *J Gen Virol* 2000; **81**: 853–79.

36 Anand K, Ziebuhr J, Wadhwani P, Mesters JR, Hilgenfeld R. Coronavirus main proteinase (3CLpro) structure: basis for design of anti-SARS drugs. *Science* 2003; **300**: 1763–7.

37 Wathelet MG, Orr M, Frieman MB, Baric RS. Severe acute respiratory syndrome coronavirus evades antiviral signaling: role of nsp1 and rational design of an attenuated strain. *J Virol* 2007; **81**: 11620–33.

38 Frieman M, Ratia K, Johnston RE, Mesecar AD, Baric RS. Severe acute respiratory syndrome coronavirus papain-like protease ubiquitin-like domain and catalytic domain regulate antagonism of IRF3 and NF-kappaB signaling. *J Virol* 2009; **83**: 6689–705.

39 Huang C, Zhang Q, Guo X-K, *et al.* Porcine reproductive and respiratory syndrome virus nonstructural protein 4 antagonizes beta interferon expression by targeting the NF-κB essential modulator. *J Virol* 2014; **88**: 10934–45.

40 Wang D, Fang L, Shi Y, *et al.* Porcine Epidemic Diarrhea Virus 3C-Like Protease Regulates Its Interferon Antagonism by Cleaving NEMO. *J Virol* 2016; **90**: 2090–101.

41 Jauregui AR, Savalia D, Lowry VK, Farrell CM, Wathelet MG. Identification of residues of SARS-

CoV nsp1 that differentially affect inhibition of gene expression and antiviral signaling. *PLoS One* 2013; **8**: e62416.

42 Jimenez-Guardeño JM, Regla-Nava JA, Nieto-Torres JL, *et al.* Identification of the Mechanisms Causing Reversion to Virulence in an Attenuated SARS-CoV for the Design of a Genetically Stable Vaccine. *PLOS Pathogens* 2015; **11**: e1005215.

43 Wilder-Smith A, Chiew CJ, Lee VJ. Can we contain the COVID-19 outbreak with the same measures as for SARS? *Lancet Infect Dis* 2020; published online March 5.

44 Liu Y, Ning Z, Chen Y, *et al.* Aerodynamic Characteristics and RNA Concentration of SARS-CoV-2 Aerosol in Wuhan Hospitals during COVID-19 Outbreak. DOI:10.1101/2020.03.08.982637

Appendix

Confronting the COVID-19 Pandemic with Systems Biology

Erica Teixeira Prates, Ph.D.[1‡], Michael R. Garvin, Ph.D.[1‡], Mirko Pavicic, Ph.D.[1], Piet Jones, M.S.[1,2], Manesh Shah, M.S.[1], Christiane Alvarez[1], David Kainer, Ph.D.[1], Omar Demerdash, M.D.[1], B Kirtley Amos, M.S.[5], Armin Geiger, M.S.[1,2], Kang Jin, B.S.[3,4], Alexis Mitelpunkt, M.D.[3,4], Eric Bardes, B.S.[3,4], Bruce Aronow, Ph.D.[3,4], Daniel Jacobson, Ph.D.[1,2,6 *]

[*]Corresponding author: Daniel Jacobson, Ph.D., jacobsonda@ornl.gov, 865 574 6134

[‡]Contributed equally

Appendix 1: Supplementary Texts

1.1. ACE2 is not highly expressed in normal lung tissue

As mentioned previously, both the Genotype-Tissue Expression (GTEx) and Proteomics DataBase (Proteomics DB) indicate ACE2 is either expressed at very low levels or is not detectable in tissues that are currently thought to be major entry points for SARS-CoV-1 and SARS-CoV-2 such as lung. High expression of ACE2 is based on several reports that can be divided into those that use mRNA and those that use protein.[1–7] The two earliest reports that measured ACE2 mRNA in lung tissue showed either no expression at all or low levels [7,8] whereas ACE is found at moderate to high levels; neither were quantitative and both were based on small sample sizes. The most recent reports used data from single cell RNA-Seq experiments (and those results agree with data from the GTEx project) and show that ACE2 is expressed at very low levels in a small fraction of cells (ACE was not reported).[5,6] ACE vs ACE2 in single cell atlas datasets from human tissues.

The earliest reports of ACE2 expression in the lung were based on protein expression using a peptide-derived polyclonal antibody and immunohistochemistry.[1,2] Aside from being non-quantitative, the results are difficult to interpret and could be due to the non-specificity of these types of antibodies. The three remaining reports of ACE2 expression in lung or lung-derived epithelium were based on improved antibody immunohistochemistry and western blots.[3,4,9] Although still not quantitative, the reports are again consistent with GTEx and Proteome DB; ACE2 is either not expressed or found at very low levels in normal lung tissue or normal lung-derived epithelial cells. However, it is induced either by exposing cells grown in culture to air (air-liquid interface) or by infection with the SARS-CoV-1 virus, suggesting that compromised lung epithelium or cells that are already infected with virus are routes of infection in this tissue.

1.2. ACE may mediate SARS-CoV-1 infection

The involvement of ACE in the system that was used to identify ACE2 as the SARS-CoV-1 receptor is unknown because they did not overexpress it as was done with ACE2.[10] However, a subsequent study using pseudovirus typing demonstrated a correlation between ACE2 expression and infectivity in several cell lines and although ACE2 was much higher, overexpression of ACE was shown to increase infectivity in some cell lines.[11] Notably, ACE2-mediated infection was highest in kidney- and colon-derived cell lines and much less efficient in those from the lung. This same pattern in kidney-, colon-, and lung-derived cell lines were confirmed in a more recent analysis of beta-coronaviruses.[12] From these combined results we conclude that while ACE2 is involved in SARS-CoV-1 infection in kidney-derived cells, the involvement of ACE2 in lung-mediated viral entry and ACE in viral infection in general is inconclusive. It should be noted that the majority of positive results were garnered in kidney- and colon-derived cell lines, which express higher levels of ACE2 than ACE. In the lung and other respiratory tissues, the reverse is true; ACE is expressed at higher levels than ACE2. From *in vitro* work it appears that ACE may be a less efficient receptor for SARS-CoV-1 (and perhaps SARS-CoV-2) but, given an environment in which levels of ACE are an order of magnitude higher than ACE2, this may be a potential route of infection in respiratory tissue.

1.3. Gene Expression of BAL Samples from COVID-19 patients

Interestingly, in addition to the high representation of the raft-127 genes in those that were differentially expressed, the vast majority (98%) of the 2,057 downregulated genes were protein coding whereas only 41% of the upregulated genes coded for protein and 40% were non-coding RNAs (ncRNA). One of the most upregulated ncRNA (LOC10537114, a 3,171 fold increase) is associated with lung cancer susceptibility.[13] A Gene Ontology (GO) biological enrichment analysis identified no significant enrichment within the upregulated protein coding genes but the downregulated genes demonstrated significant enrichment of the adaptive and innate immune responses, including the NF-κB pathway, which is known to be inhibited by coronavirus during infection.[14] In addition, trafficking among the Endoplasmic Reticulum (ER), Golgi, and endosomal compartments was highly enriched for many categories. Other enriched terms implicated lipid rafts, apoptosis, protein production, and respiration.

Appendix 2: Supplementary Materials and Methods

2.1. Gene Expression analysis

FASTQ files were downloaded from the Sequence Read Archive on NCBI and trimmed using the default parameters in CLC Genomics Workbench (20.0.3). RNA-Seq analysis was performed using the latest version of the human transcriptome (GRCh38_latest_rna.fna) with the reference genome for the SARS-CoV-2 virus appended (NC_045512). Transcripts per million (TPM) were used for comparisons across groups and the *Differential Expression for RNA-Seq* tool from CLC, which uses log-transformed TPM fitted to a GLM was used to identify globally differentially significantly expressed genes.[15] Expression was significant if the Bonferroni correction was less than 0.01, the fold-change was greater than 2, and the mean TPM for either the control or the experimental groups was greater than 25. Significant enrichment of the Gene Ontology terms (geneontology.org) was determined with an FDR <0.05. GO terms and their enrichment values were analyzed with REVIGO (revigo.irb.hr) to reduce redundancy of terms. A similarity score of 0.4 was used to be conservative within the homo sapiens GO database.
For nasal brushings and the BAL samples, the TPM were $\log_2+1$ transformed, tested for differential expression and then filtered for genes that were found to be highly correlated with ACE2 expression in human lung. For those genes, we extracted the human gene expression data for lung from GTEx (N=578 individuals) and collapsed gene transcripts to genes (N=38340). We calculated Pearson correlation coefficient for all pairs of genes and used those that had correlation of +0.80 and -0.65 for ACE2 as the filter.
ToppCell carries out extensive differential gene expression analysis between cell classes, subclasses, and tissues and allows for between-dataset and between-sample of origin evaluation of gene expression patterns. This enabled the comparison of ACE vs. ACE2 expression across different tissues and datasets.

2.2. Population level SARS-CoV-2 analysis

FASTA files and meta data were downloaded from the Global Initiative on Sharing All Influenza Data (GISAID) website. Sequences were aligned with CLC Genomics Workbench (20.0.3), trimmed to include only coding sequence, and then translated to protein. A Neighbor-Joining Tree was generated using Jukes-Cantor protein distance measure.

2.3. Molecular dynamics simulations

The solved structures of the complexes of SARS-CoV-1 and SARS-CoV-2 receptor binding domains (RBDs) bound to angiotensin-converting enzyme 2 (ACE2), referred as RBD1-ACE2 and RBD2-ACE2, respectively, were used as starting configuration for atomistic molecular dynamics (MD) simulations (PDB ids: 6m17 and 2ajf).[16,17] MD simulations were performed with the 2020 version of GROMACS.[18] Each system was simulated in triplicate. The CHARMM36 protein force field[19] was used with TIP3P water.[20] For glycosylated proteins, the CHARMM carbohydrate force field was used.[21]
Proteins were solvated with a 15 Å water layer in an octahedral box under periodic boundary conditions. Sodium and chloride ions were added at a concentration of ~0.16 M to ensure electroneutrality of the system. All ions present in the original crystallographic structures were kept for the simulations.
The system was first energy minimized for 5000 steps with steepest descent. This was followed by 2 phases of equilibration. In the first phase, 6 ns passed while applying positional constraints on alpha carbons except for those at binding interfaces. The temperature was gradually raised to 298.15 K. In the second phase, 20 ns passed during which positional constraints were applied to alpha carbons of the C-terminal domains of ACE and ACE2 and a few residues in the core of the spike protein's receptor-binding domain. In each simulation of the triplicate, atomic velocities were reinitialized from a Maxwell-Boltzmann distribution at 10 ns, using a different random seed in each case. Equilibration and production were run in the NPT ensemble, and a 2 fs time step was used. Temperature was maintained using velocity rescaling via a stochastic term that properly generates constant pressure-constant temperature ensembles;[22] and a coupling constant of 1.0 ps. The Berendsen barostat[23] was used with a coupling constant of 1.0 ps to maintain a pressure of 1 atm. All systems were run with a total of 300 ns of production.

2.4. Simulation analyses

Inter-residue contacts: Time evolution of number of contacts was computed using the mindist utility of GROMACS and a distance cutoff within alpha carbons of 8 Å. The probability density of contacts was computed using the timeline plugin of VMD.[24]
Visual analysis and the design of figures were made using the version 1.93 of VMD.

2.5. Ensemble workflow for protein structure prediction

To date, partial or full structures of five proteins from SARS-CoV-2 have been experimentally solved. In view of the urgency to understand the molecular machinery of SARS-CoV-2, we used an ensemble workflow to generate structural models of all unsolved structural and mature nonstructural viral proteins. Due to the performance of methods for protein structure prediction varying by complexity, protein sequences were carefully analyzed to optimize the combination of the state-of-the-art methods of protein structure prediction. As such, the resulting models have the highest possible resolution and maximum information with regards to the overall shape of each protein. Here, we provide a synopsis for each of the 27 mature viral proteins along with their structural models and additional important information, such as variability relative to SARS-CoV-1 and potential functional relevance to SARS-CoV-2.

Case-by-case protocols were generated based on a profile extracted from each sequence, consisting of two main factors:

1. *Primary sequence-based information*. Residues within conserved domains (Pfam[25]) and intrinsically disordered regions were identified using IuPred2[26], which relies on the composition of amino acid segments and their tendency to form stable structural motifs. TMHMM[27] was used to predict the helical transmembrane protein regions based on a hidden Markov model. No β-barrel transmembrane proteins are coded for in SARS-CoV-2.
2. *Availability of experimentally determined structures*. PSI-BLAST was used to identify homologous with partial or full structures available in the Protein Data Bank (PDB) that could be used as templates for modeling.

Several SARS-CoV-1 proteins that are highly conserved have been solved experimentally and are available for our analysis. In order to maximize accurate translation of information from these structures, amino acid substitutions were analyzed to identify those that likely impact protein conformation. Examples of changes that affect protein structure are a hydrophobic side chain being replaced by a charged amino acid at the protein core or a substitution to proline (a helix "breaker") within a helical structure. In case such substitutions are not found, and the protein has more than 70% identity to the template, loops and substitutions are locally modeled (LM) using the Rosetta remodel[28] and fixbb[29,30]applications, respectively. The comparison of recently released crystallographic structures with the models generated using carefully analyzed protein sequences and using LM for selected regions appears to be an effective approach (table S5). Achieving high local resolution, especially in sites of substrate/ligand binding, can considerably enhance the results of subsequent studies for small molecule candidate identification using molecular docking. Although ensemble docking approaches are often applied to contend with the conformational flexibility of the protein target, refining the binding site based on structural information from homologs in the *holo* form, if available, is more suitable for identifying functional complexes.

Homology-based modeling is typically the optimal approach for cases in which the identity to the template is above 30%. The fragment-based (FB) approach of the I-TASSER[43] workflow was used in cases where the range of identity was 30-70%, and to provide an alternative model to LM in regions of proteins harboring substitutions that would be expected to significantly affect protein conformation. In order to predict structures for proteins that do not have a crystal structure of a homolog available, we applied the trRosetta[31] workflow. Based on benchmarks of the Critical Assessment of Techniques for Protein Structure Prediction (CASP13), trRosetta was designed to achieve sound performance for modeling novel folds by using a deep residual network for predicting inter-residue distance and orientation that guides energy minimization.

We use the analysis of nsp3, the largest mature protein of SARS-CoV-2, as an example for the workflow (figure S7): 1. The first model was generated using residues 5-111 via LM (78% identity to PDB 2GRI). This region partially overlaps with a conserved domain (DUF3655), which is predicted by ANCHOR2 to be a disordered binding region; 1b. A fragment-based model that expands to the end of this domain (segment 2-169) suggests that it undergoes disorder-to-order transition upon interaction with the N-terminal (figure 6B). 2. The flexible linker following DUF3655 was skipped and the next segment (205-374), corresponding to the ADP ribose phosphatase domain of nsp3 was modeled based on the alignment with a highly similar template (73% identity, PDB id 2fav, SARS-CoV-1). A comparison with a recently solved crystal structure of this region (PDB id 6w02) reveals that the generated model is good (RMSD 1.5 Å); 3. The preliminary analysis of the substitutions in the segment composed of residues 413-676 relative to the template from SARS-CoV-1 (2w2g) demonstrates that most of the non-conservative substitutions occur in superficial regions, but several suggest that conformational divergence should be considered. In order to account for local rearrangements in this region, which includes the SUD-M domain, a FB model was generated as an alternative to the LM model, since the I-TASSER workflow includes an atomistic refinement phase. figure 6C shows the amino acid substitutions in the LM model aligned to the template. The magnified section depicts the conformational change captured by the FB model due to the loss of the disulfide bridge present on the template. 4. Residues 680-743

correspond to PL2^{pro}, the coronavirus polyprotein cleavage domain. Although a psi-blast search identifies a solved homolog covering this region, this domain is missing in the structure. Due to the lack of any template, this domain was modeled *ab initio*, with the trRosetta workflow; 5. Residues 746-1062 correspond to the papain-like viral protease, which acts in the proteolytic processing of viral replicase. It was locally modeled using the homolog from SARS-CoV-1 (PDB id 5tl6), which is 82% identical; 6. Similarly, the subsequent residues 1089-1203 correspond to the putative nucleic acid binding domain of nsp3, was modeled via LM using the homolog from SARS-CoV-1 (PDB id 2k87), which is 81% identical. No other conserved regions were identified in nsp3. The C-terminal region, which is predicted to contain transmembrane segments, was not modeled, since the soluble functional domains of this protein are of highest interest, and as there is no previous structural information to ensure the generation of a reliable model for it.

2.6. Molecular docking

Molecular docking of the nuclear factor-kappa B essential modulator (NEMO) targeting SARS-CoV-2 and PEDV 3CL^{pro} proteins was performed using Autodock Vina.[32] Autodocktools was used to prepare the inputs.[33] The search space was defined as a box with dimensions 20 x 20 x 20 Å, encompassing the side chains of the full catalytic site of these enzymes. Grid space 1·0 Å was used, and exhaustiveness parameter was set 20. The N-Cα and Cα-C bonds in the segment Gln^{229}-Ala^{233} and bonds in the side chains of Leu^{227}, Leu^{230}, Val^{232} and Ala^{233} were set as flexible, except for those forming π-conjugated systems. The remaining bonds were fixed in the conformation of NEMO in the crystal structure of PEDV nsp5-NEMO (PDB id 5zqg), as the results of Vina are often more accurate for a number of active bonds lower than 15.[34]

2.7 Tracking SARS-CoV-2 molecular evolution

Resources are available to track and infer functional consequences of mutations as the virus spreads globally. We aligned the full protein coding sequences of all available SARS-CoV-2 isolates (N=1,495 as of this publication) and identified the frequency and geographic associations of the new mutations as well as their proteome location. The most frequent mutation (72%) at position 84 in the ORF8 protein from an ancestral serine to leucine has been proposed to increase the aggressiveness of the virus, allowing it to spread further.[35] This idea seems to also be consistent with the less effective spread of SARS-CoV-1, in which ORF8b has a threonine in this site (it has similar chemical properties of serine).

Another important mutation is the double substitution in the nsp13 helicase (Pro^{504}Leu, Tyr^{541}Cys), frequently found in Washington State, USA. Interestingly, nsp13 is absolutely conserved within SARS-like coronaviruses.[36] Both substitutions are located in the 2A (a RecA like) domain of the protein, with Tyr^{541} predicted to be part of a critical region for nucleic acid binding.[37] In fact, in previous work, it was shown that a the double substitution Ser^{539}Ala/Tyr^{541}Ala decreased helicase unwinding activity.[38] Based on those results, it is likely that the loss of the tyrosine, a bulky aromatic amino acid, will impact nucleic acid binding efficiency. The binding region is also associated with the RNA triphosphatase activity of nsp13, thus the mutation may affect the viral 5' RNA capping, and thereby viral replication.[39]

Finally, we also find a frequent (41%) Asp to Gly substitution at position 614 in the end of S1 subunit of the spike protein. We hypothesize that Asp^{614} counterbalances the attraction of Arg^{646} to Glu^{868} in the subunit S2 of the adjacent monomer, and, therefore, this mutation may enhance the trimer packing by releasing Arg^{646} to mainly interact with Glu^{868} (figure S5). This site is linked to a second mutation (Pro^{323}Leu) in nsp12. Other notable findings are a Scandanavian-specific strain defined by a Thr to Met substitution in the membrane protein and a Leu to Phe in the nsp6 gene of the viral type infecting individuals in Japan.

2.8 Single-cell RNA-seq Analysis

Mouse cell Atlas and Human Cell Landscape Single-cell RNA-seq data were downloaded from Tabula Muris and Human Cell Landscape platforms. Total UMI counts per cell were normalized to one million to compare across tissues. Cell classes and subclasses were drawn from metadata in Tabula Muris and Human Cell Landscape compendium and were amended ---by cell annotations of scRNA-Seq data from other well-annotated datasets (*i.e.* LungMap) for further labelling. Cells were collapsed into bins by their identities, including tissue, lineage, class and subclass to support visualization. The 75 percentile of expression values of cells in the bins was used as their expression levels. Using the annotations, a hierarchy was defined for each bin and gene lists were created to represent the hierarchical gene expression modules. Differential expression analysis was carried out by t-test and the top two hundred genes of each module were used to create a heatmap using Morpheus. Rows were collapsed to the mean expression of ACE and

ACE2 across modules and cell subclasses from Mouse Cell Atlas. Gene modules with ACE or ACE2 in their gene lists were represented for Human Cell Landscape.

scRNA-seq data of respiratory tissues were downloaded from GSE12260 (including samples from nasal brushing, nasal turbinate and bronchial biopsy) and Tissue Stability Cell Atlas (including samples from parenchyma). Mouse Tongue scRNA-seq data were extracted from Mouse Cell Atlas. We used the package SCANPY[40] to filter out cells with less than 500 total UMI counts or more than 15% mitochondrial reads. Total UMI counts were normalized to one million per cell. The top 3,000 highly variable genes were calculated, and normalized values were scaled to unit variance. PCA was carried out on highly variable genes and the top 50 principle components were selected. Louvain clustering and UMAP visualization were used to cluster the cells and gain a better view of the cell distribution. We used known marker genes to label major cell types and refined our annotations by mapping signatures from the well-annotated scRNA-seq datasets. Gene expression levels of ACE and ACE2 were shown on the same UMAP and compared across respiratory tissues.

Appendix 3: Supplementary Figures

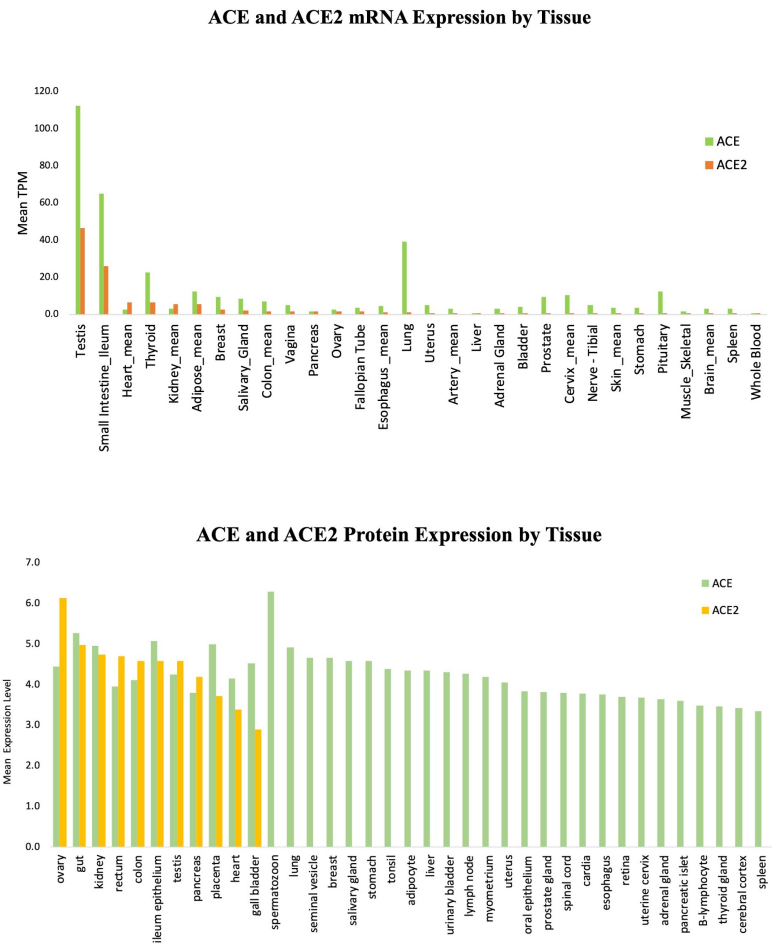


Figure S1. Expression values for ACE and ACE2 from GTEx and Proteome DB databases. mRNA are given as mean TPM (transcripts per million).

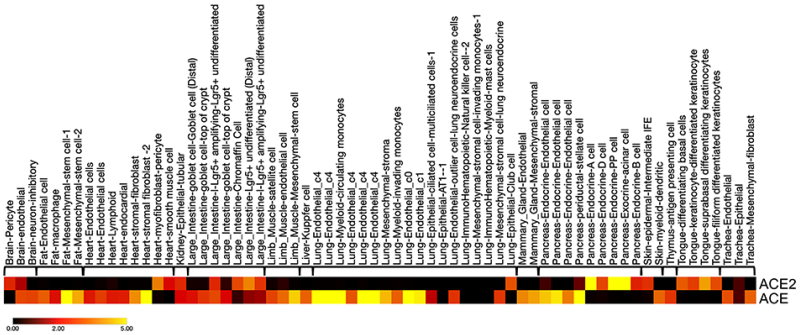


Figure S2. ACE2 vs. ACE expression in Mouse Tabula Muris. Like human tissues and cell types, ACE2 exhibits highly selective expression patterns in mice and is expressed in select epithelial cells of the tongue and GI tract, including pancreatic cells and enterocytes. In the lung, only Club cells have significant ACE2 expression, albeit at 4-fold or lower levels vs pancreatic or enterocyte subtypes. Pericytes in the brain and heart are notably positive for ACE2, but lung pericytes do not demonstrate significant expression. Similar to the nasal epithelium brushing samples, mouse tongue expression of ACE2 corresponds to developing and differentiating keratinocytes. Color scales are Log2(TPM+1).

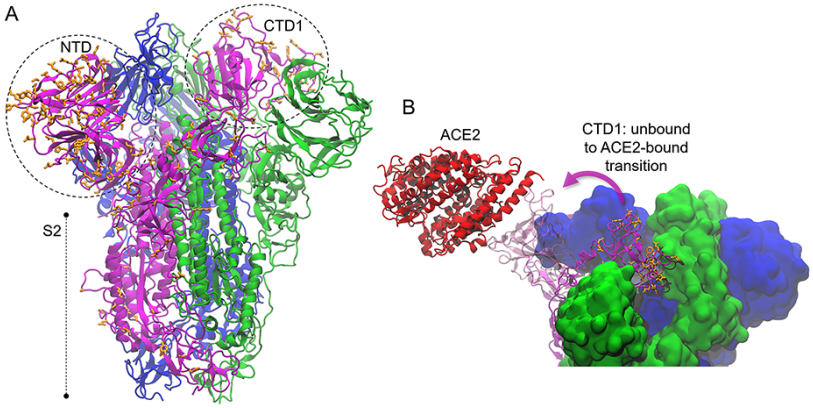


Figure S3. A) Local modeling-derived SARS-CoV-2 spike glycoprotein. B) Conformational transition of the receptor-binding domain of the S1 subunit of the spike glycoprotein and association with ACE2 receptor. Non-conservative substitutions relative to SARS-CoV-1 spike are depicted in orange. N-terminal (NTD) and C terminal domains (CTD) are identified.

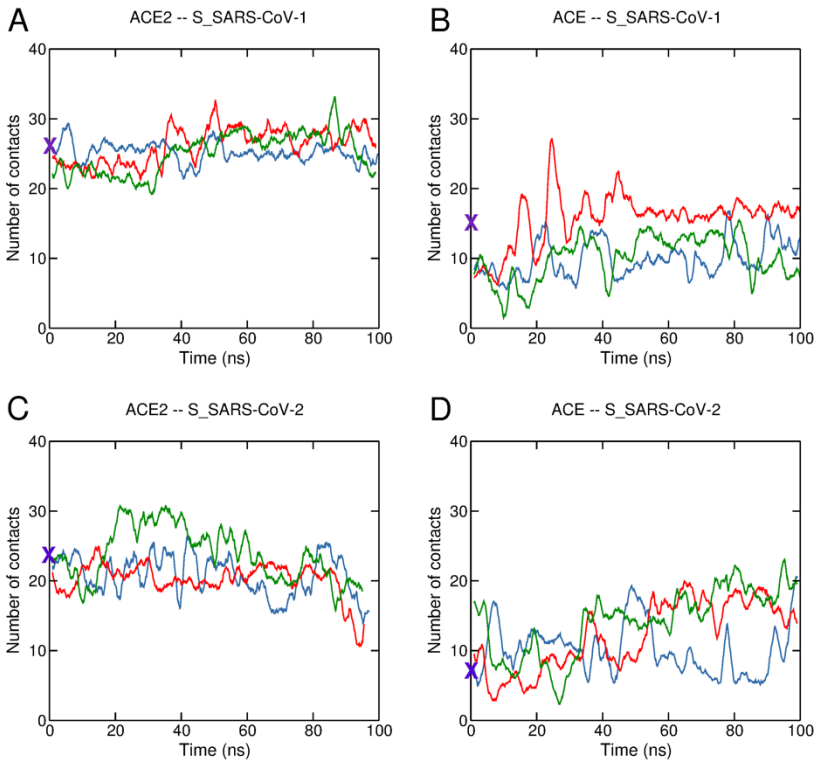


Figure S4. Time evolution of the number of contacts within the receptor binding domain (SARS-CoV-1 and SARS-CoV-2 spike glycoproteins) and the putative receptors ACE2 and ACE. A contact was considered for pairs of residues with C-alpha less than 8 Å distant.

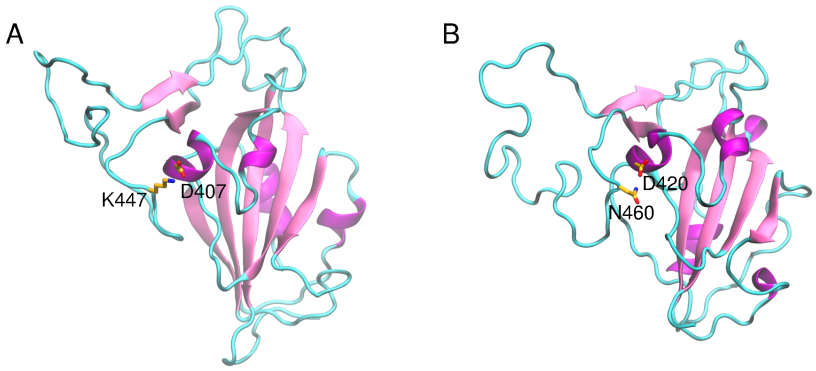


Figure S5. Substitution of Lys447 in SARS-CoV-1 S protein (A) by Asn^{460} in SARS-CoV-2 S (B) protein results in loss of a salt bridge connecting a long loop to RBD core.

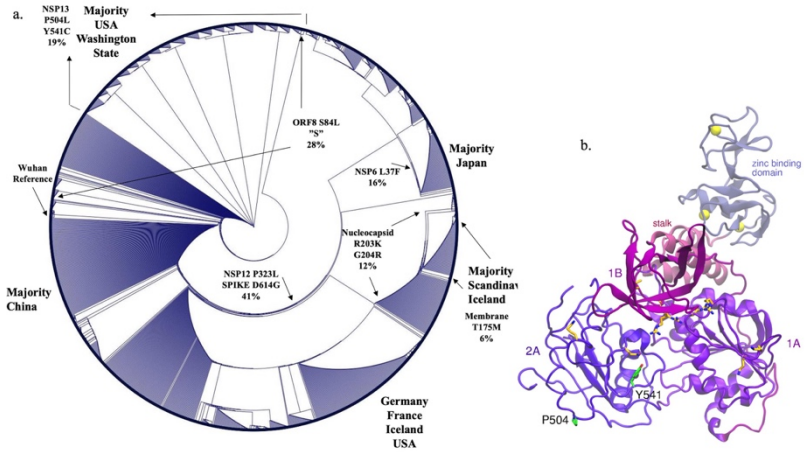


Figure S6. A) Neighbor-joining circular dendrogram of 1495 protein sequences for SARS-CoV-2 from GSAID. Frequent mutations that define branch points are given as well as some geographic-specific clades (e.g., the Thr to Meth substitution in the Membrane protein at site 175 defines a sub-group of viruses in Scandinavia). Similarly, a linked set of mutations is currently spreading in Washington State in the USA, which may have functional implications on the (B) Predicted structure of SARS-CoV-2 nsp13. Non-conservative substitutions relative to SARS-CoV-1 nsp13 are depicted in orange. Mutations observed in SARS-CoV-2 nsp13 are depicted in green. Zinc ions are represented as yellow spheres.

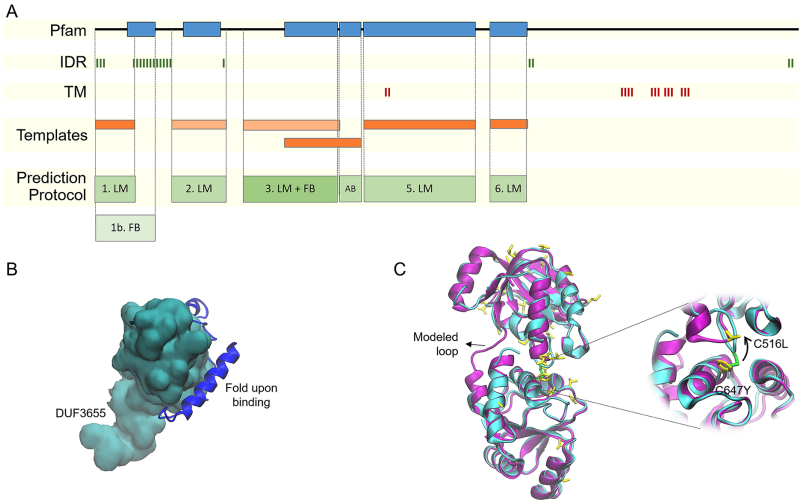


Figure S7. Ensemble workflow for structure prediction of SARS-CoV-2 nsp3. Case-by-case protocols of structure prediction are determined by finely parsing each protein sequence using information about the position of intrinsically disordered regions (IDR), transmembrane regions (TM), signal peptides, and templates (figure S1A). The method applied to SARS-CoV-2 nsp3 defined six regions to be modeled by the local modeling (LM), fragment-based (FB) and/or *ab initio* (AB) approaches. The first region, 2-169, of nsp3 sequence was modeled using the combination LM+FB (figure S1B), so that the structured region (5-111) could be determined in high resolution via LM of the side chain using a highly similar template (PDB 2gri, identity 76%). The bound conformation of intrinsically disordered segments was predicted using FB, a more flexible method. The choice of prediction method for protein regions with templates of high identity includes a thorough structural analysis of the template. For example, the region 413-676 of nsp3 is aligned with a high identity template (2w2g, 76%), so that modeling only variant side chains was considered. However, due to the predicted loss of a disulfide bridge, the FB approach was used since it allows larger conformational changes (C).

Appendix 4: Supplementary Tables

Table S1. Allele frequencies for the *Alu* insertion (I) and deletion (D) in the ACE gene across 199 populations from the ALlele FREquency Database (ALFRED) at the Yale Center for Medical Informatics.

Population	I	D
Africa_Kung_San	0·35	0·65
Africa_Nguni	0·42	0·58
Africa_San	0·31	0·69
Africa_Sotho-Tswana	0·32	0·68
Africa_Tsonga	0·22	0·78
Africa_African_Pygmies	0·52	0·48
Africa_Afro-Caribbeans	0·26	0·74
Africa_Biaka	0·44	0·56
Africa_Hausa	0·41	0·59
Africa_Ibo	0·38	0·62
Africa_Mbuti	0·35	0·65
Africa_Yoruba	0·37	0·63
Africa_Chagga	0·20	0·80
Africa_Jews_Ethiopian	0·38	0·62
Africa_Masai	0·46	0·54
Africa_African_Americans	0·41	0·59
Africa_Berber	0·29	0·71
Africa_Moroccans	0·27	0·73
Africa_Nigerians	0·33	0·68
Africa_Tunisian	0·33	0·67
Asia_Anatolian	0·29	0·71
Asia_Arabs	0·48	0·52
Asia_Armenian	0·22	0·78
Asia_Azerbaijani	0·35	0·65
Asia_Bahraini	0·50	0·50
Asia_Druze	0·33	0·67
Asia_Emirati	0·40	0·61
Asia_Jews_Yemenite	0·38	0·63
Asia_Jordanian	0·44	0·56

Asia_Samaritans	0·28	0·73
Asia_Saudi	0·00	1·00
Europe_Abazians	0·55	0·45
Europe_Adygei	0·47	0·53
Europe_Albanian	0·43	0·57
Europe_Aromanian	0·39	0·62
Europe_Azorian	0·52	0·49
Europe_Bashkir	0·39	0·61
Europe_Basque	0·43	0·57
Europe_Bosniaks	0·38	0·62
Europe_Bretons	0·37	0·63
Europe_Canarian	0·30	0·70
Europe_Cherkess	0·54	0·46
Europe_Chuvash	0·46	0·54
Europe_Croatian	0·37	0·63
Europe_Cypriot_Greek	0·35	0·65
Europe_Cypriot_Turkish	0·50	0·50
Europe_Czech	0·42	0·58
Europe_Danes	0·17	0·83
Europe_Darghinian	0·45	0·55
Europe_English	0·51	0·49
Europe_European_Americans	0·40	0·60
Europe_Europeans_Mixed	0·50	0·50
Europe_Finns	0·42	0·58
Europe_French	0·46	0·54
Europe_Gaguazes	0·40	0·60
Europe_Galician	0·35	0·65
Russia_Georgians	0·31	0·69
Europe_Greeks	0·34	0·66
Europe_Ingush	0·43	0·57
Europe_Irish	0·34	0·66
Europe_Italians	0·34	0·66
Europe_Jews_Ashkenazi	0·27	0·73

Europe_Kabardinian	0·61	0·39
Europe_Kalmyks	0·53	0·48
Europe_Karachay	0·38	0·62
Europe_Kumyk	0·48	0·52
Europe_Macedonian	0·31	0·69
Europe_Mari	0·48	0·52
Europe_Moldovan	0·70	0·30
Europe_Mordvin	0·46	0·54
Europe_Nogay	0·37	0·63
Europe_Portuguese	0·45	0·55
Europe_Roma	0·43	0·57
Europe_Romanian	0·57	0·43
Europe_Russians	0·28	0·72
Europe_Sardinian	0·39	0·61
Europe_Serb	0·41	0·59
Europe_Spaniards	0·38	0·62
Europe_Swiss	0·40	0·61
Europe_Tatar	0·54	0·46
Europe_Udmurt	0·40	0·60
Europe_Ukrainian	0·62	0·38
Asia_Burusho	0·57	0·43
Asia_Kalash	0·44	0·56
Asia_Pakistani	0·54	0·46
Asia_Pashtun	0·60	0·40
Asia_Dungan	0·56	0·44
Asia_Indian_Mixed	0·55	0·45
Asia_Kazakh	0·85	0·15
Asia_Khanty	0·59	0·41
Asia_Komi-Zyrian	0·63	0·37
Asia_Kyrgyz	0·44	0·56
Asia_Tajik	0·44	0·56
Asia_Turks	0·57	0·43
Asia_Uzbek	0·42	0·58

Asia_Agharia	0·48	0·52
Asia_Ambalakarer	0·46	0·54
Asia_Badaga	0·68	0·32
Asia_Bagdi	0·55	0·45
Asia_Brahmin	0·70	0·30
Asia_Chakma	0·70	0·30
Asia_Chamar	0·60	0·40
Asia_Gaud	0·61	0·39
Asia_Gond	0·65	0·35
Asia_Halba	0·68	0·32
Asia_Irula	0·68	0·32
Asia_Jamatiya	0·64	0·36
Asia_Kamar	0·61	0·39
Asia_Kapu	0·68	0·32
Asia_Kashmiri	0·61	0·39
Asia_Khonda_Dora	0·59	0·41
Asia_Kotas	0·55	0·45
Asia_Kshatriya	0·81	0·19
Asia_Kurumba	0·86	0·14
Asia_Lodha	0·69	0·31
Asia_Madiga	0·56	0·44
Asia_Mahishya	0·69	0·31
Asia_Mala	0·58	0·42
Asia_Mog	0·59	0·41
Asia_Moor_Sri_Lanka	0·64	0·36
Asia_Munda	0·53	0·47
Asia_Muria	0·58	0·42
Asia_Nepalese	0·65	0·35
Asia_Pallan	0·56	0·45
Asia_Parsi	0·54	0·46
Asia_Rajput	0·55	0·45
Asia_Relli	0·62	0·38
Asia_Riang	0·47	0·53

Asia_Santal	0·62	0·38
Asia_Sri_Lankan	0·65	0·35
Asia_Tamil	0·45	0·55
Asia_Tanti	0·60	0·40
Asia_Tharu	0·60	0·41
Asia_Tipperah	0·47	0·53
Asia_Toda	0·62	0·38
Asia_Vanniyar	0·66	0·34
Asia_Vellala	0·44	0·56
Asia_Vysya	0·53	0·47
Asia_Yadava	0·55	0·45
EastAsia_Ami	0·62	0·38
EastAsia_Atayal	0·70	0·31
Asia_Ewenki	0·61	0·39
EastAsia_Han	0·42	0·58
EastAsia_Japanese	0·58	0·43
EastAsia_Koreans	0·58	0·42
EastAsia_Sibo	0·50	0·50
EastAsia_Taiwanese	0·57	0·43
EastAsia_Uyghur	0·79	0·21
EastAsia_Cambodians_Khmer	0·53	0·47
EastAsia_Filipino	0·68	0·32
EastAsia_Hakka	0·62	0·39
EastAsia_Moluccas	0·63	0·38
EastAsia_Nusa_Tengarras	0·56	0·44
EastAsia_Vietnamese	0·86	0·14
EastAsia_Javanese	0·84	0·16
EastAsia_Malaysians	0·48	0·52
Oceania_New_Zealander	0·70	0·30
Oceania_Papuan_New_Guinean		
Oceania_Australian	1·00	0·00
Oceania_Melanesian_Nasioi	0·59	0·41
Oceania_Micronesians	0·64	0·36

Siberia_Buryat	0·78	0·22
Siberia_Chukchi	0·80	0·20
Siberia_Siberian_Eskimo	0·55	0·45
Siberia_Tuva	0·63	0·37
Siberia_Yakut	0·52	0·48
NorthAmerica_Alaskan_Natives		
NorthAmerica_Cheyenne	0·49	0·51
NorthAmerica_French_Acadians		
NorthAmerica_Hispanic_Americ		
NorthAmerica_Inuit_Canadian	0·57	0·43
NorthAmerica_Inuit_Greenland		
NorthAmerica_Muscogee	0·74	0·26
NorthAmerica_Na-Dene	0·89	0·11
NorthAmerica_Navajo	0·56	0·44
NorthAmerica_Pima_Arizona	0·54	0·46
NorthAmerica_Pima_Mexico	0·77	0·23
NorthAmerica_Zuni	0·68	0·33
NorthAmerica_Maya_Yucatan	1·00	0·00
SouthAmerica_Ache	0·63	0·37
SouthAmerica_Ayoreo	0·50	0·50
SouthAmerica_Bari	0·42	0·58
SouthAmerica_Brazilian	0·73	0·27
SouthAmerica_Chimila	0·81	0·19
SouthAmerica_Cinta_Larga	0·93	0·07
SouthAmerica_Gaviao	0·83	0·17
SouthAmerica_Guarani	0·88	0·12
SouthAmerica_Guihiba	0·54	0·46
SouthAmerica_Kaingang	0·95	0·06
SouthAmerica_Karitiana	0·98	0·02
SouthAmerica_Parakana	0·73	0·27
SouthAmerica_Quechua	0·92	0·08
SouthAmerica_Surui	0·81	0·19
SouthAmerica_Ticuna	0·98	0·02

SouthAmerica_Wai-Wai	0·68	0·32
SouthAmerica_Xavante	0·75	0·25
SouthAmerica_Yanomami	0·94	0·06
SouthAmerica_Yuco	096	0·04
SouthAmerica_Zoro	0·00	1·00

Table S2. Raft-127 genes identified in Kowalski et al. 2007 and their differential expression in BAL samples

Symbol	Lipid rafts of cells infected with *P. aeruginosa*	DEGs
SPC25	microsomal signal peptide 25 kDa subunit	UP
TF	serotransferrin	UP
AGTRAP	angiotensin II receptor-associated protein	DOWN
AHNAK	Ahnak	DOWN
ARF1	ADP ribosylation factor 1	DOWN
BANF1	barrier to autointegration factor 1	DOWN
CD81	CD81	DOWN
CLIC1	nuclear chloride ion channel 27	DOWN
CLPTM1	cleft lip and palate associated transmembrane	DOWN
COX5A	cytochrome c oxidase subunit Va	DOWN
CTSD	cathepsin d	DOWN
CYCS	cytochrome c	DOWN
FLNA	filamin a	DOWN
FLOT2	flotillin 2	DOWN
GLUD1	glutamate dehydrogenase 1	DOWN
GOLGA7	Golgi autoantigen, golgin subfamily a, 7	DOWN
HINT1	histidine triad nucleotide binding protein 1	DOWN
HM13	minor histocompatibility antigen h13	DOWN
HNRNPF	hnrnp f	DOWN
IMPDH2	inosine monophosphate dehydrogenase 2	DOWN
IQGAP1	ras GTPase-activating-like protein IQGAP1	DOWN
KPNB1	importin beta 1	DOWN
LAMTOR5	hepatitis b virus x interacting protein	DOWN
LDHA	L-lactate dehydrogenase	DOWN
LMNA	lamin a	DOWN

MDH1	malate dehydrogenase	DOWN
MGST1	microsomal glutathione S-transferase 1	DOWN
MGST3	microsomal glutathione s-transferase 3	DOWN
MIF	macrophage migration inhibitory factor	DOWN
MSN	Moesin	DOWN
MVP	major vault protein	DOWN
MYDGF	C19orf10 protein	DOWN
MYH9	myosin heavy chain, nonmuscle, differentially spliced	DOWN
NDUFS3	NADPH-ubiquinone oxidoreductase 30 kDa	DOWN
NME1	nucleoside diphosphate kinase a	DOWN
NME2	nucleoside diphosphate kinase b	DOWN
OTUB1	ubiquitin-specific protease otubain 1	DOWN
PDIA6	protein disulfide isomerase-associated 6	DOWN
PHB2	prohibitin 2	DOWN
PPIF	peptidylprolyl isomerase F	DOWN
PRDX1	peroxiredoxin 1	DOWN
RAB11B	rab11b	DOWN
RAB5C	rab5c	DOWN
RAP1A	rap1a	DOWN
RAP1B	rap1B	DOWN
RER1	Rer1	DOWN
RHOA	rhoa	DOWN
RPS17	40s ribosomal protein s17	DOWN
S100A8	calgranulin B	DOWN
SOD1	superoxide dismutase	DOWN
ST13	hsc70-interacting protein	DOWN
SYPL1	pantophysin	DOWN
TFRC	transferrin receptor protein 1	DOWN

TMED10	transmembrane protein tmp21	DOWN
TOMM20	mitochondrial import receptor subunit	DOWN
TPM3	tropomyosin 3	DOWN
VAMP2	vesicle associated membrane protein 2	DOWN
AK2	adenylate kinase isoenzyme 2	na
ALDOC	fructose bisphosphate aldolase c	na
AOP2	antioxidant protein 2	na
ARF2	ADP-ribosylation factor 4	na
ASPH	aspartyl/asparaginyl beta-hydroxylase	na
BKCO1	transmembrane protein 32	na
CARD11	caspase recruitment domain protein 11	na
CD3D	CD3e-associated protein	na
CDH13	Cadherin 13	na
CISD1	cells protein MDS029	na
CLPP	putative atp-dependent clp protease proteolytic	na
CLTH	clathrin heavy chain 1	na
CSNK1A1	casein kinase I	na
CYP51	cytochrome p450 51	na
DHCR7	7 dehydrocholesterol reductase	na
DNAJC5	DNA homologue subfamily c member 5	na
DNASE1L1	dnase I-like	na
EBP	emopamil binding protein	na
EERG28	potential membrane protein C14orf1	na
EPRS	bifunctional aminoacyl-trna synthase	na
FASN	fatty acid synthase	na
FLNB	beta filamin	na
GANAB	glucosidase II alpha-subunit	na
GCN1	similar to yeast translation activator GCN1	na

GLG1	golgi apparatus protein 1	na
HACD3	HSPC121	na
HCC1	nuclear protein hcc-1	na
HNRNPH2	heterogeneous nuclear ribonucleoprotein H2	na
HRAS	h-ras-1	na
IER3	homologue of mouse immediate early response 3	na
ITGA2	integrin alpha-2	na
ITIH5	inter-alpha (globulin) inhibitor H5-like	na
KRT15	cytokeratin 15	na
LAMTOR3	late endosomal/lysosomal mp1 interacting protein	na
LBR	lamin b receptor	na
LDLR	low-density lipoprotein receptor	na
LSS	lanosterol synthase	na
MAOA	amine oxidase a	na
MARCHF5	ring finger protein 153	na
MAVS	KIAA1271	na
MYL2	myosin regulatory light chain 2	na
NECTIN1	poliovirus receptor related protein 1	na
NUP155	nuclear pore complex protein nup155	na
PAG1	phosphoprotein associated with glycosphingolipid-enriched	na
PDIA4	protein disulfide isomerase a4	na
PEX13	peroxisomal membrane protein 20	na
PFK6	6-phosphofructokinase	na
PSE1	importin beta 3	na
PSEN1	presenilin 1	na
RAB18	rab18	na
RAB33B	rab33b	na
RAB35	rab35	na

RAC1C	isoform Rac1c	na
RAC1C	ras-related C3 botulinum toxin substrate 1	na
RALA	Ras-related protein ral-a	na
RRAS	r-ras	na
RTN4	reticulon 4	na
SC2	synaptic glycoprotein sc2	na
SDH	succinate dehydrogenase	na
SLC39A10	SLC39A10 protein	na
SPCS1	microsomal signal peptidase 12kDa subunit	na
SPTB	beta spectrin	na
TENM4	transmembrane protein 4	na
TIMM50	homologue of yeast Tim50	na
TMEM33	db83	na
TPD52L2	tumor protein d54	na
TRAP1	TRAP-1	na
UQCR	ubiquinol cytochrome c reductase complex	na
VMA5	vacuolar atp synthase subunit c	na
XPO1	exportin 1	na

Table S3. Enrichment analysis in Gene Ontology (GO) for differentially expressed genes in BAL samples from COVID-19 patients and asthma controls.

GO term ID	Description	Fold-enriched	Biological Category
GO:1990928	response to amino acid starvation	3.75	Apoptosis
GO:2001242	regulation of intrinsic apoptotic signaling pathway	3.51	Apoptosis
GO:0038066	p38MAPK cascade	7.66	Cellular signaling
GO:0060071	Wnt signaling pathway, planar cell polarity pathway	4.21	Cellular signaling
GO:0048010	vascular endothelial growth factor receptor signaling pathway	3.7	Cellular signaling
GO:0006890	retrograde vesicle-mediated transport, Golgi to ER	3.4	ER/Golgi/Endosome
GO:1901998	toxin transport	3.67	ER/Golgi/Endosome
GO:0006984	ER-nucleus signaling pathway	3.76	ER/Golgi/Endosome
GO:0071985	multivesicular body sorting pathway	4.81	ER/Golgi/Endosome
GO:0045056	transcytosis	4.86	ER/Golgi/Endosome
GO:1904380	endoplasmic reticulum mannose trimming	5.11	ER/Golgi/Endosome
GO:0061738	late endosomal microautophagy	5.37	ER/Golgi/Endosome
GO:1903513	endoplasmic reticulum to cytosol transport	5.67	ER/Golgi/Endosome
GO:0036258	multivesicular body assembly	5.79	ER/Golgi/Endosome
GO:0045324	late endosome to vacuole transport	6.81	ER/Golgi/Endosome
GO:1904903	ESCRT III complex disassembly	7.15	ER/Golgi/Endosome
GO:0006907	pinocytosis	7.43	ER/Golgi/Endosome
GO:0007042	lysosomal lumen acidification	7.66	ER/Golgi/Endosome
GO:0006614	SRP-dependent targeting to membrane	9.25	ER/Golgi/Endosome
GO:0016601	Rac protein signal transduction	3.89	Immune
GO:0050860	negative regulation of T cell receptor signaling pathway	3.89	Immune
GO:0042535	positive regulation of tumor necrosis factor biosynthetic process	4.08	Immune
GO:0098761	cellular response to interleukin-7	4.08	Immune
GO:0090382	phagosome maturation	4.27	Immune
GO:0048143	astrocyte activation	4.3	Immune
GO:0010759	positive regulation of macrophage chemotaxis	4.77	Immune
GO:0038061	NIK/NF-kappaB signaling	4.86	Immune
GO:0019882	antigen processing and presentation	4.92	Immune
GO:0097178	ruffle assembly	5.11	Immune

GO:0045657	positive regulation of monocyte differentiation	5.57	Immune
GO:0032930	positive regulation of superoxide anion generation	6.13	Immune
GO:0035722	interleukin-12-mediated signaling pathway	6.22	Immune
GO:0002479	antigen processing via MHC class I, TAP-dependent	6.99	Immune
GO:0019076	viral release from host cell	8.51	Immune
GO:2001188	regulation of T cell activation via MHC - antigen presenting cell	8.51	Immune
GO:0002399	MHC class II protein complex assembly	10.21	Immune
GO:0071226	cellular response to molecule of fungal origin	10.21	Immune
GO:0071826	ribonucleoprotein complex subunit organization	3.46	LipidRaft
GO:0034381	plasma lipoprotein particle clearance	3.57	LipidRaft
GO:1990776	response to angiotensin	4.16	LipidRaft
GO:0046479	glycosphingolipid catabolic process	5.5	LipidRaft
GO:0055094	response to lipoprotein particle	4.16	Lipoprotein
GO:0071402	cellular response to lipoprotein particle stimulus	4.58	Lipoprotein
GO:0010288	response to lead ion	3.68	Metal ion
GO:0033572	transferrin transport	5.11	Metal ion
GO:0000715	nucleotide-excision repair, DNA damage recognition	4.44	Nuclease
GO:0032069	regulation of nuclease activity	6.03	Nuclease
GO:0032075	positive regulation of nuclease activity	8.93	Nuclease
GO:0000084	mitotic S phase	5.34	Proliferation
GO:1902410	mitotic cytokinetic process	5.83	Proliferation
GO:0006457	protein folding	3.42	Protein Folding
GO:0042982	amyloid precursor protein metabolic process	4.3	Protein Folding
GO:1903332	regulation of protein folding	5.11	Protein Folding
GO:0097242	beta-amyloid clearance	5.77	Protein Folding
GO:0042987	amyloid precursor protein catabolic process	5.96	Protein Folding
GO:1900223	positive regulation of beta-amyloid clearance	7.29	Protein Folding
GO:0018279	protein N-linked glycosylation via asparagine	4.02	Protein Metabolism
GO:0006517	protein deglycosylation	4.08	Protein Metabolism
GO:0051444	negative regulation of ubiquitin-protein transferase activity	4.2	Protein Metabolism
GO:0006521	regulation of cellular amino acid metabolic process	6.06	Protein Metabolism
GO:0006413	translational initiation	7.71	Protein Metabolism
GO:0006091	generation of precursor metabolites and energy	3.48	Respiration

GO:2000191	regulation of fatty acid transport	3.65	Respiration
GO:0098780	response to mitochondrial depolarisation	4.08	Respiration
GO:0045730	respiratory burst	5.35	Respiration
GO:0019646	aerobic electron transport chain	6.13	Respiration
GO:0019322	pentose biosynthetic process	10.21	Respiration
GO:0043620	regulation of DNA-templated transcription in response to stress	4.7	RNA metabolism
GO:0061418	regulation of transcription in response to hypoxia	5.89	RNA metabolism
GO:1902415	regulation of mRNA binding	6.13	RNA metabolism
GO:1904874	positive regulation of telomerase RNA localization to Cajal body	6.81	RNA metabolism
GO:0000184	mRNA catabolic process, nonsense-mediated decay	7.66	RNA metabolism
GO:0021762	substantia nigra development/14-3-3 family of proteins	4.22	Serine phosphorylation
GO:0070262	peptidyl-serine dephosphorylation	4.77	Serine phosphorylation

Table S4. Classification of conservative substitutions considered in this study. Amino acids in brackets can be considered to be part of a group after structural analysis.

Group	Amino acid
Aromatic	Y, W, [H], [F]
Non-polar	A, I, V, L, M, [F], [P]
Polar, hydrogen bond interacting	S, T, [C]
Amidic, hydrogen bond interacting	N, Q
Acidic	D, E
Basic	R, H, K
Glycine	G

Table S5. Average root mean-square deviation (RMSD) of models built with the workflow relative to experimentally determined structures (*Workflow*). In parenthesis, the region modeled is defined. *Exp*.: Resolution of experimentally determined structure; *Other*: average RMSD of models generated with C-I-TASSER pipeline.[41] Structure alignment was conducted using Lovoalign.[42]

Protein	Exp. / Å	Workflow / Å	Other / Å
nsp5 (full)	2.0	1.5	2.8
nsp3 (205-374)	1.5	1.5	-
nsp15 (27-332)	2.2	0.7	0.5

Appendix 5: Structural Analysis of the SARS-CoV-2 Proteome

5.1 Nonstructural protein 1 (nsp1)

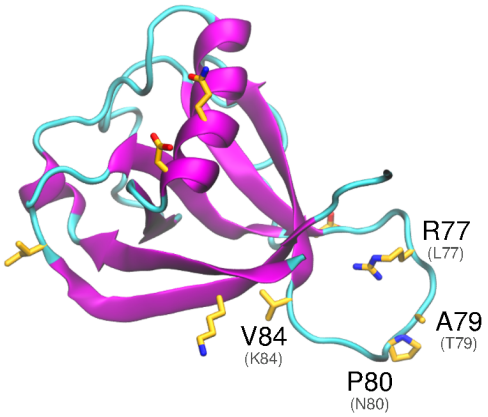


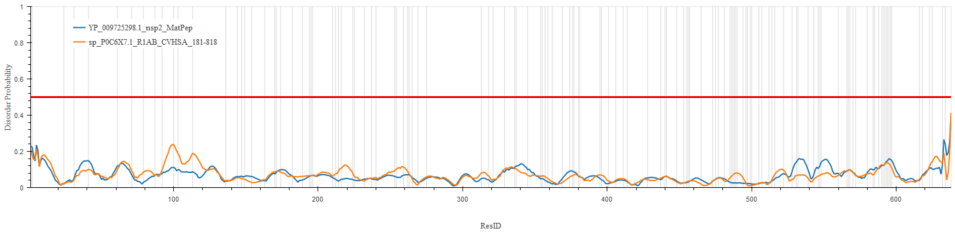
Figure S7. Predicted structure of SARS-CoV-2. Non-conservative substitutions relative to SARS-CoV-1 are depicted in *orange*.

The nonstructural protein 1 (nsp1) is the first nonstructural protein in the ORF1a/ORF1ab gene. Nsp1 is found within α and β coronaviruses (figure S7). Experiments *in vitro* suggest that SARS-CoV-1 nsp1 disrupts the host interferon defense response by potentially affecting the downstream defense signaling.[44–46] The nsp1 protein binds the 40S Ribosomal subunit, which has been associated with degradation of host mRNA and suppression of host mRNA translation, leaving the viral RNA unaffected.[46,47] The resultant complex cleaves the 5' UTR of host mRNAs, inhibiting translation.

5.2. Nonstructural protein 2 (nsp2)

Positive RNA viruses create viroplasms, subcellular compartments that are thought to protect the viral machinery against host-defense strategies, and even facilitate replication.[48] They often contain double stranded RNA, as well as the replication complex of the virus. For SARS-CoV-1, the replicase proteins assemble at the cytosol side of the ER, after which invagination occurs. There is no evidence of a pore maintaining the connection of the invaginated space to the cytosol, thus no maintained sphererule has been observed, though vesicles do form. A network of double-membrane-vesicles (DMVs) is eventually formed, anchored to the ER by a convoluted-membrane (CM) compartment. Replication appears to happen in the CM, while DMVs directly connected to the CM contain replicase proteins and viral proteins. Those DMVs not directly connected to CM form vesicle packets containing viral genomes.[48,49]
In Hagemeijer *et al*.[50], it was shown that the nonstructural protein 2 (nsp2) localized to the membranes of CM and DMV, and it is exposed to the cytoplasm. It appears that during formation of the replication complex, nsp2 is recruited, but once formed no further exchange of nsp2 occurs. Mutants of SARS-CoV-1 with the deleted nsp2 coding sequence are still capable of viral replication, although with decreased growth.[51] The specific role of nsp2 has not been established; it may be that nsp2 assists other viral proteins in performing their function, such as regulating the autophagy defense response or promoting mitochondrial dysfunction, thereby helping viral replication or effecting disease severity. It has been shown that nsp2 interacts directly with prohibitin 1 (PHB1) and PHB2[52] host proteins with a wide variety of functions. Kathiria et al.[53] have shown that PHB knockdown resulted in increased ROS, mitochondrial depolarization, and induced autophagy. In[54] several phenotypes of PHB are reviewed, including the role of PHB in mitochondrial stability. PHB has also been implicated in inflammatory response in both the lung and gut.[54,55] In von Brunn et al.[56], it was shown that nsp2 displayed co-immunoprecipitation (CoIP) interaction with N-terminal side of nsp3, nsp6, nsp8, nsp11, nsp16 and ORF3a, and co-localization with nsp8 and nsp3, where nsp8 almost always co-localized with the microtubule protein, LC3, an autophagy marker protein.[57] These results suggest that nsp2 may be involved in mitochondrial dysfunction through PHB1 and 2, and association with LC3 may be worth investigating. Another potential for nsp2's association with mitochondrial dysfunction may lie with its CoIP interaction with ORF3a, as the SARS protein ORF3a is known to disrupt mitochondrial stability.[58]

A



B

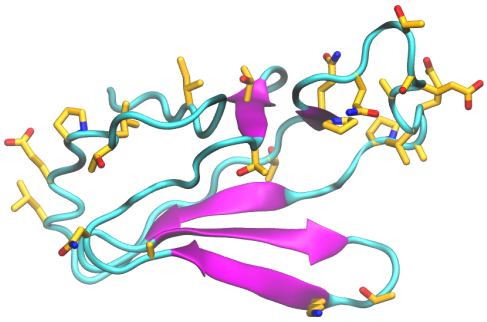


Figure S8. A) Profile of disordered/structured regions of SARS-CoV-2 nsp2, predicted using DisEMBL[59], using the *hot loop* definition. The red line indicates the 50% threshold for disorder, comparing nsp2 SARS-CoV-1 (*yellow* line), to nps2 SARS-CoV-2 (*blue* line). The grey vertical lines indicate the location of the non-conserved substitutions in SARS-CoV-1 nsp2 relative to SARS-CoV-2 nsp2. B) C-terminal domain of SARS-CoV-2 nsp2. Non-conservative substitutions relative to SARS-CoV-1 are depicted in *orange*.

Structural analysis and comparison with SARS-CoV-1 nsp2 - Structure information about nsp2 is scarce. SARS-CoV-2 nsp2 is 68% identical to SARS-CoV-1 nsp2. Among the substitutions, 130 are non-conservative and highly concentrated in the C-terminal end of the protein (figure S8A). An *ab initio* model was generated for this region (556-633, figure S8B). Fewer mutations appear at the N-terminal side, and a highly conserved region appears in the middle of the protein sequence.

5.3. Nonstructural protein 3 (nsp3) - Papain-like proteinase

The nonstructural protein 3 (nsp3) acts as a phosphatase and its catalytic domain is conserved, sharing homology with the Ymx7 protein in yeast, AF1521 in the archaea (*Archeoglobus fulgidus*), and Er58 in *E. coli*.[60] The protein comprises at least six domains: 1) an N-terminal ubiquitin-like (Ubl1) domain followed by glu-rich acidic domain; 2) an X domain with a predicted Appr-100-p processing activity; 3) a SUD domain (SARS-specific unique domain); 4) a peptidase C-16 domain that contains the papain-like protease (abbreviated PLnc); 5) a transmembrane domain; and 6) the Y domain.

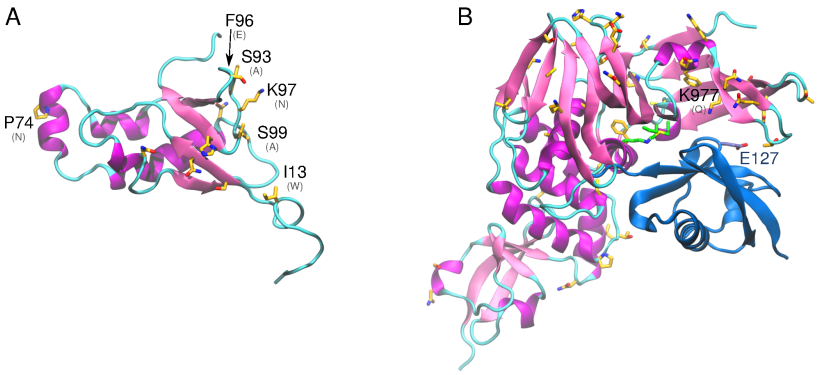


Figure S9. A) Predicted structure of N-terminal ubiquitin-like domain, and (B) papain-like protease domain of SARS-CoV-2 nsp3 bound to human interferon-stimulated gene product 15 (ISG15). Non-conservative substitutions relative to SARS-CoV-1 are depicted in *orange*. Key residues identified to bind ISG15 are conserved (in *green*). The complex with ISG15 was built based on PDB id 5tl6.

Structural analysis and comparison with SARS-CoV-1 nsp3 - Ub1 binds the nucleocapsid (N) protein and the interaction involves acidic residues of Ubl1 helix α2 and the serine and arginine rich region of the N protein.[61] Ub1 also binds to the 5' untranslated region of coronavirus RNA. Ub1 has significant structural homology with Ras effector

proteins, and it is thought that Ub1 may interact with and modulate the activity of Ras in the host. This interaction potentially affects growth and cell cycle cascades and could be a potential mechanism for cell mortality.[62] Significantly, there are differences in residue identity between SARS-CoV-1 and SARS-CoV-2 in residues 41 to 63, 83 to 87, 88 to 94, 95 to 98 of the Ub1 region (figure S9A). These regions correspond to structural homology with Ras effectors and these residue changes may affect interactions with host Ras and contribute to pathogenicity divergence.[62]

The glu-rich acidic region is also known as the hypervariable region (HVR). It is intrinsically disordered and its interaction partners are not entirely clear. One study by yeast-2-hybrid demonstrates interaction with nsp6, but a GST pull-down assay also revealed interactions with nsp8, nsp 9, and 3 regions of nsp3 itself (nucleic-acid binding domain, betacoronavirus-specific marker domain, and transmembrane region 1).[61] This region is significantly elongated, relative to the SARS-CoV-1 counterpart, with 16 additional amino acids, including several potential sites of glycosylation.

The X domain binds ADP-ribose, the structure of this complex was recently solved for SARS-CoV-2 (PDB id 6w02), and there are several studies and structures available in the Protein Data Bank of homolog complexes (PDB id: 5hol, 5dus, and 2fav). ADP-ribosylation is a type of post-translational modification (PTM), which may be implicated in inhibiting the immune response of the host through PTM of proteins related to expression of interleukin-6 and interferon-beta.[61] A GST-pulldown assay showed interaction of the X domain of nsp3 with nsp12. Several non-conservative substitutions occur in this domain relative to SARS-CoV-1, but those are located in superficial regions and do not significantly affect protein conformation (RMSD 1.5 Å, relative to 2fav). The papain-like protease domain (papain-like protease 2; PL2pro)[61] contains two ubiquitin binding sites. It is implicated in suppression of the host immune response, but its targets and, more generally, which immune-related signal transduction cascades are affected is unclear. It is implicated in inhibiting components of NF-κB, interferon-beta, and p53. In a structural study, $PL2^{pro}$ was found to bind ubiquitin-like interferon-stimulated gene product 15 (ISG15)[63], the latter an important post-translational modification of host proteins, including cytokines like interferon. It is believed that cleaving these posttranslational modifications of cytokine proteins by PL2pro disrupts the host immune response.[63] Importantly, ISG15 has significant interspecies variability, potentially contributing to its very different virulence patterns among species. This region is mostly conserved within SARS-CoV-1 and SARS-CoV-2, including residues that were identified as critical for the interaction with ISG15, namely Arg^{911}, Met^{953} and Pro^{992}. However, the substitution $Gln^{977}Lys$ likely intensifies the interaction with ISG15 by forming a salt bridge with Glu^{127}, suggesting an important mechanism for variable virulence (figure S9B).

5.4. Nonstructural protein 4 (nsp4)

Working in coordination with nsp3 and nonstructural protein 6 (nsp6), the nonstructural protein 4 (nsp4) of SARS-CoV-1 is essential for membrane rearrangements during viral replication.[64,65] Data suggests that coexpression of nsp3 with nsp4 results in host membrane rearrangement and the formation of double membrane vesicles and convoluted membranes.[66] Nsp4 in mouse hepatitis virus, another coronavirus, is a glycosylated protein and is demonstrated to be involved in virally induced membrane rearrangement, replication complex assembly, and assembly of double membrane vesicles.[67–71] Aberrant double membrane vesicles and impaired RNA replication can be observed when mouse hepatitis virus nsp4 is lacking a glycosylation site.[68,72] Prevention of interaction between nsp4 and nsp3 eliminated viral replication.[64]

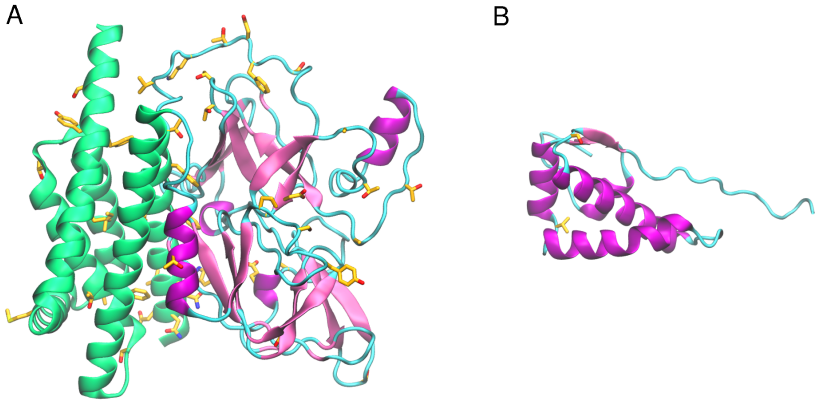


Figure S10. A) *Ab initio* model of region 1-402 of SARS-CoV-2 nsp4. B) Predicted structure of the conserved C-terminal domain of SARS-CoV-2. Transmembrane regions are shown in *green*, soluble domains in *pink* and non-conservative substitutions relative to SARS-CoV-1 nsp4 in *orange*.

Structural analysis and comparison with SARS-CoV-1 nsp4 - Nsp4 of SARS-CoV-1 is a nonstructural protein derived from the replicase polyprotein and is thought to be a tetra spanning transmembrane protein.[69] Disruption in glycosylation sites within the luminal loop between transmembrane domains 1 and 2 give rise to aberrant double membrane vesicles.[68] Key amino acid residues His^{120} and Phe^{121} of nsp4 of SARS-CoV-1, which are conserved in SARS-CoV-2, are essential for interaction and binding of nsp4 to nsp3 as well as viral propagation.[64]

Structural information about nsp4 is scarce. A low-resolution model was generated for the full nsp4 using the *ab initio* protocol (figure S10A). The soluble C-terminal domain is solved for a homologue (3vcb, MHV nsp4), enabling the generation of a higher resolution model for this region using the fragment-based modeling protocol (figure S10B). Despite the likely inaccurate orientation of domains in the full protein model, the predicted secondary structures are consistent with the expected positions of the transmembrane and soluble domains. This model is a useful resource for general structural analysis.

The sequence of SARS-CoV-2 nsp4 is 80% identical to SARS-CoV-1 nsp4. Among the non-conservative substitutions, the predicted model reveals the substitution of four cysteines in the predicted TM helices by the bulky amino acids, Trp^{6}, Phe^{7}, Phe^{24} and Phe^{390}. These substitutions may affect the arrangement and packing of transmembrane segments. Several substitutions are predicted to occur on the surface of the modeled soluble region spanning amino acids 33-275. The cytoplasmic C-terminal domain is highly conserved.

5.5. Nonstructural protein 5 (nsp5 or 3CL^{pro}) - 3C-like proteinase

The nonstructural protein 5 (nsp5, also known as 3CL^{pro}), is the main protease of the coronavirus genome, exhibiting a main role in the cleavage of the polyproteins translated from the viral RNA.[73–75] This protein is highly conserved relative to SARS-CoV-1 (96% identity) and among RNA+ viruses (Nidovirales) in general, making it an attractive target for pan-antiviral drugs,[76–78] In addition, it has been shown that loss of nsp3 and nsp10 substantially reduces 3CL^{pro} activity[79,80] and therefore therapeutics that target these proteins could indirectly inhibit 3CL^{pro}.

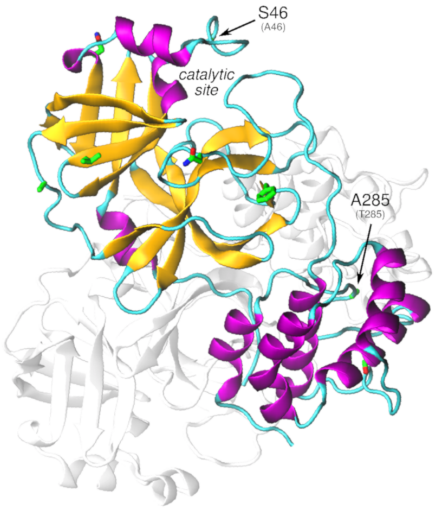


Figure S11. SARS-CoV-2 nsp5 dimer modeled using the crystal structure of the monomer and PDB 3m3v as dimer template. Non-conservative substitutions relative to SARS-CoV-1 nsp5 are depicted in *green*.

Structural analysis and comparison with SARS-CoV-1 nsp5 - Studies with SARS-CoV-1 show that dimerization is essential to stabilize the productive conformation of 3CL^{pro} catalytic site. The recently solved structure of 3CLpro of SARS-CoV-2 (PDB id: 6y2e) confirms the dimer as its biological state. The dimer interface is highly conserved within SARS-CoV-1 and -CoV-2, as well as other residues that indirectly affect dimerization, such as Ser^{144}, Ser ^{147} and Asn^{28}.[81] However, a relevant substitution Thr^{285}Ala is found on the interface. Based on previous studies with SARS-CoV-1 3CL^{pro}, this replacement is thought to favor the hydrophobic pack within monomers, and it was recently associated with a slightly higher catalytic efficiency of SARS-CoV-2 compared to SARS-CoV-2.[82] The analysis of the evolutionary tree from the aligned sequences of coronavirus from all available species reveals that alanine at site 285 defines the SARS-CoV-2 clade and three bat coronaviruses from mainland China (figure S12). In contrast, the many of the beta coronaviruses that infect mammals have a cysteine at this location. Given the proximity with the cysteine in the opposite monomer, it is possible that a disulfide bridge is formed in these proteases, which may result in a more tightly bound dimer and increased catalytic efficiency. Further exploration of this site is warranted.

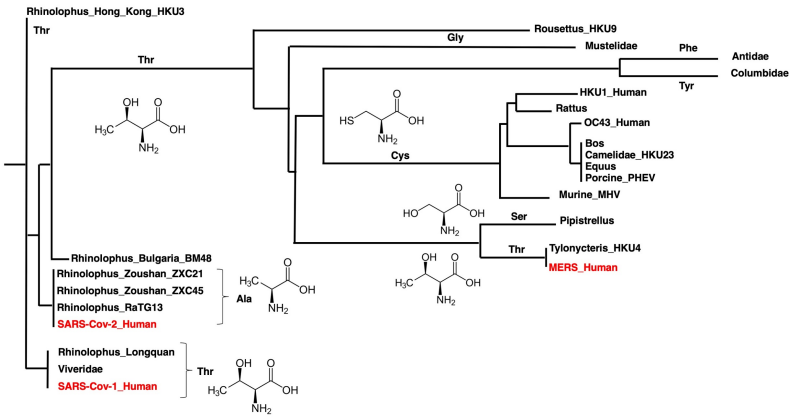


Figure S12. Phylogenetic tree of all known beta coronaviruses using the full protein sequence for 3CL^{pro}. *Red* text indicates viruses of concern for human health. The amino acid changes for the site at 285, which may affect dimerization are shown along branches. An alanine at this site defines the SARS-CoV-2 clade with the horseshoe bats from mainland China.

There are 7 non-conservative amino acid substitutions between SARS-CoV-1 and SARS-CoV-2, besides $Thr^{285}Ala$. Most of those are located in regions that are not clearly associated with protease function, except by $Ala^{46}Ser$, positioned at the catalytic cleft entrance, which may affect substrate affinity/selectivity. Moreover, the phosphorylation of serines in $3CL^{pro}$ of rotaviruses was shown to be essential for protease activity. The substitutions $Ala^{46}Ser$, $Ser^{65}Asn$, $Ser^{94}Ala$, $Ala^{267}Ser$ and $Thr^{285}Ala$ may also affect the phosphorylation pattern of $3CL^{pro}$ in SARS-CoV-2.

5.6. Nonstructural protein 6 (nsp6)

Upon infection, corona viruses create signature membrane rearrangements (e.g. DMVs) so that the viral RNA replication complex may anchor to host machinery.[83] Nsp6, along with nsp3 and nsp4, plays a critical role in this membrane rearrangement.[65] In SARS-CoV-1 nsp6 is known to activate autophagy[86] by inducing perinuclear vesicles localized around the microtubule organization center.[65]

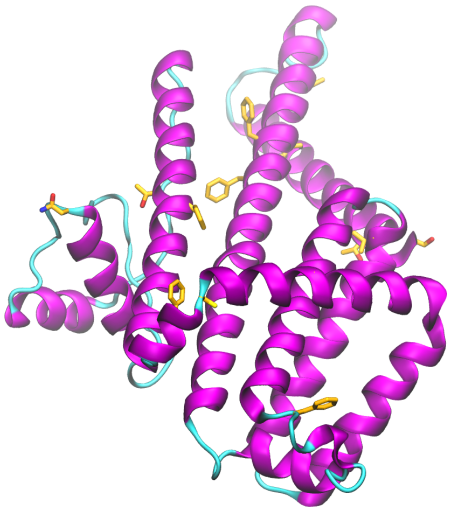


Figure S13. *Ab initio* predicted structure of SARS-CoV-2 nsp6. Non-conservative substitutions relative to SARS-CoV-1 nsp6 are depicted in *orange*.

Structural analysis and comparison with SARS-CoV-1 nsp6 - SARS-CoV-2 nsp6 is 87% identical to SARS-CoV-1 nsp6. Non-conservative mutations are located on the protein surface (figure S13).

5.7. Nonstructural proteins 7, 8 and 12 (nsp7, nsp8, nsp12) - Replication complex

After infection the virus assembles a multi-subunit RNA-synthesis complex consisting of several nonstructural proteins (nsps), namely nsp7, 8, 9, 10, and 12.[87] This complex is responsible for the replication and transcription of the viral genome. The complex of nsp12, 7, and 8 (total 160 kDa) constitutes the minimum set of nsps required for nucleotide polymerization.[88] Nsp12 encodes the RNA-dependent RNA polymerase (RdRp) domain. Nsp7 and nsp8 are nsp12 cofactors. Nsp7 forms a hexadecameric complex with nsp8 and together these may act as a processivity clamp for the RNA polymerase.[89] In the replication complex there are two well defined regions of nsp8: nsp8(I), forming the heterodimer with nsp7, and nsp8(II), binding to nsp12.[88]

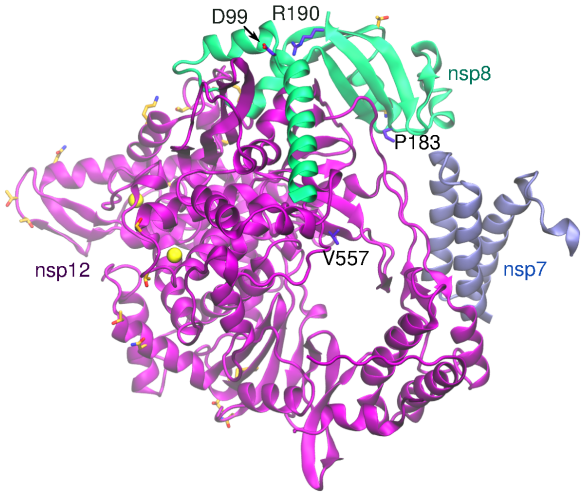


Figure S14. Local modeling-derived structure of SARS-CoV-2 nsp7-nsp8-nsp12 complex. Non-conservative substitutions are depicted in *orange*. Some key functional residues identified are depicted in *purple*. Zinc ions are represented as *yellow* spheres. PDB 6nur was used as the template to assemble the complex.

Structural analysis and comparison with SARS-CoV-1 nsp7-nsp8-nsp12 complex- A study conducted by Kirchdoerfer and Ward[88] showed that the polymerase domain of SARS-CoV-1 nsp12 **(**a.a 398-919**)** consists of the fingers domain (a.a 398–581, 628–687), palm domain (a.a. 582–627, 688–815) and a thumb domain (a.a. 816–919). The Nidovirus-unique N-terminal extension (aka NiRAN) is situated between a.a. 1–397. There are two metal (Zn) binding sites in nsp12: The first is in the NiRAN extension and is coordinated by residues His^{295}, Cys^{301}, Cys^{306}, and Cys^{310}. The second site is in the fingers domain and is coordinated by Cys^{487}, His^{642},Cys^{645}, and Cys^{646}. The binding site for the nsp7-nsp8 heterodimer overlaps with the conserved regions of polymerase functional domains (fingers (a.a 398–581, 628–687) and thumb domains (a.a. 816–919). The binding site between nsp8(II) and nsp12 is on the N-terminal region (77–126) of nsp8(II).

Several studies have pointed to functional residues in nsp7, 8, and 12 sequences. The study conducted by Lehmann et al.[90], shows conserved sequence motifs in the NiRAN domain named A_N, B_N, and C_N. These motifs are conserved throughout all members of the order *Nidovirales*, of which CoVs are members.[90] Another example is covalent modification of the Lys^{73} residue in nsp12 (using GTP or UTP) that reduced viral growth and recovery in the equine arterivirus.[90] *In vitro* polymerase activity assays further enabled the identification of other key functional sites, Lys^{7}, His^{36}, and Asn^{37}, which, being replaced by alanine, are associated with decreased polymerase activity.[91] Four other mutations in nsp8 were detrimental to polymerase activity namely, Pro^{183}Ala, Asp^{99}Ala, Pro^{116}Ala, and Arg^{190}Ala. These mutations are associated with a defective fold of nsp8 and disruption of nsp8(II)-nsp12 binding.[91]

The nucleoside analog GS-5734 is capable of impairing CoV RNA synthesis by targeting the viral RNA synthesis machinery. The side chain in a motif of the fingers domain in nsp12 is involved in GS-5734 interaction through Val^{557} residue.[92,93]

All the mentioned residues are fully conserved in SARS-CoV-2. All non-conservative substitutions are located in the complex surface, mostly in nsp12 (figure S14). Several of these substitutions involve residue that are potential sites for post-translational modifications (e.g., cysteines, serines, threonines, asparagines and tyrosines), indicating a variation in post-translational patterns relative to the SARS-CoV-1 RNA polymerase complex.

5.8. Nonstructural protein 9 (nsp9)

The nonstructural protein 9 (nsp9) is cleaved from the viral polyproteins by the $3CL^{pro}$ protease.[87,94] SARS-CoV-1 nsp9 is able to form dimers (figure S15) that bind either ssDNA and ssRNA[95] and is thought to protect the coronavirus genome from degradation during replication.[94,96] Deletion of nsp9 in the mouse β-coronavirus (hepatitis virus , MHV), impairs viral RNA synthesis and viral infection, where dimerization is particularly important.[97,98]

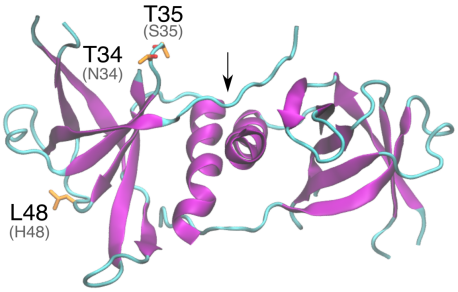


Figure S15. Cartoon representation of SARS-CoV-2 nsp9 dimer. Arrow indicates dimer interface and mutations are highlighted in orange.

Structural analysis and comparison with SARS-CoV-1 nsp9 - Overall, the structure of nsp9 is very well conserved among β-coronaviruses.[94,99] In SARS-CoV-1, nsp9 consists of a small globular protein (113 amino acid residues) with seven antiparallel β-sheets and one α-helix.[94] Parallel dimers occur between two nsp9 monomers interacting in their C-terminal α-helixes and their N-fingers.[99] In SARS-CoV-1, Gly^{100} and Gly^{104} residues are shown to be in the core of the dimer interface.[94] Mutation of these conserved glycines in the nsp9 α-helix impairs ssDNA binding, suggesting that dimerization is essential for nucleic acid binding.[99] Site-specific mutagenesis studies in the alpha-coronavirus Porcine Epidemic Diarrhea (PEDV) nsp9 identified other key residues for dimerization. The substitutions Lys^{10}Ala, Arg^{68}Ala, Lys^{69}Ala and Arg^{106}Ala decreases binding affinity to form the dimer 7·2-fold relative to wild-type nsp9, while Tyr^{82}Ala enhance dimer stability, by a 8·0-fold increase in binding affinity.[99] Additionally, Zeng et al. (2018), studied Porcine Epidemic Diarrhea Virus (PEDV) nsp9 residues involved in ssDNA binding activity and they identified five mutations that affected this process, namely, Lys^{10}Ala, Arg^{68}Ala, Lys^{69}Ala, Arg^{106}Ala and Tyr^{82}Ala. The equivalents in SARS-CoV-2 identified from a sequence alignment are Arg^{10}Ala, Arg^{74}Ala, Arg^{111}Ala, and Tyr^{87}Ala, which are conserved in SARS-CoV-1.

Because SARS-CoV-1 and SARS-CoV-2 sequences are highly conserved (sequence identity is 97%), much structural information from SARS-CoV-1 is transferable to SARS-CoV-2. The three substitutions in SARS-CoV-2 nsp2 relative to SARS-CoV-1, Asn^{34}Thr, Ser^{35}Thr and His^{48}Leu, are located on the nsp9 surface, distant from the dimerization region (figure S15). Particularly, substitution Asn^{34}Thr can result in an additional phosphorylation site, but phosphorylation of nsp9 has not yet been reported. As Thr^{34} and Thr^{35} are close to a potential ubiquitination site (Lys^{36}), we hypothesize that their phosphorylation may prevent nsp9 ubiquitination.

5.9. Nonstructural protein 10 (nsp10)

Nsp10 is one of sixteen nonstructural proteins of the SARS-CoV-2 proteome. It interacts with nsp14 and nsp16 to perform 3'-5' exoribonuclease and 2'-O-methyltransferase activities, respectively.[100,101] Disturbance in the interaction between nsp10 and nsp16 has been shown to be crippling to the virus, although still functional.[102] Nsp14, nsp16, and nsp10 have been associated with the coronavirus wide replication and transcription complex.[103,104] Crystal structures of SARS-CoV-1 nsp10-nsp16 have been reported.[105]

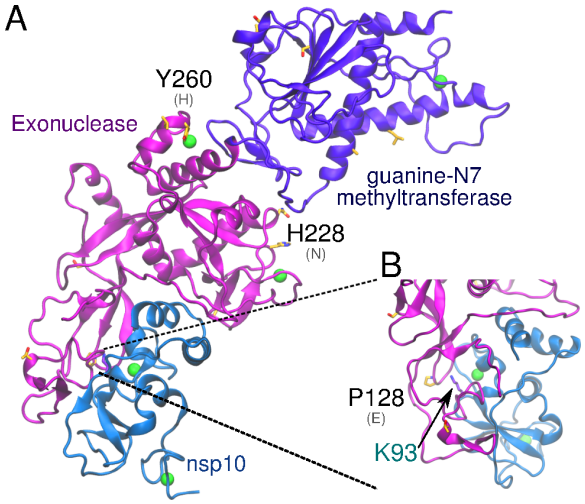


Figure S16. A) Local modeling-derived structure of SARS-CoV-2 nsp10-nsp14 complex. B) Closer view of residues Pro^{128} and Lys^{93} in nsp14 and nsp10 respectively. Non-conservative substitutions relative to SARS-CoV-1 nsp10 and nsp14 are depicted in *orange*. Zinc ions are represented as *green* spheres. PDB 5c8s was used as the template to assemble the complex.

Structural analysis and comparison with SARS-CoV-1 nsp9 - SARS-CoV-1 and SARS-CoV-2 nsp10 proteins are highly conserved (identity 97%), with no non-conservative substitutions between them. These proteins are composed of 139 a.a. residues and fold forming two zinc fingers.[102] This tight interaction with nsp14, involving multiple hydrogen bonds, salt bridges and hydrophobic packing, suggests that nsp10 might be necessary to maintain the integrity of nsp14 exoribonuclease domain.[106,107]

5.10. Nonstructural protein 11 (nsp11)

Nsp11 is a short protein only 13 a.a. in length and is translated from both polyproteins ORF1a and ORF1ab. In SARS-CoV-1 nsp11 has been implicated in RNA synthesis.[108] SARS-CoV-2 nsp11 has 85% sequence identity to SARS-CoV-1 nsp11, with two substitutions between them ($Ser^{5}Gln$ and $Thr^{6}Ser$).

5.11. Nonstructural protein 13 (nsp13) - helicase

The nonstructural protein 13 (nsp13), encoded from ORF1ab as part of the pp1ab polyprotein, is involved in a number of functions, such as NTPase, dNTPase, RTpase, RNA helicase, and DNA helicase activity.[39] It also interacts with nsp12, which is thought to enhance its helicase activity.[39,109]

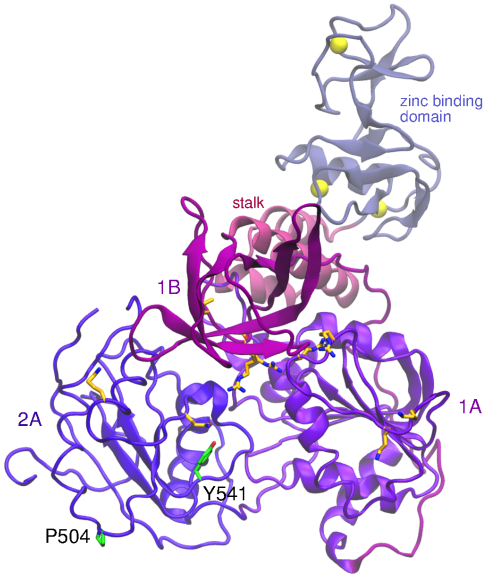


Figure S17. Predicted structure of SARS-CoV-2 nsp13. Non-conservative substitutions relative to SARS-CoV-1 nsp13 are depicted in orange. Mutations observed in SARS-CoV-2 nsp13 are depicted in green. Zinc ions are represented as yellow spheres.

Structural analysis and comparison with SARS-CoV-1 nsp13 - The helicase nsp13 contains 5 domains, an N-terminal zinc-binding domain, stalk domain, RecA-like domain 1A, 1B, and RecA-like domain 2A, all together consisting of 603 residues. It forms a triangular structure with the 2A and 1A as the base.[38]

The zinc-binding domain consists of conserved cysteine residues, which have shown to be essential for the various enzymatic activities of nsp13.[110] The stalk domain links the zinc-binding domain to the rest of the helicase, where the stalk domain has been argued to serve as an essential signal transduction mechanism.[38] A single mutation in the stalk domain of SARS-CoV-1 nsp13, Arg[132]Pro, resulted in a dramatic decrease in viral infectivity *in vitro*.[111] The essential nature of nsp13 in the viral replication cycle, and its multifunctional nature has made it an attractive target for vaccine research.[112113,114] SARS-CoV-2 nsp13 (figure S17) is fully conserved relative to SARS-CoV-1.

5.12. Nonstructural protein 14 (nsp14) - 3' to 5'- exonuclease

The nonstructural protein nsp14 plays a critical role in coronaviruses replication and transcription. The exonuclease domain of nsp14 is imperative for replication fidelity within RNA viruses and has been shown to function as a proofreading exoribonuclease. Nsp14 associates with other nonstructural proteins from ORF1a and ORF1ab, including nsp7, nsp8, nsp12, and nsp10.[91,106,115–117] The association with nsp10 provides the ability to excise mismatched nucleotides[107], and disruption of this heterodimer was shown to decrease replication fidelity.[118]

Structural analysis and comparison with SARS-CoV-1 nsp14 - Similar to SARS-CoV-1, SARS-CoV-2 nsp14 protein contains two catalytic domains: (1) The N-terminal exonuclease domain and (2) the C-terminal guanine-N7 methyltransferase domain. These domains, N-terminal and C-terminal, span a.a. sequences 1-287 and 288-527, respectively. The exonuclease domain also has conserved residues Asp, Glu, Asp and Asp (DEDD) that are key indicators of the DEDDh superfamily of DNA and RNA exonucleases.[87,119] The guanine-N7 methyltransferase domain has been shown to be necessary for viral mRNA capping[106], hiding viral mRNA from host mRNA degradation machinery.

Nsp14 contains two zinc fingers within the N-terminal domain. The first is composed of Cys[207], Cys[210], Cys[226] and His[229], while the second is composed of His[257], Cys[261], His[264], and Cys[279] [106]**.** Ala[1]-Arg[76] and Ala[119]-Asp[145] of nsp14 interact with nsp10. The third zinc finger of nsp14 is within the C-terminal domain and is formed by Cys[452], Cys[477], Cys[484], and His[487].[106]

SARS-CoV-2 nsp14 is 73% identical to the SARS-CoV-1 counterpart. Among the 14 non-conservative substitutions, Glu[128]Pro is the only substitution located close to the interface with nsp10 (figure S16B). The proximity with the

positively charged Lys[93] of nsp10 in the heterodimer model suggests that this substitution may slightly affect binding of nsp10-nsp14 (figure S16B). The substitutions His[260]Tyr and Asn[228]His may also be relevant as they are located in close proximity to the zinc fingers in the N-terminal exonuclease domain (figure S16A).

5.13. Nonstructural protein 15 (nsp15) - endoRNAse

Nsp15 is a Nidoviral RNA uridylate-specific endoribonuclease (NendoU) and its C-terminal is a catalytic domain belonging to the EndoU family of enzymes.[120–122] EndoU enzymes have RNA endonuclease activity producing 2'-3' cyclic phosphodiester and 5'-hydroxyl termini.[123] A study in 2017 proposed that the NendoU activity interferes with the innate immune response,[124] however this finding was disputed in a 2019 study[125] that showed that the interference was independent of the endonuclease activity.

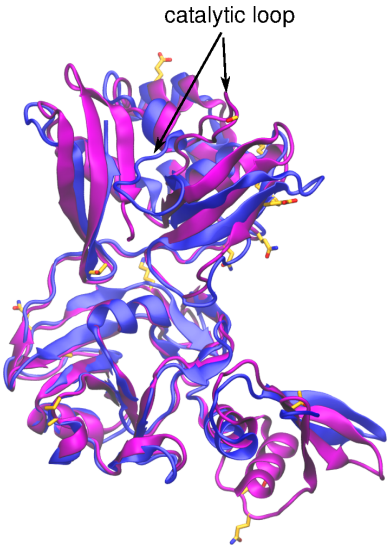


Figure S18. Superimposed crystal structures of SARS-CoV-1 (*blue*) and SARS-CoV-2 (*purple*) nsp15. Non-conservative residues in SARS-CoV-2 nsp15 relative to SARS-CoV-1 are depicted in *orange*.

Structural analysis and comparison with SARS-CoV-1 nsp15 - The nsp15 monomer (346 a.a.) forms hexamers. Each monomer consists of three domains, namely the N-terminal, middle domain, and catalytic NendoU domain at the C-terminal.[120] Within SARS-CoV-1 and SARS-CoV-2, nsp15 is very conserved (89% identity). However, the recently solved crystal structure of the nsp15 monomer, shows a significant conformational variation in the region of the catalytic site. Although several non-conservative substitutions occur around this region, the conformation of the catalytic loop (a.a. 234-249) is known to greatly change upon protein oligomerization (figure S18).

5.14. Nonstructural protein 16 (nsp16) - 2'-O-ribose methyltransferase

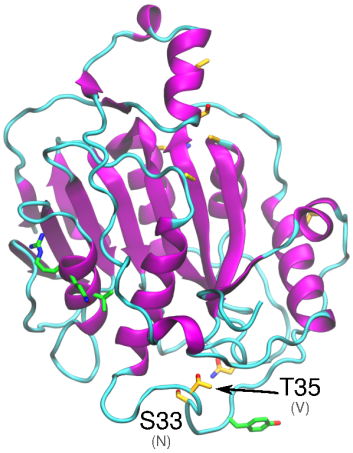


Figure S19. Local modeling-derived structure of SARS-CoV-2 nsp16. Non-conservative substitutions relative to SARS-CoV-1 nsp16 and depicted in *orange* and hey functional residues are depicted in *green*.

The nonstructural protein nsp16 is involved in capping of viral mRNA to protect it from host degradation, and it has been demonstrated that it has to be associated with nsp10 to be active. One potential method of ablating nsp16 activity is to disrupt the binding interface between nsp16 and nsp10. In SARS-CoV-1, the following interfacial mutations in nsp16 were shown to ablate nsp16 2'-O-methyltransferase activity: Ile^{40}Ala, Met^{41}Ala, Val^{44}Ala, Val^{78}Ala, Arg^{86}Ala, Val^{104}Gly, Leu^{244}Ala and Met^{247}Ala.[105] In a separate study, a double mutant (His^{83}Ala/Pro^{84}Ala) and a triple mutant (Tyr^{76}Ala/Cys^{77}Ala/Arg^{78}Ala) of nsp10 abolished SAM and m7GppA-RNA binding to nsp16[126], further supporting the essential role of nsp10 in activating nsp16. In terms of the physicochemical interfacial properties that are essential for activating nsp16, it appears that certain hydrophobic interactions play a critical role and that enhancement of these may increase nsp16. This is exemplified by the Tyr^{96}Phe mutation in nsp10 that increases methyltransferase activity by increasing hydrophobic interaction between nsp10 and the hydrophobic pocket of nsp16, which includes the sidechain of Val^{84} and the mainchain portions of Gln^{87} and Arg^{86}.[126] In SARS-Cov-1, residue Tyr^{30} in the RNA binding site of nsp16, when mutated to either alanine or phenylalanine, abolished methlyltransferase activity, suggesting the critical role of this residue in RNA binding and, consequently, the enzymatic function.[105] All these key amino acids are conserved in SARS-CoV-2 nsp16, but non-conservative substitutions in the vicinity of Tyr^{30}, namely Asn^{33}Ser and Val^{35}Thr may affect RNA binding, as well as conservative substitutions (Glu^{32}Adp and Ile^{36}Leu) that may have a steric effect (figure S19).

5.15. Spike glycoprotein (discussed along the main text)

5.16. ORF3a

ORF3a is the largest of all SARS-CoV accessory proteins and it is expressed in SARS-CoV-1 infected cells, incorporated in viral particles.[127] ORF3a is an integral transmembrane protein, localized mostly in the Golgi complex, cytoplasm, and cell surface.[128–130] Deletion of several accessory proteins including ORF3a do not dramatically decrease viral replication, suggesting they are not essential for viral growth.[131] However, the most significant decrease in viral growth was observed when ORF3a was deleted.[131] ORF3a is able to form homodimers via disulfide bridges, which, in turn, form homotetramers through non-covalent interactions.[130] The homotetramers have ion channel properties able to transport Na^+, K^+, and Ca^{+2} ions.[132] Although ion channel activity of ORF3a was shown to not be required for viral replication in mouse lungs[132], if ORF3a is silenced, viral release decays proportionally to the amount of siRNA used.[130]

ORF3a has been demonstrated to interact with M, N and S viral proteins.[133] For instance, it forms disulfide bonds with S protein via the cysteine rich motif.[134] A number of mutations in ORF3a correlates with mutation in the S protein in different SARS-CoV-1 isolates, suggesting that its function is linked to the S protein.[134] Additionally, ORF3a and the E protein are both required for maximum viral growth.[132] Deletion of both ORF3a and E protein binding modules yielded no viable viruses, while the presence of only one of them was sufficient to revert the phenotype.[132]

Regarding pathogenic effects, ORF3a can trigger inflammatory responses by inducing the expression of NB-κB and IL-8.[135] Besides, ORF3a has been shown to be involved in inflammatory responses by promoting activation of NLRP3

inflammasome.[136] NLRP3 activation is achieved by the ubiquitination of apoptosis-associated speck-like protein containing a caspase recruitment domain (ASC) facilitated by TNF receptor-associated factor 3 (TRAF3) and ORF3a direct interaction.[136] Furthermore, ORF3a triggers unfolded protein response (UPR) in the endoplasmic reticulum via PKR-like ER kinase (PERK) pathway, thereby increasing IFN alpha-receptor subunit 1 (IFNAR1) phosphorylation, ubiquitination and lysosomal degradation.[137] Thus, ORF3a may weaken INF responses and innate immunity responses. ORF3a has apoptotic and cell cycle regulation functions via caspase pathways and by arresting cells at G1 stage.[58,138] Further, if either tyrosine (YXXφ), di-acidic (EXD), or cysteine rich motifs are mutated the apoptotic effects of ORF3a are reduced.[139] Furthermore, ORF3a potassium binding activity is blocked, apoptosis is also blocked.[139] From a clinical perspective, some patients recovered from SARS showed joint pain and bone damage. Consequently, the study conducted by Obitsu et al.[140] showed that ORF3a is able to induce osteoclastogenesis and thus bone abnormalities observed in SARS survivors. Further, some patients developed pulmonary thrombosis during SARS-CoV-1 infection[141], while ORF3a was reported to increase expression of all subunits of fibrinogen.[142] Taken together, ORF3a may contribute to the formation of blood clots in some patients.

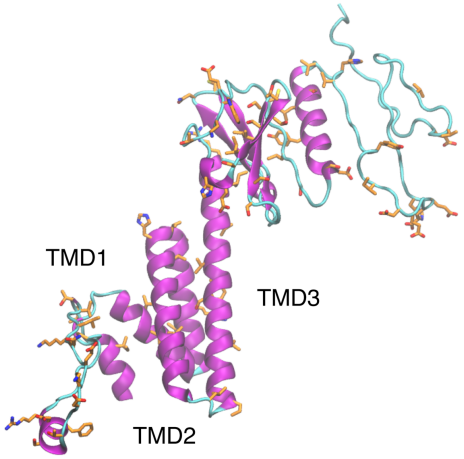


Figure S20. *Ab initio* predicted structure of SARS-CoV-2 ORF3a. Non-conservative substitutions relative to SARS-CoV-1 ORF3a are depicted in *orange*.

Structural analysis and comparison with SARS-CoV-1 ORF3a - ORF3a protein is the biggest of all SARS-CoV-1 accessory proteins with 274 a.a. Structurally, ORF3a possess a signal peptide (1 - 15) and three transmembrane domains TMD1 (34 - 56), TMD2 (77 - 99), and TMD3 (103 - 125) and a cytoplasmic domain at its C-terminal of 149 residues.[134,143,144] Several important residues have been identified for ORF3a function. At ORF3a cytoplasmic domain, there is a predicted calcium binding domain at residues 209 - 264 that goes through conformational changes upon binding calcium *in vitro*.[143] Besides, a cysteine rich motif (Cys-Trp-Leu-Cys-Trp-Lys-Cys) from residues 127 - 133, located at the end of TMD3 region is essential for ORF3a and S protein interaction.[134] The tyrosine motif Tyr-Asn-Ser-Val (160 - 163) was found essential for a proper ORF3a export to the cell surface.[129] Mutations at Tyr-Asn-Ser-Val motif, increase ORF3a retention at Golgi complex and lysosome degradation.[129] Further, the motifs Tyr-Asn-Ser-Val, di-acidic (Glu-Gly-Asp, 172 - 173) and cysteine rich motifs, have been associated with apoptotic function of ORF3a.[139] Mutations at 176 - 274 residues pointed this region to be important to induce cell cycle arrest at G1 stage.[138] Regarding subcellular transport, yeast two-hybrid assays showed that caveoline-1 interacts directly to ORF3a N-terminal domain.[145] ORF3a contains three caveolin binding sites (Trp-Gln-Leu-Ala-Leu-Tyr-Lys-Gly-Phe, Tyr-Leu-Tyr-Ala-Leu-Ile-Tyr-Phe, and Tyr-Asp-Ala-Asn-Tyr-Phe-Val-Cys-Trp).[145] if one or two are mutated ORF3a colocalize still with caveolin-1, suggesting that only one site is enough for this interaction to take place.[145] Like M protein, ORF3a is glycosylated and substitutions Ser^{27}Gly, Thr^{28}Ala, Thr^{32}Ala and Thr^{34}Ala prevents its glycosylation.[146] Further substitutions at TMD2 Tyr^{91}Ala and His^{93}Ala, and TMD3 Tyr^{109}Ala eliminated ORF3a channel activity.[132] Besides, residues 125 - 200 are predicted to be β-sheet and are able to bind at the 5'-UTR the viral genomic RNA, which may in turn help with virion assembly.[147] Finally, ORF3a contains a PDZ domain-binding motif (PBM, Ser-Val-Pro-Leu, 271 - 274) required for maximum viral yield.[132]

SARS-CoV-2 ORF3a is a protein of 275 residues with three transmembrane domains (TMD) and a C-terminal cytoplasmic domain (figure S20). Equal to SARS-CoV-1, the predicted transmembrane positions for SARS-CoV-2 ORF3a are located at 34 - 56 (TMD1), 77 - 99 (TMD2) and 103 - 125 (TMD3). In total 76 mutations created a sequence identity of 72·36% between SARS-CoV-1 and SARS-CoV-2. The mutations are distributed along the whole ORF3a

sequence, but some conservated zones were observed at TMD1, TMD2, TMD3, cysteine-rich domain and the last portion of C-terminal domain (244 - 275). TMD2 Tyr^{91} and His^{93} as well as TMD3 Tyr^{109} were conserved, reinforcing the idea that these residues are important for ORF3a ion channel activity. One mutation was observed in the cysteine rich motif Cys^{127}Leu (Leu-Trp-Leu-Cys-Trp-Lys-Cys). This substitution may affect ORF3a and S protein interaction, but it is likely that the other two cysteines are enough to establish the interaction. The tyrosine motif Tyr-Asn-Ser-Val was fully conserved likely due to its importance for ORF3a function. Besides, the Glu^{171}Ser was found in the di-acidic domain (Ser-Gly-Asp). This mutation may have a reductive effect in ORF3a apoptotic feature.[139] Truncation of residues 176 - 274 impacts the role of ORF3a in cell cycle arrest.[138] Ten mutations were detected at this zone in SARS-CoV-2 ORF3a, but also a conserved motif Asp-Tyr-Gln-Ile-Gly-Gly-Tyr-Thr-Glu with only Ser^{190}Thr substitution was observed. Thus, this conserved motif might be enough to carry out cell cycle arrest, but this hypothesis requires proper testing.

The three caveoline-1 binding sites were highly similar (Trp-Gln-Leu-Ala-Leu-Ser-Lys-Gly-Val, Tyr-Leu-Tyr-Ala-Leu-Val-Tyr-Phe and Tyr-Asp-Ala-Asn-Tyr-Phe-Leu-Cys-Trp) with only one non-conservative substitution Tyr^{74}Ser and two conservative substitutions, Ile^{112}Val and Val^{147}Leu. The Tyr^{74}Ser substitution led us to hypothesize that residue could be a potential phosphorylation site. Although *in silico* analysis of phosphorylation yielded 29 putative sites, phosphorylation of ORF3a has not been demonstrated yet. Non-conservative substitutions were found in two out of four putative glycosylation sites (Ser^{27}Asp and Thr^{28}Phe), meanwhile Thr^{32} and Thr^{34} were conserved. This observation suggests that glycosylation of ORF3a may occur in these residues. Furthermore, Thr^{32} and Thr^{34} are found in a conserved motif (SARS-CoV-1 Val-His-Ala-Thr-Ala-Thr-Ile-Pro and SARS-CoV-2 Val-Arg-Ala-Thr-Ala-Thr-Ile-Pro) where only one conserved substitution Arg^{30}His was found.

The ORF3a area from 125 to 200 is required for binding to the 5'-UTR of viral genomic RNA.[147] Although residues 125 - 200 is a well conserved region, several mutations were observed in this area. To address the question if any of these mutations may have an effect on ORF3a RNA binding, we paid attention to know nucleic acid binding residues arginine, lysine and histidine. Thus, one conservative Lys^{134}Arg and non-conservative His^{152}Asn, Lys^{179}Ile, Lys^{181}Glu, Arg^{193}Trp and His^{194}Glu were found. However, close to these mutations were found also others in their proximity that could replace their function. For instance, close to His^{152}Asn there is a conserved histidine at position 150, next to Lys^{179}Ile and Lys^{181}Glu there is Glu^{182}His, and next to Arg^{193}Trp and His^{194}Glu a Asp^{192}Lys substitution was observed. These new substitutions suggest that ORF3a could retain its RNA binding properties. However, this hypothesis requires empirical validation. Finally, the PBM was also found in SARS-CoV-2 ORF3a with identical sequence to SARS-CoV-1.

5.17. Envelope protein

The envelope (E) protein is one of the four structural proteins of coronaviruses. The other structural proteins are the membrane (M), nucleocapsid (N) and spike (S) proteins. The E protein is a type 1 transmembrane protein that is able to form pentamers by associating with other E proteins.[148] This pentamer forms a membrane pore that is able to transport ions.[148] The pore structure is called viroporin and it is present in many common viruses. The ion channel activity can be inhibited in SARS-CoV-1 E protein by the drug HMA .[148] Only a few E proteins are present in each viral particle, but it is highly expressed in the host cells.[149,150] The E protein has been proposed to initiate membrane curvature together with the M protein via the interaction between their C-terminal domains, but this mechanism remains largely unknown.[151,152] The M protein alone seems to not trigger a proper membrane curvature for virion production. The M protein alone can produce virions, but the absence of the E protein cripples virion production, morphology, and plaque shape.[150,152–154] The E protein localizes mainly in the ER and Golgi apparatus, where it participates in assembly, budding, and intracellular trafficking of newly formed virions.[150,152]

The E protein has been shown to interact with other viral proteins. Tandem affinity purification assays established interaction between the S and the E proteins, but the mechanism of how this happens was not pursued further.[155] This study also shows that the E protein interacts with the nsp3 and they suggest that nsp3 mediates E protein ubiquitination. Besides, the E protein co-immunoprecipitates with the N protein, but the function of this interaction remains unclear.[156] Furthermore, a yeast two-hybrid system and an *in vitro* pull down assay showed interaction between E proteins and ORF7a, but its importance is yet to be identified.[157,158] On the other hand, the E protein PDZ-Binding Motif (PBM) interacts with PDZ domains of host proteins. For instance, the interaction of the E protein and a protein associated with *Caenorhabditis elegans,* lin-7 protein 1 (PALS1), a PDZ-containing protein, disrupts tight junctions in the lungs to reach the alveolar wall and develops into a systemic infection.[159] Further, the interaction with syntenin caused it to concentrate in the cytoplasm, triggering an overexpression of inflammatory cytokines, which activates an exaggerated immune response, resulting in lung tissue damage, edema accumulation, and leading to acute respiratory distress syndrome.[160] Interaction between the E protein and the Bcl-xL protein caused lymphopenia.[161]

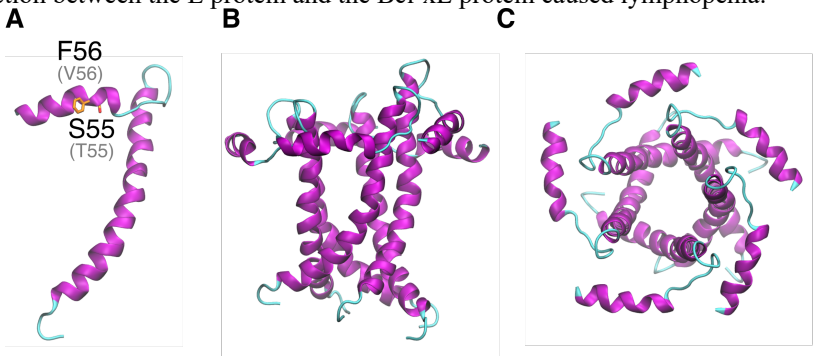


Figure S21. Cartoon representation of SARS-CoV-2 E protein. E protein monomer and substitutions in the C-terminal relative to SARS-CoV-1s (A). E protein pentamer front (B) and top views (C).

Structural analysis and comparison with SARS-CoV-1 envelope protein - Several functionally important residues in E protein have been identified in SARS-CoV-1. For instance, there is a conserved proline residue in the C-terminal in the β-coil-β motif, that if mutated changes localization of the E protein from the Golgi complex to the plasma membrane.[148,162] Besides, the mutations Asn^{15}Ala and Val^{25}Phe inhibit ion channel activity of SARS-CoV-1 E protein. After several passages through cell cultures, this function is restored by the addition of new mutations, suggesting that the E protein confers a selective advantage to the virus.[163] It is also important to note that the C-terminal of the E protein interacts with the C-terminal of the M protein in the cytoplasmic side, and this is required to virus envelope formation.[152] As being highly conserved relative to SARS-CoV-1 (96% identical), this information is likely applicable to SARS-CoV-2 E protein.

SARS-CoV-2 E protein has a predicted short 11 aa N-terminal tail, a 25 aa transmembrane region and a 37 aa C-terminal cytoplasmic region, including PBM (DLLV). Four variations relative to SARS-CoV-1 E protein were verified, all located in the C-terminal end (figure S21A), including in one conservative substitution, two non-conservative substitutions (Val^{56}Phe, Glu^{69}Arg), and a deletion. Similar to SARS-CoV-1, SARS-CoV-2 E protein is likely to assemble as a pentamer to form a viroporin (figure S21B-C), and the three C-terminal substitutions are likely exposed in the pentamer to the cytoplasm, thereby they could be involved in modulating E protein interaction with other proteins.

5.18. Membrane protein

Membrane (M) proteins are the most abundant of all coronavirus structural proteins, and their presence and conformation settle the virion shape.[164,165] M protein has been shown to interact with viral proteins N, S, ORF3a and ORF7a.[133,158,166,167] The general structure of M proteins are that they contain a short N-terminal ectodomain, three transmembrane domains (TMD), and a C-terminal endodomain.[168] M proteins may work as homodimers that can take compact and long conformations.[169] The long conformation is associated with high S protein density and efficient virus budding, while compact conformation is associated with patchy and low S protein presence and inefficient budding.[169] Only M and E proteins are required to form virus-like particles (VLP), but absence of S proteins makes VLPs appear larger, suggesting they may contribute to virus formation.[169] SARS-CoV-1 M protein has been shown to interact with N protein in mammalian two-hybrid assays.[170] Only RNAs interacting with M protein can be found in viral particles in mouse hepatitis virus (MHV).[164]

The M protein has been demonstrated to be a pathogenic factor in coronaviruses. For instance, overexpression of M protein *in vitro* and *in vivo* induces apoptosis in human cell culture and *Drosophila*, respectively.[171] Further, SARS-CoV-1 M protein was found to interact with IKKβ, impairing NF-κB signaling and reducing the expression of cyclooxygenase-2 (COX-2).[172] COX-2 is responsible for the production of prostaglandins, which in turn trigger inflammatory responses. Furthermore, interferon (INF) are a group of antiviral proteins produced by cells in response to a viral infection. M protein impairs the production of type-I INF by interfering with TRAF3-TANK-TBK1/IKKε complex formation, thus preventing IRF3 phosphorylation.[173] Nonetheless, antigenic regions of M protein are detected and production of INF is triggered.[174] These antigenic properties were observed in patients after a year of recovery from SARS-CoV-1 infection, showing persistent IFNγ release from CD4+ and CD8+ lymphocytes when cocultured with M peptides.[175] Thus, SARS-CoV-1 infection triggered the production of memory cells and M protein could be an interesting vaccine candidate.

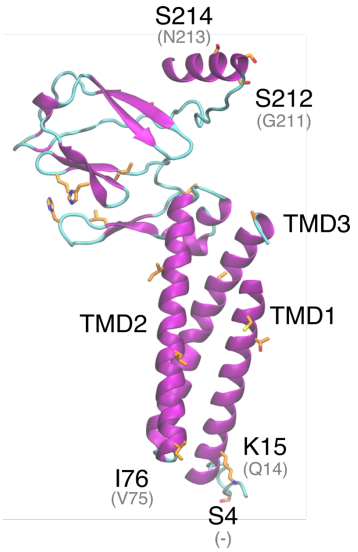


Figure S22. Ab initio predicted structure of the SARS-CoV- 2 M protein. Conserved predicted regions and non-conservative substitutions (*black*) relative to SARS-CoV-1 (*grey*) are depicted in *orange*.

Structural analysis and comparison with SARS-CoV-1 membrane protein - Several key functional residues have been identified for the M protein. Transmissible gastroenteritis virus (TEGV) showed reduced IFN induction when mutations occur in the N-terminal residues 6 - 22.[174] Alteration of the Asn-Ser-Thr motif in this region prevented glycosylation of the M protein, suggesting its importance in INF induction.[174] In another study, after several subculture passages in immortalized proximal tubular epithelial cells (PTEC), the SARS-CoV-1 M protein generated a glutamic acid to alanine point mutation at position 11.[176] This mutation enhances virus replication and persistence in PTEC.[176] Tseng *et al*.[177] identified other key residues in the SARS-CoV-1 M protein with different functions in replication. The Leu-Leu motif (Leu^{218} and Leu^{219}) is required to incorporate the N protein into VLPs. Substitution Cys^{158}Ser reduces secretion of the M protein, but not Cys^{63}Ser or Cys^{85}Ser, and Cys^{158} is involved in the interaction with the N protein. Non conservative substitutions in the motif Ser-Trp-Trp-Ser-Phe-Asn-Pro-Glu reduced the production of VLPs. Conversely, substitutions of the aromatic residues in this motif by analogs did not affect the VLPs production, suggesting that the conservation of these aromatic amino acids is critical for M protein function.

The SARS-CoV-2 M protein is predicted to have three transmembrane domains, a cytoplasmic and a non-cytoplasmic domain. This is similar to the SARS-CoV-1 M protein and other coronavirus M proteins.[168,178] InterProScan[179] defined these regions as follows: non-cytoplasmic region (1 - 19), TMD1 (20 - 40), cytoplasmic region (41 - 51), TMD2 (52 - 73), non-cytoplasmic region (74 - 78), TMD3 (79 - 103) and cytoplasmic region (103 - 222). The SARS-CoV and SARS-CoV-2 M proteins are very similar, with 90·54% identity. Among 21 variations, 12 are non-conservative substitutions and one is an insertion. Within the substitutions, $Gln^{15}Lys$ may have an effect in antigenic properties of the M protein, by modifying its ectodomain (figure S22). Besides, Ser^{4} insertion is very close to a predicted N-glycosylation site at the N-terminal, and can constitute an additional glycosylation site. Modifications to this glycosylation site have been associated with induction of INF production.[174] Most of the non-conservative mutations occur in the cytoplasmic domain at the C-terminal domain. These substitutions involve several sites that can potentially undergo post-translational modification ($Ser^{155}His$, $Thr^{189}Gly$, $Asn^{197}Ser$, $Ala^{211}Ser$, $Gly^{212}Ser$ and $Asn^{214}Ser$).

5.19. ORF6

ORF6 is an auxiliary protein in SARS coronaviruses that is not required for virus replication.[131,180] However, it can increase virus replication when expressed in a heterologous system or at low multiplicity of infection.[181] ORF6 localizes in the perinuclear and ER zones, associated with membranes and colocalized with M , S, and N structural proteins.[182,183] ORF6 is known to interact with the viral proteins ORF9b and nsp8.[184,185] Furthermore, ORF6 and nsp8 colocalize in cell culture, suggesting ORF6 may play a role in virus replication.[184]

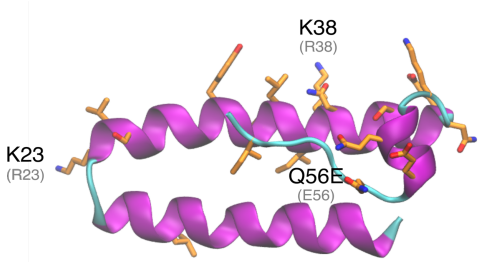


Figure S23. Cartoon representation of SARS-CoV-2 ORF6. Substitutions (*balck*) relative to SARS-CoV-1 (*grey*) are shown in *orange*.

Structural analysis and comparison with SARS-CoV-1 ORF6 - SARS-CoV-1 ORF6 is a small protein of 63 amino acids and 7·53 kDa. Several important residues have been related to SARS-CoV-1 ORF6 structure, function, and location. For instance, ORF6 sequence suggests a membrane association in residues 7 - 37.[183] Even though some residues in this region are charged (Glu^{13}, Arg^{20} and Lys^{23}, in SARS-CoV-2 ORF6), this was enough for membrane association.[183]

ORF6 prevents import of the host signal transducer and activator of transcription 1-alpha/beta (STAT1) to the nucleus by tethering karyopherin subunit alpha-2 (KPNA2) and, therefore, KPNA1 to the ER and golgi apparatus membrane, thereby impairing the activation of INF-induced genes.[186] Tests with specific mutations in the C-terminal region of ORF6 revealed specific residues that affect disruption of KPNA2 ER and Golgi arrest in ORF6 transfected cells.[186] Further, the ORF6 C-terminal was shown to interact with Nmi protein, mediating Nmi ubiquitination and proteasome degradation, thus suppressing INF signaling.[187] ORF6 N-terminal, in turn, was shown to induce membrane rearrangements typically observed in virus-infected cells and to have critical importance to prevent STAT1 translocation to the nucleus.[188]

SARS-CoV-1 and SARS-CoV-2 ORF6 proteins are 68·85% identical. Most of the substitutions are located in the C-terminal helix (figure SX). Lysine substitutions $Arg^{23}Lys$, $Arg^{38}Lys$ and $Asn^{38}Lys$ in SARS-CoV-2 ORF6 suggest the introduction of new putative ubiquitination sites. ORF6 has been shown to be involved in proteasomal degradation of Nmi, but if ORF6 itself is regulated by the proteasome system is unknown.

Two studies pointed to residues 53 - 63 to be important for SARS-CoV-1 ORF6 function.[186,188] In this region, one non-conservative substitution ($Glu^{56}Gln$) and two deletions (Tyr^{62} and Pro^{63}) are verified.

5.20. ORF7a

ORF7a is an accessory protein of coronaviruses and it is not essential for viral replication *in vitro*.[189] ORF7a is a type I transmembrane protein, localized mainly in Golgi apparatus and in the cell surface.[190,191] Besides, ORF7a colocalizes with calnexin (ER marker), showing that it is also localized at ER.[190]

ORF7a is an antagonist of bone marrow stromal antigen 2 (BST-2/CD317/tetherin).[191] BST-2 is a pre-B-cell growth promoter that inhibits virus release by tethering budding virions to the host cell membrane.[192] A greater virion tethering to the cell membrane is observed when ORF7a is not present.[191] ORF7a is usually located in the Golgi apparatus and it relocates to the plasma membrane when BST-2 is expressed, colocalizing it.[191] ORF7a binds to BST-2 and reduces restriction activity of BTS2 by preventing its glycosylation.[191] Further *in vitro* experiments showed that ORF7a is able to induce apoptosis in a caspase dependent manner.[189]

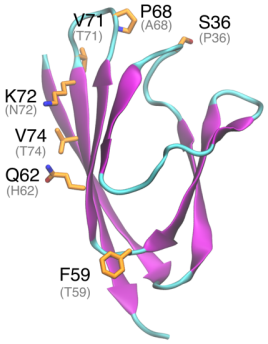


Figure S24. Cartoon representation of SARS-CoV-2 ORF7a luminal domain. Substitutions (*black*) relative to SARS-CoV-1 (*grey*) are shown in *orange*.

Structural analysis and comparison with SARS-CoV-1 ORF7a - SARS-CoV-2 ORF7a protein structure has a 15 amino acid (a.a.) N-terminal signal peptide, a 80-a.a. luminal domain, a 21-a.a. transmembrane domain, and a 5-a.a. cytoplasmic tail.[190] Residues Lys^{117}, Asn^{118} and Lys^{119} form a motif that has been described as being recognized by COPII vesicular system implicated in the transport of proteins from endoplasmic reticulum (ER) to Golgi and they are required to exit the ER.[193]

ORF7a structure is composed of seven antiparallel β-sheets that altogether make a β-sandwich (figure S24). The ORF7a is very conserved between SARS-CoV-1 and SARS-CoV-2 with 85% sequence identity. A total of 18 variations are verified, with 11 non-conservative substitutions and a deletion. Substitutions in the luminal domain are located in the protein surface (figure S24). Therefore, the ORF7a fold is conserved, but these substitutions may affect ORF7a interaction with other viral or host proteins. Mostly conservative mutations are verified at ORF7a C-terminal (83 - 121), except by Ile^{111}Thr.

5.21. ORF7b

ORF7b is a small accessory protein expressed in SARS-CoV-1 and SARS-CoV-2 infected cells with no sequence homology with other viral proteins.[194] It is translated from a bicistronic open reading frame and encoded in the subgenomic RNA 7 and translated by ribosome leaky scanning.[194] Like other accessory proteins, ORF7b is incorporated in viral particles and detected in purified virions.[194] ORF7b is an integral transmembrane protein and it localizes in the cis- and trans- Golgi.[195] Deletion of gene 7 does not affect replication kinetics in vitro, suggesting that ORF7b is not essential for virus replication.[196]

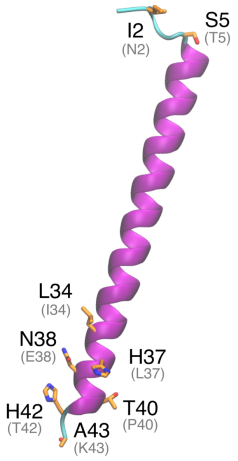


Figure S25. Cartoon representation of SARS-CoV-2 ORF7b protein. Substitutions (*balck*) relative to SARS-CoV-1 (*grey*) are shown in *orange*.

Structural analysis and comparison with SARS-CoV-1 ORF7b - SARS-CoV-1 ORF7b is a small protein of 44 amino acids with a transmembrane domain (TMD) of 22 residues.[195] Mutations in the TMD affect ORF7b cellular localization.[195] Alanine scanning experiments, identified that residues 13-15 and 19-22 are critical for ORF7b retention in the Golgi complex. SARS-CoV-2 ORF7b is 81% identical to SARS-CoV-1 ORF7b. All substitutions are found in the terminals, meaning the TMD is fully conserved (figure S25).

5.22. ORF8

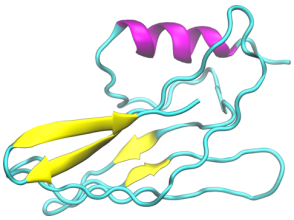


Figure S26. *Ab initio* predicted structure of SARS-CoV-2 ORF8.

SARS-CoV-2 ORF8 is most similar to sequences from Bat-SARS-like coronavirus and a bat coronavirus RaTG13.[197] Remarkably, removal of a N-terminal 29 nucleotide sequence from SARS-CoV-1 ORF8 in early tests showed a decrease in viral replication by up to 23-fold.[198] This sequence was observed to be removed from the virus in human cases later in the outbreak along with additional and different mutations of ORF8.[199,200] The exact purpose of this mutation and ORF8's function in regards to viral benefit within human hosts remains unknown. With the deletion, SARS-CoV-1 encodes ORF8a and ORF8b, while SARS-CoV-2 ORF8 is intact.[201–204] SARS-CoV-2 ORF8 (figure S26) is 127 a.a. long and 45% identical to SARS-CoV-1 ORF8b.

5.23. Nucleocapsid protein

Highly abundant in infected cells, SAR-CoV-1 nucleocapsid (N) packages the viral RNA into a helical ribonucleocapsid and plays a key role in viral assembly.[205] The N protein is assembled and organized in a modular fashion and has been shown to bind to viral RNA at multiple sites in an allosteric manner.[206] The modularity of the N protein is thought to increase selectivity through coupled allosteric binding of individual nucleocapsid domains, aid in regulation and functional expression, and independently evolve binding sites.[206]

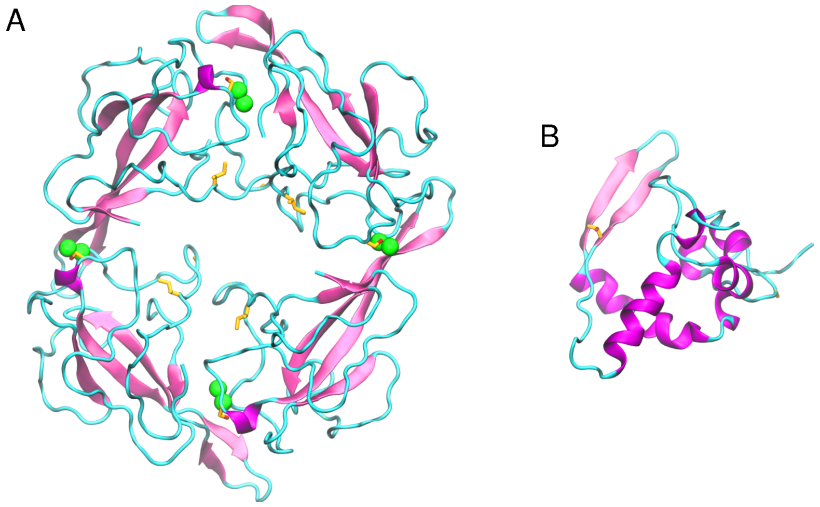


Figure S27. A) Crystal structure of SARS-CoV-2 nucleocapsid RNA-binding domain. B) Local modeling-derived structure of SARS-CoV-2 nucleocapsid C-terminal. Non-conservative substitutions relative to SARS-CoV-1 nucleocapsid are depicted in *orange*.

Structural analysis and comparison with SARS-CoV-1 nucleocapsid - SARS-CoV-2 N protein is a 419 a.a. long protein and is 89% identical to SARS-CoV-1 nucleocapsid. In SARS-CoV-1, a Ser-Arg rich motif (Ser-Ser-Arg-Ser-Ser-Ser-Arg-Ser-Arg-Gly-Asn-Ser-Arg) was found to be important for the oligomerization of N proteins.[167] This motif is mostly conserved in SARS-CoV-2, except by the substitutions Gly^{192}Asn and Asn^{193}Ser. The RNA binding domain of the N protein is located at the N-terminal, within residues 47 and 180 of SARS-CoV-2. The structure of this domain recently solved (PDB 6vyo) (figure S27A). Residues within the N-terminal of the SARS-CoV-1 nucleocapsid, namely, Tyr^{87}, Tyr^{110}, Tyr^{112}, Y^{113}, Leu^{122}, and Ala^{13}, are hypothesized to play a role in ribonucleocapsid packaging.[206] All of these residues are fully conserved in SARS-CoV-2 N. A long intrinsically disordered region follows these residues (181-246), and is predicted to fold upon binding (ANCHOR2 prediction [26]).
The oligomerization domain of nucleocapsids is situated in the C-terminal of SARS-CoV-1 nucleocapsid. Residues Trp^{302}, Ile^{305}, Pro^{310}, Phe^{315}, and Phe^{316} are thought to be associated with highly hydrophobic interactions between two helices within the nucleocapsid[206], and they are fully conserved in SARS-CoV-2 N (figure S27B).

5.24. ORF9

ORF9 is an accessory protein synthesized from an alternative reading frame in the N gene. This accessory protein is integrated in viral particles and therefore it can be considered a structural protein. This incorporation occurs in presence of E and M proteins, suggesting a potential interaction among these proteins.[100] ORF9 localizes in mitochondria outer membrane, cytoplasm, nucleus, endoplasmic reticulum and lipid vesicles.[185,207–209] At mitochondrial level, ORF9 promotes the ubiquitination and degradation of dynamin-like protein (DRP1), thereby causing mitochondria to show an elongated phenotype.[207] Here, ORF9 also interacts with mitochondrial antiviral signaling protein (MAVS) and poly(rC) binding protein 2 (PCBP2).[207] Thus, PCBP2 facilitates the ubiquitination of MAVS by AIP4, suppressing the activation of INF regulatory factors and NF-κB.[207] Besides, interaction and colocalization between ORF9 and ORF6 have been suggested[185], but the relevance of this interaction remains unknown.

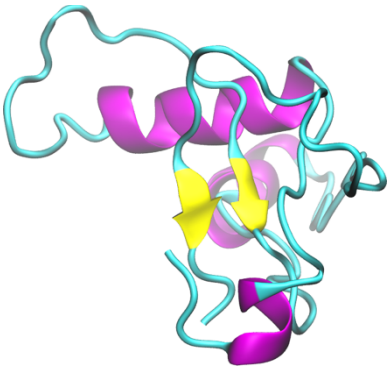


Figure S28. Predicted structure of SARS-CoV-2 ORF9.

Structural analysis and comparison with SARS-CoV-1 ORF9 - SARS-CoV-1 ORF9 is a small protein of 98 residues length. Meier et al.[209] showed that ORF9 forms a symmetric dimer where the monomer interactions resemble a handshake. Dimer assembling creates a hydrophobic tunnel that can accommodate a long fatty acid chain. The authors suggest that ORF9 could anchor itself to a lipidic membrane by internalizing one or more lipidic tails. SARS-CoV-2 ORF9 (figure S28) shares high homology with SARS-CoV-1 (72·45% protein identity).

5.25. ORF10

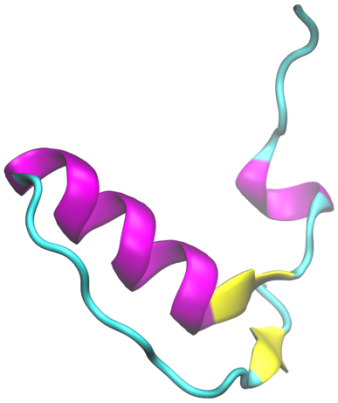


Figure S29. Predicted structure of ORF10.

Within SARS-CoV-2, ORF10 is thought to not have any functional protein purpose as it was not found within the sub genomic mRNA sequencing.[210] It's hypothesized that this protein may act itself or as a precursor of additional RNAs in roles concerning gene expression, controlling cellular antiviral pathways, or within viral replication.[210] Another hypothesis is that this protein, if not an artifact, could be involved with regulating molecular function or contributing

to response stimulus based on DeepGOPlus Gene Ontology.[211] In another study, based on the SARS-CoV-2 interactome, ORF10 is suggested to interact with a Cullin 2 RING E3 ligase complex.[204] Potentially, ORF10 might bind specifically to Cullin 2 ZYG11B complex and hijack this complex for ubiquitination and degradation. It appears to localize only within the extracellular region of the host.[211]

Structural analysis of SARS-CoV-2 ORF10 - ORF10 of SARS-CoV-2 is the last predicted coding sequence upstream of the poly-A tail and is the shortest predicted coding sequence, composed of 38 a.a.[210] ORF10 is predicted to harbor a long helix and a pair of β-strands. ORF10 is not found within the SARS-CoV-1 proteome.[210] It appears to be unique to SARS-CoV-2 (figure S29).

6. Supplementary References

1 Hamming I, Timens W, Bulthuis MLC, Lely AT, Navis GJ, van Goor H. Tissue distribution of ACE2 protein, the functional receptor for SARS coronavirus. A first step in understanding SARS pathogenesis. *J Pathol* 2004; **203**: 631–7.

2 Ren X, Glende J, Al-Falah M, *et al.* Analysis of ACE2 in polarized epithelial cells: surface expression and function as receptor for severe acute respiratory syndrome-associated coronavirus. *J Gen Virol* 2006; **87**: 1691–5.

3 Tseng C-TK, Tseng J, Perrone L, Worthy M, Popov V, Peters CJ. Apical entry and release of severe acute respiratory syndrome-associated coronavirus in polarized Calu-3 lung epithelial cells. *J Virol* 2005; **79**: 9470–9.

4 Sims AC, Baric RS, Yount B, Burkett SE, Collins PL, Pickles RJ. Severe acute respiratory syndrome coronavirus infection of human ciliated airway epithelia: role of ciliated cells in viral spread in the conducting airways of the lungs. *J Virol* 2005; **79**: 15511–24.

5 Zhang H, Kang Z, Gong H, *et al.* The digestive system is a potential route of 2019-nCov infection: a bioinformatics analysis based on single-cell transcriptomes. *BioRxiv* 2020.

6 Zhao Y, Zhao Z, Wang Y, Zhou Y, Ma Y, Zuo W. Single-cell RNA expression profiling of ACE2, the putative receptor of Wuhan 2019-nCov. *BioRxiv* 2020.

7 Harmer D, Gilbert M, Borman R, Clark KL. Quantitative mRNA expression profiling of ACE 2, a novel homologue of angiotensin converting enzyme. *FEBS Letters* 2002; **532**: 107–10.

8 Donoghue M, Hsieh F, Baronas E, *et al.* A Novel Angiotensin-Converting Enzyme–Related Carboxypeptidase (ACE2) Converts Angiotensin I to Angiotensin 1-9. *Circulation Research* 2000; **87**. e1–9.

9 Jia HP, Look DC, Shi L, *et al.* ACE2 receptor expression and severe acute respiratory syndrome coronavirus infection depend on differentiation of human airway epithelia. *J Virol* 2005; **79**: 14614–21.

10 Li W, Moore MJ, Vasilieva N, *et al.* Angiotensin-converting enzyme 2 is a functional receptor for the SARS coronavirus. *Nature* 2003; **426**: 450–4.

11 Nie Y, Wang P, Shi X, *et al.* Highly infectious SARS-CoV pseudotyped virus reveals the cell tropism and its correlation with receptor expression. *Biochem Biophys Res Commun* 2004; **321**: 994–1000.

12 Letko M, Marzi A, Munster V. Functional assessment of cell entry and receptor usage for SARS-CoV-2 and other lineage B betacoronaviruses. *Nat Microbiol* 2020; published online Feb 24.

13 Xu G, Chen J, Pan Q, *et al.* Long noncoding RNA expression profiles of lung adenocarcinoma ascertained by microarray analysis. *PLoS One* 2014; **9**: e104044.

14 Wang D, Fang L, Shi Y, *et al.* Porcine Epidemic Diarrhea Virus 3C-Like Protease Regulates Its Interferon Antagonism by Cleaving NEMO. *J Virol* 2016; **90**: 2090–101.

15 Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* 2008; **5**: 621–8.

16 Yan R, Zhang Y, Li Y, Xia L, Guo Y, Zhou Q. Structural basis for the recognition of the SARS-CoV-2 by full-length human ACE2. *Science*. 2020; **367**:1444–8.

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

18 Lindahl, Abraham, Hess, Spoel V der. GROMACS 2020 Source code. 2020.

19 Huang J, MacKerell AD Jr. CHARMM36 all-atom additive protein force field: validation based on comparison to NMR data. *J Comput Chem* 2013; **34**: 2135–45.

20 Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML. Comparison of simple potential functions for simulating liquid water. *J Chem Phys* 1983; **79**: 926–35.

21 Guvench O, Mallajosyula SS, Raman EP, *et al.* CHARMM additive all-atom force field for carbohydrate derivatives and its utility in polysaccharide and carbohydrate-protein modeling. *J Chem Theory Comput* 2011; **7**: 3162–80.

22 Bussi G, Donadio D, Parrinello M. Canonical sampling through velocity rescaling. *J Chem Phys* 2007; **126**: 014101.

23 Berendsen HJC, Postma JPM, van Gunsteren WF, DiNola A, Haak JR. Molecular dynamics with coupling to an external bath. *J Chem Phys* 1984; **81**: 3684–90.

24 Humphrey W, Dalke A, Schulten K. VMD: visual molecular dynamics. *J Mol Graph* 1996; **14**: 33–8, 27–8.

25 Finn RD, Bateman A, Clements J, *et al.* Pfam: the protein families database. *Nucleic Acids Res* 2014; **42**: D222–30.

26 Mészáros B, Erdos G, Dosztányi Z. IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding. *Nucleic Acids Res* 2018; **46**: W329–37.

27 Krogh A, Larsson B, von Heijne G, Sonnhammer EL. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol* 2001; **305**: 567–80.

28 Huang P-S, Ban Y-EA, Richter F, *et al.* RosettaRemodel: a generalized framework for flexible backbone protein design. *PLoS One* 2011; **6**: e24109.

29 Hu X, Wang H, Ke H, Kuhlman B. High-resolution design of a protein loop. *Proc Natl Acad Sci U S A* 2007; **104**: 17668–73.

30 Kuhlman B, Baker D. Native protein sequences are close to optimal for their structures. *Proc Natl Acad Sci U S A* 2000; **97**: 10383–8.

31 Yang J, Anishchenko I, Park H, Peng Z, Ovchinnikov S, Baker D. Improved protein structure prediction using predicted interresidue orientations. *Proc Natl Acad Sci U S A* 2020; **117**: 1496–503.

32 Trott O, Olson AJ. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J Comput Chem* 2010; **31**: 455–61.

33 Morris GM, Huey R, Lindstrom W, *et al.* AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *J Comput Chem* 2009; **30**: 2785–91.

34 Wang Z, Sun H, Yao X, *et al.* Comprehensive evaluation of ten docking programs on a diverse set of protein-ligand complexes: the prediction accuracy of sampling power and scoring power. *Phys Chem Chem Phys* 2016; **18**: 12964–75.

35 Tang X, Wu C, Li X, *et al.* On the origin and continuing evolution of SARS-CoV-2. *National Science Review*. 2020.

36 Wu A, Peng Y, Huang B, *et al.* Genome Composition and Divergence of the Novel Coronavirus (2019-nCoV) Originating in China. *Cell Host Microbe* 2020; **27**: 325–8.

37 Velankar SS, Soultanas P, Dillingham MS, Subramanya HS, Wigley DB. Crystal structures of complexes of PcrA DNA helicase with a DNA substrate indicate an inchworm mechanism. *Cell* 1999; **97**: 75–84.

38 Jia Z, Yan L, Ren Z, *et al.* Delicate structural coordination of the Severe Acute Respiratory Syndrome coronavirus Nsp13 upon ATP hydrolysis. *Nucleic Acids Res* 2019; **47**: 6538–50.

39 Ivanov KA, Ziebuhr J. Human coronavirus 229E nonstructural protein 13: characterization of duplex-unwinding, nucleoside triphosphatase, and RNA 5'-triphosphatase activities. *J Virol* 2004; **78**: 7833–8.

40 Wolf FA, Angerer P, Theis FJ. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol* 2018; **19**: 15.

41 Zheng W, Li Y, Zhang C, Pearce R, Mortuza SM, Zhang Y. Deep-learning contact-map guided protein structure prediction in CASP13. *Proteins* 2019; **87**: 1149–64.

42 Martínez L, Andreani R, Martínez JM. Convergent algorithms for protein structural alignment. *BMC Bioinformatics* 2007; **8**: 306.

43 Yang J, Yan R, Roy A, Xu D, Poisson J, Zhang Y. The I-TASSER Suite: protein structure and function prediction. *Nat Methods* 2015; **12**: 7–8.

44 Narayanan K, Ramirez SI, Lokugamage KG, Makino S. Coronavirus nonstructural protein 1: Common and distinct functions in the regulation of host and viral gene expression. *Virus Research* 2015; **202**: 89–100.

45 Narayanan K, Huang C, Lokugamage K, *et al.* Severe acute respiratory syndrome coronavirus nsp1 suppresses host gene expression, including that of type I interferon, in infected cells. *J Virol* 2008; **82**: 4471–9.

46 Züst R, Cervantes-Barragán L, Kuri T, *et al.* Coronavirus non-structural protein 1 is a major pathogenicity factor: implications for the rational design of coronavirus vaccines. *PLoS Pathog* 2007; **3**: e109.

47 Kamitani W, Huang C, Narayanan K, Lokugamage KG, Makino S. A two-pronged strategy to suppress host protein synthesis by SARS coronavirus Nsp1 protein. *Nature Structural & Molecular Biology* 2009; **16**: 1134–

40.

48 Netherton CL, Wileman T. Virus factories, double membrane vesicles and viroplasm generated in animal cells. *Curr Opin Virol* 2011; **1**: 381–7.

49 Knoops K, Kikkert M, van den Worm SHE, *et al.* SARS-coronavirus replication is supported by a reticulovesicular network of modified endoplasmic reticulum. *PLoS Biol* 2008; **6**: e226.

50 Hagemeijer MC, Verheije MH, Ulasli M, *et al.* Dynamics of coronavirus replication-transcription complexes. *J Virol* 2010; **84**: 2134–49.

51 Graham RL, Sims AC, Brockway SM, Baric RS, Denison MR. The nsp2 replicase proteins of murine hepatitis virus and severe acute respiratory syndrome coronavirus are dispensable for viral replication. *J Virol* 2005; **79**: 13399–411.

52 Cornillez-Ty CT, Liao L, Yates JR, Kuhn P, Buchmeier MJ. Severe Acute Respiratory Syndrome Coronavirus Nonstructural Protein 2 Interacts with a Host Protein Complex Involved in Mitochondrial Biogenesis and Intracellular Signaling. *Journal of Virology* 2009; **83**: 10314–8.

53 Kathiria AS, Butcher LD, Feagins LA, Souza RF, Boland CR, Theiss AL. Prohibitin 1 modulates mitochondrial stress-related autophagy in human colonic epithelial cells. *PLoS One* 2012; **7**: e31231.

54 Hernando-Rodríguez B, Artal-Sanz M. Mitochondrial Quality Control Mechanisms and the PHB (Prohibitin) Complex. *Cells* 2018; **7**: 238.

55 Agrawal T, Gupta GK, Agrawal DK. Vitamin D deficiency decreases the expression of VDR and prohibitin in the lungs of mice with allergic airway inflammation. *Experimental and Molecular Pathology* 2012; **93**: 74–81.

56 von Brunn A, Teepe C, Simpson JC, *et al.* Analysis of intraviral protein-protein interactions of the SARS coronavirus ORFeome. *PLoS One* 2007; **2**: e459.

57 Prentice E, McAuliffe J, Lu X, Subbarao K, Denison MR. Identification and characterization of severe acute respiratory syndrome coronavirus replicase proteins. *J Virol* 2004; **78**: 9977–86.

58 Padhan K, Minakshi R, Towheed MAB, Jameel S. Severe acute respiratory syndrome coronavirus 3a protein activates the mitochondrial death pathway through p38 MAP kinase activation. *J Gen Virol* 2008; **89**: 1960–9.

59 Linding R, Jensen LJ, Diella F, Bork P, Gibson TJ, Russell RB. Protein disorder prediction: implications for structural proteomics. *Structure* 2003; **11**: 1453–9.

60 Saikatendu KS, Joseph JS, Subramanian V, *et al.* Structural Basis of Severe Acute Respiratory Syndrome Coronavirus ADP-Ribose-1″-Phosphate Dephosphorylation by a Conserved Domain of nsP3. *Structure* 2005; **13**: 1665–75.

61 Lei J, Kusov Y, Hilgenfeld R. Nsp3 of coronaviruses: Structures and functions of a large multi-domain protein. *Antiviral Res* 2018; **149**: 58–74.

62 Serrano P, Johnson MA, Almeida MS, *et al.* Nuclear magnetic resonance structure of the N-terminal domain of nonstructural protein 3 from the severe acute respiratory syndrome coronavirus. *J Virol* 2007; **81**: 12049–60.

63 Daczkowski CM, Dzimianski JV, Clasman JR, Goodwin O, Mesecar AD, Pegan SD. Structural Insights into the Interaction of Coronavirus Papain-Like Proteases and Interferon-Stimulated Gene Product 15 from Different Species. *J Mol Biol* 2017; **429**: 1661–83.

64 Sakai Y, Kawachi K, Terada Y, Omori H, Matsuura Y, Kamitani W. Two-amino acids change in the nsp4 of SARS coronavirus abolishes viral replication. *Virology* 2017; **510**: 165–74.

65 Angelini MM, Akhlaghpour M, Neuman BW, Buchmeier MJ. Severe acute respiratory syndrome coronavirus nonstructural proteins 3, 4, and 6 induce double-membrane vesicles. *MBio* 2013; **4**: e00524–13.

66 Hagemeijer MC, Monastyrska I, Griffith J, *et al.* Membrane rearrangements mediated by coronavirus nonstructural proteins 3 and 4. *Virology* 2014; **458-459**: 125–35.

67 Clementz MA, Kanjanahaluethai A, O'Brien TE, Baker SC. Mutation in murine coronavirus replication protein nsp4 alters assembly of double membrane vesicles. *Virology* 2008; **375**: 118–29.

68 Gadlage MJ, Sparks JS, Beachboard DC, *et al.* Murine hepatitis virus nonstructural protein 4 regulates virus-induced membrane modifications and replication complex function. *J Virol* 2010; **84**: 280–90.

69 Oostra M, te Lintelo EG, Deijs M, Verheije MH, Rottier PJM, de Haan CAM. Localization and membrane topology of coronavirus nonstructural protein 4: involvement of the early secretory pathway in replication. *J Virol* 2007; **81**: 12323–36.

70 Sparks JS, Lu X, Denison MR. Genetic analysis of Murine hepatitis virus nsp4 in virus replication. *J Virol* 2007; **81**: 12554–63.

71 Beachboard DC, Lu X, Baker SC, Denison MR. Murine hepatitis virus nsp4 N258T mutants are not temperature-sensitive. *Virology* 2013; **435**: 210–3.

72 Beachboard DC, Anderson-Daniels JM, Denison MR. Mutations across murine hepatitis virus nsp4 alter virus fitness and membrane modifications. *J Virol* 2015; **89**: 2080–9.

73 Perlman S, Netland J. Coronaviruses post-SARS: update on replication and pathogenesis. *Nat Rev Microbiol* 2009; **7**: 439–50.

74 Ziebuhr J, Snijder EJ, Gorbalenya AE. Virus-encoded proteinases and proteolytic processing in the Nidovirales. *J Gen Virol* 2000; **81**: 853–79.

75 Anand K, Ziebuhr J, Wadhwani P, Mesters JR, Hilgenfeld R. Coronavirus main proteinase (3CLpro) structure: basis for design of anti-SARS drugs. *Science* 2003; **300**: 1763–7.

76 Zhang L, Lin D, Kusov Y, *et al.* α-Ketoamides as Broad-Spectrum Inhibitors of Coronavirus and Enterovirus Replication: Structure-Based Design, Synthesis, and Activity Assessment. *J Med Chem* 2020; published online Feb 24.

77 Nukoolkarn V, Lee VS, Malaisree M, Aruksakulwong O, Hannongbua S. Molecular dynamic simulations analysis of ritonavir and lopinavir as SARS-CoV 3CL(pro) inhibitors. *J Theor Biol* 2008; **254**: 861–7.

78 Dayer MR, Taleb-Gassabi S, Dayer MS. Lopinavir; A Potent Drug against Coronavirus Infection: Insight from Molecular Docking Study. *Archives of Clinical Infectious Diseases* 2017; **12**.

79 Donaldson EF, Graham RL, Sims AC, Denison MR, Baric RS. Analysis of Murine Hepatitis Virus Strain A59 Temperature-Sensitive Mutant TS-LA6 Suggests that nsp10 Plays a Critical Role in Polyprotein Processing. *J Virol* 2007; **81**: 7086–98.

80 Stokes HL, Baliji S, Hui CG, Sawicki SG, Baker SC, Siddell SG. A new cistron in the murine hepatitis virus replicase gene. *J Virol* 2010; **84**: 10148–58.

81 Barrila J, Bacha U, Freire E. Long-range cooperative interactions modulate dimerization in SARS 3CLpro. *Biochemistry* 2006; **45**: 14908–16.

82 Zhang L, Lin D, Sun X, *et al.* Crystal structure of SARS-CoV-2 main protease provides a basis for design of improved α-ketoamide inhibitors. *Science* 2020; published online March 20.

83 den Boon JA, Ahlquist P. Organelle-like membrane compartmentalization of positive-strand RNA virus replication factories. *Annu Rev Microbiol* 2010; **64**: 241–56.

84 Oostra M, Hagemeijer MC, van Gent M, *et al.* Topology and Membrane Anchoring of the Coronavirus Replication Complex: Not All Hydrophobic Domains of nsp3 and nsp6 Are Membrane Spanning. *J Virol* 2008; **82**: 12392–405.

85 Baliji S, Cammer SA, Sobral B, Baker SC. Detection of Nonstructural Protein 6 in Murine Coronavirus-Infected Cells and Analysis of the Transmembrane Topology by Using Bioinformatics and Molecular Approaches. *J Virol* 2009; **83**: 6957–62.

86 Cottam EM, Maier HJ, Manifava M, *et al.* Coronavirus nsp6 proteins generate autophagosomes from the endoplasmic reticulum via an omegasome intermediate. *Autophagy* 2011; **7**: 1335–47.

87 Smith EC, Denison MR. Coronaviruses as DNA wannabes: a new model for the regulation of RNA virus replication fidelity. *PLoS Pathog* 2013; **9**: e1003760.

88 Kirchdoerfer RN, Ward AB. Structure of the SARS-CoV nsp12 polymerase bound to nsp7 and nsp8 co-factors. *Nat Commun* 2019; **10**: 2342.

89 Zhai Y, Sun F, Li X, *et al.* Insights into SARS-CoV transcription and replication from the structure of the nsp7–nsp8 hexadecamer. *Nat Struct Mol Biol* 2005; **12**: 980–6.

90 Lehmann KC, Gulyaeva A, Zevenhoven-Dobbe JC, *et al.* Discovery of an essential nucleotidylating activity associated with a newly delineated conserved domain in the RNA polymerase-containing protein of all nidoviruses. *Nucleic Acids Res* 2015; **43**: 8416–34.

91 Subissi L, Posthuma CC, Collet A, *et al.* One severe acute respiratory syndrome coronavirus protein complex integrates processive RNA polymerase and exonuclease activities. *Proc Natl Acad Sci U S A* 2014; **111**: E3900–9.

92 Agostini ML, Andres EL, Sims AC, *et al.* Coronavirus Susceptibility to the Antiviral Remdesivir (GS-5734) Is Mediated by the Viral Polymerase and the Proofreading Exoribonuclease. *mBio* 2018; **9**. e00221–18

93 Gordon CJ, Tchesnokov EP, Feng JY, Porter DP, Gotte M. The antiviral compound remdesivir potently inhibits RNA-dependent RNA polymerase from Middle East respiratory syndrome coronavirus. *J Biol Chem* 2020; published online Feb 24.

94 Sutton G, Fry E, Carter L, *et al.* The nsp9 replicase protein of SARS-coronavirus, structure and functional insights. *Structure* 2004; **12**: 341–53.

95 Egloff M-P, -P. Egloff M, Ferron F, *et al.* The severe acute respiratory syndrome-coronavirus replicative protein nsp9 is a single-stranded RNA-binding subunit unique in the RNA virus world. *Proc Natl Acad Sci U S A* 2004; **101**: 3792–6.

96 Ponnusamy R, Moll R, Weimar T, Mesters JR, Hilgenfeld R. Variable oligomerization modes in coronavirus non-structural protein 9. *J Mol Biol* 2008; **383**: 1081–96.

97 Deming DJ, Graham RL, Denison MR, Baric RS. Processing of Open Reading Frame 1a Replicase Proteins nsp7 to nsp10 in Murine Hepatitis Virus Strain A59 Replication. *J Virol* 2007; **81**: 10280–91.

98 Miknis ZJ, Donaldson EF, Umland TC, Rimmer RA, Baric RS, Schultz LW. Severe acute respiratory syndrome coronavirus nsp9 dimerization is essential for efficient viral growth. *J Virol* 2009; **83**: 3007–18.

99 Zeng Z, Deng F, Shi K, *et al.* Dimerization of Coronavirus nsp9 with Diverse Modes Enhances Its Nucleic Acid Binding Affinity. *J Virol* 2018; **92**.

100 Bouvet M, Debarnot C, Imbert I, *et al.* In vitro reconstitution of SARS-coronavirus mRNA cap methylation. *PLoS Pathog* 2010; **6**: e1000863.

101 Wang Y, Sun Y, Wu A, *et al.* Coronavirus nsp10/nsp16 Methyltransferase Can Be Targeted by nsp10-Derived Peptide In Vitro and In Vivo To Reduce Replication and Pathogenesis. *J Virol* 2015; **89**: 8416–27.

102 Bouvet M, Lugari A, Posthuma CC, *et al.* Coronavirus Nsp10, a critical co-factor for activation of multiple replicative enzymes. *J Biol Chem* 2014; **289**: 25783–96.

103 Sawicki SG, Sawicki DL, Younker D, *et al.* Correction: Functional and Genetic Analysis of Coronavirus Replicase-Transcriptase Proteins. *PLoS Pathog* 2006; **2**: e17.

104 Sawicki SG, Sawicki DL, Younker D, *et al.* Functional and genetic analysis of coronavirus replicase-transcriptase proteins. *PLoS Pathog* 2005; **1**: e39.

105 Decroly E, Debarnot C, Ferron F, *et al.* Crystal structure and functional analysis of the SARS-coronavirus RNA cap 2'-O-methyltransferase nsp10/nsp16 complex. *PLoS Pathog* 2011; **7**: e1002059.

106 Ma Y, Wu L, Shaw N, *et al.* Structural basis and functional analysis of the SARS coronavirus nsp14-nsp10 complex. *Proc Natl Acad Sci U S A* 2015; **112**: 9436–41.

107 Bouvet M, Imbert I, Subissi L, Gluais L, Canard B, Decroly E. RNA 3'-end mismatch excision by the severe acute respiratory syndrome coronavirus nonstructural protein nsp10/nsp14 exoribonuclease complex. *Proc Natl Acad Sci U S A* 2012; **109**: 9372–7.

108 Su D, Lou Z, Sun F, *et al.* Dodecamer structure of severe acute respiratory syndrome coronavirus nonstructural protein nsp10. *J Virol* 2006; **80**: 7902–8.

109 Adedeji AO, Marchand B, Te Velthuis AJW, *et al.* Mechanism of nucleic acid unwinding by SARS-CoV helicase. *PLoS One* 2012; **7**: e36521.

110 Seybert A, Posthuma CC, van Dinten LC, Snijder EJ, Gorbalenya AE, Ziebuhr J. A complex zinc finger controls the enzymatic activities of nidovirus helicases. *J Virol* 2005; **79**: 696–704.

111 Fang S, Chen B, Tay FPL, Ng BS, Liu DX. An arginine-to-proline mutation in a domain with undefined functions within the helicase protein (Nsp13) is lethal to the coronavirus infectious bronchitis virus in cultured cells. *Virology*. 2007; **358**: 136–47.

112 Wang Z, Huang J-D, Wong K-L, *et al.* On the mechanisms of bananin activity against severe acute respiratory syndrome coronavirus. *FEBS J* 2011; **278**: 383–9.

113 Briguglio I, Piras S, Corona P, Carta A. Inhibition of RNA Helicases of ssRNA Virus Belonging to Flaviviridae, Coronaviridae and Picornaviridae Families. *Int J Med Chem* 2011; **2011**: 1–22.

114 Adedeji AO, Singh K, Kassim A, *et al.* Evaluation of SSYA10-001 as a replication inhibitor of severe acute respiratory syndrome, mouse hepatitis, and Middle East respiratory syndrome coronaviruses. *Antimicrob Agents*

Chemother 2014; **58**: 4894–8.

115 Eckerle LD, Lu X, Sperry SM, Choi L, Denison MR. High fidelity of murine hepatitis virus replication is decreased in nsp14 exoribonuclease mutants. *J Virol* 2007; **81**: 12135–44.

116 Eckerle LD, Becker MM, Halpin RA, *et al.* Infidelity of SARS-CoV Nsp14-exonuclease mutant virus replication is revealed by complete genome sequencing. *PLoS Pathog* 2010; **6**: e1000896.

117 Denison MR, Graham RL, Donaldson EF, Eckerle LD, Baric RS. Coronaviruses: an RNA proofreading machine regulates replication fidelity and diversity. *RNA Biol* 2011; **8**: 270–9.

118 Smith EC, Case JB, Blanc H, *et al.* Mutations in coronavirus nonstructural protein 10 decrease virus replication fidelity. *J Virol* 2015; **89**: 6418–26.

119 Zuo Y, Deutscher MP. Exoribonuclease superfamilies: structural analysis and phylogenetic distribution. *Nucleic Acids Res* 2001; **29**: 1017–26.

120 Kim Y, Jedrzejczak R, Maltseva NI, *et al.* Crystal structure of Nsp15 endoribonuclease NendoU from SARS-CoV-2. *bioRxiv* 2020.

121 Johnson MA, Chatterjee A, Neuman BW, Wüthrich K. SARS Coronavirus Unique Domain: Three-Domain Molecular Architecture in Solution and RNA Binding. *Mol Biol* 2010; **400**: 724–42.

122 Perlman S. Faculty of 1000 evaluation for Ribose 2'-O-methylation provides a molecular signature for the distinction of self and non-self mRNA dependent on the RNA sensor Mda5. F1000 - Post-publication peer review of the biomedical literature 2011.

123 Ulferts R, Ziebuhr J. Nidovirus ribonucleases: Structures and functions in viral replication. *RNA Biol* 2011; **8**: 295–304.

124 Deng X, Hackbart M, Mettelman RC, *et al.* Coronavirus nonstructural protein 15 mediates evasion of dsRNA sensors and limits apoptosis in macrophages. *Proc Natl Acad Sci U S A* 2017; **114**: E4251–60.

125 Liu X, Fang P, Fang L, *et al.* Porcine deltacoronavirus nsp15 antagonizes interferon-β production independently of its endoribonuclease activity. *Molecular Immunology* 2019; **114**: 100–7.

126 Chen Y, Su C, Ke M, *et al.* Biochemical and structural insights into the mechanisms of SARS coronavirus RNA ribose 2'-O-methylation by nsp16/nsp10 protein complex. *PLoS Pathog* 2011; **7**: e1002294.

127 Shen S, Lin P-S, Chao Y-C, *et al.* The severe acute respiratory syndrome coronavirus 3a is a novel structural protein. *Biochem Biophys Res Commun* 2005; **330**: 286–92.

128 Yuan X, Li J, Shan Y, *et al.* Subcellular localization and membrane association of SARS-CoV 3a protein. *Virus Res.* 2005; **109**: 191–202.

129 Minakshi R, Padhan K. The YXXΦ motif within the severe acute respiratory syndrome coronavirus (SARS-CoV) 3a protein is crucial for its intracellular transport. *Virol J* 2014; **11**: 75.

130 Lu W, Zheng B-J, Xu K, *et al.* Severe acute respiratory syndrome-associated coronavirus 3a protein forms an ion channel and modulates virus release. *Proc Natl Acad Sci U S A* 2006; **103**: 12540–5.

131 Yount B, Roberts RS, Sims AC, *et al.* Severe acute respiratory syndrome coronavirus group-specific open reading frames encode nonessential functions for replication in cell cultures and mice. *J Virol* 2005; **79**: 14909–22.

132 Castaño-Rodriguez C, Honrubia JM, Gutiérrez-Álvarez J, *et al.* Role of Severe Acute Respiratory Syndrome Coronavirus Viroporins E, 3a, and 8a in Replication and Pathogenesis. *mBio* 2018; **9**. e02325–17

133 Tan Y-J, Teng E, Shen S, *et al.* A novel severe acute respiratory syndrome coronavirus protein, U274, is transported to the cell surface and undergoes endocytosis. *J Virol* 2004; **78**: 6723–34.

134 Zeng R, Yang R-F, Shi M-D, *et al.* Characterization of the 3a protein of SARS-associated coronavirus in infected vero E6 cells and SARS patients. *J Mol Biol* 2004; **341**: 271–9.

135 Kanzawa N, Nishigaki K, Hayashi T, *et al.* Augmentation of chemokine production by severe acute respiratory syndrome coronavirus 3a/X1 and 7a/X4 proteins through NF-κB activation. *FEBS Letters* 2006; **580**: 6807–12.

136 Siu K-L, Yuen K-S, Castaño-Rodriguez C, *et al.* Severe acute respiratory syndrome coronavirus ORF3a protein activates the NLRP3 inflammasome by promoting TRAF3-dependent ubiquitination of ASC. *FASEB J* 2019; **33**: 8865–77.

137 Minakshi R, Padhan K, Rani M, Khan N, Ahmad F, Jameel S. The SARS Coronavirus 3a protein causes endoplasmic reticulum stress and induces ligand-independent downregulation of the type 1 interferon receptor.

PLoS One 2009; **4**: e8342.

138 Yuan X, Yao Z, Wu J, *et al.* G1 phase cell cycle arrest induced by SARS-CoV 3a protein via the cyclin D3/pRb pathway. *Am J Respir Cell Mol Biol* 2007; **37**: 9–19.

139 Chan C-M, Tsoi H, Chan W-M, *et al.* The ion channel activity of the SARS-coronavirus 3a protein is linked to its pro-apoptotic function. *Int J Biochem Cell Biol* 2009; **41**: 2232–9.

140 Obitsu S, Ahmed N, Nishitsuji H, *et al.* Potential enhancement of osteoclastogenesis by severe acute respiratory syndrome coronavirus 3a/X1 protein. *Archives of Virology* 2009; **154**: 1457–64.

141 Ng KHL. Pulmonary artery thrombosis in a patient with severe acute respiratory syndrome. *Postgraduate Medical Journa*. 2005; **81**: e3–e3.

142 Tan Y-J, Tham P-Y, Chan DZL, *et al.* The severe acute respiratory syndrome coronavirus 3a protein up-regulates expression of fibrinogen in lung epithelial cells. *J Virol* 2005; **79**: 10083–7.

143 Minakshi R, Padhan K, Rehman S, Hassan MI, Ahmad F. The SARS Coronavirus 3a protein binds calcium in its cytoplasmic domain. *Virus Res* 2014; **191**: 180–3.

144 Rota PA, Oberste MS, Monroe SS, *et al.* Characterization of a novel coronavirus associated with severe acute respiratory syndrome. *Science* 2003; **300**: 1394–9.

145 Padhan K, Tanwar C, Hussain A, *et al.* Severe acute respiratory syndrome coronavirus Orf3a protein interacts with caveolin. *J Gen Virol* 2007; **88**: 3067–77.

146 Oostra M, de Haan CAM, de Groot RJ, Rottier PJM. Glycosylation of the Severe Acute Respiratory Syndrome Coronavirus Triple-Spanning Membrane Proteins 3a and M. Journal of Virology. 2006; **80**: 2326–36.

147 Sharma K, Surjit M, Satija N, Liu B, Chow VTK, Lal SK. The 3a accessory protein of SARS coronavirus specifically interacts with the 5'UTR of its genomic RNA, Using a unique 75 amino acid interaction domain. *Biochemistry* 2007; **46**: 6488–99.

148 Li Y, Surya W, Claudine S, Torres J. Structure of a conserved Golgi complex-targeting signal in coronavirus envelope proteins. *J Biol Chem* 2014; **289**: 12535–49.

149 Nieto-Torres JL, Dediego ML, Alvarez E, *et al.* Subcellular location and topology of severe acute respiratory syndrome coronavirus envelope protein. *Virology* 2011; **415**: 69–82.

150 Vennema H, Godeke GJ, Rossen JW, *et al.* Nucleocapsid-independent assembly of coronavirus-like particles by co-expression of viral envelope protein genes. *EMBO J* 1996; **15**: 2020–8.

151 Schoeman D, Fielding BC. Coronavirus envelope protein: current knowledge. *Virol J* 2019; **16**: 69.

152 Lim KP, Liu DX. The missing link in coronavirus assembly. Retention of the avian coronavirus infectious bronchitis virus envelope protein in the pre-Golgi compartments and physical interaction between the envelope and membrane proteins. *J Biol Chem* 2001; **276**: 17515–23.

153 Fischer F, Stegen CF, Masters PS, Samsonoff WA. Analysis of Constructed E Gene Mutants of Mouse Hepatitis Virus Confirms a Pivotal Role for E Protein in Coronavirus Assembly. Journal of Virology. 1998; **72**: 7885–94.

154 Kuo L, Masters PS. The Small Envelope Protein E Is Not Essential for Murine Coronavirus Replication. Journal of Virology. 2003; **77**: 4597–608.

155 Alvarez E, DeDiego ML, Nieto-Torres JL, Jiménez-Guardeño JM, Marcos-Villar L, Enjuanes L. The envelope protein of severe acute respiratory syndrome coronavirus interacts with the non-structural protein 3 and is ubiquitinated. *Virology* 2010; **402**: 281–91.

156 Maeda J, Maeda A, Makino S. Release of Coronavirus E Protein in Membrane Vesicles from Virus-Infected Cells and E Protein-Expressing Cells. *Virology* 1999; **263**: 265–72.

157 Pan J 'an, Peng X, Gao Y, *et al.* Genome-wide analysis of protein-protein interactions and involvement of viral proteins in SARS-CoV replication. *PLoS One* 2008; **3**: e3299.

158 Fielding BC, Gunalan V, Tan THP, *et al.* Severe acute respiratory syndrome coronavirus protein 7a interacts with hSGT. *Biochem Biophys Res Commun* 2006; **343**: 1201–8.

159 Teoh K-T, Siu Y-L, Chan W-L, *et al.* The SARS Coronavirus E Protein Interacts with PALS1 and Alters Tight Junction Formation and Epithelial Morphogenesis. *Mol Biol Cell* 2010; **21**: 3838–52.

160 Jimenez-Guardeño JM, Nieto-Torres JL, DeDiego ML, *et al.* The PDZ-binding motif of severe acute respiratory syndrome coronavirus envelope protein is a determinant of viral pathogenesis. *PLoS Pathog* 2014; **10**: e1004320.

161 Yang Y, Xiong Z, Zhang S, *et al.* Bcl-xL inhibits T-cell apoptosis induced by expression of SARS coronavirus E protein in the absence of growth factors. *Biochemical Journal* 2005; **392**: 135–43.

162 Cohen JR, Lin LD, Machamer CE. Identification of a Golgi complex-targeting signal in the cytoplasmic tail of the severe acute respiratory syndrome coronavirus envelope protein. *J Virol* 2011; **85**: 5794–803.

163 Nieto-Torres JL, DeDiego ML, Verdiá-Báguena C, *et al.* Severe acute respiratory syndrome coronavirus envelope protein ion channel activity promotes virus fitness and pathogenesis. *PLoS Pathog* 2014; **10**: e1004077.

164 Narayanan K, Makino S. Cooperation of an RNA Packaging Signal and a Viral Envelope Protein in Coronavirus RNA Packaging. *J Virol* 2001; **75**: 9059–67.

165 Kuo L, Masters PS. Genetic evidence for a structural interaction between the carboxy termini of the membrane and nucleocapsid proteins of mouse hepatitis virus. *J Virol* 2002; **76**: 4987–99.

166 Huang Y, Yang Z-Y, Kong W-P, Nabel GJ. Generation of synthetic severe acute respiratory syndrome coronavirus pseudoparticles: implications for assembly and vaccine production. *J Virol* 2004; **78**: 12557–65.

167 He R, Dobie F, Ballantine M, *et al.* Analysis of multimerization of the SARS coronavirus nucleocapsid protein. *Biochem Biophys Res Commun* 2004; **316**: 476–83.

168 Armstrong J, Niemann H, Smeekens S, Rottier P, Warren G. Sequence and topology of a model intracellular membrane protein, E1 glycoprotein, from a coronavirus. *Nature* 1984; **308**: 751–2.

169 Neuman BW, Kiss G, Kunding AH, *et al.* A structural analysis of M protein in coronavirus assembly and morphology. *J Struct Biol* 2011; **174**: 11–22.

170 He R, Leeson A, Ballantine M, *et al.* Characterization of protein–protein interactions between the nucleocapsid protein and membrane protein of the SARS coronavirus. *Virus Research* 2004; **105**: 121–5.

171 Chan C-M, Ma C-W, Chan W-Y, Chan HYE. The SARS-Coronavirus Membrane protein induces apoptosis through modulating the Akt survival pathway. *Arch Biochem Biophys* 2007; **459**: 197–207.

172 Fang X, Gao J, Zheng H, *et al.* The membrane protein of SARS-CoV suppresses NF-kappaB activation. *J Med Virol* 2007; **79**: 1431–9.

173 Siu K-L, Kok K-H, Ng M-HJ, *et al.* Severe acute respiratory syndrome coronavirus M protein inhibits type I interferon production by impeding the formation of TRAF3.TANK.TBK1/IKKepsilon complex. *J Biol Chem* 2009; **284**: 16202–9.

174 Laude H, Gelfi J, Lavenant L, Charley B. Single amino acid changes in the viral glycoprotein M affect induction of alpha interferon by the coronavirus transmissible gastroenteritis virus. *J Virol* 1992; **66**: 743–9.

175 Yang L, Peng H, Zhu Z, *et al.* Persistent memory CD4+ and CD8+ T-cell responses in recovered severe acute respiratory syndrome (SARS) patients to SARS coronavirus M antigen. *J Gen Virol* 2007; **88**: 2740–8.

176 Pacciarini F, Ghezzi S, Canducci F, *et al.* Persistent replication of severe acute respiratory syndrome coronavirus in human tubular kidney cells selects for adaptive mutations in the membrane protein. *J Virol* 2008; **82**: 5137–44.

177 Tseng Y-T, Chang C-H, Wang S-M, Huang K-J, Wang C-T. Identifying SARS-CoV membrane protein amino acid residues linked to virus-like particle assembly. *PLoS One* 2013; **8**: e64013.

178 Hu Y, Wen J, Tang L, *et al.* The M Protein of SARS-CoV: Basic Structural and Immunological Properties. *Genomics, Proteomics & Bioinformatics* 2003; **1**: 118–30.

179 Jones P, Binns D, Chang H-Y, *et al.* InterProScan 5: genome-scale protein function classification. *Bioinformatics* 2014; **30**: 1236–40.

180 Huang C, Peters CJ, Makino S. Severe acute respiratory syndrome coronavirus accessory protein 6 is a virion-associated protein and is released from 6 protein-expressing cells. *J Virol* 2007; **81**: 5423–6.

181 Zhao J, Falcón A, Zhou H, *et al.* Severe acute respiratory syndrome coronavirus protein 6 is required for optimal replication. *J Virol* 2009; **83**: 2368–73.

182 Geng H, Liu Y-M, Chan W-S, *et al.* The putative protein 6 of the severe acute respiratory syndrome-associated coronavirus: expression and functional characterization. *FEBS Lett* 2005; **579**: 6763–8.

183 Pewe L, Zhou H, Netland J, *et al.* A severe acute respiratory syndrome-associated coronavirus-specific protein enhances virulence of an attenuated murine coronavirus. *J Virol* 2005; **79**: 11335–42.

184 Kumar P, Gunalan V, Liu B, *et al.* The nonstructural protein 8 (nsp8) of the SARS coronavirus interacts with

its ORF6 accessory protein. *Virology* 2007; **366**: 293–303.

185 Calvo E, DeDiego ML, García P, López JA, Pérez-Breña P, Falcón A. Severe acute respiratory syndrome coronavirus accessory proteins 6 and 9b interact in vivo. *Virus Res* 2012; **169**: 282–8.

186 Frieman M, Yount B, Heise M, Kopecky-Bromberg SA, Palese P, Baric RS. Severe acute respiratory syndrome coronavirus ORF6 antagonizes STAT1 function by sequestering nuclear import factors on the rough endoplasmic reticulum/Golgi membrane. *J Virol* 2007; **81**: 9812–24.

187 Cheng W, Chen S, Li R, Chen Y, Wang M, Guo D. Severe acute respiratory syndrome coronavirus protein 6 mediates ubiquitin-dependent proteosomal degradation of N-Myc (and STAT) interactor. *Virol Sin* 2015; **30**: 153–61.

188 Zhou H, Ferraro D, Zhao J, *et al.* The N-terminal region of severe acute respiratory syndrome coronavirus protein 6 induces membrane rearrangement and enhances virus replication. *J Virol* 2010; **84**: 3542–51.

189 Schaecher SR, Touchette E, Schriewer J, Buller RM, Pekosz A. Severe acute respiratory syndrome coronavirus gene 7 products contribute to virus-induced apoptosis. *J Virol* 2007; **81**: 11054–68.

190 Nelson CA, Pekosz A, Lee CA, Diamond MS, Fremont DH. Structure and intracellular targeting of the SARS-coronavirus Orf7a accessory protein. *Structure* 2005; **13**: 75–85.

191 Taylor JK, Coleman CM, Postel S, *et al.* Severe Acute Respiratory Syndrome Coronavirus ORF7a Inhibits Bone Marrow Stromal Antigen 2 Virion Tethering through a Novel Mechanism of Glycosylation Interference. *J Virol* 2015; **89**: 11820–33.

192 Sauter D, Specht A, Kirchhoff F. Tetherin: holding on and letting go. *Cell* 2010; **141**: 392–8.

193 Bickford LC, Mossessova E, Goldberg J. A structural view of the COPII vesicle coat. *Curr Opin Struct Biol* 2004; **14**: 147–53.

194 Schaecher SR, Mackenzie JM, Pekosz A. The ORF7b protein of severe acute respiratory syndrome coronavirus (SARS-CoV) is expressed in virus-infected cells and incorporated into SARS-CoV particles. *J Virol* 2007; **81**: 718–31.

195 Schaecher SR, Diamond MS, Pekosz A. The transmembrane domain of the severe acute respiratory syndrome coronavirus ORF7b protein is necessary and sufficient for its retention in the Golgi complex. *J Virol* 2008; **82**: 9477–91.

196 Pekosz A, Schaecher SR, Diamond MS, Fremont DH, Sims AC, Baric RS. Structure, expression, and intracellular localization of the SARS-CoV accessory proteins 7a and 7b. *Adv Exp Med Biol* 2006; **581**: 115–20.

197 Hu D, Zhu C, Ai L, *et al.* Genomic characterization and infectivity of a novel SARS-like coronavirus in Chinese bats. *Emerg Microbes Infect* 2018; **7**: 154.

198 Muth D, Corman VM, Roth H, *et al.* Attenuation of replication by a 29 nucleotide deletion in SARS-coronavirus acquired during the early stages of human-to-human transmission. *Sci Rep* 2018; **8**: 15177.

199 Chiu RWK, Chim SSC, Tong Y-K, *et al.* Tracing SARS-coronavirus variant with large genomic deletion. *Emerg Infect Dis* 2005; **11**: 168–70.

200 Chinese SARS Molecular Epidemiology Consortium. Molecular evolution of the SARS coronavirus during the course of the SARS epidemic in China. *Science* 2004; **303**: 1666–9.

201 Guan Y, Zheng BJ, He YQ, *et al.* Isolation and characterization of viruses related to the SARS coronavirus from animals in southern China. *Science* 2003; **302**: 276–8.

202 Chinese SARS Molecular Epidemiology Consortium. Molecular evolution of the SARS coronavirus during the course of the SARS epidemic in China. *Science* 2004; **303**: 1666–9.

203 Lau SKP, Woo PCY, Li KSM, *et al.* Severe acute respiratory syndrome coronavirus-like virus in Chinese horseshoe bats. *Proc Natl Acad Sci U S A* 2005; **102**: 14040–5.

204 Gordon DE, Jang GM, Bouhaddou M, *et al.* A SARS-CoV-2-Human Protein-Protein Interaction Map Reveals Drug Targets and Potential Drug-Repurposing. *bioRxiv* 2020.

205 Chang C-K, Hsu Y-L, Chang Y-H, *et al.* Multiple nucleic acid binding sites and intrinsic disorder of severe acute respiratory syndrome coronavirus nucleocapsid protein: implications for ribonucleocapsid protein packaging. *J Virol* 2009; **83**: 2255–64.

206 Chang C-K, Hou M-H, Chang C-F, Hsiao C-D, Huang T-H. The SARS coronavirus nucleocapsid protein –

Forms and functions. *Antiviral Research*. 2014; **103**: 39–50.

207 Shi C-S, Qi H-Y, Boularan C, *et al.* SARS-coronavirus open reading frame-9b suppresses innate immunity by targeting mitochondria and the MAVS/TRAF3/TRAF6 signalosome. *J Immunol* 2014; **193**: 3080–9.

208 Moshynskyy I, Viswanathan S, Vasilenko N, *et al.* Intracellular localization of the SARS coronavirus protein 9b: evidence of active export from the nucleus. *Virus Res* 2007; **127**: 116–21.

209 Meier C, Aricescu AR, Assenberg R, *et al.* The crystal structure of ORF-9b, a lipid binding protein from the SARS coronavirus. *Structure* 2006; **14**: 1157–65.

210 Taiaroa G, Rawlinson D, Featherstone L, *et al.* Direct RNA sequencing and early evolution of SARS-CoV-2. *bioRxiv* 2020.

211 Alam I, Kamau A, Kulmanov M, *et al.* Functional pangenome analysis suggests inhibition of the protein E as a readily available therapy for COVID-2019. *bioRxiv* 2020.