JBC ARTICLE

SARS-CoV-2 (COVID-19) structural and evolutionary dynamicome: Insights into functional evolution and human genomics

Received for publication, June 22, 2020, and in revised form, June 23, 2020 Published, Papers in Press, June 25, 2020, DOI 10.1074/jbc.RA120.014873

Ruchir Gupta[1,2], Jacob Charron[1,2,3], Cynthia L. Stenger[4], Jared Painter[4], Hunter Steward[1], Taylor W. Cook[5], William Faber[5], Austin Frisch[1], Eric Lind[1], Jacob Bauss[1], Xiaopeng Li[1], Olivia Sirpilla[1,2,6], Xavier Soehnlen[1,2,6], Adam Underwood[6], David Hinds[1,7], Michele Morris[7], Neil Lamb[7], Joseph A. Carcillo[8], Caleb Bupp[9], Bruce D. Uhal[10], Surender Rajasekaran[1,11,12], and Jeremy W. Prokop[1,2,*]

From the [1]Department of Pediatrics and Human Development, College of Human Medicine, Michigan State University, Grand Rapids, Michigan, USA, the Departments of [2]Pharmacology and Toxicology and [10]Physiology, Michigan State University, East Lansing, Michigan, USA, [3]Calvin University, Grand Rapids, Michigan, USA, the [4]Department of Mathematics, University of North Alabama, Florence, Alabama, USA, [5]Grand Rapids Community College, Grand Rapids, Michigan, USA, [6]Walsh University, North Canton, Ohio, USA, the [7]HudsonAlpha Institute for Biotechnology, Huntsville, Alabama, USA, the [8]Department of Critical Care Medicine and Pediatrics, Children's Hospital of Pittsburgh, University of Pittsburgh School of Medicine, Pittsburgh, Pennsylvania, USA, [9]Spectrum Health Medical Genetics, Grand Rapids, Michigan, USA, the [11]Pediatric Intensive Care Unit, Helen DeVos Children's Hospital, Grand Rapids, Michigan, USA, and the [12]Office of Research, Spectrum Health, Grand Rapids, Michigan, USA

Edited by Craig E. Cameron

The pandemic caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has challenged the speed at which laboratories can discover the viral composition and study health outcomes. The small ~30-kb ssRNA genome of coronaviruses makes them adept at cross-species spread while enabling a robust understanding of all of the proteins the viral genome encodes. We have employed protein modeling, molecular dynamics simulations, evolutionary mapping, and 3D printing to gain a full proteome- and dynamicome-level understanding of SARS-CoV-2. We established the Viral Integrated Structural Evolution Dynamic Database (VIStEDD at RRID:SCR_018793) to facilitate future discoveries and educational use. Here, we highlight the use of VIStEDD for nsp6, nucleocapsid (N), and spike (S) surface glycoprotein. For both nsp6 and N, we found highly conserved surface amino acids that likely drive protein–protein interactions. In characterizing viral S protein, we developed a quantitative dynamics cross-correlation matrix to gain insights into its interactions with the angiotensin I–converting enzyme 2 (ACE2)–solute carrier family 6 member 19 (SLC6A19) dimer. Using this quantitative matrix, we elucidated 47 potential functional missense variants from genomic databases within ACE2/SLC6A19/transmembrane serine protease 2 (TMPRSS2), warranting genomic enrichment analyses in SARS-CoV-2 patients. These variants had ultralow frequency but existed in males hemizygous for ACE2. Two ACE2 noncoding variants (rs4646118 and rs143185769) present in ~9% of individuals of African descent may regulate ACE2 expression and may be associated with increased susceptibility of African Americans to SARS-CoV-2. We propose that this SARS-CoV-2 database may aid research into the ongoing pandemic.

The current SARS-CoV-2 outbreak has become a global pandemic. There is an urgent need to understand the proteins coded by SARS-CoV-2 and how they can be targeted for intervention. Coronaviruses belong to the *Orthocoronavirinae* subfamily, which lies under the *Coronaviridae* family. Their 26–33-kb genome consists of positive-sense, single-stranded RNA, coding for nonstructural and structural proteins. To date, seven coronaviruses have been discovered that are capable of human-to-human transmission. Four of these cause the common cold (HKU1, NL63, OC43, and 229E), whereas the other three (MERS-CoV, SARS-CoV, SARS-CoV-2) can cause more severe respiratory illnesses resulting in multisystem organ failure and death (1). SARS-CoV-2 shares 79% genomic similarity with SARS-CoV, linking it to the bat and human SARS-CoV annotation. Both SARS-CoV and SARS-CoV-2 bind to the ACE2 receptor through the spike (S) protein (2, 3). The ubiquitous presence of Coronaviridae in many animals and its relatively small genome makes these an ideal infective agent as it adapts and evolves into a highly effective pathogen. In the case of the SARS-CoV-2 genome, 96.2% of the genome is shared with a bat coronavirus, suggesting a zoonotic origin (4, 5). Initial disease propagation was detected in Wuhan, China, a major transportation hub with over 11 million people. The population density and the heavy traffic into and out of the city made for a large outbreak that spread quickly throughout the world (6, 7). This combination has given rise to a once-in-a-generation pandemic.

Insights can be gathered from SARS-CoV and MERS-CoV regarding SARS-CoV-2 illness severity. Studies in mouse models have led to speculation that SARS-CoV and MERS-CoV infections cause a delayed type I interferon response, which allows for early uncontrolled viral replication. This leads to an influx of neutrophils and monocytes/macrophages, resulting in a hyperproduction of cytokines causing pneumonia, acute respiratory distress syndrome, and global sepsis. In SARS-CoV-2, patients experience similar changes in neutrophils and lymphocytes, indicating that SARS-CoV-2 infection severity may closely depend on a delayed response of the innate immune

This article contains supporting information.

* For correspondence: Jeremy W Prokop, jprokop54@gmail.com.

system (8). Early Chinese data provide remarkable insights into how SARS-CoV-2 drives lethality through a sepsis-driven multiple-organ failure model (9), including early spike in ferritin, cytokine storm, and injury to the cardiac system (10, 11). Understanding of the SARS-CoV-2 genome could provide critical insights into the complex interplay with the host genome driving disease progression in severe risk group patients, allowing for the identification of potential target sites for intervention.

Several *Coronaviridae* proteins have been highly studied as targets of intervention to prevent infection spread. One of these proteins, the N protein, is of particular interest as it interacts with host ribosomal subunits and has been shown to suppress nonsense-mediated decay of viral mRNA by the host cell (12, 13). Enzymes encoded by SARS-CoV-2, such as the 3-chymotrypsin (3C)-like protease, RNA-dependent RNA polymerase, and papain-like protease, are potential targets of drugs (14). Proteins of the virus, including N, nsp9, nsp13, nsp15, ORF3a, and ORF6, have been shown to target innate immune signaling pathways (13). The S protein, surface-expressed, enters cells through ACE2/SLC6A19 membrane receptor contact similarly to SARS-CoV, but with higher binding affinity (2). This interaction has the potential to be therapeutically inhibited through antibody neutralization (15). ACE2 is highly expressed in the heart, small intestine, kidney, thyroid, breast, arterial walls, adipose, and testis, with a lower number of cells in lung, oral/nasal cavity, pancreas, and liver (16). Nasal epithelial cells play a critical role in SARS-CoV-2 (17). In the lung, ACE2 is expressed in the type 2 pneumocytes, putative progenitor cells of alveolar epithelia linked to lung fibrosis pathways (18, 19). The spike-ACE2 complex is further processed by TMPRSS2 for internalization of the virus, which has been postulated to be a target site with protease inhibitors (20). Currently, it is not well-understood how the spike-ACE2 complex is impacted by viral or human variants, suggesting a need for further research.

As we learn more about viral pathogenesis, hopefully the current outbreak will be brought under control and future outbreaks prevented. In this current work, we developed the Viral Integrated Structural Evolution Dynamic Database (VIStEDD) for the SARS-CoV-2 proteome, enriching VIStEDD using evolutionary insights of viruses and human variant mapping for potential functional outcomes.

Results

SARS-CoV-2 dynamicome database

A total of 24 proteins of SARS-CoV-2 (Table 1) were run through our standardized workflow (Fig. 1), consisting of protein structure assessment, setting of protein protonation at pH 7.4, energy minimization in water with NaCl, 20 ns of molecular dynamics simulations, analysis of the movement trajectories, sequence identification from the nonredundant (nr) database, mapping of conservation onto structure/dynamics, and assessment of interactions with known binding partners. We extracted 55,390 sequences homologous to the SARS-CoV-2 proteins that consist of 9,701 amino acids (Table 1). The uneven sequence depth for each of the proteins made it necessary to utilize a z-score conservation calculation for each of the amino acids in the proteins so as to normalize, using the average (z-score of 0) as the starting point (*yellow*), followed by z-scores of 0–0.5 (*yellow*), 0.5–1 (*bright orange*), 1–1.5 (*orange*), 1.5–2 (*dark orange*), and >2 (*red*) (Fig. 2). The dynamics and evolution for each protein were integrated together into VIStEDD, available at RRID:SCR_018793. VIStEDD has been built for the addition of future viruses. Within the SARS-CoV-2 page is a list of each of the 24 proteins in the format of Table 1, where each protein can be clicked to assess data. On the page for each protein is a link to the individual protein data folder system, a video of the protein rotating with conservation, details of the protein function, a widget to purchase a 3D print of the protein at cost of production, the amino acid movement from molecular dynamics simulations (mds), and the table of data for each amino acid of the protein. If protein interactions structures are known, information is present with a link to protein–protein interaction (PPI) data. For example, within the nsp10 data (RRID:SCR_018793, SARS-CoV-2, nsp10), structures of nsp10 interacting with either nsp14 or nsp16 are available, both of which show highly conserved contact sites of interaction. As we continue to advance VIStEDD, we anticipate the addition of more material within each page.

The raw data of each protein is the strength of VIStEDD (drive.google.com/drive/folders/1dXBJpLo3bay1JQ9BckUsVcTViv6P0w1q?usp=sharing). For each protein present in VIStEDD, we have generated in the root folder of the protein a fasta sequence file, PDB file of the protein structure, protein models with conservation (sce = YASARA scene, pse- PyMOL scene), high-resolution image of conservation, molecular video of the conservation rotating around the *y* axis (mpg and mp4), and compiled conservation and dynamics data for each amino acid (csv or tab-delineated). Five folders of data are also present: 1) 3D, containing a vrml 3D printing file (also zipped) of conservation mapped on the protein; 2) genomics, containing aligned reads of the species sequences extracted; 3) mds, containing all of the trajectory files for the mds; 4) report, containing all of the analysis files from YASARA assessment of mds; and 5) tab, containing all of the tab-delineated analysis files of the mds. All of the 3D files can be ordered from Shapeways with the web links provided in Table S1. Another folder (PPI) contains data for all of the mds performed on protein–protein interactions, including spike-ACE2-SLC6A19, TMPRSS2, the polymerase complex, the N complex, and ACE2_S_Database consisting of 235 species ACE2 sequences modeled with spike interaction, energy-minimized, and binding or potential energy calculated. A tab-delineated file is in the ACE2_S_Database to label the species for each numbered complex.

From the mds of all proteins, a total of 6,594,981 atoms including water, there is an average movement per amino acid (root mean square fluctuation (RMSF)) of 3.2 Å with 3 amino acids correlated per residue >0.9 based on dynamics cross-correlation calculation. The secondary structures of the proteins are relatively similar from the beginning of the simulation compared with the end (Fig. 3*A*), with the largest percentage coiled (C, 45.02%) followed by helix (H, 25.42%), β-sheet (E, 15.08%), turns (T, 13.44%), and three-turn helix (G, 1.05%). On their own, before PPI, the nsp7, nsp8, nsp9, E, 3C-proteinase, and RNA-directed RNA polymerase are the most helical and β-sheet–containing proteins, whereas protein 3a, ORF8, nsp2,

Table 1
SARS-CoV-2 proteins analyzed

Shown for each protein are the gene, accession number, tool used to model the protein, known PDB files used for modeling, number of amino acids in the protein, bound molecules, number of sequences for evolution, average RMSF per residue, average number of amino acids correlated with each amino acid greater than 0.9 (DCCM), and percentage of each protein's secondary structure (C, coil; H, helix; E, β-sheet; T, turn; G, three-turn helix).

Gene	SARS protein	Accession	Modeling tool	PDB files	Amino acids	Bound molecules	Sequences	RMSF	DCCM >0.9	C	H	E	T	G
rep	nsp1	QHD43415	YASARA	2GDT	180		250	3.6	3.1	43.3	21.1	22.2	11.1	2.2
rep	nsp2	QHD43415	ITASSER		638		246	6.0	11.8	67.1	19.1	0.6	12.5	0.6
rep	Papain-like proteinase	QHD43415	YASARA	Multiple	1945	Zn	3,180	3.1	8.8	50.5	24.6	13.1	11.3	0.5
rep	nsp4	QHD43415	ITASSER		500		3,325	2.8	3.0	57.8	35.0	0.0	7.2	0.0
rep	3C-like proteinase	QHD43415	YASARA	Multiple	306	Dimer	3,397	1.4	0.6	26.5	23.9	31.7	16.7	1.3
rep	nsp6	QHD43415	ITASSER		290		2,558	2.6	0.7	62.8	14.1	1.4	21.7	0.0
rep	nsp7	QHD43415	YASARA	6NUR, 2AHM, 3UB0	83		3,256	2.9	3.0	8.4	81.9	0.0	9.6	0.0
rep	nsp8	QHD43415	YASARA	6NUR, 3UB0, 5XOG	198		3,339	5.9	8.9	19.2	52.0	19.7	9.1	0.0
rep	nsp9	QHD43415	YASARA	Multiple	113	Dimer	3,386	1.5	0.1	20.4	14.2	47.8	17.7	0.0
rep	nsp10	QHD43415	YASARA	Multiple	139	Zn	3,344	1.8	0.4	44.6	28.1	7.2	20.1	0.0
rep	RNA-directed RNA polymerase	QHD43415	YASARA	6NUR	932	Zn	5,086	2.2	2.8	30.5	42.0	11.5	13.3	2.8
rep	Helicase	QHD43415	YASARA	5WWP, 6JYT	601	Zn	5,598	1.5	1.5	35.8	26.6	23.5	14.1	0.0
rep	Guanine-N7 methyltransferase	QHD43415	YASARA	5NFY, 5C8T	527	Zn	2,794	2.1	2.4	36.1	20.1	24.1	17.5	2.3
rep	Uridylate-specific endoribonuclease	QHD43415	YASARA	Multiple	346		2,489	1.3	0.4	37.9	23.4	29.8	7.8	1.2
rep	2′-O-methyltransferase	QHD43415	YASARA	2XYV, 5YNN	298	Na, Mg, SAH	2,495	1.3	0.0	40.3	23.8	18.1	13.4	4.4
S	Spike glycoprotein	QHD43416.1	YASARA	6CRW, 6NB6, 5X58	1273	Trimer/Membrane	6,612	3.0	5.4	40.4	23.0	25.3	10.4	0.9
3a	Protein 3a	QIM47458.1	ITASSER		275		65	2.7	0.6	73.8	12.0	2.5	11.6	0.0
E	E	QIM47459.1	ITASSER		75		94	4.8	2.2	18.7	58.7	0.0	16.0	6.7
M	Membrane protein	QIM47460.1	ITASSER		222		1,507	2.7	0.5	51.8	20.7	7.7	19.8	0.0
6	ORF6	QIM47461.1	ITASSER		61		31	9.2	8.2	57.5	6.4	9.1	26.0	1.0
7a	7a	QHD43421.1	YASARA	1XAK, 1YO4	121		42	6.5	6.5	33.1	17.4	32.2	17.4	0.0
8	ORF8	QIM47463.1	ITASSER		121		35	2.2	0.4	69.4	10.7	0.0	19.8	0.0
N	Nucleoprotein	QHD43423.2	YASARA/ITASSER	Multiple	419		2,261	2.7	0.2	47.5	26.2	0.0	19.7	6.6
10	ORF10	QIM47465.1	ITASSER		38			2.6	0.0	42.1	31.6	26.3	0.0	0.0

nsp6, and nsp4 are the most disordered with coil composition (Fig. 3*B*). Several of the proteins contain high movement of the structure (ORF6, ORF7a, nsp8, and nsp2), and two of the proteins (papain-like proteinase and spike glycoprotein) have more overall lower movement indicative of hydrophobic collapse, with a high number of correlated amino acids per residue (Fig. 3*C*).

The largest proteins of SARS-CoV-2 contain the highest number of sequences extracted for homology, including the papain-like proteinase, RNA-directed RNA polymerase, spike glycoprotein, and helicase (Fig. 3*D*). Several of the proteins, including E, protein 3a, protein 7a, ORF8, and ORF6, have a low number of mapped sequences (Fig. 3*D*), whereas ORF10 has no other identified sequences. Plotting the conservation relative to the mds-based movement, RMSF, for each of the 9,701 amino acids of SARS-CoV-2 can be used to identify critical sites of proteins under high selection that might be targeted (Fig. 3*E*). Highly dynamic and conserved amino acids are the prime location for critical PPI. Therefore, we mapped sites of high movement, >5 Å, with conservation 1–1.5 (*gray*), 1.5–2 (*orange*), or >2 (*red*) S.D. values higher than the mean of the protein. Clustering these sites to the percentage of amino acid reveals a likely high selection of dynamic amino acids (Fig. 3*F*). The nsp2 protein with low coverage of species has some suggested PPI contacts conserved in the range of 1–1.5 S.D. conservation. Proteins nsp7 and nsp8, which are known to contact the RNA-dependent RNA polymerase (Fig. 1), have several amino acids conserved in the 1.5–2 S.D. range with high dynamics, as the mds of PDB 6m71 and 7btf show stability and correlation to PPI within VIStEDD.

nsp6 and conserved

The papain-like protease and nsp6 have conserved sites >2 S.D. values (Fig. 3*F*). The SARS-CoV-2 nsp6 protein is known to interact with multiple ATPases of vesicle trafficking (21) and interacts with nsp3 and nsp4 to induce double-membrane vesicles (22). Nsp6 protein also interacts with the Sigma receptor, which is thought to regulate ER stress response (21) and blocks ER-induced autophagosome/autolysosome vesicle that restricts viral production, leading to the generation of small autophagosome vesicles, thereby limiting their expansion (23). To date, no structures of nsp6 have been solved even though the protein is present in 2,558 *Coronaviridae* genomes. We identify two regions with minimal conserved hydrophobic collapse (Fig. 3*G*) consisting of mostly coiled secondary structure (Fig. 3*B*). These two regions of conservation (Fig. 3*G*) cluster together with multiple charged and aromatic amino acids that would tend to drive PPI (Fig. 3, *H* and *I*). Moving forward, these nsp6 sites could be critical regions to target with therapeutics for the broad *Coronaviridae* proteins.

Nucleocapsid (N) data insights

Both the S and N proteins have multiple sites conserved >2 S.D. with high dynamics (Fig. 3*F*), with N having around 4% of its amino acids falling into this category. These two proteins are further dissected below. The SARS-CoV-2 N protein has been shown to interact with multiple RNA processing and

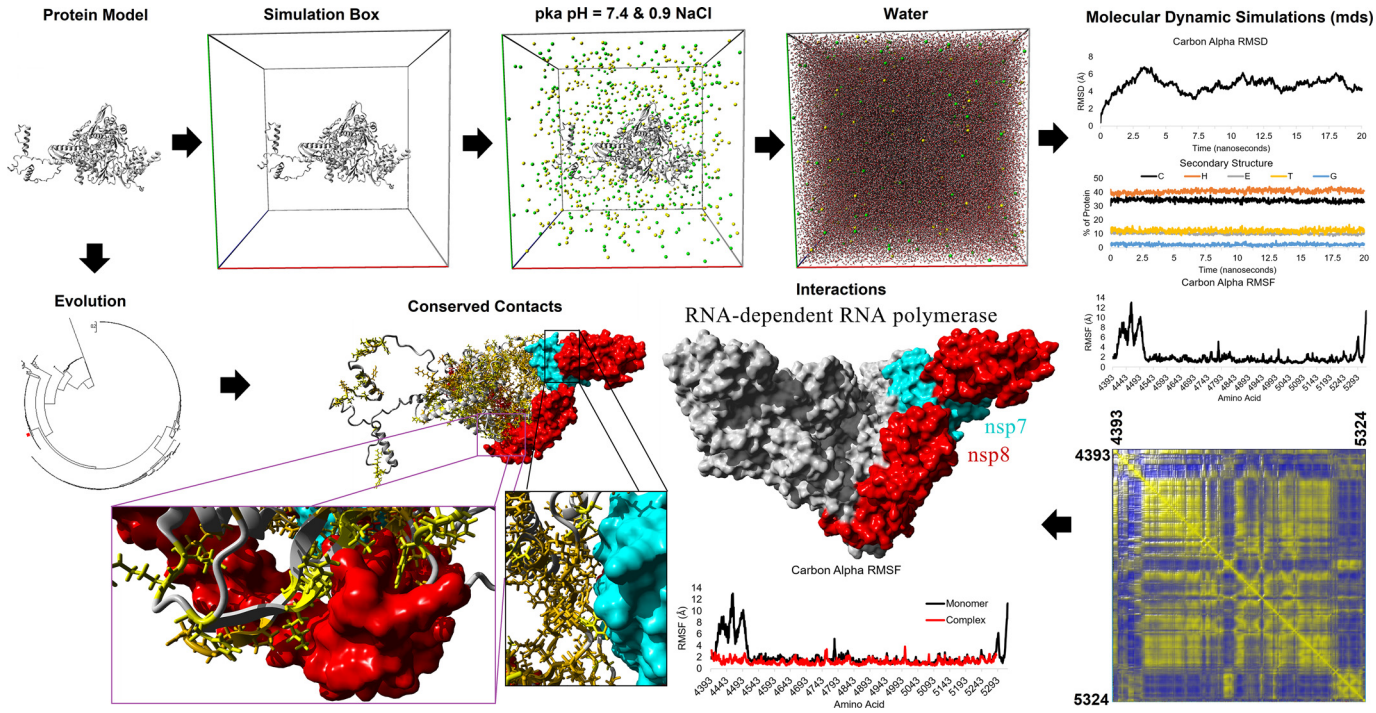

Figure 1. SARS-CoV-2 structural/evolution dynamicome workflow. Shown are data for the RNA-directed RNA polymerase.

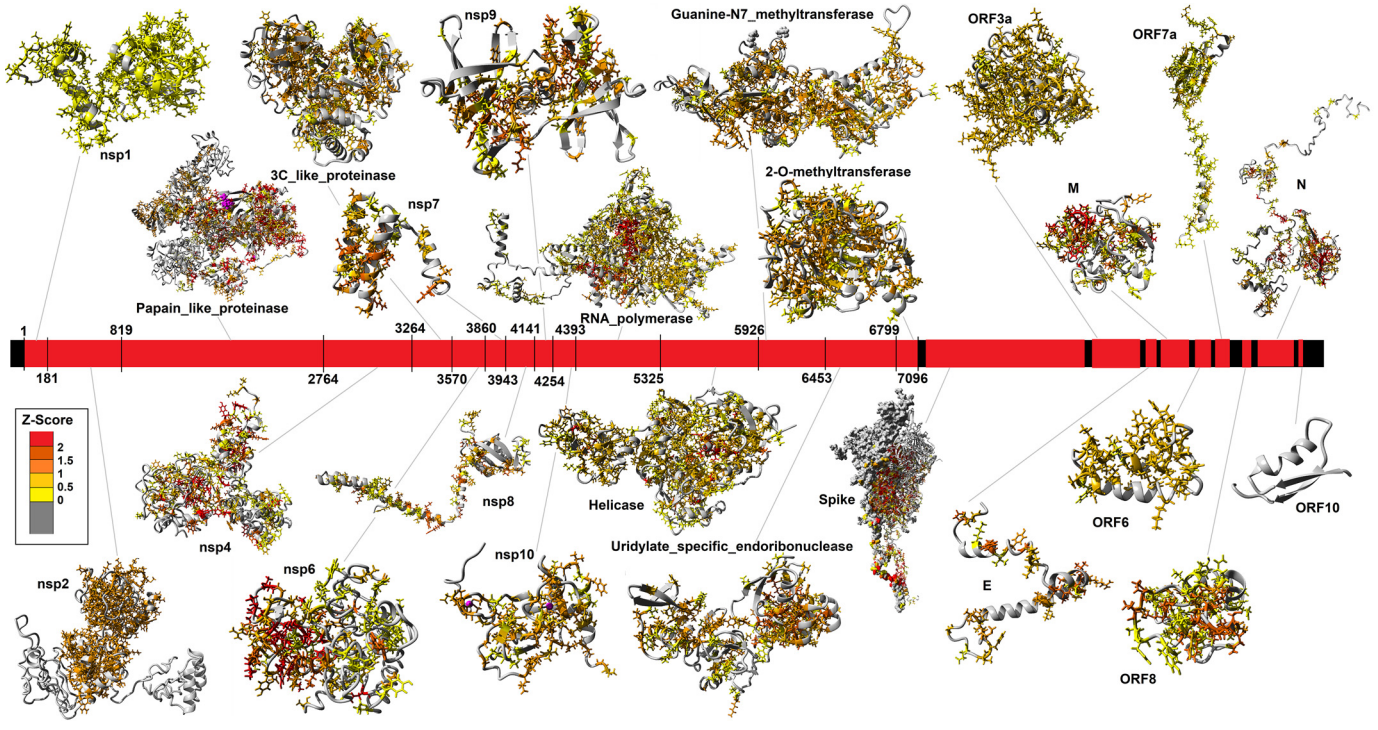

Figure 2. SARS-CoV-2 structural/evolution models. Shown in the middle is the viral RNA with protein-coding genes in *red*. *Gray lines* connect the RNA region to each protein. *Colors* on the *models* are based on z-score levels of conservation for each protein based on extracted sequences (Table 1). *Gray* amino acids fall below the average conservation (value <0) for each protein; *yellow*, 0–0.5; *light orange*, 0.5–1; *orange*, 1–1.5; *dark orange*, 1.5–2; *red*, >2.

stress granule proteins (21). Using 2,261 sequences (Fig. 4*A*), we mapped four highly conserved regions of N (Fig. 4*B*). The conservation of region 1 contributes to a highly conserved hydrophobic, aromatic core consisting of several β strands (Fig. 4, *C* and *D*). Conserved region 3 consists of several serine amino acids that are likely phosphorylated and a potential 14-3-3 binding motif (Fig. 4*D*). Molecular dynamics of N suggests three regions of structural organization, with region 1 corresponding to a domain with structural folding (Fig. 4*E*). Amino acids 331–333 were the most conserved yet dynamic site of N (Fig. 4*F*). The C-terminal region of N is known to form a multimer complex (Fig. 4*G*) with amino acids 331–333 fitting into

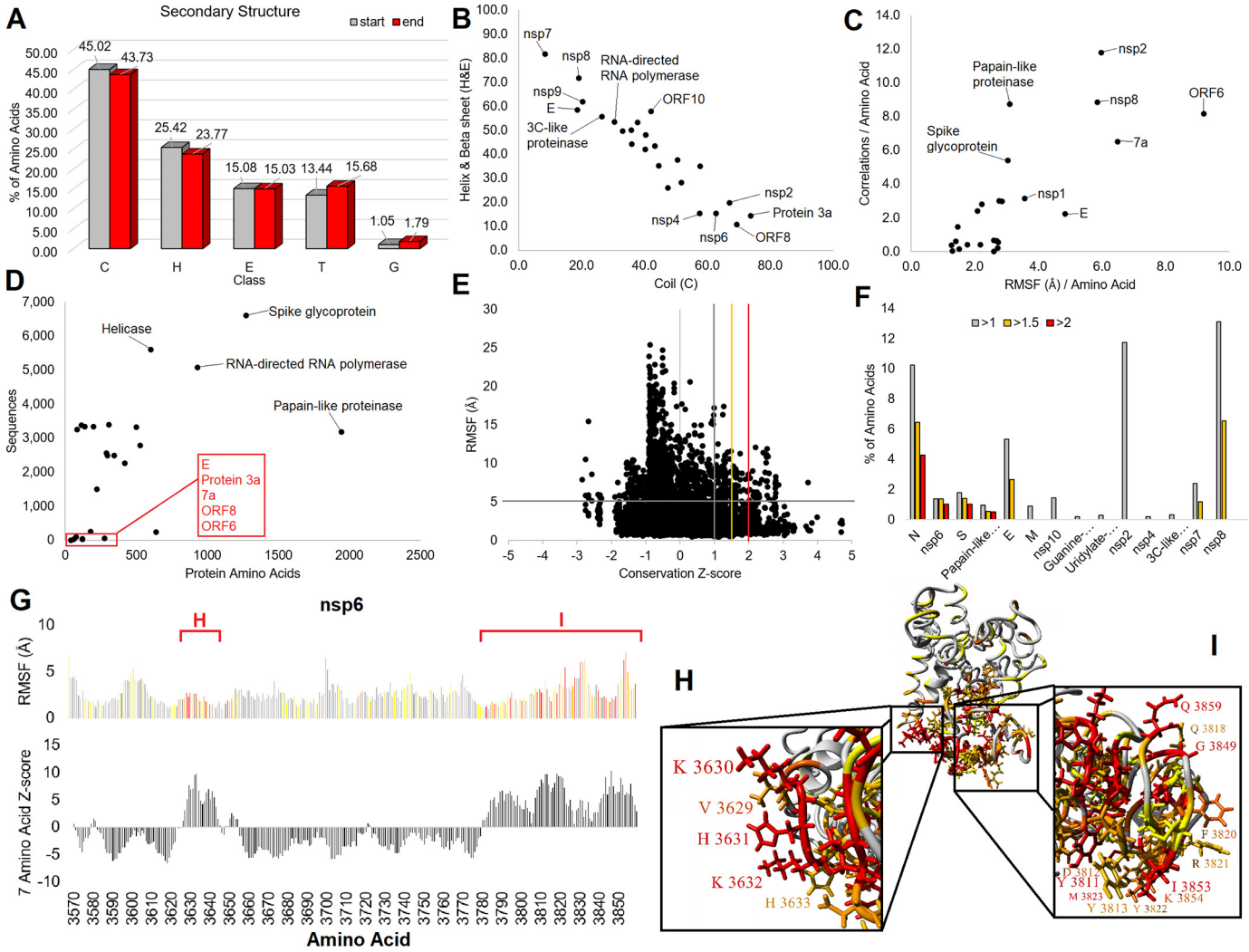

Figure 3. SARS-CoV-2 structural/evolution statistics. *A*, the percentage of secondary structure for all proteins at the start of molecular dynamics simulations (*gray*) and at the end (*red*). Classes consist of coil (*C*), helix (*H*), β-sheet (*E*), turn (*T*), and 3-turn helix (*G*). *B*, breakdown for each protein for secondary structure percentage that is coil (*x* axis) *versus* helix/β-sheet (*y* axis). Those with the most helix/beta sheet or coiled are *labeled*. *C*, breakdown of molecular dynamics simulation data for each protein showing the average amino acid RMSF (Å, *x* axis) *versus* the average number of correlated amino acids per residue (*y* axis). Proteins with high movement are *labeled*. *D*, plot of the number of amino acids in each protein *versus* the number of BLAST-extracted sequences. *Labeled* are those that have high numbers of identified sequences (*black*) and those with only a few (*red*). *E*, the conservation z-score (*x* axis) *versus* the RMSF (*y* axis) for all amino acids analyzed in SARS-CoV-2. The *lines* represent cutoffs used for *F*, with those >5 Å for RMSF and 1–1.5 (*gray*), 1.5–2 (*orange*), or >2 (*red*) value for z-score cutoffs. *F*, the percentage of each protein's amino acids that fall into identified groups from *E*, representing identified highly dynamic and conserved amino acids. *G*, conservation/dynamics of nsp6 amino acids. Shown at the *top* is the RMSF of each amino acid of nsp6 with *colors* corresponding to cutoffs of *E* and *F*. Shown at the *bottom* is a sliding window calculation of 7 amino acids for additive z-scores to map two highly conserved sites (shown in *H* and *I*). *H* and *I*, protein model of nsp6 with z-score *coloring* of Fig. 2. Shown are the two sites of high conservation with amino acids *labeled*.

contacts of the subunits (Fig. 4*H*). From the multimer N structure, conserved amino acids Val-270, Phe-274, Arg-277, Asn-285, Gly-287, Phe-286, and Asp-288 are clustered and surface-exposed (Fig. 4*I*). These sites are likely to contribute to PPI and warrant future investigations.

Posttranslational modification analysis

Building on our insights for N, we expanded to a systematic analysis of posttranslational modifications for SARS-CoV-2 proteins that was integrated into our amino acid matrix insights (Fig. 5). Current literature on SARS-CoV-2 modifications focuses on conserved and novel glycosylation and phosphorylation sites of the S protein (24). With host–pathogen interaction being regulated by modifications, it is now a target for pharmacotherapy. Data from SARS-CoV suggest modifications in the N, M, E, 3a, nsp4, and nsp9 proteins that have yet to be explored in SAR-CoV-2 (25–27). Specifically, the N protein was shown to undergo extensive acetylation, phosphorylation, sumoylation, cleavage, and ADP-ribosylation (28, 29). Inhibition of several cellular kinases (CK2 and CDK) was shown to impact proper localization of the N protein, trapping it within the nucleus of the host cell, further highlighting the need for N protein phosphorylation to carry out proper function (30). With these PTMs playing a vital role in proper virion assembly, they must be further explored and understood in SARS-CoV-2 to elucidate all options for targeted pharmacotherapy.

With high filters on each tool, we identify 186 NetPhos3.1 (phosphorylation), 15 SUMO1.0 (Sumo binding or SUMOylation), 27 SNO 1.0 (*S*-nitrosylation), 28 YNO21.0 (tyrosine

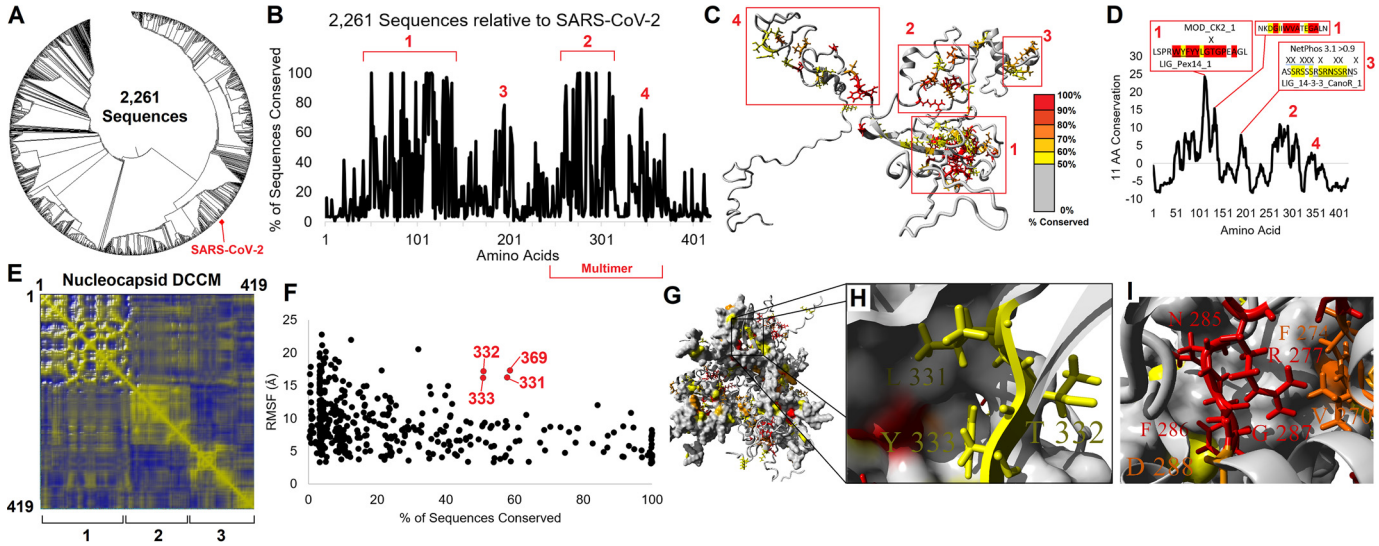

Figure 4. SARS-CoV-2 protein N (nucleocapsid) conserved dynamic amino acids. *A*, phylogenetic tree of 2,261 sequences extracted from the nr protein sequence database for N. *B*, conservation of amino acids based on sequences from *A*. Four regions of conserved function are identified in *red*. *C*, model of N with conservation (percentage of 2,261 sequences) *colored* (<50% (*gray*), 50–60% (*yellow*), 60–70% (*light orange*), 70–80% (*orange*), 80–90% (*dark orange*), and >90% (*red*)). The four regions of *B* are *boxed* and *labeled* in *red*. *D*, top conserved motifs in N. The z-scores for N conservation were placed on an 11-codon sliding window to identify regions of interest in the four conserved regions of *B*. *E*, dynamics cross-correlation matrix (DCCM) of amino acids on the nucleocapsid protein assessed with molecular dynamics simulations. *F*, intersection of conservation and dynamics of the protein with highly dynamic and conserved amino acids in *red*. *G*, multimeric model of N with *coloring* based on *E*. *H*, *zoom-in view* of *black box* from *F* showing amino acids 331–333 identified in *D*. *I*, conserved region 2 from *B* identifying the top amino acids on the multimer model.

nitration), 273 CCD1.0 (calpain cleavage), 36 Polo1.0 (polo-like kinases), 34 PUP1.0 (pupylation), 20 TSP1.0 (tyrosine sulfation), 25 PAIL2.0 (lysine acetylation), and 41 Lipid1.0 (lipidation) predicted modifications to SARS-CoV-2 viral proteins (Fig. 5*A*) with data available within our VIStEDD tools. Few modification predictions occur at highly conserved amino acids (Fig. 5*B*) or within 5-amino acid conserved motifs (Fig. 5*C*) based on our evolutionary analysis. Moreover, there are also very few unique modification predictions to the SARS-CoV-2 sequence not found throughout our evolutionary conservation. We have identified 22 modification predictions that are highly conserved and 28 sites poorly conserved, including multiple phosphorylation, lipidation, and acetylation events (Fig. 5, *D–F*). The most highly conserved motif predicted to be modified is Ser-816 of spike (Fig. 5*G*), where the amino acid is found surface-exposed on a loop of the protein (Fig. 5, *H* and *I*). Future work is desperately needed to further refine the modification sites within SARS-CoV-2.

SARS-CoV-2 S interaction with host ACE2, SLC6A19, TMPRSS2 genomics

The most highly researched protein of SARS-CoV-2, the S surface glycoprotein, is of most interest as it is the only portion outside of the virus that could be recognizable by B and T cells. The coronavirus S protein is the primary determinant of viral tropism and is responsible for receptor binding and membrane fusion. It is a large (~180-kDa) glycoprotein that is present on the viral surface as a prominent trimer, and it is composed of two domains, S1 and S2 (31). The S protein interacts with ACE2 to enter into cells, forming contacts with the ACE2-SLC6A19 dimer complex (Fig. 6*A*), where the presence of SLC6A19 is not required for spike-ACE2 binding. ACE2 servers a chaperone function on SLC6A19 and is able to stabilize the full ACE2 dimer complex in crystallization (2). We elected to build a full complex model for analysis to allow for simultaneous matrix generation for spike binding and SLC6A19 chaperone screening of genomic variants. This protein complex model was built through the integration of PDB structures 6CRW, 6NB6, and 5X58 for the trimer of spike proteins with 6M17, which shows the interaction of a fragment of spike with the ACE2-SLC6A19 dimer of dimers. From 236 species of ACE2, we determined the conservation of ACE2 amino acids, suggesting poor conservation of the S-ACE2 contact across all of vertebrate evolution (Fig. 6*B*). The lack of conservation at this site also suggests that ACE2 does not likely have a conserved interaction with another human protein that would compete with spike for function. Structural mapping of human variants from 141,456 people of the gnomADv2 database for ACE2 suggests a few possible variants at the interface of S-ACE2 interaction (Fig. 6*C*). To go from qualitative mapping to quantitative insights into human variants, we utilized mds of S-ACE2-SLC6A19 complex, determining amino acids that correlate in movement between the proteins (Fig. 6*D*). From these correlations we calculated the amino acids contributing to S-ACE interaction (*red* in Fig. 6, *E* and *F*), ACE2 dimerization (*blue* in Fig. 6, *E* and *G*), and ACE2-SLC6A19 interaction (*magenta/yellow* in Fig. 6, *E*, *H*, and *I*).

From this mds data, along with functional variant prediction tools (PolyPhen2, Provean, SIFT, Align-GVGD, and our conservation analysis), we systematically assessed functional human variants for ACE2, SLC6A19, and TMPRSS2. TMPRSS2 is involved in cleaving the complex for internalization (20). Of these three proteins, ACE2 is the only one found on a sex chromosome (X-chromosome), linking it to male hemizygous status that elevates the impact of genomic variants. Variants included are linked to protein function

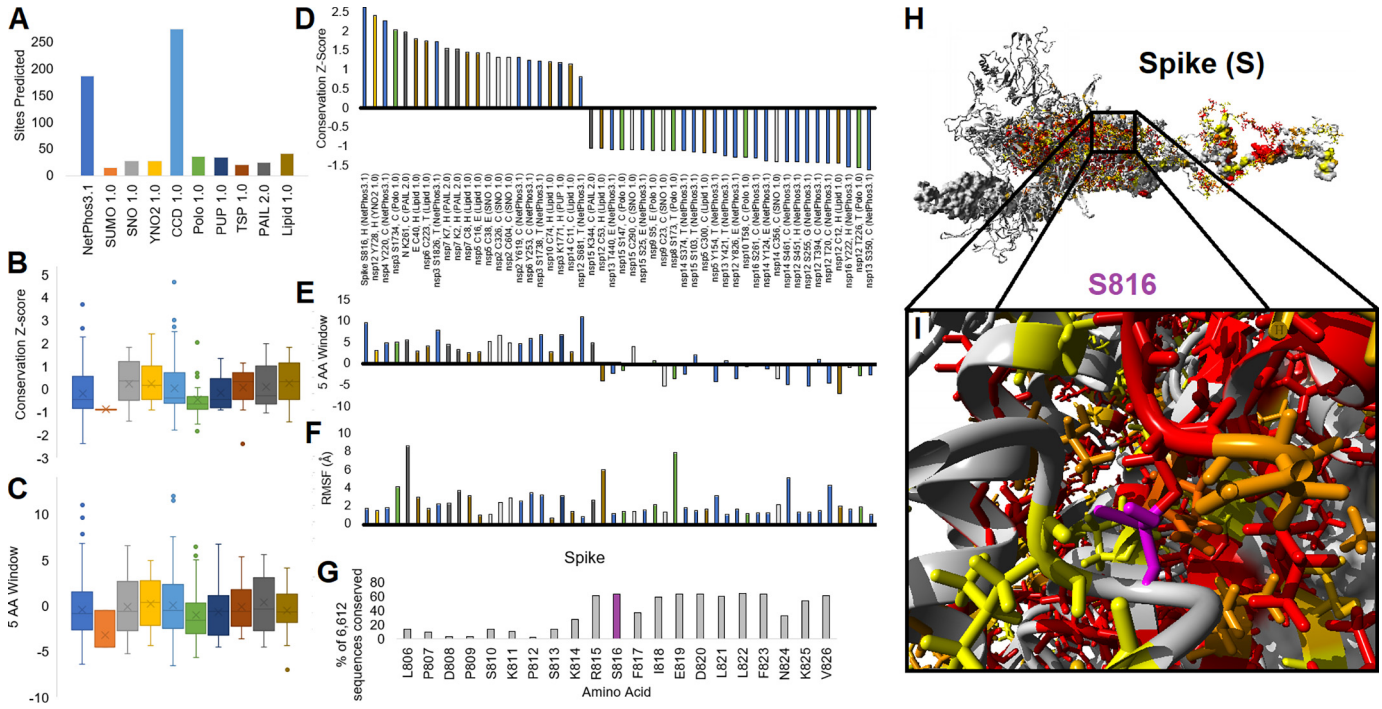

Figure 5. Posttranslational modification screening of SARS-CoV-2. *A*, ranking of functional predicted sites using 10 different tools. *B* and *C*, conservation z-score (*B*) and z-scores put on a five-codon sliding window for additive motif conservation (*C*) of each site from *A* for each tool shown as a *box and whisker plot*. *D–F*, top conserved sites (*left*) and least conserved sites that are unique to SARS-CoV-2 (*right*) for functional predictions. *Colors* of *bars* correspond to the tools of *A*. Each functional amino acid is labeled (protein, variant, secondary structure, and tool annotating PTM). Data are shown for the z-score of conservation (*D*), the 5-amino acid sliding window (*E*), and structural movement (*F*). *G*, the highest conserved site throughout the database found on spike at Ser-816 with conservation of amino acids around the site. *H* and *I*, highlighting the phosphorylation prediction (*magenta*) on spike for Ser-816 (*H*) with a *zoom-in view* of the site on one of the monomers (*I*).

(Table 2, Inclusion group = Functional), S contact (Table 2, Inclusion group = Spike contact), SLC6A19-ACE2 contact (Table 2, Inclusion group = ACE2 contact), posttranslational modifications (Table 2, Inclusion group = Glycosylation, Disulfide bond, or Phosphoserine), or known active-site amino acids (Table 2, Inclusion group = Zinc binding or Active site).

47 variants are ranked by the maximum allele frequency within the subpopulations of gnomAD. The ACE2 variant K26R has the highest allele frequency of any variant within the table, but the conserved polar basic amino acid at the spike contact likely does not impact binding. This means that there are only ultra-rare variants in these proteins, with SLC6A19 having the highest-impact variants at 29, ACE2 with 13, and TMPRSS2 with 5. Outside of the European non-Finnish population, the East Asian population carries 10 of these variants, "other" (those individuals not falling into other populations) with 9, African with 6, Latino with 6, and South Asian with 4. A total of 31 of the variants are predicted with a score of ≥4 (of 6 maximum) to be functional variants, 6 at the spike contact of ACE2, and 5 that would alter a glycosylation signal. The most interesting ACE2 variant, H378R (rs142984500), is found in 0.019% of European non-Finnish individuals and has been observed as hemizygous in 6 males of gnomADv2, is one of the critical residues of the Zn binding that drives the enzyme's function, and has not been previously published.

ACE2's hemizygous nature warrants an investigation of noncoding variants that might influence expression. The ACE2 gene contains a 5′ region that suggests most gene regulation for ACE2 to occur within this region (chrX:15,612,899-15,641,393, hg19). We extracted all noncoding variants followed by an assessment with RegulomeDB (Table S2). Five total variants are identified to potentially alter transcriptional regulation by RegulomeDB score. Two of these variants (rs4646118 and rs143185769) are found in ~9% of African individuals with hundreds of male hemizygotes identified within gnomADv3 whole-genome data. This supports the potential for noncoding variants of ACE2, with a higher allele frequency than that of coding variants, as contributors to increased susceptibility and within at-risk populations (African and Male) to SARS-CoV-2 infection.

Discussion

SARS-CoV-2 represents a generational challenge to science, racing the clock next to a global pandemic that kills ~2% of those infected. The need to understand the viral structure is urgent regarding therapeutic targets, repurposing compounds, understanding zoonotic spread, and identifying gene variant risk factors in the human host that interact with the pathogen to increase spread and pathogenicity. In future studies, investigations of known human PPI to SARS-CoV-2, similar to ACE2, can determine how genetic variations of host proteins impact the disease course and susceptibility to the virus. Further insight into the genetics could provide useful information about the susceptibility or prognosis of those exposed to or infected with SARS-CoV-2. In this paper, we generated a

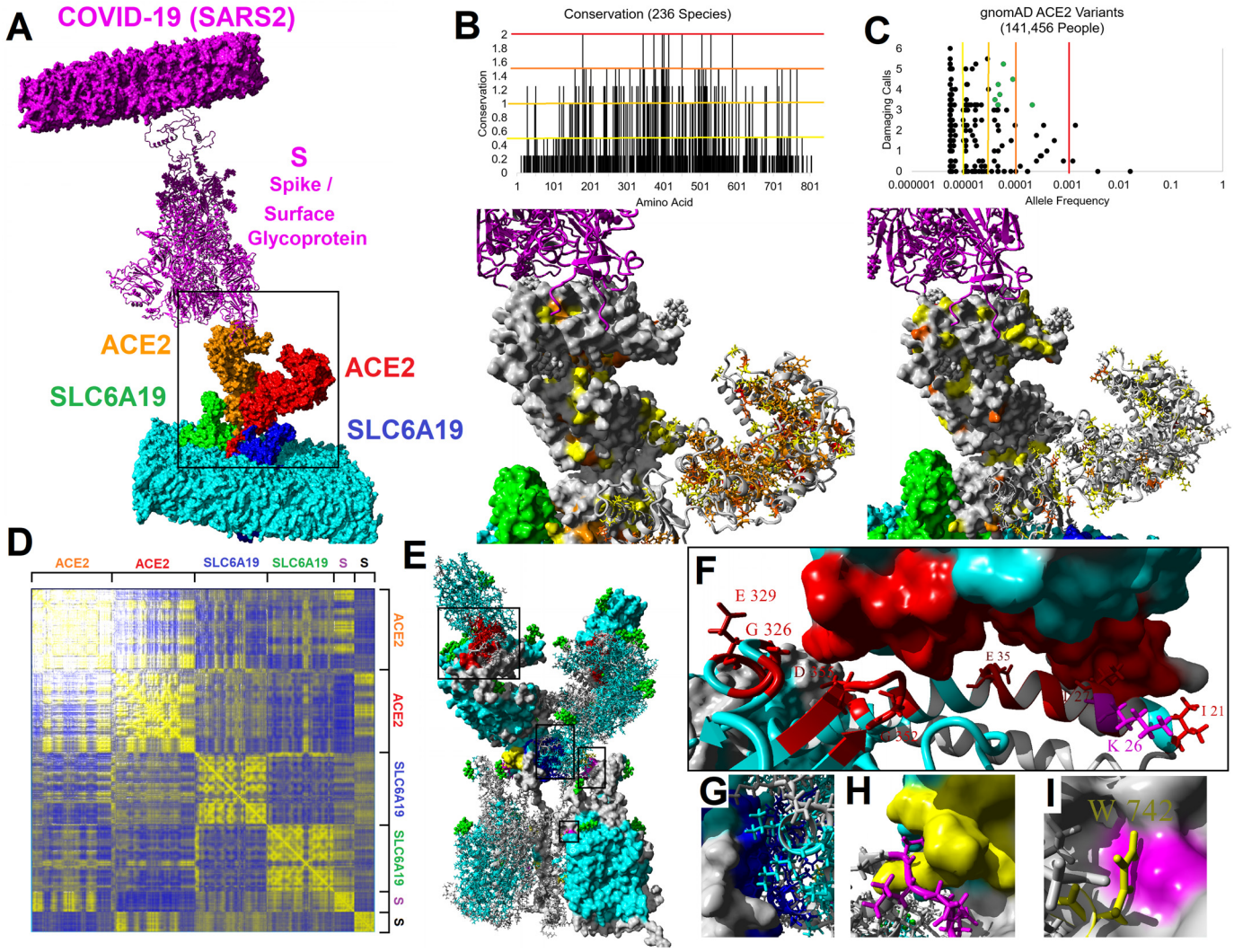

Figure 6. Spike-ACE2-SLC6A19 dynamics/evolution to human variants. *A*, structural model of the spike (*magenta*), ACE2 (*red/orange*), and SLC6A19 (*green/blue*) in lipid membranes. *Black box*, region *zoomed in* in *B* and *C*. *B*, conserved amino acids in 236 species ACE2. Cutoff *colors* are as follows: 2 (highest) (*red*), 1.5 (*dark orange*), 1 and 1.25 (*light orange*), and 1 and 0.5 (*yellow*). *C*, missense variants from 141,456 people for ACE2. Cutoffs for allele frequency are shown on the model. Several *yellow* low frequency variants can be seen at the ACE2/spike contact. *D*, dynamics correlation matrix of interacting sites throughout simulation of the dimer complex. High correlations are shown in *yellow*. *E*, amino acids in *red* are those that correlate in *D* between S and ACE2; those shown as side chains and *labeled* have human variants. The *four boxes* are the sites for *F–I*. *F*, *zoom-in view* of variants in ACE2 predicted to alter spike contact (*red*). Lys-26 is *labeled* in *magenta* with its uncertain but relatively common variant. *G*, contact sites for ACE2 dimerization (*blue*). *H* and *I*, contact sites between SLC6A19 (*magenta*) and ACE2 (*yellow*) at two different regions.

SARS-CoV-2 structural dynamicome database, integrating structural/dynamic insights with viral evolution for 24 proteins coded by SARS-CoV-2. VIStEDD has elucidated insights that include potential druggable targets, educational material describing each protein, and human variants that may impact the viral life cycle.

We show here two highly dynamic protein regions that have high conservation, indicative of PPI sites critical to viral infection and spreading. The first is the largely understudied role of the nsp6 conserved amino acids that interact with ATPases of vesicle trafficking (21). Evolutionary conservation throughout thousands of *Coronaviridae* sequences is found in two regions of the protein that are likely found near each other in 3D space, yet to date, no protein structures have ever been solved of nsp6 and publicly shared, representing a challenge to the structural biology community. ATPases are required for both endocytic and exocytic portions of the viral infections (32) and integral to the release of viral RNA into the cell (33). The conserved amino acids are found on surfaces exposed on the I-TASSER–generated predicted structure of nsp6, including multiple charged or aromatic residues. The conserved sites of nsp6 and the ATPases they interact with can likely be therapeutically targeted (21). We also show here that the N protein has several highly conserved amino acids that contribute to a multimer structural organization with surface-exposed conserved amino acids and the N-terminal region of the protein that may reflect sites of targeting to alter the ribosomal control of the protein.

Second, this new database presents many opportunities for use in education. VIStEDD was generated through a team partnership known as Characterizing our DNA Exceptions (CODE) with the intent of bringing the mds and evolutionary data to undergraduate students across the United States. The data can

Table 2

Top functional genomic variants of ACE2, SLC6A19, and TMPRSS2

The inclusion group consists of protein contacts based on molecular dynamics simulation correlations, protein modifications based on UniProt, and functional predictions. Damaging calls are a maximum of 6 based on conservation, PolyPhen2, Provean, SIFT, and Align-GVGD. Hemizygote count, maximum allele frequency, and population are from gnomADv2.

Inclusion	Protein	Chromosome	rsID	AA	Damaging calls	Hemizygote count	Maximum allele frequency	Maximum population
Spike contact	ACE2	X	rs4646116	K26R	0	282	0.011876795	Ashkenazi Jewish
Functional	SLC6A19	5	rs781039193	W530C	5	0	0.000868709	Latino
ACE2 contact	SLC6A19	5	rs374559483	P351S	1.5	0	0.000640872	African
Functional	SLC6A19	5	rs374527866	R513S	4.5	0	0.000603136	East Asian
Functional	SLC6A19	5	rs141487939	T180M	5.5	0	0.000601878	African
Functional	SLC6A19	5	rs200783817	P73L	5.25	0	0.000451309	East Asian
Functional	SLC6A19	5	rs752378013	P109L	5.5	0	0.000437158	East Asian
Functional	SLC6A19	5	rs141497538	P230S	5.5	0	0.000417246	Other
Functional	SLC6A19	5	rs552867213	R328H	5	0	0.000401091	European (Finnish)
Glycosylation	ACE2	X	rs761944150	N546D	2.5	1	0.00036784	African
Functional	SLC6A19	5	rs374976872	A271V	5.5	0	0.000300933	East Asian
Functional	SLC6A19	5	rs199795977	A329T	5	0	0.000289519	Ashkenazi Jewish
Functional	SLC6A19	5	rs1253340252	Y209S	4	0	0.000278242	Other
Functional	SLC6A19	5	rs139182948	R95W	5	0	0.000277701	Other
Functional	SLC6A19	5	rs369804798	S314L	5.5	0	0.000276932	Other
Functional	SLC6A19	5	rs142164435	R328C	5.5	0	0.000247927	European (non-Finnish)
Functional	SLC6A19	5	rs762989809	R57C	5.25	0	0.000218126	East Asian
Glycosylation	ACE2	X	rs143158922	N103H	0.5	0	0.000210615	African
Zinc binding	ACE2	X	rs142984500	H378R	4.5	6	0.000194968	European (non-Finnish)
Spike contact	ACE2	X	rs143936283	E329G	0	1	0.000190476	Other
Functional	SLC6A19	5	rs1403845937	W56R	5	0	0.000164582	Other
Functional	SLC6A19	5	rs201936518	R95Q	5	0	0.00016372	Other
Functional	SLC6A19	5	rs375879452	T256M	5.5	0	0.000150708	East Asian
Functional	TMPRSS2	21	rs762844469	G391E	5.25	0	0.000144601	Latino
Spike contact	ACE2	X	rs1348114695	E35K	0	1	0.000144436	East Asian
Functional	SLC6A19	5	rs757679627	G93R	5.5	0	0.00013885	Other
Functional	TMPRSS2	21	rs1306483136	S460R	5.25	0	0.000138351	Other
Functional	SLC6A19	5	rs202220597	T243M	5.5	0	0.000130693	South Asian
Functional	SLC6A19	5	rs200745023	L242P	5	0	0.000130685	South Asian
Functional	SLC6A19	5	rs778015723	P149L	5.5	0	0.000130651	South Asian
Functional	ACE2	X	rs200745906	P263S	5.25	0	0.000125798	European (non-Finnish)
Functional	SLC6A19	5	rs765501634	E405K	5.25	0	0.000120318	African
Functional	SLC6A19	5	rs748703513	T330I	5	0	0.000114404	European (non-Finnish)
Active site	ACE2	X	rs1395782023	E375D	3.25	1	0.000109585	Latino
Functional	SLC6A19	5	rs769457402	E165K	5.25	0	0.000108755	East Asian
Functional	TMPRSS2	21	rs145171279	T459I	5.5	0	0.000108731	East Asian
Functional	SLC6A19	5	rs756920378	I325T	5.25	0	0.000108731	East Asian
Glycosylation	TMPRSS2	21	rs1326192818	N213K	1.5	0	8.67553E−05	Latino
Spike contact	ACE2	X	rs781255386	T27A	0	0	7.30327E−05	Latino
Glycosylation	SLC6A19	5	rs766784542	N158T	2.5	0	6.53381E−05	South Asian
Functional	ACE2	X	rs766319182	M270V	4	2	6.44787E−05	European (non-Finnish)
Disulfide bond	TMPRSS2	21	rs906113408	C297Y	6	0	5.99988E−05	Latino
Glycosylation	SLC6A19	5	rs146176472	N258S	1.5	0	4.06207E−05	African
Functional	ACE2	X	rs756358940	I291K	5.25	2	3.75756E−05	European (non-Finnish)
Phosphoserine	SLC6A19	5	rs754779609	S17C	2	0	3.57373E−05	European (non-Finnish)
Spike contact	ACE2	X	rs961360700	D355N	4.25	0	2.59141E−05	European (non-Finnish)
Spike contact	ACE2	X	rs778030746	I21V	0	2	2.44636E−05	European (non-Finnish)

be used by anyone as the full data set is publicly available (drive.google.com/drive/folders/1dXBJpLo3bay1JQ9BckUsVcTViv6P0w1q?usp=sharing). From these data, we provide high-resolution figures of conservation-mapped, structural files that can be opened in either YASARA or PyMOL tools and molecular videos of the molecules rotating with conservation. For the 3D files, we have provided a vrml file for each protein that can be fed to any 3D printer, with our file containing colors for conservation as well. To expedite 3D printing, we have provided all of the vrml files to Shapeways to allow at-cost printing of the proteins (Table S1), where the files can be used for education.

For this work, we utilized an integration of known PDB-based structures using YASARA modeling tools followed by energy minimization within a physiological environment. This allows for reduction of crystal packing forces and merging the structural knowledge of the PDB into a single starting protein. Where no templates exist in the PDB (9 of 24 proteins), we utilized the already existing database of I-TASSER SARS-CoV-2 proteins.

From these structures, our primary goal was to move qualitative structures into a quantitative matrix that can be integrated with evolutionary data. This strategy allows for a single amino acid matrix of functional data for every protein of the SARS-CoV-2. To do that, we utilized 20 ns of molecular dynamics simulations, allowing for tracking of the constraints and correlations of each amino acid using the starting structure. These 20 ns of time all provide dynamic equilibrium as can be observed within the source data (tab-delineated folder, file x_analysisres.tab). Whereas 20 ns of time is not enough to give insights into allosteric movement of the protein, it provides robust quantitative maps of amino acid constraints of the initial protein templates. In the future, we plan to integrate the growing wealth of structural knowledge of inhibitor-bound enzymes, protein–protein interactions, and SARS-CoV-2 protein allosteric structures into our integrated amino acid matrix using additional starting points of the protein structures for 20-ns simulations. The five protein–protein interaction simulations shown within this paper were the beginning of that work.

The biophysical and structural evidence suggested that SARS-CoV-2 may bind ACE2 with a much higher affinity than SARS-CoV (34). Our group had previously investigated the evolution of ACE2 throughout species, including mapping variants within rat populations (35), a model system for studying the renin-angiotensin aldosterone system (known as the RAAS). We have expanded those tools here, generating ACE2 models for 235 species from mammals to birds/fish, where each of the models is energy-minimized with the S protein interaction. That database can be used by groups to investigate species where SARS-CoV-2 may be able to enter cells, with variants that enhance or inhibit binding. With the S-ACE2-SLC6A19 complex resolved and mds available, a systematic quantitative map of human variants was created. Few functional variants were identified in ACE2, SLC6A19, or TMPRSS2. Moreover, none of the variants identified are common, all falling below 1% of the global population. In SLC6A19 or TMPRSS2, these rare variants would have minimal outcomes as they would rarely, if ever, reach homozygosity to cause 100% of the proteins to be influenced by variants. However, ACE2 falls on the X-chromosome, linking it to male-specific hemizygous influence. Several of the top rare ACE2 variants identified within this paper have been seen to have hemizygous variants, suggesting a dominant outcome. Whereas functional missense variants in ACE2 are ultra-rare, noncoding variants reach slightly higher allele frequencies, namely rs4646118 and rs143185769. This would suggest that as genome sequencing of patients with SARS-CoV-2 occurs, we should focus on analysis of ultra-rare missense variants listed within this paper and more importantly on several likely functional noncoding variants.

Through the quantitative dissection of the SARS-CoV-2–encoded proteins and their interaction partners, we have developed a database (VIStEDD) of information that can be used to advance our knowledge. The amino acid quantitative matrix knowledge generated from this work can be used to pinpoint the many human protein interaction partners, mechanisms for PTMs, screening of SARS-CoV-2 functional mutational drift, and drug development to sites that are unique to SARS-CoV-2 or conserved across the coronavirus family. From the ability to regulate these interactions with pharmaceutical intervention to understanding how host genomics can influence the viral biology, VIStEDD will allow for more robust insight into SARS-CoV-2 biology.

Experimental procedures

Protein modeling and molecular dynamics simulations

Similar to our laboratory's analysis of human variants, we have assessed the SARS-CoV-2 proteins using our previous established workflow (36). Protein modeling was performed by utilizing YASARA homology modeling (37, 38) when a structural template was available that matched the sequences listed in Table 1. Homology modeling is the preferred platform as it allows molecules associated with the proteins to be included in the protein structure, including Zn ions critical to folding of Zn fingers within the papain-like proteinase, nsp10, RNA-directed RNA polymerase, helicase, and guanine-N7 methyltransferase. The transmembrane portions of S were manually cleaned and clustered, allowing for insertion into a phosphatidylethanolamine membrane before mds using the YASARA md_runmembrane macro. For those proteins without structural homologs, we utilized models that are part of the I-TASSER SARS-CoV-2 database (39). Each of these models was then fed through homology modeling in YASARA to normalize energetic predictions to the homology models. mds were performed on each of the proteins in YASARA (38) using the AMBER14 force field (40), 0.997 g/ml explicit water, NaCl at 0.9 mass fraction, a pH of 7.4 for protonation predictions, saving trajectory files every 25 ps for 20 ns total. The trajectory files were analyzed with the YASARA md_analyze and md_analyzeres macros, generating an HTML file present in each of the protein report folders. If this report folder is downloaded and the HTML file opened, it generates a full report of the protein dynamics, including multiple figures of analysis. Additionally, all of the tab-delineated analysis files are within the tab folder of VIStEDD proteins and the trajectory files, allowing for reanalysis of trajectory.

Generation of database information

From the models and mds, we generated files for VIStEDD. Sequences (within the genomics folder of each protein) were extracted using the sequences listed in Table 1 with BLASTp against the nr protein sequences and aligned using ClustalW (41). An amino acid matrix of all sequences was generated in MEGA (42), followed by calculating the percentage of all amino acids at each spot that are the same as in SARS-CoV-2. The conservation of each protein was normalized using a z-score ((value − mean)/S.D.) for comparison across all proteins. The generated models were loaded into YASARA (z-score 0–0.5 (yellow), 0.5–1 (*color 166*), 1–1.5 (*color 157*), 1.5–2 (*color 145*), >2 (*red*)) or PyMOL (z-score 0–0.5 (*yellow*), 0.5–1 (*bright orange*), 1–1.5 (*orange*), 1.5–2 (*warm pink*), >2 (*red*)), *colored* based on conservation, and saved as respective scene files for the tool. The YASARA colored molecule was saved as a *y* axis rotation in mpg and converted into mp4, which is more amenable to PowerPoint. Within PyMOL, the structure was also exported as a vrml file for 3D printing. Motifs and posttranslational modification predictions for the N protein were generated using ELM (43) and NetPhos3.1 (44).

Human variant analysis

Models for ACE2 and SLC6A19 were homology-modeled using PDB 6M17 (S-ACE2-SLC6A19) followed by alignment back onto the complex. TMPRSS2 was homology-modeled using YASARA followed by manual correction of the transmembrane helix. Each of the two models was embedded into a phosphatidylethanolamine lipid membrane using the YASARA macro md_runmembrane followed by mds as done on SARS-CoV-2 proteins. Vertebrate sequences of the three proteins were extracted using NCBI orthologs for the transcript, open reading frames were assessed using Transdecoder (45), and sequences were aligned using ClustalW codons in MEGA. Conservation was performed on the data as previously published (46). Genomic missense variants were extracted from gnomADv2 for each of the three genes followed by assessment using PolyPhen2 (47), Provean (48), SIFT (49), and

Align-GVGD (50). The ACE2 regulatory region was identified using the Roadmap Epigenomics 18-state model (51), followed by the extraction of all gnomADv3 variants and assessment with RegulomeDB (52).

Data availability

All data published in this paper are available at 10.6084/m9.figshare.12298790.v1 or within our VIStEDD tools (RRID: SCR_018793).

Author contributions—R. G., N. L., C. P. B., B. D. U., S. R., and J. W. P. conceptualization; R. G., J. C., C. L. S., O. S., D. H., M. M., and J. W. P. data curation; R. G., J. C., C. L. S., J. P., H. S., T. W. C., W. F., A. F., E. L., J. B., X. L., O. S., X. S., A. U., D. H., and J. W. P. formal analysis; R. G., J. C., H. S., A. F., E. L., J. B., O. S., J. A. C., S. R., and J. W. P. writing-original draft; R. G., J. C., C. L. S., H. S., T. W. C., W. F., A. F., E. L., J. B., X. L., O. S., X. S., A. U., D. H., M. M., N. L., J. A. C., C. P. B., B. D. U., S. R., and J. W. P. writing-review and editing; C. L. S., A. U., M. M., N. L., J. A. C., C. P. B., and J. W. P. supervision; N. L., J. A. C., S. R., and J. W. P. funding acquisition; J. W. P. resources; J. W. P. investigation; J. W. P. methodology; J. W. P. project administration.

Funding and additional information—This study was supported by National Institutes of Health Grants R01-GM108618 (to J. A. C.) and K01-ES025435 (to J. W. P.), Michigan State University, and Spectrum Health. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Conflict of interest—The authors declare that they have no conflicts of interest with the contents of this article.

Abbreviations—The abbreviations used are: SARS, severe acute respiratory syndrome; CoV-2, coronavirus 2; MERS, Middle East respiratory syndrome; 3C, 3-chymotrypsin; VIStEDD, Viral Integrated Structural Evolution Dynamic Database; nr, nonredundant; mds, molecular dynamics simulation(s); RMSF, root mean square fluctuation; PPI, protein–protein interaction; 3D, three-dimensional; PDB, Protein Data Bank; DCCM, dynamics cross-correlation matrix; S, spike; N, nucleocaspid.

References

1. Kasmi, Y., Khataby, K., Souiri, A., and Ennaji, M. M. (2020) Coronaviridae: 100,000 years of emergence and reemergence. In *Emerging and Reemerging Viral Pathogens* (Ennaji, M. M., ed) pp. 127–149, Academic Press, Inc., New York
2. Yan, R., Zhang, Y., Li, Y., Xia, L., Guo, Y., and Zhou, Q. (2020) Structural basis for the recognition of SARS-CoV-2 by full-length human ACE2. *Science* **367,** 1444–1448 CrossRef Medline
3. Kuba, K., Imai, Y., Rao, S., Gao, H., Guo, F., Guan, B., Huan, Y., Yang, P., Zhang, Y., Deng, W., Bao, L., Zhang, B., Liu, G., Wang, Z., Chappell, M., *et al.* (2005) A crucial role of angiotensin converting enzyme 2 (ACE2) in SARS coronavirus-induced lung injury. *Nat. Med.* **11,** 875–879 CrossRef Medline
4. Ceraolo, C., and Giorgi, F. M. (2020) Genomic variance of the 2019-nCoV coronavirus. *J. Med. Virol.* **92,** 522–528 CrossRef Medline
5. Andersen, K. G., Rambaut, A., Lipkin, W. I., Holmes, E. C., and Garry, R. F. (2020) The proximal origin of SARS-CoV-2. *Nat. Med.* **26,** 450–452 CrossRef Medline
6. Guarner, J. (2020) Three emerging coronaviruses in two decades: the story of SARS, MERS, and now COVID-19. *Am. J. Clin. Pathol.* **153,** 420–421 CrossRef Medline
7. Dong, E., Du, H., and Gardner, L. (2020) An interactive web-based dashboard to track COVID-19 in real time. *Lancet Infect. Dis.* **20,** 533–534 CrossRef Medline
8. Prompetchara, E., Ketloy, C., and Palaga, T. (2020) Immune responses in COVID-19 and potential vaccines: lessons learned from SARS and MERS epidemic. *Asian Pac. J. Allergy Immunol.* **38,** 1–9 CrossRef Medline
9. Zhou, F., Yu, T., Du, R., Fan, G., Liu, Y., Liu, Z., Xiang, J., Wang, Y., Song, B., Gu, X., Guan, L., Wei, Y., Li, H., Wu, X., Xu, J., *et al.* (2020) Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. *Lancet* **395,** 1054–1062 CrossRef Medline
10. Mehta, P., McAuley, D. F., Brown, M., Sanchez, E., Tattersall, R. S., and Manson, J. J. and HLH Across Speciality Collaboration, UK (2020) COVID-19: consider cytokine storm syndromes and immunosuppression. *Lancet* **395,** 1033–1034 CrossRef Medline
11. Shi, S., Qin, M., Shen, B., Cai, Y., Liu, T., Yang, F., Gong, W., Liu, X., Liang, J., Zhao, Q., Huang, H., Yang, B., and Huang, C. (2020) Association of cardiac injury with mortality in hospitalized patients with COVID-19 in Wuhan, China. *JAMA Cardiol.* CrossRef CrossRef Medline
12. Wada, M., Lokugamage, K. G., Nakagawa, K., Narayanan, K., and Makino, S. (2018) Interplay between coronavirus, a cytoplasmic RNA virus, and nonsense-mediated mRNA decay pathway. *Proc. Natl. Acad. Sci. U. S. A.* **115,** E10157–E10166 CrossRef Medline
13. Malle, L. (2020) A map of SARS-CoV-2 and host cell interactions. *Nat. Rev. Immunol.* **20,** 351 CrossRef Medline
14. Wu, C., Liu, Y., Yang, Y., Zhang, P., Zhong, W., Wang, Y., Wang, Q., Xu, Y., Li, M., Li, X., Zheng, M., Chen, L., and Li, H. (2020) Analysis of therapeutic targets for SARS-CoV-2 and discovery of potential drugs by computational methods. *Acta Pharm. Sin. B* **10,** 766–788 CrossRef Medline
15. Walls, A. C., Park, Y.-J., Tortorici, M. A., Wall, A., McGuire, A. T., and Veesler, D. (2020) Structure, function, and antigenicity of the SARS-CoV-2 spike glycoprotein. *Cell* **181,** 281–292.e6 CrossRef Medline
16. Hamming, I., Timens, W., Bulthuis, M. L. C., Lely, A. T., Navis, G. J., and van Goor, H. (2004) Tissue distribution of ACE2 protein, the functional receptor for SARS coronavirus: a first step in understanding SARS pathogenesis. *J. Pathol.* **203,** 631–637 CrossRef Medline
17. Sungnak, W., Huang, N., Bécavin, C., Berg, M., Queen, R., Litvinukova, M., Talavera-López, C., Maatz, H., Reichart, D., Sampaziotis, F., Worlock, K. B., Yoshida, M., Barnes, J. L., and HCA Lung Biological Network (2020) SARS-CoV-2 entry factors are highly expressed in nasal epithelial cells together with innate immune genes. *Nat. Med.* **26,** 681–687 CrossRef Medline
18. Uhal, B. D., Dang, M., Dang, V., Llatos, R., Cano, E., Abdul-Hafez, A., Markey, J., Piasecki, C. C., and Molina-Molina, M. (2013) Cell cycle dependence of ACE-2 explains downregulation in idiopathic pulmonary fibrosis. *Eur. Respir. J.* **42,** 198–210 CrossRef Medline
19. Barkauskas, C. E., Cronce, M. J., Rackley, C. R., Bowie, E. J., Keene, D. R., Stripp, B. R., Randell, S. H., Noble, P. W., and Hogan, B. L. M. (2013) Type 2 alveolar cells are stem cells in adult lung. *J. Clin. Invest.* **123,** 3025–3036 CrossRef Medline
20. Hoffmann, M., Kleine-Weber, H., Schroeder, S., Krüger, N., Herrler, T., Erichsen, S., Schiergens, T. S., Herrler, G., Wu, N.-H., Nitsche, A., Müller, M. A., Drosten, C., and Pöhlmann, S. (2020) SARS-CoV-2 cell entry depends on ACE2 and TMPRSS2 and is blocked by a clinically proven protease inhibitor. *Cell* **181,** 271–280.e8 CrossRef Medline
21. Gordon, D. E., Jang, G. M., Bouhaddou, M., Xu, J., Obernier, K., O'Meara, M. J., Guo, J. Z., Swaney, D. L., Tummino, T. A., Huettenhain, R., Kaake, R. M., Richards, A. L., Tutuncuoglu, B., Foussard, H., Batra, J., *et al.* (2020) A SARS-CoV-2-human protein-protein interaction map reveals drug targets and potential drug-repurposing. *bioRxiv* CrossRef CrossRef
22. Angelini, M. M., Akhlaghpour, M., Neuman, B. W., and Buchmeier, M. J. (2013) Severe acute respiratory syndrome coronavirus nonstructural proteins 3, 4, and 6 induce double-membrane vesicles. *mBio* **4,** CrossRef CrossRef Medline

23. Cottam, E. M., Whelband, M. C., and Wileman, T. (2014) Coronavirus NSP6 restricts autophagosome expansion. *Autophagy* **10,** 1426–1441 CrossRef Medline
24. Xiao, X., Chakraborti, S., Dimitrov, A. S., Gramatikoff, K., and Dimitrov, D. S. (2003) The SARS-CoV S glycoprotein: expression and functional characterization. *Biochem. Biophys. Res. Commun.* **312,** 1159–1164 CrossRef Medline
25. Fung, T. S., and Liu, D. X. (2018) Post-translational modifications of coronavirus proteins: roles and function. *Future Virol.* **13,** 405–430 CrossRef Medline
26. Nal, B., Chan, C., Kien, F., Siu, L., Tse, J., Chu, K., Kam, J., Staropoli, I., Crescenzo-Chaigne, B., Escriou, N., van der Werf, S., Yuen, K.-Y., and Altmeyer, R. (2005) Differential maturation and subcellular localization of severe acute respiratory syndrome coronavirus surface proteins S, M and E. *J. Gen. Virol.* **86,** 1423–1434 CrossRef Medline
27. Oostra, M., de Haan, C. A M., de Groot, R. J., and Rottier, P. J. M. (2006) Glycosylation of the severe acute respiratory syndrome coronavirus triple-spanning membrane proteins 3a and M. *J. Virol.* **80,** 2326–2336 CrossRef Medline
28. Fan, Z., Zhuo, Y., Tan, X., Zhou, Z., Yuan, J., Qiang, B., Yan, J., Peng, X., and Gao, G. F. (2006) SARS-CoV nucleocapsid protein binds to hUbc9, a ubiquitin conjugating enzyme of the sumoylation system. *J. Med. Virol.* **78,** 1365–1373 CrossRef Medline
29. Surjit, M., Kumar, R., Mishra, R. N., Reddy, M. K., Chow, V. T. K., and Lal, S. K. (2005) The severe acute respiratory syndrome coronavirus nucleocapsid protein is phosphorylated and localizes in the cytoplasm by 14-3-3-mediated translocation. *J. Virol.* **79,** 11476–11486 CrossRef Medline
30. Surjit, M., and Lal, S. K. (2008) The SARS-CoV nucleocapsid protein: a protein with multifarious activities. *Infect. Genet. Evol.* **8,** 397–405 CrossRef Medline
31. Belouzard, S., Millet, J. K., Licitra, B. N., and Whittaker, G. R. (2012) Mechanisms of coronavirus cell entry mediated by the viral spike protein. *Viruses* **4,** 1011–1033 CrossRef Medline
32. Palokangas, H., Metsikkö, K., and Väänänen, K. (1994) Active vacuolar H^+ATPase is required for both endocytic and exocytic processes during viral infection of BHK-21 cells. *J. Biol. Chem.* **269,** 17577–17585 Medline
33. Hinton, A., Bond, S., and Forgac, M. (2009) V-ATPase functions in normal and disease processes. *Pflugers Arch.* **457,** 589–598 CrossRef Medline
34. Wrapp, D., Wang, N., Corbett, K. S., Goldsmith, J. A., Hsieh, C.-L., Abiona, O., Graham, B. S., and McLellan, J. S. (2020) Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. *Science* **367,** 1260–1263 CrossRef Medline
35. Prokop, J. W., Petri, V., Shimoyama, M. E., Watanabe, I. K. M., Casarini, D. E., Leeper, T. C., Bilinovich, S. M., Jacob, H. J., Santos, R. A. S., Martins, A. S., Araujo, F. C., Reis, F. M., and Milsted, A. (2015) Structural libraries of protein models for multiple species to understand evolution of the renin-angiotensin system. *Gen. Comp. Endocrinol.* **215,** 106–116 CrossRef Medline
36. Prokop, J. W., Lazar, J., Crapitto, G., Smith, D. C., Worthey, E. A., and Jacob, H. J. (2017) Molecular modeling in the age of clinical genomics, the enterprise of the next generation. *J. Mol. Model.* **23,** 75 CrossRef Medline
37. Krieger, E., Joo, K., Lee, J., Lee, J., Raman, S., Thompson, J., Tyka, M., Baker, D., and Karplus, K. (2009) Improving physical realism, stereochemistry, and side-chain accuracy in homology modeling: four approaches that performed well in CASP8. *Proteins* **77,** 114–122 CrossRef Medline
38. Krieger, E., and Vriend, G. (2015) New ways to boost molecular dynamics simulations. *J. Comput. Chem.* **36,** 996–1007 CrossRef Medline
39. Zhang, C., Zheng, W., Huang, X., Bell, E. W., Zhou, X., and Zhang, Y. (2020) Protein structure and sequence reanalysis of 2019-nCoV genome refutes snakes as its intermediate host and the unique similarity between its spike protein insertions and HIV-1. *J. Proteome Res.* **19,** 1351–1360 CrossRef Medline
40. Duan, Y., Wu, C., Chowdhury, S., Lee, M. C., Xiong, G., Zhang, W., Yang, R., Cieplak, P., Luo, R., Lee, T., Caldwell, J., Wang, J., and Kollman, P. (2003) A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. *J. Comput. Chem.* **24,** 1999–2012 CrossRef Medline
41. Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., McGettigan, P. A., McWilliam, H., Valentin, F., Wallace, I. M., Wilm, A., Lopez, R., Thompson, J. D., Gibson, T. J., and Higgins, D. G. (2007) Clustal W and Clustal X version 2.0. *Bioinformatics* **23,** 2947–2948 CrossRef Medline
42. Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., and Kumar, S. (2011) MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol. Biol. Evol.* **28,** 2731–2739 CrossRef Medline
43. Dinkel, H., Michael, S., Weatheritt, R. J., Davey, N. E., Van Roey, K., Altenberg, B., Toedt, G., Uyar, B., Seiler, M., Budd, A., Jödicke, L., Dammert, M. A., Schroeter, C., Hammer, M., Schmidt, T., *et al.* (2012) ELM–the database of eukaryotic linear motifs. *Nucleic Acids Res.* **40,** D242–D251 CrossRef Medline
44. Blom, N., Sicheritz-Pontén, T., Gupta, R., Gammeltoft, S., and Brunak, S. (2004) Prediction of post-translational glycosylation and phosphorylation of proteins from the amino acid sequence. *Proteomics* **4,** 1633–1649 CrossRef Medline
45. Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J., Couger, M. B., Eccles, D., Li, B., Lieber, M., MacManes, M. D., Ott, M., Orvis, J., Pochet, N., Strozzi, F., *et al.* (2013) *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat. Protoc.* **8,** 1494–1512 CrossRef Medline
46. Prokop, J. W., Yeo, N. C., Ottmann, C., Chhetri, S. B., Florus, K. L., Ross, E. J., Sosonkina, N., Link, B. A., Freedman, B. I., Coppola, C. J., McDermott-Roe, C., Leysen, S., Milroy, L.-G., Meijer, F. A., Geurts, A. M., *et al.* (2018) Characterization of coding/noncoding variants for SHROOM3 in patients with CKD. *J. Am. Soc. Nephrol.* **29,** 1525–1535 CrossRef Medline
47. Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., Kondrashov, A. S., and Sunyaev, S. R. (2010) A method and server for predicting damaging missense mutations. *Nat. Methods* **7,** 248–249 CrossRef Medline
48. Choi, Y., and Chan, A. P. (2015) PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels. *Bioinformatics* **31,** 2745–2747 CrossRef Medline
49. Ng, P. C., and Henikoff, S. (2003) SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res.* **31,** 3812–3814 CrossRef Medline
50. Tavtigian, S. V., Deffenbaugh, A. M., Yin, L., Judkins, T., Scholl, T., Samollow, P. B., de Silva, D., Zharkikh, A., and Thomas, A. (2006) Comprehensive statistical study of 452 BRCA1 missense substitutions with classification of eight recurrent substitutions as neutral. *J. Med. Genet.* **43,** 295–305 CrossRef Medline
51. Roadmap Epigenomics Consortium, Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M. J., Amin, V., Whitaker, J. W., Schultz, M. D., Ward, L. D., *et al.* (2015) Integrative analysis of 111 reference human epigenomes. *Nature* **518,** 317–330 CrossRef Medline
52. Boyle, A. P., Hong, E. L., Hariharan, M., Cheng, Y., Schaub, M. A., Kasowski, M., Karczewski, K. J., Park, J., Hitz, B. C., Weng, S., Cherry, J. M., and Snyder, M. (2012) Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res.* **22,** 1790–1797 CrossRef Medline