CoV3D: A database of high resolution coronavirus protein structures

Ragul Gowthaman[1,2], Johnathan D. Guest[1,2], Rui Yin[1,2], Jared Adolf-Bryfogle[3,4,5,6,7], William R. Schief[3,4,8], and Brian G. Pierce[1,2,*]

[1]University of Maryland Institute for Bioscience and Biotechnology Research, Rockville, MD 20850, USA

[2]Department of Cell Biology and Molecular Genetics, University of Maryland, College Park, MD 20742, USA

[3]Department of Immunology and Microbiology, The Scripps Research Institute, La Jolla, CA 92037, USA

[4]IAVI Neutralizing Antibody Center, The Scripps Research Institute, La Jolla, CA 92037, USA

[5]Institute for Protein Innovation, Boston, MA 02115, USA

[6]Division of Hematology/Oncology, Boston Children's Hospital, Boston, MA 02115, USA

[7]Department of Pediatrics, Harvard Medical School, Boston, MA 02115, USA

[8]The Ragon Institute of Massachusetts General Hospital, Massachusetts Institute of Technology, and Harvard University, Cambridge, MA 02139, USA

*To whom correspondence should be addressed. Email: pierce@umd.edu,

Abstract

The SARS-CoV-2 virus is the cause of the current COVID-19 pandemic, and exemplifies the general threat to global health posed by coronaviruses. The urgent need for effective vaccines and therapies is leading to a rapid rise in the number of high resolution structures of SARS-CoV-2 proteins that collectively reveal a map of virus vulnerabilities. To assist structure-based design of vaccines and therapeutics against SARS-CoV-2 and other coronaviruses, we developed CoV3D, a database and molecular viewer of SARS-CoV-2 and other coronavirus protein structures updated weekly. CoV3D annotates structures of coronavirus proteins and their complexes with antibodies, receptors, and small molecules. Additionally, CoV3D provides information on spike glycoprotein sequence variability and polymorphisms, and maps these features onto the spike structure in the integrated molecular viewer. In order to further aid structure-based design and analysis, CoV3D includes viewable and downloadable spike glycoprotein structures with modeled glycosylation from Rosetta. CoV3D is available at: https://cov3d.ibbr.umd.edu.

Introduction

Coronaviruses (CoVs) have been responsible for several outbreaks over the past two decades, including SARS-CoV in 2002-2003, MERS-CoV in 2012 [1], and the current COVID-19 pandemic, caused by SARS-CoV-2, which began in late 2019 [2]. The scale of the COVID-19 pandemic has led to unprecedented efforts by the research community to rapidly identify and test therapeutics and vaccines, and to understand the molecular basis of SARS-CoV-2 entry, pathogenesis, and immune targeting.

From February to May 2020, a large number of SARS-CoV-2 protein structures were released in the Protein Data Bank (PDB) [3]: 20 spike glycoprotein structures, over 100 main protease structures, and over 50 structures of other SARS-CoV-2 proteins, including nucleocapsid and RNA dependent RNA polymerase, the target of the drug remdesivir. These high-resolution protein structures are of immense importance for understanding viral assembly and to aid rational vaccine and therapeutic design. The first structures of the SARS-CoV-2 trimeric spike glycoproteins (the major target of SARS-CoV-2 vaccines and antibody therapeutics) were reported in February and early March 2020 [4, 5]. Previously determined spike glycoprotein structures have enabled advances including rational stability optimization of SARS-CoV and MERS-CoV spikes, yielding improved protein expression and immunogenicity [6]. Given that the rapid rate of coronavirus protein structural determination and deposition is likely to continue, a simple and updated resource detailing these structures would provide a useful reference.

Here we describe a new database of experimentally determined coronavirus protein structures, CoV3D. CoV3D is updated automatically on a weekly basis, as new structures are released in the

PDB. Structures are classified by CoV protein, as well as bound molecule, such as monoclonal antibody, receptor, and small molecule ligand. To enable insights into the spike glycoprotein, we also include information on SARS-CoV-2 residue polymorphisms, overall coronavirus sequence diversity of betacoronaviruses mapped onto spike glycoprotein structures, and structures of spike glycoproteins with Rosetta modeled glycans. This resource can aid in efforts for rational vaccine design, targeting by immunotherapies, biologics, and small molecules, and basic research into coronavirus structure and recognition. CoV3D is publicly available at https://cov3d.ibbr.umd.edu.

Results

Database contents

The main components of the CoV3D database are interrelated tables, datasets, and tools for coronavirus protein structures and spike glycoprotein sequences. A schematic of the CoV3D input, organization, and contents is shown in **Figure 1**. The structure portion of the database includes dedicated pages and tables for:

- Spike glycoprotein structures
- Main protease structures
- Other SARS-CoV-2 protein structures
- Peptide-Major Histocompatibility Complex (MHC) structures
- Spike structures with modeled glycans from Rosetta

Spike glycoprotein structures are annotated by binding partner(s), including bound antibody and receptor, as well as domains present in the structure. Browser-based viewers are available in CoV3D to view individual structures and multiple spike complex structures (multiple complex

viewer is shown in **Figure 2A**), polymorphic sites on the SARS-CoV-2 spike (examples in **Figure 3**), and spike structures with modeled glycosylation (examples in **Figure 4**). Main protease structures include annotation of bound ligand, for convenience of those investigating protease inhibitors computationally or experimentally. The sequence portion of CoV3D focuses on spike glycoprotein variability, and includes:

- Spike glycoprotein residue polymorphisms
- Spike glycoprotein sequence logos based on sets of betacoronavirus sequences (example in **Figure 3C**)
- Visualization of conservation on spike glycoprotein structures

Additional features of the database include a page with recently released structures, and a downloads page for users to download structural data, sequence data, and tables. Summary statistics are also given, providing metrics such as the number of spike structures released for SARS-CoV, MERS-CoV, and SARS-CoV-2.

Example usage: viewing sarbecovirus conservation in an RBD antibody epitope

In addition to a table of all known structures of CoV spike proteins and their interactions, CoV3D provides a viewer with SARS-CoV-2 spike RBD interactions with antibodies and the ACE2 receptor superposed by RBD in a common reference frame (**Figure 2A**). This viewer permits users to assess overlapping or shared RBD binding modes among antibodies and ACE2. One antibody of interest is S309, which is a human antibody that was cloned from an individual who was infected with SARS-CoV in 2003, and potently neutralizes SARS-CoV-2 [7]. Comparing its RBD binding with ACE2 in the RBD interaction viewer (**Figure 2A**) highlights how S309 engages the RBD at a distinct site from ACE2, in accordance with their lack of

observed RBD binding competition [7]. The epitope targeted by S309 is dominated by RBD residues 334-346 (circled in **Figure 2A**). On the CoV3D sequence logo generator page, users can enter spike residue ranges, and generate sequence logos that represent positional residue propensities and conservation for regions of interest, based on sets of CoV spike sequences. A sequence logo generated from a reference set of spike proteins from 18 SARS-like CoVs (**Figure 2B**) is shown in **Figure 2C**, indicating the conservation of the N-glycosylation site (NxS/NxT sequon) at position 343, and amino acid variability at position 340, which is glutamic acid (E) in SARS-CoV-2. Both the N343 glycan and the E340 side chain are directly engaged by S309 in the recently determined complex structure [7].

Example usage: location of polymorphisms

CoV3D provides easily accessible information on identified SARS-CoV-2 spike polymorphisms, and their mapping onto the spike glycoprotein structures. Under “Sequences”, users can navigate to “Spike Polymorphisms” where a table is shown with observed single and multiple substitutions in the spike glycoprotein, along with the counts of sequences containing them. The D614G variant is currently the most prevalent, with more occurrences than the reference sequence that is represented in current spike glycoprotein structures (1553 sequences, versus 920 sequences, as of May 26, 2020). Viewing the position of this substitution on the spike glycoprotein can provide some indication of its possible structural and functional impact. By clicking “view” next to this substitution, users can visualize the position of this substitution, showing that it is located on the spike surface, and is not located within or adjacent to the receptor binding domain (RBD; **Figure 3A**). However, this site is closer to the RBD than the site of another less prevalent SARS-CoV-2 spike variant (T791I; **Figure 3B**).

Example usage: visualizing modeled glycosylation

N-glycosylation of viral glycoproteins can play a key role by masking the glycoprotein from the immune system, or affecting function [8]. Experimentally reported protein structures often lack N-glycans, or large portions thereof, due to limitations from resolution or intrinsic glycan dynamics or heterogeneity. To enable visualization and additional analysis or modeling of glycosylated spike glycoproteins, CoV3D includes sets of structures with modeled N-glycans at all predicted glycosylation sites, with N-glycans built onto the glycoprotein structures and refined using new tools in Rosetta based on RosettaCarbohydrates [9]. Examples of glycosylated structures that can be visualized in CoV3D are shown in **Figure 4**, and these can be downloaded directly by users for further processing. This permits users to view features such as the N-glycosylation present on the ACE2 surface (**Figure 4A**) and the relatively high glycosylation of the spike glycoprotein stem (bottom in **Figure 4B**). Presently, these structures include uniform oligomannose glycoforms that were found to be prevalent on the SARS-CoV spike based on previous mass spectroscopy analysis [10], with five branched mannose sugars. In the future, we plan to include more spike structures with modeled glycans, and to include glycoforms that reflect recent mass spectroscopy experimental characterization of the SARS-CoV-2 spike [11].

Discussion

We have constructed the CoV3D database as a reference for the research community, providing a simple and updated interface to all coronavirus 3D structures, with integrated molecular viewers and other useful features. This will allow researchers to identify and analyze new coronavirus protein structures as they are released, particularly for SARS-CoV-2, while allowing

insights into sequence features, polymorphisms, as well as glycosylation. A recent study combining coronavirus structural and sequence analysis revealed insights regarding the determinants of ACE2 recognition [12], and recent SARS-CoV-2 spike-antibody complex structures have shown mechanisms of cross-reactive CoV spike targeting and neutralization [7, 13, 14]. Importantly, as new spike-antibody structures continue to delineate determinants of effective immune recognition of the spike glycoprotein, we will update our multi-structure viewer to include those structures to facilitate comparison with previously determined ones, and will explore adding annotations including structure-based classification of spike-antibody interactions based on their features. This database should be helpful to virologists, computational biologists, immunologists, and those interested in learning about and targeting SARS-CoV-2 proteins with small molecules and antibodies, as well as those engaged in vaccine design.

Methods

Web and database implementation

CoV3D is implemented using the Flask web framework (https://flask.palletsprojects.com/) and the SQLite database engine (https://www.sqlite.org/).

Structure identification, visualization, and glycan modeling

Structures are identified from the PDB on a weekly basis using NCBI BLAST command line tools [15], with coronavirus protein reference sequences from SARS-CoV, MERS-CoV, and SARS-CoV-2 as queries. The spike glycoprotein reference sequences (GenBank identification NP_828851.1, YP_009047204.1 and QHD43416.1 for SARS-CoV, MERS-CoV and SARS-CoV-2 virus respectively) are used as queries to identify all available spike glycoprotein

structures. Peptide-MHC structures containing coronavirus peptides are identified in the PDB through semi-manual searches of the PDB site and literature, though future automated updates are planned in conjunction with an expanded version of the TCR3d database [16]. Structural visualization is performed using NGL viewer [17]. N-glycans are modeled onto spike glycoprotein structures using a glycan modeling and refinement protocol in Rosetta [9]. An example command line and Rosetta Script for this glycan modeling protocol is provided as Supplemental Information.

Sequence data

SARS-CoV-2 spike glycoprotein sequences and sequence information were downloaded from NCBI Virus [18], followed by filtering out sequences with missing residues. Sequence polymorphism information was obtained by BLAST search using a reference SARS-CoV-2 spike glycoprotein sequence (QHD43416.1). To develop spike glycoprotein alignments, betacoronavirus spike glycoprotein sequences were downloaded from NCBI Virus [18] and aligned with Clustal Omega [19] in SeaView [20]. Sequences that were redundant (>95% similarity) or contained missing residues were removed, with the remaining 70 sequences forming the pan-betacoronavirus alignment. A subset of 18 sequences from the pan-betacoronavirus alignment was used to generate the SARS-like sequence alignment, which contains every sequence from the pan-betacoronavirus alignment with >70% sequence similarity to the SARS-CoV-2 spike. Sequence logos are generated dynamically for user-specified residue ranges using the command-line version of WebLogo [21]. Phylogenetic trees representing spike reference sequences were generated using ClustalX [22], and visualized using the APE package [23] in R (www.r-project.org).

Acknowledgements

We are grateful to the structural biologists and researchers whose work resulted in the structures enabling the development of this database. We are additionally thankful to Ghazaleh Taherzadeh and Stefan Ivanov (University of Maryland Institute for Bioscience and Biotechnology Research), as well as the RosettaCommons community, for helpful discussions and suggestions. We also thank John Moult (University of Maryland Institute for Bioscience and Biotechnology Research) for useful comments regarding the non-spike SARS-CoV-2 structures. The Institute for Bioscience and Biotechnology Research computing facility and staff, including Christian Presley, provided resources and assistance with server implementation and web hosting. This work was supported in part by NIH R01 GM126299 (to BGP).

Author contributions

B.G.P., R.G., J.D.G., and R.Y. conceived, designed, and conducted the study. J.A.B. and W.R.S. contributed software for the database. B.G.P. and R.Y. wrote the manuscript, and all authors participated in editing the manuscript.

REFERENCES

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

[2] Tse LV, Meganck RM, Graham RL, Baric RS. The Current and Future State of Vaccines, Antivirals and Gene Therapies Against Emerging Coronaviruses. Frontiers in microbiology. 2020;11:658.

[3] Rose PW, Beran B, Bi C, Bluhm WF, Dimitropoulos D, Goodsell DS, et al. The RCSB Protein Data Bank: redesigned web site and web services. Nucleic acids research. 2011;39:D392-401.

[4] Wrapp D, Wang N, Corbett KS, Goldsmith JA, Hsieh CL, Abiona O, et al. Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. Science. 2020;367:1260-3.

[5] Walls AC, Park YJ, Tortorici MA, Wall A, McGuire AT, Veesler D. Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein. Cell. 2020;181:281-92 e6.

[6] Pallesen J, Wang N, Corbett KS, Wrapp D, Kirchdoerfer RN, Turner HL, et al. Immunogenicity and structures of a rationally designed prefusion MERS-CoV spike antigen. Proc Natl Acad Sci U S A. 2017;114:E7348-E57.

[7] Pinto D, Park YJ, Beltramello M, Walls AC, Tortorici MA, Bianchi S, et al. Cross-neutralization of SARS-CoV-2 by a human monoclonal SARS-CoV antibody. Nature. 2020.

[8] Lavie M, Hanoulle X, Dubuisson J. Glycan Shielding and Modulation of Hepatitis C Virus Neutralizing Antibodies. Front Immunol. 2018;9:910.

[9] Labonte JW, Adolf-Bryfogle J, Schief WR, Gray JJ. Residue-centric modeling and design of saccharide and glycoconjugate structures. J Comput Chem. 2017;38:276-87.

[10] Ritchie G, Harvey DJ, Feldmann F, Stroeher U, Feldmann H, Royle L, et al. Identification of N-linked carbohydrates from severe acute respiratory syndrome (SARS) spike glycoprotein. Virology. 2010;399:257-69.

[11] Watanabe Y, Allen JD, Wrapp D, McLellan JS, Crispin M. Site-specific glycan analysis of the SARS-CoV-2 spike. Science. 2020.

[12] Wan Y, Shang J, Graham R, Baric RS, Li F. Receptor Recognition by the Novel Coronavirus from Wuhan: an Analysis Based on Decade-Long Structural Studies of SARS Coronavirus. J Virol. 2020;94.

[13] Yuan M, Wu NC, Zhu X, Lee CD, So RTY, Lv H, et al. A highly conserved cryptic epitope in the receptor binding domains of SARS-CoV-2 and SARS-CoV. Science. 2020;368:630-3.

[14] Wu Y, Wang F, Shen C, Peng W, Li D, Zhao C, et al. A noncompeting pair of human neutralizing antibodies block COVID-19 virus binding to its receptor ACE2. Science. 2020.

[15] Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC bioinformatics. 2009;10:421.

[16] Gowthaman R, Pierce BG. TCR3d: The T cell receptor structural repertoire database. Bioinformatics. 2019;35:5323-5.

[17] Rose AS, Hildebrand PW. NGL Viewer: a web application for molecular visualization. Nucleic Acids Res. 2015;43:W576-9.

[18] Hatcher EL, Zhdanov SA, Bao Y, Blinkova O, Nawrocki EP, Ostapchuck Y, et al. Virus Variation Resource - improved response to emergent viral outbreaks. Nucleic Acids Res. 2017;45:D482-D90.

[19] Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol. 2011;7:539.

[20] Gouy M, Guindon S, Gascuel O. SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. Mol Biol Evol. 2010;27:221-4.

[21] Crooks GE, Hon G, Chandonia JM, Brenner SE. WebLogo: a sequence logo generator. Genome Res. 2004;14:1188-90.

[22] Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, et al. Clustal W and Clustal X version 2.0. Bioinformatics. 2007;23:2947-8.

[23] Paradis E, Claude J, Strimmer K. APE: Analyses of Phylogenetics and Evolution in R language. Bioinformatics. 2004;20:289-90.

[24] Liu J, Sun Y, Qi J, Chu F, Wu H, Gao F, et al. The membrane protein of severe acute respiratory syndrome coronavirus acts as a dominant immunogen revealed by a clustering region of novel functionally and structurally defined cytotoxic T-lymphocyte epitopes. The Journal of infectious diseases. 2010;202:1171-80.

[25] Wang Q, Zhang Y, Wu L, Niu S, Song C, Zhang Z, et al. Structural and Functional Basis of SARS-CoV-2 Entry by Using Human ACE2. Cell. 2020.

FIGURE LEGENDS

Figure 1. Implementation and organization of CoV3D. The database combines curated and annotated CoV structural data from the PDB, updated automatically on a weekly basis, as well as CoV sequence data from NCBI. Structures include immunologically relevant complexes; shown are an antibody in complex with the SARS-CoV-2 spike receptor binding domain (PDB code 7BZ5) [14], and a SARS-CoV membrane protein peptide in complex with the human major histocompatibility complex (MHC) protein HLA-A2 (PDB code 3I6K) [24]. Also shown is a sequence logo generated on CoV3D for a spike glycoprotein subsequence.

Figure 2. Viewing the structure, targeting, and sequence diversity of SARS-CoV-2 spike receptor binding domain (RBD). (A) Visualization of the spike RBD complexes with antibody S309 (PDB code 6WPS) [7] and ACE2 (PDB code 6LZG) [25] in CoV3D, with spike colored magenta, S309 cyan, ACE2 green, and an interacting region of interest on the RBD (residues 334-346) circled for reference. For clarity, only the S309-RBD region of the S309-spike complex structure is shown. (B) Unrooted phylogenetic tree generated using 18 reference spike sequences from SARS-like CoVs used in the CoV3D sequence logo generator. Figure generated using the APE package in R [23], and scale bar is shown in lower left, representing 5% sequence divergence. (C) CoV3D sequence logo generation interface, and logo for spike residues 334-346. The logo was generated using 18 SARS-related spike glycoprotein reference sequences and the command-line version of WebLogo [21].

Figure 3. Visualization of SARS-CoV-2 spike polymorphisms in CoV3D. Shown are (A) residue 614, the site of the D614G substitution, and (B) residue 791, site of the T791I

substitution. Polymorphisms are shown in the context of a prefusion spike structure [4] (PDB code 6VSB) (gray), with sites in spacefill representations of the wild-type residue on each spike trimer subunit, colored by atom type (cyan: carbon, blue: nitrogen, red: oxygen). The receptor binding domains (residues 331-529) are shown in magenta, and red dashed circles denote mutation sites, for reference.

Figure 4. Spike structures with modeled oligomannose N-glycans. (A) The SARS-CoV-2 spike RBD-ACE2 complex [25] (PDB code 6LZG) with RBD and ACE2 shown as red and blue cartoons, respectively, with modeled N-glycans shown as gray, red, and blue spheres. One N-glycan is modeled on the spike RBD and six are modeled on ACE2. (B) A trimeric spike structure in RBD-closed conformation [5] (PDB code 6VXX) with spike shown as gray surface and modeled N-glycans shown as gray, red, and blue spheres. 48 modeled N-glycans are present on the spike structure.

Figure 1.

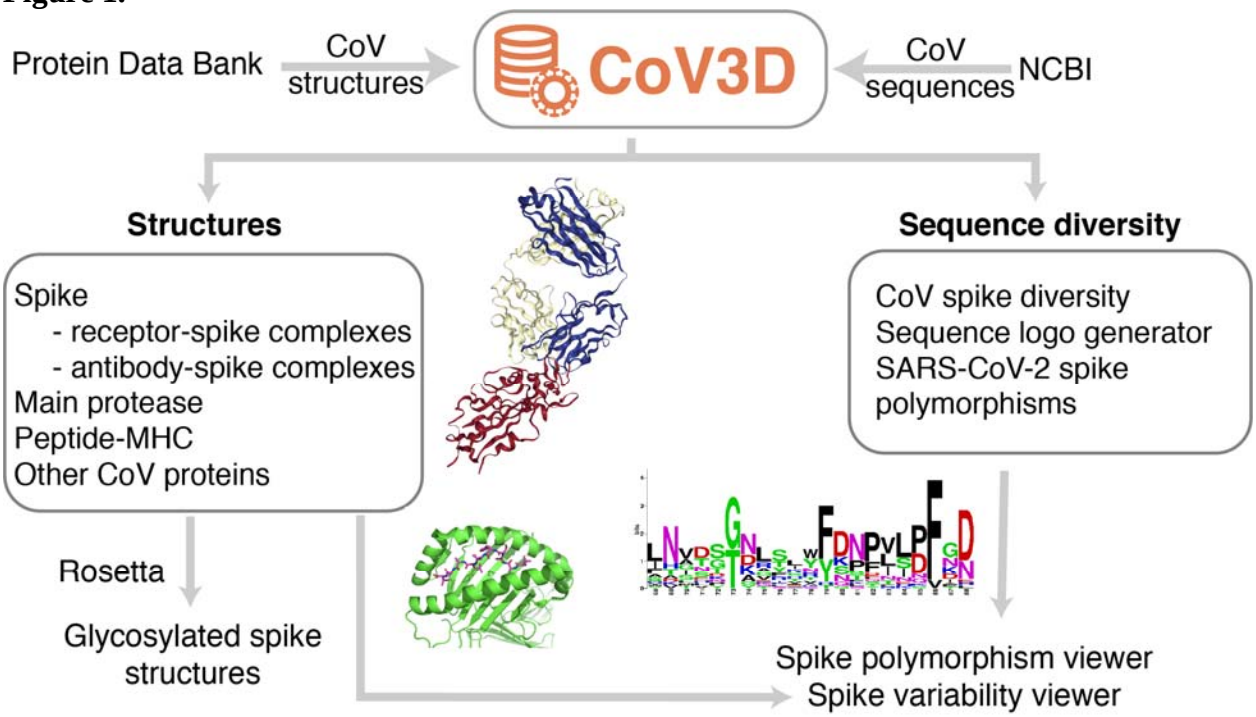

Figure 2.

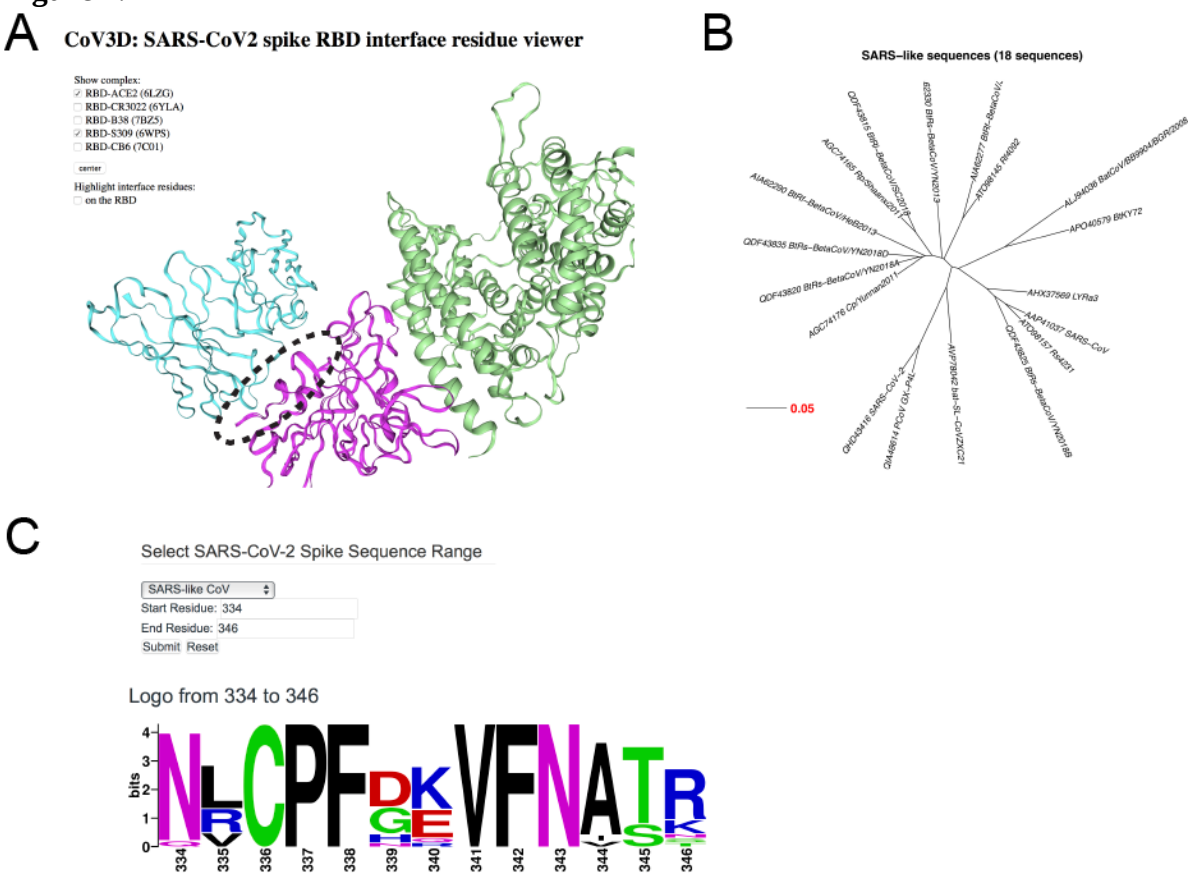

Figure 3.

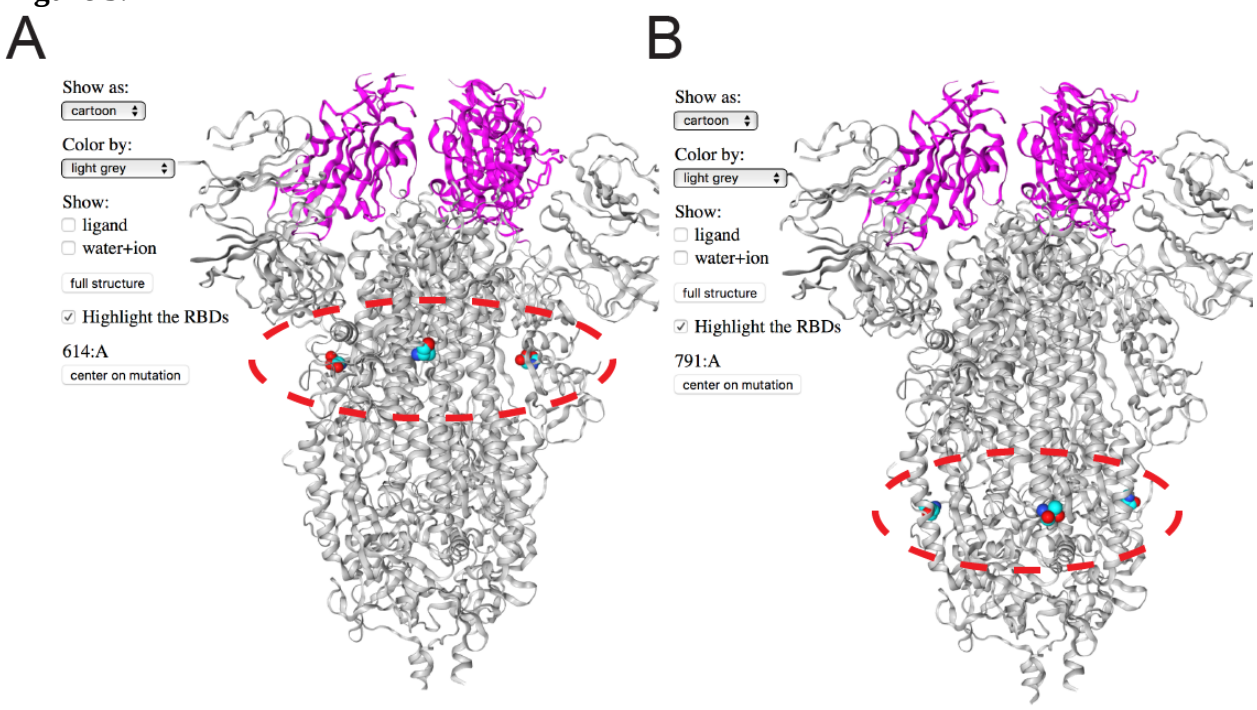

Figure 4.