# Exploring potential super infection in SARS-CoV2 by genome-wide analysis and receptor–ligand docking

Chuanjun Shu*[1,2], Xuan Huang*[3],Juergen Brosius[4], Cheng Deng[2,#]

[1]Department of Bioinformatics, School of Biomedical Engineering and Informatics, Nanjing Medical University, Nanjing, 211166, China.

[2]Jiangsu Key Laboratory for Biodiversity and Biotechnology, College of Life Sciences, Nanjing Normal University, Nanjing, 210023, China.

[3]Reproductive Medical Center, Jinling Hospital Affiliated to The Medical School of Nanjing University, Nanjing, 210002, China.

[4]Institutes for Systems Genetics, West China Hospital, Sichuan University, Chengdu, 610041, China.

*Contributed equally.

[#]corresponding information: Cheng Deng, dengcheng@njnu.edu.cn.

## Abstract

SARS-CoV2 (corona virus) has spread globally at an unprecedented rate; so far, increasing SARS-CoV2-infected individuals have been identified. Although the situation in China is improving and is currently under control, the outbreak in other countries and its pandemic management is only beginning to develop. Based on 154 SARS-CoV2 genome sequence analyses, we used receptor–ligand docking to identify one potential point mutation (V354F) on the spike structure which enhances spike binding to ACE2 receptors underlying potential super infection. Importantly, the V354F site on spike S1 had been identified in 5/10 infected French patients living in Paris, who sharing 100% identical SARS-CoV2 genomes. With Covid-19 cases increasing rapidly in France that could lead to a new explosion, we suggest that the French government should identify all potential super spreaders and treat them accordingly. In summary, our study provides on of the measures to avoid the potential second worldwide explosion of SARS-CoV2.

**Key words：** SARS-CoV2, spike, receptor–ligand docking, super infection

# Main text

In December 2019, a pneumonia outbreak associated with the 2019 novel coronavirus (2019-nCoV, also named as SARS-CoV2) occurred in Wuhan, Hubei Province, China (Pan et al., 2020; D. Wang et al., 2020; Zhu et al., 2020a, 2020b). Subsequently, the new coronavirus pneumonia (NCP) - termed COVID-19 has rapidly spread around the world(Guan et al., 2020; Huang et al., 2020; Wu, Leung, & Leung, 2020). Up to now-March 18, 2020, 198,032 (81,151-China and 116,881-other counties) have been identified(Guan et al., 2020; D. Wang et al., 2020). Through the coordinated efforts, the situation tends to improve and be under control in China. Thus far, more than half of the patients in China have been recovered and the total number of patients continues to decrease. However, the potential danger of a second worldwide spread is still imminent.

Currently, there are no effective therapeutics for treatment of the SARS-CoV2 available. SARS-CoV2 is a single-stranded RNA beta-coronavirus(Woo et al., 2005; A. Wu et al., 2020). The SARS-CoV2 genome encodes non-structural proteins (such as papain-like protease, 3-chymotrypsin-like protease (also known as main protease Mpro), RNA-dependent RNA polymerase, and helicase), structural proteins (such as spike glycoprotein, membrane protein, envelope protein, and nucleocapsid protein) and accessory proteins (such as ORF3a, ORF8)(A. Wu et al., 2020). The four non-structural proteins are critical enzymes for the viral life cycle(Shimamoto et al., 2016). Spike binds to the human cellular receptor angiotensin-converting enzyme 2 (ACE2) and mediates fusion between the viral and cellular membranes(Wong & K., 2003). The receptor binding domain (RBD) for ACE2 is distributed in the S1 region of spike (spike-S1)(Kirchdoerfer et al., 2016) (S. Wang et al., 2008). These five proteins were recognized as attractive targets to develop antiviral agents against human coronavirus, such as SARS-CoV2, SARS and MERS. Therefore, amino acid substitutions in SARS-CoV2 could bear a major challenge for outbreak prevention and treatment. Especially, amino acid substitutions in spike-S1 have the potential for generating SARS-CoV2 variants with super infection ability. In this study, our

objectives were to investigate potential super infection based on evolution analysis of SARS-CoV2 and structural pharmacological analysis of receptor-ligand (ACE2-Spike) binding ability.

We compared 154 complete genomes of SARS-CoV2 from the GISAID EpiFlu™ database (access date 22 February 2020) and CNCB/BIG (https://bigd.big.ac.cn/ncov, access date 22 February 2020) to explore variations of SARS-CoV2. The aligned genome size of SARS-CoV2 is 29,674 bp, and is consisted of ten coding sequence (CDS) regions (Fig 1A). The SARS-CoV2-Wuhan01 virus (GenBank reference ID: MN908947.3) was firstly compared to bat-CoV-RaTG13 (GISAID ID: EPI_ISL_402131), pangolin-CoV-Guangxi (GISAID ID: EPI_ISL_410538), and pangolin-CoV-Guangdong (GISAID ID: EPI_ISL_410721) for exploring the potential origin of SARS-CoV2 before it infected a human. Compared with these three coronaviruses, sequence similarities of whole genome for SARS-CoV2-Wuhan01 variation from 0.85 to 0.96 (bat-CoV-RaTG13: 0.96, pangolin-CoV-Guangxi: 0.85, pangolin-CoV-Guangdong: 0.90) is larger than that between SARS-CoV2 and SARS (~0.80). There were 360/755, 1215/3003, and 804/2033 synonymous/non-synonymous substitution sites for bat-CoV-RaTG13, pangolin-CoV-Guangxi, and pangolin-CoV-Guangdong, respectively, when compared to SARS-CoV2-Wuhan01. As shown in Fig 1B, non-synonymous substitution sites are mainly distributed in the ORF1ab and spike regions. These two regions have a higher substitution rate compared to other regions (Fig 1B). These results indicated that bat-CoV is more closely related to SARS-CoV2 than to pangolin-CoV, and that pangolin could not be the only intermediate host. Furthermore, to explore the evolution and transmissions of SARS-CoV2 in the recent three months, the genomic variations of SARS-CoV2 were also calculated. Substitution sites in five proteins, i.e., 3-chymotrypsin-like protease (Mpro), papain-like protease, helicase, RNA-dependent RNA polymerase, and spike-S1, were calculated, since they were considered to be the potential drug targets to develop anti-viral agents against SARS and MERS. Subsequently, we identified 138 substitution sites (synonymous sites: 52, non-synonymous sites: 86) in ten CDS regions, and all substitutions can be classified

as single nucleotide polymorphism. The substitution sites were mainly distributed in the ORF1ab and spike regions (Fig 1C). But ORF8a, ORF3a, ORF10, and spike-S1 had a higher substitution speed in SARS-CoV2 (Fig 1D). With the ratio of non-synonymous sites/synonymous sites >1, our results indicated that SARS-CoV2 probably underwent adaptive evolution after infection of human hosts.

These non-synonymous substitutions could result in a SARS-CoV2 with super infectivity. To address potential super infection in SARS-CoV2, amino acid mutations caused by non-synonymous substitutions of SARS-CoV2-Wuhan01 against SARS-CoV2 mutants were analyzed and the corresponding amino acids in bat-CoV-RaTG13, pangolin-CoV, and SARS viruses were then used as controls. As shown in Fig 1E, most amino acid substitutions between SARS-CoV2 viruses always are conserved residues in bat-CoV-RaTG13, pangolin-CoV, and SARS viruses, except amino acid substitutions in spike-S1. This suggested that most amino acid substitutions probably are random events. However, 10 amino acid substitutions recurred in SARS-CoV2, i.e. I789V of papain-like protease, H36Y and V354F of spike-S1, G251V of ORF3a, D209H of membrane glycoprotein, V62L, L84S, and P85S of ORF8, and S194L, S202N, and P344S of nucleocapsid phosphoprotein (Fig 1E). Spike protein, binds to the human cellular receptor-ACE2 and responsible for virus infection, and according to WHO data, the basic reproduction number of the infection ($R_0$) for SARS-CoV2 is about 1.4 to 2.5. Hence, we should pay more attention to amino acids substitutions occurring in RBD region of spike-S1 or substitutions repeated more than three times. In addition, epidemiology and evolution data for H36Y, N341D, D351Y, V354F of spike-S1 and G251V of ORF3a, V62L and L84S of ORF8, S194L, S202N of nucleocapsid phosphoprotein were analyzed. The amino acid substitution of ORF8, i.e., L84S, had the highest substituted times (value=36) among the 10 mutant residues (Fig 1E). However, this site is the reverse mutation compared to its ancestor (Fig 1E). We selected the whole genome of these 36 SARS-CoV2 viruses and their ancestors (bat-CoV and pangolin-CoV viruses) to construct a phylogenetic tree (neighbor-joining, bootstrap=500) (Fig 1F). As shown in Fig 1F, we found that the SARS-CoV2 mutant with L84S in ORF8 probably had a

closer phylogenetic relationship between this mutant and its ancestors than that of SARS-CoV2-Wuhan01 (Fig 1F). According to these phylogenetic results, we proposed a hypothesis that Wuhan, China is an early substitutions site of the SARS-CoV2 outbreak but it is probably not the original site for the human infection. Hence, this SARS-CoV2 mutant-L84S could be a more ancestral virus rather than a virus with super infectivity.

Another SARS-CoV2 mutant with high frequency of occurrence, i.e. G251V of ORF3a, showed similar result (Fig 1E). The amino acid substitution of ORF3a, i.e., G251V, had the second highest substitution times (value=17) in the 10 mutant residues (Fig 1E). These SARS-CoV mutants were mainly isolated from patients of France and Singapore. An interesting observation is that five patients from France with 100% identical protein sequences (Supplementary materials Table 1) shared the same mutation - V354F and this amino acid substitution was in the RBD region of spike-S1 (V354F). Based on the recorded epidemiology data, only one patient came from Wuhan, China. Furthermore, these five patients with V354F lived in the same area (Location: Europe / France / Ile-de-France / Paris). The *Ro* (basic reproduction number) of the SARS-CoV2 mutant that was isolated from France probably reached 4. This *Ro* was far beyond the WHO predicted value of SARS-CoV2. Taken together, these results suggested that the SARS-CoV2 mutant which has two mutations (G251V of ORF3a and V354F of spike-S1) probably exhibits a super infectivity.

Since the function of coronavirus ORF3a is still unknown, we explored whether it might contribute to super infectivity of SARS-CoV2 mutants based on receptor–ligand (ACE2-spike) docking. As shown in Fig 2A and B, amino acid substitution in the RBD region of spike-S1 theoretically could affect the infection ability of SARS-CoV2. Therefore, three mutants found in spike-S1 (N341D, D351Y, and V354F) were utilized to investigate docking to ACE2. Based on sequence, physical and chemical parameters, i.e. theoretical pI (isoelectric point), GRAVY (grand average of hydropathicity), and instability index, for wild type and mutants were analyzed by using ProtParam tool (https://web.expasy.org/protparam/). As shown in Fig 2C, we found that these three amino acid substitutions could affect the physical

and chemical parameters of spike protein, and the mutant V354F had the smallest impact among these three mutants. We further tested whether these three mutations can result in different interaction models between ACE2 and spike. With the three-dimensional structure of ACE2 downloaded from the PDB database (PDB ID: 6acj, structures of wild type spike-S1 and its mutants were obtained by utilizing I-TASSER (Iterative Threading Assembly Refinement) algorithm. Complexes of ACE2-spike were obtained by a combination of using Discovery studio and Rosetta package. As shown in Fig2D, V354F had the lowest energy and interaction score in these ACE2-spike-S1 complexes, based on structural comparison and protein-protein interaction analysis (Fig 2D, supplement materials Table S1). These results further suggested that spike with the V354F amino acid substitution had a higher affinity for ACE2 than wild type spike and potentially leads to super infectivity. Meanwhile, in France SARS-CoV2 cases accumulate faster than in many other affected countries and could be one of the centers of a new explosion (Fig 2 E). Therefore, SARS-CoV2 mutant-V354F with a potential super infectivity should deserve more attention by government health institutions in France by epidemic surveillance and thus reducing the worldwide threat of the potential second explosion of SARS-CoV2.

In this study, we showed that SARS-CoV2 underwent adaptive evolution during infection from Wuhan patients to non-Wuhan patients (Fig. 1). Although genomic variations are still low for SARS-CoV2, some non-synonymous substitutions probably lead to a SARS-CoV2 variant with super infectivity for humans. For example, SARS-CoV2 with the V354F substitution of spike-S1 probably is a novel SARS-CoV2 with both super infectivity and a high *Ro*. In summary, evolutionary analysis of SARS-CoV2 and structural pharmacological analysis of receptor-ligand (ACE2-Spike) binding ability could aid in monitoring epidemiological situations to prevent a potential second explosion. We strongly suggest that the French government and other counties should identify these potential super spreaders, particularly V354F, and treat them accordingly. Patient material from other countries with potential second explosion, e.g. Korea, Italy, Iran and Japan, should be collected and further investigated by this method.

**Supplementary Material**

Supplementary Materials are available online.

## Methods and Materials

### Data preparation

Whole genome sequences of SARS-CoV2, SARS, bat-CoV, and pangolin-CoV were downloaded from CNCB/BIG (https://bigd.big.ac.cn/ncov), the GISAID EpiFlu™ database and NCBI. According to gene locations, the nucleotide sequences of RBD regions of coronavirus and their corresponding amino acids were acquired by BioEdit software (Helge-Friedrich, 2004). Three-dimensional spatial structure of ACE2 was downloaded from PDB database (Sussman et al., 2010).

### Genetic and phylogenetic analysis

Synonymous/non-synonymous substitution sites between coronavirus were calculated by DnaSP and BioEdit package(Rozas et al., 2017). The nucleotide/amino acid sequences of proteins and the similarities between sequences were aligned by BioEdit software. Unrooted tree topology based on multiple alignments of amino acids was established with the neighbor-joining method in MEGA 6.06 (Lewis, Kumar, Tamura, & Nei, 1995). Consistency of branching was tested using a bootstrap analysis with 500 resamplings of the data in MEGA 6.06.

### Structure analysis

I-TASSER was utilized to construct structure spike-S1 of SARS-CoV2(Roy, Kucukural, & Zhang). We utilized ProtParam tool (https://web.expasy.org/protparam/) to analysis physical and chemical parameters of proteins(Garg, Avashthi, & Tiwari, 2016). The largest possible binding pocket of these proteins (ACE2, spike and its

mutants) were predicted by Discovery Studio 3.0 software(Gao & Huang, 2011). These predicted pockets were utilized to construct an initial coarse model of the spike-S1-ACE2 complexes. The complex structures were refined by Rosetta software (RosettaDock and FelxPepDock module)(Rohl, Strauss, Misura, & Baker, 2003). The final structure was obtained based on energy scores. The interactions scores were calculated by Rosetta. High-quality 3-D images of the proteins were drawn by PyMOL(Ordog, 2008).

## Acknowledgement

We thank Dr. Jürgen Brosius immeasurable help for this project. This work was supported by the National Key Research and Development Program of China (2018YFD0900602) and National Natural Science Foundation of China (31970388, 31701234), the Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD), the Natural Science Foundation of the Jiangsu Higher Education Institutions (17KJB180006), the Natural Science Foundation from Jiangsu Province BK20160043, BK20151546, 15KJA180004 and BK20171035, Jiangsu Distinguished Professor Funding.

## References

Gao, Y. D., & Huang, J. F. (2011). An extension strategy of Discovery Studio 2.0 for non-bonded interaction energy automatic calculation at the residue level. *Zoological Research*.

Garg, V. K., Avashthi, H., & Tiwari, A. (2016). MFPPI – Multi FASTA ProtParam Interface. *12*(2), 74-77.

Guan, W.-j., Ni, Z.-y., Hu, Y., Liang, W.-h., Ou, C.-q., He, J.-x., . . . Zhong, N.-s. (2020). Clinical characteristics of 2019 novel coronavirus infection in China. *medRxiv*, 2020.2002.2006.20020974. doi:10.1101/2020.02.06.20020974

Helge-Friedrich, T. (2004). Analysis for free: Comparing programs for sequence analysis. *Briefings in Bioinformatics*(1), 1.

Huang, C., Wang, Y., Li, X., Ren, L., Zhao, J., Hu, Y., . . . Cao, B. (2020). Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet*. doi:10.1016/S0140-6736(20)30183-5

Kirchdoerfer, R. N., Cottrell, C. A., Wang, N., Pallesen, J., Yassine, H. M., Turner, H. L., . . . Ward, A. B. (2016). Pre-fusion structure of a human coronavirus spike protein. *Nature, 531*(7592), 118-121.

Lewis, P. O., Kumar, S., Tamura, K., & Nei, M. (1995). MEGA: Molecular Evolutionary Genetics Analysis, Version 1.02. *Systematic Biology, 44*(4).

Ordog, R. (2008). PyDeT, a PyMOL plug-in for visualizing geometric concepts around proteins. *Bioinformation, 2*(8), 346-347.

Pan, Y., Guan, H., Zhou, S., Wang, Y., Li, Q., Zhu, T., . . . Xia, L. (2020). Initial CT findings and temporal changes in patients with the novel coronavirus pneumonia (2019-nCoV): a study of 63

patients in Wuhan, China. *Eur Radiol*. doi:10.1007/s00330-020-06731-x

Rohl, C. A., Strauss, C. E., Misura, K. M., & Baker, D. (2003). Protein structure prediction using Rosetta. *383*(383), 66.

Roy, A., Kucukural, A., & Zhang, Y. I-TASSER: a unified platform for automated protein structure and function prediction. *Nature Protocols, 5*(4), 725-738.

Rozas, J., Ferrer-Mata, A., Sánchez-DelBarrio, J. C., Guirao-Rico, S., Librado, P., Ramos-Onsins, S. E., & Sánchez-Gracia, A. (2017). DnaSP 6: DNA Sequence Polymorphism Analysis of Large Data Sets. *Mol Biol Evol, 34*(12), 3299-3302. doi:10.1093/molbev/msx248

Shimamoto, Y., Hattori, Y., Kobayashi, K., Teruya, K., Sanjoh, A., Nakagawa, A., . . . Akaji, K. (2016). Fused-ring structure of decahydroisoquinolin as a novel scaffold for SARS 3CL protease inhibitors. *Bioorg Med Chem, 23*(4), 876-890.

Sussman, J. L., Lin, D., Jiang, J., Manning, N. O., Prilusky, J., Ritter, O., & Abola, E. E. (2010). Protein Data Bank (PDB): Database of Three-Dimensional Structural Information of Biological Macromolecules . *Acta Crystallographica, 54*(6-1), 1078-1084.

Wang, D., Hu, B., Hu, C., Zhu, F., Liu, X., Zhang, J., . . . Peng, Z. (2020). Clinical Characteristics of 138 Hospitalized Patients With 2019 Novel Coronavirus-Infected Pneumonia in Wuhan, China. *JAMA*. doi:10.1001/jama.2020.1585

Wang, S., Guo, F., Liu, K., Wang, H., Rao, S., Yang, P., & Jiang, C. (2008). Endocytosis of the receptor-binding domain of SARS-CoV spike protein together with virus receptor ACE2. *Virus Research, 136*(1-2), 0-15.

Wong, & K., S. (2003). A 193-Amino Acid Fragment of the SARS Coronavirus S Protein Efficiently Binds Angiotensin-converting Enzyme 2. *Journal of Biological Chemistry, 279*(5), 3197-3201.

Woo, P. C., Lau, S. K., Chu, C. M., Chan, K. H., Tsoi, H. W., Huang, Y., . . . Yuen, K. Y. (2005). Characterization and complete genome sequence of a novel coronavirus, coronavirus HKU1, from patients with pneumonia. *J Virol, 79*(2), 884-895. doi:10.1128/JVI.79.2.884-895.2005

Wu, A., Peng, Y., Huang, B., Ding, X., Wang, X., Niu, P., . . . Jiang, T. (2020). Genome Composition and Divergence of the Novel Coronavirus (2019-nCoV) Originating in China. *Cell Host Microbe*. doi:10.1016/j.chom.2020.02.001

Wu, J. T., Leung, K., & Leung, G. M. (2020). Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: a modelling study. *Lancet, 395*(10225), 689-697. doi:10.1016/S0140-6736(20)30260-9

Zhu, N., Zhang, D., Wang, W., Li, X., Yang, B., Song, J., . . . Research, T. (2020a). A Novel Coronavirus from Patients with Pneumonia in China, 2019. *N Engl J Med*. doi:10.1056/NEJMoa2001017

Zhu, N., Zhang, D., Wang, W., Li, X., Yang, B., Song, J., . . . Research, T. (2020b). A Novel Coronavirus from Patients with Pneumonia in China, 2019. *N Engl J Med, 382*(8), 727-733. doi:10.1056/NEJMoa2001017

## Figure legend

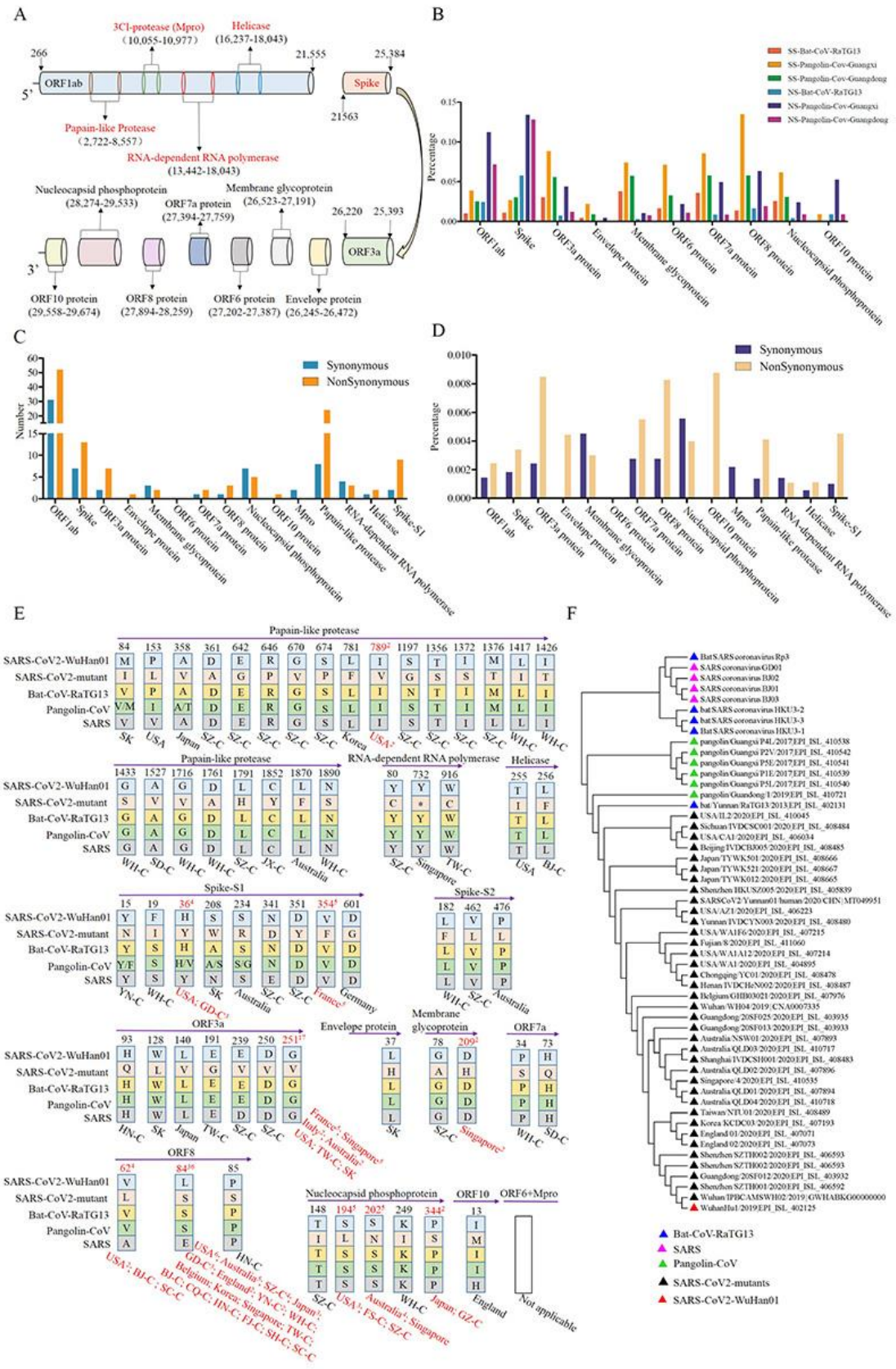


Figure 1. Summary information for 86 non-synonymous sites in ten CDS regions of SARS-CoV2. (A) Genomic organization of SARS-CoV2. Red represents the coding regions for proteins that are potential drug targets. (B) Compared to

SARS-CoV2-Wuhan01, percentages of synonymous/non-synonymous substitution sites in ten CDS regions for bat-CoV-RaTG13, pangolin-CoV-Guangxi, and pangolin-CoV-Guangdong. (C) and (D) represent numbers and percentages of synonymous/non-synonymous substitution sites in ten CDS regions and five potential drug targets for SARS-CoV2, respectively. (E) Amino acid substitutions in SARS-CoV2 mutants, the corresponding residues in bat-CoV-RaTG13, pangolin-CoV, and SARS viruses. Red indicates that amino acid substitution times are greater than 2. (F) The phylogenetic tree of SARS-CoV2 mutants (L84S in ORF8).

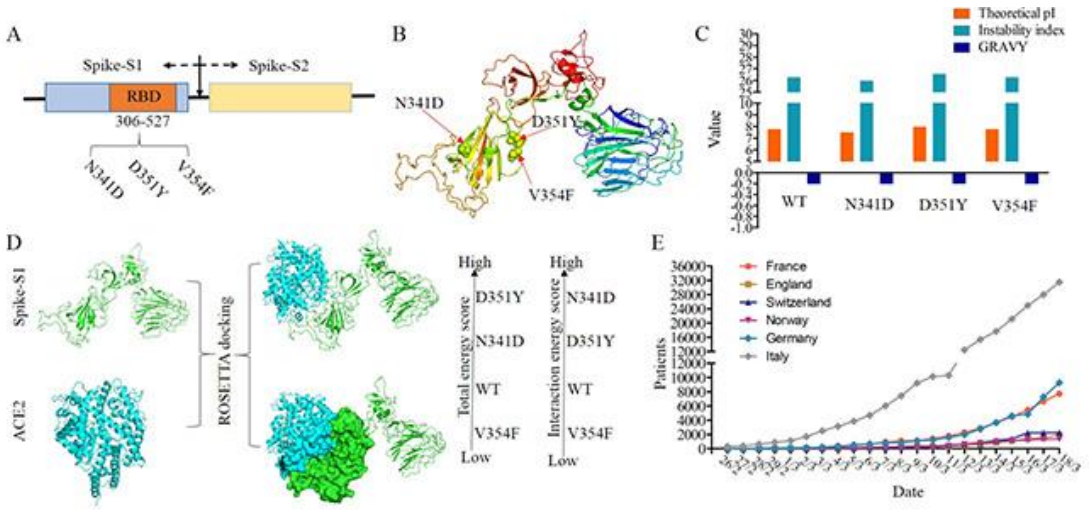


Figure 2. Structural pharmacology analysis among the three spike protein mutants. (A) RMD domain in spike. (B) Distribution of three mutations for spike, N341D, D351Y, and V354F, in its three-dimensional structure. (C) Physical and chemical parameters for spike and its mutants. (D) The energy score and interaction score for complex structure of ACE2-spike-S1 and its mutants. (E) Number of COVID-19 patients in Europe countries.