bioRxiv preprint doi: https://doi.org/10.1101/2020.03.15.991844; this version posted March 23, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

RBD mutations from circulating SARS-CoV-2 strains enhance the structural stability and human ACE2 affinity of the spike protein

Junxian Ou[1†], Zhonghua Zhou[2†], Jing Zhang[3], Wendong Lan[1], Shan Zhao[1], Jianguo Wu[3], Donald Seto[4], Gong Zhang[2]*, Qiwei Zhang[1,3]*

[1] Guangdong Provincial Key Laboratory of Tropical Disease Research, School of Public Health, Southern Medical University, Guangzhou, Guangdong 510515, China

[2] Key Laboratory of Functional Protein Research of Guangdong Higher Education Institutes, Institute of Life and Health Engineering, College of Life Science and Technology, Jinan University, Guangzhou, Guangdong 510632, China.

[3] Guangdong Provincial Key Laboratory of Virology, Institute of Medical Microbiology, Jinan University, Guangzhou, Guangdong 510632, China

[4] Bioinformatics and Computational Biology Program, School of Systems Biology, George Mason University, Manassas, VA 20110, USA

†These authors contributed equally to this work.

***Correspondence**:

Qiwei Zhang, Guangdong Provincial Key Laboratory of Tropical Disease Research, School of Public Health, Southern Medical University, Guangzhou, Guangdong 510515, China, Tel: 86-20-61648649, Fax: 86-20-61648324, Email: zhangqw@smu.edu.cn ;

Gong Zhang, Key Laboratory of Functional Protein Research of Guangdong Higher Education Institutes and MOE Key Laboratory of Tumor Molecular Biology, Institute of Life and Health Engineering, College of Life Science and Technology, Jinan University, Guangzhou, Guangdong 510632, China, Tel: +86-20-85224031 Email: zhanggong@jnu.edu.cn

Running title: RBD mutations enhance the stability and affinity of SARS-CoV-2

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.15.991844; this version posted March 23, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

RBD mutations from circulating SARS-CoV-2 strains enhance the structural stability and human ACE2 affinity of the spike protein

Abstract

A novel coronavirus SARS-CoV-2 is associated with the current global pandemic of Coronavirus Disease 2019 (COVID-19). Bats and pangolins are suspected as the reservoir and the intermediate host. The receptor binding domain (RBD) of the SARS-CoV-2 S protein plays the key role in the tight binding to human receptor ACE2 for viral entry. Here, we analyzed the worldwide RBD mutants and found 18 mutant strains fell into 8 mutation types under high positive selection pressure during the spread. The equilibrium dissociation constant (K_D) of three types of RBD mutants emerging in Wuhan, Shenzhen, Hong Kong and France were two orders of magnitude lower than the prototype Wuhan-Hu-1 strain due to the stabilization of the beta-sheet scaffold of the RBD. This indicated that the mutated viruses may have evolved to acquire remarkably increased infectivity. Five France isolates and one Hong Kong isolate shared the same RBD mutation enhancing the binding affinity, which suggested that they may have originated as a novel sub-lineage. The K_D values for the bat and the pangolin SARS-like CoV RBDs indicated that it would be difficult for bat SARS-like CoV to infect humans; however, the pangolin CoV is potentially infectious to humans with respect to its RBD. These analyses of critical RBD mutations provide further insights into the evolutionary trend of SARS-CoV-2 under high selection pressure. The enhancement of the SARS-CoV-2 binding affinity to ACE2 reveals a possible higher risk of more severe virus transmissions during a sustained pandemic of COVID-19 if no effective precautions are implemented.

Keywords: SARS-CoV-2, ACE2, RBD, spike glycoprotein protein, mutations

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.15.991844; this version posted March 23, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Introduction

A novel coronavirus SARS-CoV-2 has caused the outbreaks of Coronavirus Disease 2019 (COVID-19) all over the world since the first appearance in mid-December 2019 in Wuhan, Central China[1–4]. As of March 20, 2020, SARS-CoV-2 has infected 209,839 people world-wide and caused 8778 deaths with the fatality rate of 4.18%[5]. The pandemic of COVID-19 has been the most serious threat to the public health.

The origin of SARS-CoV-2 remains elusive. However, the initial cases were largely associated with the seafood market, which indicated this were potential zoonotic infections[2]. Although bats and pangolins are most likely the reservoir hosts and the intermediate hosts in the wild, more evidences are in need to support the zoonotic infections and track the origin of this new coronavirus[6–8].

The angiotensin-converting enzyme 2 (ACE2) has been proven to the cellular receptor of SARS-CoV-2, which is the same receptor of SARS-CoV. The spike glycoprotein protein (S) of SARS-CoV-2 recognizes and attaches ACE2 when the viruses infect the cells. S protein consists of a receptor-binding subunit S1 and a membrane-fusion subunit S2. Previous studies revealed that the S1 binds to a receptor on the host cell surface for viral attachment, and S2 fuses the host and viral membranes, allowing viral genomes enter host cells[9–12].

The receptor binding domain (RBD) of the subunit S1 directly interact with ACE2, while the other part of the S protein does not. This RBD alone is sufficient for tight binding to the peptidase domain of ACE2. Therefore, RBD is the critical determinant of virus-receptor interaction and thus of viral host range, tropism and infectivity[9,13,14].

Meanwhile, S protein participates in antigen recognition expressed on its protein surface, likely to be immunogenic as for carrying both T-cell and B-cell epitopes. The potential antibody binding sites that have been identified indicates RBD has important B-cell epitopes. The main antibody binding site substantially overlaps with RBD, and the antibody binding to this surface is likely to block viral entry into cells[15,16].

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.15.991844; this version posted March 23, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

The amino acid mutations and recombination in the RBD of different host origin coronaviruses are deemed to be associated with the host adaption and across species infection. Previous bioinformatic analysis indicated that the shared identity of critical amino acid sites between SARS-CoV-2 and Pangolin CoV might be due to random mutations coupled with natural selection [17]. Recent research found that the recombination and a cleavage site insertion in the RBD might increase the virus infectivity and replication capacity[7]. Although the RBD sequences of different SARS-CoV-2 viruses spreading in the world are conserved, mutations in RBD still appeared, which might relate to the progression of the infectivity of this virus.

To invest whether these mutations in RBD have enhanced or weakened the receptor binding activity and whether the viruses are becoming more infectious and spreading more quickly, we investigated and compared the exact receptor binding dynamics between the SARS-CoV-2 RBDs of all the newly mutated strains and human ACE2 as well as their potential hosts such as bats and pangolins.

Materials and methods

Genome sequence dataset in this study

Full-length protein sequences of S protein RBD were downloaded from the NCBI GenBank Database, China 2019 Novel Coronavirus Resource (https://bigd.big.ac.cn/ncov) and GISAID EpiFluTM Database (http://www.GISAID.org). 662 SARS-CoV-2 and SARS-like CoV full genome sequences were downloaded and the sequences with mutations in S protein and RBD region were screened. The genome sequences used in dynamics analyses are as follow: SARS-CoV-2 (NC_045512.2, EPI_ISL_407071, EPI_ISL_412028, EPI_ISL_411220, EPI_ISL_411219, EPI_ISL_410720, EPI_ISL_406597, EPI_ISL_406596, EPI_ISL_408511, EPI_ISL_406595, EPI_ISL_413522, EPI_ISL_413602, EPI_ISL_415655, EPI_ISL_413650, EPI_ISL_413651, EPI_ISL_415596, EPI_ISL_414618,EPI_ISL_413652, EPI_ISL_415605); Bat SARS-like CoV RaTG13: MN996532; pangolin SARS-like CoV GD 01: EPI_ISL_410721.

Sequences alignment and polymorphism analyses

Alignment of S protein sequences from different sources and comparison of ACE2 proteins among

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.15.991844; this version posted March 23, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

different species were accomplished by MAFFT version 7 online serve with default parameter （https://mafft.cbrc.jp/alignmeloadnt/server/）and Bioedit[18,19]. Polymorphism and divergence were analyzed by DnaSP6 (version 6.12.03)[20]. Analyses were conducted using the Nei-Gojobori model[21]. All positions containing gaps and missing data were eliminated. Evolutionary analyses were conducted in Mega X（version 10.0.2）[22].

Molecular dynamics (MD) simulation

The complex structure of the SARS-CoV-2 S-protein RBD domain and human ACE2 was obtained from Nation Microbiology Data Center (ID: NMDCS0000001) (PDB ID: 6LZG). Mutated amino acids of the SARS-CoV-2 RBD mutants were directly replaced in the model, and the bat/pangolin CoV RBD domain was modelled using SWISS-MODEL[23]. Molecular dynamics simulation was performed using GROMACS 2019 with the following options and parameters: explicit solvent model, system temperature 37℃, OPLS/AA all-atoms force field, LINCS restraints. With 2fs steps, each simulation was performed 10ns, and each model was simulated 3 times to generate 3 independent trajectory replications. Binding free energy (ΔG) was calculated using MM-PBSA method (software downloaded from GitHub: https://github.com/Jerkwin/gmxtool) with the trajectories after structural equilibrium assessed using RMSD (Root Mean Square Deviation)[24]. The formula $\Delta G = RTlnK_D$ was used to calculate between equilibrium dissociation constant (K_D) and ΔG. The estimated ΔG of the RBD mutants were normalized using the ΔG of the prototype strain which was derived from experimental data[10].

Results

The profile of SARS-CoV-2 S protein RBD mapping the mutants

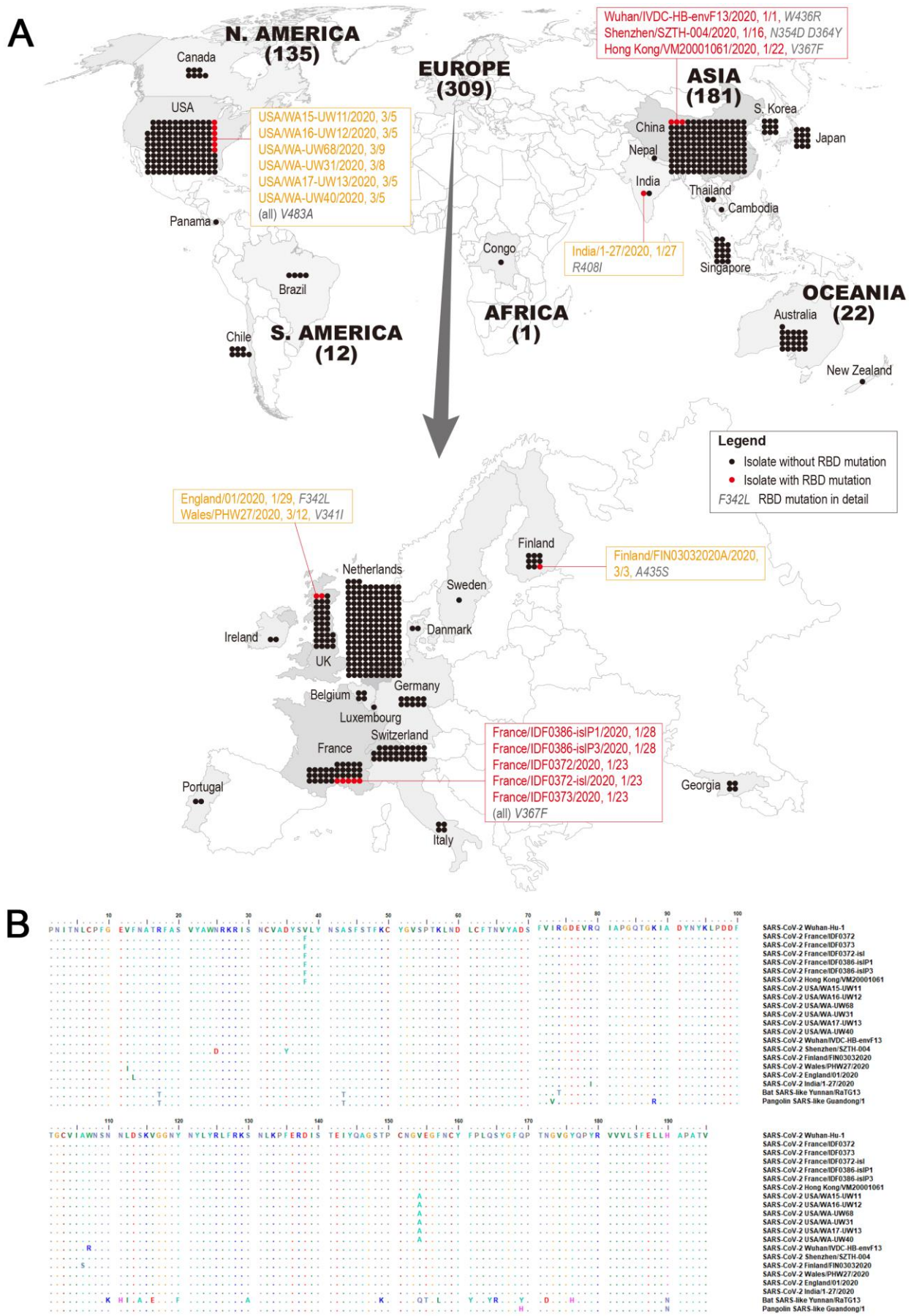
Among the 660 SARS-CoV-2 strains in the public databases with whole genome sequences available, only 18 strains contained amino acid mutations in the RBD (Table S1). These mutants were isolated from multiple locations in the world, including Wuhan, Shenzhen, Hong Kong, England, Finland, France and India (Fig. 1A). 17 out of 18 mutants deviate from the firstly reported strain (SARS-COV-2 Wuhan-Hu-1) for only one amino acid, while the Shenzhen-SZTH-004 strain contain two amino acids substitutions (Fig. 1B). These 18 mutants fall into 9 mutation types.

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.15.991844; this version posted March 23, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Mutation V367F was found in six individual isolates from four adult patients: three in France and one in Hong Kong, China, which suggested that these strains may have originated as a novel sub-lineage. The same also applied to the 6 isolates from the US, which all contained V483A mutation.

To be noted, none of the mutations in SARS-CoV-2 mutants were found in the Bat SARS-like CoV-RaTG013 or in the Pangolin SARS-like CoV-GD-1. This demonstrated that these mutations were not recombinants from the animal-originated virus, at least in the RBD, but rather naturally selected during spreading and circulating among human beings.

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.15.991844; this version posted March 23, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



155

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.15.991844; this version posted March 23, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Fig. 1: The SARS-CoV-2 mutated strains in RBD of the S protein. (A) The geographic distribution of the RBD mutated isolates. The strains with names in red are mutants with the enhanced binding affinity. The strains with names in yellow are mutants with similar binding affinity. **(B)** Multiple alignments of the RBD amino acid sequences. SARS-CoV-2 Wuhan-Hu-1, the first isolated strain, is used as reference. The bat and pangolin SARS-like virus are also included. Amino acid substitutions are marked.

Nucleotide diversity indicates strong positive selective pressure in RBD

The protein mutations are originated from the mutated RNA genome sequence, which is the nature of RNA virus. Since RBD is the only domain to bind human ACE2 to initiate the invasion, it is thought that the RBD should be highly conserved. However, our nucleotide diversity analysis of the entire S gene showed that the RBD domain is as diverse as the other regions of the S protein (Fig. 2). The peak signals for diversity distribute in the entire S protein, and the multiple peaks in the RBD also reached the Pi value of ~0.0002-0.0005. Since the RBD function is essential for the virus, we hypothesize that the mutation-prone RBD should be selected to maintain or even improve the binding affinity against human ACE2.

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.15.991844; this version posted March 23, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

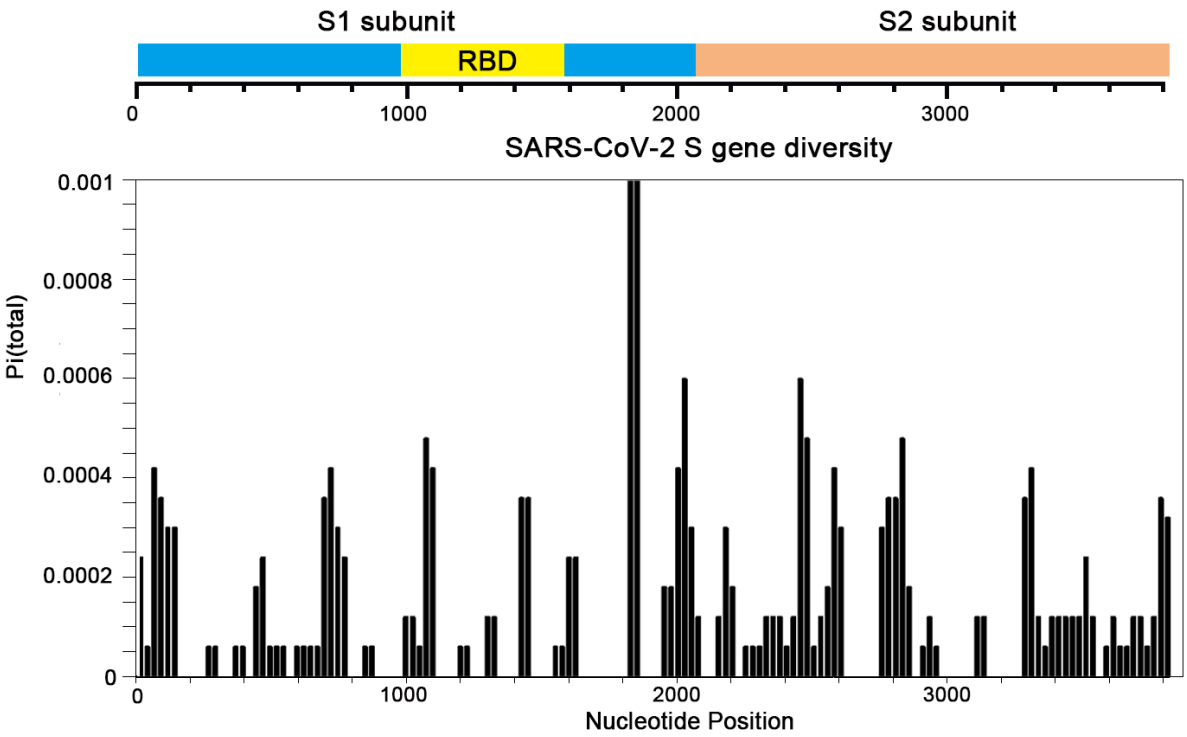


Fig. 2: Polymorphism and divergence graph of SARS-CoV-2 S gene. Structural domains are annotated. The Pi values are calculated with window size: 50 nt, step size: 10.

To further test this hypothesis, we investigated the selective pressures of the S gene by calculating nonsynonymous/synonymous substitution rate ratios (dN/dS ratios) for various segments of the S gene in the 660 SARS-CoV-2 strains. In accordance to our hypothesis, the entire S gene exhibited a dN/dS of 4.6526, remarkably greater than 1, showing that the S gene is under positive selective pressure (Table 1). Surprisingly, the S1 subunit showed a much higher dN/dS value of 10.8094. Therefore, S1 is the major contributor of positive selective pressure to the S gene. The high dN/dS (4.3966) of RBD located in S1 subunit indicated that the high selective pressure was applied to this functionally essential domain. Therefore, the functional relevance of these RBD mutations can be postulated.

Table 1: Estimates of Average Codon-based Evolutionary Divergence over S gene Pairs.

The numbers of nonsynonymous and synonymous differences per sequence from averaging over all sequence pairs are shown. Analyses were conducted using the Nei-Gojobori model. The analysis involved 660 SARS-CoV-2 nucleotide sequences. All positions containing gaps and missing data were discarded.

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.15.991844; this version posted March 23, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Gene	Length(bp)	Mean Non-syonymous Substations/site	Mean Syonymous Substations/site	dN/dS
S	3822	0.7233	0.1555	4.6526
S1	2043	0.5885	0.0544	10.8094
S1-RBD	585	1.0514	0.2391	4.3966
S2	1779	0.1348	0.1010	1.3342

Three types of mutants bind ACE2 with higher affinity

To estimate the functional alteration caused by the RBD mutations, we performed molecular dynamics simulation for the prototype SARS-CoV-2 (Wuhan-Hu-1 strain) and the RBD mutants to assess their binding energy to human ACE2. Each model was simulated in triple replicates. All trajectories reached plateau of RMSD after 2~5ns (Fig. 3A), indicating that their structure reached an equilibrium. Therefore, all the subsequent computation on thermodynamics was based on the 5~10ns trajectories. Three types of RBD mutants (N354D and D364Y, V367F, W436R) exhibited significantly lowered ΔG, suggesting their significantly increased affinity to human ACE2; the other mutants showed similar ΔG compared to the prototype (Fig. 3B). The ΔG of these three mutation types were all around -58 kJ/mol, approximately 25% lower than the prototype strain (-46.5 kJ/mol, calculated from the experimentally measured K_D) (Fig. 3B). Comparing to the K_D = 14.7 nM of the prototype RBD[9], the equilibrium dissociation constant (K_D) of these three mutants are calculated as 0.12 nM for N354D and D364Y, 0.11 nM for V367F, and 0.13 nM for W436R (Fig. 3C), two orders of magnitude lower than the prototype strain, indicating a remarkably increased affinity of these mutated viruses.

Only one mutant isolated from Shenzhen possesses dual amino acids mutation (N354D, D364Y). We also made models of single amino acids respectively and performed molecular dynamics simulation to investigate their individual influence to the affinity. The N354D substitution decreased the affinity, while the D364Y single mutation reached even higher affinity than the dual mutant (Fig. 3B). This indicated that the D364Y is the major contributor to the enhanced affinity.

In comparison, the bat CoV RaTG13 showed only minor binding affinity to human ACE2, while the

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.15.991844; this version posted March 23, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

pangolin CoV discovered in Guangdong showed remarkable ΔG, but slightly higher than the prototype human SARS-CoV-2 Wuhan-Hu-1 strain. The K_D of the SARS-CoV RBD of bats and pangolins to human ACE2 are estimated as 1.17mM and 1.89μM, respectively (Fig. 3C). Considering that the SARS-CoV RBD binds to human ACE2 at an affinity of K_D = 0.326μM [9], these data indicated that bat SARS-like CoV RaTG13, which was the closest bat CoV to human SARS-CoV-2, may be hardly infectious to humans. However, the K_D of pangolin CoV is only 5.8 times higher than the SARS-CoV. This indicated that the pangolin CoV is potentially infectious for humans by unprotected close contact with the virus-rich media, such as body fluid of the infected animal. This was consistent with the situation in the seafood market.

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.15.991844; this version posted March 23, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

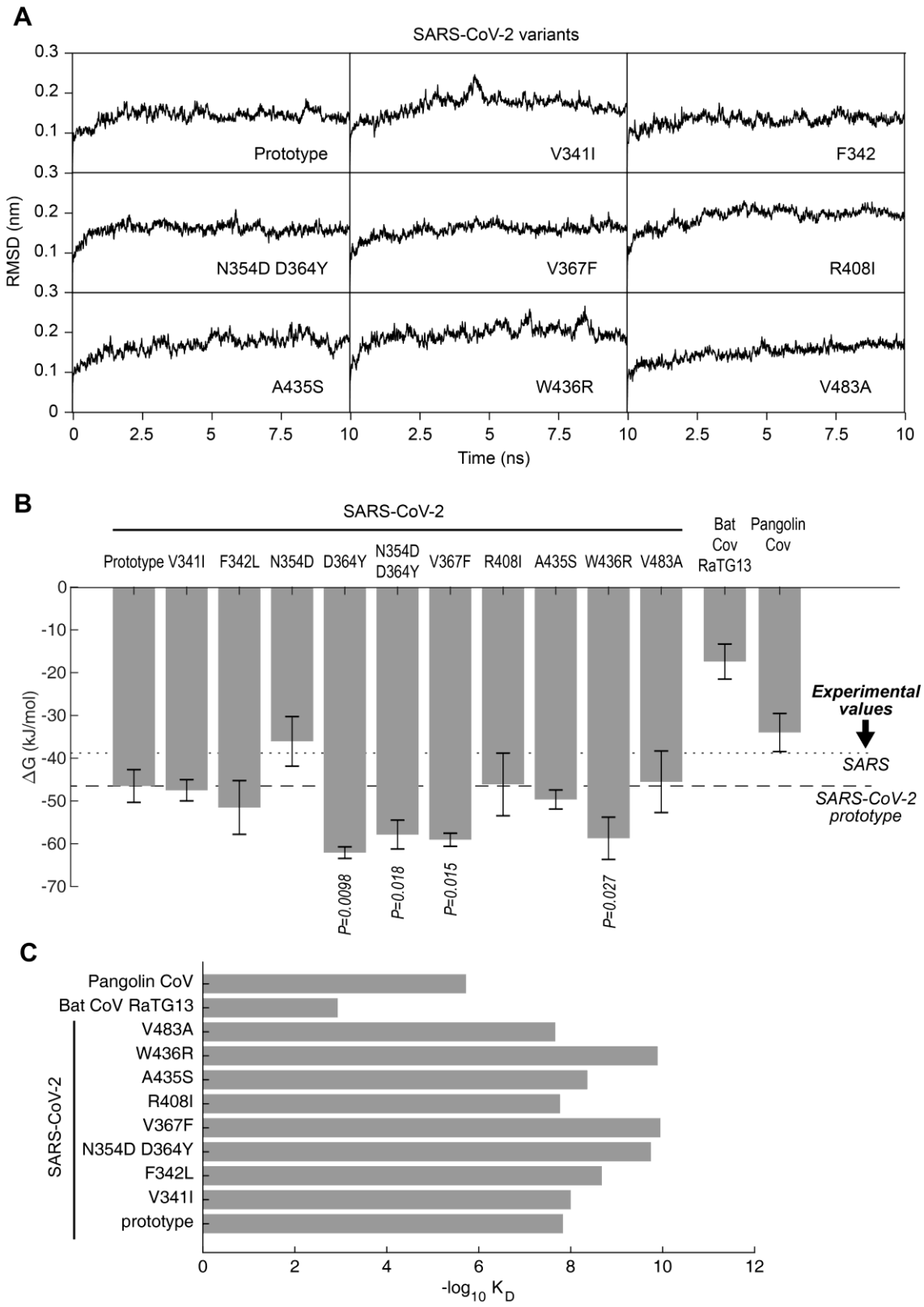


Fig. 3: Binding free energy of the SARS-CoV-2 S-RBD to human ACE2. (A) RMSD of typical MD trajectories of SARS-CoV-2 prototype and mutants. **(B)** Binding free energy (ΔG) of the RBDs and the human ACE2. Lower ΔG means higher affinity. Data are presented as mean±SD. *P*-values were calculated using single-tailed student t-test. The *P*-values are shown for those with $P < 0.05$.

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.15.991844; this version posted March 23, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

The ΔG calculated from experimental K_D values of SARS and SARS-CoV-2 prototype are marked in dotted and dashed lines, respectively. **(C)** The equilibrium dissociation constant (K_D) calculated according to the ΔG.

Structural basis of the increased affinity

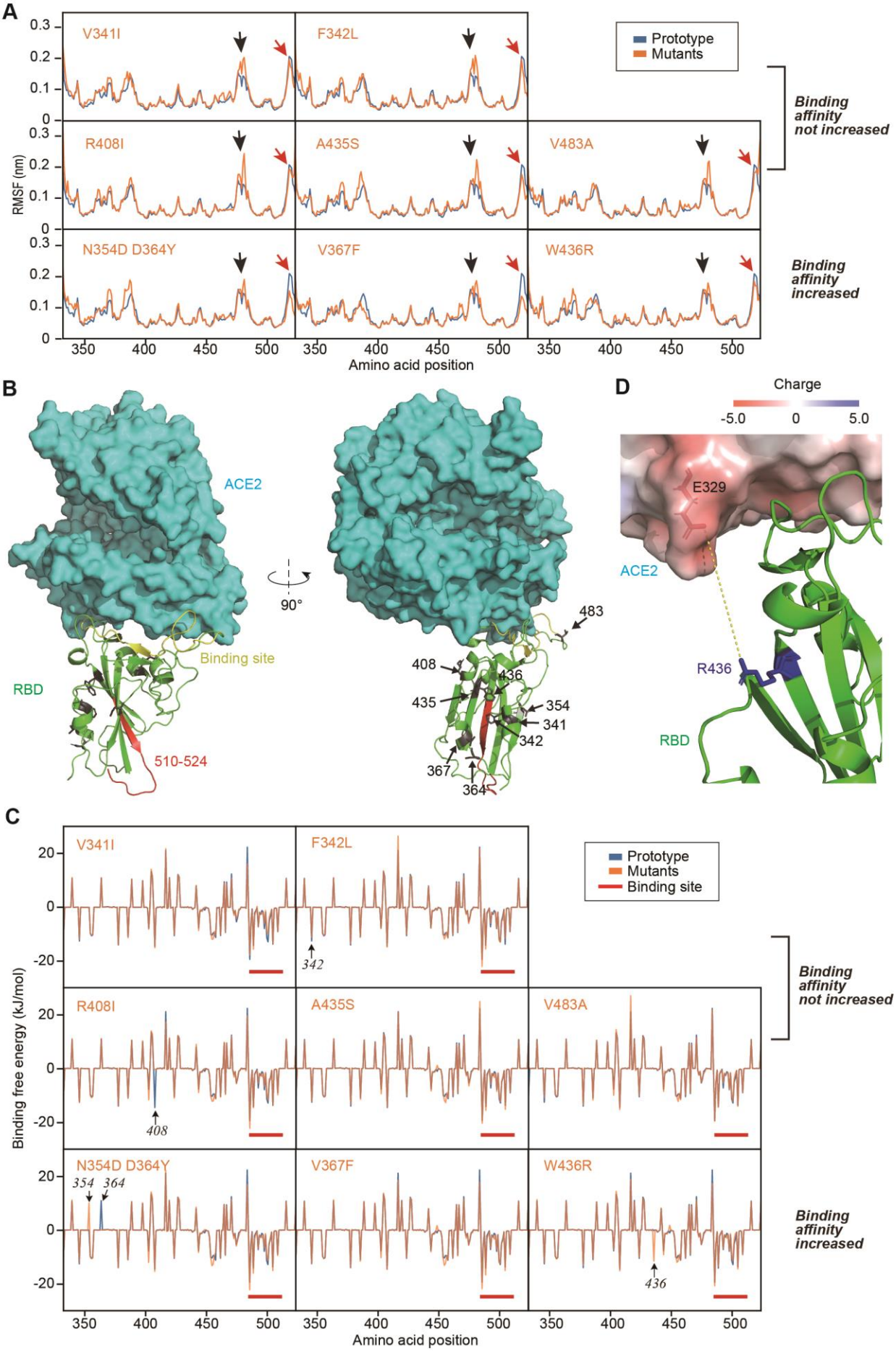
To explain the structural basis of the increased affinity, we investigated deeper into the dynamics of the residues of these structures. The 8 mutant types were divided into two groups: the "similar affinity" group (V341I, F342L, R408I, A435S, V483A), whose affinity is not significantly increased, and the "higher affinity" group (N354D D364Y, V367F, W436R), whose affinity is significantly increased. We compared the RMSF (Root Mean Square of Fluctuation) of the mutants to the prototype Wuhan-Hu-1 strain (Fig. 4A). It is notable that in the C-terminal of the RBD domain, namely the amino acids 510-524, the "higher affinity" mutants showed considerable decrease of the RMSF at this region, but not in the "similar affinity" mutants. Coincidently, the mutated amino acids which caused the affinity increase (D364Y, V367F, W436R) are all located near this fragment, while the mutated amino acids which did not increase the affinity (V341I, F342L, N354D, R408I, A435S, V483A) are away from this fragment (Fig. 4B). This explains the structural influence. Lower fluctuation reflects more rigid structure. The fragment 510-524 is the center of the beta-sheet structure (Fig. 4B, marked as red), which is the center scaffold of the RBD domain. To be noted, the binding surface of the RBD to ACE2 is largely in random-coil conformation, which lacks structural rigidity. In this case, a firm scaffold should be necessary to maintain the conformation of the interaction surface and thus may facilitate the binding affinity.

There are two features supporting this hypothesis. The first support lays in the RMSF curves. The residues 475-485 is a random coil near the binding site. The "similar affinity" group mutants showed remarkable increase of RMSF at this region, indicating a remarkable flexibility. In contrast, the "higher affinity" group mutants showed similar RMSF at this region to the prototype (Fig. 4A). The second support lays in the contribution of each amino acids to the binding free energy. In the binding site region, the "similar affinity" group mutants did not show an obvious decrease in ΔG, while the "higher affinity" group mutants exhibited a general decrease of ΔG in this region (Fig. 4C). The substitution R408I itself caused a remarkable increase of ΔG and thus weakened the

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.15.991844; this version posted March 23, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

affinity. In addition, the D364Y and W436R themselves directly contributed to the ΔG decrease. In contrast, the N354D mutation directly elevated the ΔG, which coincides its consequence (Fig. 4B). The positively charged arginine of the W436R is in the proximity of the highly negative charged ACE2 surface. The electrostatic attraction may additionally contribute to the affinity (Fig. 4D).

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.15.991844; this version posted March 23, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



267

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.15.991844; this version posted March 23, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Fig. 4: Structural analysis of RBD mutants on their affinity. (A) RMSF of the 5 mutants compared to the prototype. Red arrows denote the fragment of residues 510-524. Black arrows denote the fragment of residues 475-485. **(B)** Spatial location of the mutated amino acids and the fragment 510-524. **(C)** Contribution of each amino acids to the binding free energy. Red bars denote the binding site. (D) The charge state of the interaction surface of ACE2. The arginine of the W436R mutant is in the proximity. The electrostatic surface charge of the ACE2 are calculated using Pymol. The charge unit is K_bT/e_c, according to the Pymol manual.

Discussion

Due to the pandemic and constant mutations of the SARS-CoV-2 virus all over the world, the evolution of the virus infectivity is one of the most interested questions by the public. Alterations of virus infectivity may severely influence the quarantine policies. Our work tried to unravel the functional aspect of the RBD mutants.

Firstly, we investigated the polymorphism and diversity among the available SARS-CoV-2 S gene sequences. Among them, several diversity hot spots in S protein have been found in the whole gene, i.e., in both S1 and S2 subunits, including RBD domain which was related to receptor binding and antigen cognition. The high non-synonymous and synonymous mutation rate ratio revealed the strong selective pressure of S gene, especially in S1 subunit gene.

By the detailed alignment of all the S gene sequences available in the databases, two groups of amino acid mutations in SARS-CoV-2 RBD domain were identified: the “similar affinity” group (V341I, F342L, R408I, A435S, V483A) and the “higher affinity” group (N354D D364Y, V367F, W436R). Mutations F342L, N354D D364Y, and W436R were only discovered in single isolate. However, mutation V367F was discovered in six isolates. It was firstly discovered in one Hong Kong isolate, later appeared in five French isolates, across the continent. As RBD is conserved in SARS-CoV-2, the coincidence of six strains with the same mutation V367F in RBD in both France and Hong Kong is presumed significant for the virus transmission. It also indicates that these isolates may have originated as a novel sub-lineage, which has been circulating in the world,

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.15.991844; this version posted March 23, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

considering the close isolation dates (January 22 and 23, respectively). Similar scenario may also apply to the 6 mutated isolates in US. More epidemiological data are needed to confirm their potential relatedness.

It is interesting that only a fraction of RBD mutants showed increased affinity to human ACE2 and thus potentially increased the infectivity. It seems that the "higher affinity" mutants were discovered either in China, or could be potentially linked to China, where strict lockdown quarantine measures were effectively performed. Under these quarantine measures, the chances for virus to reach a susceptible person. This applies a high positive selection pressure on the virus: if the infectivity cannot be increased, the virus will soon become extinct. This explained the high positive selection pressure observed in Fig.2. Such a pressure would select the mutants with higher affinity to ACE2 for higher infectivity, just like the antibiotics pressure would select pan-resistant bacteria. The "similar affinity" mutants were all isolated from the countries without strict quarantine policy at the time of strain isolation. Without effective segregation, no such stress was applied, and the mutation towards higher infectivity is not essential. Although low quarantine measures would delay the emergence of "super-infectious" mutants, a country must risk the pandemic of the virus and multiple deaths. In contrast, strict quarantine measures in mainland China have been proven to effectively and robustly trap the mutants with higher infectivity to extinction: the Wuhan and Shenzhen isolates was not observed again.

The origination of the virus is a constant hot topic since the virus outbreak. Due to the high homology of the bat SARS-like CoV genome and pangolin CoV RBD to the SARS-CoV-2, these wild animals, especially the ones which were illegally on sale in the Wuhan Huanan Seafood Market, were thought to initiate the infection in human. Our results provided more clues on this postulation. In one aspect, the binding energy of the bat SARS-like CoV RBD is too high to directly bind human ACE2 (K_D in millimolar range). In contrast, the pangolin CoV showed a K_D to human ACE2 at micromolar range, just ~6x higher than that of the human SARS virus (Fig. 4), which indicates that the pangolin CoV can potentially infect human in close contact. The highly homologous pangolin CoV has been widely detected among the illegally transported Malayan pangolins in recent years in multiple provinces in China[7,8], which means that the wild pangolins are

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.15.991844; this version posted March 23, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

frequent carriers of the CoV in the nature. This indicates that the risk of zoonotic infection from wild animals to human constantly and widely exists. In another aspect, however, the sequence pattern suggested that this outbreak of SARS-CoV-2 was not directly originated from the pangolin CoV infection. The pangolin CoV deviate from human SARS-CoV-2 for 6 amino acids in RBD, but none of the 244 SARS-CoV-2 strains circulating in the world contain any of these 6 amino acids (Fig. 1B). Alignment of the genomic sequences of SARS-CoV-2 and pangolin CoV viruses indicated the evidence for recombination events in RBD domain between pangolin and bat viruses[6,8].

Our analysis of molecular dynamics simulation indicates the remarkable enhancement of the affinity efficiency of mutated S protein. Compared to the prototype strain Wuhan-Hu-1, the ΔG of mutants decreased ~25%. Mutants bind ACE2 more stably due to the enhancement of the base rigidity. Potential and recent animal-to-human transmission events of SARS-CoV-2, may explain the strong positive selection and enhancement of the affinity during the pandemic. The viruses have been adapting to transmission and replication in humans; mutation or recombination events in RBD may boost the binding affinity and cause the basic reproduction number (R0) to climb in theory, i.e., the human to human transmission more easily. Although these binding energies were estimated via *in silico* MD simulation besides the experimentally measured parameters of the prototype S protein, the dielectric estimation error in the MM/PBSA method would not affect accurate ranking prediction of the binding[25]. Namely, although the absolute value of ΔG might be refined by further approaches, the reproducible computational results can reveal the trend of affinity alterations.

The S protein is also important for antigen cognition. Fortunately, only a few amino acid mutations occurred in the RBD domain of the S protein, which showed the conservativeness of this domain. Judging from this point, the vaccines which focus on the RBD of S protein may still work for the SARS-CoV-2. However, the continuous surveillance of RBD variation is of critical importance because some mutants may enhance the infectivity or change the antigenicity.

In summary, our study identified two groups of amino acid mutations in SARS-CoV-2 RBD domain: the “similar affinity” group and the “higher affinity” group. The “higher affinity” group included

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.15.991844; this version posted March 23, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

the amino acids that were located at the firm scaffold, which facilitated the receptor binding. The four mutations of RBD under the positive selective pressure enhanced the affinity efficiency of the SARS-CoV-2 S protein. Knowing the structural binding mechanism will support the vaccine development and facilitate prevention countermeasure development. Although the biological outcomes of these mutations have not been confirmed by wet bench, the mutation analysis of RBD provides the insights into the evolutional trend of SARS-CoV-2 under high selection pressure. Combined with the epidemiology data, mutation surveillance is of critical importance, which can reveal more exact spreading routes of the epidemics and provide early warning for the possible outbreaks. Enhancement of SARS-CoV-2 binding affinity to human ACE2 reveals the higher risk of more severe virus transmissions during a sustained pandemic of COVID-19 if no effective precautions are implemented. The emergence of RBD mutations in Hong Kong, France and other countries which enhanced the RBD affinity to ACE2 receptor, requires special attention by all the countries.

Reference

1. Zhu N, Zhang D, Wang W, et al. A Novel Coronavirus from Patients with Pneumonia in China, 2019. *N Engl J Med*. 2020:727-733. doi:10.1056/nejmoa2001017
2. Li Q, Guan X, Wu P, et al. Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus–Infected Pneumonia. *N Engl J Med*. 2020:1-9. doi:10.1056/nejmoa2001316
3. Wang D, Hu B, Hu C, et al. Clinical Characteristics of 138 Hospitalized Patients with 2019 Novel Coronavirus-Infected Pneumonia in Wuhan, China. *JAMA - J Am Med Assoc*. 2020:1-9. doi:10.1001/jama.2020.1585
4. Chan JFW, Yuan S, Kok KH, et al. A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster. *Lancet*. 2020;395(10223):514-523. doi:10.1016/S0140-6736(20)30154-9
5. Coronavirus disease (COVID-2019) situation reports. 2020;49(3):e99-e100. （https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200319-sitrep-59-covid-19.pdf?sfvrsn=c3dcdef9_2）
6. Zhou P, Yang X-L, Wang X-G, et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature*. 2020. doi:10.1038/s41586-020-2012-7

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.15.991844; this version posted March 23, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

7. Xiao K, Zhai J, Feng Y, et al. Isolation and Characterization of 2019-nCoV-like Coronavirus from Malayan Pangolins. *bioRxiv*. January 2020:2020.02.17.951335. doi:10.1101/2020.02.17.951335

8. Lam TT-Y, Shum MH-H, Zhu H-C, et al. Identification of 2019-nCoV related coronaviruses in Malayan pangolins in southern China. *bioRxiv*. January 2020:2020.02.13.945485. doi:10.1101/2020.02.13.945485

9. Wrapp D, Wang N, Corbett KS, et al. Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. *Science (80-)*. 2020;367(6483):1260 LP - 1263. doi:10.1126/science.abb2507

10. Walls AC, Tortorici MA, Bosch B, et al. Cryo-electron microscopy structure of a coronavirus spike glycoprotein trimer Alexandra. 2016;531(7592):114-117. doi:10.1038/nature16988.Cryo-electron

11. Reese JB, Bober SL, Daly MB, et al. Prefusion structure of a human coronavirus spike protein Robert. *Nature*. 2016;531(7592):118-121. doi:10.1002/cncr.31084.Talking

12. Walls AC, Tortorici MA, Snijder J, et al. Tectonic conformational changes of a coronavirus spike glycoprotein promote membrane fusion. *Proc Natl Acad Sci U S A*. 2017;114(42):11157-11162. doi:10.1073/pnas.1708727114

13. Chen Y, Guo Y, Pan Y, Zhao ZJ. Structure analysis of the receptor binding of 2019-nCoV. *Biochem Biophys Res Commun*. 2020;2(xxxx):0-5. doi:10.1016/j.bbrc.2020.02.071

14. Wan Y, Shang J, Graham R, Baric RS, Li F. Receptor recognition by novel coronavirus from Wuhan: An analysis based on decade-long structural studies of SARS. *J Virol*. 2020;(January). doi:10.1128/jvi.00127-20

15. Fast E, Chen B. Potential T-cell and B-cell Epitopes of 2019-nCoV. 2020:1-13.

16. Ahmed SF, Quadeer AA, McKay MR. Preliminary identification of potential vaccine targets for 2019-nCoV based on SARS-CoV immunological studies. *Viruses*. 2020;(February):2020.02.03.933226. doi:10.1101/2020.02.03.933226

17. Tang X, Wu C, Li X, et al. On the origin and continuing evolution of SARS-CoV-2. *Natl Sci Rev*. March 2020. doi:10.1093/nsr/nwaa036

18. Kuraku S, Zmasek CM, Nishimura O, Katoh K. aLeaves facilitates on-demand exploration of metazoan gene family trees on MAFFT sequence alignment server with enhanced

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.15.991844; this version posted March 23, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

interactivity. *Nucleic Acids Res*. 2013;41(Web Server issue):22-28. doi:10.1093/nar/gkt389

19. Katoh K, Rozewicki J, Yamada KD. MAFFT online service: Multiple sequence alignment, interactive sequence choice and visualization. *Brief Bioinform*. 2018;20(4):1160-1166. doi:10.1093/bib/bbx108

20. Rozas J, Ferrer-Mata A, Sanchez-DelBarrio JC, et al. DnaSP 6: DNA sequence polymorphism analysis of large data sets. *Mol Biol Evol*. 2017;34(12):3299-3302. doi:10.1093/molbev/msx248

21. Nei M, Gojoborit T. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol Biol Evol*. 1986;3(5):418-426. doi:10.1093/oxfordjournals.molbev.a040410

22. Kumar S, Stecher G, Li M, Knyaz C, Tamura K. MEGA X: Molecular evolutionary genetics analysis across computing platforms. *Mol Biol Evol*. 2018;35(6):1547-1549. doi:10.1093/molbev/msy096

23. Waterhouse A, Bertoni M, Bienert S, et al. SWISS-MODEL: Homology modelling of protein structures and complexes. *Nucleic Acids Res*. 2018;46(W1):W296-W303. doi:10.1093/nar/gky427

24. Homeyer N, Gohlke H. Free energy calculations by the Molecular Mechanics Poisson-Boltzmann Surface Area method. *Mol Inform*. 2012;31(2):114-122. doi:10.1002/minf.201100135

25. Sanders JM, Wampole ME, Thakur ML, Wickstrom E. Molecular Determinants of Epidermal Growth Factor Binding : A Molecular Dynamics Study. 2013;8(1):8-10. doi:10.1371/journal.pone.0054136

Funding

This work was supported by grants from the National Key Research and Development Program of China (2017YFA0505001/2018YFC0910200/2018YFE0204503), National Natural Science Foundation of China (81730061), Guangdong Key Research and Development Program (2019B020226001), Natural Science Foundation of Guangdong Province (2018B030312010) as well as the Guangzhou Healthcare Collaborative Innovation Major Project (201803040004 and 201803040007).

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.15.991844; this version posted March 23, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Conflict of interest

The authors declare that they have no conflicts of interest.

Acknowledgments

We gratefully acknowledge the authors, originating and submitting laboratories of the sequences from GISAID's EpiFlu™ Database on which this research is based. All submitters of data may be contacted directly via www.gisaid.org.

Appendix:

Table.S1 Meta data of the isolates with mutations in spike glycoprotein RBD

GISAID Virus name	RBD mutation	Collection date	Location	Gender	Age	Specimen source	Additional information	Accession ID
hCoV-19/Wuhan/IVDC-HB-envF13/2020	W436R	2020-1-1	Asia/China/Hubei/Wuhan	-	-	Environment	Huanan Seafood Market	EPI_ISL_408511
hCoV-19/Shenzhen/SZTH-004/2020	N354D, D364Y	2020-1-16	Asia/China/Guandong/Shenzhen	Male	63	Alveolarlavage fluid	-	EPI_ISL_406595
hCoV-19/HongKong/VM20001061/2020	V367F	2020-1-22	Asia/HongKong	Male	39	Nasopharyngeala spirate&Throat swab	-	EPI_ISL_412028
hCoV-19/France/IDF0372/2020	V367F	2020-1-23	Europe/France/Ile-de-France/Paris	Female	31	Oro-Pharyngeal swab	-	EPI_ISL_406596
hCoV-19/France/IDF0372-isl/2020	V367F	2020-1-23	Europe/France/Ile-de-France/Paris	Female	31	Oro-Pharyngeal swab	-	EPI_ISL_410720
hCoV-19/France/IDF0373/2020	V367F	2020-1-23	Europe/France/Ile-de-France/Paris	Male	32	Orao-pharungeal swab	-	EPI_ISL_406597
hCoV-19/India/1-27/2020	R408I	2020-1-27	Asia/India/Kerala	Female	20	Throat swab	Travel history to China	EPI_ISL_413522
hCoV-19/France/IDF0386-islP1/2020	V367F	2020-1-28	Europe/France/Ile-de-France/Paris	Female	30	Naso-pharyngeal swab	Related to EPI_ISL_406596	EPI_ISL_411219
hCoV-19/France/IDF0386-islP3/2020	V367F	2020-1-28	Europe/France/Ile-de-France/Paris	Female	30	Naso-pharyngeal swab	Related to EPI_ISL_406596	EPI_ISL_411220
hCoV-19/England/01/2020	F342L	2020-1-29	Europe/England	Female	50	swab	England cluster patient1	EPI_ISL_407071
hCoV-19/Finland/FIN03032020A/20	A435S	2020-3-3	Europe/Finland/Helsinki	Male	40	-	-	EPI_ISL_41

20								3602
hCoV-19/USA/WA15-UW11/2020	V483A	2020-3-5	NorthAmerica/USA/Washington	-	-	-	Collection dates may be off by +/-3days	EPI_ISL_413650
hCoV-19/USA/WA16-UW12/2020	V483A	2020-3-5	NorthAmerica/USA/Washington	-	-	-	Collection dates may be off by +/-3days	EPI_ISL_413651
hCoV-19/USA/WA17-UW13/2020	V483A	2020-3-5	NorthAmerica/USA/Washington	-	-	-	-	EPI_ISL_413652
hCoV-19/USA/WA-UW40/2020	V483A	2020-3-5	NorthAmerica/USA/Washington	-	-	-	-	EPI_ISL_415605
hCoV-19/USA/WA-UW31/2020	V483A	2020-3-8	NorthAmerica/USA/Washington	-	-	-	-	EPI_ISL_414618
hCoV-19/USA/WA-UW68/2020	V483A	2020-3-9	NorthAmerica/USA/Washington	-	-	-	-	EPI_ISL_415596
hCoV-19/Wales/PHW27/2020	V341I	2020-3-12	Europe/UnitedKingdom/Wales	Male	49	-	-	EPI_ISL_415655