Polymorphism and selection pressure of SARS-CoV-2 vaccine and diagnostic antigens: implications for immune evasion and serologic diagnostic performance

Running title: SARS-CoV-2 antigen polymorphism

Eric Dumonteil, Claudia Herrera

Department of Tropical Medicine, Vector-Borne and Infectious Disease Research Center, School of Public Health and Tropical Medicine, Tulane University, New Orleans, LA, USA.

Corresponding Author: Eric Dumonteil, Department of Tropical Medicine, School of Public Health and Tropical Medicine, Vector-Borne and Infectious Disease Research Center, Tulane University, 1440 Canal St., New Orleans, LA, 70112, USA. E-mail: edumonte@tulane.edu

Abstract

The ongoing SARS-CoV-2 pandemic has triggered multiple efforts for serological tests and vaccine development. Most of these tests and vaccines are based on the Spike glycoprotein (S) or the Nucleocapsid (N) viral protein. Conservation of these antigens among viral strains is critical to ensure optimum diagnostic test performance and broad protective efficacy, respectively. We assessed N and S antigen diversity from 17,853 SARS-CoV-2 genome sequences and evaluated selection pressure. Up to 6-7 incipient phylogenetic clades were identified for both antigens, confirming early variants of the S antigen and identifying new ones. Significant diversifying selection was detected at multiple sites for both antigens. Some sequence variants have already spread in multiple regions, in spite of their low frequency. In conclusion, the N and S antigens of SARS-CoV-2 are well conserved antigens, but new clades are emerging and may need to be included in future diagnostic and vaccine formulations.

Keywords: antigenic drift, diversifying selection, coronavirus, immune evasion, diversity

Introduction

The emergence and rapid spread of a novel Coronavirus, referred to as SARS-CoV-2, is resulting in one of the worst pandemic in the world, causing an unprecedented health and economic crisis. About seven months after the first cases were identified, over 8 million cases have been reported worldwide, with over 400,000 deaths according to the Johns Hopkins Coronavirus Resource Center.

The pandemic has triggered multiple efforts at developing serological tests, able to detect both acute infections by detecting virus-specific IgM, as well as recovered individuals by detecting virus-specific IgG. Several immunochromatographic rapid tests are already available (1), and several more will become available in the next few months. Such tools would be critical to increase testing for the accurate and rapid identification of cases and their isolation to limit further transmission of the virus. However, their performance needs to be evaluated, and initial testing suggested variable performance of these tests (1, 2). Test performance relies in part on the antigen used, and its conservation among virus strains circulating in the population being tested. Currently, most of these tests are based on the Spike glycoprotein (S) or the Nucleocapsid (N) viral proteins (1). The receptor-binding domain (RBD) of the S protein, which mediates binding to the angiotensin-converting enzyme 2 (ACE2) receptor in human cells (3), is also widely used as a diagnostic antigen.

Similarly, vaccine development efforts have been very intense and a growing number of vaccine candidates are being quickly moved into clinical trials. These are based on different technological platforms, ranging from recombinant proteins, RNA and DNA vaccines, or recombinant viral vectors (4, 5). A first RNA vaccine candidate

recently completed clinical phase 1 evaluation, and is expected to move into Phase 2 shortly. Most of these vaccine candidates are based on the viral S protein, or the RBD as antigen. Multiple potential vaccine epitopes have also been identified in the S as well as in the N viral proteins (6). As for diagnostics, conservation of these vaccine antigens among viral strains is critical to ensure broad protection and avoid immune evasion by the virus.

As an RNA virus, SARS-CoV-2 is prone to frequent mutations, in spite of some proof-reading abilities of its RNA polymerase complex (7, 8). An early assessment of genomic changes SARS-CoV-2 showed a mutation hot-spot in the virus RNA dependent RNA polymerase (RdRp), but a few mutations were also detected in other parts of the viral genome, including the N and S proteins (9). The growing availability of a large number of complete genome sequences gathered since the beginning of the pandemic provides a unique tool to assess the extent of viral antigen polymorphisms, and potential selection pressures on these. A first analysis of polymorphisms in the S glycoprotein until early April 2020 identified a handful of variant sites, including D614G, S943P, and possibly L5F and L8V (10). Variant sites V367F, G476S, and V483A were also identified in the RBD. We analyzed here the sequence variation in a broader set of viral proteins N and S, which represent the main diagnostic and vaccine antigens to date. We examined the implications of the identified sequence variants on vaccine and serological diagnostic performance.

Experimental procedures

Viral sequence data

Whole genome sequences from 18,247 SARS-CoV-2 virus were obtained from GISAID (Supplemental Table 1), covering virus isolates from multiple continents, including Asia, Africa, Europe, Oceania, and America. These sequences included those from initial human cases in Wuhan, China from December 2019 up to sequences from May 11, 2020.

Sequence analysis

Viral genome sequences were aligned using MAFFT (11) as implemented in Geneious 11, and alignments were edited to exclude partial or low quality sequences. A final alignment including 17,853 quality sequences were used to construct phylogenetic trees using FastTree (12) for a global analysis of viral diversity across the world. FastTree infers approximately-maximum-likelihood phylogenetic trees. Sequence conservation across genome alignment was calculated using s sliding window of one in Geneious.

Separate analyses were then performed using S and N genes, as well as the RBD from the S protein (positions 319-540 within the S protein). For these, translated sequences were aligned with the MAFFT algorithm using Blossum62 matrix and the frequency of variants at each site was calculated. Unique sequences from these proteins were then selected and phylogenetic trees were constructed using FastTree as above. Predicted epitopes from these antigens (6, 13) were mapped in the alignments, as well as glycosylation sites (14) to assess their conservation among viral sequences. Finally, evolutionary selection pressures on the antigens were analyzed using the Fast,

Unconstrained Bayesian AppRoximation (FUBAR), as implemented in HyPhy (15) and statistical significance was considered at a threshold of $P<0.1$.

Results

Analysis of over 17,000 genome sequences confirmed the SARS-CoV-2 is a fast evolving virus, as it is rapidly accumulating mutations. Indeed, in the less than 5 months that viral sequences have been available, we detected sequence variants scattered throughout the viral genome, rather than clustered in specific genes (Figure 1A and B) and some virus circulating now in multiple countries has somewhat diverged from some of the isolates initially sequenced in December 2019 in Wuhan, China (Figure 1C). Importantly, some sequence variation could be detected within both the N and S genes.

These genes were then analyzed in detail and separately. For the N protein, we included a dataset of 16,656 sequences, and significant sequence diversity was detected, with up to 326 distinct protein sequences. For a clearer assessment of their phylogenetic relationship, these variant sequences were analyzed independently (Figure 2A). Notably, a structuring including up to seven incipient clades was found emerging, with sequences from the first virus from Wuhan, China included in Clade 1 (Figure 2C). There was no specific geographic clustering of the sequence variants, illustrating the widespread multidirectional spreading of the virus across the world. A notable exception was observed for Clade 3, which included mostly sequences from Europe. Analysis of sequence variation along the protein sequence indicated that about half of the protein on the amino side was mostly conserved, except in two regions at sites 13 and 203-204,

respectively (Figure 2B). On the other hand, the carboxy half of the protein appeared more variable, but this also reflected some sequencing ambiguities.

A total of 178/419 (42.5%) sites presented variation in the N protein. This included seven sites with four variant amino acids, seven sites with three variant amino acids, and 13 sites with two variant amino acids that were found under significant diversifying selection pressure (26/419 (6.2%), Table 1). Because of these changes, the N protein is slowly diverging from the sequence from some of the early virus, belonging to Clade 1, and up to six additional major clades (Clades 2-7) are emerging for the N antigen (Figure 2C). Site D144 that can be substituted by E, H, Y or N may disrupt a predicted epitope (ALNTPKDHI 138-146). Importantly, most variants were still found at relatively low frequency among the viral population (0.018 to 0.541%), with only R203X and P13X variants detected at higher frequency (18.108 and 1.589%, respectively, Table 1), indicating an overall high level of conservation of the N protein. Nonetheless, many of the low frequency variants were found to have already dispersed in multiple countries and regions. This is for example the case of S202X variants, which were detected in 90 cases from Australia, China, Democratic Republic Congo, England, Ghana, India, The Netherland, Russia, Saudi Arabia, Senegal, Turkey, and the USA, or D22X, detected in 53 cases from Australia, England, Taiwan, Uruguay and Wales. On the other hand, a few variants were likely more associated with limited clusters of infection, such as A208G variants, which were mostly limited to the US so far.

A similar analysis of the S antigen was performed, based on 17,802 sequences. It revealed even greater sequence diversity, with up to 681 unique S protein variants, and most of this diversity was observed in the most recent months of March and April 2020

compared to January and February 2020 (Figure 3A). Furthermore, up to six emerging clades could be defined, that present a clear divergence from Clade 1, which includes some of the first sequences from Wuhan, China (Figure 3B). While some sequences from Clades 3 and 6 could be detected as early as February 2020, sequences from Clades 2, 4 and 5 appeared in March 2020 and expanded in April 2020. At the same time, sequences from Clade 1 appeared less frequent with time.

Analysis of sequence variation along the protein sequence indicated that amino acid variants were spread along most of the S glycoprotein (Figure 3C), although a few regions of lower sequence conservation could be detected at positions 260-320 just before the RBD, at position 445-515 in the carboxy end of the RBD, and at site 614. Further analysis of each major clade revealed that each had amino acid substitutions that concentrated in different domains of the proteins, except for Clade 1, which accumulated the greatest number of substitutions across the entire protein (Figure 4). For example, Clade 2 had more substitutions between sites 850-970, Clade 3 between 550-750 and 1150-1250, Clade 4 between 250-320, Clade 5 between 140-250 and 420-500, and Clade 6 between 750-800.

A total of 362/1273 (28.4%) sites presented variation in the S protein, of which 32 sites (2.5%) were found under significant diversifying selection pressure (Table 2). These included one site with five variant amino acids, one site with four variants, seven sites with three variants, 11 sites with two variants, and 12 sites with a single variant amino acid. Further more, different selection patterns were identified in each major clade, and only a few notable sites had substitutions in more that one clade (Table 2 and Figure 4). For example sequences from Clade 1 are clearly defined by site D614, which is under

strong diversifying selection, together with sites V615, G476, V483 and H519 and their corresponding variants. Most sequences from Clades 2-6 have a D614G substitution, together with clade specific variants. Thus, Clade 2 is characterized by a cluster of substitutions around sites 936-943, with specific sites D936, S940 and S943 and their variants under strong diversifying selection. Clade 3 is characterized by variants sites V622, A653, A684, A703 and their variants, and Clade 5 by sites D215, S221 and Q238 and their variants (Table 2 and Figure 4). A single predicted epitope may be affected by diversifying selection and substitutions at site 1078. The furin cleaveage region (671-692), and particularly the cleavage site were well conserved, although two sites, Q675 and A684 are under diversifying selection. Substitutions at these less conserved sites may thus not affect furin cleavage, which is unique to SARS-Cov-2 (3). Similarly, none of the sites with N-linked glycosylation (14) were found under diversifying selection, allowing for the conservation of the glycosylation pattern of the S protein across its diversity.

With the exception of the D614G substitution which has taken over and is now widespread in virus populations across the globe (over 63% of sequences carry this substitution), the other variants under selection still represent a low proportion of viral sequences, ranging from 0.017 to 0.586% (Table 3). A few of these variants likely correspond to limited clusters of infections, as they come from a single geographic region and are grouped in time. This is the case for the G1124V variant, which is limited to 50 cases from Victoria, Australia, between March 20-27, 2020. Similarly, the N439K variant is limited to 40 cases from Scotland, identified between March 16-April 5, 2020. However, most of the other variants have already spread to multiple countries and regions, such as Q675X, which has been found in Denmark, England, Finland, Iceland,

Norway, Scotland, Spain, and the USA over March and April 2020. Similarly, L5F variants have been found on 102 cases from Australia, Belgium, Canada, England, France, Iceland, India, Italy, Japan, Netherlands, Portugal, Scotland, Singapore, Taiwan, Thailand, USA, and Wales and H49X variants have been found in 36 cases from Australia, China, England, Mexico, Taiwan, and the USA, for example.

As mentioned above, some of the sequence variation affecting the S protein was detected within the RBD, which is a key functional domain of the protein and one of the most used targets for serological diagnostic. We thus analyzed in detail its polymorphism. Sequence analysis of RBD revealed that it represented a highly conserved region of the S protein. Nonetheless, up to 54 RBD sequence variants were identified, with again some significant divergence from the first sequences from Wuhan, China (Figure 5). Importantly, divergence seemed to increase with time as more variants accumulate and become established. A total of seven sites from the RBD were found under significant diversifying selection pressure, and variants sites within the RBD were observed in each of the major clades of the S protein (Table 2). Nonetheless, while possible RBD clades are emerging, these do not match the S protein major clades described above.

Discussion

Antigen polymorphism from pathogens has the potential to impair serological diagnostic test performance, as well as vaccine efficacy. It is thus of key importance to consider these aspects for serological test and vaccine development, to ensure their usefulness and broad efficacy. This is commonly done for influenza vaccines for example, that are updated each year based on circulating viral strains, as cross protection

among strains is still elusive (16). We investigated here the sequence diversity of two major antigens of the novel SARS-CoV-2 virus, the N and S proteins. Importantly, a significant level of sequence diversity was detected for both antigens, with incipient clades emerging as multiple sites were found under significant diversifying selection pressure.

The N protein, mostly used in serological diagnostic tests (1) had a large number of sequence variants, and 6.2% of its residues were found under diversifying selection. Overall up to seven major sequence clades have been emerging in recent months for this antigen, and these did not show any geographic clustering. A notable exception was Clade 3 of the N protein, which appeared over-represented in sequences from Europe so far. Importantly, predicted epitopes appeared conserved so far, although a more detail epitope mapping is still needed for this antigen. Nonetheless, N protein variants diverging from the initial sequences from Wuhan, China are now circulating in most geographic regions. While these changes are so far limited to a relatively small proportion of sequences (23.4%) and may not interfere with protein antigenicity, the inclusion of some of the variants in serological tests would ensure optimum sensitivity of tests, particularly if some of these variants become more frequent.

The S glycoprotein is the main vaccine candidate currently tested in multiple vaccine platforms/formulation (4, 5). Compared to the N antigen, it is more conserved and only 2.5% of its sites were found under diversifying selection pressure. We confirmed the importance of most of the variant sites previously identified in this antigen. These include D614G, S943P, as well as L5F and L8V and variant sites V367F, G476S, and V483A in the RBD (10). However, multiple additional variants were also identified

here, leading to the identification of up to six major clades of the S glycoprotein that are emerging. Most of these variants appeared in the past weeks/months and may be slowly replacing the virus presenting sequences similar to that of the initial isolates from Wuhan, China. Indeed, while most of the variants still have a low frequency in the viral population, several have already spread to multiple countries and regions, where they may reach higher frequencies in the near future if they are successfully transmitted. Importantly, none of the substitutions identified affected the glycosylation pattern of the S protein, and none of the predicted epitopes appear affected. While the functional impact of these variants is unknown, the D614G mutation has been associated with potential increased viral transmission and/or fitness (10), which may explain why it became so frequent. A recent comparison of functional properties of the S proteins with aspartic acid (SD614) and glycine (SG614) confirmed a greater infectivity correlated with less S1 shedding and greater incorporation of the S protein into the pseudovirion with the SG614 variant (17). Similar functional studies of the additional variants identified here may help evaluate their impact on virus fitness. Future studies will also provide data on how the different clades identified here may be successfully transmitted or go extinct.

While the RBD is particularly well conserved, some sequence variation was also detected in this region within the S glycoprotein, with up to 54 sequence variants. Because these differ by only 1-2 amino acids, the overall antibody recognition of the RBD can be expected to be mostly preserved so far, but some specific epitopes may nonetheless be lost. Also, our phylogenetic analysis suggested that possible clades may be emerging within the RBD as well, and newer sequences may diverge further from the sequence from the initial isolates from Wuhan.

In conclusion, we found that the N and S antigens of SARS-CoV-2 are so far highly conserved, so that both are good antigens for both diagnostic and vaccine development. However, some sequence variation is also emerging and 6-7 phylogenetic clades could be identified for both antigens. Some of these sequence variants have already spread in multiple countries and regions, in spite of their low frequency. Sequence variants may arise by random substitutions in the viral genome during replication, but the significant diversifying selection detected at multiple sites in both antigens suggests that immune selection pressure and adaptation to human hosts may be driving some of these changes, which may lead to the establishment of some of these variants. New variants are also likely to emerge with time. The recent identification of potential co-infections with more than one viral strain suggests that recombination could also contribute to the generation of SARS-CoV-2 genetic diversity (18). Therefore, further monitoring of antigen drift over time will be needed to ensure that diverging antigens can be identified in a timely manner and included in future diagnostic and vaccine formulations.

References

1. Whitman JD, Hiatt J, Mowery CT, Shy BR, Ruby Yu R, Yamamoto TN, et al. Test performance evaluation of SARS-CoV-2 serological assays. medRxiv: medRxiv; 2020.

2. Alger J, Cafferata ML, Alvarado T, Ciganda A, Corrales A, Desale H, et al. Using prenatal blood samples to validate COVID-19 rapid serologic tests. Research Square: Research Square; 2020.

3. 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 Apr 16;181(2):281-92 e6.

4. Chen WH, Strych U, Hotez PJ, Bottazzi ME. The SARS-CoV-2 Vaccine Pipeline: an Overview. Curr Trop Med Rep. 2020 Mar 3:1-4.

5. Cohen J. Vaccine designers take first shots at COVID-19. Science. 2020;368(6486):14-6.

6. Lee CH, Koohy H. In silico identification of vaccine targets for 2019-nCoV. F1000Res. 2020;9:145.

7. Smith EC, Blanc H, Surdel MC, Vignuzzi M, Denison MR. Coronaviruses lacking exoribonuclease activity are susceptible to lethal mutagenesis: evidence for proofreading and potential therapeutics. PLoS pathogens. 2013 Aug;9(8):e1003565.

8. Bouvet M, Imbert I, Subissi L, Gluais L, Canard B, Decroly E. RNA 3'-end mismatch excision by the severe acute respiratory syndrome coronavirus nonstructural protein nsp10/nsp14 exoribonuclease complex. Proceedings of the

National Academy of Sciences of the United States of America. 2012 Jun 12;109(24):9372-7.

9. Pachetti M, Marini B, Benedetti F, Giudici F, Mauro E, Storici P, et al. Emerging SARS-CoV-2 mutation hot spots include a novel RNA-dependent-RNA polymerase variant. J Transl Med. 2020 Apr 22;18(1):179.

10. Korber B, Fischer WM, Gnanakaran S, Yoon H, Theiler J, Abfalterer W, et al. Spike mutation pipeline reveals the emergence of a more transmissible form of SARS-CoV-2. bioRxiv: bioRxiv; 2020.

11. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Molecular biology and evolution. 2013 Apr;30(4):772-80.

12. Price MN, Dehal PS, Arkin AP. FastTree 2--approximately maximum-likelihood trees for large alignments. PloS one. 2010 Mar 10;5(3):e9490.

13. Ahmed SF, Quadeer AA, McKay MR. Preliminary Identification of Potential Vaccine Targets for the COVID-19 Coronavirus (SARS-CoV-2) Based on SARS-CoV Immunological Studies. Viruses. 2020 Feb 25;12(3).

14. Watanabe Y, Allen JD, Wrapp D, McLellan JS, Crispin M. Site-specific glycan analysis of the SARS-CoV-2 spike. Science. 2020 May 4.

15. Kosakovsky Pond SL, Poon AFY, Velazquez R, Weaver S, Hepler NL, Murrell B, et al. HyPhy 2.5-A Customizable Platform for Evolutionary Hypothesis Testing Using Phylogenies. Molecular biology and evolution. 2020 Jan 1;37(1):295-9.

16. Vemula SV, Sayedahmed EE, Sambhara S, Mittal SK. Vaccine approaches conferring cross-protection against influenza viruses. Expert review of vaccines. 2017 Nov;16(11):1141-54.

17. Zhang I, Jackson CB, Mou H, Ojha A, Rangarajan ES, Izard T, et al. The D614G mutation in the SARS-CoV-2 spike protein reduces S1 shedding and increases infectivity. BioRxiv: BioRxiv; 2020.

18. Sashittal P, Luo Y, Peng J, El-Kebir M, Azou M. Characterization of SARS-CoV-2 viral diversity within and across hosts. bioRxiv: bioRxiv; 2020.

Supplemental material

Supplemental Table 1: List of SARS-CoV-2 sequences used in the study

Table 1. Amino acids of SARS-CoV-2 N protein under diversifying selection pressure

Reference	Position	Variants				Proportion	%
P	13	L	T	R	S	264/16353	1.589
G	34	E	V	L	W	15/16630	0.090
D*	144	E	H	Y	N	12/16642	0.072
S	180	I	G	C	R	21/16620	0.126
R	191	G	C	L	S	14/16635	0.084
R	209	K	I	T	del	29/16616	0.174
A	381	V	T	P	S	9/16519	0.054
Q	28	H	E	R		12/16637	0.072
P	151	L	S	H		13/16633	0.078
R	185	H	L	C		25/16623	0.15
R	203	K	S	T		3004/13585	18.108
A	208	S	G	del		90/16558	0.541
S	232	I	R	T		5/16440	0.030
D	377	Y	H	G		17/16505	0.103
P	20	S	L			9/16630	0.054
D	22	G	Y			53/16588	0.318
T	24	N	I			32/16610	0.192
A	119	S	V			32/16609	0.192
S	190	G	I			32/16608	0.192
S	202	I	N			80/16561	0.481
T	205	I	del			51/16586	0.307
A	218	S	V			3/16635	0.018
H	300	Q	Y			9/16469	0.055
P	344	S	L			20/16551	0.121
D	348	H	Y			6/16607	0.036
E	378	Q	K			7/16516	0.042
A	397	S	V			4/16513	0.024

* indicate site(s) included in predicted epitope(s).

Table 2. Amino acids of SARS-CoV-2 S protein under diversifying selection pressure

Reference	Position	Variants					Clade	Proportion	%
D	215	N	H	G	S	Y	5	12/17556	0.068
Q#	675	K	R	H	S		1, 5	30/17709	0.169
V	615	I	L	F			1, 3	16/17799	0.090
S	221	P	L	W			5	16/17560	0.091
Q	239	K	R	H			5	16/17577	0.091
V	**483**	**A**	**F**	**I**			**1, 5**	**33/17025**	**0.194**
V	622	I	F	L			3	16/17792	0.090
S	943	T	I	P			2	28/17689	0.158
A*	1078	S	V	T			1, 4	27/17782	0.152
H	49	Y	Q				1, 2	36/17722	0.203
N	**354**	**K**	**D**				**1**	**5/17668**	**0.028**
H	**519**	**Q**	**P**				**1, 2**	**3/17014**	**0.018**
A	653	S	V				3	3/17734	0.017
A#	684	T	V				3	5/17701	0.028
A	771	S	V				6	9/17732	0.051
A	892	S	V				1	6/17761	0.034
D	936	Y	H				2	95/17750	0.535
S	940	F	T				2	7/17747	0.039
G	1167	S	V				1	5/17755	0.028
K	1192	N	Q				1	8/17729	0.045
L	5	F					2, 6	102/17416	0.586
L	8	V					1	54/17445	0.309
A	288	S					4	4/16067	0.025
E	309	Q					1, 4	6/15759	0.038
V	**367**	**F**					**1, 3**	**21/17577**	**0.119**
N	**439**	**K**					**5**	**40/17282**	**0.280**
G	**476**	**S**					**1, 5**	**10/17027**	**0.059**
S	**494**	**P**					**5**	**6/17023**	**0.035**
D	614	G					1, 2, 3, 6	11326/17744	63.830
A	706	V					1,3	12/17695	0.068
A	771	V					1	9/17729	0.051
G	1124	V					1, 3	50/17747	0.282

Sites in bold are localized within the RBD. * indicate site(s) included in predicted epitopes, and # sites included in the furin cleavage region.

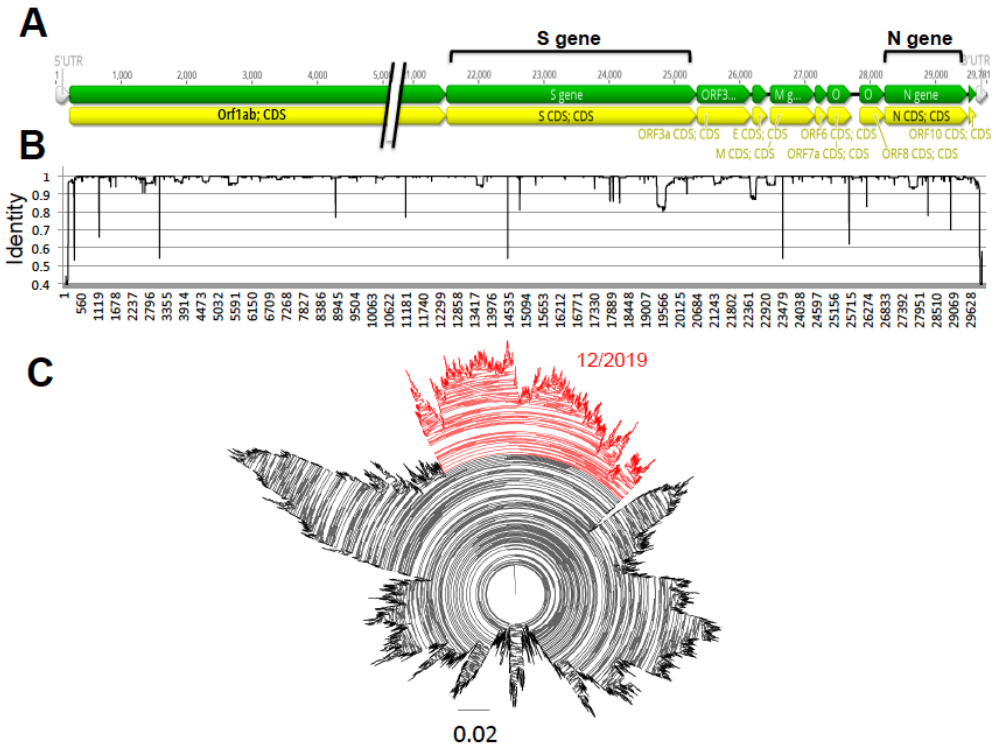

Figure 1. Diversity of SARS-CoV-2 genome.
(**A**) Diagram of SARS-CoV-2 genome organization. Position of the S and N genes is highlighted. (**B**) Sliding window analysis of nucleotide identity along SARS-CoV-2 genome. (**C**) Phylogenetic analysis of 17,853 SARS-CoV-2 genomes. Highlighted in red is the clade that includes the earliest sequences derived from human cases in December 2019.

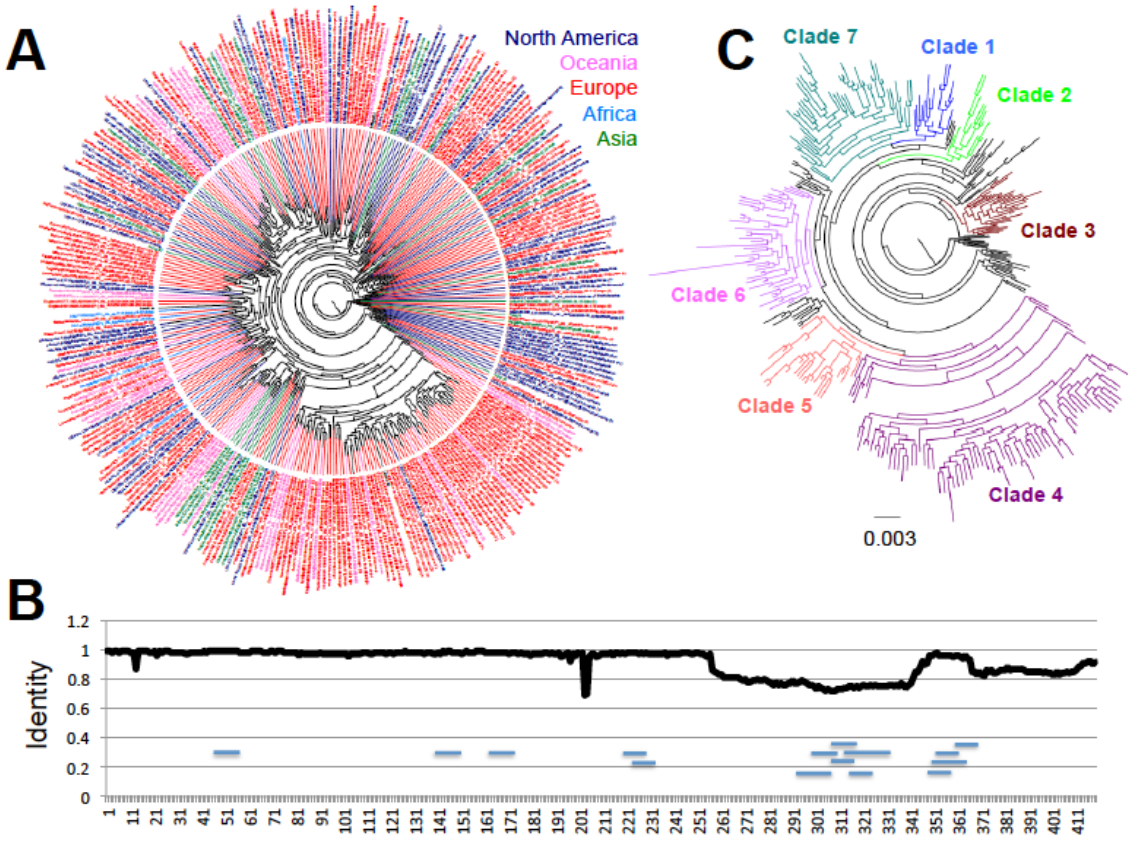

Figure 2. Sequence diversity of SARS-CoV-2 N antigen
(**A**) Phylogenetic analysis of 326 unique N protein sequence variants, color-coded according to region of origin. (**B**) Sliding window analysis of sequence identity along the N protein sequence. Small horizontal lines within the sequence indicate the position of predicted epitopes. (**C**) Phylogenetic analysis showing the identified incipient clades. Early sequences from Wuhan, China from December 2019 are included in Clade 1.

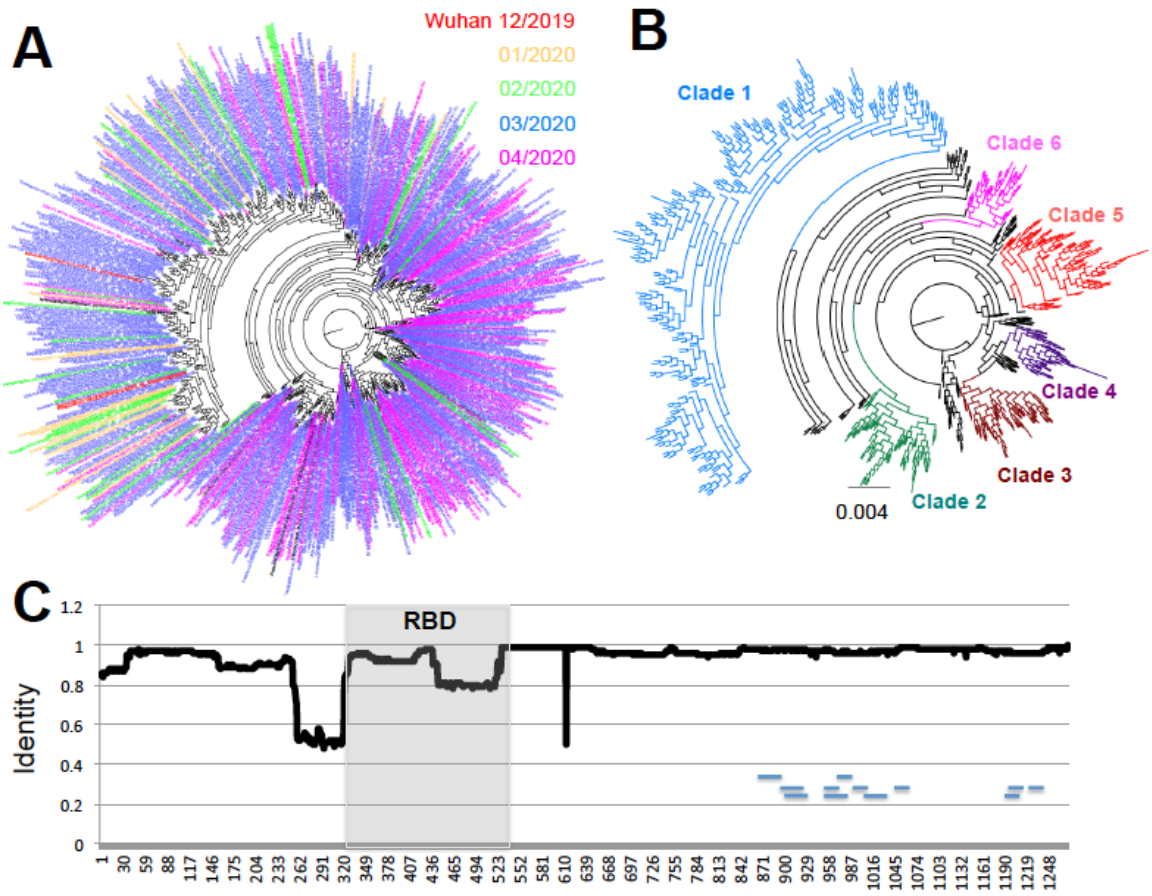

Figure 3. Sequence diversity of SARS-CoV-2 S antigen
(**A**) Phylogenetic analysis of 681 unique S antigen sequence variants, color-coded according to date of identification. (**B**) Phylogenetic analysis showing the identified incipient clades. Early sequences from Wuhan, China from December 2019 are included in Clade 1. (**C**) Sliding window analysis of sequence identity along the S protein sequence. Small horizontal lines within the sequence indicate the position of predicted epitopes. The RBD is highlighted in gray.

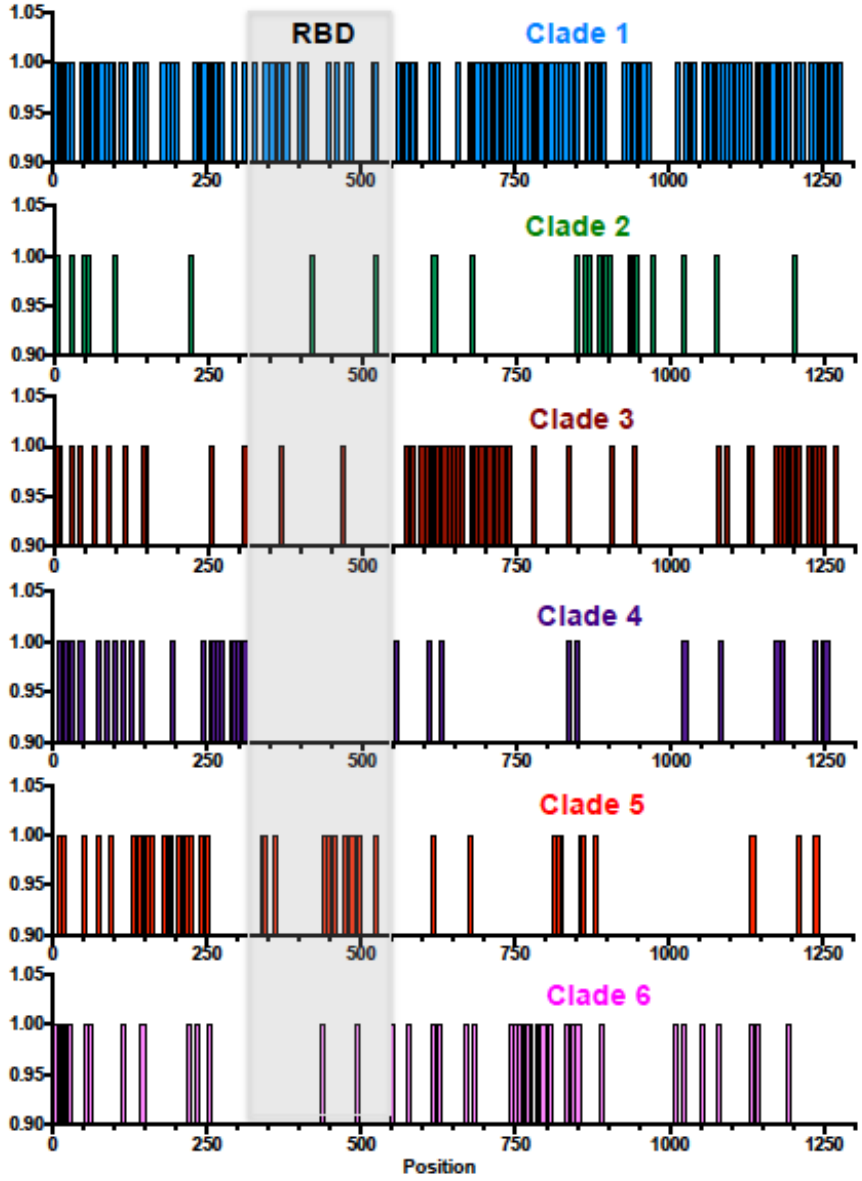

Figure 4. Location of variant sites along the S antigen sequence for each clade. Each vertical bar indicates a variant site. The position of the RBD within the S antigen is highlighted in light gray.

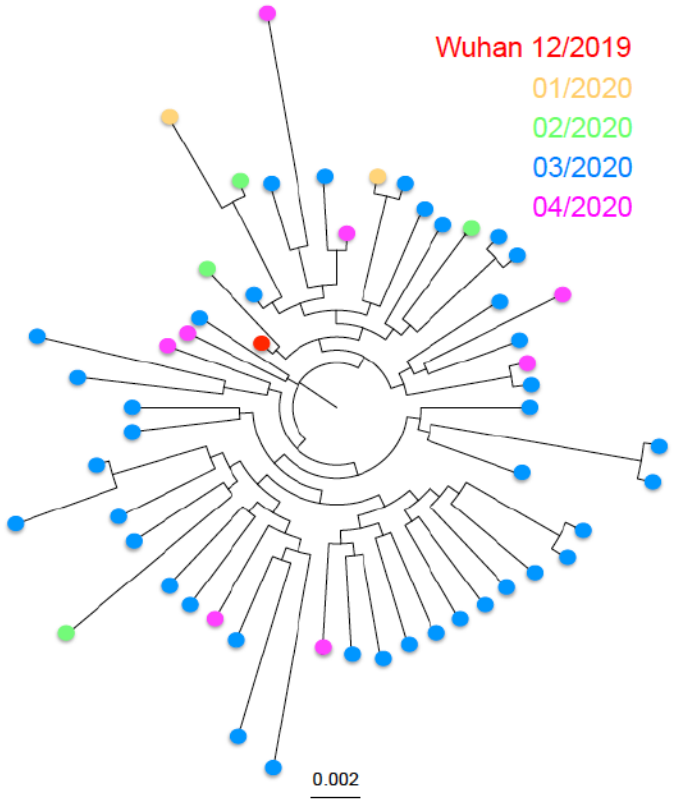

Figure 5. Phylogenetic analysis of RBD sequence variants.
Variants are color-coded according to the date of isolation of the sequence.