Snapshot of the evolution and mutation patterns of SARS-CoV-2

Jin Zhao^{1,a}, Jiumeng Sun^{1,a}, Wan-Ting He^{1,a}, Xiang Ji³, Qi Gao¹, Xiaofeng Zhai¹, Marc A. Suchard⁵, Samuel L. Hong², Guy Baele², Shuo Su¹, Jiyong Zhou⁴ and Michael Veit⁶

1. Institute of Immunology, College of Veterinary Medicine, Nanjing Agricultural University, Nanjing, China.
2. Department of Microbiology, Immunology and Transplantation, Rega Institute, Laboratory for Clinical and Epidemiological Virology, KU Leuven, Leuven, Belgium
3. Department of Mathematics, Tulane University, New Orleans, USA
4. Key laboratory of Animal Virology of Ministry of Agriculture, Zhejiang University, Hangzhou, China.
5. Departments of Biomathematics and of Human Genetics, David Geffen School of Medicine at UCLA, and Department of Biostatistics, UCLA Fielding School of Public Health, University of California, Los Angeles, USA.
6. Institute for Virology, Center for Infection Medicine, Veterinary Faculty, Free University Berlin, Germany.

^a co-first author

The manuscript is under reviewing and any suggestions regarding the article should be sent to Prof. Jiyong Zhou (E-mail: jyzhou@zju.edu.cn) or Jin Zhao (E-mail: zhaojinvet@163.com), who represent all authors.

Abstract

The COVID-19 pandemic is the most important public health threat in recent history. Here we study how its causal agent, SARS-CoV-2, has diversified genetically since its first emergence in December 2019. We have created a pipeline combining both phylogenetic and structural analysis to identify possible human-adaptation related mutations in a data set consisting of 4,894 SARS-CoV-2 complete genome sequences. Although the phylogenetic diversity of SARS-CoV-2 is low, the whole genome phylogenetic tree can be divided into five clusters/clades based on the tree topology

and clustering of specific mutations, but its branches exhibit low genetic distance and bootstrap support values. We also identified 11 residues that are high-frequency substitutions, with four of them currently showing some signal for potential positive selection. These fast-evolving sites are in the non-structural proteins nsp2, nsp5 (3CL-protease), nsp6, nsp12 (polymerase) and nsp13 (helicase), in accessory proteins (ORF3a, ORF8) and in the structural proteins N and S. Temporal and spatial analysis of these potentially adaptive mutations revealed that the incidence of some of these sites was declining after having reached an (often local) peak, whereas the frequency of other sites is continually increasing and now exhibit a worldwide distribution. Structural analysis revealed that the mutations are located on the surface of the proteins that modulate biochemical properties. We speculate that this improves binding to cellular proteins and hence represents fine-tuning of adaptation to human cells. Our study has implications for the design of biochemical and clinical experiments to assess whether important properties of SARS-CoV-2 have changed during the epidemic.

Introduction

Coronavirus disease 2019 (COVID-19), caused by a *Betacoronavirus* named Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2), was initially reported in China in late December 2019^{1,2}. As of the 20th of Jun 2020, 8,525,042 infections and 456,973 deaths had been reported worldwide (data from WHO Situation Report). SARS-CoV-2 is a positive stranded enveloped RNA virus with a genome size of approximately 30 kb^{1,3}. The virus genome comprises 14 open reading frames (ORFs) encoding 27 proteins. Virus particles are composed of four structural proteins: the Spike (S) protein, which is required for virus entry, the main determinant of host tropism and the target for neutralizing antibodies, the envelope (E) protein (an ion channel) and the matrix (M) protein are inserted into the viral envelope, whereas the nucleocapsid (N) protein is internal and enwraps the viral genome⁴. Some of the 8 accessory proteins (ORFs 3a, 3b, p6, 7a, 7b, 8b, 9b, 14) might also be virus

components, but their functions are largely unexplored³. The 16 non-structural proteins (nsp1 to nsp16) are generated by proteolytic cleavage from two polyprotein precursors, pp1a and pp1ab, which are encoded by ORF1a and ORF1ab, respectively. ORF1ab accounts for more than 2/3 of the whole genome and is generated by ribosomal frame shifting during translation. The enzymes required for polyprotein processing are encoded by nsp3, a large multifunctional protein having a papain-like protease domain, and by the main protease (nsp5, 3C-like protease), which cleaves pp1a and pp1ab at eleven sites^{5,6}. Most nsp proteins assemble into a membrane-bound replication and transcription complex. Anchoring of the complex to double-membrane vesicles that are derived from the endoplasmic reticulum membrane is achieved by the transmembrane proteins nsp4 and nsp6. The main enzymes for RNA processing are the RNA-dependent RNA polymerase (RdRp, nsp12) and the helicase (nsp13), but they are assisted by nsp7 to nsp16. Nsp1 and nsp2 are thought to modulate host cell responses^{7,8}. Thus, most of the proteins encoded by ORF1a and ORF1ab are essential for virus replication and - at least for *Deltacoronaviruses* - for adaption of the virus to a new host⁹.

RNA viruses are known to be rapidly evolving, which could lead to the accumulation of amino acid mutations that might affect the transmissibility of the virus, its cell tropism and pathogenicity. Fortunately, until now, the observed diversity among pandemic SARS-CoV-2 sequences is low, but its rapid global spread may provide the possibility of positive natural selection. The lack of pre-existing immunity in the population and the high transmissibility of SARS-CoV-2 render its spontaneous disappearance unlikely^{10,11}. Furthermore, it is not known whether SARS-CoV-2 is already fully adapted for efficient growth in human cells after its host-jump from bats or from a putative intermediate host^{2,12}. This is usually associated with adaptive mutations as the analysis of genomes from early, middle and late cases during the SARS-CoV-1 epidemic revealed^{13,14}. For avian H5N1 Influenza viruses it was shown that transmissibility of the virus is mainly associated with mutations in one polymerase protein and in the hemagglutinin (HA); the latter increased the stability of

the molecule and its receptor binding affinity^{15,16}. HA is also the target of antibodies, but during the flu season mutations slowly accumulate in HA and antibody escape mutants are generated that become the predominant strain¹⁷. Therefore, insights from the evolution of SARS-CoV-2 will provide a better understanding of the epidemic and offer important implications for the development of vaccines.

So far, some studies have reported that SARS-CoV-2 has evolved into new genotypes /subtypes, but those analyses of the phylogenetic tree were based on only a few hundred sequences¹⁸⁻²⁰. Here, we developed an early warning pipeline to study SARS-CoV-2 evolution within the context of the pandemic. We used 4,894 SARS-CoV-2 sequences available from GISAID and NCBI GenBank (data release: April 11, 2020) to perform phylogenetic analysis based on whole genomes. In particular, we identified and characterized the geographic and temporal patterns of high-frequency mutations. Some of the mutations exhibit a dynamically changing pattern in their frequencies at different geographic regions which might indicate a potential signal of positive selection in the genes encoding ORF1ab and the structural proteins.

Materials and Methods

Sequence collection

A total of 6,193 complete genomic sequences of SARS-CoV-2 covering a time range from December 24, 2019 (the first isolate) to April 11, 2020 were downloaded from GISAID (<https://www.gisaid.org/>), including detailed sequence information (accession ID, virus name, location and collection date). Sequences with more than 300 unidentified bases were considered as low quality and were filtered out as well as sequences from cell passage isolates. The remaining 4,894 high-quality sequences which are at least 28,000bp in length, exhibit high coverage and were subsequently used for our analyses.

Sequence Alignment

The 4,894 complete genome sequences of SARS-CoV-2 were aligned with MAFFT v7²¹ using the G-INS-i strategy and manually revised by using MEGA 7.0²². After aligning these genomes, we extracted the ORF1a (13218bp), ORF1ab (8072bp), S (3822bp), ORF3a (828bp), E (228bp), M (669bp), ORF6 (186bp), ORF7a (366bp), ORF7b (132bp), ORF8 (366bp), N (1260bp) and ORF10 (117bp) from each genome based on the reference strain (Accession NC_045512.2), and translated them into the corresponding amino acid sequences. For the next analysis, the twelve amino acid sequence sets were aligned by MEGA 7.0 using MUSCLE (Codons).

Phylogenetic analysis of genome sequences, ORF1ab and S gene of SARS-CoV-2

The 4,894 complete genome, ORF1ab and S genes were used to reconstruct the maximum-likelihood (ML) phylogenetic tree using IQ-TREE v2.0.2²³. The most suitable nucleotide substitution model was selected by the automatic model selection functionality in IQ-TREE²⁴. In order to perform a more thorough search through tree space, we used both default IQ-TREE settings as well as additional parameters to intensify the search algorithm (allnni, ntop=100 and nbest=20). Multiple independent IQ-TREE analyses with these settings were performed and the tree with the highest (log) likelihood among these runs was retained. UltraFast bootstrap analyses were performed using 1,000 bootstrap trees. Trees were displayed by iTOL (<https://itol.embl.de/>) and colored by continent and country of each sequence²⁵. The clusters were manually divided according to the topology of the whole genome phylogenetic tree, and then the genetic distance between clusters was calculated by MEGA 7.0 using the function “Compute Between Group Mean Distance” with the following parameters: Variance Estimation Method is none, the Substitutions Type is nucleotide, the Model/Method is p-distance and other parameters were default.

Evolutionary analysis by determination of the ratio of non-synonymous versus synonymous nucleotide substitutions

The HyPhy software package was used to estimate the ratio of non-synonymous substitutions versus synonymous substitutions (dN/dS) and to identify the sites that are subjected to potential positive selection²⁶. Phylogenetic trees of ORF1ab and S gene were used as reference trees in selection analysis, because ORF1ab has the most high-frequency mutation sites, and plays, besides S, an important role for virus replication. The following methods were used for selection analysis: SLAC (Single Likelihood Ancestry Counting), FEL (Fixed Impact Probability), MEME (Mixed Evolutionary Model of Evolution) and FUBAR (Fast Unconstrained Bayesian Approximation)²⁷⁻³⁰. We set stricter standards than usual for statistical significance, the *p* values of SLAC, FEL and MEME ≤ 0.05 and FUBAR posterior probability ≥ 0.95 .

Structural analysis

The software PyMol (<https://pymol.org/2/>) was used to create the figures from the pdb files. The integrated measuring wizard was used to determine the distance between two atoms. The electrostatic surface potential was calculated with the APBS plugin. Visualization of hydrophobicity and charge on protein surfaces was done with the YRB script³¹. Exchange of certain amino acids was performed with the integrated mutagenesis tool. Among the different possible rotamers of the mutated amino acid side chain the one was chosen that exhibits no clashes with neighboring amino acids. Prediction of phosphorylation sites was done with the NetPhos 3.1 tool (<http://www.cbs.dtu.dk/services/NetPhos/>).

Result

1. Evolution of SARS-COV-2 and high-frequency amino acid mutations in SARS-COV-2 structural and non-structural proteins

We obtained 4,894 whole genome sequences from 60 countries and 6 continents on which we performed phylogenetic analysis (Fig. 1 and Fig. S1). The whole genome phylogenetic tree can be divided into five clusters/clades based on the tree topology and clustering of mutations. In these five clusters, the high-frequency mutation sites (with mutations greater than 1% in our data set of 4894 sequences) in each cluster are almost entirely different (Fig. S1). Each cluster also has a different geographical distribution: cluster I mainly consists of sequences from Asia and Europe, cluster II is mainly composed of sequences from North America and Asia. In cluster III, sequences from North America and Europe are alternately distributed, whereas clusters IV and V are dominated by European sequences. We found that the sequences from China gathered in clusters I and II, and the sequences from the United States clustered in clusters II and III. Sequences from European countries, such as the United Kingdom exist in each cluster, but are more prominently found in clusters I, IV and V. This indicates that there are differences in epidemic isolates of different regions. Of the five clusters, clusters IV and V have the smallest genetic distance to each other in the range of 10^{-5} substitutions per site. Cluster I contains the early Chinese sequences and all sequences from 2019, and exhibits the largest genetic distance to cluster II, which, however, is also small, in the range of 10^{-4} substitutions per site. It is worth noting that although the clustering of the whole genome phylogenetic tree is relatively clear visually and each cluster has its own characteristic mutation sites, the between group mean p-distance of five clusters is too small (Table S1) and its bootstrap value is too low to support genotyping, therefore, we did not perform more meticulous genotyping (Fig. S1).

The ORF1ab and whole genome phylogenetic trees have a lot in common. Comparing the structures of the trees revealed that the ORF1ab phylogenetic tree (Fig. 2) is visually similar to the whole genome phylogenetic tree, but the S protein phylogenetic

tree is quite different and exhibits very low diversity and also low bootstrap support on internal nodes (Fig. S2). In terms of national distribution, the whole genome and ORF1ab phylogenetic tree show country aggregation, whereas the S protein phylogenetic tree is relatively scattered. This may be due to the high homogeneity of the S protein sequences, resulting in poor diversity. Like the whole genome tree, the ORF1ab and S gene phylogenetic trees do not support genotyping because of the low bootstrap values. It needs to be explained that the S protein phylogenetic tree exhibits low bootstrap support and low diversity, and has no significant features except for the Asp614Gly mutation site that has been reported many times. This shows that there is limited genetic variation in the currently sampled viruses, but more recent ones are showing more divergence in location as expected for fast evolving RNA viruses.

Since nonsynonymous changes are more likely to associate with functional adaptations of viruses, we searched for amino acids substitutions in our collection of 4,894 whole genome sequences. We identified 11 amino acids which are mutated with high frequency (more than 10%) in SARS-CoV-2 sequences. Two of the sites are located on ORF1a (nsp2, nsp6), three on ORF1ab (nsp12, nsp13), one on S, two on ORF3a, one on ORF8 and two on N (Table 1). On the ORF1ab phylogenetic tree, five high-frequency mutant strains show branch aggregation and one site (nsp5 Lyr90Arg, mutation frequency more than 1%) shows a scattered distribution (Fig. 2). Whereas four of the five high-frequency mutant strains can almost cover one cluster and the other (nsp12 Pro323Leu) is very widely distributed, covering almost three clusters. Note that the two mutations in nsp13 are clustered together and occurred mainly in North America, the other mutations are present in different clusters of various geographic origin.

Figure 3 shows a more detailed temporal and spatial analysis: the percentage of a certain mutation in all sampled sequences and in sequences from Asia, Europe and North America is plotted against the time period when the sequences were deposited.

Some mutant viruses appeared with a high but declining frequency in January in certain regions that presented pronounced local heterogeneity. For example, ORF3a

Gly251Val was highly abundant in Europe in the second half of January (71% versus 5-11% in Asia and North America), whereas the mutation ORF8 Leu84Ser was very frequent in North America and in Asia in the second half of January (40-60%), but absent in Europe (0%). The mutation in nsp6 revealed a more fluctuating pattern with local variation (71% in Asia, but only 14% in Europe and North America), but this isolate is now of low abundance throughout the world.

Other mutations occurred first in February with declining frequencies and notable local heterogeneity. Both mutations in nsp13 appeared in North America in the second half of February at high frequency (~46%), which increased to >55% and then dropped sharply to 11%. These mutations in nsp13 were observed only at very low frequency in Europe (<1%) and not at all in Asia.

Interestingly, some mutations presented an increasing abundance over time. For example, the mutations in S and nsp12 are highly abundant in all regions (>68%) in the second half of March as also observed in another study published very recently³². An increase in the frequency of the mutation in NSP12 and S first occurred in China in the second half of January. The mutations in ORF3a Gln57His and nsp2 Thr85Ile reached a high frequency in North America (>60%) in March but remained lower in Europe and Asia (9-16%). Both mutations in the N-protein are also increasing in frequency, first in Europe (up to 30% already in February) and more recently in Asia (26% in March).

Note that the graphs for S and nsp12 are similar to each other in all regions, and to the graphs of both mutations in the N-protein. When analyzing individual sequences, we found that the two mutations in the N-protein as well as the mutations in S and nsp12 very often appeared together in one genome. The strong association of the mutation sites S D614G and nsp12 P323L (together with a C-to-T mutation at position 241 of the 5' UTR and a silent C-to-T nucleotide exchange at position 3,037) has been described also in another study published very recently³². We also observed quadruple mutations: viruses with the mutation in S and nsp12 acquired the two mutations in the N gene, but the additional mutations were not always retained. The same pattern is

consistent for the mutations T85I in nsp2 and Q57H in ORF3a that often occurred together in one genome and also as a quadruple mutation together with the mutations in S and nsp12. Thus, some of the mutations are inherited together explaining their similar local and spatial occurrence.

There are 17 additional amino acid exchanges if we lower the threshold value ($>1\%$), present in all genes, except for the genes encoding E, ORF7a, ORF7b, and ORF10. The additional mutation sites are mostly located in ORF1ab with one in ORF2, one in ORF8 and two in the M and N genes (Table S2). Of note, except for the highly abundant Asp614Gly exchange, there were no further changes identified in the spike protein. Further tracking the occurrence of these mutations in the ongoing pandemic might provide additional insights into their roles.

In addition, we found that ten of the eleven high-frequency amino acid substitutions were caused by a one nucleotide change, 7x at the second position, 2x at the third position and 1x at the first position of a codon. Only one mutant exhibits an exchange of two nucleotides: Arg encoded by AGG changed to AAA specifying Lys at amino acid position 203 of the N gene. Since the viral polymerase makes at most one nucleotide substitution per replication cycle, a double nucleotide exchange is unusual.

In all 12 mutation sites the variable residue was always exchanged to the same amino acid, although more replacements would be possible by a one nucleotide exchange. Only four of the amino acid mutations are conservative replacements, such as the exchange of the basic residues Arg and Lys and of the hydrophobic residues Leu to Phe. Five exchanges involve the unusual amino acids Gly and Pro, which destabilize secondary structures, and are not able to interact with other amino acids.

In order to explore a potential adaptation of SARS-CoV-2 to humans, we estimated dN/dS ratios for the ORF1ab and S protein phylogenetic trees using four algorithms (Fig. 2 and Fig. S2). Interestingly, some high-frequency mutations exhibit dN/dS >1 : Thr85Ile in nsp2, Leu37Phe in nsp6, Asp614Gly in S and Pro323Leu in nsp12. Note that the frequency of three of these sites is increasing in almost all continents since

their first appearance, and that only Leu37Phe in nsp6 is decreasing in frequency (Fig. 3). Another site, Lys90Arg in nsp5 was also identified by the dN/dS analysis, which has a mutation frequency higher than 1% (Fig. 2b). Thus, this suggests that a certain number of sites within SARS-CoV-2 have potentially undergone positive selection. However, the increase in the frequency may coincide with the transmission pattern of the virus in different regions and further analysis is needed to substantiate this preliminary conclusion.

2. Structural analysis of frequently exchanged sites in non-structural proteins

Nsp2, a protein that modulates host responses, contains one site with exchanges greater than 15%, the exchange Thr85Leu, the site might be positively selected, but no structural information is available for nsp2 of any β -coronavirus to specify any function for this site.

One high-frequency site for which some evidence of selection is available can be found in **nsp6**, a membrane protein with six transmembrane spans which induces the formation of autophagosomes from the ER membrane³³. The mutation we found (Leu37Phe) is located at the end of the first transmembrane region directed to the lumen of the ER. The physiological role of this exchange is obscure since residue 37 is already preceded by three Phe, i.e. the amino acid sequence changes from LFFFLY to LFFFFY.

One site with mutation frequency greater than 50% was identified in **nsp12**, the RNA-dependent RNA polymerase (RdRp)⁷. Nsp12 forms a complex with one copy of nsp7 and two copies of nsp8 (Fig. 4a). It is divided into the catalytic part common to all RNA-viruses (green colours) containing the canonical “palm”, “finger” and “thumb” domains and N-terminal extension unique to Nidoviruses, which is composed of a domain with nucleotide transferase activity (NiRan, magenta) and an interface domain (blue)³⁴. The selected site, residue 323, is located in the interface domain at the surface of a cavity at the backside of the polymerase complex. The back

wall of the cavity is built by the “finger” domain, its left and right part are formed by nsp8 and the interface region, respectively (Fig. 4b). The opening of the cavity is flanked by two basic amino acids and its back contains several hydrophobic residues (Fig. 4c). Proline 323 is located in a short, twisted α -helix; its side chain is exposed at the surface of the cavity (Fig. 4d). Exchanging proline with the hydrophobic leucine is increasing the hydrophobic character of the cavity.

Two identified high-frequency mutation sites were located on **nsp13**, a helicase that unwinds DNA or RNA duplexes⁸. We used the structure of the highly similar nsp13 of MERS to highlight the location of the residues³⁵. Nsp13 is composed of a N-terminal CH domain, which is connected by a stalk domain to two RecA domains. The two exchanged residues, proline 504 and tyrosine 541, are located at the surface of the RecA2 domain (Fig. 5a). The side chain of Tyr 541 runs parallel to the surface and forms a hydrogen bond with Gln531, which is lost if substituted by a cysteine. The main chain of Pro 504 interacts with Ser507. This interaction is preserved when Pro504 is exchanged by a Leu, but the substitution exposes a hydrophobic moiety at the surface of NSP13. Thus, both exchanges might alter the surface properties of nsp13, perhaps to modify binding to other viral or cellular factors.

One site which is possibly positively selected, but exhibits a low mutation frequency, is located in **nsp5**, the main protease of coronaviruses. Nsp5 exists as a dimer and dimer formation is essential for catalytic activity³⁶. Fig. 6c shows the structure of an nsp5 monomer from SARS-CoV-2 with the active center (His41, Cys145) shown in red and other residues involved in substrate binding as orange sticks³⁴. They are located in domain II which is connected by a long loop to domain III which contains amino acids important for dimerization (shown as blue sticks). The selected site, lysine 90, is located in domain I at the other side of the molecule. The arginine is part of a β -sheet, its side chain is exposed at the surface of the molecule where it forms a hydrogen bond with Asp34 (Fig. 6d). The exchange of Lys90 by the other basic amino acid Arg does not change the electrostatic properties of the molecule's surface, but the hydrogen bond to Asp34 might be lost, since the side chain of Arg is longer (Fig. 6e).

3. Structural analysis of frequently exchanged sites in accessory proteins and virus components

Two frequent exchanges were identified in **ORF3a**, a homotetrameric ion channel of unknown structure that also activates inflammatory responses of the host^{37,38}. The first exchange Gln57His is located at the cytosolic end of the first transmembrane region and the second exchange Gly251Val in the cytoplasmic region.

ORF8, a small glycoprotein of unknown function, contains one frequently exchanged site at position 84, a hydrophilic serine is exchanged by hydrophobic leucine. Note that ORF8 often acquires mutations after the host jump of a SARS virus. A 382-nucleotide deletion covering almost the entire ORF8 was reported for SARS-CoV-2 isolates from eight hospitalized patients³⁹. Likewise, for ORF8 of SARS-CoV-1 it has been reported that a 23-nucleotide deletion acquired early during human-to-human transmission attenuates virus replication in cell culture⁴⁰.

Two sites were identified in the **N protein**, which packages the viral RNA into ribonucleoparticles. N consists of a N-terminal RNA-binding domain and a C-terminal oligomerization domain, which are connected by an unstructured linker region that harbours the two penultimate exchanges⁴¹. It contains an Arg/Ser rich region with various phosphorylation sites, but the exchanges do not change the probability for phosphorylation of any of the sites.

The S protein contains one selected exchange (Asp614Gly), which is located in the S1 subunit between the receptor-binding domain (RBD) and the furin cleavage site (Fig 6a). S2 harbours the fusion machinery of the viral spike: a hydrophobic fusion peptide adjacent to the second proteolytic cleavage site (S2) and a heptad repeat regions (HR) (Fig. 7b). Asp614 forms the apex of a loop that links a short β -strand to an unresolved, hence probably flexible part of S1. The loop contains a used N-glycosylation site, Asn 616. It is covalently connected to another loop by a disulphide-linkage between Cys 617 and Cys 649 and via interaction of the main chain atoms of Asp614 with Ala647. Both loops are at the interface between two monomers of the trimeric spike protein. Interestingly, the carboxyl group of Asp 614

forms a hydrogen bond with the -OH group of Thr859, which is part of a short β -strand in the S2 subunit between the fusion peptide and the heptad region 1 and is (Fig. 7b). S of SARS-CoV-1 also exhibits a similar interaction between monomers, Asp 600 forms a hydrogen bond with Thr841 and the other elements characteristic for this region, are also present (Fig. 7c).

Individual monomers of the prefusion structure of S exists in several conformations: The RBD could be either hidden from receptor engagement (“down” conformation) or exposed by a hinge-like movement at the top of the molecule (“up”) that would allow receptor-binding. Interestingly, the distance between the carboxyl group of Asp and the -OH group of threonine is the shortest (2.7Å) if Asp is a component of an up-monomer (Fig. 7d). The distance becomes larger (3.4Å) if Thr is present in an up monomer (Fig. 7e) or if both monomers are in the down conformation (4.4Å) (Fig. 7f). Analysis of resolved structures of S of SARS-CoV-1 showed that binding to the ACE2 receptor or to antibodies, which both move the RBD into the up conformation, increased the distance between Asp and Thr to more than 5 Å^{42,43}, which is too large to maintain the hydrogen bond between Asp and Thr⁴⁴. This suggests that exposing the RBD at the top of the molecule is accompanied by small movements of the S1 and S2 subunits against each other. The exchange of Asp614 by a glycine would prevent hydrogen bond formation and thus might facilitate exposure of the RBD.

In the S-protein, we also searched for sites which are exchanged in only ten or more sequences. However, none of the amino acids making up the epitopes of neutralizing monoclonal antibodies generated from memory B-cells of human SARS-CoV-1 or SARS-CoV-2 patients were detected^{45,46}. Likewise, there was just one mutation in the whole receptor binding domain, residue 483 changed from Val to Ala in 16 sequences. However, this residue does not directly contact the ACE receptor⁴⁷. Thus, there is no evidence that SARS-CoV-2 can easily generate antibody escape mutants or is adapting to recognize the human receptor more efficiently.

Discussion

The evolution and epidemiology of a new virus at different stages after its spread into a new host are often considered to be critical for its success in creating and sustaining an epidemic. As of 20th June 2020 more than 49,000 complete or near-complete genomic sequences of SARS-CoV-2 are publicly available and the number rises daily. These genomes provide invaluable insights into the ongoing evolution of the virus during the pandemic, which might be helpful to eventually mitigate and control the virus spread. Until now, there is no official definition of the SARS-CoV-2 phylogenetic lineages, but a few articles have proposed methods of classification. Although the classification methods are diverse, the method of Rambaut et al [52] is the current mainstream which clearly defines the principles of lineage classification. We reconstructed phylogenetic trees based on early 4,894 full genomes, ORF1ab and S gene sequences, which is by now the one of the most comprehensive analysis of the early epidemic/mutations characteristic of SARS-COV-2. Using full genomes we identified five major clusters in the phylogenetic tree which is similar to other observations

(<https://virological.org/t/year-letter-genetic-clade-naming-for-sars-cov-2-on-nextstain-org/498>), but it is too early to denote new genotypes of SARS-COV-2, because sequences are highly homogeneous and the phylogenetic trees exhibit low bootstraps values and no significant changes in clinical symptoms or transmissibility have been reported so far. Although some manuscripts report that SARS-COV-2 has produced new genotypes¹⁸⁻²⁰, our phylogenetic analysis shows that the genetic diversity of SARS-COV-2 is relatively low during this early stage of the epidemic, the viral genome is largely stable and the virus did not evolve rapidly after its emergence in humans. However, each of the five clusters has its own mutations that may serve as targets for fast genotyping of samples from patients. Our nomenclature may complement the system suggested by Rambaut et al⁴⁸ which proposed a dynamic system for labeling transient lineages that have local epidemiological significance.

In addition, on the amino acid level we identified eleven mutations that occur in more

than 10%, two of them even in more than 50% of the analyzed SARS-CoV-2 sequences (Table 1). Although some mutations have been reported before⁴⁹⁻⁵², most of the high frequency mutations in ORF1ab are new findings.

We analyzed these mutations further by using four algorithms to calculate dN/dS for each site. We found that four of them exhibit a dN/dS larger than 1, three are in ORF1ab (Thr85Ile in nsp2, Leu37Phe in nsp6, Pro323Leu in nsp12) and one in the S protein (Asp641Gly) (Table S3). Other sites also exhibit a dN/dS value greater than 1, but the frequency of these exchanges is currently very low. Although these sites have selection signals, the selection analysis has some limitations and hence we consider the results as preliminary. It has been reported that some mutations that seemingly arise multiple times along the phylogenetic tree may be caused by sequencing error and/or are the result of either artefactual lab recombination, or potential hypermutation⁵³ (<https://virological.org/t/issues-with-sars-cov-2-sequencing-data/473>). In addition, during an ongoing pandemic, purifying selection signals occur frequently and recombination events between viruses might obscure the signal⁵³. Furthermore, we cannot exclude that some of the exchanges are the result of non-selective conditions, i.e. due to a “genetic bottleneck”. A patient releases droplet that contain only a limited virus population which does not representative of the whole virus “swarm” replicating in his body. One droplet that by chance does not contain a single particle representing the original strain might then infect another person where it creates a new and different virus population by a ‘founder’ effect. However, since the pandemic continues and new viruses will be sequenced, it is worthwhile to analyze whether one of the sites become positively selected.

In summary, our study revealed that in the early large genome of SARS-CoV-2 only a few amino acids are exchanged and hence the selection pressure is low, which is consistent with the conclusion in the good sequence analysis reported by MacLean et al⁵³.

We also performed a spatial and temporal analysis that revealed the time and region at which a certain mutation appeared first and how it spread to other continents together

with its frequency change in the sampled population. The mutations in nsp6, nsp13, ORF3a 251 and ORF8 exhibit high-frequency peaks early in the epidemic and only in certain regions, but then showed a sharp drop and are of low abundance now. In contrast, the frequency of the mutations in nsp2, nsp12, ORF3a 57, N and S increased since their introduction into the viral genome. They are present at high abundance in the second half of March, either worldwide (S, nsp12, >68%) or in certain continents (N in Europe and Asia, ~25%, nsp2 and ORF3a in North America, >60%).

For the viral proteins where 3D structural information is available, we performed an in-depth analysis of possible consequences of these amino acid exchanges. One conclusion is that none of the exchanges is presumed to disturb the folding of the protein. The selected amino acids are located either in loops (Asp614 in S) that are flexible and could probably accommodate many amino acids or at the edge of β -sheets (Lys90 in nsp5, Tyr451 in nsp13).

Another similarity between the identified sites is that they are located at or very close to the surface of the molecule. The crystal structures revealed that the side chains of three of the sites interact with other amino acid in close proximity. Lys90 in nsp5 forms a salt bridge with Asp34, whereas Tyr 541 in nsp13 and Asp614 in S form hydrogen bonds with Gly531 and Thr859, respectively. In all three cases the interaction is destroyed by the substituted amino acid.

Two other frequently exchanged residues are prolines, which do not have a side chain that can interact with other amino acids. They are substituted in nsp12 and nsp13 by a leucine residue, which then exposes a new hydrophobic moiety at the molecule's surface. Especially interesting is the exchange in nsp12, the RNA-dependent RNA polymerase, since it is potentially positively selected and occurs in more than 50% of SARS-CoV-2 sequences. Pro 323 is located on the back side of a hydrophobic cavity, its exchange by a leucine increases its hydrophobic character.

In conclusion, the exchanges in nsp2, nsp5, nsp13 and S alter the local biophysical properties at the molecule's surface and expose new amino acids, which could

potentially interact with other molecules. One might thus speculate that the identified residues are elements of an interaction surface with cellular polypeptides, e. g. one of the 323 human proteins that have been identified to interact with SARS-CoV-2 proteins⁵⁴. Since the virus has been only recently introduced into the human population, likely from bats via an intermediate host ², it is reasonable to assume that they are not perfectly adapted to human cells. The mutations thus might increase the affinity for a human polypeptide that differs in the interaction surface from the ortholog protein from the previous hosts of SARS-CoV-2. The physiological role might be to better antagonize the innate immune response of the host, but the Coronavirus proteins that fulfil this tasks have not been carefully explored ⁵⁵. In addition, improved interaction with cellular proteins might facilitate gene expression (nsp5, nsp13), RNA processing (N), signaling (nsp13, N) and especially vesicle trafficking (nsp2, nsp6, nsp13, ORF3a, ORF8), the latter perhaps to facilitate the reconfiguration of ER/Golgi membranes to create the replication and transcription complex⁵⁴.

Bat viruses are also adapted to grow at a wide range of temperatures, ambient temperature during sleeping and up to 41°C during flight ⁵⁶. Thus, proteins from bat viruses must withstand higher temperatures, which occur in humans only as part of the antiviral defense mechanism. Thermostability of proteins arises from the simultaneous effect of the forces between its amino acids⁵⁷. Interestingly, two of the exchanges we identified (Tyr541Cys in nsp13 and Asp614Gly in S) likely cause a loss of a salt bridge or hydrogen bond, whereas the opposite effect, a gain of a new interaction by the substituted amino acid could not be predicted. The loss of bonds increases the local flexibility of a polypeptide chain, which might enhance protein-protein interactions ⁵⁸.

A more specific speculation can be offered for the potentially positively selected and high frequency site in the spike protein S. Asp614 in the S1 subunit forms a hydrogen bond with Thr859 in the S2 subunit of an adjacent monomer. The substitution to a glycine abolishes the hydrogen bond and thus might destabilize the trimeric spike

protein. This in turn might affect exposure of the receptor-binding domain at the top of the molecule, which is a prerequisite for binding to ACE2. The mutation might also affect the fusion activity of S, which requires dissociation of the S1 subunit from the spike before the canonical conformational change of S2 can occur that executes membrane fusion. The trigger for the conformational change is unknown, but it seems possible that a slightly destabilized trimer might be more prone to refolding compared to S having the Asp residue at position 614.

Functional consequences of the Asp614Gly exchange have recently been published. Two pseudoviruses having a Gly at position 614 grow to 2-9-fold higher titers in various cell types compared to pseudoviruses having the Asp. In addition, COVID-19 patients infected with a virus with the G614 mutation tend to exhibit higher virus load, but no association with disease severity was found. Furthermore, both pseudoviruses are neutralized with equal efficiency with convalescent sera suggesting that this locus is not critical for antibody-mediated immunity³². Therefore, although S Gly614 has a unique phenotype, more research is needed to decide whether it increases virus entry and/or transmissibility between humans.

Here, by studying 4,894 sequences from an early stage of the epidemic, we have created an analysis pipeline to track potential adaptive mutations in SARS-CoV-2. Whether any of the identified exchanges alter clinical properties of SARS-CoV-2, such as transmissibility, cell tropism, replication rate and pathogenicity needs further explorations. Usually, such important traits are controlled by multiple genes⁵⁹. Since essentially no exchanges occurred in the receptor binding domain of the spike protein and in the epitopes recognized by neutralizing antibodies, there is no evidence that SARS-CoV-2 adapts to improve receptor binding and/or to escape from the adaptive immune response. Nevertheless, it seems appropriate to follow further the evolution of SARS-CoV-2 during the ongoing pandemic.

Acknowledgements

SS, WTH, JZ, JMS, XFZ are financially supported by the National Key Research and Development Program of China (Grant no. 2016YFD0500402), National Ten Thousand Talent Program, the Fundamental Research Funds for the Central Universities (Grant no. Y0201900459), Six Talent Peaks Project of Jiangsu Province of China (Grant no. NY-045) and the Bioinformatics Center of Nanjing Agricultural University. XJ and MAS are partially supported through National Institutes of Health grant U19 AI135995. GB acknowledges support from the Interne Fondsen KU Leuven / Internal Funds KU Leuven under grant agreement C14/18/094, and the Research Foundation – Flanders (‘Fonds voor Wetenschappelijk Onderzoek – Vlaanderen’, G0E1420N).

Tables

Table 1. Frequently exchanged amino acids in SARS-COV-2 genomes

Gene	Protein	Position	Nucleotide mutation	Amino acid mutation	Mutation frequency	Geographic Sample and Major Country
ORF1a	nsp2	85 [#]	ACC→ATC	Thr→Ile	15.47%	23 Country, USA, IS, FR, AU, NL
	nsp6	37 [#]	TTG→TTT	Leu→Phe	14.02%	36 Country, UK, AU, JP, NL, USA
ORF1ab	nsp12, RdRp	323	CCT→CTT	Pro→Leu	56.10%	49 Country, USA, UK, NL, IS, BE
	nsp13	504	CCT→CTT	Pro→Leu	10.08%	5 Country, USA, AU, CA, IS, UK
		541	TAT/TAC→TGT	Tyr→Cys	10.38%	5 Country, USA, AU, CA, IS, UK
S	S1	614 [#]	GAT→GGT	Asp→Gly	56.49%	49 Country, USA, UK, NL, IS, BE
ORF3a		57	CAG→CAT	Gln→His	19.06%	28 Country, USA, FR, IS, AU, UK

	ORF3a	251	GGT→GTT	Gly→Val	10.25%	27 Country, UK, AU, NL, CH, IS
ORF8	ORF8	84	TTA→TCA	Leu→Ser	16.13%	27 Country, USA, CH, AU, CA, ES
N	N	203	AGG→AAA	Arg→Lys	16.30%	27 Country, UK, NL, BE, IS, AU
		204	GGA→CGA	Gly→Arg	16.27%	40 Country, UK, NL, BE, IS, AU

Amino acids with a mutation frequency greater than 10%.

“#” indicates that the site might be a positively selected site (see table S3)

Table S1. Between group mean p-distance of five clusters in whole genome phylogenetic tree

	Cluster I	Cluster II	Cluster III	Cluster IV	Cluster V
Cluster I					
Cluster II	1.61E-04				
Cluster III	1.04E-04	1.53E-04			
Cluster IV	1.03E-04	1.52E-04	9.52E-05		
Cluster V	8.65E-05	1.35E-04	7.83E-05	7.75E-05	

Table S2. Amino acids in the SARS-COV-2 genome exchanged with a frequency of more than 1%

gene	protein	position	original amino acid	mutant amino acid	frequency of mutation
ORF1a	nsp2	198	Val	Ile	2.02%
		212	Gly	Asp	3.09%

		559	Ile	Val	3.25%
		585	Pro	Ser	3.56%
	nsp3	58	Ala	Thr	2.87%
	nsp4	308	Phe	Tyr	1.23%
	nsp5	15	Gly	Ser	1.02%
		90	Lys	Arg	1.17%
	nsp7	25	Ser	Leu	1.37%
ORF1ab	nsp13	392	Arg	Cys	1.00%
	nsp14	320	Ala	Val	1.14%
ORF3	ORF3	196	Gly	Val	1.19%
M	M	3	Asp	Gly	1.25%
		175	Thr	Met	4.03%
ORF8	ORF8	62	Val	Leu	1.09%
N	N	193	Ser	Ile	1.68%
		197	Ser	Leu	1.21%

Table S3. Selection analysis of ORF1ab and S of SARS-CoV-2

Gene	protein	position	mutation	SLAC	FUBAR	FEL	MEME
ORF1a	nsp2	85 [#]	Thr→Ile	<0.01	>0.99	<0.01	<0.01
	nsp2	339	Gly→Ser	0.05	0.97	0.05	>0.05
	nsp3	142	Leu→Phe	>0.05	0.96	0.04	0.05

	nsp3	444	Asn→Thr	>0.05	0.99	0.02	0.03
	nsp5	90	Lys→Arg	>0.05	0.99	0.04	>0.05
	nsp6	37 [#]	Leu→Phe	<0.01	0.99	<0.01	<0.01
ORF1ab	nsp12	323 [#]	Pro→Leu	0.03	0.99	0.034	0.02
S	S1	614 [#]	Asp→Gly	<0.01	>0.99	<0.01	<0.01

“ # ” indicates that the site is also a high frequently exchanged amino acid.

p values for SLAC, FEL and MEME ≤ 0.01 , FUBAR posterior probability ≥ 0.99 are considered as putative positive selection sites.

Figure Legends

Figure 1: Geographical distribution and whole genome phylogenetic trees. Whole genome phylogenetic tree. The outer circle represents the origin of a sequence from a certain continent, and the inner circle from a certain country as specified in the inset.

Figure 2: Phylogenetic tree of ORF1ab including the location of putatively positively selected sites and sites exchanged in more than 10% of SARS-CoV-2 sequences. The outer circle represents the origin of a sequence from a certain continent, and the inner circle from a certain country, as specified at the bottom. The triangle-shaped shadow represents the position of the mutant strains in the phylogenetic tree. Different colors represent different mutation sites.

A: T85I in nsp2, a site, with a mutation frequency of 15.47%.

B: K90R in nsp5 are sites $dN/dS > 1$, but occur only in more than 1% of sequences.

C: L37F in nsp6, a site with a mutation frequency of 14.02%.

D: P323L in RdRp in nsp12, a site with a mutation frequency of 56.10%.

E: P504L in nsp13 a site, with a mutation frequency of 10.08%.

F: Y541C in nsp13 a site, with a mutation frequency of 10.38%.

Figure 3: Temporal and spatial analysis of high-frequency mutation sites. The frequency of a certain mutation in three continents (Asia, Europe, North America) and in all sequences (world) is plotted against the time period these sequences were submitted to the database. The graph at the bottom right shows the number of submitted nucleotide sequences for each time period.

Figure 4: Amino acid frequently exchanged in the polymerase complex

A: Crystal structure of the nsp12 polymerase in complex with one copy of nsp7 (coloured wheat) and two copies of nsp8 (grey). The NiRan domain of nsp12 is coloured in magenta and the interface region in blue. The location of residue 322 is shown as red sphere. The catalytic domain is coloured from green to light blue indicating the palm, finger and thumb domains. The catalytic amino acids are shown as red sticks.

B: Surface representation of the same complex. Pro 322 is coloured red. Colouring of the domains is as in A.

C: Hydrophobic cavity with the location of Pro 322 (labelled P). Three hydrophobic surface patches in the back are in yellow (from left to right: V398, A399; V535, I536; V320, F321) and two basic amino acids (K391, R654) at the entrance are in blue.

D: Detail of the structure showing the location of Pro 323 and adjacent amino acids as sticks. Figures were created with PyMol from pdb file 6M71³⁴. Inspection of the other available nsp12 structures either bound to remdesivir (pdb 7BV2,⁵⁴) or of a replicating SARS-CoV-2 polymerase (pdb 6YYT, unpublished) revealed the same results.

Figure 5: Amino acids frequently exchanged in nsp13

A: Structure of nsp13 from MERS has 72% identical and 85% similar amino acids compared to nsp13 of SARS-CoV-2. The individual domains are coloured and labelled, CH domain in blue, the stalk in light blue, the two RecA domains in light and dark green. The active site contains a sulfate (yellow sphere). Pro504 and Tyr541 are shown as red spheres.

B: Surface representation of nsp13 with Pro504 and Tyr541 and their interaction partners indicated as white sticks. The main chain of Pro504 interacts with Ser507, whereas the side chain of Tyr541 forms a hydrogen bond with Gln 531. Note that position 507 is an Arg in nsp13 of SARS-CoV-2, but this exchange is unlikely to disturb the interactions with the main chain of Pro504.

C: Exchange of Tyr541 with a cysteine destroys the hydrogen bond whereas substitution of Pro504 by Leu exposes a hydrophobic side chain at the molecule's surface. All figures were created from pdb file 5WWP³⁵. Homology modelling of nsp13 from SARS-CoV-2 using swiss model produced the same results.

Figure 6: Amino acid under positive selection in nsp3 and nsp5

A: Dimer of mac 2 and mac 3 domains of nsp3 of SARS-CoV-1, which exhibits 75% identical and 86% similar amino acids compared to the corresponding polypeptides of SARS-CoV-2. Domains which belong to one monomer are labelled blue or green. The location of the selected amino acid Asn 444 (which is a Lys at position 420 in SARS-CoV-1) is shown as red stick. Basic amino acids which are involved in poly(G) binding are shown as magenta sticks and their number is indicated in one monomer.

B: Electrostatic surface potential of a groove between monomers that runs perpendicular to the poly(G) binding cleft. Asn 444 is located at the entrance of the cleft. The bottom of the cleft contains the basic amino acids Lys462, Lys463 and Arg473. Nsp3 from SARS-CoV-2 also has basic amino acids at the positions mentioned in A and B. Figures were created with Pymol from pdb file 2W2G⁶⁰. Homology modelling of nsp3 from SARS-CoV-2 using swiss model (<https://swissmodel.expasy.org/>) produced the same results.

C: Crystal structure of a nsp5 monomer. The catalytic residues (His41, Cys145) are shown as red sticks and residues involved in substrate binding (His163, Met165, Glu166, His172) as orange sticks. Amino acids important for dimerization (Arg4, Glu290, and Arg298) are shown as blue sticks. The selected site, Lys 90, is shown as white stick.

D: Detail of the structure showing the location of Lys90. Its side chain is exposed at the surface of the molecule where it forms salt bridge (distance 3.1Å) with Glu34. Created with Pymol from pdb file 6Y2E³⁴. Another available structure of nsp5 from SARS-CoV-2 (pdb file 7BQY³⁴) also revealed a hydrogen bond between Lys90 with Glu 34.

E: Exchange of Lys90 by Arg90 increases the distance to Glu34 to 4.8Å.

Figure 7: Amino acid frequently exchanged in the spike protein

A: Primary structure of S showing the location of positions 614 and 859. The S1 subunit is in blue and contains the N-terminal domain (NTD) and the receptor binding domain (RBD). The S2 subunit is in green except for the heptad repeat region 1 (HR1), which is in orange and the fusion peptide F, which is in blue. The proteolytic cleavage sites S1/2 and S2' are indicated by red arrows.

B: Location of D614 in the trimeric Cryo-EM structure of S. Protein domains in one monomer are coloured as in A. The RBD of the coloured monomer is in the up conformation, the other two in the down conformation. Pdb file: 6VSB⁶¹)

C: D614 in S1 forms a hydrogen bond with Thr859 in S2 of another monomer. Protein domains are coloured as in A. Pdb file: 6VSB.

D: Uncleaved S of SARS-CoV-1 with all RBDs in the down conformation: D614 in S1 forms a hydrogen bond with Thr841 in S2 of another monomer (pdb file 5WRG⁴). Protein domains are coloured as in A.

E: Distance between D614 in the up monomer and Thr 859 in the down monomer.

F: Distance between D614 in the down monomer and Thr 859 in the up monomer.

G: Distance between D614 and Thr 859 both in the down monomer. E to G were created with PyMol from pdb file 6VSB. Note that in the other available Cryo-EM structures of S (pdb files 6VXX and 6VYB⁶²) the distance between Glu614 and Thr 859 is larger. However, since not all side chains are properly displayed in PyMol, this might be due to a lower resolution in this region.

Figure S1: Whole genome phylogenetic tree showing the distribution of the high-frequency mutation sites in each cluster. The five clusters are color coded, as shown in the inset. The semi-circular arches show the distribution of the high frequency mutation sites in the clusters, each mutation is highlighted with a different colour, as shown in the inset. The bootstrap values at each node of the tree are also shown.

Figure S2: Phylogenetic tree of S including the location of the high frequency mutation site D614G which revealed some signal for positive selection. The outer circle represents the origin of a sequence from a certain continent, and the inner circle from a certain country, as specified at the left. The triangle-shaped shadow represents the position of the mutant strains in the phylogenetic tree.

References

- 1 Wu, F. *et al.* A new coronavirus associated with human respiratory disease in China. *Nature* **579**, 265-269, doi:10.1038/s41586-020-2008-3 (2020).
- 2 Zhou, P. *et al.* A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* **579**, 270-273, doi:10.1038/s41586-020-2012-7 (2020).
- 3 Wu, A. *et al.* Genome Composition and Divergence of the Novel Coronavirus (2019-nCoV) Originating in China. *Cell Host Microbe* **27**, 325-328, doi:10.1016/j.chom.2020.02.001 (2020).
- 4 Gui, M. *et al.* Cryo-electron microscopy structures of the SARS-CoV spike glycoprotein reveal a prerequisite conformational state for receptor binding. *Cell Res* **27**, 119-129, doi:10.1038/cr.2016.152 (2017).
- 5 Baez-Santos, Y. M., St John, S. E. & Mesecar, A. D. The SARS-coronavirus papain-like protease: Structure, function and inhibition by designed antiviral

- compounds. *Antiviral Research* **115**, 21-38, doi:10.1016/j.antiviral.2014.12.015 (2015).
- 6 Lei, J., Kusov, Y. & Hilgenfeld, R. Nsp3 of coronaviruses: Structures and functions of a large multi-domain protein. *Antiviral Res* **149**, 58-74, doi:10.1016/j.antiviral.2017.11.001 (2018).
- 7 Posthuma, C. C., Te Velhuis, A. J. W. & Snijder, E. J. Nidovirus RNA polymerases: Complex enzymes handling exceptional RNA genomes. *Virus Res* **234**, 58-73, doi:10.1016/j.virusres.2017.01.023 (2017).
- 8 Snijder, E. J., Decroly, E. & Ziebuhr, J. The Nonstructural Proteins Directing Coronavirus RNA Synthesis and Processing. *Advances in virus research* **96**, 59-126, doi:10.1016/bs.aivir.2016.08.008 (2016).
- 9 Fang, P. *et al.* Porcine Deltacoronavirus Accessory Protein NS6 Antagonizes Interferon Beta Production by Interfering with the Binding of RIG-I/MDA5 to Double-Stranded RNA. *J Virol* **92**, doi:10.1128/JVI.00712-18 (2018).
- 10 Boni, M. F., Gog, J. R., Andreasen, V. & Feldman, M. W. Epidemic dynamics and antigenic evolution in a single season of influenza A. *Proc Biol Sci* **273**, 1307-1316, doi:10.1098/rspb.2006.3466 (2006).
- 11 Zhang, Y. Z. & Holmes, E. C. A Genomic Perspective on the Origin and Emergence of SARS-CoV-2. *Cell* **181**, 223-227, doi:10.1016/j.cell.2020.03.035 (2020).
- 12 Zhou, H. *et al.* A Novel Bat Coronavirus Closely Related to SARS-CoV-2 Contains Natural Insertions at the S1/S2 Cleavage Site of the Spike Protein. *Current biology : CB*, doi:10.1016/j.cub.2020.05.023 (2020).

- 13 Molecular Evolution of the SARS Coronavirus During the Course of the SARS Epidemic in China. *Science* **303**, 1666, doi:10.1126/science.1092002 (2004).
- 14 Menachery, V. D., Graham, R. L. & Baric, R. S. Jumping species-a mechanism for coronavirus persistence and survival. *Curr Opin Virol* **23**, 1-7, doi:10.1016/j.coviro.2017.01.002 (2017).
- 15 Herfst, S. *et al.* Airborne transmission of influenza A/H5N1 virus between ferrets. *Science* **336**, 1534-1541, doi:10.1126/science.1213362 (2012).
- 16 Imai, M. *et al.* Experimental adaptation of an influenza H5 HA confers respiratory droplet transmission to a reassortant H5 HA/H1N1 virus in ferrets. *Nature* **486**, 420-428, doi:10.1038/nature10831 (2012).
- 17 Chattopadhyay, I., Kiciman, E., Elliott, J. W., Shaman, J. L. & Rzhetsky, A. Conjunction of factors triggering waves of seasonal influenza. *Elife* **7**, doi:10.7554/eLife.30756 (2018).
- 18 Tang, X. *et al.* On the origin and continuing evolution of SARS-CoV-2. *National Science Review*, doi:10.1093/nsr/nwaa036 (2020).
- 19 Forster, P., Forster, L., Renfrew, C. & Forster, M. Phylogenetic network analysis of SARS-CoV-2 genomes. *Proc Natl Acad Sci U S A* **117**, 9241-9243, doi:10.1073/pnas.2004999117 (2020).
- 20 Zhang, X. *et al.* Viral and host factors related to the clinical outcome of COVID-19. *Nature*, doi:10.1038/s41586-020-2355-0 (2020).
- 21 Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* **30**, 772-780,

- doi:10.1093/molbev/mst010 (2013).
- 22 Kumar, S., Stecher, G. & Tamura, K. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Molecular Biology and Evolution* **33**, 1870-1874, doi:10.1093/molbev/msw054 (2016).
- 23 Nguyen, L. T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evo***32**, 268-274, doi:10.1093/molbev/msu300 (2015).
- 24 Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A. & Jermiin, L. S. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat Methods* **14**, 587-589, doi:10.1038/nmeth.4285 (2017).
- 25 Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res* **47**, W256-W259, doi:10.1093/nar/gkz239 (2019).
- 26 Delport, W., Poon, A. F. Y., Frost, S. D. W. & Kosakovsky Pond, S. L. Datamonkey 2010: a suite of phylogenetic analysis tools for evolutionary biology. *Bioinformatics (Oxford, England)* **26**, 2455-2457, doi:10.1093/bioinformatics/btq429 (2010).
- 27 Smith, M. D. *et al.* Less is more: an adaptive branch-site random effects model for efficient detection of episodic diversifying selection. *Molecular biology and evolution* **32**, 1342-1353, doi:10.1093/molbev/msv022 (2015).
- 28 Ben, M. *et al.* FUBAR: A Fast, Unconstrained Bayesian AppRoximation for Inferring Selection. 5 (2013).
- 29 Kosakovsky, P. S. L., Frost, S. D. W. J. M. B. & Evolution. Not So Different After All: A Comparison of Methods for Detecting Amino Acid Sites Under Selection. 5 (2005).

- 30 Murrell, B. *et al.* Detecting individual sites subject to episodic diversifying selection. (2012).
- 31 Hagemans, D., van Belzen, I. A., Moran Luengo, T. & Rudiger, S. G. A script to highlight hydrophobicity and charge on protein surfaces. *Front Mol Biosci* **2**, 56, doi:10.3389/fmolb.2015.00056 (2015).
- 32 Korber, B. *et al.* Tracking changes in SARS-CoV-2 Spike: evidence that D614G increases infectivity of the COVID-19 virus. *Cell*, doi:10.1016/j.cell.2020.06.043.
- 33 Oostra, M. *et al.* Topology and membrane anchoring of the coronavirus replication complex: not all hydrophobic domains of nsp3 and nsp6 are membrane spanning. *J Virol* **82**, 12392-12405, doi:10.1128/jvi.01219-08 (2008).
- 34 Gao, Y. *et al.* Structure of the RNA-dependent RNA polymerase from COVID-19 virus. *Science* **368**, 779-782, doi:10.1126/science.abb7498 (2020).
- 35 Hao, W. *et al.* Crystal structure of Middle East respiratory syndrome coronavirus helicase. *PLoS Pathog* **13**, e1006474, doi:10.1371/journal.ppat.1006474 (2017).
- 36 Ahmed, S. F., Quadeer, A. A. & McKay, M. R. Preliminary Identification of Potential Vaccine Targets for the COVID-19 Coronavirus (SARS-CoV-2) Based on SARS-CoV Immunological Studies. *Viruses* **12**, 254, doi:10.3390/v12030254 (2020).
- 37 Yang, Y. *et al.* Antigen-specific memory in B-1a and its relationship to natural immunity. **109**, 5388-5393 (2012).
- 38 Siu, K. L. *et al.* Severe acute respiratory syndrome coronavirus ORF3a protein activates the NLRP3 inflammasome by promoting TRAF3-dependent ubiquitination of ASC. *FASEB journal : official publication of the Federation of American Societies for*

- Experimental Biology* **33**, 8865-8877, doi:10.1096/fj.201802418R (2019).
- 39 Su, Y. C. *et al.* Discovery of a 382-nt deletion during the early evolution of SARS-CoV-2. *bioRxiv*, 2020.2003.2011.987222, doi:10.1101/2020.03.11.987222 (2020).
- 40 Muth, D. *et al.* Attenuation of replication by a 29 nucleotide deletion in SARS-coronavirus acquired during the early stages of human-to-human transmission. *Sci Rep* **8**, 15177, doi:10.1038/s41598-018-33487-8 (2018).
- 41 McBride, R., van Zyl, M. & Fielding, B. C. The coronavirus nucleocapsid is a multifunctional protein. *Viruses* **6**, 2991-3018, doi:10.3390/v6082991 (2014).
- 42 Song, W., Gui, M., Wang, X. & Xiang, Y. Cryo-EM structure of the SARS coronavirus spike glycoprotein in complex with its host cell receptor ACE2. *PLoS Pathog* **14**, e1007236, doi:10.1371/journal.ppat.1007236 (2018).
- 43 Walls, A. C. *et al.* Unexpected Receptor Functional Mimicry Elucidates Activation of Coronavirus Fusion. *Cell* **176**, 1026-1039.e1015, doi:10.1016/j.cell.2018.12.028 (2019).
- 44 Hubbard & Roderick, E. *Encyclopedia of Life Sciences // Hydrogen Bonds in Proteins: Role and Strength*. (2001).
- 45 Yuan, M. *et al.* A highly conserved cryptic epitope in the receptor binding domains of SARS-CoV-2 and SARS-CoV. *Science* **368**, 630-633, doi:10.1126/science.abb7269 (2020).
- 46 Wu, Y. *et al.* A noncompeting pair of human neutralizing antibodies block COVID-19 virus binding to its receptor ACE2. *Science*, doi:10.1126/science.abc2241 (2020).

- 47 Lan, J. *et al.* Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor. *Nature* **581**, 215-220, doi:10.1038/s41586-020-2180-5 (2020).
- 48 Rambaut, A. *et al.* A dynamic nomenclature proposal for SARS-CoV-2 to assist genomic epidemiology. *bioRxiv*, 2020.2004.2017.046086, doi:10.1101/2020.04.17.046086 (2020).
- 49 Brufsky, A. Distinct Viral Clades of SARS-CoV-2: Implications for Modeling of Viral Spread. *Journal of Medical Virology* **na**, doi:10.1002/jmv.25902 (2020).
- 50 Ortega, J. T., Serrano, M. L., Pujol, F. H. & Rangel, H. R. Role of changes in SARS-CoV-2 spike protein in the interaction with the human ACE2 receptor: An in silico analysis. *EXCLI journal* **19**, 410-417, doi:10.17179/excli2020-1167 (2020).
- 51 Dabravolski, S. A. & Kavalionak, Y. K. SARS-CoV-2: Spike glycoprotein structural diversity, phylogeny and potential animal host identification. *Journal of Medical Virology* **na**, doi:10.1002/jmv.25976 (2020).
- 52 Jin, Z. *et al.* Structure of M(pro) from SARS-CoV-2 and discovery of its inhibitors. *Nature*, doi:10.1038/s41586-020-2223-y (2020).
- 53 MacLean, O. A. *et al.* Evidence of significant natural selection in the evolution of SARS-CoV-2 in bats, not humans. *bioRxiv*, 2020.2005.2028.122366, doi:10.1101/2020.05.28.122366 (2020).
- 54 Gordon, D. E. *et al.* A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. *Nature*, doi:10.1038/s41586-020-2286-9 (2020).
- 55 Kindler, E., Thiel, V. & Weber, F. Interaction of SARS and MERS Coronaviruses with the Antiviral Interferon Response. *Adv Virus Res* **96**, 219-243,

- doi:10.1016/bs.aivir.2016.08.006 (2016).
- 56 O'Shea, T. J. *et al.* Bat flight and zoonotic viruses. *Emerg Infect Dis* **20**, 741-745, doi:10.3201/eid2005.130539 (2014).
- 57 Teilum, K., Olsen, J. G. & Kragelund, B. B. Protein stability, flexibility and function. *Biochim Biophys Acta* **1814**, 969-976, doi:10.1016/j.bbapap.2010.11.005 (2011).
- 58 Teilum, K., Olsen, J. G. & Kragelund, B. B. Functional aspects of protein flexibility. *Cell Mol Life Sci* **66**, 2231-2247, doi:10.1007/s00018-009-0014-6 (2009).
- 59 Grubaugh, N. D., Petrone, M. E. & Holmes, E. C. We shouldn't worry when a virus mutates during disease outbreaks. *Nature microbiology* **5**, 529-530, doi:10.1038/s41564-020-0690-4 (2020).
- 60 Tan, J. *et al.* The SARS-unique domain (SUD) of SARS coronavirus contains two macrodomains that bind G-quadruplexes. *PLoS Pathog* **5**, e1000428, doi:10.1371/journal.ppat.1000428 (2009).
- 61 Wrapp, D. *et al.* Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. *Science* **367**, 1260-1263, doi:10.1126/science.abb2507 (2020).
- 62 Walls, A. C. *et al.* Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein. *Cell* **181**, 281-292 e286, doi:10.1016/j.cell.2020.02.058 (2020).

Figure 1

Figure 2

Figure 3

Figure 4

Figure 5

A

B

C

Figure 6

Figure 7