The genomic variation landscape of globally-circulating clades of SARS-CoV-2 defines a genetic barcoding scheme

Qingtian Guan[1], Mukhtar Sadykov[1], Raushan Nugmanova[1], Michael J. Carr[2,3], Stefan T. Arold[4,5], Arnab Pain[1,3,6*]

[1]King Abdullah University of Science and Technology (KAUST), Pathogen Genomics Laboratory, Biological and Environmental Science and Engineering (BESE), Thuwal-Jeddah, 23955-6900, Saudi Arabia;

[2]National Virus Reference Laboratory (NVRL), School of Medicine, University College Dublin, Belfield, Dublin 4, Ireland;

[3]Research Center for Zoonosis Control, Global Institution for Collaborative Research and Education (GI-CoRE); Hokkaido University, N20 W10 Kita-ku, Sapporo, 001-0020 Japan;

[4]King Abdullah University of Science and Technology (KAUST), Computational Bioscience Research Center (CBRC), Biological and Environmental Science and Engineering (BESE), Thuwal-Jeddah, 23955-6900, Saudi Arabia;

[5]Centre de Biochimie Structurale, CNRS, INSERM, Université de Montpellier, 34090 Montpellier, France;

[6]Nuffield Division of Clinical Laboratory Sciences (NDCLS), The John Radcliffe Hospital, University of Oxford, Headington, Oxford, OX3 9DU, United Kingdom.

***Correspondence:** Arnab Pain Email: arnab.pain@kaust.edu.sa. King Abdullah University of Science and Technology, Jeddah, Saudi Arabia. Phone: (+966) 54 470 0687

ABSTRACT

We describe fifteen major mutation events from 2,058 high-quality SARS-CoV-2 genomes deposited up to March 31st, 2020. These events define five major clades (G, I, S, D and V) of globally-circulating viral populations, representing 85.7% of all sequenced cases, which we can identify using a 10 nucleotide genetic classifier or barcode. We applied this barcode to 4,000 additional genomes deposited between March 31st and April 15th and classified successfully 95.6% of the clades demonstrating the utility of this approach. An analysis of amino acid variation in SARS-CoV-2 ORFs provided evidence of substitution events in the viral proteins involved in both host-entry and genome replication. The systematic monitoring of dynamic changes in the SARS-CoV-2 genomes of circulating virus populations over time can guide therapeutic and prophylactic strategies to manage and contain the virus and, also, with available efficacious antivirals and vaccines, aid in the monitoring of circulating genetic diversity as we proceed towards elimination of the agent. The barcode will add the necessary genetic resolution to facilitate tracking and monitoring of infection clusters to distinguish imported and indigenous cases and thereby aid public health measures seeking to interrupt transmission chains without the requirement for real-time complete genomes sequencing.

INTRODUCTION

SARS-CoV-2 has now reached 185 countries across all continents, except Antarctica. With over 2.5 million cases currently confirmed globally[1], a pandemic was declared by the World Health Organisation (WHO) on March 11th, 2020. SARS-CoV-2 belongs to the

Coronaviridae family, genus Betacoronavirus, which are enveloped positive-sense, single-stranded RNA viruses, of zoonotic origin. Among human RNA viruses, coronaviruses have the largest known genome (~30 kb), which consists of the structural proteins (spike, envelope, membrane and nucleocapsid), nonstructural proteins (nsp1-16), and accessory proteins (ORF3a, ORF6, ORF7a and b, ORF8, ORF10). Structural proteins are required for host cell entry, viral assembly and exit[2,3]. Nonstructural proteins are involved in genome replication-transcription and formation of vesicles[4], whereas accessory proteins interfere with host innate defense mechanisms[5,6,7]. During replication within the host, the virus acquires genome mutations, which can be passed on to virus progeny in subsequently infected individuals. Systematically tracking mutations in SARS-CoV-2 genomes is therefore important as it allows monitoring of the molecular epidemiology of circulating viral sequences nationally and internationally. Here, we have investigated the genomic variation landscape of a large set of globally-derived SARS-CoV-2 genomes and defined major mutation events. This analysis allowed us to produce a first-generation genetic classifier, or 'barcode', defining the major clades of the virus circulating up to April 15th, 2020. Notably, this barcode allowed reliable tracking of the spatial distribution and prevalence of these viral clades over time. While most of the nonsynonymous mutation events appear neutral with respect to protein function and stability, we also found evidence of mutations in the spike protein that may modulate the interaction between SARS-CoV-2 and the host.

RESULTS

Five clades of SARS-CoV-2 are characterised by 15 major mutational events across the globe.

The SARS-CoV-2 genome is genetically most closely related (96%) to a bat SARS-related coronavirus (SARSr-CoV) RaTG13 and also to SARSr-CoVs from pangolins[8] (Supplementary Figure 1). We analysed single nucleotide polymorphisms (SNPs) of 2,058 high-quality complete genomes downloaded from GISAID[9] (March 31st). We studied their chronological occurrences during the spread of the SARS-CoV-2 across human populations, which allowed us to define several major and minor clades of the virus that share unique SNPs (Supplementary Table 1, Supplementary Figures 2, 3). We observed 1,221 SNPs in the current dataset with 753 missense, 452 silent, 12 nonsense and 4 intergenic substitutions. We defined five major clades (Figure 1) compared to the prototype (MN908947.3), covering 85.7% of the global set of SARS-CoV-2 high-quality genomes publicly available up to March 31st. The clades were named by the amino acid mutation: S (*Orf8*, L84S), V (*Orf3a*, G251V), I (*Orf1ab*, V378I), D (*Orf1ab*, G392D) and G (*S*, D614G).

These 5 clades are characterised by 15 major nucleotide substitution events in the SARS-CoV-2 genome (Figure 1, Figure 2 and Supplementary Figure 3) representing 1,763 genomes (85.7%) from 45 countries. To obtain a global picture of the regional distribution of the clades over time during a 14 week period, we plotted the relative proportions of the major and minor clades and their cumulative trend (Figure 3). We observed major differences in the apparent spread for individual clades: Clade G represents 46.2% of all the sequenced viral sequences, followed by S (25.4%), V (9.4%), I (2.6%) and D (2.1%). The remaining 14.3% were not assigned to a major clade. Clade G is widely distributed in Africa, Europe, West Asia and South America; whereas Clade S represents 63% of North

American sampled genomes, and nearly a quarter of those from Oceania. Clade I represents around one-third of genomes derived from South and West Asia, and Oceania. Southeast Asia and South Asia have the greatest number of unassigned genomes (56.9%). For these genomes that cannot be assigned to a major clade, we identified nine minor clades which were named for the amino acid mutation or nucleotide substitution (the latter shown in bold): H (*Orf1ab*, Q676H); H2 (*M*, D209H); L2 (*N*, S194L, we name it L2 to avoid confusion with the previously defined clade[8]); S2 (*N*, P344S); G11410A (*Orf1ab*, **G11410A**); Y (*S*, H49Y); C17373A (*Orf1ab*, **C17373A**); I2 (*Orf1ab*, T6136I) and K (*Orf1ab*, T2016K). The minor clades represent any monoclades with $n \geq 5$ which covers 3.2% of the total 2,058 cases. The global and regional cumulative trends were plotted over time, with the majority of the trends revealing the increasing dominance of one or two clades in each geographic region. For example, the Asian and Oceanian genomes are largely clade I whereas European genomes are predominantly clade G with clade S predominating in North America cases. This is likely attributable to founder effects during the early phases of the seeding of the local epidemics from imported cases and subsequent dissemination in the regions.

The genetic barcoding method assigns new SARS-CoV-2 cases to clades with high sensitivity.

We defined a 10 nucleotide genetic barcode of SARS-CoV-2 that identified with high sensitivity the five major clades of the circulating viral genomes available on March 31st,

based on the GISAID data (Figure 4A, Table 1). Given that SARS-CoV-2 evolves at an average rate/genome of nearly 8 x 10^{-4} nucleotide substitutions/site/year[10], and is subject to back-mutations, we further tested the sensitivity and specificity of these clade-defining SNPs, which are all above 90% (Table 1). We then applied the 10 nucleotide barcode to the 4,000 SARS-CoV-2 genomes that became available in GISAID between March 31st and April 15th, 2020 in the early stages of the pandemic. Among the 4,000 globally-circulating genomes from 66 countries, we could assign ~96% to one of the 5 major clades (Figure 4B). The remaining unassigned 4% of genomes were typified by either errors in their sequences (such as 'N's in the genome assemblies), or single cases those could not be reliably assigned to any of the major clades. The increase in the correct clade assignment for the new genomes compared to the initial validation of the methodology on 2,058 genomes (from which the barcode was established) was indicative of the rising dominance of a few major clades, and suggested that the 10 nucleotide barcode will retain its predictive power as new genomes are obtained. While this barcode represents a snapshot of the early phases of the genetic diversity of the virus during the first 16 weeks of its global spread and is expected to change over time, a barcoding strategy to monitor the progress of virus elimination after vaccines become widely available will be strategically useful to monitor decreases in viral genetic diversity. In addition, our barcode could serve as a reference for setting the baseline for global genomic diversity analysis at the beginning of the pandemic.

Mutations are not equally distributed across the SARS-CoV-2 genome.

Based on our analysis of the 2,058 available genomes from GISAID (March 31th, 2020) we observed that the genes *S*, *N* and *Orf3a*, accumulated markedly more mutations than

135 expected solely by random drift (Supplementary Figure 4) (Real/Expected ratio: *S*: 1.21;
136 *N*: 1.99; *ORF3a*: 1.82). This mutation rate may indicate adaptation to the human host
137 following recent spill over from an, as yet unknown, animal reservoir. Conversely, several
138 nonstructural proteins showed a lower-than-expected mutation rate (Real/Expected ratio:
139 nsp1: 0.22; nsp3: 0.77; nsp5: 0.70; nsp7: 0.77: nsp12: 0.88; nsp14: 0.68; nsp15: 0.76).
140 These proteins are predicted to be involved in evading host immune defenses, in enhancing
141 viral expression and in cleavage of the replicase polyprotein, based on prior studies of
142 related betacoronaviruses[11,12] . Hence, this lower mutation rate may indicate purification
143 selection to maintain these functions essential for efficient immune evasion and subsequent
144 viral dissemination. Indeed, structural proteins in coronaviruses undergo a greater degree
145 of antigenic variation which increases the fitness of the virus by means of adaptation to the
146 host and by facilitating immune escape[13].
147
148 Structural protein modeling confirmed that most of the nonsynonymous mutations in
149 the nonstructural proteins were neutral (Supplementary Figure 5-13). Conversely, several
150 nonsynonymous mutations in the spike protein might have functional consequences:
151 notably, the G clade–defining mutation D614G is located in subdomain 1 (SD1; Figure 5,
152 Supplementary Figure 5). In the trimeric S, D612 engages stabilising interactions within
153 SD1 (R646 or the backbone of F592, depending on the chain) and with the S1 of the
154 adjacent chain (T859 and K854). Replacement of D614 with a glycine would entail losing
155 these stabilising electrostatic interactions and increase the dynamics in this region. Notably,
156 V483A (found in 13 cases in Washington state, USA), V367F (in 6 cases: 5 in Paris, France;
157 1 in Hong Kong) and G476S (in 7 cases: 4 in Washington state, USA; 1 in Idaho state,

158 USA; 1 in Oregon state, USA; and 1 in Braine-l'Alleud, Belgium) are localised in the
159 receptor binding domain (RBD) of the spike protein which mediates binding to the host
160 receptor angiotensin-converting enzyme 2 (ACE2) (Figure 5)[3]. All of the viral genomes
161 harbouring V483A and G476S mutations belong to Clade S. Interestingly, the V367F
162 mutation has appeared independently in Clade V and Clade S, suggesting that this mutation
163 contributes to viral fitness. We found the V483A substitution in 13 closely-related cases
164 from Washington State. An equivalent amino acid substitution located in a similar position
165 within the RBD in the MERS-CoV spike protein reduces its binding to its cognate receptor
166 DPP4/CD26[14] (Supplementary Table 2). However, V483 is more than 10Å away from
167 ACE2 and could affect receptor binding by SARS-CoV-2 only indirectly by altering the
168 structural dynamics of the RBD loop it is a part of. The V367F mutation is located in an
169 even greater distance from ACE2 (Figure 5, Supplementary Figure 5D). The exchange of
170 the small hydrophobic residue valine with a bulky hydrophobic phenylalanine might
171 influence the efficiency of glycosylation of the nearby N343, or the positioning of the
172 sugars. The substitution G476S would lead to possible clashes with predicted interacting
173 ACE2 residues and with the RBD residue N487. However, minor reorientation of the side
174 chains might allow an additional hydrogen bond to be formed between S476 and ACE2
175 Q24 and E23, thus enhancing the affinity (Supplementary Figure 5D) to ACE2. An
176 equivalent amino acid substitution located in an analogous position within the RBD in the
177 SARS-CoV spike protein was associated with neutralisation escape from monoclonal
178 antibodies, together with other mutations observed previously[15] (Supplementary Table 2).
179 Given that V483A and V367F are solvent exposed and markedly alter the surface
180 characteristics of the RBD, they might also facilitate antibody evasion. Escape mutations

in the RBD of the SARS-CoV S protein (T332I, F460C and L443R) were identified previously[15]. These mutations negatively impact viral fitness through reducing the affinity to the host receptor[16] (Supplementary Table 2). The nonsynonymous mutations in the N protein, which have key roles in viral assembly, might also have functional implications. The hotspot mutations S202N, R203K and G204R all cluster in a linker region where they might potentially enhance RNA binding and alter the response to serine phosphorylation events (Supplementary Figure 6). The clade I–defining mutation in the nucleocapsid protein, which has key roles in viral assembly, is synonymous. However, we observed nonsynonymous mutations in the nucleocapsid protein that are predicted to have functional implications. The hotspot mutations S202N, R203K and G204R all cluster in a linker region where they might potentially enhance RNA binding and alter the response to serine phosphorylation events (Supplementary Figure 6).

DISCUSSION

In this study, we have defined 5 major clades (G, I, S, D and V) and 9 minor clades (H2, L2, G1110A, H, Y, C17373A, S2, I2 and K) which covers ~89% of the 2,058 high quality genomes available until March 31st in GISAID database. The clustering of these genomes revealed the spread of clades to diverse geographical regions (Figure 1, Figure 3). This pattern contrasts with those observed for other epidemic coronaviruses, such as MERS-CoV, which display distinct geographical clustering[17]. For example, clade G, which was first detected on January 28th, has reached 40 countries and 130 cities, within a span of 10 weeks (Supplementary Table 1). This pattern demonstrates efficient viral transmission through frequent intercontinental travel during the period when international

204 travel restrictions were only present sporadically, which has enabled the virus to spread to
205 multiple distant locations within a short period of time. This observation reinforces the
206 importance of curtailing international travel and imposing restrictions early in pandemics
207 and imposing social distancing in order to contain the global spread of viruses.
208
209 We have observed a distinct distribution of the major clades in different parts of the
210 world (Figure 3). Most of the viral genomes that have not been assigned to a major clade
211 are found in Asia and have earlier detection times in January and February at the start of
212 the epidemic in China (Supplementary Figure 2). We observed a decrease in the genetic
213 diversity of the virus over time following dissemination from China, especially in Europe
214 and North America that each notably now has a predominant clade type, which we believe
215 to be associated with a founder effect whereby a single clade was introduced and
216 subsequently disseminated (Figure 3). An important caveat of the present study is that the
217 current sampling of available public genomes does likely not represent the extant genetic
218 diversity of virus populations in circulation due to biases of genome data deposits from the
219 sequencing laboratories based mainly in the northern hemisphere and new datasets may
220 define new clades in the near future from regions, including Africa, the Indian subcontinent
221 and Latin America with comparably few genomes available at present. In this case,
222 additional identifiers within an evolving barcode scheme can be added to track and monitor
223 future emerging clades with higher resolution. On the other hand, the genetic stability of
224 SARS-CoV-2[18] may result in the continuing circulation of a limited number of clades until
225 such time as mitigation measures including the isolation of vulnerable populations and the
226 availability of efficacious antivirals and vaccines reduces the genetic diversity in

227 circulation. This molecular genotyping approach has been demonstrated for other viruses
228 (e.g. measles, poliovirus, rotavirus and human papillomaviruses) with herd-immunity
229 vaccination programmes working to eliminate pathogens from endemic circulation in
230 humans[19,20,21,22] . The availability of a barcoding scheme that rapidly generates a SNP
231 allowing clade assignment will be critical in this elimination phase when widespread
232 availability of vaccines permits eradication of SARS-CoV-2 from endemicity in humans.
233
234 Our work provides a baseline global genomic epidemiology of SARS-CoV-2 prior to
235 introduction of therapeutic and prophylactic approaches. The mutational landscape of
236 global populations of over SARS-CoV-2 6,000 genomes provides an evidence-based
237 framework for tracking the clades that comprise the pandemic on different continents.
238 However, due to the bias in the representation of countries depositing the SARS-CoV-2
239 genomes with over-representation of North American and European genomes (28.3% and
240 47.2% respectively) and the available genome data representing only a minute proportion
241 of the total COVID-19 positive cases from each of these regions (America, 0.27%; Europe,
242 0.31%; China, 0.31%. Data collected from GISAID[9] and COVID-19 dashboard[1]), the
243 genetic barcode described hererin may need to be updated in order to be globally
244 representative, once sufficient numbers of genomes covering less represented parts of the
245 world are eventually sequenced and deposited to publicly-available database. We envisage
246 a qPCR-based allelic discrimination approach, such as PCR allele competitive extension
247 (PACE), which would enable rapid turnaround in real-time following the identification of
248 a laboratory-confirmed case. This would allow viral genetic epidemiological data to be
249 added to contact tracing information to allow efficient detection of circulating SARS-CoV-

2 clades that will allow discrimination of autochthonous and imported clades to aid progressive elimination of the genetic diversity and, ultimately, eradication in all regions. A robust genetic barcoding scheme for SARS-CoV-2 can facilitate this molecular tracking of larger numbers of laboratory-confirmed cases and by implementing such a facile genotyping approach upstream of next-generation sequencing will allow whole genome sequencing to be performed on selected cases. This is of particular relevance when the available genomes represent only a small sample of the over 2.5 million total COVID-19 cases globally to date.

METHODS

Phylogenomic Analysis

1,427 coronavirus genomes were downloaded from Virus Pathogen Database and Analysis Resource (ViPR)[23] on February 14th 2020, including 329 SARS, 35 SARS-CoV-2, 61 NL63, 521 MERS, 52 HKU1, 170 OC43, 97 bovine coronaviruses and 61 mouse hepatitis viruses. Sequence alignment was performed using MAFFT(version 7.407)[24] software and then trimmed by trimAL(version 1.4.1)[25]. Phylogenetic analyses of the complete genomes were done with FastTree(version 2.1.10)[26] software with default parameters, and iTOL(version 5)[27] was used for phylogenetic tree visualisation.

2,127 complete SARS-CoV2 genomes were downloaded from the GISAID9 (March 31th 2020). 69 of those genomes were removed due to poor assembly quality resulting in 2,058 complete genomes that were subsequently used for analysis. Sequence alignment was performed with MAFFT(version 7.407)[24] and trimmed by trimAL(version 1.4.1)[25].

Phylogenetic analyses of the complete genomes were performed with RAxML[28] (version 8.2.12) with 1,000 bootstrap replicates, employing the general time-reversible nucleotide substitution model. iTOL(version 5)[27] was used for the phylogenetic tree visualisation. SNPs from each of the genomes were called by Parsnp (version 1.2) from the Harvest suite[29] using MN908947.3 as the reference genome, and the SNPs were further annotated by SnpEff(version 4.3m)[30]. The monoclades associated with SNPs with a frequency ≥40 were defined as major clades and monoclades associated with SNPs with a frequency ≥5 were defined as minor clades.

Protein Structural Analysis

Experimentally determined protein structures were obtained from the Protein Data Bank (PDB). SwissModel[31], I-Tasser[32], RaptorX[33] and an in-house modelling pipeline was used to produce protein structure homology models. Phobius[34] was used for prediction of trans-membrane regions. RaptorX was also used for predicting secondary structure, protein disorder and solvent exposure of amino acids. Pymol(version 1.8.6.2) was used for visualization.

ACKNOWLEDGEMENTS

This work was supported by funding from King Abdullah University of Science and Technology (KAUST), Office of Sponsored Research (OSR), under award number FCC/1/1976-25-01. Work in AP's laboratory is supported by the KAUST faculty baseline fund (BAS/1/1020-01- 01) and research grants from the Office for Sponsored Research (OSR-2015-CRG4-2610, OCRF-2014-CRG3-2267). We thank all laboratories which have

296 contributed sequences to the GISAID database. We thank Olga Douvropoulou, Raeece
297 Naeem Mohamed Ghazzali and Sharif Hala for their support during the work. We also
298 thank Richard Culleton (Nagasaki University, Japan) and Gabo Gonzalez (UCD, Ireland)
299 for their critical comments on the manuscript draft.

300

301 AUTHOR CONTRIBUTIONS

302 AP conceived the study and supervised the work; AP and QG designed the analysis. QG,
303 MS and SA performed the data analysis and prepared the initial draft of the manuscript,
304 followed by edits from AP, MC and RN. All authors have commented on various sections
305 of the manuscript.

306

307 COMPETING INTERESTS

308 The authors have no conflicts of interest to declare.

309 REFERENCES:

310 1. Dong, E., Du, H. & Gardner, L. An interactive web-based dashboard to track
311 COVID-19 in real time. *Lancet. Infect. Dis.* **3099**, 19–20 (2020).

312 2. Bárcena, M. *et al.* Cryo-electron tomography of mouse hepatitis virus: Insights
313 into the structure of the coronavirion. *Proc. Natl. Acad. Sci. U. S. A.* (2009).
314 doi:10.1073/pnas.0805270106

315 3. Hoffmann, M. *et al.* SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2
316 and Is Blocked by a Clinically Proven Protease Inhibitor Article SARS-CoV-2
317 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically
318 Proven Protease Inhibitor. *Cell* **181**, 1–10 (2020).
319 4. Hagemeijer, M. C. *et al.* Membrane rearrangements mediated by coronavirus
320 nonstructural proteins 3 and 4. *Virology* (2014). doi:10.1016/j.virol.2014.04.027
321 5. Wu, A. *et al.* Genome Composition and Divergence of the Novel Coronavirus
322 (2019-nCoV) Originating in China. *Cell Host Microbe* (2020).
323 doi:10.1016/j.chom.2020.02.001
324 6. Wu, F. *et al.* A new coronavirus associated with human respiratory disease in
325 China. *Nature* (2020). doi:10.1038/s41586-020-2008-3
326 7. Liu, D. X., Fung, T. S., Chong, K. K. L., Shukla, A. & Hilgenfeld, R. Accessory
327 proteins of SARS-CoV and other coronaviruses. *Antiviral Research* (2014).
328 doi:10.1016/j.antiviral.2014.06.013

8. Tang, X. *et al.* On the origin and continuing evolution of SARS-CoV-2. *Natl. Sci. Rev.* **6**, (2020).

9. Shu, Y. & McCauley, J. GISAID: Global initiative on sharing all influenza data – from vision to reality. *Eurosurveillance* (2017). doi:10.2807/1560-7917.ES.2017.22.13.30494

10. Hadfield, J. *et al.* NextStrain: Real-time tracking of pathogen evolution. *Bioinformatics* (2018). doi:10.1093/bioinformatics/bty407

11. Báez-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* (2015). doi:10.1016/j.antiviral.2014.12.015

12. Posthuma, C. C., te Velthuis, A. J. W. & Snijder, E. J. Nidovirus RNA polymerases: Complex enzymes handling exceptional RNA genomes. *Virus Research* (2017). doi:10.1016/j.virusres.2017.01.023

13. Walls, A. C. *et al.* Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein. *Cell* (2020). doi:10.1016/j.cell.2020.02.058

14. Kleine-Weber, H. *et al.* Mutations in the Spike Protein of Middle East Respiratory Syndrome Coronavirus Transmitted in Korea Increase Resistance to Antibody-Mediated Neutralization. *J. Virol.* (2018). doi:10.1128/jvi.01381-18

15. Rockx, B. *et al.* Escape from Human Monoclonal Antibody Neutralization Affects In Vitro and In Vivo Fitness of Severe Acute Respiratory Syndrome Coronavirus. *J. Infect. Dis.* (2010). doi:10.1086/651022

16. Tang, X. C. *et al.* Identification of human neutralizing antibodies against MERS-CoV and their role in virus adaptive evolution. *Proc. Natl. Acad. Sci. U. S. A.* (2014). doi:10.1073/pnas.1402074111

17. Kim, J. Il *et al.* The recent ancestry of Middle East respiratory syndrome coronavirus in Korea has been shaped by recombination. *Sci. Rep.* (2016). doi:10.1038/srep18825

18. Jia, Y. *et al.* Analysis of the mutation dynamics of SARS-CoV-2 reveals the spread history and emergence of RBD mutant with lower ACE2 binding affinity. *bioRxiv* (2020).

19. Brown, K. E. *et al.* Genetic characterization of measles and rubella viruses detected through global measles and rubella elimination surveillance, 2016-2018. *Morb. Mortal. Wkly. Rep.* (2019). doi:10.15585/mmwr.mm6826a3

20. Mankertz, A. *et al.* Spread of measles virus D4-Hamburg, Europe, 2008-2011. *Emerg. Infect. Dis.* (2011). doi:10.3201/eid1708.101994

21. Grassly, N. C. The final stages of the global eradication of poliomyelitis. *Philosophical Transactions of the Royal Society B: Biological Sciences* (2013). doi:10.1098/rstb.2012.0140

22. Soares-Weiser, K. *et al.* Vaccines for preventing rotavirus diarrhoea: vaccines in use. in *Cochrane Database of Systematic Reviews* (2012). doi:10.1002/14651858.cd008521.pub3

23. Pickett, B. E. *et al.* ViPR: An open bioinformatics database and analysis resource for virology research. *Nucleic Acids Res.* (2012). doi:10.1093/nar/gkr859

372 24. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software
373 version 7: Improvements in performance and usability. *Mol. Biol. Evol.* (2013).
374 doi:10.1093/molbev/mst010
375 25. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: A tool for
376 automated alignment trimming in large-scale phylogenetic analyses.
377 *Bioinformatics* **25**, 1972–1973 (2009).
378 26. Price, M. N., Dehal, P. S. & Arkin, A. P. Fasttree: Computing large minimum
379 evolution trees with profiles instead of a distance matrix. *Mol. Biol. Evol.* (2009).
380 doi:10.1093/molbev/msp077
381 27. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL): An online tool for
382 phylogenetic tree display and annotation. *Bioinformatics* (2007).
383 doi:10.1093/bioinformatics/btl529
384 28. Stamatakis, A. RAxML version 8: A tool for phylogenetic analysis and post-
385 analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).

386 29. Treangen, T. J., Ondov, B. D., Koren, S. & Phillippy, A. M. The harvest suite for
387 rapid core-genome alignment and visualization of thousands of intraspecific
388 microbial genomes. *Genome Biol.* **15**, (2014).
389 30. Cingolani, P. *et al.* A program for annotating and predicting the effects of single
390 nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila
391 melanogaster strain w1118; iso-2; iso-3. *Fly (Austin)*. (2012).
392 doi:10.4161/fly.19695
393 31. Arnold, K., Bordoli, L., Kopp, J. & Schwede, T. The SWISS-MODEL workspace:
394 A web-based environment for protein structure homology modelling.
395 *Bioinformatics* (2006). doi:10.1093/bioinformatics/bti770
396 32. Yang, J. *et al.* The I-TASSER suite: Protein structure and function prediction.
397 *Nature Methods* (2014). doi:10.1038/nmeth.3213
398 33. Källberg, M., Margaryan, G., Wang, S., Ma, J. & Xu, J. Raptorx server: A
399 resource for template-based protein structure modeling. *Methods Mol. Biol.*
400 (2014). doi:10.1007/978-1-4939-0366-5_2

401 34. Käll, L., Krogh, A. & Sonnhammer, E. L. L. Advantages of combined

402 transmembrane topology and signal peptide prediction-the Phobius web server.

403 *Nucleic Acids Res.* **35**, (2007).

404 35. Senior, A. W. *et al.* Improved protein structure prediction using potentials from

405 deep learning. *Nature* (2020). doi:10.1038/s41586-019-1923-7

406 36. John Jumper, Tunyasuvunakool, K., Kohli, P. & Hassabis, D. Computational

407 predictions of protein structures associated with COVID-19. (2020). Available at:

408 https://deepmind.com/research/open-source/computational-predictions-of-protein-

409 structures-associated-with-COVID-19.

410

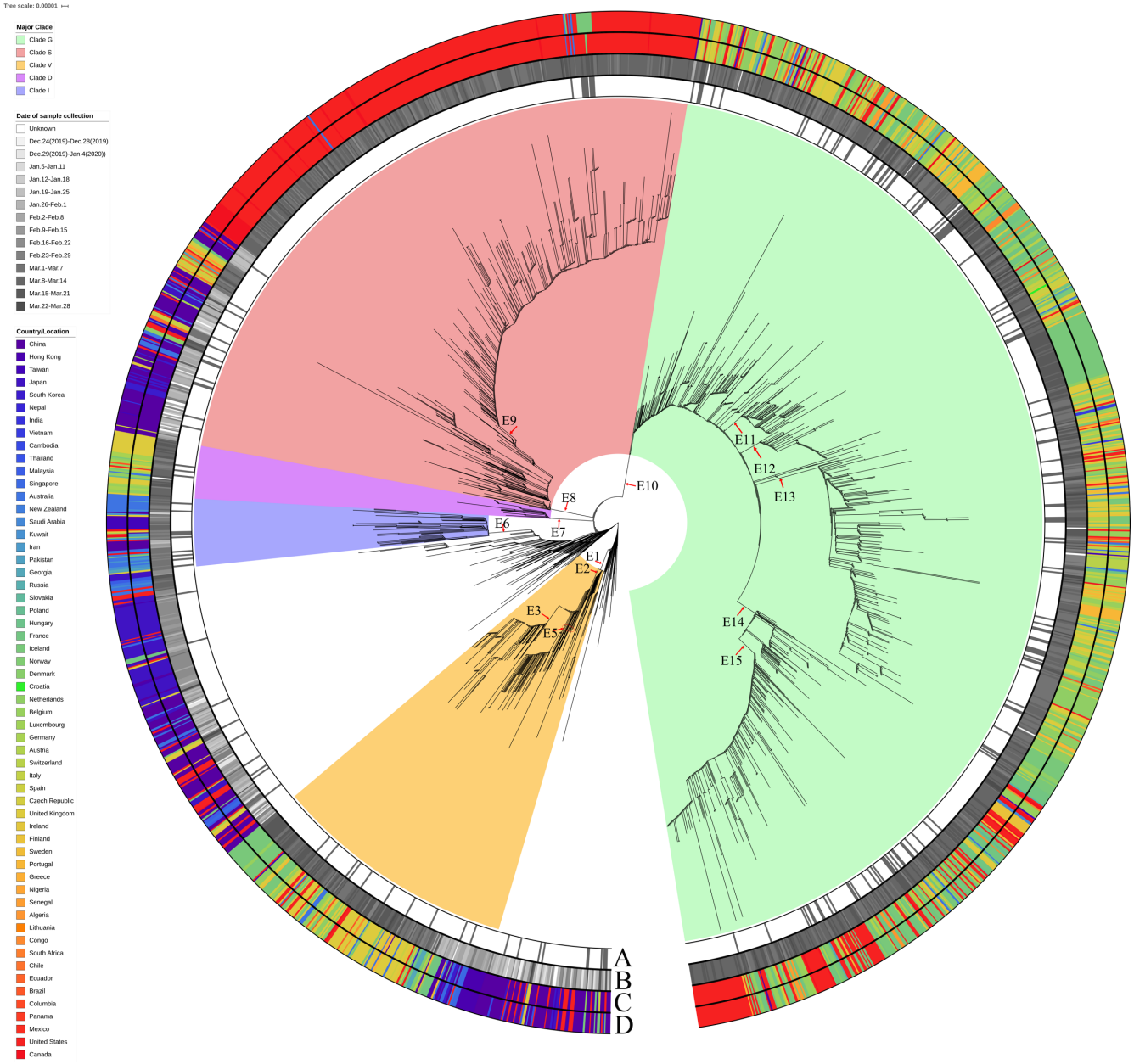

Tree scale: 0.00001
Major Clade
Clade G
Clade S
Clade V
Clade D
Clade I
Date of sample collection
Unknown
Dec.24(2019)-Dec.28(2019)
Dec.29(2019)-Jan.4(2020))
Jan.5-Jan.11
Jan.12-Jan.18
Jan.19-Jan.25
Jan.26-Feb.1
Feb.2-Feb.8
Feb.9-Feb.15
Feb.16-Feb.22
Feb.23-Feb.29
Mar.1-Mar.7
Mar.8-Mar.14
Mar.15-Mar.21
Mar.22-Mar.28
Country/Location
China
Hong Kong
Taiwan
Japan
South Korea
Nepal
India
Vietnam
Cambodia
Thailand
Malaysia
Singapore
Australia
New Zealand
Saudi Arabia
Kuwait
Iran
Pakistan
Georgia
Russia
Slovakia
Poland
Hungary
France
Iceland
Norway
Denmark
Croatia
Netherlands
Belgium
Luxembourg
Germany
Austria
Switzerland
Italy
Spain
Czech Republic
United Kingdom
Ireland
Finland
Sweden
Portugal
Greece
Nigeria
Senegal
Algeria
Lithuania
Congo
South Africa
Chile
Ecuador
Brazil
Columbia
Panama
Mexico
United States
Canada
E9
E11
E12
E13
E10
E8
E6
E7
E1
E2
E3
E5
E14
E15
A
B
C
D

Figure 1. Major substitutions events in SARS-CoV-2. A global SNP-based radial phylogeny of SARS-CoV-2 genomes defining five major clades (S, G, V, D and I) and several subclades based on nucleotide substitution events. E1-E15 represents the 15 major evolutionary events. Branches with >70% bootstrap values are shown and all the main clade-defining branches have bootstrap values >90%. E4, which corresponds to a G11083A mutation in ORF1a was found in two independent branches of the tree and for illustration purposes it is shown in Supplementary Figure 2. The labeling for the concentric outer circles I as follows: A, the imported cases which the country of exposure differs from the country of isolation; B, the collection date of each case with one week resolution; C, collection locations of the cases; D exposure locations of the cases.

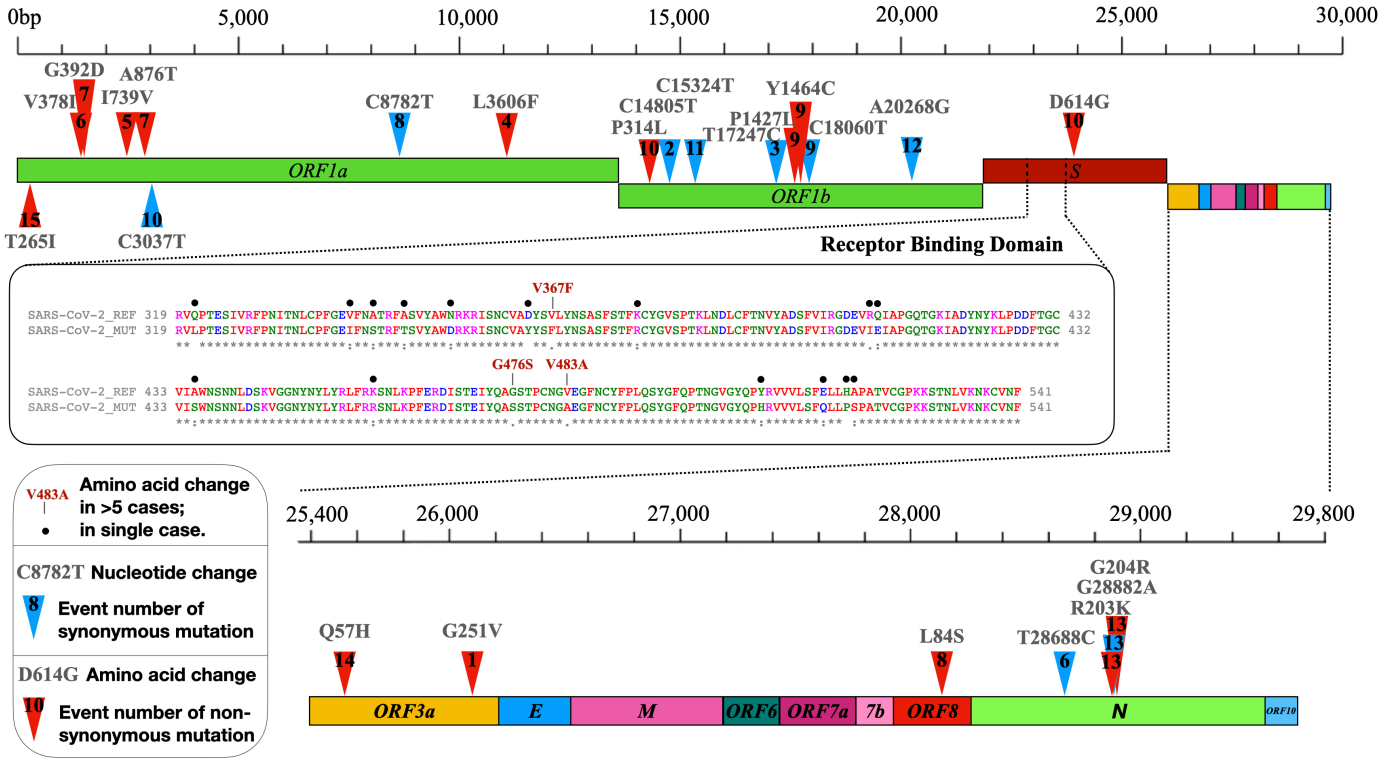


Figure 2. Major mutations and associated variation in globally circulating SARS-CoV-2 genomes (n=2,058). Genomic localisation of major mutation events as defined within our study. SARS-CoV-2 mutations in the receptor-bing domain (RBD) sequence contain all amino acid substitutions from 2,058 genomes available up until March 31st 2020.

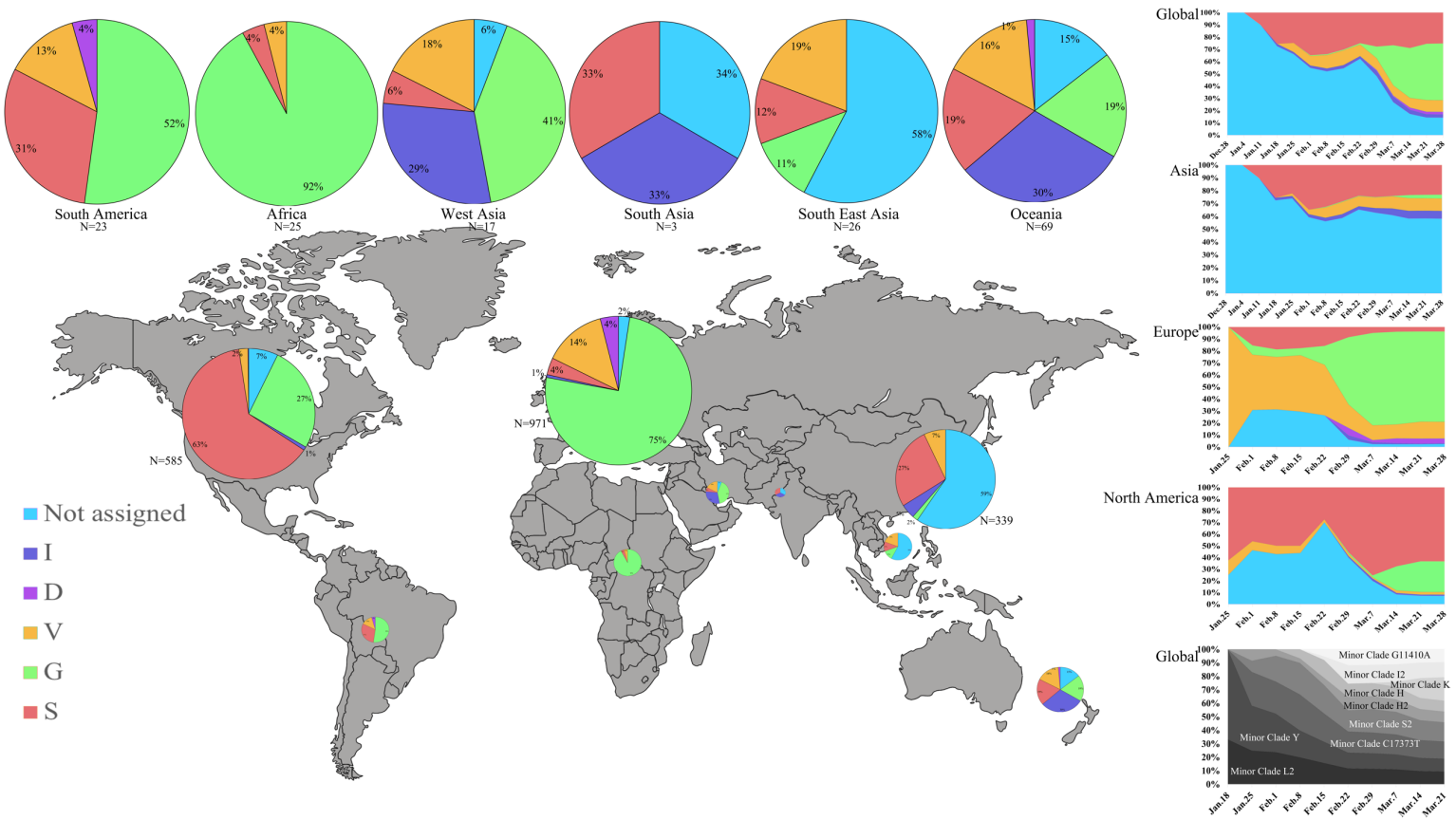


Figure 3. Global distribution of various major and minor clades of SARS-CoV2 genomes and their relative prevalence over a 14 week time period from December 24th 2019 to March 28th 2020 from the outbreak and early stages of the pandemic. The size of each pie chart is proportional to the numbers within each respective clade (Europe, max=971; South Asia, min=3). The cumulative

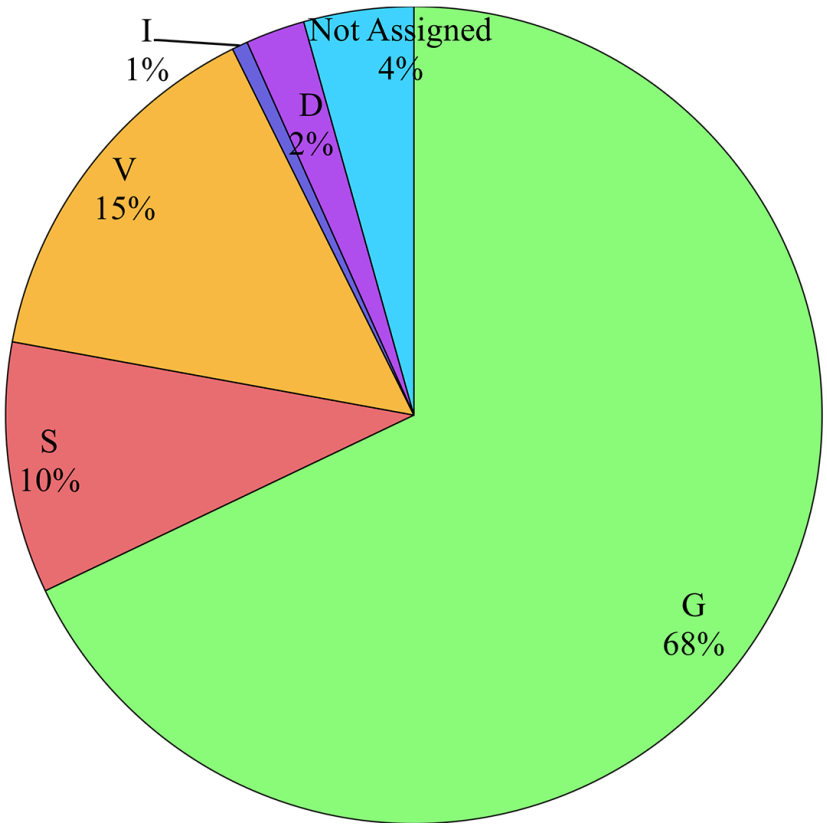
trend of the clades is shown on the right and the span of time indicates the first and last observed case in each particular clade, until March 28th 2020.

429

A

Position	8782(ORF1a)	28144(ORF8)	26144(ORF3a)	3037(ORF1a)	14408(ORF1a)	23403(S)	1397(ORF1a)	28688(N)	1440(ORF1a)	2891(ORF1a)
Clade S	T	C	G	C	C	A	G	T	G	G
Clade V	C	T	T	C	C	A	G	T	G	G
Clade G	C	T	G	T	T	G	G	T	G	G
Clade I	C	T	G	C	C	A	A	C	G	G
Clade D	C	T	G	C	C	A	G	T	A	A

B



430

431 **Figure 4. Major mutations and associated variation in globally circulating SARS-**

432 **CoV-2 genomes (n=6,058) available until April 15th 2020.** (A) A 10 nucleotide SNP

433 genetic barcode that defines the 5 clades of SARS-CoV-2. (B) By applying this barcode
434 we are able to classify 95.6% of the cases published between March 31st and April 15th
435 2020 during the early phase of the pandemic. Pie charts showing the percentage of each
436 major clade by applying a 10 nucleotide genetic classifier to 4,000 new genomes
437 downloaded from GISAID.

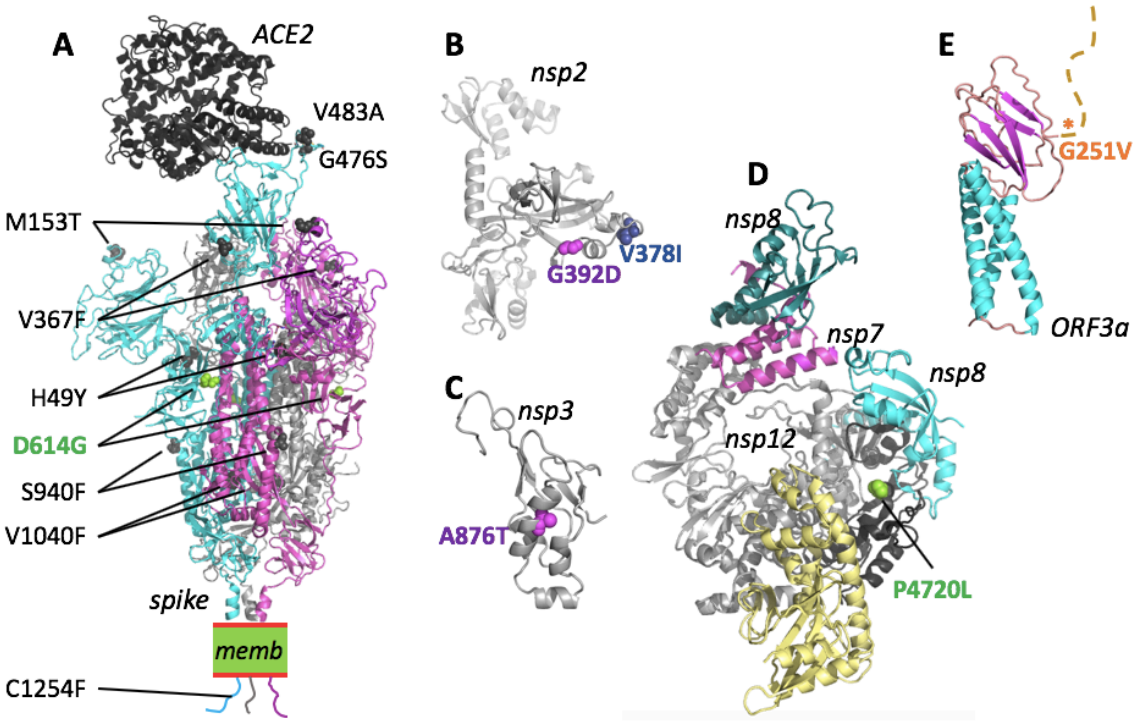


Figure 5: Mapping of SARS-CoV-2 clade-defining mutations onto the proteins. Nonsynonymous mutations for proteins where the 3D structure was experimentally determined (spike, nsp12/7/8) or can be inferred with reasonable confidence. Mutations are colour-coloured as for the corresponding clades in Figure 3 (D: magenta; G: light green; I: blue; V: orange). For a detailed analysis, see Supplementary Figures 5-13. (A) The structure of the SARS-CoV-2 spike trimer in its open conformation (chains are cyan, magenta and grey) bound to the human receptor ACE2 (black) modeled based on PDB accessions 6m17 and 6vyb. Identified nonsynonymous mutations are shown as spheres in the model. For reasons of visibility only mutations of two of the three spike chains are labeled. memb. indicates the plasma membrane. (B) Fragment comprising residues 180-534 of nsp2, modelled by AlphaFold[35]. Both clade-defining mutations are located in solvent-exposed regions and would not lead to steric clashes. (C) The substitution A876T (corresponding to residue A58 in the nsp3 cleavage product numbering) is situated in the

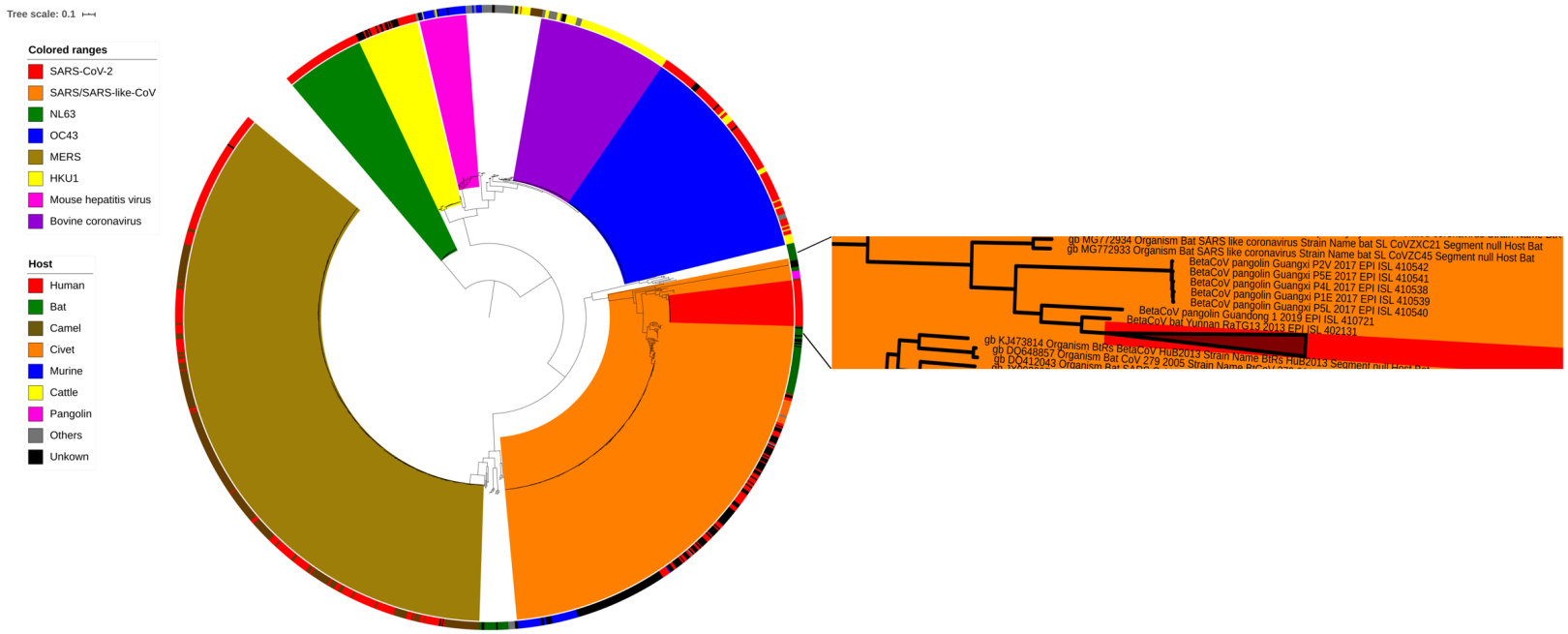
N-terminal ubiquitin-like domain of nsp3. The structure of this domain can be inferred based on the 79% identical structure of residues 1-112 from SARS-CoV (PDB id 2idy). The substitution A876T can be accommodated with only minor structural adjustments and is not expected to have a substantial influence on the proteins stability or function. (D) The structure shows the nsp12 in complex with nsp7 (magenta) and nsp8 (cyan and teal), based on PDB 7btf. P4720 (P323 in nsp12 numbering) is located in the 'interface domain' (black). In this position, the P323L substitution is not predicted to disrupt the folding or protein interactions and hence is not expected to have strong effects. (E) A theoretical model for the Orf3a monomer has been proposed by AlphaFold[36]. The structure-function relationship of this protein remains to be clarified. The mutation G251V is located C-terminal to the β-sandwich domain and the tail (marked by an asterisk).

Table 1. Sensitivity and specificity of the SARS-CoV-2 clade-defining SNP based on 2,058 genomes covering the first 14 weeks of the COVID-19 outbreak.

Clade Defined	Event	ORF	Nucleotide Substitution	Amino Acid Substitution	True Positive	False Positive	False Negative	True Negative	Specificity	Sensitivity
V	E1	*Orf3a*	G26144T	Gly251Val	195	0	1	1862	1.000	0.995
I	E6	*Orf1ab*	G1397A	Val378Ile	54	0	0	2004	1.000	1.000
		N	T28688C		53	1	2	2002	1.000	0.964
D	E7	*Orf1ab*	G1440A	Gly392Asp	42	0	0	2016	1.000	1.000
		Orf1ab	G2891A	Ala876Thr	39	0	3	2016	1.000	0.929
S	E8	*Orf1ab*	C8782T		521	4	2	1531	0.997	0.996
		Orf8	T28144C	Leu84Ser	523	1	0	1534	0.999	1.000
G	E10	*Orf1ab*	C3037T		946	1	4	1107	0.999	0.996
		Orf1ab	C14408T	Pro4720Leu	948	1	2	1107	0.999	0.998
		S	A23403G	Asp614Gly	946	1	4	1107	0.999	0.996

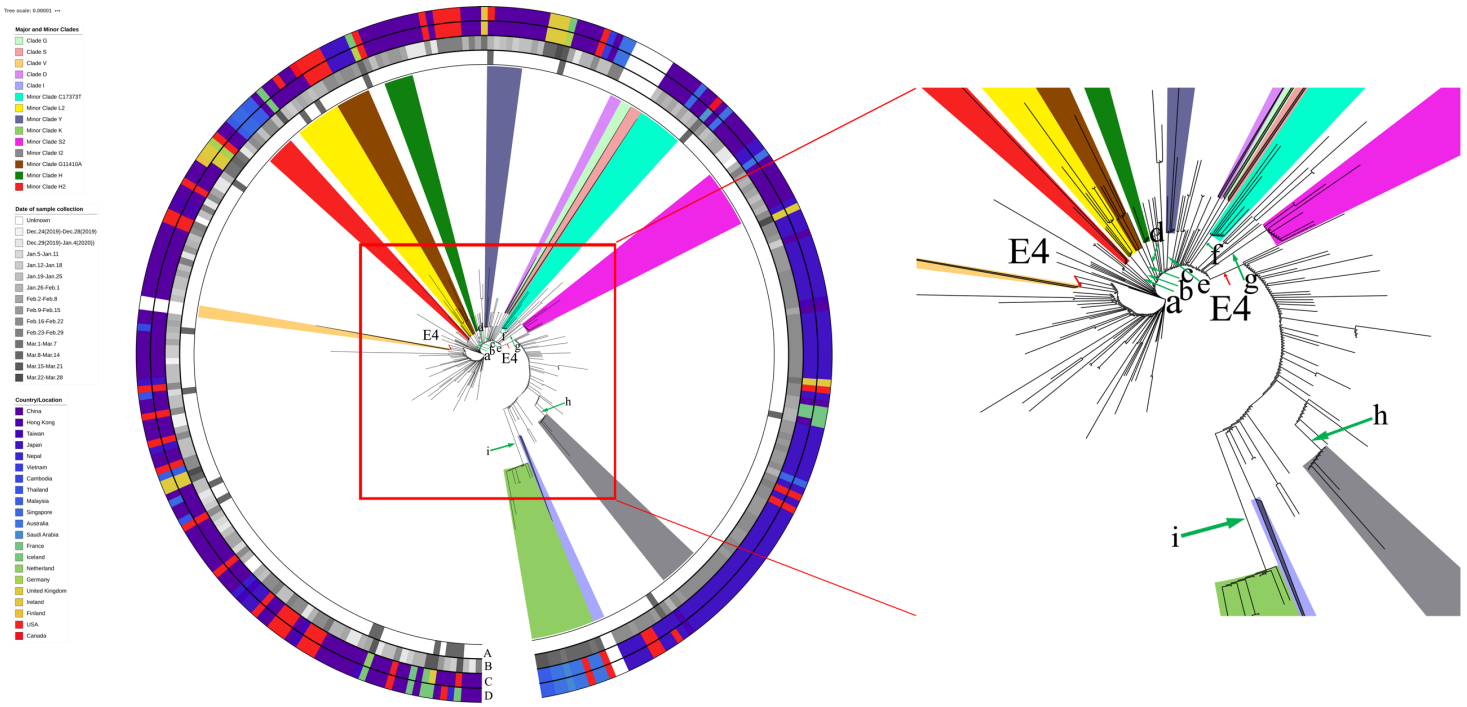
The genomic variation landscape of globally-circulating clades of SARS-CoV-2 defines a genetic barcoding scheme

SUPPLEMENTARY FIGURES



Supplementary Figure 1. A maximum likelihood phylogenomic tree based on whole genomes from SARS-CoV-2 and other coronaviruses. All available complete genomes of an alphacoronavirus (NL63) and several major groups of betacoronavirus species were used to construct the phylogenetic tree. The colour-coded outer circle represents the host of each specific genome

476 sequence. The SARS-CoV-2 part of the phylogenetic tree has been expanded for better resolution of the tree with closely related
477 viral species. A total of 1,427 SARS-CoV-2 whole genomes were illustrated in the phylogeny.

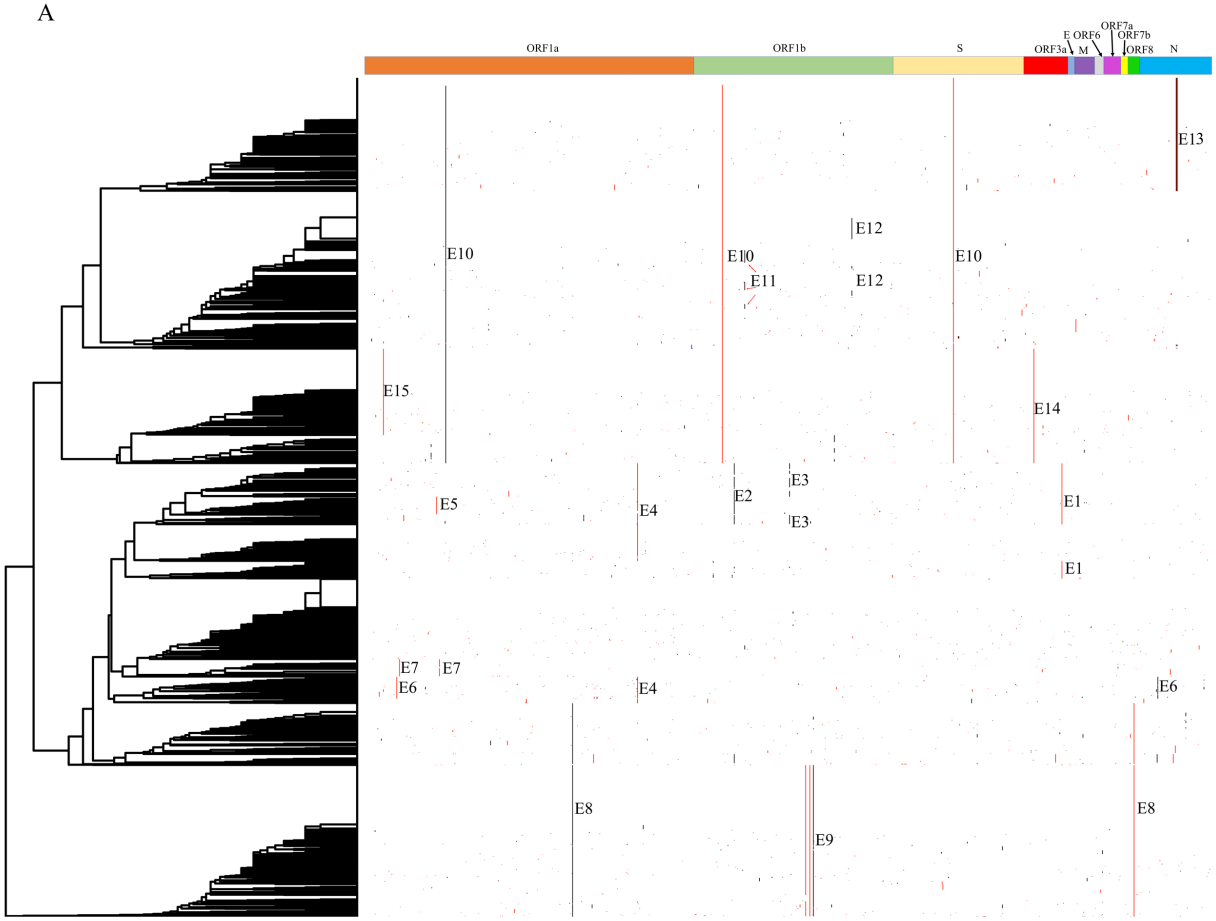


Event	Minor Clades	Nucleotide substitution	ORF	Amino Acid Substitution	First Date detected	Location first detected	Reference: EPI_ISL
a	H2	G27147C	M	p.Asp209His	2020/1/25	Singapore	407987
b	L2	C28854T	N	p.Ser194Leu	2020/1/16	China, Shenzhen	406594
c	G11410A	G11410A	ORF1a		2020/2/15	Japan, Diamond Princess cruise ship	416584
d	H	G2293T	ORF1a	p.Gln676His	2020/2/4	China,Shanghai	416353
e	Y	C21707T	S	p.His49Tyr	2020/1/17	China,Zhuhai	403936
f	C17373T	C17373T	ORF1b		2020/1/22	China,Foshan	406536
g	S2	C15324T	ORF1b		2020/1/22	China,Guangzhou	406533
		C29303T	N	p.Pro344Ser			
h	I2	C18656T	ORF1b	p.Thr6136Ile	2020/2/15	Japan, Diamond Princess cruise ship	416565
i	K	C6312A	ORF1a	p.Thr2016Lys	2020/3/5	Australia, Sydney	417388
		C13730T	ORF1b	p.Ala4494Val			
		C23929T	S				
		C28311T	N	p.Pro13Leu			

478

Supplementary Figure 2: Minor clades SARS-CoV-2 . For illustration purposes, the major clades have been collapsed and the central part of the figure has been enlarged. Most of the genomes that cannot be assigned to any major clade are derived from Asian countries indicative of high genetic diversity in the relatively early stages of the local epidemics and that a founder effect likely explains the observations of single predominant clades in North America and Europe which exhibit a reduction in genetic diversity. These genomes obtained were further analysed and assigned to minor clades. A minor clade was defined with n≥5 cases of SARS-CoV-2 based on phylogenetic and SNP analysis and the events associated with them are indicated by the green arrow. The information contained in the concentric circles: A, the imported cases; B, the collection date of each genome at one-week resolution; C, collection locations of the genomes; D exposure locations of the genomes.

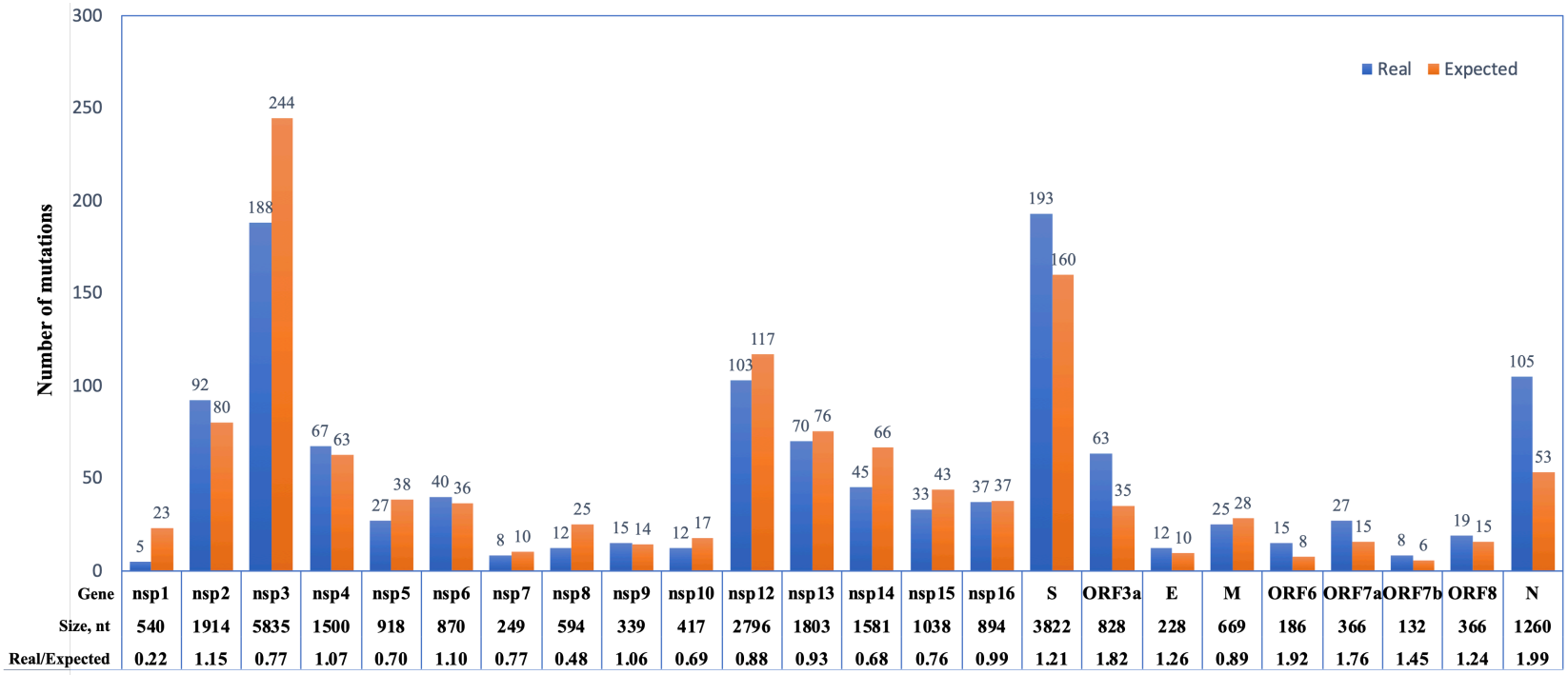
A



B

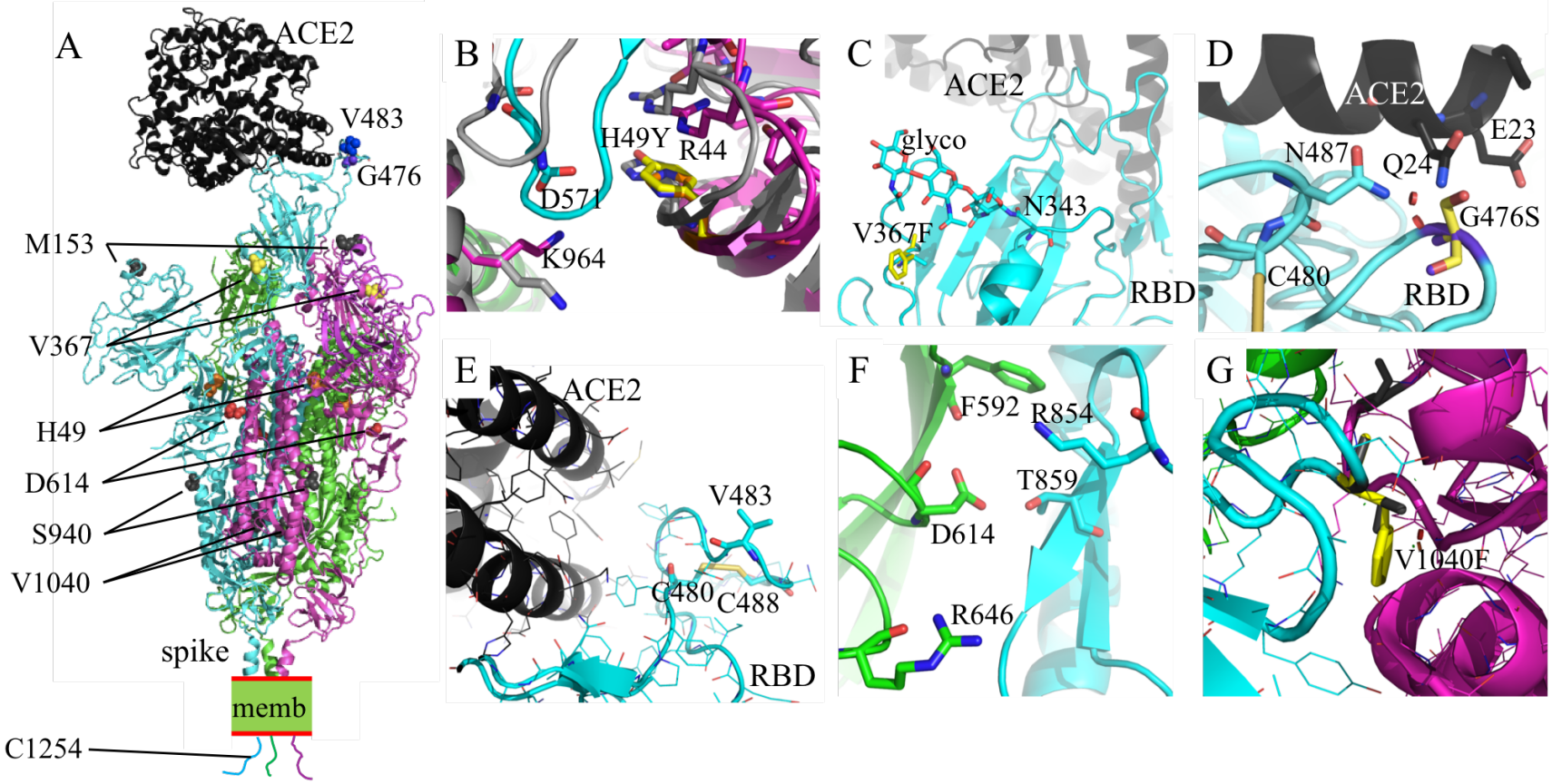
Event	Date first detected	Location first detected	Reference: EPI_ISL
E1	22-Jan-20	USA, CA	406036
E2	25-Jan-20	China, Hangzhou	415709
E3	26-Feb-20	UK, England	414011
E4	17-Jan-20	China, Yunnan	408480
E5	28-Feb-20	UK, England	414006
E6	18-Jan-20	China, Wuhan	412981
E7	25-Feb-20	Germany, North Rhine-Westphalia	414497
E8	5-Jan-20	China, Wuhan	406801
E9	20-Feb-20	USA, WA	413456
E10	28-Jan-20	Germany, Munich	406862
E11	22-Jan-20	France, Auvergne-Rhône-Alpes	417333
E12	27-Feb-20	Switzerland, Geneva	414019
E13	25-Feb-20	Germany, Baden Wuerttemberg	412912
E14	21-Feb-20	France, Hauts de France	418218
E15	21-Feb-20	France, Hauts de France	418218

Supplementary Figure 3. The genomic position and details of SNPs in 2,058 SARS-CoV-2 genomes and the amino acid substitution eventss. (A) Nonsynonymous mutations are marked in red, and synonymous mutations are labeled in black. (B) The first appearance and the location of each event



Gene	nsp1	nsp2	nsp3	nsp4	nsp5	nsp6	nsp7	nsp8	nsp9	nsp10	nsp12	nsp13	nsp14	nsp15	nsp16	S	ORF3a	E	M	ORF6	ORF7a	ORF7b	ORF8	N
Size, nt	540	1914	5835	1500	918	870	249	594	339	417	2796	1803	1581	1038	894	3822	828	228	669	186	366	132	366	1260
Real/Expected	0.22	1.15	0.77	1.07	0.70	1.10	0.77	0.48	1.06	0.69	0.88	0.93	0.68	0.76	0.99	1.21	1.82	1.26	0.89	1.92	1.76	1.45	1.24	1.99

Supplementary Figure 4. Number of unique mutations in the genome of SARS-CoV-2. The number of mutations (blue) were taken from 2,058 available genomes from GISAID (March 31st 2020). The number of expected mutations (orange) were calculated based on the assumption that there are no purifying or selection pressures present against different categories of point mutations, and all of the mutations occurring randomly according to the size of each gene.

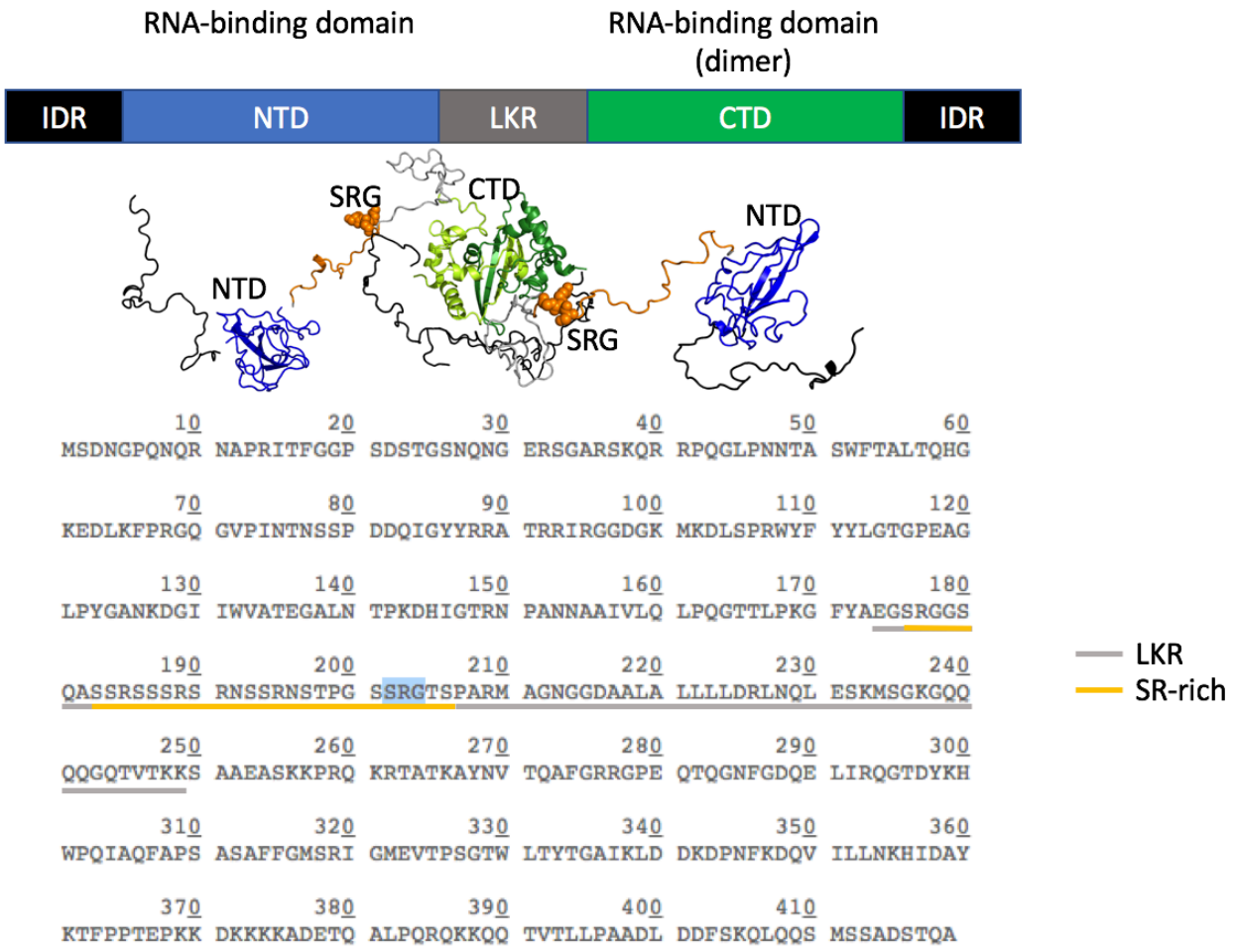


Supplementary Figure 5. Mutations in the SARS-CoV-2 spike protein. (A) An overview of an S trimer. The structure of the SARS-CoV-2 S trimer in its open conformation (chains are cyan, magenta and green) bound to the human receptor ACE2 (black) has been

504 modeled based on PDB accessions 6m17 and 6vyb. Identified nonsynonymous mutations are shown as sphere models. For reasons of
505 visibility only mutations of two of the three S chains are labeled. Memb indicates the membrane. The cytoplasmic tails are depicted in
506 the same colours as the three spike chains. B-E) Magnified regions of the mutations. (B) H49Y: H49 is located in the N-terminal domain
507 (NTD) which is in contact with the sub-domain 2 (SD2) of a symmetry-related molecule. The relative position of NTD and SD2 changes
508 considerably upon going from the closed (grey coloured) to the open state (cyan and magenta). H49 stabilises the rim of the NTD
509 through cation-π-stacking with R44. A tyrosine in position 49 would be able to perform the same role and does not lead to clashes.
510 However, the substitution may slightly alter the stability of this interaction and the interaction between the NTD and SD2, which, in
511 turn, might have subtle effects on the stability and equilibrium of open and closed conformations of S. (C) V367F: V367 is part of the
512 receptor binding domain (RBD), however located too far away from the ACE2-binding site to directly affect receptor binding. V367 is
513 surface exposed, and its substitution would not create clashes with other protein regions. However, the exchange of a small with a bulky
514 hydrophobic residue would alter the surface characteristics of this region, which might influence the efficiency of glycosylation (stick
515 model) of the nearby N343, or the positioning of the sugars. Additionally, the altered RBD surface could potentially interfere with
516 antibody recognition. (D) G476S: G476 is located in the RBD. It is positioned solvent-exposed in a SARS-CoV-2-specific loop. This
517 loop is stabilised by a disulphate bridge (C480:C488; C480 is shown as stick model). In the open, ACE2-bound conformation of the
518 RBD, G476 is close to ACE2 Q24 and E23. The substitution G476S would lead to light clashes with these ACE2 residues (indicated as
519 red spheres) and with the RBD residues N487. However, minor reorientation of the side chains might allow an additional hydrogen bond
520 to be formed between S476 and ACE2 Q24 and E23, thus enhancing the affinity. (E) V483A: V483 is also located in the RBD, solvent-
521 exposed in the same SARS-CoV-2-specific loop as G476 (C480:C488 are shown as stick models). In the open, ACE2-bound
522 conformation of the RBD, V483 is more than 10Å away from the ACE2 receptor, and hence does not contribute to direct binding or
523 stability. In the closed conformation, this loop is not modeled in the EM structures (PDB 6vyb), inferring it is flexible in the absence of

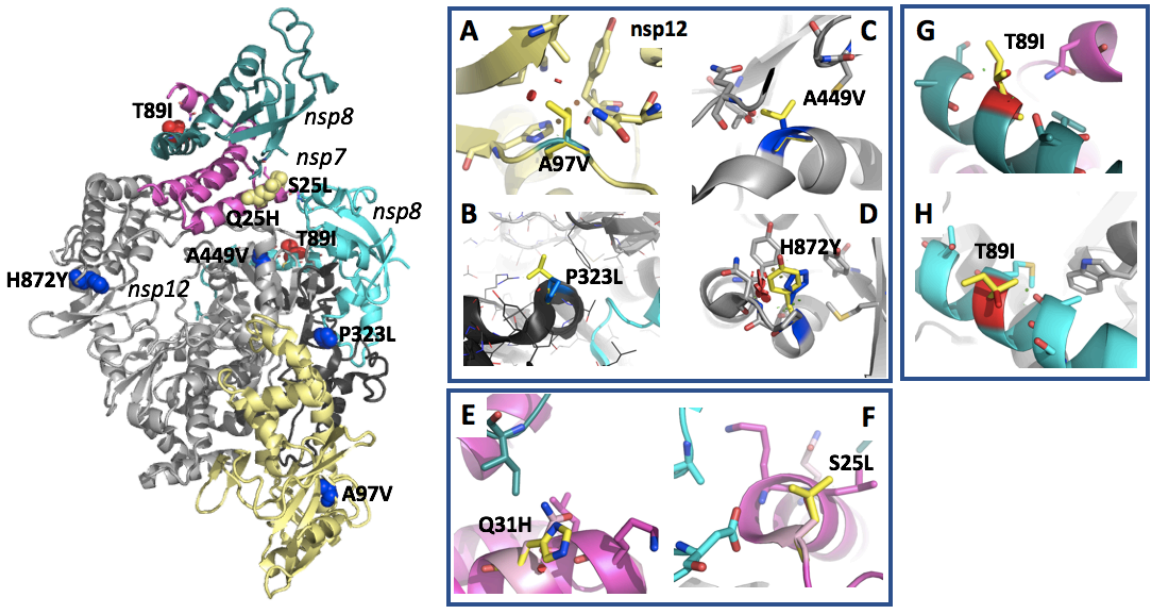
ACE2. Superimposition of the ACE2-bound conformation of the RBD onto RBDs in a closed conformation shows that this loop region would stick out into the solvent. Substitution of V483 is consequently not predicted to have a strong impact on receptor binding or protein stability. By lowering the hydrophobic surface, this substitution might however reduce the non-specific stickiness of this loop region, and/or affect binding of antibodies. (F) D614G: D614 is located in the SD1. In the trimeric S, D612 engages stabilising interactions within the SD1 (R646 or the backbone of F592, depending on the chain) and with the S1 of the adjacent chain (T859 and K854). Replacement of D614 with a glycine would entail losing these stabilising interactions and increase the dynamics in this region. (G) V1040F: V1040 is located in a loop region that makes hydrophobic contacts between stalk regions of the spike trimer. The V1040F substitution is possible without steric hindrance, and would slightly increase the hydrophobic contacts between the chains. The other mutations are not shown in detail, but are evaluated as follows. M153T: M153 is located in the NTD in a solvent-exposed loop. N electron density was modeled for this loop in the cryo-EM structure (6vyb) suggesting it is flexible. The substitution is expected to be neutral. S940F: S940 is located in the stalk region, in a solvent-exposed turn. Introducing the bulkier phenylalanine in this position would not destabilise the structure but locally change the surface characteristics. C1254F: C1254 is the last cysteine in a cysteine-rich unstructured short cytoplasmic region. This region is required for efficient membrane fusion. The exact mechanism remains to be elucidated, and hence we cannot assess the exact impact of the C1254F mutation.

538


RNA-binding domain
RNA-binding domain
(dimer)
IDR
NTD
LKR
CTD
IDR
SRG
CTD
NTD
NTD
SRG
10 20 30 40 50 60
MSDNGPQNQR NAPRITFGGP SDSTGSNQNG ERSGARSKQR RPQGLPNNTA SWFTALTQHG
70 80 90 100 110 120
KEDLKFPRGQ GVPINTNSSP DDQIGYYRRA TRRIRGGDGK MKDLSPRWYF YYLGTGPEAG
130 140 150 160 170 180
LPYGANKDGI IWVATEGALN TPKDHIGTRN PANNAAIVLQ LPQGTTLPKG FYAEGSRGGS
190 200 210 220 230 240
QASSRSSSRS RNSSRNSTPG SSRGTSPARM AGNGGDAALA LLLLDRLNQL ESKMSGKGQQ
250 260 270 280 290 300
QQGQTVTKKS AAEASKKPRQ KRTATKAYNV TQAFGRRGPE QTQGNFGDQE LIRQGTDYKH
310 320 330 340 350 360
WPQIAQFAPS ASAFFGMSRI GMEVTPSGTW LTYTGAIKLD DKDPNFKDQV ILLNKHIDAY
370 380 390 400 410
KTFPPTEPKK DKKKKADETQ ALPQRQKKQQ TVTLLPAADL DDFSKQLQQS MSSADSTQA
LKR
SR-rich

539

Supplementary Figure 6. Mutations in the SARS-CoV-2 nucleocapsid. The SARS-CoV-2 nucleoprotein (N) compacts the viral genome into a helical ribonucleocapsid and hence has key roles in viral assembly[1]. N contains two folded domains, the N-terminal and the dimeric C-terminal domain (NTD; blue, and CTD; green, respectively). Both domains are flanked by flexible regions, namely a flexible linker between them (LKR, underlined in the bottom panel sequence) and ~40 residue intrinsically disordered regions (IDRs) as tails. The NTD and CTD are RNA-binding regions, but the LKR and IDRs also affect RNA-binding of N[2]. The LKR contains a serine-arginine (SR)-rich region, which contributes to RNA binding and N oligomerization. The arginines may promote RNA-binding through electrostatic interactions, whereas the serines are putative phosphorylation sites that would counteract RNA binding and possibly favour oligomerisation upon phosphorylation[3]. The hotspot mutations S202N, R203K and G204R are all within the SR region of the LKR. S202N would delete a putative phosphorylation site. Given the large number of alternative serine phosphorylation sites, and the presence of many asparagines within the linker, this S/N substitution might not have a very strong influence. The homologous substitution R203K is also expected to have only minor effects. G204R might increase binding to RNA, and/or affect other homo- or heterologous interactions. In summary, the substitutions might potentially enhance RNA binding and alter the response to serine phosphorylation events. The structural figure consists of homology models for the NTD and CTD, linked by arbitrary but stereochemically-plausible linker regions.

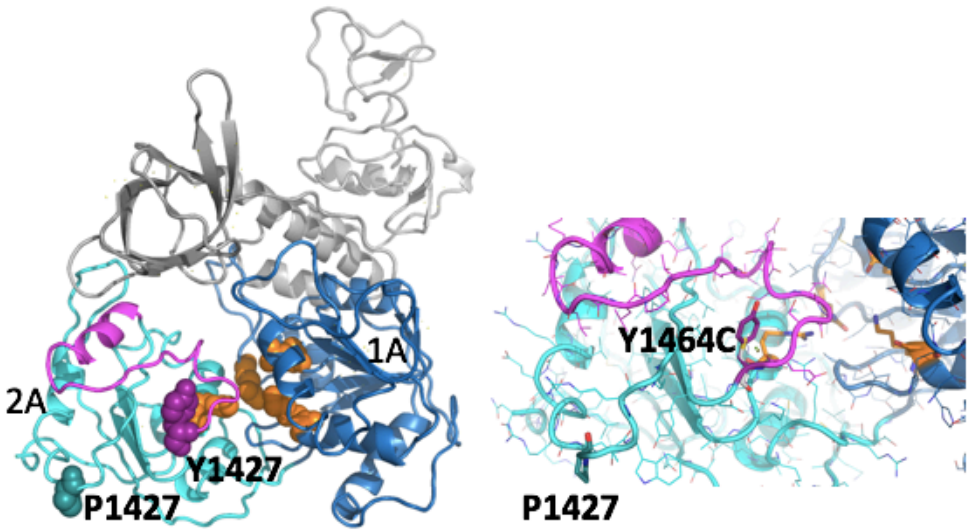


Supplementary Figure 7. Mutations in SARS-CoV-2 *ORF1ab*, nsp12: The nsp12 RNA-dependent RNA polymerase forms a complex with one nsp7 and two nsp8, which markedly enhances its polymerase activity[4]. The shown SARS-CoV-2 nsp12 structure is composed of a nidovirus-unique N-terminal extension (pale yellow), a linker domain (black) and the RNA-dependent RNA polymerase (RdRp) domain (grey). The structure shows the nsp12 in complex with nsp7 (magenta) and nsp8 (cyan and teal), based on PDB 7btf. (A) A4494 (A97 in nsp12 numbering; shown as a blue sphere models) is located in the N-terminal extension of the polymerase[5,6]. A4494 is sealing the hydrophobic core of the N-terminal lobe, and its side chain is not solvent exposed. Its replacement with the hydrophobic but slightly bigger valine (yellow) only leads to minor clashes (small red discs) that would not have a significant impact on the function or stability. (B) P4720 (P323 in nsp12 numbering) is located in the 'interface domain' (black). In this position, the P323L substitution (yellow) is

not predicted to disrupt the folding or protein interactions and hence is not expected to have strong effects. (C) A4846V (residue A449 in nsp12 numbering; blue) is located in the finger domain, with its side chain pointing inwards, contributing to a hydrophobic interaction with the adjacent beta strand. The substitution of A4846 by a valine (yellow) is tolerated in this context. Leading only to minor clashes, it might slightly improve the stability of this region. (D) H5269 (H872 in nsp12 numbering; blue) is located in the thumb domain. It is at the tip of a solvent exposed turn, where its replacement by a tyrosine (yellow) would be tolerated without functional impact.

Nsp7: S2884 and Q3890 (S25 and Q31 in nsp7 numbering; both in light yellow) are solvent exposed on a helix that makes contact with both nsp8 and nsp12. (E) S25 is capping the N-terminal end of this helix. Its substitution with a leucine (yellow) does not cause steric problems, but would lead to loss of the capping hydrogen bond. However, D163 from one of the nsp8 molecules also performs a capping function in the complex, and hence S25L would only have a slightly destabilising effect. (F) Q31 is located on the surface of the helix. Although Q31 is close to nsp8, its substitution with a histidine (yellow) would not influence this interaction measurably.

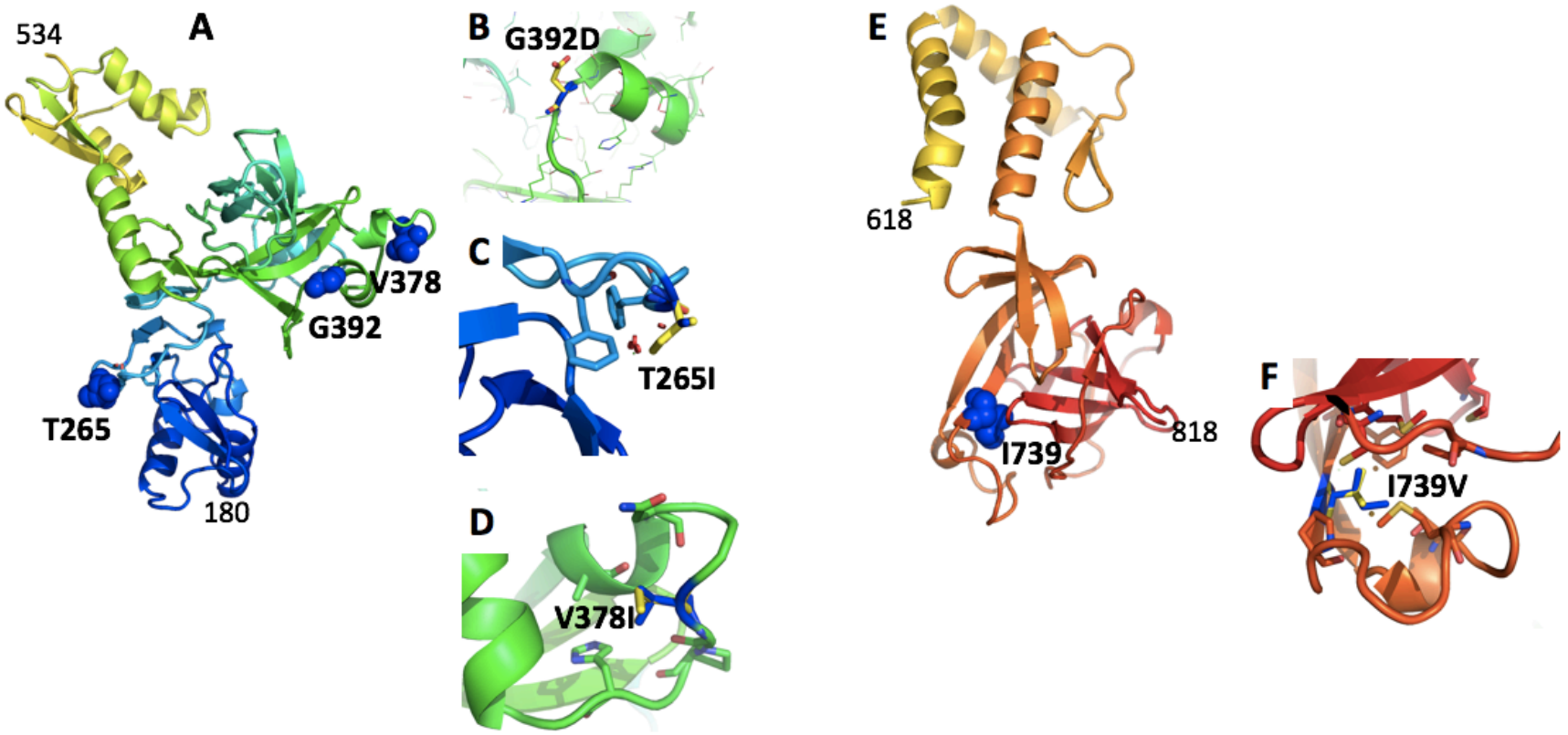
Nsp8: Only T4031 (T89I in nsp8 numbering; red) is included in the structural model. In both nsp8 molecules, T89 is located solvent exposed on a helix. In neither of the two nsp8 molecules would the substitution T89I create steric problem or affect the interaction with nsp12 or nsp7.



583

584 **Supplementary Figure 8. Mutations in SARS-CoV-2 *ORF1ab,* nsp13:** P1427L and Y1464C are both in the helicase nsp13, which
585 catalyses the unwinding of duplex oligonucleotides into single strands. Both residues are located on the same side of the 2A domain of
586 the helicase. The 2A and adjacent 1A domains coordinate together to complete the final unwinding process. The helicase has been
587 modelled based on the 99.8% identical SARS nsp13 (PDB id 6jyt). The domains 1A and 2A are coloured in blue and cyan, respectively.
588 The residues involved in NTP binding are shown in orange. Residues on domain 2A that are involved in RNA binding are shown in
589 magenta. The other domains are grey. *Left*: overview of the complete structure. Key residues are shown as sphere models. *Right*: zoom
590 into the mutated area. Key residues are shown as stick models. For Y1464 the *in silico* mutated cysteine is shown in white. P1427 is

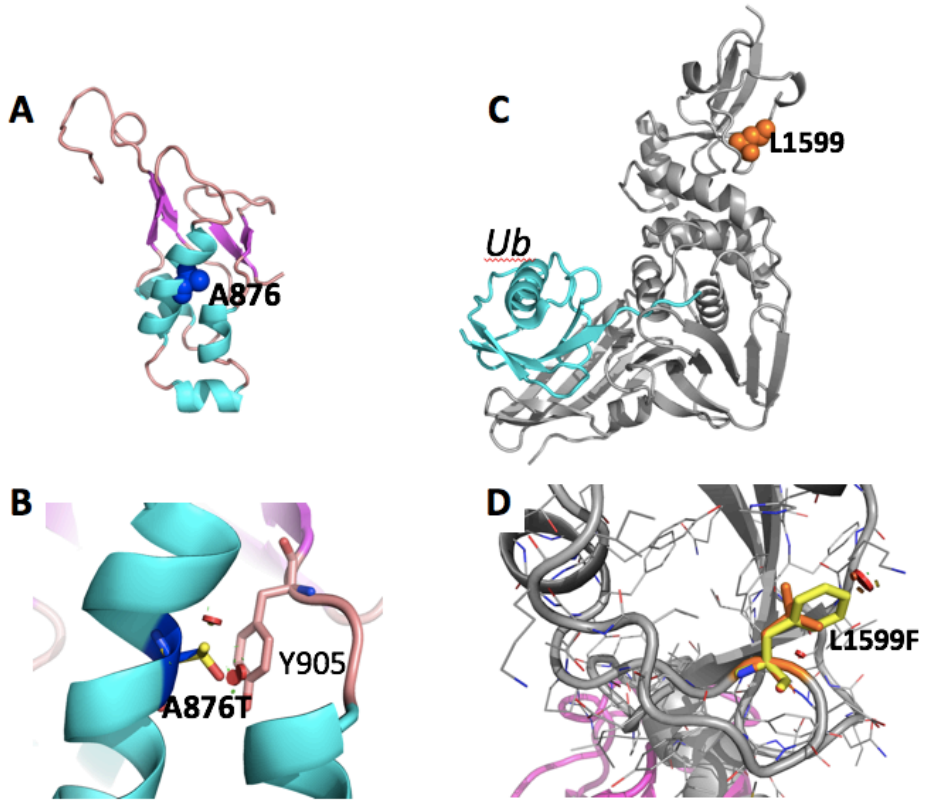
591 located in a solvent-exposed loop region that has not yet reported to be involved in nucleotide binding. Its substitution with a leucine is
592 not expected to create noticeable effects. Y1464 is part of a region that contributes to binding and unwinding of duplex oligonucleotides[7].
593 Its substitution by a cysteine would decrease the stability and enhance the dynamics of this particular region, and might affect RNA
594 binding and processing.



Supplementary Figure 9. Mutations in SARS-CoV-2 *ORF1a*: nsp2.

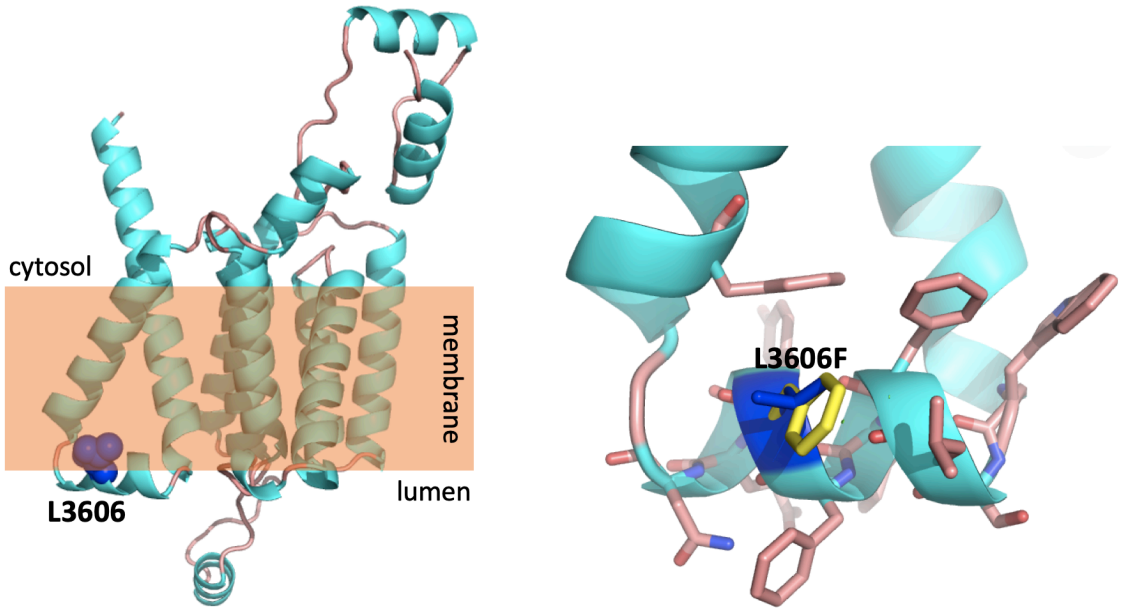
No experimental structure of sufficiently close homologue is known for nsp2. This structural analysis is therefore based on the proposed theoretical AlphaFold model. (A) Shown is a model for residues 180-534, colour ramped from blue to yellow. T265 and G392 are shown as blue sphere models. (B) T265 (blue stick model, corresponding to residue T58 in the nsp2 cleavage product numbering) is located at the tip of a loop that is pinned to the core of the N-terminal domain (blue) via hydrophobic residues (two phenylalanines are shown as

stick models). Being exposed to the solvent, the T265I substitution (shown in yellow) is not expected to have significant impact on protein fold or function. (C) G392 (G212 in nsp2 numbering; shown in blue) is placed in a solvent accessible loop, according to the AlphaFold model. In this position even the non-conservative substitution with an aspartic acid (yellow) is not predicted to have significant impact on protein fold or function. (D) V378 is located in a surface exposed loop. Its substitution by an isoleucine will not create steric clashes, nor lead to loss of hydrophobic interactions. (E) A second nsp2 fragment is shown, colour ramped from yellow to red, comprising residues 618 to 818. I793 is shown as blue sphere model. (F) I739V (I559 in nsp2 numbering; shown in blue) is part of a hydrophobic core of a small C-terminal domain. Its substitution with an only slightly smaller hydrophobic valine (yellow) would only have insignificant destabilising effects on this region.

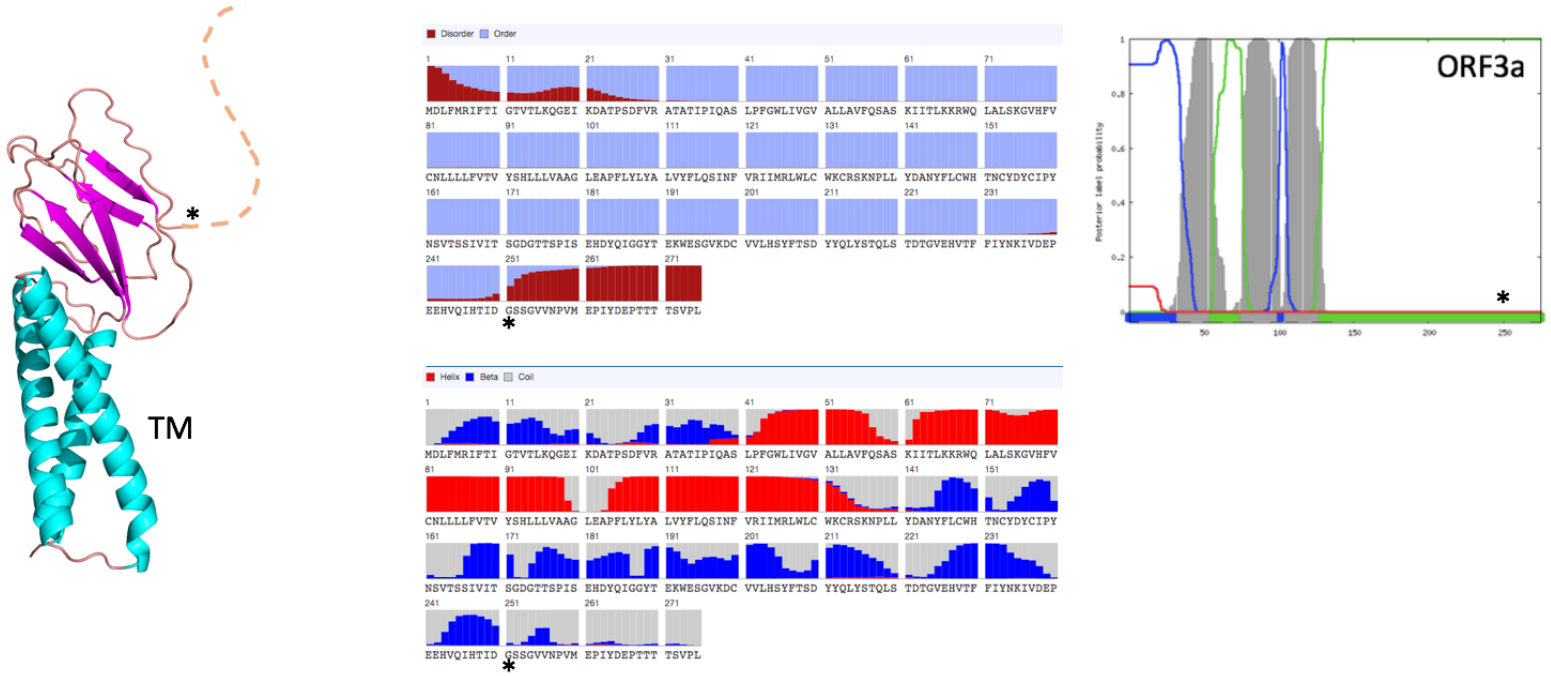


Supplementary Figure 10. Mutations in SARS-CoV-2 *ORF1a*: nsp3. The substitution A876T (corresponding to residue A58 in the nsp3 cleavage product numbering) is situated in the N-terminal ubiquitin-like domain of nsp3. (A) The structure of this domain can be inferred based on the 79% identical structure of residues 1-112 from SARS-CoV (PDB id 2idy). Helices are coloured in cyan, strands in magenta and loops in pale orange. A876 is shown as blue sphere model. (B) Zoom onto A876. A876 (blue stick model) is placed within a helix, engaging hydrophobic contacts with Y905 (stick model). The substitution into a slightly larger threonine (yellow stick

615 model) can be accommodated with only minor structural adjustments (clashes are shown as red spheres) and hence the A876T mutation
616 is not expected to have a substantial influence on the proteins stability and function. (C) L1599F is located in the papain-like protease.
617 The structure (grey) has been modelled based on the ~83% identical SARS-CoV structures in complex with ubiquitin (Ub, cyan, based
618 on 4m0w). L1599 is remote from the active site in the Ubl domain. (D) Its substitution does not lead to clashes and is expected to be
619 neutral in terms of function and protein stability. No templates have been identified for V378I.

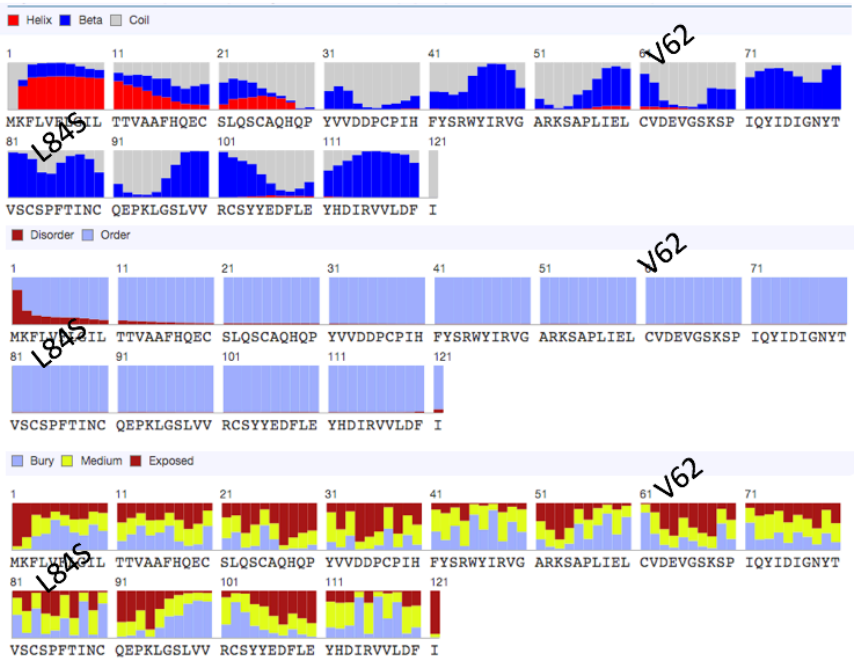


Supplementary Figure 11. Mutations in SARS-CoV-2 *ORF1a*: nsp6. L3606 (corresponding to residue L37 in the nsp6 cleavage product) is a multi-pass transmembrane protein for which no experimental structure of sufficiently close homologue is known. This structural analysis is therefore based on the proposed theoretical AlphaFold model (*Left*). Helices are colored in cyan, loops in pale orange. L3606 is shown as blue sphere model. The endoplasmic reticulum membrane is shown in orange; lumen and cytoplasmic sides are indicated. *Right*: close-up view of the location of L3606 (shown as blue stick model). L3606 is located in a predicted helical region that is partly submerged in the membrane, lying parallel to its luminal surface. According to the structural model, L3606 is exposed to the membrane, and hence its substitution with a larger but still hydrophobic phenylalanine would not impact structure or function.



629

630 **Supplementary Figure 12. Mutations in SARS-CoV-2 *ORF3a*.** ORF3a is a viroporin that forms a pentameric potassium-sensitive ion
631 channel. ORF3a activates the inflammasome which facilitates viral release and aggravates disease symptoms[8]. ORF3a contains a 3-pass
632 α-helical TM region (cyan in left panel, and greyed regions in TM prediction, right panel) and a domain predicted to have a β-sandwich
633 fold (magenta; see also middle panel and right panel). The ORF3a N- and C-terminal tails are predicted to be disordered (middle panel,
634 top). A theoretical model for the Orf3a monomer has been proposed by AlphaFold[9]. The structure-function relationship of this protein
635 remains to be clarified. The mutation G251V is located C-terminal to the β-sandwich domain and the tail (marked by an asterisk). In
636 this position, the substitution is not expected to affect the protein fold or function significantly.



Supplementary Figure 13. Mutations in SARS-CoV-2 *ORF8*. ORF8 has an N-terminal *sec*-pathway signal peptide with a cleavage site after residue 15, suggesting that it is secreted into the extracellular space. Following signal peptide cleavage, the ORF8 protein core is predicted to consist largely of β-strands and features seven cysteines (top panel). We predict that this protein adopts a cysteine disulfide-bond stabilised β-sandwich structure inferring that ORF8 also functions as a ligand binding module. Homology modeling or *ab initio* servers failed to produce a model consistent with the proteins' secondary structure predictions. V62 and L84 are predicted to be partially solvent-exposed (bottom), and located at the end (V62) or in the middle (L84) of secondary structural elements (depending on the prediction server, the region surrounding V62 is predicted as strand or helix; L84 is in a β strand). Hence, we can only speculate that the mutations V62L and L84S might have little effect on the stability of the disulphate-stabilised fold, but might have minor (or no) effects on ligand binding.

SUPPLEMENTARY TABLES

Supplementary Table 2. Experimentally validated amino acid substitutions resulting in escape mutants or having functional significance in N and S proteins of SARS-CoV and MERS-CoV.

Protein	Residue	Virus	Substitution in SARS-CoV-2	GISAID (EPI_ISL)	Functional role of the residue	Reference
N	T149	SARS-CoV	T148I	406595	T149 is an epitope that interacts with neutralising antibody.	[10]
	K250	SARS-CoV	K249I	408515	K250 is an essential motif for binding SN5-25 antibody.	[10]
	S328	SARS-CoV	S327L	413557	S328 forms inter-dimer disulfide bonds connecting the two β-strands.	[11]
S	R395	SARS-CoV	R408I	413522	R395 interacts with one of the most potent neutralising antibody m396.	[12]
	D463	SARS-CoV	G476S	417085, 417380, 418055, 418077, 417353, 41708, 416447	D463G is associated with neutralisation escape from MAbs S224.1.	[13]
	I529	MERS-CoV	V483A	414618, 415596, 415605, 417072, 417073, 417075, 417076, 417111, 417139, 417159, 417160, 418029, 418071	I529T reduces affinity of RBD to the DPP4/CD26 receptor.	[14]

SUPPLEMENTARY REFERENCES:

1. Chang, C. K., Hou, M. H., Chang, C. F., Hsiao, C. D. & Huang, T. H. The SARS coronavirus nucleocapsid protein - Forms and functions. *Antiviral Research* (2014). doi:10.1016/j.antiviral.2013.12.009
2. Chang, C.-K. *et al.* Multiple Nucleic Acid Binding Sites and Intrinsic Disorder of Severe Acute Respiratory Syndrome Coronavirus Nucleocapsid Protein: Implications for Ribonucleocapsid Protein Packaging. *J. Virol.* (2009). doi:10.1128/jvi.02001-08
3. Peng, T. Y., Lee, K. R. & Tarn, W. Y. Phosphorylation of the arginine/serine dipeptide-rich motif of the severe acute respiratory syndrome coronavirus nucleocapsid protein modulates its multimerization, translation inhibitory activity and cellular localization. *FEBS J.* **275**, 4152–4163 (2008).
4. Subissi, L. *et al.* One severe acute respiratory syndrome coronavirus protein complex integrates processive RNA polymerase and exonuclease activities. *Proc. Natl. Acad. Sci. U. S. A.* (2014). doi:10.1073/pnas.1323705111
5. Gao, Y. *et al.* Structure of the RNA-dependent RNA polymerase from COVID-19 virus. *Science* **7498**, 1–9 (2020).
6. Kirchdoerfer, R. N. & Ward, A. B. Structure of the SARS-CoV nsp12 polymerase bound to nsp7 and nsp8 co-factors. *Nat. Commun.* (2019). doi:10.1038/s41467-019-10280-3
7. Jia, Z. *et al.* Delicate structural coordination of the Severe Acute Respiratory Syndrome coronavirus Nsp13 upon ATP hydrolysis. *Nucleic Acids Res.* (2019). doi:10.1093/nar/gkz409
8. Siu, K. L. *et al.* Severe acute respiratory syndrome Coronavirus ORF3a protein activates the NLRP3 inflammasome by promoting TRAF3-dependent ubiquitination of ASC. *FASEB J.* (2019). doi:10.1096/fj.201802418R

9. John Jumper, Tunyasuvunakool, K., Kohli, P. & Hassabis, D. Computational predictions of protein structures associated with COVID-19. (2020). Available at: https://deepmind.com/research/open-source/computational-predictions-of-protein-structures-associated-with-COVID-19.
10. Lin, Y. *et al.* Identification of an epitope of SARS-coronavirus nucleocapsid protein. *Cell Res.* (2003). doi:10.1038/sj.cr.7290158
11. Chang, C. ke, Chen, C. M. M., Chiang, M. hui, Hsu, Y. lan & Huang, T. huang. Transient Oligomerization of the SARS-CoV N Protein - Implication for Virus Ribonucleoprotein Packaging. *PLoS One* (2013). doi:10.1371/journal.pone.0065045
12. Tian, X. *et al.* Potent binding of 2019 novel coronavirus spike protein by a SARS coronavirus-specific human monoclonal antibody. *Emerging Microbes and Infections* (2020). doi:10.1080/22221751.2020.1729069
13. Rockx, B. *et al.* Escape from Human Monoclonal Antibody Neutralization Affects In Vitro and In Vivo Fitness of Severe Acute Respiratory Syndrome Coronavirus. *J. Infect. Dis.* (2010). doi:10.1086/651022
14. Kim, Y. *et al.* Spread of mutant middle east respiratory syndrome coronavirus with reduced affinity to human CD26 during the south Korean outbreak. *MBio* (2016). doi:10.1128/mBio.00019-16