Evolution to increased positive charge on the viral spike protein may be part of the adaptation of SARS-CoV-2 to human transmission.

Matthew Cotten*, My V.T. Phan

Author affiliations: Medical Research Council–University of Glasgow Centre for Virus Research, Glasgow, Scotland, UK (MC), UK Medical Research Council–Uganda Virus Research Institute and London School of Hygiene and Tropical Medicine Uganda Research Unit, Entebbe, Uganda (MC, MVTP); UK

*Correspondence to Matthew Cotten, email: matthew.cotten@lshtm.ac.uk

Abstract

The severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), the causative agent of the coronavirus disease 2019 (COVID-19) epidemic, continues to evolve and infect individuals. The physical properties of the surface proteins of the virus such as charge, size, hydrophobicity and folding may influence survival and promoting movement through the environment and within the human airway to reach target cells to bind, enter and establish the next round of infection. The exterior surface of the SARS-CoV-2 virion is dominated by the spike protein and the current work examined spike protein biochemical features that have changed during the 2 years that SARS-CoV-2 has infected humans. In addition to selective pressure to avoid immune recognition of viral proteins, we hypothesised that SARS-CoV-2 emerged from an animal reservoir capable of human infection and transmission but in a sub-optimum state and a second level of selective pressure acts on these biochemical features. Our analysis identified a striking change in spike protein charge occurring across the 2 years of human host evolution of SARS-CoV-2.

Introduction

The severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), the causative agent of the coronavirus disease 2019 (COVID-19) epidemic, continues to evolve and infect individuals. Virion biochemical properties play an important role in controlling virus transmission. After replication in an infected individual and release from an infected cell, onward transmission requires survival of the virion to reach susceptible cells in a new host individual initiating the next round of infection. The physical properties of the surface proteins of the virus such as charge, size, hydrophobicity and folding may influence movement of the virion through the environment, promoting or limiting binding of the virion to the external surfaces. Once reaching a susceptible individual, virion physical properties may influence movement within the human airway and determine the ability of an infecting virion to reach target cells to bind, enter and replicate (Adamczyk et al. 2021). The exterior surface of the SARS-CoV-2 virion is dominated by the spike protein and the current work examines simple spike protein features have changed during the 2 years that SARS-CoV-2 has been infecting humans. In addition to selective pressure to avoid immune recognition of viral proteins, we hypothesise that SARS-CoV-2 emerged from an animal reservoir capable of human infection and transmission but in a sub-optimum state. Additionally, there is a second level of selective pressure to adjust to the physical transmission between humans. Evidence for this adaptation can be found in changes in the spike protein over recent evolution. With over 11 million SARS-CoV-2 genomic sequences generated from across the pandemic, many of these sequences have intact spike gene sequences which can be used to monitor change across the 2 years of human host evolution of this virus.

Much of the observed spike protein substitutions may be in response to the developing immune response to this new pathogen, which is reflected in substitutions occurring in the immune-exposed S1 domain of the spike protein and there is ample evidence that these many of the spike protein changes allow escape from host immunity (Tzou et al. 2022)(Greaney, Loes, et al.

2021)(Greaney, Starr, et al. 2021)(Greaney et al. 2022) (Cao et al. 2022) (Dejnirattisai et al. 2022) (DeGrace et al. 2022). There may also be evolutionary selection for changes that improve host interactions apart from immune evasion and these include altering spike/receptor binding kinetics, protease cleavage events, tertiary structure (S1/S2 interactions after cleavage) or the physical properties of the virion (charge, hydrophobicity, and protein folding or secondary structure) in ways that might improve transmission. To explore the role of the biochemical features of the spike protein in human transmission, we monitored changes in spike biochemical features over the two years that SARS-CoV-2 has been evolving in humans and report a surprising change in the spike protein charge.

Results.

The SARS-CoV-2 spike protein physical features were calculated from spike protein sequences from across 2 years of the COVID-19 epidemic. Features that could be quantitated from protein sequence were used (see Methods), including charge at pH 7.4, Kyle and Doolittle GRAVY score (Kyte & Doolittle 1982) (a measure of hydrophobicity), an instability index derived from dipeptide content (Guruprasad et al. 1990), properties influencing protein folding (percent helix, fold or sheet as predicted from amino acid content), individual amino acid total fraction and di-amino acid total fraction.

A dominant pattern of SARS-CoV-2 evolution during the two years of human adaptation has been the regular appearance and the subsequent regional and then global dominance of lineages. These lineages typically encode a small set of amino acid changes from earlier lineages, many of which are likely to provide temporary or long-term advantage for the viral lineage. A principal component analysis was performed to identify spike physical features most strongly linked with SARS-CoV-2 lineage from all available genome entries retrieved from GISAID (up to 15 June 2022) (Figure 1). Intact spike protein sequences were extracted and the physical features of each protein were collected into a matrix. The top feature distinguishing SARS-CoV-2 lineage were identified with charge as the most important feature (Figure 1A). A principal component analysis using the top 8 features (charge, gravy, fraction T, fraction R, instability, fraction G, fraction D and fraction K), provided clustering of spike sequences by lineage (Figure 1B). The results support the idea that spike protein charge (among other features) is an important determinant of the lineages that have evolved during the first two years of the COVID-19 epidemic.

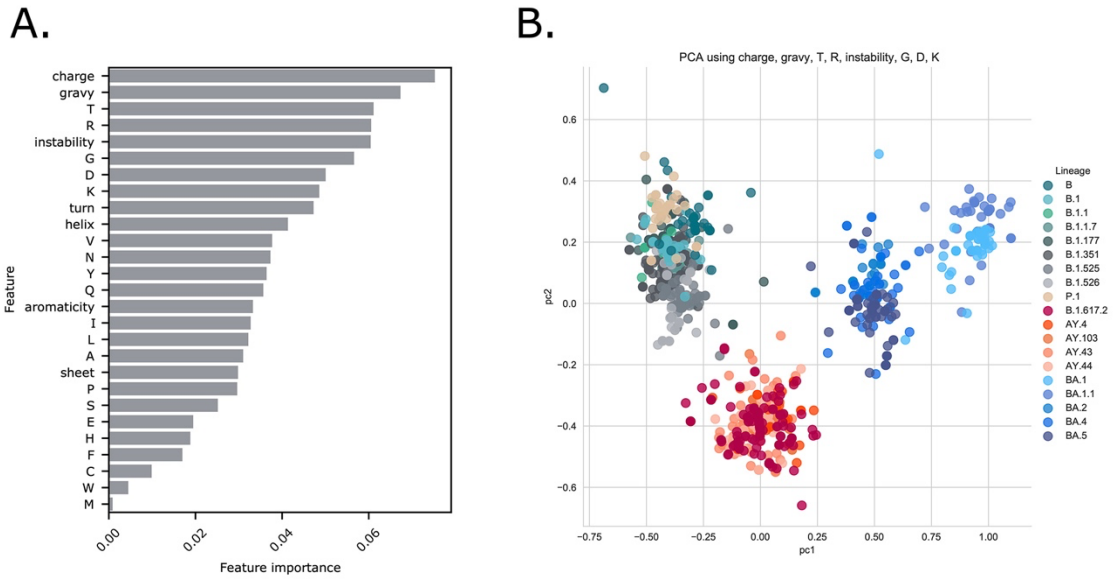

Figure 1. Identification of spike protein charge association with SARS-CoV-2 lineage. **Panel A:** A set of 300 spikes sequences extracted from the first 300 SARS-CoV-2 genomes per lineage (by date of collection) was analyzed, features for each sequence were collected (see Methods). SKLearn feature selection (Pedregosa, F. and Varoquaux, G. and Gramfort, A. and Michel, V. et al. 2011) was used to identify features that most accurately identified the sequence lineage. The importance of features were ranked in order. **Panel B:** The top 8 features (charge, gravy, fraction T, fraction R, instability, fraction G, fraction D, fraction K) were further used in a principal component analysis to cluster the same set of SARS-CoV-2 spike sequences. Each node represents a single spike sequence, nodes were coloured by Pangolin lineage assigned to the genome from which the spike sequence was obtained. Lineage colouring is explained in the figure legend to the right**.**

Changes in charge of spike protein across the epidemic were investigated. Plotting total spike charge for all genomes per month of the epidemic showed a clear pattern of the increase in charge over two years of evolution (Figure 2, panel A). Median spike charge was -8.3 in the original SARS-CoV-2 viruses reported in late 2019 to early 2020, but by March 2020, an increase in positive charge to -7.28 was observed. Subsequently, an additional increase in positive charge occurred in mid-2021 to -3.28, and most recently a charge increase occurred in late 2020/early 2021 to around -1.26.

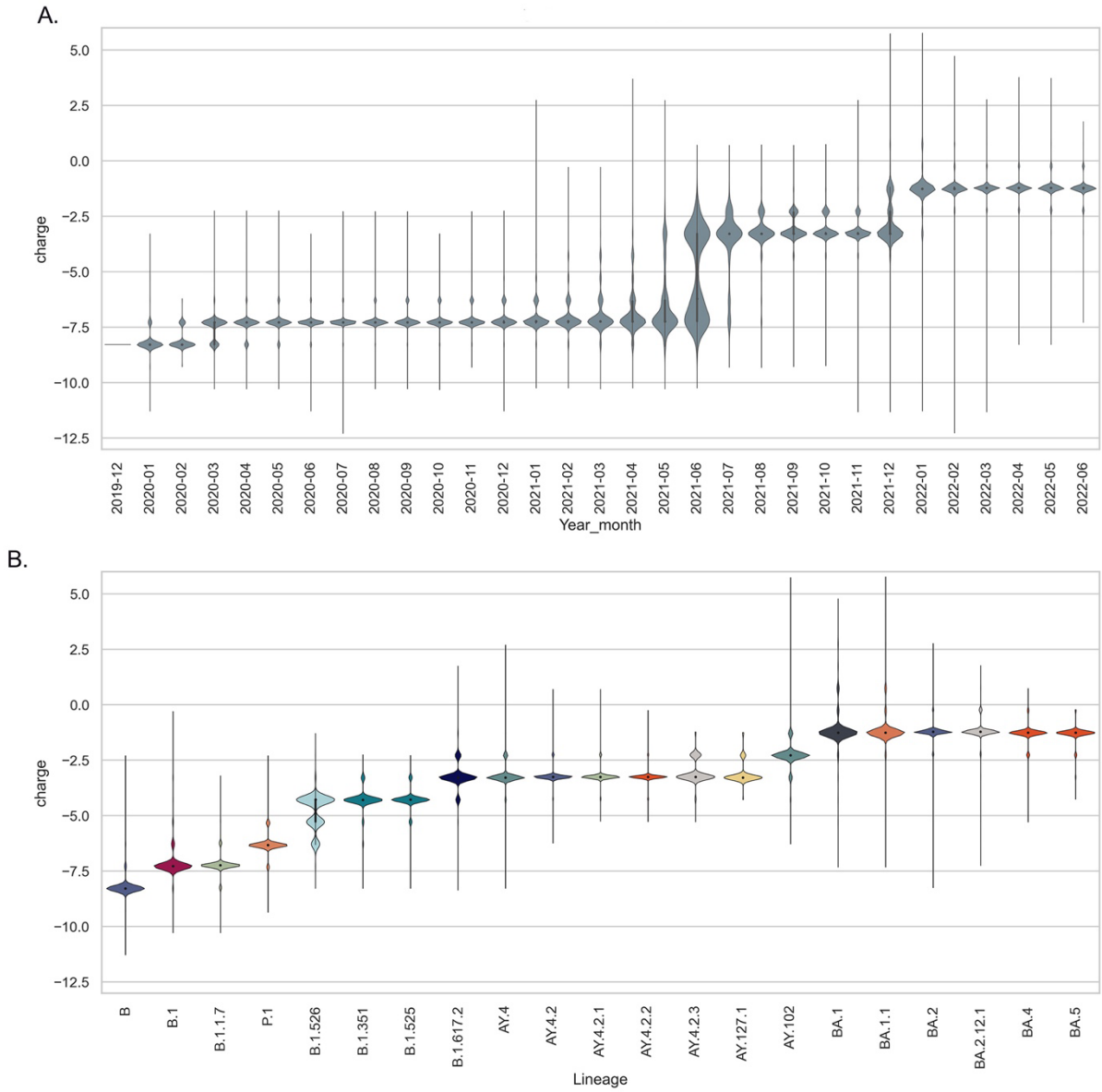

Figure 2.Panel A. Total SARS-CoV-2 spike charge per epidemic month. All available SARS-CoV-2 genomes up to 15 June 2022 were retrieved from GISAID (GISAID 2020) and the spike protein sequence was extracted (if intact). Total charge at pH 7.4 was calculated and values were plotted using a violin plot by month of sample collection. Values were plotted using a violin plot by month of sample collection. **Panel B. Spike charge in major SARS-CoV-2 lineages.** For each lineage, all available spike sequences were collected (up to 15 June 2022), total charge was measured and violin plots prepared to show the charge distribution by lineage. Lineages (indicated at bottom of chart) were ordered by their appearance in the epidemic.

These spike protein charge increases can be attributed to the major successful lineages reported over time (Figure 2B). The B.1, B.1.1 and B.1.1.7 (Alpha) lineages that dominated the first year of the epidemic encoded spike proteins with charges between -8 and -6 while the B.1.351 (Beta) and B.1.525 (Eta) lineages showed a further increase in charge to around -4.5. The B.1.617.2 (Delta) lineage and sublineages (AY.x) displayed further increase in charge. Most recently, the Omicron variants (including BA.1, BA.1.1, BA.2, BA.2.12,BA.3, BA.4 and BA.5) show further spike

charge increases with the majority of Omicron encoded spike proteins showing charge around -1.3 (Figure 2, panel B).

Some indication of functional consequences of the observed changes in spike charge can be obtained from the location on the substitutions in the spike protein. Sets of spike sequences (extracted from the first 300 reported genomes per select lineage) were processed to illustrate the changes to more negative charge (blue) or more positive charge (orange/red) in the protein relative to the initial Lineage B genome sequences. The initial change in charge was a substitution of an aspartic acid residue (D, -1) with a glycine (G, neutral). In some early lineages (e.g. A.23.1), proline (P) 681was substituted with the positively charged arginine (R),or Q680 was substituted with a partially charged histidine H residue. The P681R positive substitution promotes furin cleavage and activation of the spike protein for cell fusion (Lubinski et al. 2022)(Liu et al. 2022). The Delta lineage spikes proteins encoded additional positive charge in the ACE2 binding region, as well as in the far amino terminal region and near the heptad repeat (HR1) which may also enhance membrane fusion activity. More recently, a number of positive substitutions have occurred in the Omicron lineage virus spike proteins with predominance of positive charge changes in the receptor binding domain (Figure 3), suggesting a role of increased charge in spike/receptor interactions.

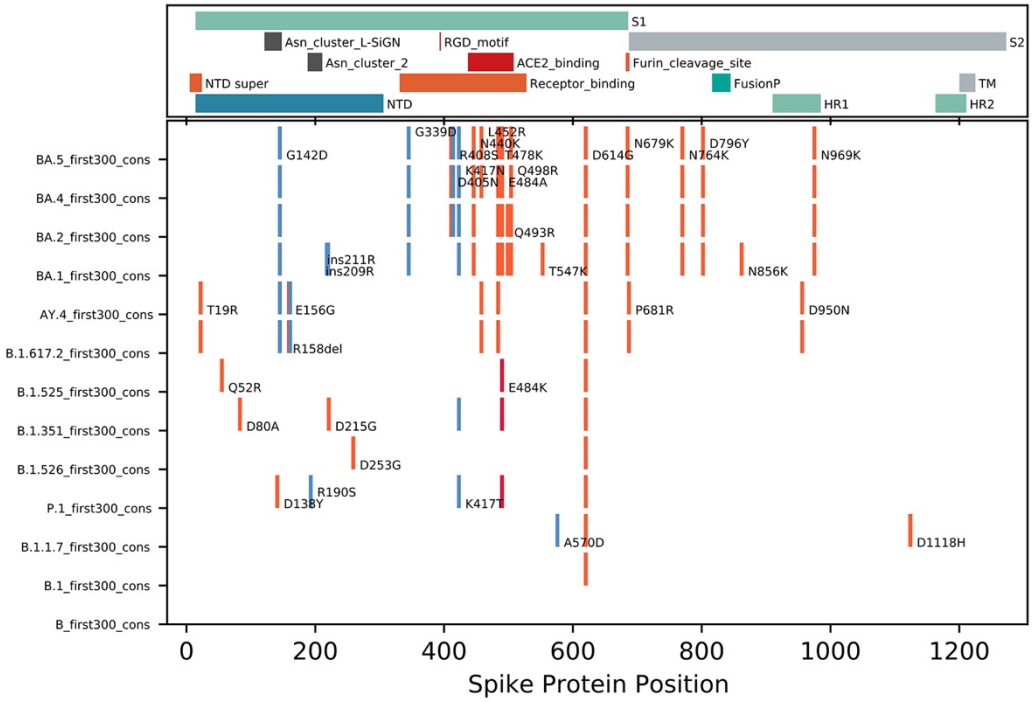

Figure 3. Location of charged amino acid changes in the spike protein. The spike protein sequences encoded by the first 300 reported genomes for the indicated SARS-CoV-2 lineages were collected, and charged amino acid changes from the original B lineage spike sequence were plotted. Charge changes were colored from dark blue (change from positive to negative charged amino acid (aa)), blue change from neutral to negative charged aa), orange (change from neutral to positive charged aa) and red (change from negative to positive charged aa). Substitutions are indicated by original aa/position in reference sequence spike/novel aa. The GenBank NC_045512

genome was used as reference. Key spike protein features of the SARS-CoV-2 spike protein are shown in the upper panel of the figure.

We investigated if a similar pattern of spike charge evolution could be observed in other coronaviruses that have made a transition to human transmission. In recent history, several additional coronaviruses have been observed to move between an animal reservoir and humans. For example, coronavirus 229E is commonly detected in humans and a very close coronavirus was identified in camels and humans suggesting recent movement of the virus between hosts (Corman et al. 2016) (Sabir et al. 2016). All available coronavirus 229E full genomes sequences were retrieved from GenBank, the spike coding region was extracted from the genomes, translated and total charge was calculated. A difference of almost 9 charge units was observed in spike median charge comparing 229E viruses from camel vs. human infections (Figure 4A).

Infection with coronavirus OC43 is common in humans and closely related viruses are found in cattle (bovine coronavirus, BCoV) (Vijgen et al. 2006) and pigs (porcine hemagglutinating encephalomyelitis virus, PHEV) (Vijgen et al. 2005) (Vijgen et al. 2006). Comparing the three OC43-type virus groups, the human virus OC43 has an increased charge of ca. 5 units compared to PHEV and ca.7 units compared to BCoV (Figure 4B).

The commonly known host for the Middle East Respiratory Syndrome coronavirus (MERS-CoV) is dromedary camels; however, zoonosis and serious human infections occur frequently (Cotten et al. 2013) (Cotten et al. 2014) (Memish et al. 2014) (Zhou et al. 2021) (So et al. 2019) reviewed in (Peiris & Perlman 2022). From 698 full MERS-CoV genomes available in GenBank, there was no strong difference in the encoded spike charge of virus sequences derived from human vs. camels infections (Figure 4C).

Considering the location of the charge differences in the spike proteins, for coronavirus 229E, the charge increases occurred throughout the protein, although there is a slightly higher number of positive changes in the receptor binding region of the human infection derived viruses (Figure 4D). For OC43, the porcine and human viruses also show increases in positive charge throughout the spike protein, the porcine PHEV also showed a slight enrichment in positive charge in the receptor binding region (Figure 4E).

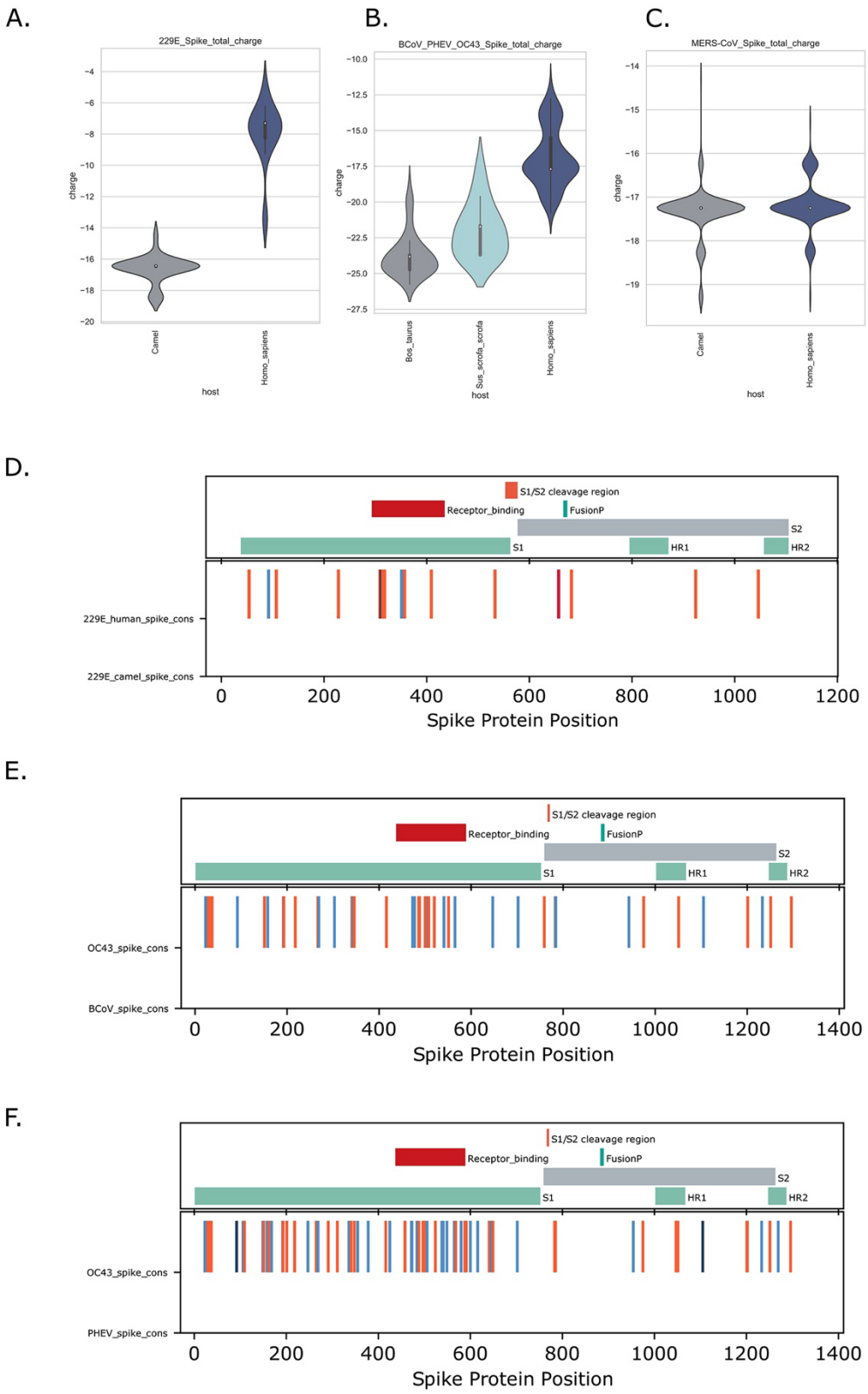

Figure 4. Spike charges from select groups of coronaviruses that have moved into humans (Panels A-C). All available full genomes for the indicated coronaviruses were retrieved from

GenBank, the spike coding region was identified and translated into protein and total charge at ph 7.4 was calculated. The data included 217 OC43 full genome sequences, 109 BCoV full genome sequences and 14 PHEV full genome sequences. Violin plots indicate the charges of each collection of spike proteins, median values are indicated by the open square. **Panel A**. Coronavirus 229E from camel or human infections, **Panel B**. BCoV (from bovine infections) PHEV (from porcine infections) and OC43 (from human infection), **Panel C.** MERS-CoV from camel or human infection. **Panel D-F: Consensus spike protein sequences were generated from the indicated virus groups and charged amino acid changes were determined.** Charge changes were colored from dark blue (change from positively to negatively charged amino acid (aa)), blue change from neutral to negatively charged aa), orange (change from neutral to positively charged aa) and red (change from negative to positively charged aa). **Panel D:** 229E spike from human infections compared to 229E spike from camel infections, **Panel E**: Human OC43 spike compared to BCoV spike, **Panel F**: Human OC43 spike compared to PHEV spike. Key spike protein features of each group's spike protein are shown in the upper portion of each panel.

Discussion

After more than two years of the COVID-19 pandemic and with the availability of >11 million SARS-CoV-2 genome sequences, a trend of SARS-CoV-2 spike charge can be observed, with successive lineages showing an increase in positive charge over earlier lineages. Over the course of the pandemic, SARS-CoV-2 spike protein has evolved from a protein with a total charge of -8.28 in the original Lineage A and B viruses to a protein with a total charge of -1.26 in the currently circulating Omicon lineage viruses. This pattern has been noted and commented on previously (Pawłowski 2021) (Nie et al. 2022). We expand on these observations, and document sites of change in the spike protein and explore similar phenomena of evolution to more positive charge in two other coronaviruses (OC43 and 229E) that have moved between animals and humans.

This study does not identify a mechanistic basis for the increased spike charge although there are several possible transmission steps that might be promoted by increasing charge. Exposed, positively charged spike protein should promote interactions with negatively charged cellular structures. Negatively charged heparin interactions have been reported with SARS-CoV-2 spike (Kim et al. 2020) and negatively charged sialylated glycans are reported to promote entry of SARS-CoV-2 (Nguyen et al. 2022). The upper respiratory tract is coated by and protected by mucins, frequently modified with sialic acid or phosphorylated, high mannose N-glycans (Byrd-Leotis et al. 2021) which present a negatively charged matrix that could either promote or protect against viral transmission. The SARS-CoV-2, OC43, and BCoV virions display binding to negatively charged carbohydrate structures found in the airway (Byrd-Leotis et al. 2021) and the ionic environment of the human upper respiratory tract may favour binding and transmission of viruses with increased positive charge. Perhaps it is not surprising that both OC43 and 229E coronaviruses exhibited increases in spike positive charge after moving from animal hosts (cow, pig and camel) to human hosts (Figure 4). A similar change in MERS-CoV was not observed, however MERS-CoV currently shows only limited human to human transmission with most known transmission chains ending after 2 to 3 human to human transmission events

224 (e.g. (Assiri et al. 2013) (Cotten et al. 2013)). The MERS-CoV might not yet have undergone the
225 same level of selection that OC43, 229E and SARS-CoV-2 have experienced. For all examples of
226 coronaviruses moving to humans, the broad location of the positive changes across the spike
227 protein sequence suggested that positive charge may be promoting several functions including
228 receptor binding, furin cleavage, cell fusion as well as antigenic changes or less specific changes to
229 avoid or promote ionic interactions during transmission.
230 There is likely a limit to the accumulation of positively charged residues in the SARS-CoV-2
231 spike protein. Functional constraints exist, there may also be penalties associated with non-specific
232 binding due to excess positive charge, and there are certainly charge influences on protein folding
233 and higher order protein interactions. Our prediction is that the SARS-CoV-2 protein will reach some
234 upper limit of charge defined by these constraints. Indeed, we observe that the various Omicron
235 lineages have now settled into the range of charge around -1, after more than 6 months of evolution
236 (Figures 2A and 2B) and it is not clear if additional positive charge will be beneficial or even tolerated.
237 The next months of SARS-CoV-2 evolution will be informative on this point.
238 Could these changes in spike charge have occurred by chance and not be a response to
239 selective pressure? Of the 20 standard amino acids (aa), only 2 aa have negatively charged side
240 chains, 2 aa have positive side chains and the remaining 16 aa are neutral at pH 7.4. Assuming equal
241 probability of any aa change, there is an 18/20 chance of a negative aa being substituted by a
242 neutral or positively charged aa and the majority of change opportunities would result in loss of
243 negative charge. However, natural selection is more complex, because the genetic code uses 3
244 adjacent nucleotides to encode an aa, there are multiple encoding possibilities for each aa, the
245 codon redundancy is not identical for each aa and the number of nucleotide changes required to
246 produce any particular aa change can be 1, 2 or 3. This has resulted in an evolved protein stability in
247 the genetic code (Chan et al. 2020) with aa changes that maintain rather than change physical
248 properties (negative, positive, polar, non-polar, aromatic) more likely based on the codon array
249 (Livingstone & Barton 1993) and the nucleotide changes required for an aa change. For example, the
250 probability of a negative aa to negative aa change is 0.333 while the probabilities of change of a
251 negative aa to a non-polar, aromatic, polar or positive aa are 0.051, 0.044, 0.028 and 0.044
252 respectively, with changes away from a negative charged aa nearly 10-fold less likely to occur than
253 conserving the negative charge at that position (Livingstone & Barton 1993). For these reasons, it
254 appears that the accumulation of positive charge on spike protein has not occurred by chance and is
255 likely providing some selective advantage for the virus. It should also be noted that the observed
256 charge changes in exposed virion proteins seem to be limited to spike. Two additional SARS-CoV-2
257 proteins are externally exposed, the E protein (ORF4) and the M protein (ORF5), showing no
258 consistent change in the charge of either of these proteins across the 2 years of the epidemic.
259 Obermeyer et al. documented aa substitutions associated with SARS-CoV-2 fitness
260 (Obermeyer et al. 2022). Consistent with the idea that the increase in positive charges is not by
261 chance, of the top 20 substitutions increasing SARS-CoV-2 fitness, 14 were in Spike and 4 of the 14

262 were changes that increased positive charge while only 1 of 14 introduced a negative charge in spike
263 (Obermeyer et al. 2022).
264 Natural selection could be acting on multiple features of the spike protein. The necessity to
265 avoid host immune responses is likely to be the major selective force acting on the virus. This results
266 in the amino acid changes, which in turn are determined by epitopes. The selection for increased
267 charge in the spike protein is probably occurring in the background, not as a major shift needed to
268 bypass immune responses. However, the increase in charge may improving survival and
269 transmission in humans in subtle ways, and this advantage, when multiplied over the millions of
270 infections can provide some of the growth and infection advantages seen by new SARS-CoV-2
271 variants. It is proposed that the N764K, N856K and N969K substitutions (all increasing spike positive
272 charge) may enhance S1/S2 subunit interactions after proteolytic processing of the spike protein,
273 resulting in reduced S1 shedding and improving transmission (Martin et al. 2022) Increased charge
274 may also alter receptor interactions. In the Omicron (BA.1) spike protein, the Q493R and Q498R
275 substitutions are predicted to allow two additional salt bridges with ACE2 receptor position
276 35Glu and 38Glu (McCallum et al. 2022). Indeed, looking at the timing of charge shifts in each major
277 lineage, the changes to more positive charge accumulate later than the changes that first allow a
278 lineage to emerge and dominate global infections. In this model, the primary spike changes are
279 driven by immune selection and allow a new lineage to bypass existing immune responses. Once a
280 successful new variant emerges, the large number of new infections may promote the accumulation
281 of beneficial positive charge changes. The similar pattern of increased positivity of spike protein in
282 other coronaviruses that have moved between animals and humans (OC43, 229E, Figure 4) suggest
283 that the change in surface protein charge may be a more general phenomenon with coronaviruses
284 and might be a useful parameter to examine when monitoring zoonosis.
285 Data and observations from this study warrant further investigation of the spike protein with
286 regards to coronavirus evolution. Additionally modelling may help add predictions in future
287 emergence of new SARS-CoV-2 variants taking advantage of more positively charged changes in
288 spike protein, as well as the periodic timing of such new variant emergence. This study provides a
289 framework to monitor viral evolution through changes in biochemical properties, which can be easily
290 applied to other viruses important to public and global health.
291
292 **Methods**
293 Full alignments of SARS-CoV-2 genomes were obtained from GISAID (GISAID 2020) with
294 collection dates to 15 June 2022. All spaces in fasta IDs were removed using sed (sed -i -e 's/ /_/g'
295 msa_xxxx.fasta), the alignment was dealigned ("-" characters removed) and genomes were classified
296 using Pangolin (Áine O'Toole et al. 2020) with the most recent database updates (pangolin v4.1.1,
297 pangolin-data v1.11
298 constellations v0.1.10 and scorpio v0.3.17). The spike coding region from each genome (if present
299 and intact (no Ns)) was translated into protein. Features of the protein that could be quantitated from

the spike protein sequence were determined using the ProteinAnalysis functions from BioPython (Cock et al. 2009). These features included charge at pH 7.4, Kyle and Doolittle GRAVY score (Kyte & Doolittle 1982) (a measure of hydrophobicity), an instability index derived from dipeptide content (Guruprasad et al. 1990), the total percent helix, fold or sheet properties of the protein and the total fractions of individual amino acids and fractions of di-amino acids. A matrix of all spike protein features plus collection date, and lineage was prepared and used for analysis. Similar analyses were performed for other coronaviruses such as 229E, OC43 and MERS-CoV by retrieving all complete genomes available from GenBank (15 June 2022). The spike protein was also extracted using the same method as aforementioned. Additional details are provided in the figure legends. The python code used for the analyses is available here: https://github.com/mlcotten13/SARS-CoV-2_spike_charge .

Acknowledgements

We thank all global SARS-CoV-2 sequencing groups for the open sharing of sequence data and to the GISAID platform and team for making these data available. We are grateful to Andrew Rambaut, Áine O'Toole and the Pangolin team for the Pangolin typing tool and resources.

Funding

This work was supported by the UK Medical Research Council (MRC/UK Research and Innovation) and the UK Department for International Development (DFID) under the MRC/DFID Concordat agreement (grant agreement no. MC_PC_20010) and Wellcome Trust, UK FCDO-Wellcome Epidemic Preparedness-Coronavirus (grant agreement no. 220977/Z/20/Z).

References

Adamczyk, Z., Batys, P., & Barbasz, J. (2021). 'SARS-CoV-2 virion physicochemical characteristics pertinent to abiotic substrate attachment', *Current Opinion in Colloid & Interface Science*, 55: 101466. DOI: 10.1016/j.cocis.2021.101466

Áine O'Toole et al. (2020). 'Phylogenetic Assignment of Named Global Outbreak LINeages (PANGOLIN)',.

Assiri, A., McGeer, A., Perl, T. M., Price, C. S., Al Rabeeah, A. A., Cummings, D. A. T., Alabdullatif, Z. N., et al. (2013). 'Hospital Outbreak of Middle East Respiratory Syndrome Coronavirus', *New England Journal of Medicine*, 369/5: 407–16. DOI: 10.1056/NEJMoa1306742

Byrd-Leotis, L., Lasanajak, Y., Bowen, T., Baker, K., Song, X., Suthar, M. S., Cummings, R. D., et al. (2021). 'SARS-CoV-2 and other coronaviruses bind to phosphorylated glycans from the human lung', *Virology*, 562: 142–8. DOI: 10.1016/j.virol.2021.07.012

Cao, Y., Wang, J., Jian, F., Xiao, T., Song, W., Yisimayi, A., Huang, W., et al. (2022). 'Omicron escapes the majority of existing SARS-CoV-2 neutralizing antibodies', *Nature*, 602/7898: 657–63. DOI: 10.1038/s41586-021-04385-3

Chan, K.-F., Koukouravas, S., Yeo, J. Y., Koh, D. W.-S., & Gan, S. K.-E. (2020). 'Probability of change in life: Amino acid changes in single nucleotide substitutions', *Biosystems*, 193–194: 104135. DOI: 10.1016/j.biosystems.2020.104135

Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., et al. (2009). 'Biopython: freely available Python tools for computational molecular biology and bioinformatics', *Bioinformatics*, 25/11: 1422–3. DOI: 10.1093/bioinformatics/btp163

Corman, V. M., Eckerle, I., Memish, Z. A., Liljander, A. M., Dijkman, R., Jonsdottir, H., Juma Ngeiywa, K. J. Z., et al. (2016). 'Link of a ubiquitous human coronavirus to dromedary camels', *Proceedings of the National Academy of Sciences*, 113/35: 9864–9. DOI: 10.1073/pnas.1604472113

Cotten, M., Watson, S. J., Kellam, P., Al-Rabeeah, A. A., Makhdoom, H. Q., Assiri, A., Al-Tawfiq, J. A., et al. (2013). 'Transmission and evolution of the Middle East respiratory syndrome coronavirus in Saudi Arabia: a descriptive genomic study', *Lancet (London, England)*, 382/9909: 1993–2002. DOI: 10.1016/S0140-6736(13)61887-5

Cotten, M., Watson, S. J., Zumla, A. I., Makhdoom, H. Q., Palser, A. L., Ong, S. H., Al Rabeeah, A. A., et al. (2014). 'Spread, circulation, and evolution of the Middle East respiratory syndrome coronavirus', *mBio*, 5/1: e01062-13. DOI: 10.1128/mBio.01062-13

DeGrace, M. M., Ghedin, E., Frieman, M. B., Krammer, F., Grifoni, A., Alisoltani, A., Alter, G., et al. (2022). 'Defining the risk of SARS-CoV-2 variants on immune protection', *Nature*, 605/7911: 640–52. DOI: 10.1038/s41586-022-04690-5

Dejnirattisai, W., Huo, J., Zhou, D., Zahradník, J., Supasa, P., Liu, C., Duyvesteyn, H. M. E., et al. (2022). 'SARS-CoV-2 Omicron-B.1.1.529 leads to widespread escape from neutralizing antibody responses', *Cell*, 185/3: 467-484.e15. DOI: 10.1016/j.cell.2021.12.046

GISAID. (2020). 'The GISAID Initiative',.

Greaney, A. J., Loes, A. N., Crawford, K. H. D., Starr, T. N., Malone, K. D., Chu, H. Y., & Bloom, J. D. (2021). 'Comprehensive mapping of mutations in the SARS-CoV-2 receptor-binding domain that affect recognition by polyclonal human plasma antibodies', *Cell Host & Microbe*, 29/3: 463-476.e6. DOI: 10.1016/j.chom.2021.02.003

Greaney, A. J., Starr, T. N., Barnes, C. O., Weisblum, Y., Schmidt, F., Caskey, M., Gaebler, C., et al. (2021). 'Mapping mutations to the SARS-CoV-2 RBD that escape binding by different classes of antibodies', *Nature Communications*, 12/1: 4196. DOI: 10.1038/s41467-021-24435-8

Greaney, A. J., Starr, T. N., & Bloom, J. D. (2022). 'An antibody-escape estimator for mutations to the SARS-CoV-2 receptor-binding domain', *Virus Evolution*, 8/1: veac021. DOI: 10.1093/ve/veac021

Guruprasad, K., Reddy, B. V. B., & Pandit, M. W. (1990). 'Correlation between stability of a protein and its dipeptide composition: a novel approach for predicting *in vivo* stability of a protein from its primary sequence', *Protein Engineering, Design and Selection*, 4/2: 155–61. DOI: 10.1093/protein/4.2.155

Kim, S. Y., Jin, W., Sood, A., Montgomery, D. W., Grant, O. C., Fuster, M. M., Fu, L., et al. (2020). 'Characterization of heparin and severe acute respiratory syndrome-related coronavirus 2 (SARS-CoV-2) spike glycoprotein binding interactions', *Antiviral Research*, 181: 104873. DOI: 10.1016/j.antiviral.2020.104873

Kyte, J., & Doolittle, R. F. (1982). 'A simple method for displaying the hydropathic character of a protein', *Journal of Molecular Biology*, 157/1: 105–32. DOI: 10.1016/0022-2836(82)90515-0

Liu, Y., Liu, J., Johnson, B. A., Xia, H., Ku, Z., Schindewolf, C., Widen, S. G., et al. (2022). 'Delta spike P681R mutation enhances SARS-CoV-2 fitness over Alpha variant', *Cell Reports*, 39/7: 110829. DOI: 10.1016/j.celrep.2022.110829

Livingstone, C. D., & Barton, G. J. (1993). 'Protein sequence alignments: a strategy for the hierarchical analysis of residue conservation', *Bioinformatics*, 9/6: 745–56. DOI: 10.1093/bioinformatics/9.6.745

Lubinski, B., Frazier, L. E., Phan, M. V. T., Bugembe, D. L., Cunningham, J. L., Tang, T., Daniel, S., et al. (2022). 'Spike Protein Cleavage-Activation in the Context of the SARS-CoV-2 P681R Mutation: an Analysis from Its First Appearance in Lineage A.23.1 Identified in Uganda', *Microbiology Spectrum*, e0151422. DOI: 10.1128/spectrum.01514-22

Martin, D. P., Lytras, S., Lucaci, A. G., Maier, W., Grüning, B., Shank, S. D., Weaver, S., et al. (2022). 'Selection analysis identifies clusters of unusual mutational changes in Omicron lineage BA.1 that likely impact Spike function', (K. Crandall, Ed.)*Molecular Biology and Evolution*, msac061. DOI: 10.1093/molbev/msac061

McCallum, M., Czudnochowski, N., Rosen, L. E., Zepeda, S. K., Bowen, J. E., Walls, A. C., Hauser, K., et al. (2022). 'Structural basis of SARS-CoV-2 Omicron immune evasion and receptor engagement', *Science*, 375/6583: 864–8. DOI: 10.1126/science.abn8652

399 Memish, Z. A., Cotten, M., Meyer, B., Watson, S. J., Alsahafi, A. J., Al Rabeeah, A. A., Corman, V. M.,
400 et al. (2014). 'Human infection with MERS coronavirus after exposure to infected camels,
401 Saudi Arabia, 2013', *Emerging Infectious Diseases*, 20/6: 1012–5. DOI:
402 10.3201/eid2006.140402
403 Nguyen, L., McCord, K. A., Bui, D. T., Bouwman, K. M., Kitova, E. N., Elaish, M., Kumawat, D., et al.
404 (2022). 'Sialic acid-containing glycolipids mediate binding and viral entry of SARS-CoV-2',
405 *Nature Chemical Biology*, 18/1: 81–90. DOI: 10.1038/s41589-021-00924-1
406 Nie, C., Sahoo, A. K., Netz, R. R., Herrmann, A., Ballauff, M., & Haag, R. (2022). 'Charge Matters:
407 Mutations in Omicron Variant Favor Binding to Cells', *ChemBioChem*. DOI:
408 10.1002/cbic.202100681
409 Obermeyer, F., Jankowiak, M., Barkas, N., Schaffner, S. F., Pyle, J. D., Yurkovetskiy, L., Bosso, M.,
410 et al. (2022). 'Analysis of 6.4 million SARS-CoV-2 genomes identifies mutations associated
411 with fitness', *medRxiv: The Preprint Server for Health Sciences*, 2021.09.07.21263228. DOI:
412 10.1101/2021.09.07.21263228
413 Pawłowski, P. (2021). 'SARS-CoV-2 variant Omicron (B.1.1.529) is in a rising trend of mutations
414 increasing the positive electric charge in crucial regions of spike protein S.', *Acta Biochimica*
415 *Polonica*. DOI: 10.18388/abp.2020_6072
416 Pedregosa, F. and Varoquaux, G. and Gramfort, A. and Michel, V., and Thirion, B. and Grisel, O. and
417 Blondel, M. and Prettenhofer, P., and Weiss, R. and Dubourg, V. and Vanderplas, J. and
418 Passos, A. and, & Cournapeau, D. and Brucher, M. and Perrot, M. and Duchesnay, E. (2011).
419 'Scikit-learn: Machine Learning in Python', *Journal of Machine Learning Research*, 12: 2825–
420 30.
421 Peiris, M., & Perlman, S. (2022). 'Unresolved questions in the zoonotic transmission of MERS',
422 *Current Opinion in Virology*, 52: 258–64. DOI: 10.1016/j.coviro.2021.12.013
423 Sabir, J. S. M., Lam, T. T.-Y., Ahmed, M. M. M., Li, L., Shen, Y., E. M. Abo-Aba, S., Qureshi, M. I., et
424 al. (2016). 'Co-circulation of three camel coronavirus species and recombination of MERS-
425 CoVs in Saudi Arabia', *Science*, 351/6268: 81–4. DOI: 10.1126/science.aac8608
426 So, R. T. Y., Chu, D. K. W., Miguel, E., Perera, R. A. P. M., Oladipo, J. O., Fassi-Fihri, O., Aylet, G., et
427 al. (2019). 'Diversity of Dromedary Camel Coronavirus HKU23 in African Camels Revealed
428 Multiple Recombination Events among Closely Related Betacoronaviruses of the Subgenus
429 Embecovirus', *Journal of Virology*, 93/23: e01236-19. DOI: 10.1128/JVI.01236-19
430 Tzou, P. L., Tao, K., Pond, S. L. K., & Shafer, R. W. (2022). 'Coronavirus Resistance Database (CoV-
431 RDB): SARS-CoV-2 susceptibility to monoclonal antibodies, convalescent plasma, and
432 plasma from vaccinated persons', (J. Bhattacharya, Ed.)*PLOS ONE*, 17/3: e0261045. DOI:
433 10.1371/journal.pone.0261045
434 Vijgen, L., Keyaerts, E., Lemey, P., Maes, P., Van Reeth, K., Nauwynck, H., Pensaert, M., et al.
435 (2006). 'Evolutionary history of the closely related group 2 coronaviruses: porcine
436 hemagglutinating encephalomyelitis virus, bovine coronavirus, and human coronavirus
437 OC43', *Journal of Virology*, 80/14: 7270–4. DOI: 10.1128/JVI.02675-05
438 Vijgen, L., Keyaerts, E., Moës, E., Thoelen, I., Wollants, E., Lemey, P., Vandamme, A.-M., et al.
439 (2005). 'Complete genomic sequence of human coronavirus OC43: molecular clock analysis
440 suggests a relatively recent zoonotic coronavirus transmission event', *Journal of Virology*,
441 79/3: 1595–604. DOI: 10.1128/JVI.79.3.1595-1604.2005
442 Zhou, Z., Hui, K. P. Y., So, R. T. Y., Lv, H., Perera, R. A. P. M., Chu, D. K. W., Gelaye, E., et al.
443 (2021). 'Phenotypic and genetic characterization of MERS coronaviruses from Africa to
444 understand their zoonotic potential', *Proceedings of the National Academy of Sciences of the*
445 *United States of America*, 118/25: e2103984118. DOI: 10.1073/pnas.2103984118
446