Unique protein features of SARS-CoV-2 relative to other *Sarbecoviruses*

Matthew Cotten[1,2], David L. Robertson[2], My V.T. Phan[1]

Affiliations

1. MRC/UVRI & LSHTM Uganda Research Unit, Entebbe, Uganda

2. MRC-University of Glasgow Centre for Virus Research, Glasgow, UK

Correspondence: Matthew Cotten (Matthew.Cotten@lshtm.ac.uk)

Keywords: SARS-CoV-2, proteome changes, *Sarbecovirus* evolution, spike protein changes

Abstract

Defining the unique protein features of SARS-CoV-2, the viral agent causing Coronavirus Disease 2019, may guide efforts to control this pathogen. We examined proteins encoded by the *Sarbecoviruses* closest to SARS-CoV-2 using profile Hidden Markov Model similarities to identify features unique to SARS-CoV-2. Consistent with previous reports, a small set of bat and pangolin-derived *Sarbecoviruse*s show the greatest similarity to SARS-CoV-2. The analysis provided a measure of total proteome similarity and showed that a small subset of bat *Sarbecoviruses* are closely related but unlikely to be the direct source of SARS-CoV-2. Spike analysis reveals that the current SARS-CoV-2 variants of concern have sampled only 36% of the possible spikes changes which have occurred historically in *Sarbecovirus* evolution. It is likely that new SARS-CoV-2 variants with changes in these regions are compatible with virus replication and are to be expected in the coming months, unless global viral replication is severely reduced.

Introduction

Since the first report of Coronavirus Disease 2019 (COVID-19) caused by SARS-CoV-2 in December 2019 in Wuhan city, China(1)(2) and the World Health Organisation declaring COVID-19 a global pandemic in March 2020, the disease has continued to affect every part of the world. The SARS-CoV-2 virus belongs to the *Coronaviridae* family of enveloped positive-sense single-stranded RNA viruses, *Betacoronavirus* genus, *Sarbecovirus* subgenus. Other *Sarbecoviruses* include SARS-CoV (the coronavirus causing the SARS outbreak in 2002-2004) and a large number of SARS-like bat viruses that have been identified. The genomes of *Sarbecoviruses* are 30kb in length, encoding >14 open reading frames (ORFs). Among the structural proteins, the spike protein plays a crucial role in the virus cell tropism, host range, cell entrance and infectivity and is considered the main protein target for the host immune response. Other ORFs encode for accessory proteins, many of which modulate host responses to infection. Investigation of the evolutionary history of SARS-CoV-2 show a clear link to *Sarbecoviruses* from bats although no direct animal precursor for SARS-CoV-2 has been identified (3) (4) (5) (6) (7). We sought to identify unique peptide regions of SARS-CoV-2 compared to all available *Sarbecoviruses* to determine the features that might have allowed SARS-CoV-2 to replicate and

transmit efficiently in humans. Comparative analyses of viral proteins would aid in better understanding of virus biology and pathology, providing insights into the origin of the virus and the conditions that led to its zoonosis to humans, efficient spread without the need for adaptation, as well as providing leads for drug and immune targets for effective treatments.

Results and Discussion

Protein domains and profile hidden Markov models. We have explored the genomes across the *Sarbecovirus* subgenus using profile hidden Markov models (pHMMs) based on viral proteins which provide a statistical description of the properties of amino acid sequences of viral proteins. The differences in pHMMs between virus genomes can reveal differences among viral proteins (8) (9). Efficient tools for preparing and comparing pHMMs are available with HMMER-3 (10) and the method is useful for comparing large or difficult to align genomes and for identifying changes in functional regions of the genomes. We have recently used these methods to identify and classify diverse coronaviruses in the *Coronaviridae* family (11) and to explore large and unwieldy genomes such as those from the African Swine Fever Virus (12). These methods were employed here to explore the relationship between SARS-CoV-2 and the other known *Sarbecoviruses* to gain understanding of their evolutionary history.

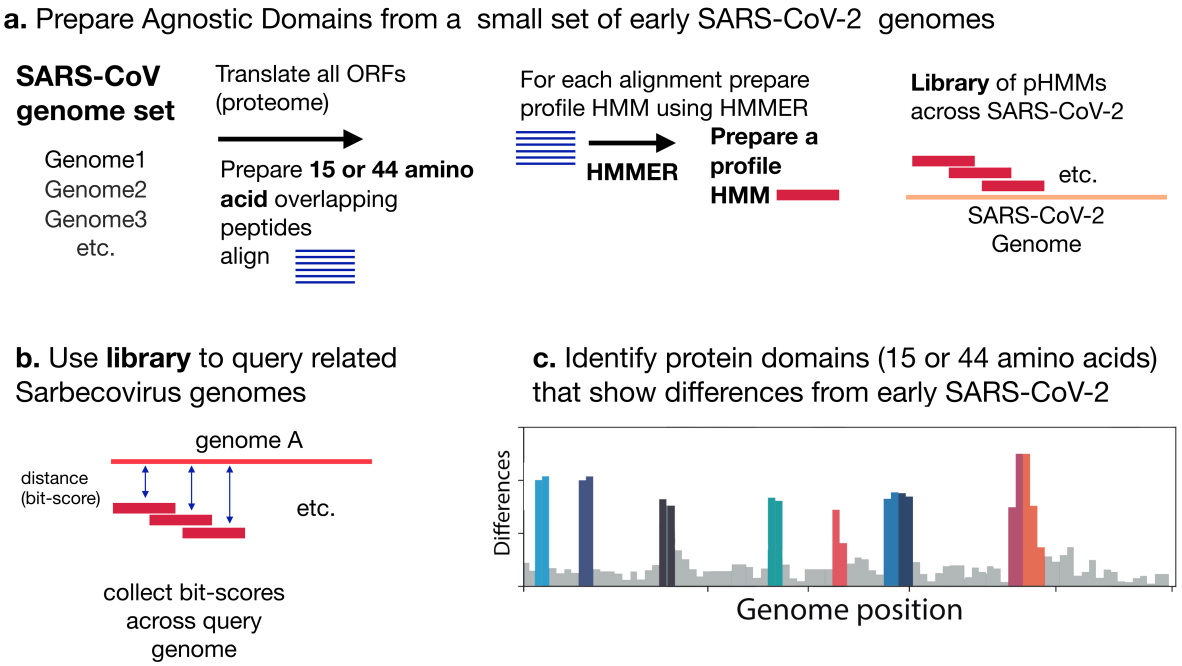


Figure 1. Analysis scheme. (A) Profile Hidden Markov Model (pHMM) domains were generated from a set of 35 early lineage B SARS-CoV-2 genome sequences. All open reading frames were translated and then sliced into either 44 amino acid peptides with a step size of 22 amino acids or 15 amino acid peptides with a step size of 8 amino acid. The peptides were clustered using Uclust (13), aligned with MAFFT (14) and then each alignment was built into a pHMM using HMMER-3 (10). **(B)** The set of pHMMs were used to query *Sarbecovirus* genome sequences, bit scores were collected as a measure of similarity between each pHMM and the query sequence. **(C)** Bit-scores were gathered an analyzed to detect regions that differ between early SARS-CoV-2 genomes and query genomes.

Genome scans using custom pHMM domains. We generated overlapping 44 or 15 amino acid (aa) alignments of all SARS-CoV-2 encoded proteins derived from a set of 35 lineage B SARS-

CoV-2 genomes from early in the pandemic and then prepared pHMMs from each alignment (see Figure 1). The resulting libraries of pHMMs were then used to survey variations across all *Sarbecoviruses* relative to the initial SARS-CoV-2 found in human cases in late 2019/early 2020.

Using this method, ten early SARS-CoV-2 genomes were compared to a representative subset of 29 *Sarbecovirus* genomes obtained from humans, bats, civet cats and pangolins (Figure 2) selected after applying the same analysis to all available *Betacoronavirus* genomes to ensure that we were not missing any unexpectedly close viral genome regions (results not shown). For each query genome, the bit-score from the corresponding pHMM from early B lineage SARS-CoV-2 were accumulated and hierarchical clustering based on the normalized domain bits-scores was performed. The bit-score describes the similarity between the 44aa (or 15aa) query sequence and the related sequence from early SARS-CoV-2. The central region of the *Sarbecovirus* genome is conserved across the genome set with all domains marked as dark or light grey in the Figure 2A clustermap indicating normalized bit-scores close to 1. This is not unexpected as this central genomic region spans nsp5 to nsp16 proteins encoding the viral polymerase, other enzymes and non-surface exposed structural proteins of the virus, which are presumably more constrained and less likely to change than other regions of the virus. In contrast, the domains displayed in yellow, orange and red in Figure 2A indicate more increasingly divergent regions between SARS-CoV-2 and the query *Sarbecovirus* genomes (much lower normalized bit scores). For each domain across the proteome, a measure of the difference between the query and SARS-CoV-2 domain (1-mean bit-score) was plotted to identify regions that differed extensively across the *Sarbecoviruses* genome set (Figure 2B). Further details on the relationship between amino acid changes across a genome or gene set and the metric 1-mean bit-score are found in the Supplemental Figure 1. The higher 1-mean bit-score values ($\geq$0.3) were found in nsp2, nsp3, nsp4, spike and orf8 and orf9N (Figure 2B), indicating that these proteins differ in the set of SARS-CoV-2 and closely related *Sarbecoviruses*. These regions may be dispensable, flexible and can tolerate change and may be products of neutral evolution they may differ because selective pressure has resulted in new functional roles or immune escape.

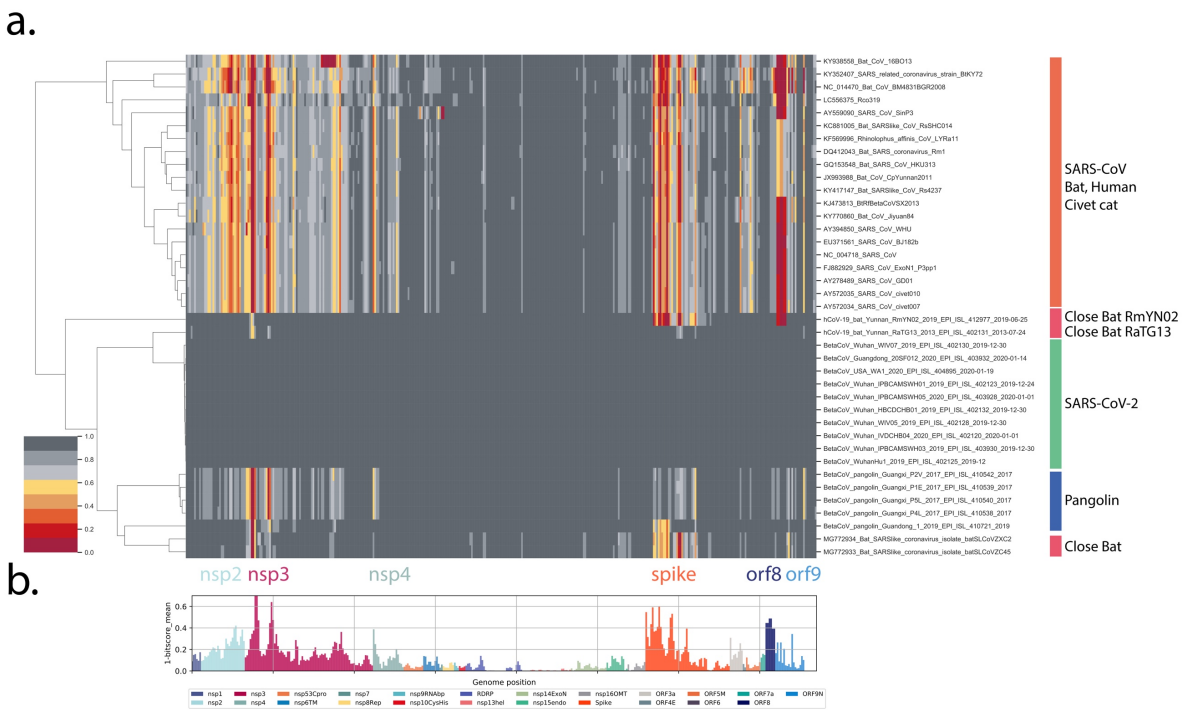


Figure 2. Proteome differences in SARS-CoV-2 vs close Bat, Human and Civet cat *Sarbecoviruses*. All forward open reading frames from the 35 early lineage B SARS-CoV-2 genomes were translated, and processed into 44 aa peptides (with 22 aa overlap), clustered at 0.65 identity using Uclust (11), aligned with MAAFT (12) and converted into pHMMs using HMMER-3 (10). The presence of these domains was sought in a set of *Sarbecovirus* genomes plus the SARS-CoV-2 genomes and genomes were then clustered using hierarchical clustering based on the normalized domain bit-scores (e.g. the similarity of the identified query domain to the reference lineage B SARS-CoV-2 domain). Each row represents a genome, each column represents a domain. Domains are displayed in their order across the SARS-CoV-2 genome, Red = low normalized domain bit-score (lower similarity to lineage B SARS-CoV-2) = distant from SARS-CoV-2, Darkest grey = normalized domain bit-score = 1 = highly similar to lineage B SARS-CoV-2. Groups of coronaviruses were indicated to the right of the figure. **(A)** Domain differences across the *Sarbecovirus* subgenus. **(B)** For each domain the mean bit-score was calculated across the entire set of *Sarbecovirus genomes* and the value 1-mean bit-score was plotted for each domain. Domains are coloured by the proteins from which they were derived with the colour code indicated below the figure.

The role of pangolins as an amplifying intermediate host of SARS-CoV-2 is important to document securely, to guide efforts to prevent or prepare for future zoonotic events. A small number of *Sarbecoviruses* have been identified in samples from trafficked pangolins in China (15) (7) (16), yet there is no direct evidence that pangolins host the virus in their natural environment. It remains possible that the positive pangolins identified in China were infected by viruses encountered after transport to China. Five CoV sequences from pangolins were included in this analysis (Figure 2), including four (EPI_ISL_410538,-39,-40,-42) generated by Lam *et al.* (7) after sequencing the original samples described by Liu *et al.*(15); a 5th genome (Guandong_1_2019_EPI_ISL_410721_2019) deposited by Xiao *et al.* (16) was included despite unclear progeny for this sequence (17). Four additional CoV sequences from pangolins from GISAID were excluded due to excessive gaps in the sequence which precluded the domain analysis. Overall, the pangolin CoVs show many genomic regions of very close

domain homology although the overall similarity across the entire genome is not close enough to support of direct pangolin-human jump as the source of SARS-CoV-2 (see below).

A number of closely related *Sarbecoviruses* have been reported in bats (18) (19). We focused on the four bat *Sarbecoviruses* with closest similarity to SARS-CoV-2 (close bats: RaTG13, RmYn02, MG772933,34). The clustermap and variance analysis (Figure 2A) showed higher similarity across most of the genome (dark grey sectors) with three proteins (nsp3, spike and orf9) displaying high reduced bit-scores from the SARS-CoV-2 sequences (Figure 2A, close bats, yellow/red domains). These limited regions that differ between the 4 closely related bat *Sarbecoviruses* and SARS-CoV-2 indicate virus elements that needed to adjust to human transmission. The spike differences have been explored in detail (refs and see below) however it may be important to consider nsp3 and orf9 changes in the complete analysis.

Spike changes with 15 amino acid domains. To provide more detailed resolution, 15 amino acid pHMMs across the early lineage B spike protein were prepared and used to examine changes across the set of close *Sarbecoviruses* and SARS-CoV-2 (Figure 3). The regions of change observed in the spike protein show an expanded version of the total proteome pattern observed in Figure 2 plus additional details. Adding the current knowledge of SARS-CoV-2 spike evolution reveals some important patterns. The S1 domain of spike (the amino-terminal half of the protein) tolerates at large amount of change with most of the low score domains (red) concentrated here (Figure 3A), consistent with the surface exposure of the S1 region, and driven by pressure to avoid immune responses which in turn require the virus to maintain a malleable structure in the exposed S1. Except for the Guangdong_1 pangolin and RatG13 spikes, the central ACE2 receptor binding region is very different between the close *Sarbecoviruses* and SARS-CoV-2. The furin cleavage site at the junction between the S1 and S2 domains is also a region showing a lot of malleability in the *Sarbecovirus* spikes (Figure 3A) and is completely unique to SARS-CoV-2. This has been discussed in detail (20) and is also a site of frequent change in the current SARS-CoV-2 Variant of Concern (VOC) spike sequence with Q677, P681 and T717 flanking the furin site showing changes (Figure 3B).

Thirdly, the amino acid changes that have accumulated in the VOC spike protein are marked in color (Figure 3B) and cluster in regions with high variation (1-mean bit-score > 0.3) suggesting that *Sarbecoviruses* have made changes in these regions in previous evolutionary periods and are continuing to change in SARS-CoV-2 evolution. If all domains with 1-mean bit-score of 0.3 and above tolerate change in the current SARS-CoV-2 spike protein, this suggests that SARS-CoV-2 has many additional changes available for future immune evasion. Important regions that show high levels of historical change are RBD, the furin cleavage site and flanking regions and the NTD.

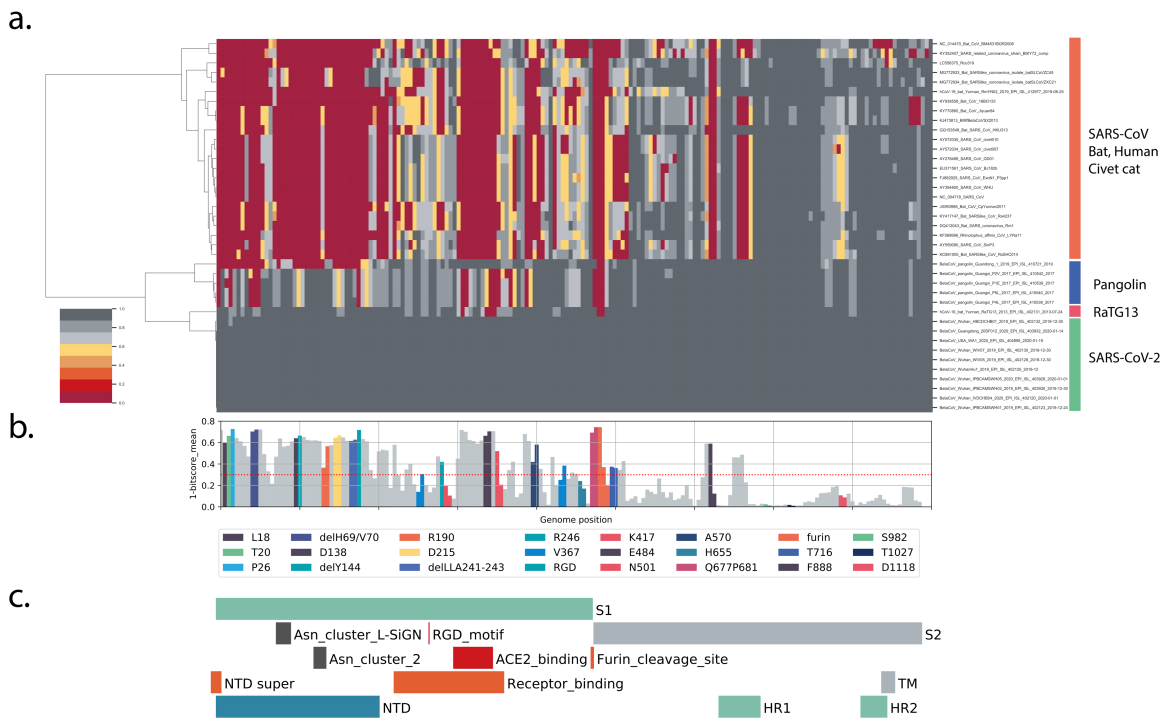


Figure 3. Spike differences in SARS-CoV-2 vs close Bat, Human and Civet cat *Sarbecoviruses*. All forward spike open reading frames from the 35 early lineage B SARS-CoV-2 genomes were translated, and processed into 15 aa peptides (with 8 aa overlap) and processed into an pHMM library as described in Figure 2. **(A)** shows a hierarchical clustering of 15 amino acid domain bit-scores. **(B)** shows the 1-mean of each domain bit-scores across the genome set, domain values, individual domains that span known amino acid changes in the 5 VOC are colored (see key below panel B). **(C)** The locations of important spike protein features are indicated. NTD: N-terminal domain, RBD: receptor-binding domain, S1: spike 1, S1: Spike 2, TM: transmembrane domain, HR1: helical repeat 1, HR2: helical repeat 2, NTD super: N-terminal domain supersite.

Global proteome similarities. A measure of the total protein distance between the SARS-CoV-2 and any query *Sarbecovirus* can be obtained by summing the normalized bit-scores (SNBS) across the entire query proteome. The encoded proteomes from more distant viruses will show lower SNBS values. We examined SNBSs grouped by virus host for both the 44 amino acid total genome analysis as well as the 15 amino acid spike gene analysis. Overall, the patterns showed that the similarity to SARS-CoV-2 follows the order Pangolin > Bat > human=civet cat (Figure 4) although the small set of close bat *Sarbecoviruses* show some exceptions. Of special interest, the bat coronavirus genome RaTG13 (GenBank MN996532.1) was identified as closely related to the SARS-CoV-2 lineage {Citation} and supports a bat coronavirus being the zoonotic source of the epidemic, although the genetic distance is too far for RaTG13 itself to be a direct source of the pandemic SARS-CoV-2 virus (21)(3). Another close bat coronavirus RmYN02, shows some regions of close identity to SARS-CoV-2 (6), yet overall it is more distant from SARS-CoV-2 than the strain RaTG13 (Figure 4A) due its possible recombinant nature. A single pangolin derived SARS-CoV-2 (Guangdong_1) showed an SNBS value that was also elevated but not as high as the RaTG13 (Figure 4a), the 15 aa spike analysis showed

similar patterns except that only the RaTG13 spike displayed the high similarity to SARS-CoV-2 (Figure 4b).

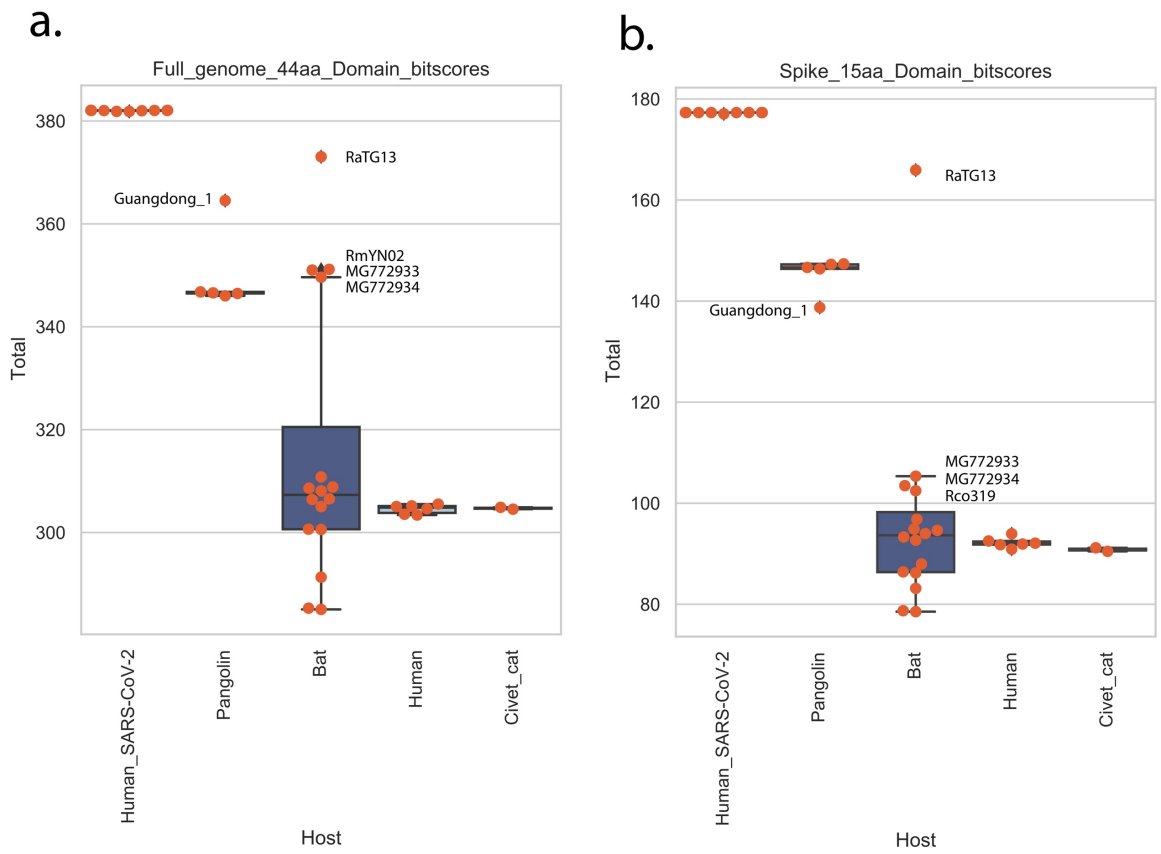


Figure 4. Total domain distances between virus groups. Normalized bit-score sums (NBSS) grouped into SARS-CoV-2 and *Sarbecoviruses* from pangolin, bat, human and civet cat NBSS for all domains for each genome were summed. The boxplot shows individual values marked in orange, median values indicated by horizontal black lines, 1st interquartile ranges marked with a box. The identities of several high scoring bat and pangolin genomes are indicated. **(A)** NBSS for 44 aa domains across the entire coronavirus genome. **(B)** NBSS for 15 aa domains across the spike protein.

Conclusions

What is special about SARS-CoV-2? Spike changes in SARS-CoV-2 compared to a large set of known *Sarbecovirus* indicate that the immediate zoonotic source of SARS-CoV-2 is yet to be identified and reinforces the unique nature of the SARS-CoV-2 genome. The more global analysis of spike regions in SARS-CoV-2 genomes (Figure 3) revealed the changes that have occurred across the *Sarbecovirues*. Combined with the current VOC spike changes, the patterns suggest that SARS-CoV-2 has a great deal of evolutionary possibilities to avoid immune pressure. This suggests that genomic variant surveillance should continue and vaccine producers should be prepared to accommodate such spike changes in the next generation of vaccine updates. In addition to the spike protein, additional regions of high variance were observed in the nsp3 across all *Sarbecoviruses* (Figure 2) in close bat and pangolins (Figure 3). The high variance regions flanked and partially overlapped the Macro domain, which is frequently associated with ADP-deribosylase activity (22),(23). Variance observed in the ORF8

changes across the set was due to frequent deletion of this ORF, suggesting that the encoded protein may be dispensable for human infection. Similar loss of ORF8 was observed with the original SARS-CoV (24) (25) and has been observed in several SARS-CoV-2 lineages as the virus adapted to humans (26) (27) (28). The ORF9 (N protein) variance observed across *Sarbecoviruses* and the changes in this protein in VOC strains suggest an additional region that may be adapting to human replication. The regions of variance identified here may indicate either functional changes in SARS-CoV-2 proteins or amino acid positions that can be changed without impairing the necessary functions of the protein. The relatively high mutation rate of SARS-CoV-2 combined with the unprecedented number of SARS-CoV-2 infections in the world is resulting in massive viral adaptation. Additional experiments are required to distinguish true functional changes from neutral evolution.

Finally, the detailed spike analysis of Figure 3 reveals 82 domains of 15-aa that have shown high variation (1-mean bit-score>= 0.3) in the *Sarbecoviruses* while 29 of these domains show changes in VOC relative to early lineage B SARS-CoV-2. In broad terms, the SARS-CoV-2 evolution observed in the current VOC has sampled only 36% (29/82) of the possible spikes changes which have occurred historically in *Sarbecovirus* evolution. It is highly likely that a large number of new SARS-CoV-2 variants with changes in these regions are possible, compatible with virus replication and expected in the coming months, unless global viral replication is severely reduced.

Acknowledgements

We thank all global SARS-CoV-2 sequencing groups for their open and rapid sharing of sequence data and GISAID for providing an effective platform for making these data available. The study is funded by the Wellcome, DFID - Wellcome Epidemic Preparedness – Coronavirus (AFRICO19, grant agreement number 220977/Z/20/Z) awarded to MC. DLR and MC receive funding from the MRC (MC_UU_1201412).

References

1. Li,Q., Guan,X., Wu,P., Wang,X., Zhou,L., Tong,Y., Ren,R., Leung,K.S.M., Lau,E.H.Y., Wong,J.Y., *et al.* (2020) Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus–Infected Pneumonia. *N. Engl. J. Med.*, 10.1056/NEJMoa2001316.

2. Yang,X., Yu,Y., Xu,J., Shu,H., Xia,J., Liu,H., Wu,Y., Zhang,L., Yu,Z., Fang,M., *et al.* (2020) Clinical course and outcomes of critically ill patients with SARS-CoV-2 pneumonia in Wuhan, China: a single-centered, retrospective, observational study. *Lancet Respir. Med.*, 10.1016/S2213-2600(20)30079-5.

3. Andersen,K.G., Rambaut,A., Lipkin,W.I., Holmes,E.C. and Garry,R.F. (2020) The proximal origin of SARS-CoV-2. *Nat. Med.*, 10.1038/s41591-020-0820-9.

4. Boni,M.F., Lemey,P., Jiang,X., Lam,T.T.-Y., Perry,B.W., Castoe,T.A., Rambaut,A. and Robertson,D.L. (2020) Evolutionary origins of the SARS-CoV-2 sarbecovirus lineage responsible for the COVID-19 pandemic. *Nat. Microbiol.*, **5**, 1408–1417.

5. Zhang,T., Wu,Q. and Zhang,Z. (2020) Probable Pangolin Origin of SARS-CoV-2 Associated with the COVID-19 Outbreak. *Curr. Biol.*, 10.1016/j.cub.2020.03.022.

6. Zhou,H., Chen,X., Hu,T., Li,J., Song,H., Liu,Y., Wang,P., Liu,D., Yang,J., Holmes,E.C., *et al.* (2020) A novel bat coronavirus reveals natural insertions at the S1/S2 cleavage site of the Spike protein and a possible recombinant origin of HCoV-19 Microbiology.

7. Lam,T.T.-Y., Jia,N., Zhang,Y.-W., Shum,M.H.-H., Jiang,J.-F., Zhu,H.-C., Tong,Y.-G., Shi,Y.-X., Ni,X.-B., Liao,Y.-S., *et al.* (2020) Identifying SARS-CoV-2-related coronaviruses in Malayan pangolins. *Nature*, **583**, 282–285.

8. Eddy,S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.

9. Eddy,S.R. (1996) Hidden Markov models. *Curr. Opin. Struct. Biol.*, **6**, 361–365.

10. Eddy,S.R. (2011) Accelerated Profile HMM Searches. *PLOS Comput. Biol.*, **7**, e1002195.

11. Phan,M.V.T., Ngo Tri,T., Hong Anh,P., Baker,S., Kellam,P. and Cotten,M. (2018) Identification and characterization of Coronaviridae genomes from Vietnamese bats and rats based on conserved protein domains. *Virus Evol.*, **4**.

12. Masembe,C., Phan,M.V.T., Robertson,D.L. and Cotten,M. (2020) Increased resolution of African Swine Fever Virus genome patterns based on profile HMM protein domains Genomics.

13. Edgar,R.C. (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, **26**, 2460–2461.

14. Katoh,K. and Standley,D.M. (2013) MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Mol. Biol. Evol.*, **30**, 772–780.

15. Liu,P., Chen,W. and Chen,J.-P. (2019) Viral Metagenomics Revealed Sendai Virus and Coronavirus Infection of Malayan Pangolins (Manis javanica). *Viruses*, **11**, 979.

16. Xiao,K., Zhai,J., Feng,Y., Zhou,N., Zhang,X., Zou,J.-J., Li,N., Guo,Y., Li,X., Shen,X., *et al.* (2020) Isolation of SARS-CoV-2-related coronavirus from Malayan pangolins. *Nature*, **583**, 286–289.

17. Chan,Y.A. and Zhan,S.H. (2020) Single source of pangolin CoVs with a near identical Spike RBD to SARS-CoV-2 Genomics.

18. Tang,X.C., Zhang,J.X., Zhang,S.Y., Wang,P., Fan,X.H., Li,L.F., Li,G., Dong,B.Q., Liu,W., Cheung,C.L., *et al.* (2006) Prevalence and Genetic Diversity of Coronaviruses in Bats from China. *J. Virol.*, **80**, 7481–7490.

19. Li,W. (2005) Bats Are Natural Reservoirs of SARS-Like Coronaviruses. *Science*, **310**, 676–679.

20. Hoffmann,M., Kleine-Weber,H. and Pöhlmann,S. (2020) A Multibasic Cleavage Site in the Spike Protein of SARS-CoV-2 Is Essential for Infection of Human Lung Cells. *Mol. Cell*, **78**, 779-784.e5.

21. Andrew Rambaut (2020) Divergence of nCoV-2019 to closest non-human relative. *Virological.org*.

22. Frick,D.N., Virdi,R.S., Vuksanovic,N., Dahal,N. and Silvaggi,N.R. (2020) Molecular Basis for ADP-Ribose Binding to the Mac1 Domain of SARS-CoV-2 nsp3. *Biochemistry*, **59**, 2608–2615.

23. Lei,J., Kusov,Y. and Hilgenfeld,R. (2018) Nsp3 of coronaviruses: Structures and functions of a large multi-domain protein. *Antiviral Res.*, **149**, 58–74.

24. Chiu,R.W.K., Chim,S.S.C., Tong,Y., Fung,K.S.C., Chan,P.K.S., Zhao,G. and Lo,Y.M.D. (2005) Tracing SARS-Coronavirus Variant with Large Genomic Deletion. *Emerg. Infect. Dis.*, **11**, 168–170.

25. Tang,J.W., Cheung,J.L.K., Chu,I.M.T., Sung,J.J.Y., Peiris,M. and Chan,P.K.S. (2006) The Large 386-nt Deletion in SARS-Associated Coronavirus: Evidence for Quasispecies? *J. Infect. Dis.*, **194**, 808–813.

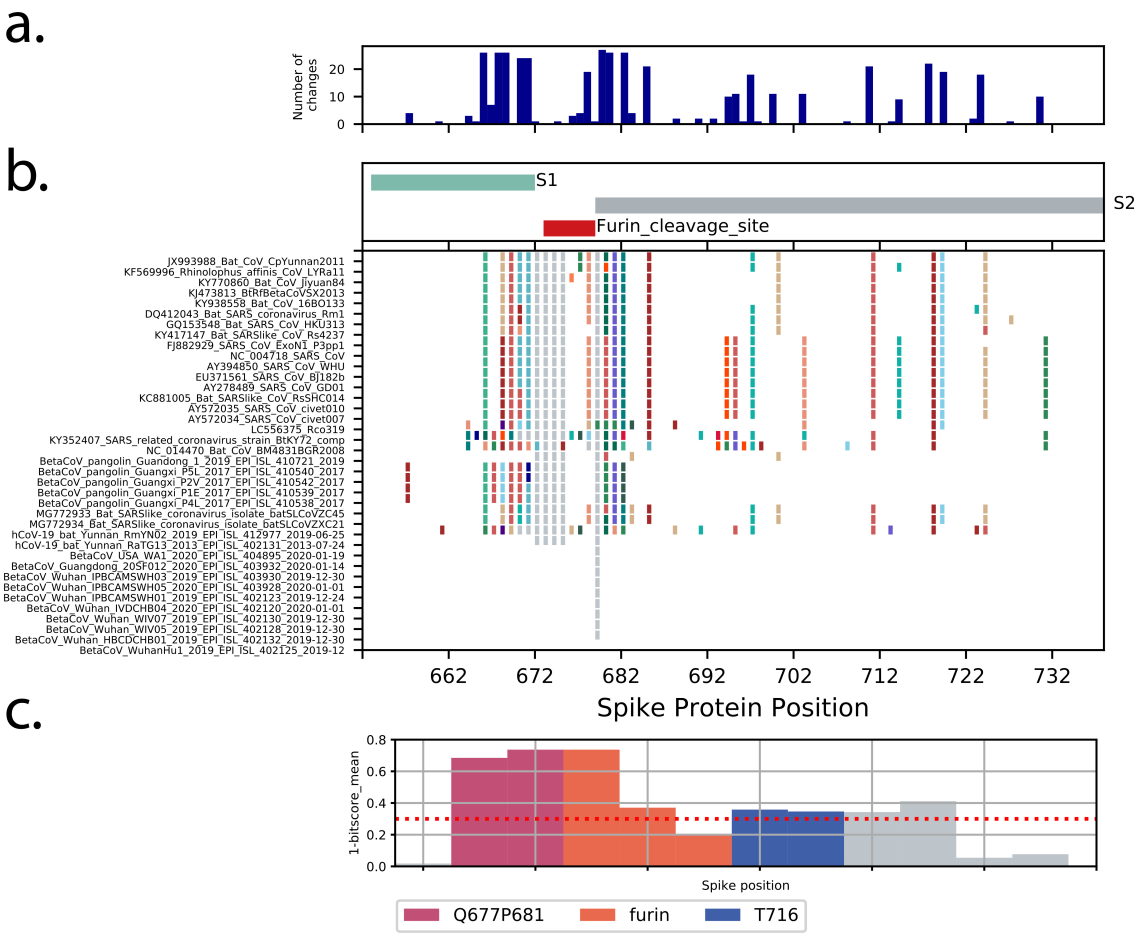
26. Su,Y.C.F., Anderson,D.E., Young,B.E., Linster,M., Zhu,F., Jayakumar,J., Zhuang,Y., Kalimuddin,S., Low,J.G.H., Tan,C.W., *et al.* (2020) Discovery and Genomic Characterization of a 382-Nucleotide Deletion in ORF7b and ORF8 during the Early Evolution of SARS-CoV-2. *mBio*, **11**, e01610-20, /mbio/11/4/mBio.01610-20.atom.

27. Gong,Y.-N., Tsao,K.-C., Hsiao,M.-J., Huang,C.-G., Huang,P.-N., Huang,P.-W., Lee,K.-M., Liu,Y.-C., Yang,S.-L., Kuo,R.-L., *et al.* (2020) SARS-CoV-2 genomic surveillance in Taiwan revealed novel ORF8-deletion mutant and clade possibly associated with infections in Middle East. *Emerg. Microbes Infect.*, **9**, 1457–1466.

28. Young,B.E., Fong,S.-W., Chan,Y.-H., Mak,T.-M., Ang,L.W., Anderson,D.E., Lee,C.Y.-P., Amrun,S.N., Lee,B., Goh,Y.S., *et al.* (2020) Effects of a major deletion in the SARS-CoV-2 genome on the severity of infection and the inflammatory response: an observational cohort study. *Lancet Lond. Engl.*, **396**, 603–611.

Supplementary Figure 1 demonstrates the relationship between the 1-mean bit-score and amino acid changes across a sequence alignment. A set of twelve, 15 aa pHMM domains across the central region of the spike protein were used to demonstrate the relationship between the total amino acid changes and the resulting pHMM 1-mean bit-scores. An alignment of 38 spike proteins from 10 early SARS-CoV-2 genomes, 4 close bat coronavirus *Sarbecoviruses* plus 24 additional close *Sarbecoviruses* was prepared and trimmed to an 87 amino acid region spanning conserved and variable sequences across the furin cleavage site. The bit-scores(10) for the twelve, 15 aa pHMMs across each query sequence were gathered, and the mean value for each domain across the set was calculated and converted to a 1-mean bit-score value for ease of visualization (higher value = greater difference from SARS-CoV-2).

The amino acid changes across the alignment are shown in Supplementary Figure 1 panel B. with each colored bar indicating an amino acid change (or gap, indicated in grey) from the spike of reference genome NC-045512, The 3' end of the spike S1, the furin cleavage site and the 5' start of the spike S2 region are indicated in Panel B. Peaks of variability are seen flanking and within the furin site and in the S2 region (Supplementary Figure 1 panel A). The 1-mean bit-score values for the 12 domains are plotted in Panel C. and show a related but more softened pattern of variation across the region. The advantage the pHMM method and the 1-mean bit-score metrics is that the score is weighted to reflect the type of amino acid change or insertion/deletions and can be determined quickly in large genome sets using the HMMER tools(10). For our comparative analysis, the cutoff of 0.3 for the 1-mean bit-score was set to reflect about 25 amino changes in the total aa positions covered by each domain query (15 aa domain across 38 sequences = 570 total position differences possible).



Supplemental Figure 1. Demonstrating relationship between amino acid changes and 1-mean bit score. (B) The spike proteins encoded by the set of *Sarbecovirus* genomes examined in Figure 2 and 3 were aligned and the amino acid changes from the early SARS-CoV-2 genomes (= NC-045512 sequence) were depicted as colored lines across each sequence. **(A)** The total number of proteins showing a change from the early SARS-CoV-2 genomes (NC-045512) were plotted for each position of the alignment. **(C)** The set of 12, 15 aa pHMMs spanning the central region of the spike (a subset of the pHMMs used for Figure 3) were used to examine the *Sarbecovirus* spike sequences. For each pHMM, the mean bit-score for the entire sequence sequencer set was calculate and the value 1-mean bit-score is plotted for each of the 12 pHMMs. Domains containing known amino acid changes in VOC (Q677, P681, T716) and the domains spanning the furin cleavage site are indicated by color (with the legend found below panel C).