Genomic determinants of pathogenicity in SARS-CoV-2 and other human coronaviruses

Ayal B. Gussow[1,*], Noam Auslander[1,*,#], Yuri I. Wolf[1], Eugene V. Koonin[1,#]

[1]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA

* These authors contributed equally
[#]For correspondence: koonin@ncbi.nlm.nih.gov; noam.auslander@nih.gov

Abstract
SARS-CoV-2 poses an immediate, urgent and major threat to public health across the globe. Here we report an in-depth molecular analysis to reconstruct the evolutionary origins of the enhanced pathogenicity of SARS-CoV-2 and other coronaviruses that are severe human pathogens. Using integrated comparative genomics and machine learning techniques, we identify key genomic features that differentiate SARS-CoV-2 and the viruses behind the two previous coronavirus outbreaks, SARS-CoV and MERS-CoV, from less pathogenic coronaviruses. The identified features could be crucial elements of coronavirus pathogenicity and possible targets for diagnostics, prognostication and interventions.

Main

The emergence of novel SARS-coronavirus-2 (SARS-CoV-2), which causes the respiratory disease COVID-19, triggered a global pandemic that has led to an unprecedented worldwide public health emergency[1]. Since it was first reported in December 2019 and as of April 4, 2020, SARS-CoV-2 has infected over a million individuals worldwide, and has led to an estimated 55,000 deaths, with its associated morbidity and mortality rates continuously rising[2]. SARS-CoV-2 is the seventh member of the *Coronaviridae* family known to infect humans[3]. SARS-CoV and MERS-CoV, two other members of this family, are the causative agents of recent outbreaks, accountable for the severe acute respiratory syndrome (SARS, 2002-2003) and Middle East respiratory syndrome (MERS, began in 2012) outbreaks[3,4], and are associated with relatively high case fatality rates (CFR, 9% and 36%, respectively). The novel SARS-CoV-2 can also cause severe disease and is appreciably more infectious than SARS-CoV or MERS-CoV, but with a lower associated CFR[4]. By contrast, the other coronaviruses infecting humans, HCoV-HKU1, HCoV-NL63, HCoV-OC43, and HCoV-229E, are endemic and cause mild symptoms, accounting for 15-29% of common colds[3]. The three coronaviruses that can cause severe diseases (hereafter high-CFR CoV) originated in a zoonotic transmission from an animal host to humans. SARS-CoV and MERS-CoV have bat reservoirs, and were transmitted to humans through an intermediate host (likely civets and camels, respectively)[4]. Similarly, the closest known relative to SARS-CoV-2 is a bat coronavirus (Fig. 1a), but the specific route of transmission from bats to humans remains unclear. These repeated, independent zoonotic transmissions and the high associated pathogenicity call for an in-depth investigation of the genomic features that contribute to coronaviruses pathogenicity and transmission, to better understand the molecular mechanisms of the high-CFR CoV pathogenicity and thus to be better prepared for any future coronavirus outbreaks.

In this work, we developed an approach combining advanced machine learning methods with well-established genome comparison techniques, to identify the potential genomic determinants of pathogenicity of the high-CFR CoV strains (Fig. 1b). Coronaviruses have positive-sense RNA genomes consisting of 6 conserved proteins, along with an additional set of strain-specific accessory proteins[5]. The conserved proteins are the polyproteins pp1a and pp1ab, spike glycoprotein (S), envelope (E), membrane glycoprotein (M) and nucleocapsid phosphoprotein (N, Fig. 1c). To detect potential genomic determinants of pathogenicity, we aligned the full genomes of all human coronaviruses (Supplementary File 1) and used support vector machines (SVM) to detect high-confidence genomic features that are predictive of the high CFR (see Methods for details). In total, our method identified 11 regions of nucleotide alignments that were reliably predictive of the high CFR of coronaviruses (Fig. 1c, Supplementary Table 1). Two proteins were significantly enriched with these predictive regions, the nucleocapsid protein and the spike glycoprotein (p-values: 4e-16 and 0.036, respectively, Fig 1d). Only four of the diagnostic regions detected in the nucleotide alignment corresponded to observable differences in the protein alignments as well, with three located in the nucleocapsid protein and one in the spike protein (Fig. 1e).

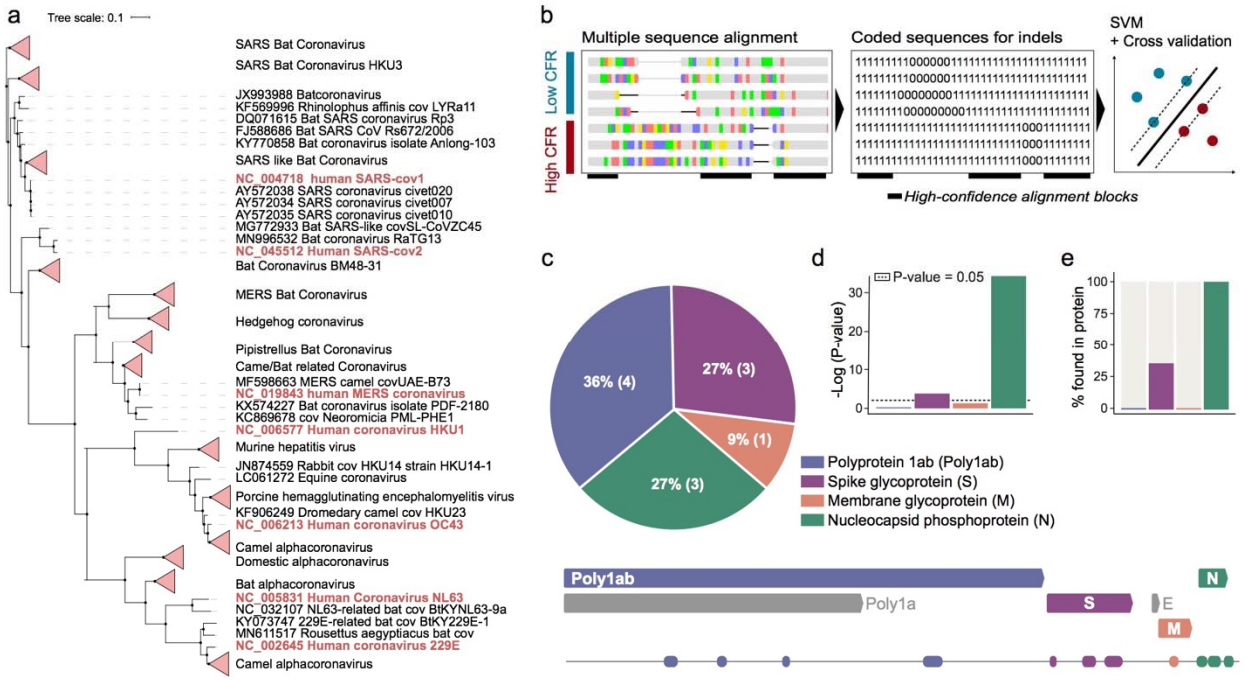

Figure 1. Searching coronavirus genomes for determinants of pathogenicity. **(a)** Phylogenetic tree of coronavirus species, based on the alignment of complete nucleotide sequences of virus genomes. **(b)** A schematic illustration of the pipeline applied for detection of genomic regions predictive of high-CFR strains. **(c)** Top panel: Pie chart showing the percentage of identified genomic determinants in each protein. Bottom panel: Map of SARS-CoV-2 genome with detected regions. **(d)** Bar plot showing the significance of the distribution of detected regions across each protein. **(e)** Percentage of detected predictive regions in each protein.

Exploring the regions identified within the nucleocapsid that predict the CFR feature, we found that these deletions and insertions result in the emergence of motifs that determine nuclear localization[6], which are found specifically in high-CFR CoV (Supplementary Figure 1a). The deletions, insertions and substitutions in the N proteins of the high-CFR CoV map to two monopartite nuclear localization signals (NLS), one bipartite NLS and a nuclear export signal (NES, Fig. 2a). In the course of the evolution of coronaviruses, these nuclear localization and export signals grow markedly stronger in the clades that include the high-CFR viruses and their relatives from animals (primarily, bats), as demonstrated by the increasing positive charge of the amino acids comprising the NLS , a known marker of NLS strength[7] (Fig. 2a). In the clade containing SARS-CoV and SARS-CoV-2, the accumulation of positive charges was observed in the monopartite NLS, the bipartite NLS and the NES, whereas in the clade including MERS-CoV, positive charges accumulated primarily in the first of the two monopartite NLS (Fig. 2a). In all cases, the strengthening of these signals is part of a clear gradual and significant trend towards the occurrence of these signals that accompanied viral evolution concomitantly with the emergence of more pathogenic strains (empirical p-value < 0.001, Fig 2a). The charge of the complete nucleocapsid protein gradually evolves towards greater positive values due specifically to the formation of the NLS, as demonstrated by sequence permutation analysis (Fig. 2b, see Methods), which implies a key role for these motifs in the function of the nucleocapsid including likely contribution to virus pathogenicity. The accumulation of positive charges resulting in strengthening of the NLS, which correlates with the

growing CFR of coronaviruses, implies that the localization pattern of the nucleocapsid proteins of high-CFR strains differs from that of the low-CFR strains and might contribute the increased pathogenicity of the high-CFR strains. Localization of the nucleocapsid protein to the nuclei, and specifically, to the nucleoli, has been previously reported in coronaviruses[8] and has been associated with increased pathogenicity in porcine coronavirus model[9–11]. The presence of both NLS and NES raises an uncertainty as to the precise effect of these motifs on the nucleocapsid protein localization, and the reports are indeed somewhat contradictory[6,9,12]. Nevertheless, the striking extent of the changes in the NLS of the high-CFR strains suggests that localization of the nucleocapsid protein is an important determinant of coronavirus pathogenicity.

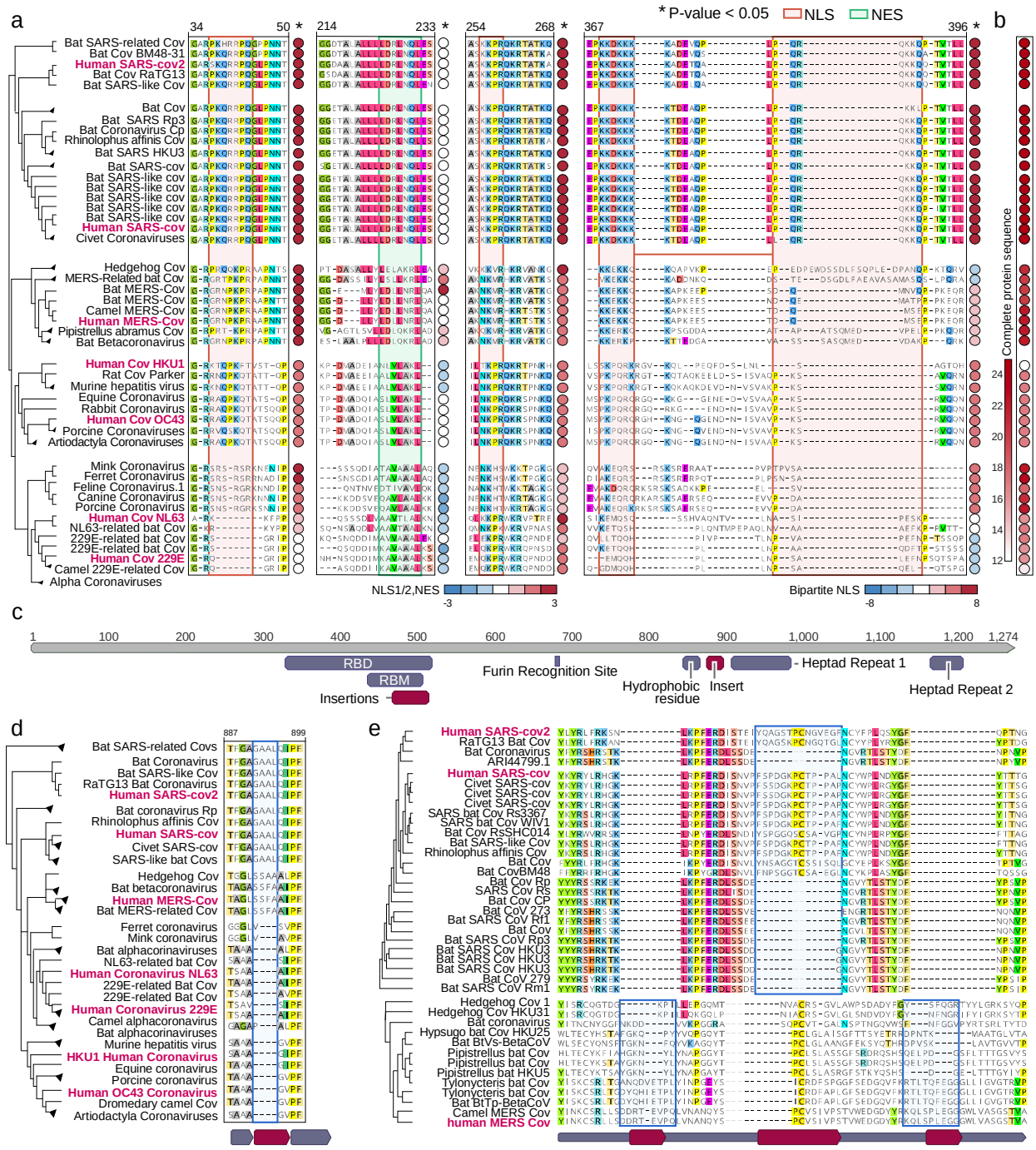

Figure 2. Putative determinants of coronavirus pathogenicity in the nucleoprotein and the spike protein. **(a)** Phylogenetic tree and protein alignment of the nucleocapsid protein across coronavirus species. NLS and NES are outlined in red and green, respectively. The circle next to each signal sequence denotes peptide charge, with red denoting higher charge and blue denoting a lower charge. **(b)** The overall charge of each full protein sequence. **(c)** Map of SARS-CoV-2 spike protein with relevant protein regions (blue) and regions detected by the analyses in this manuscript (red). Relevant regions include the receptor-binding domain (RBD), the receptor-binding motif (RBM), the furin recognition site, a hydrophobic residue preceding the first heptad repeat, and both heptad repeats. The two regions detected by the analysis are the insertions in the RBD found in pathogenic strains before the zoonotic

transmission to human, and the insertion in pathogenic strains preceding the heptad repeat. **(d)** Phylogenetic tree and protein alignment of the spike protein insertion preceding the first heptad repeat. **(e)** Phylogenetic tree and protein alignment of the spike protein insertions in the RBD of high CFR coronaviruses.

We next investigated the diagnostic feature identified within the spike glycoprotein. The SARS-CoV-2 spike protein binds ACE2, the host cell receptor that facilitates SARS-CoV-2 invasion[13], with a 10-20 fold greater affinity compared to SARS-CoV, and contains a furin cleavage site that could further enhance infectivity[4]. The spike protein consists of multiple domains[13] (Fig 2c), including two heptad repeat regions that are crucial to infection[14]. The heptad repeats are important for membrane fusion and virus entry[15], and are typically immediately adjacent to hydrophobic residues[16]. The spike protein fusion peptide is apparently located upstream of the first heptad repeat[17] (HR1), with a long connecting region between the fusion peptide and HR1 that adopts an α-helical structure. Our analysis revealed a 4 amino acid insertion in the connecting region in all high-CFR viruses but not in the low-CFR ones, with the MERS and SARS clades apparently acquiring this insertion independently as supported by the unrelated insert sequences (Fig 2c, Supplementary Figure 1b). The insertion increases the length and flexibility of the connecting region as confirmed by the examination of the spike glycoprotein structure of the SARS-CoV. The strict correlation of this insertion with CoV pathogenicity implies a direct link but how, precisely, does the insert contribute to pathogenicity, remains to be studied experimentally.

Finally, we sought to identify genomic features that might be associated with the repeated jumping of coronaviruses across the species boundaries to human, specifically, in the high-CFR strains. To this end, we aligned the genomes of all coronaviruses from different hosts (Supplementary File 2) and selected, for each human-infecting strain of the high-CFR CoV, the closest non-human infecting relatives (see Methods for details). Within each such set of human high-CFR CoV and their ancestors from animals, we searched for genomic insertions or deletions that occurred in the most proximal strains before the zoonotic jump to human. This analysis identified independent insertions in each of the three groups of viruses all of which were located within the spike glycoprotein, specifically, in the receptor binding domain, within the subdomain that binds ACE2[13,18] (the receptor binding motif, RBM) in the cases of SARS-CoV and SARS-CoV-2, and DPP4 in the case of MERS-CoV[19]. The insertions appear to have occurred independently of each other, notwithstanding the different binding targets, across all three high-CFR virus groups (Fig 2d). Notably, although the insertions occurred in slightly different locations, all the inserted segments contained a proline-cysteine (PC) amino acid doublet, which could result either from convergent evolution, indicating strong selection for these particular amino acids, or independent capture of the inserts related sources (the inserts were too short and dissimilar except for the signature amino acid doublet, to disambiguate between these possibilities). The independent insertions in the region of the spike protein that interacts with the receptor are highly likely to contribute to or even enable the zoonotic transmission to humans of the high-CFR CoV strains, and might also contribute to pathogenicity.

SARS-CoV-2 has led to the most devastating pandemic since the 1918 Spanish flu, prompting an urgent need to elucidate the evolutionary history and genomic features that led to the increased pathogenicity and rampant spread of this virus as well as those coronaviruses that caused previous deadly outbreaks. A better understanding of viral pathogenicity and zoonotic transmission is a crucial step for prediction and prevention of future outbreaks. Here, using an integrated approach of machine-learning and comparative genomics, we found three previously undetected likely determinants of pathogenicity and zoonotic transmission. The enhancement of the NLS in the high-CFR CoV nucleocapsids imply an important role of the subcellular localization of the nucleocapsid protein in the CoV pathogenicity.

Strikingly, the insertions in the spike protein appear to have been acquired independently by the SARS and MERS clades of high-CFR CoV, most likely, enhancing the pathogenicity of high-CFR viruses and contributing to their ability to zoonotically transmit to humans. All these features are shared by the high-CFR CoV and their animal (in particular, bat) infecting relatives in the same clade, which is compatible with the possibility of zoonotic transition of additional highly pathogenic strains to humans. The predictions made through this analysis unveil critical features in the mechanism of SARS-CoV-2 virulence and evolutionary history and are amenable to straightforward experimental validation.

Methods

Data

The complete nucleotide sequences of 3035 coronavirus genomes were obtained from NCBI (Supplementary File 2). Of these, 944 genomes belong to viruses that infect humans, including both viruses with low case fatality rates (CFR), NL63, 229E, OC43 and HKU1, and those with high CFR, namely, MERS, SARS-CoV-1 and SARS-CoV-2. The protein sequences that are encoded in the genomes of all human coronaviruses and closely related viruses from animals were obtained from NCBI, including the two polyproteins (1ab and 1a), spike glycoprotein, envelope, membrane glycoprotein, and nucleocapsid phosphoprotein.

Identification of genomic determinants of high-CFR coronaviruses

To identify genomic determinants of coronaviruses associated with high CFR, comparative genome analysis was combined with machine learning techniques. First, the 944 human coronavirus genomes were aligned using Mafft[20] v7.407 . Then, we identified high confidence alignment blocks within the multiple sequence alignment (MSA), which were defined as regions longer than 15 bp, containing less than 10% of gaps in each position. We searched for regions containing deletions or insertions that separate high-CFR from low-CFR viruses and are surrounded by high confidence alignment blocks because these are most likely to contain relevant differences within conserved genomic regions. To this end, the aligned sequences were recoded such that each nucleotide was coded as '1' and each gap as '0'. We then applied Support Vector Machines (using the Python library scikit-learn[21] with a linear Kernel function) to a 5bp sliding window in the identified high-confidence alignment regions, with a leave-one-out cross validation (where all samples of one of the 7 coronaviruses were left out in each round of the cross validation, for a total of 7 rounds). Finally, we selected regions that predicted the high-CFR viruses with high confidence (greater than 80% accuracy) for further evaluation.

From the 11 regions identified, 4 were in the polyprotein 1AB, 3 in the spike glycoprotein, 1 in the membrane glycoprotein and 3 in the nucleocapsid phosphoprotein. To evaluate the significance of these finding, we computed a hyper-geometric enrichment P-value, using the sizes of the identified regions and the lengths of the coding regions of each protein within the MSA. We found that the nucleocapsid phosphoprotein was most enriched with genomic differences that predict CFR (P-value = 4e-16), followed by the Spike glycoprotein (P-value = 0.036) and that the polyprotein 1AB and membrane glycoprotein were not significantly enriched with such differences. We further examined the effects of this set of genomic differences on the resulting protein sequences, and found that only 4 of the 11 differences identified were reflected in the protein alignment, of which 3 occurred in the nucleocapsid phosphoprotein and one in the spike glycoprotein.

Non-human proximal coronavirus strains

To compile a list of human and proximal non-human coronavirus strains, we first constructed a multiple sequence alignment of all 3035 collected strains using Mafft v7.407. From that alignment, we build a phylogenetic tree using FastTree[22] 2.1.10 with the "-nt" parameter, and extracted the distances between leaves of each strain from each of the reference genomes of the 7 human coronaviruses (Supplementary File 3). We then extracted the proximal strains of each human coronavirus, which were within a distance less than 1.0 to one of the human coronaviruses. To obtain a unique set of strains, we removed highly similar strains by randomly sampling one strain from each group of strains with more than 98% pairwise sequence identity (the resulting strains are provided in Supplementary File 4).

Amino acid charge calculations

To evaluate the strength of the identified NLS and NES motifs within the nucleocapsid phosphoprotein, we calculated the amino acid cumulative charge within the alignment region of each motif, and of the complete protein, for each of the selected human and proximal non-human coronavirus strains (Supplementary File 4). The charge of each region was evaluated by the number of positively charged amino acids (lysine and arginine) minus the number of negatively charged amino acids in that region (aspartic acid and glutamic acid). To evaluate the significance of the association between CFR and the charge of specific motifs within the nucleocapsid phosphoprotein, we first calculated the rank-sum P-value comparing the charges of regions in high-CFR versus low-CFR strains. Then, we applied a permutation test, by counting the fraction of similarly or more significant charge differentials values between high-CFR and low-CFR viruses within 1,000 randomly selected motifs with similar length from the alignment of the nucleocapsid phosphoprotein.

Genomic determinants of the interspecies jump

To identify genomic determinants that discriminate high-CFR viruses that made the zoonotic transmission to humans, we used the nucleotide MSA of MERS, SARS-CoV-1 and SARS-CoV-2 and the selected proximal non-human coronaviruses of each of these.

We searched regions that maximize the following function:

$$\boldsymbol{f(S,V)} = \min_{d_k:k\ in\ V} \prod_{j:\ d_j > d_k} \boldsymbol{I}_{S_j \neq S_h}$$

Where $\boldsymbol{I}_{S_j \neq S_h} = \begin{cases} \mathbf{1}\ \boldsymbol{if\ S_j \neq S_h} \\ \mathbf{0}\ \boldsymbol{else} \end{cases}$

$\boldsymbol{S}$ is a position within the encoded MSA ('1' for a nucleotide and '0' for a gap), and $\boldsymbol{V}$ is the set of strains selected for either MERS, SARS-CoV-1 or SARS-CoV-2. $\boldsymbol{d_k}$ is the distance of non-human strain $\boldsymbol{k}$ from the human strain of group $\boldsymbol{V}$, $\boldsymbol{S_j}$ is position $\boldsymbol{S}$ of non-human strain $\boldsymbol{j}$, and $\boldsymbol{S_h}$ is position $\boldsymbol{S}$ of the human strain in group $\boldsymbol{V}$.

Thus, this function aims to find, for each position, within each of the three groups of strains, the non-human strain $\boldsymbol{k}$ with the minimal distance from the human strain, such that all non-human strains that are more distant are more different than the human strain in that position (i.e. a genomic change that occurred as close as possible to the human strain). We searched for regions in which over 50% of the strains in the alignment differed from the human strain, and for which the differing strains were explicitly the most distant from human. We identified only one such location, across all three high-CFR virus groups.

Acknowledgements

The authors' research is supported through the Intramural Research Program of the National Institutes of Health (National Library of Medicine)

References

1. World Health Organization. Statement on the Meeting of the International Health Regulations Emergency Committee Regarding the 2014 Ebola Outbreak in West Africa. *World Health Organization* 1–5 (2014).
2. World Health Organization. Coronavirus disease (COVID-19) Pandemic. https://www.who.int/emergencies/diseases/novel-coronavirus-2019.
3. Su, S. *et al.* Epidemiology, Genetic Recombination, and Pathogenesis of Coronaviruses. *Trends Microbiol.* **24**, 490–502 (2016).
4. Zhang, Y.-Z. & Holmes, E. C. A Genomic Perspective on the Origin and Emergence of SARS-CoV-2. *Cell* (2020) doi:10.1016/j.cell.2020.03.035.
5. Cui, J., Li, F. & Shi, Z.-L. Origin and evolution of pathogenic coronaviruses. *Nat. Rev. Microbiol.* **17**, 181–192 (2019).
6. You, J. *et al.* Subcellular localization of the severe acute respiratory syndrome coronavirus nucleocapsid protein. *J. Gen. Virol.* **86**, 3303–3310 (2005).
7. Cokol, M., Nair, R. & Rost, B. Finding nuclear localization signals. *EMBO Rep.* **1**, 411–5 (2000).
8. Wurm, T. *et al.* Localization to the nucleolus is a common feature of coronavirus nucleoproteins, and the protein may disrupt host cell division. *J. Virol.* **75**, 9345–56 (2001).
9. McBride, R., van Zyl, M. & Fielding, B. The Coronavirus Nucleocapsid Is a Multifunctional Protein. *Viruses* **6**, 2991–3018 (2014).
10. Chen, H., Wurm, T., Britton, P., Brooks, G. & Hiscox, J. A. Interaction of the Coronavirus Nucleoprotein with Nucleolar Antigens and the Host Cell. *J. Virol.* **76**, 5233–5250 (2002).
11. Pei, Y. *et al.* Functional mapping of the porcine reproductive and respiratory syndrome virus capsid protein nuclear localization signal and its pathogenic association. *Virus Res.* **135**, 107–114 (2008).
12. Rowland, R. R. R. *et al.* Intracellular Localization of the Severe Acute Respiratory Syndrome Coronavirus Nucleocapsid Protein: Absence of Nucleolar Accumulation during Infection and after Expression as a Recombinant Protein in Vero Cells. *J. Virol.* **79**, 11507–11512 (2005).
13. Wrapp, D. *et al.* Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. *Science (80-.).* **367**, 1260–1263 (2020).
14. Bosch, B. J. *et al.* Severe acute respiratory syndrome coronavirus (SARS-CoV) infection inhibition using spike protein heptad repeat-derived peptides. *Proc. Natl. Acad. Sci.* **101**, 8455–8460 (2004).
15. Chan, W.-E., Chuang, C.-K., Yeh, S.-H., Chang, M.-S. & Chen, S. S.-L. Functional characterization of heptad repeat 1 and 2 mutants of the spike protein of severe acute respiratory syndrome coronavirus. *J. Virol.* **80**, 3225–37 (2006).
16. Chambers, P., Pringle, C. R. & Easton, A. J. Heptad Repeat Sequences are Located Adjacent to Hydrophobic Regions in Several Types of Virus Fusion Glycoproteins. *J. Gen. Virol.* **71**, 3075–3080 (1990).
17. Lai, A. L., Millet, J. K., Daniel, S., Freed, J. H. & Whittaker, G. R. The SARS-CoV Fusion Peptide Forms an Extended Bipartite Fusion Platform that Perturbs Membrane Order in a Calcium-Dependent Manner. *J. Mol. Biol.* **429**, 3875–3892 (2017).
18. Wan, Y., Shang, J., Graham, R., Baric, R. S. & Li, F. Receptor Recognition by the Novel Coronavirus from Wuhan: an Analysis Based on Decade-Long Structural Studies of SARS Coronavirus. *J. Virol.* **94**, (2020).
19. Wang, N. *et al.* Structure of MERS-CoV spike receptor-binding domain complexed with human receptor DPP4. *Cell Res.* **23**, 986–993 (2013).
20. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–80 (2013).
21. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830

(2011).
22. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. *PLoS One* **5**, e9490 (2010).

Supplementary materials

Supplementary Information

Supplementary Table 1
Supplementary Figure 1

Supplementary datasets 1-4: ftp://ftp.ncbi.nih.gov/pub/wolf/_suppl/SARSpath20/