Overview

Sequencing of 2019-nCoV revealed two notable features of its genome. We investigate these features and outline some examples for how the virus may have acquired them. We also discuss some scenarios by which these features could have arisen. **Analysis of the virus genome sequences clearly demonstrates that the virus is not a laboratory construct or experimentally manipulated virus**. We believe the features discussed, which may explain the infectiousness and transmissibility of 2019-nCoV in humans, could have arisen through selection and adaptation prior to the initial outbreak.

The two primary features of 2019-nCoV of interest were:

- Based on structural modeling and early biochemical experiments, 2019-nCoV appears to be optimized for binding to the human ACE2 receptor.
- The highly variable spike protein of 2019-nCoV has a furin cleavage inserted at the S1 and S2 boundary via the insertion of twelve in-frame nucleotides. Additionally, this event also led to the acquisition of three predicted O-linked glycans around the furin cleavage site.

Mutations in the receptor binding domain of 2019-nCoV

The receptor binding domain (RBD) in the spike protein of SARS-CoV and SARS-like coronaviruses is the most variable part of the virus genome. When aligned against related viruses, 2019-nCoV displays a similar level of diversity as predicted from previous studies, including to its most closely related virus - SARS-like CoV isolated from bats (RaTG13, which is ~96% identical to 2019-nCoV).

Six residues in the RBD have been described as critical for binding to the human ACE2 receptor and determining host range[1]. Using coordinates based on the Ubani strain of SARS-CoV, they are Y442, L472, N479, D480, T487, and Y491 (the corresponding residues in 2019-nCoV are L455, F486, Q493, S494, N501, and Y505). Five out of six of these residues are mutated in 2019-nCoV compared to the closely related virus, RaTG13 (**Figure 1**). Based on modeling[1] and early biochemical experiments[2,3], 2019-nCoV seems to have an RBD that may bind with high affinity to ACE2 from human, primate, ferret, pig, and cat, as well as other species with high receptor homology. In contrast, 2019-nCoV may bind less efficiently to ACE2 in other species associated with SARS-like viruses, including rodents, civets, and bats[1].

A phenylalanine at F486 in 2019-nCoV corresponds to L472 in the SARS-CoV Ubani strain. In cell culture experiments the leucine at position 472 mutated to phenylalanine (L472F)[4], which has been predicted to be optimal for binding of the SARS-CoV RBD to the human ACE2 receptor[5]. However, a phenylalanine in this position is also present in several SARS-like CoVs from bats (**Figure 1**). While these analyses suggest that 2019-nCoV may be capable of binding the human ACE2 receptor with high affinity, importantly, the interaction is not predicted to be optimal[1]. Additionally, several of the key residues in the RBD of 2019-nCoV are different from those previously described to be optimal for human ACE2 receptor binding as determined by both natural evolution of SARS-CoV and rational design[5]. This latter point is strong evidence *against* 2019-nCoV being specifically engineered as, presumably, in such a scenario the most optimal residues would have been introduced, which is not what we observe.

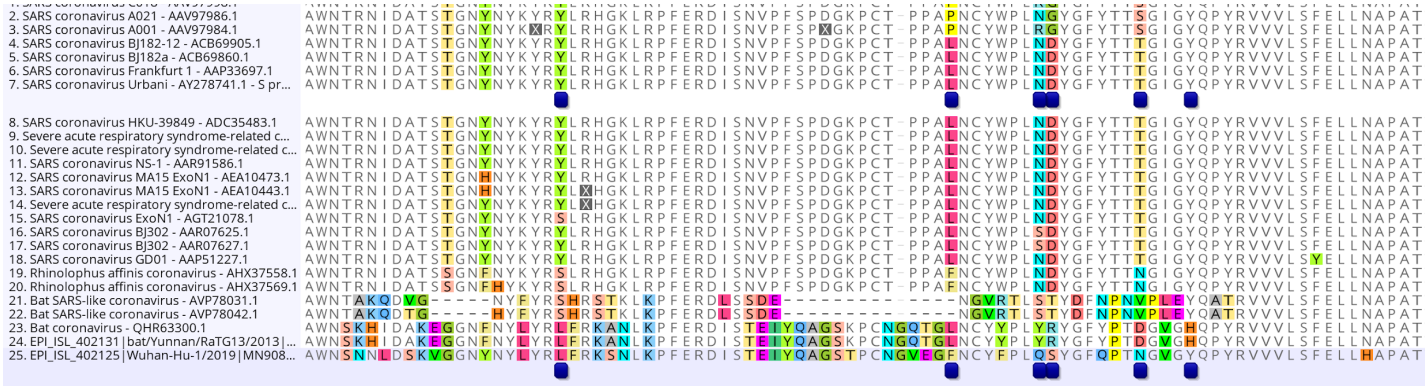

Figure 1 | Mutations in contact residues of the 2019-nCoV spike protein. The spike protein of 2019-nCoV (bottom) was aligned against the most closely related SARS and SARS-like CoVs. Key residues in the spike protein that make contact to the ACE2 receptor have been marked with blue boxes in both 2019-nCoV and the SARS-CoV Urbani strain.

Furin cleavage site and O-linked glycans

An interesting feature of 2019-nCoV is a predicted furin cleavage site in the spike protein (**Figure 2**). In addition to the furin cleavage site (RRAR), a leading P is also inserted so the fully inserted sequence becomes PRRA (**Figure 2**). A proline in this position is predicted to create three flanking O-linked glycans at S673, T678, and S686. A furin site has never before been observed in the lineage B betacoronaviruses and is a unique feature of 2019-nCoV. Some human betacoronaviruses, including HCoV-HKU1 (lineage A) have furin cleavage sites (typically RRKR), although not in such an optimal position.

Figure 2 | Acquisition of furin cleavage site and O-linked glycans. The spike protein of 2019-nCoV (bottom) was aligned against the most closely related SARS and SARS-like CoVs. The furin cleavage site is marked in grey with the three adjacent predicted O-linked glycans in blue. Both the furin cleavage site and O-linked glycans are unique to 2019-nCoV and not previously seen in this group of viruses.

While the functional consequence - if any - of the furin cleavage site in 2019-nCoV is unknown, previous experiments with SARS-CoV have shown that it enhances cell–cell fusion but does not affect virus entry[6]. Furin cleavage sites are often acquired in condition selecting for rapid virus replication and transmission (e.g., highly dense chicken populations) and are a hallmark of highly pathogenic avian influenza virus, although these viruses acquire the site in different and more direct ways[7–9]. The acquisition of furin cleavage sites have also been observed after repeated passage of viruses in cell culture (personal correspondence and NASEM call, February 3, 2020).

A potential function of the three predicted O-linked glycans is less clear, but could create a “mucin-like domain” shielding potential epitopes or key residues on the 2019-nCoV spike protein.

Origin of 2019-nCoV

As noted at the start of this document, we believe that the origin of 2019-nCoV through laboratory manipulation of an existing SARS-related coronavirus can be ruled out with a high degree of confidence. If

genetic manipulation would have been performed, one would expect that a researcher would have used one of the several reverse genetics systems available for betacoronaviruses. However, this is not the case as the genetic data clearly shows that 2019-nCoV is not derived from any previously used virus backbone, for example those described in a 2015 paper in *Nature Medicine*[10].

Instead we believe one of three main scenarios could explain how 2019-nCoV acquired the features discussed above: (1) natural selection in humans, (2) natural selection in an animal host, or (3) selection during passage.

Adaptation to humans

As the features outlined above are likely to enhance the ability of the virus to infect humans, it is possible that these are indeed adaptations to humans as a host and arose after the virus jumped from a non-human host, during the early stages of the epidemic. However, all of the genome sequences so far have the features described above and estimates of the timing of the most recent common ancestor of the currently sampled viruses support the seafood market outbreak as the zoonotic origin (i.e., in early December) and this would afford little opportunity for adaptation to occur. This may be explained by a transition to a rapid growth phase in the epidemic when the features arose and from which all current cases are derived. However this would require a prior hidden epidemic of sufficient magnitude and duration for the adaptations to occur and there is no evidence of this. We also note that these features did not emerge during the SARS epidemic, which involved extensive human to human transmission.

Selection in an animal host

Given the similarity of 2019-nCoV to bat SARS-like CoVs, particularly RaTG13, it is highly likely that bats serve as the reservoir for this virus. However, previous human epidemics caused by betacoronaviruses have involved intermediate (possibly amplifying) hosts such as civets and other animals (SARS) and camels (MERS). It is therefore likely that an intermediate host would also exist for 2019-nCoV, although it is unclear what that host may be. Given the mutations in key residues of the RBD in 2019-nCoV it seems less likely that civets would be involved, although it is impossible to say with certainty at this stage. Notably, provisional analyses reveal that Malayan pangolins (*Manis javanica*) illegally imported into Guangdong province contain CoVs that are extremely similar to 2019-nCoV[11]. Although RaTG13 remains the closest relative to 2019-nCoV across the genome as a whole, the Malayan pangolin CoVs are identical to 2019-nCoV at all six key RBD residues. Analyses of these pangolin viruses are ongoing, although they do not carry the furin cleavage site insertion.

For the virus to acquire the furin cleavage site and mutations in the spike proteins that appear to be suitable for human ACE2 receptor binding, it seems plausible that this animal host would have to have a high population density – to allow the necessary natural selection to proceed efficiently – and an ACE2 gene that is similar to the human orthologue. Since furin cleavage sites have not been observed in sarbecoviruses before, it is unclear what conditions would be required for it to be acquired in the lineage leading to 2019-nCoV.

Selection during passage

Basic research involving passage of bat SARS-like coronaviruses in cell culture and/or animal models have been ongoing in BSL-2 for many years across the world, including in Wuhan (e.g.,[12–15]). It is possible that 2019-nCoV could have acquired the RBD mutations and furin cleavage site as part of passage in cell culture, which have been observed in previous studies with e.g., SARS-CoV[4]. However, it is less clear how the O-linked glycans - if functional - would have been acquired, as these typically suggest the involvement of an immune system, which is not present *in vitro*. In this scenario, it is also unclear how the virus would be linked to the fact that the epidemic seemed to 'take off' at a particular food market, although the exact role of this locality is currently uncertain.

Limitations and recommendations

The evolution scenarios discussed above are largely indistinguishable and current data are consistent with all three. It is currently impossible to prove or disprove either, and it is unclear whether future data or analyses will help resolve this issue. Identifying the immediate non-human animal source and obtaining virus sequences from it would be the most definitive way of distinguishing the three scenarios.

The main limitation of what is described here is our clear ascertainment bias. We are looking for features or evolutionary aspects that could help explain how 2019-nCoV lead to such a rapidly expanding human epidemic, yet the specific features we are trying to find may be the exact features one would expect in a virus that could lead to an epidemic of the magnitude currently observed. Before 2019-nCoV 'took off' and started the current epidemic, it is plausible that many stuttering transmission chains of highly similar viruses could have entered the human population, but because they never took off they were never sampled. It is extremely important to keep this in mind as any inference about the plausibility of various scenarios about the evolution and/or epidemic potential of 2019-nCoV is attempted.

To further clarify the evolutionary origins and functional features of 2019-nCoV it would be helpful to obtain additional data about the virus - both genetic and functional. This includes experimental studies of receptor binding and the role of the furin cleavage site and predicted O-linked glycans. The identification of a potential intermediate host of 2019-nCoV as well as sequencing of very early cases, including those not connected to the market, could also help refute the passage scenario described above. Even in the light of such data, however, it is not guaranteed that data can be obtained to conclusively prove all aspects of the initial emergence of 2019-nCoV.

References

1. Wan, Y., Shang, J., Graham, R., Baric, R. S. & Li, F. Receptor recognition by novel coronavirus from Wuhan: An analysis based on decade-long structural studies of SARS. *J. Virol.* (2020) doi:10.1128/JVI.00127-20.
2. Letko, M. & Munster, V. Functional assessment of cell entry and receptor usage for lineage B β-coronaviruses, including 2019-nCoV. *bioRxiv* 2020.01.22.915660 (2020) doi:10.1101/2020.01.22.915660.
3. Hoffmann, M. *et al.* The novel coronavirus 2019 (2019-nCoV) uses the SARS-coronavirus receptor ACE2 and the cellular protease TMPRSS2 for entry into target cells. *bioRxiv* 2020.01.31.929042 (2020) doi:10.1101/2020.01.31.929042.
4. Sheahan, T. *et al.* Mechanisms of zoonotic severe acute respiratory syndrome coronavirus host range expansion in human airway epithelium. *J. Virol.* **82**, 2274–2285 (2008).
5. Cui, J., Li, F. & Shi, Z.-L. Origin and evolution of pathogenic coronaviruses. *Nat. Rev. Microbiol.* **17**, 181–192 (2019).
6. Follis, K. E., York, J. & Nunberg, J. H. Furin cleavage of the SARS coronavirus spike glycoprotein enhances cell-cell fusion but does not affect virion entry. *Virology* **350**, 358–369 (2006).
7. Longping, V. T., Hamilton, A. M., Friling, T. & Whittaker, G. R. A novel activation mechanism of avian influenza virus H9N2 by furin. *J. Virol.* **88**, 1673–1683 (2014).
8. Alexander, D. J. & Brown, I. H. History of highly pathogenic avian influenza. *Rev. Sci. Tech.* **28**, 19–38 (2009).
9. Luczo, J. M. *et al.* Evolution of high pathogenicity of H5 avian influenza virus: haemagglutinin cleavage site selection of reverse-genetics mutants during passage in chickens. *Sci. Rep.* **8**, 11518 (2018).
10. Menachery, V. D. *et al.* A SARS-like cluster of circulating bat coronaviruses shows potential for human emergence. *Nat. Med.* **21**, 1508–1513 (2015).
11. virological.org: http://virological.org/t/ncov-2019-spike-protein-receptor-binding-domain-shares-high-amino-acid-identity-with-a-coronavirus-recovered-from-a-pangolin-viral-metagenomic-dataset/362 (2020).
12. Ge, X.-Y. *et al.* Isolation and characterization of a bat SARS-like coronavirus that uses the ACE2 receptor. *Nature* **503**, 535–538 (2013).
13. Hu, B. *et al.* Discovery of a rich gene pool of bat SARS-related coronaviruses provides new insights into the origin of SARS coronavirus. *PLoS Pathog.* **13**, e1006698 (2017).

14. Zeng, L.-P. *et al.* Bat Severe Acute Respiratory Syndrome-Like Coronavirus WIV1 Encodes an Extra Accessory Protein, ORFX, Involved in Modulation of the Host Immune Response. *J. Virol.* **90**, 6573–6582 (2016).

15. Yang, X.-L. *et al.* Isolation and Characterization of a Novel Bat Coronavirus Closely Related to the Direct Progenitor of Severe Acute Respiratory Syndrome Coronavirus. *J. Virol.* **90**, 3253–3256 (2015).