We discussed four specific features of 2019-nCoV:

1. Highly mutated receptor binding domain (RBD), including around key residues when comparing 2019-nCoV to a highly related bat SARS-like CoV (RaTG13; 96% identity).
    a. Nothing unusual. Comparisons were made between between SARS and SARS-like CoVs that were of similar divergence as nCoV and RaTG13. Similar levels of diversity were observed in the RBD, showing that this domain in general is highly variable, which is likely due to strong positive selection for receptor binding.
2. Reversion of gain-of-function site in the RBD to that seen in SARS.
    a. Nothing unusual. The reversion of F (observed in RaTG13) to Y (observed in SARS and nCoV) is only a single base-pair change (A > T transversion). Given that the RBD is variable in general, this is not unusual.
3. Gain of BamHI restriction site in the 3' end of the spike protein of nCoV. Sequence upstream of site is somewhat variable and sequence after is conserved.
    a. Probably not unusual. The gain of the BamHI site in nCoV is the result of a single synonymous transition (T > C) that happens frequently in RNA viruses. The 3' sequence following this site is conserved not only between nCoV and RaTG13, but also more broadly across similar viruses. The site could be used to insert different versions of the spike protein gene into nCoV, but no specific data suggest that it is utilized as such.
4. Gain of furin cleavage site and O-linked glycans.
    a. Unclear. This is the first time an optimal furin cleavage site has been observed in a betacoronavirus and it is additionally coupled to a gain of O-linked glycans. Several different scenarios could explain how this was gained:
        i. Natural selection, plausibly in a non-bat reservoir / intermediate host.
        ii. Repeated passage of virus in tissue culture.
        iii. Specific engineering of the site.

In summary, after considering all things above, the only thing that remains perplexing about 2019-nCoV is the fact that it has a furin site with O-linked glycans in the spike protein between S1 and S2. It is impossible to distinguish whether this was gained due to e.g., evolution or passage, and the data is consistent with either scenario. Specific engineering is also a possibility, but appears less likely as that would require significant amounts of molecular work utilizing an uncommon virus backbone.

Below we briefly outline different scenarios for how nCoV may have originated.

1. Bioweapon. Highly unlikely and there is no data supporting this hypothesis.
2. Specific engineering. Unlikely as this would require significant amounts of work utilizing uncommon and currently unknown backbones of SARS-like bat CoVs. For this type of work, there are preexisting backbones that could have been utilized, but they clearly were not.
3. Tissue culture passage. The data is consistent with this scenario, although no specific hypothesis exists for how the furin site was gained, but could be due to passage in tissue culture. The virus could have been released via accidental infection of researcher(s).
4. Spillover from animal host. The data is consistent with this scenario, although no specific hypothesis exists for how the furin site was gained in an animal host. However, even though the furin site has

February 1, 2020 7:40pm

not been observed in these viruses previously, virus evolution coupled with strong selective pressure (possibly in an intermediate host) would be capable of creating such a domain.

February 1, 2020 7:40pm

We discussed four specific features of 2019-nCoV:

1. Highly mutated receptor binding domain (RBD), including around key residues when comparing 2019-nCoV to a highly related bat SARS-like CoV (RaTG13; 96% identity).
    a. Nothing unusual. Comparisons were made between between SARS and SARS-like CoVs that were of similar divergence as nCoV and RaTG13. Similar levels of diversity were observed in the RBD, showing that this domain in general is highly variable, which is likely due to strong positive selection for receptor binding.
2. Reversion of gain-of-function site in the RBD to that seen in SARS.
    a. Nothing unusual. The reversion of F (observed in RaTG13) to Y (observed in SARS and nCoV) is only a single base-pair change (A > T transversion). Given that the RBD is variable in general, this is not unusual.
3. Gain of BamHI restriction site in the 3' end of the spike protein of nCoV. Sequence upstream of site is somewhat variable and sequence after is conserved.
    a. Probably not unusual. The gain of the BamHI site in nCoV is the result of a single synonymous transition (T > C) that happens frequently in RNA viruses. The 3' sequence following this site is conserved not only between nCoV and RaTG13, but also more broadly across similar viruses. The site could be used to insert different versions of the spike protein gene into nCoV, but no specific data suggest that it is utilized as such.
4. Gain of furin cleavage site and O-linked glycans.
    a. Unclear. This is the first time an optimal furin cleavage site has been observed in a betacoronavirus and it is additionally coupled to a gain of O-linked glycans. Several different scenarios could explain how this was gained:
        i. Natural selection, plausibly in a non-bat reservoir / intermediate host.
        ii. Repeated passage of virus in tissue culture.
        iii. Specific engineering of the site.

In summary, after considering all things above, the only thing that remains perplexing about 2019-nCoV is the fact that it has a furin site with O-linked glycans in the spike protein between S1 and S2. It is impossible to distinguish whether this was gained due to e.g., evolution or passage, and the data is consistent with either scenario. Specific engineering is also a possibility [would make insertion really easy] [would require molecular work].

Below we briefly outline different scenarios for how nCoV may have originated.

1. Bioweapon. Highly unlikely and there is no data supporting this hypothesis.
2. Specific engineering. Unlikely as this would require significant amounts of work utilizing uncommon and currently unknown backbones of SARS-like bat CoVs. For this type of work, there are preexisting backbones that could have been utilized, but they clearly were not.
3. Tissue culture passage. The data is consistent with this scenario, although no specific hypothesis exists for how the furin site was gained, but could be due to passage in tissue culture. The virus could have been released via accidental infection of researcher(s).
4. Spillover from animal host. The data is consistent with this scenario, although no specific hypothesis exists for how the furin site was gained in an animal host. However, even though the furin site has

February 1, 2020 8:57pm

not been observed in these viruses previously, virus evolution coupled with strong selective pressure (possibly in an intermediate host) would be capable of creating such a domain.

February 1, 2020 8:57pm

Four features - already noticed or easy to discover

Changes in RBD can easily be explained

- Not quite. Residue (SARS coordinates) 472 picks up F in tissue culture from L increasing binding and infectivity (PMID: 18094188). In nCoV (position 486) F is fixed in this position - it's an L in RaTG13 and other bat viruses
- Of the 6 critical contact residues described, nCoV has mutations in 5/6 as compared to bats (https://jvi.asm.org/content/early/2020/01/23/JVI.00127-20). Most of these optimal for interaction with ACE2, including ones that mutated in SARS *during* the epidemic, leading to better binding and infectivity.
- Highly optimized for binding to human ACE2 receptor

BamHI site doesn't mean anything and is a small synonymous transition

Furin site + O-linked glycans more difficult

- Evolution, likely in non-bat reservoir
  - Selection can do amazing things
  - We're missing a lot of evolution and has never happened before in CoV
- Passage in either cells or animals as part of ongoing research on SARS-like bat CoVs
  - Selection for extremely rapid transmission
  - Could lead to acquisition of furin cleavage site
- Specific engineering as part of ongoing basic research (this 'trick' has been done in SARS)
  - Easy to introduce the site this way
  - BamHI and other sites could be used, however, many other ways to do it
  - For this type of research, investigators would have to be using a novel reverse genetics system not previously described, as opposed to those already available. This seems less likely.
- Data is consistent with all three but it is impossible to definitively prove any single scenario
  - Apart from the simplest scenario of somebody having introduced a novel gene/insert into a pre-existing virus backbone, it is difficult to see exactly what conclusive evidence would look like

Two different ways of origin of outbreak considered

- Introduction from animal reservoir - specific scenarios considered below.
- Accidental infection of researcher as part of ongoing research.
  - This type of research (including gain of function research on SARS-like bat CoVs) has been ongoing in Wuhan and other places (published papers)
  - Consideration for what containment would have been used - ranging from likely (BSL2) to unlikely (BSL4). We cannot answer this question

We discussed four specific features of 2019-nCoV:

1. Highly mutated receptor binding domain (RBD), including around key residues when comparing 2019-nCoV to a highly related bat SARS-like CoV (RaTG13; 96% identity).
   a. Nothing unusual. Comparisons were made between between SARS and SARS-like CoVs that were of similar divergence as nCoV and RaTG13. Similar levels of diversity were observed in the RBD, showing that this domain in general is highly variable, which is likely due to strong positive selection for receptor binding.
2. Reversion of gain-of-function site in the RBD to that seen in SARS.

February 2, 2020 6:29pm

    a. Nothing unusual. The reversion of F (observed in RaTG13) to Y (observed in SARS and nCoV) is only a single base-pair change (A > T transversion). Given that the RBD is variable in general, this is not unusual.

3. Gain of BamHI restriction site in the 3' end of the spike protein of nCoV. Sequence upstream of site is somewhat variable and sequence after is conserved.
    a. Probably not unusual. The gain of the BamHI site in nCoV is the result of a single synonymous transition (T > C) that happens frequently in RNA viruses. The 3' sequence following this site is conserved not only between nCoV and RaTG13, but also more broadly across similar viruses. The site could be used to insert different versions of the spike protein gene into nCoV, but no specific data suggest that it is utilized as such.
4. Gain of furin cleavage site and O-linked glycans.
    a. Unclear. This is the first time an optimal furin cleavage site has been observed in a betacoronavirus and it is additionally coupled to a gain of O-linked glycans. Several different scenarios could explain how this was gained:
        i. Natural selection, plausibly in a non-bat reservoir / intermediate host.
        ii. Repeated passage of virus in tissue culture.
        iii. Specific engineering of the site.

In summary, after considering all things above, the only thing that remains perplexing about 2019-nCoV is the fact that it has a furin site with O-linked glycans in the spike protein between S1 and S2. It is impossible to distinguish whether this was gained due to e.g., evolution or passage, and the data is consistent with either scenario. Specific engineering is also a possibility [would make insertion really easy] [would require molecular work].

Below we briefly outline different scenarios for how nCoV may have originated.

1. Bioweapon. Highly unlikely and there is no data supporting this hypothesis.
2. Specific engineering. Unlikely as this would require significant amounts of work utilizing uncommon and currently unknown backbones of SARS-like bat CoVs. For this type of work, there are preexisting backbones that could have been utilized, but they clearly were not.
3. Tissue culture passage. The data is consistent with this scenario, although no specific hypothesis exists for how the furin site was gained, but could be due to passage in tissue culture. The virus could have been released via accidental infection of researcher(s).
4. Spillover from animal host. The data is consistent with this scenario, although no specific hypothesis exists for how the furin site was gained in an animal host. However, even though the furin site has not been observed in these viruses previously, virus evolution coupled with strong selective pressure (possibly in an intermediate host) would be capable of creating such a domain.

February 2, 2020 6:29pm

Background:

Bat coronavirus RaTG13 is the closest relative to nCoV-2019. Two recombinant bat viruses are close in some regions of the genomes. Pangolin virus?

Furin cleavage site rough notes about evolutionary origins:

Avian influenza example of natural and spontaneous evolution - get references and details.

There are two scenarios by which we could imagine the furin cleavage site could evolve.

1. As a human adaptation during the initial stages of the outbreak. The appearance of the mutation may have then triggered a second phase of rapid transmission. All current genome sequences are from this second phase and thus show limited diversity.

2. Adaptation to a non-human host prior to the jump to humans. This mutation is not seen in any bat coronavirus and is thus unlikely to be adaptive in those species.

Thoughts on 1: is it likely to spontaneously appear in a relatively short amount of time (and presumably small number of infections). It didn't happen in SARS with 8000 infections over 6 months. The link to the market would then be spurious - some doubt on that already. Prediction would be that the animal/environmental samples apparently found by China CDC would not have cleavage site.

Thoughts on 2: can we suggest a host where this cleavage site would likely be advantageous. Ferrets/polecats? Rodents - bamboo rats (don't know if they are popular in China)? Circulating in wild populations so limited prior human exposure until infected individual brought to the market.

February 2, 2020 6:29pm

Overview

Sequencing of 2019-nCoV revealed two particularly notable features of its genome. We investigate these features and outline some examples for how the virus may have acquired them. As rumours have been circulating about this virus being engineered or otherwise created with intent, we wish to make it clear that our analyses show that such scenarios are largely incompatible with the data.

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 an optimal 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 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 closely related viruses, including 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 often 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 tissue 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[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.

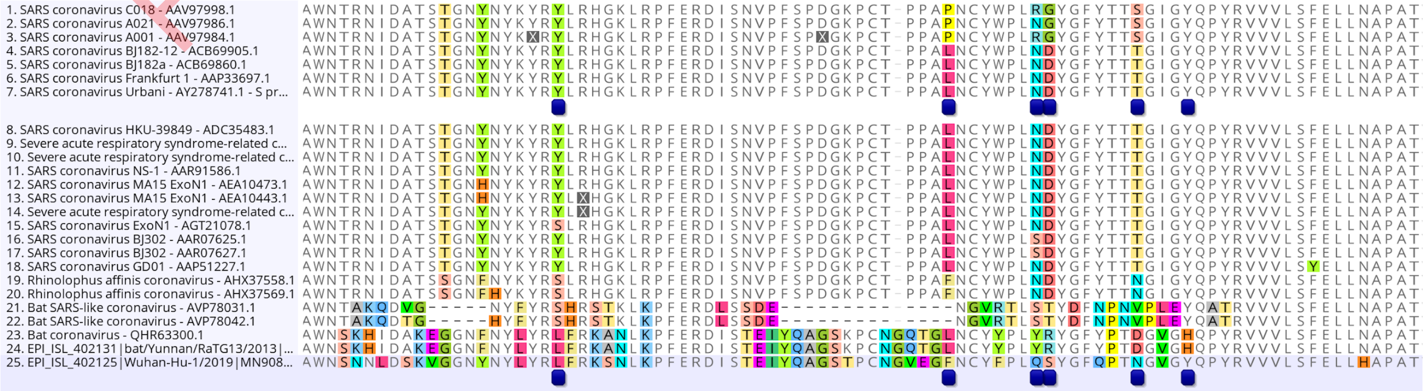

1. SARS coronavirus C018 - AAV97998.1
2. SARS coronavirus A021 - AAV97986.1
3. SARS coronavirus A001 - AAV97984.1
4. SARS coronavirus BJ182-12 - ACB69905.1
5. SARS coronavirus BJ182a - ACB69860.1
6. SARS coronavirus Frankfurt 1 - AAP33697.1
7. SARS coronavirus Urbani - AY278741.1 - S pr...
8. SARS coronavirus HKU-39849 - ADC35483.1
9. Severe acute respiratory syndrome-related c...
10. Severe acute respiratory syndrome-related c...
11. SARS coronavirus NS-1 - AAR91586.1
12. SARS coronavirus MA15 ExoN1 - AEA10473.1
13. SARS coronavirus MA15 ExoN1 - AEA10443.1
14. Severe acute respiratory syndrome-related c...
15. SARS coronavirus ExoN1 - AGT21078.1
16. SARS coronavirus BJ302 - AAR07625.1
17. SARS coronavirus BJ302 - AAR07627.1
18. SARS coronavirus GD01 - AAP51227.1
19. Rhinolophus affinis coronavirus - AHX37558.1
20. Rhinolophus affinis coronavirus - AHX37569.1
21. Bat SARS-like coronavirus - AVP78031.1
22. Bat SARS-like coronavirus - AVP78042.1
23. Bat coronavirus - QHR63300.1
24. EPI_ISL_402131|bat/Yunnan/RaTG13/2013|...
25. EPI_ISL_402125|Wuhan-Hu-1/2019|MN908...

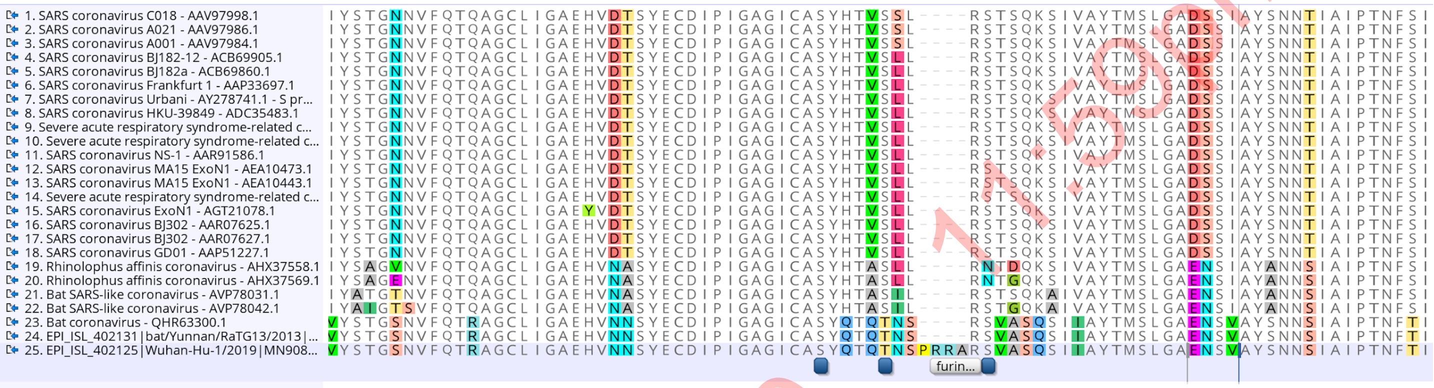
Figure 1 | Mutations in contact residues of the 2019-nCoV spike protein. The spike protein of 2019-nCoV (bottom) was

February 3, 2020 11:59pm

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.

Acquisition of furin cleavage site and O-linked glycans

An interesting feature of 2019-nCoV is the acquisition of 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**). The addition of a proline in this position is also predicted to create three O-linked glycans at S673, T678, and S686. The addition of 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.



```
1. SARS coronavirus C018 - AAV97998.1                    IYSTGNNVFQTQAGCLIGAEHVDTSYECDIPIGAGICASYHTVSSL----RSTSQKSIVAYTMSLGADSSIAYSNNTIAIPTNFSI
2. SARS coronavirus A021 - AAV97986.1                    IYSTGNNVFQTQAGCLIGAEHVDTSYECDIPIGAGICASYHTVSSL----RSTSQKSIVAYTMSLGADSSIAYSNNTIAIPTNFSI
3. SARS coronavirus A001 - AAV97984.1                    IYSTGNNVFQTQAGCLIGAEHVDTSYECDIPIGAGICASYHTVSSL----RSTSQKSIVAYTMSLGADSSIAYSNNTIAIPTNFSI
4. SARS coronavirus BJ182-12 - ACB69905.1                IYSTGNNVFQTQAGCLIGAEHVDTSYECDIPIGAGICASYHTVSLL----RSTSQKSIVAYTMSLGADSSIAYSNNTIAIPTNFSI
5. SARS coronavirus BJ182a - ACB69860.1                  IYSTGNNVFQTQAGCLIGAEHVDTSYECDIPIGAGICASYHTVSLL----RSTSQKSIVAYTMSLGADSSIAYSNNTIAIPTNFSI
6. SARS coronavirus Frankfurt 1 - AAP33697.1             IYSTGNNVFQTQAGCLIGAEHVDTSYECDIPIGAGICASYHTVSLL----RSTSQKSIVAYTMSLGADSSIAYSNNTIAIPTNFSI
7. SARS coronavirus Urbani - AY278741.1 - S pr...        IYSTGNNVFQTQAGCLIGAEHVDTSYECDIPIGAGICASYHTVSLL----RSTSQKSIVAYTMSLGADSSIAYSNNTIAIPTNFSI
8. SARS coronavirus HKU-39849 - ADC35483.1               IYSTGNNVFQTQAGCLIGAEHVDTSYECDIPIGAGICASYHTVSLL----RSTSQKSIVAYTMSLGADSSIAYSNNTIAIPTNFSI
9. Severe acute respiratory syndrome-related c...        IYSTGNNVFQTQAGCLIGAEHVDTSYECDIPIGAGICASYHTVSLL----RSTSQKSIVAYTMSLGADSSIAYSNNTIAIPTNFSI
10. Severe acute respiratory syndrome-related c...       IYSTGNNVFQTQAGCLIGAEHVDTSYECDIPIGAGICASYHTVSLL----RSTSQKSIVAYTMSLGADSSIAYSNNTIAIPTNFSI
11. SARS coronavirus NS-1 - AAR91586.1                   IYSTGNNVFQTQAGCLIGAEHVDTSYECDIPIGAGICASYHTVSLL----RSTSQKSIVAYTMSLGADSSIAYSNNTIAIPTNFSI
12. SARS coronavirus MA15 ExoN1 - AEA10473.1             IYSTGNNVFQTQAGCLIGAEHVDTSYECDIPIGAGICASYHTVSLL----RSTSQKSIVAYTMSLGADSSIAYSNNTIAIPTNFSI
13. SARS coronavirus MA15 ExoN1 - AEA10443.1             IYSTGNNVFQTQAGCLIGAEHVDTSYECDIPIGAGICASYHTVSLL----RSTSQKSIVAYTMSLGADSSIAYSNNTIAIPTNFSI
14. Severe acute respiratory syndrome-related c...       IYSTGNNVFQTQAGCLIGAEHVDTSYECDIPIGAGICASYHTVSLL----RSTSQKSIVAYTMSLGADSSIAYSNNTIAIPTNFSI
15. SARS coronavirus ExoN1 - AGT21078.1                  IYSTGNNVFQTQAGCLIGAEYVDTSYECDIPIGAGICASYHTVSLL----RSTSQKSIVAYTMSLGADSSIAYSNNTIAIPTNFSI
16. SARS coronavirus BJ302 - AAR07625.1                  IYSTGNNVFQTQAGCLIGAEHVDTSYECDIPIGAGICASYHTVSLL----RSTSQKSIVAYTMSLGADSSIAYSNNTIAIPTNFSI
17. SARS coronavirus BJ302 - AAR07627.1                  IYSTGNNVFQTQAGCLIGAEHVDTSYECDIPIGAGICASYHTVSLL----RSTSQKSIVAYTMSLGADSSIAYSNNTIAIPTNFSI
18. SARS coronavirus GD01 - AAP51227.1                   IYSTGNNVFQTQAGCLIGAEHVDTSYECDIPIGAGICASYHTVSLL----RSTSQKSIVAYTMSLGADSSIAYSNNTIAIPTNFSI
19. Rhinolophus affinis coronavirus - AHX37558.1         IYSAGVNVFQTQAGCLIGAEHVNASYECDIPIGAGICASYHTASLL----RNTDQKSIVAYTMSLGAENSIAYANNSIAIPTNFSI
20. Rhinolophus affinis coronavirus - AHX37569.1         IYSAGENVFQTQAGCLIGAEHVNASYECDIPIGAGICASYHTASLL----RNTGQKSIVAYTMSLGAENSIAYANNSIAIPTNFSI
21. Bat SARS-like coronavirus - AVP78031.1                IYATGTNVFQTQAGCLIGAEHVNASYECDIPIGAGICASYHTASIL----RSTSQKAIVAYTMSLGAENSIAYANNSIAIPTNFSI
22. Bat SARS-like coronavirus - AVP78042.1                IYAIGTSVFQTQAGCLIGAEHVNASYECDIPIGAGICASYHTASIL----RSTGQKAIVAYTMSLGAENSIAYANNSIAIPTNFSI
23. Bat coronavirus - QHR63300.1                         VYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICASYQTQTNS----RSVASQSIIAYTMSLGAENSVAYSNNSIAIPTNFTI
24. EPI_ISL_402131|bat/Yunnan/RaTG13/2013|...            VYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICASYQTQTNS----RSVASQSIIAYTMSLGAENSVAYSNNSIAIPTNFTI
25. EPI_ISL_402125|Wuhan-Hu-1/2019|MN908...              VYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICASYQTQTNSPRRARSVASQSIIAYTMSLGAENSVAYSNNSIAIPTNFTI
```

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[7–9]. The acquisition of furin cleavage sites have also been observed after repeated passage of betacoronaviruses in tissue culture (personal correspondence and NASEM call, February 3, 2020).

A potential function of the three 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.

Evolution of 2019-nCoV

Three main scenarios could explain how 2019-nCoV acquired the features discussed above: (1) natural selection in an animal host, (2) selection during passage, or (3) deliberate engineering. As described in the beginning, engineering (#3) can be ruled out with a high degree of confidence as the data is inconsistent with this scenario. In addition, if engineering would have been performed, one would also 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, including those recently posited by various conspiracy theories, based on a 2015 paper in *Nature Medicine*[10].

The other two scenarios are largely indistinguishable and current data are consistent with both. It is currently impossible to prove or disprove either, and it is unclear whether future data or analyses will help resolve this issue.

February 3, 2020 17:50pm

Selection in an animal host

Given the similarity of 2019-nCoV to bat SARS-like CoVs, particularly RaTG13, it is highly likely that bats also serve as the reservoir for this virus. However, previous human epidemics caused by betacoronaviruses have involved intermediate (possibly amplifying) hosts such as civets (SARS) and camels (MERS). It is therefore likely that an intermediate host would also exist for 2019-nCoV, although it is currently 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.

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 very 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 this group of viruses 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 tissue culture and/or animal models have been ongoing in BSL-2 for many years across the world, including in Wuhan (e.g.,[11–14]). It is possible that 2019-nCoV could have acquired the RBD mutations and furin cleavage site as part of passage in tissue 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 main limitation of what is described here is the clear ascertainment bias. We are looking for features or evolutionary aspects that could help explain how 2019-nCoV could lead to a rapidly evolving 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 detected. 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 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.

February 3, 2020 11:59pm

Background:

Bat coronavirus RaTG13 is the closest relative to nCoV-2019. Two recombinant bat viruses are close in some regions of the genomes. Pangolin virus?

Furin cleavage site rough notes about evolutionary origins:

Avian influenza example of natural and spontaneous evolution - get references and details.

There are two scenarios by which we could imagine the furin cleavage site could evolve.

1. As a human adaptation during the initial stages of the outbreak. The appearance of the mutation may have then triggered a second phase of rapid transmission. All current genome sequences are from this second phase and thus show limited diversity.

2. Adaptation to a non-human host prior to the jump to humans. This mutation is not seen in any bat coronavirus and is thus unlikely to be adaptive in those species.

Thoughts on 1: is it likely to spontaneously appear in a relatively short amount of time (and presumably small number of infections). It didn't happen in SARS with 8000 infections over 6 months. The link to the market would then be spurious - some doubt on that already. Prediction would be that the animal/environmental samples apparently found by China CDC would not have cleavage site.

Thoughts on 2: can we suggest a host where this cleavage site would likely be advantageous. Ferrets/polecats? Rodents - bamboo rats (don't know if they are popular in China)? Circulating in wild populations so limited prior human exposure until infected individual brought to the market.

February 3, 2020 11:59pm

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. Ge, X.-Y. *et al.* Isolation and characterization of a bat SARS-like coronavirus that uses the ACE2 receptor. *Nature* **503**, 535–538 (2013).
12. 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).
13. 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).
14. Yang, X.-L. *et al.* Isolation and Characterization of a Novel Bat Coronavirus Closely Related to the

February 3, 2020 11:59pm

Direct Progenitor of Severe Acute Respiratory Syndrome Coronavirus. *J. Virol.* **90**, 3253–3256 (2015).

February 3, 2020 11:59pm

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. As rumours have been circulating about this virus being engineered or otherwise created with intent, we wish to make it clear that our analyses show that such scenarios are largely incompatible with the data.

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 closely related viruses, including 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.

February 6, 2020 3:09am

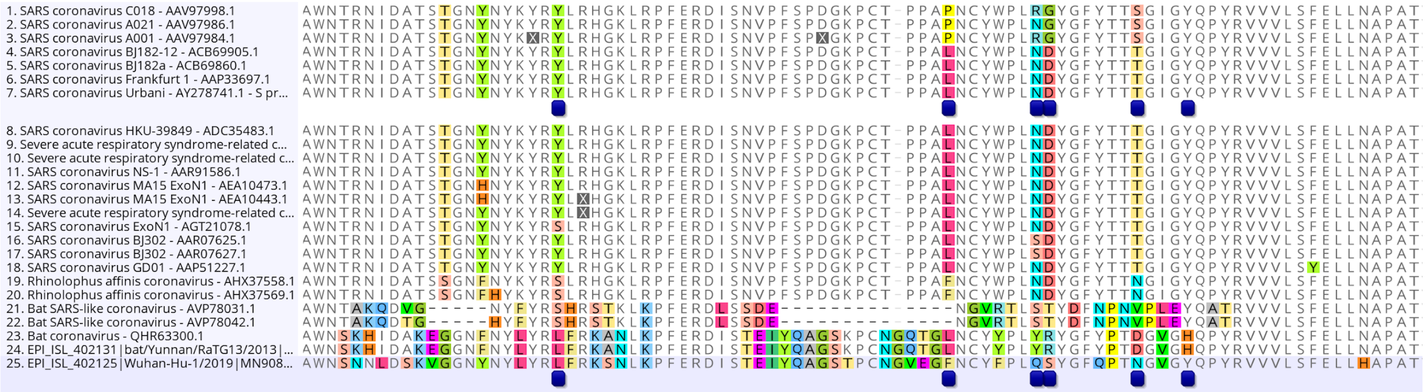

1. SARS coronavirus C018 - AAV97998.1
2. SARS coronavirus A021 - AAV97986.1
3. SARS coronavirus A001 - AAV97984.1
4. SARS coronavirus BJ182-12 - ACB69905.1
5. SARS coronavirus BJ182a - ACB69860.1
6. SARS coronavirus Frankfurt 1 - AAP33697.1
7. SARS coronavirus Urbani - AY278741.1 - S pr...
8. SARS coronavirus HKU-39849 - ADC35483.1
9. Severe acute respiratory syndrome-related c...
10. Severe acute respiratory syndrome-related c...
11. SARS coronavirus NS-1 - AAR91586.1
12. SARS coronavirus MA15 ExoN1 - AEA10473.1
13. SARS coronavirus MA15 ExoN1 - AEA10443.1
14. Severe acute respiratory syndrome-related c...
15. SARS coronavirus ExoN1 - AGT21078.1
16. SARS coronavirus BJ302 - AAR07625.1
17. SARS coronavirus BJ302 - AAR07627.1
18. SARS coronavirus GD01 - AAP51227.1
19. Rhinolophus affinis coronavirus - AHX37558.1
20. Rhinolophus affinis coronavirus - AHX37569.1
21. Bat SARS-like coronavirus - AVP78031.1
22. Bat SARS-like coronavirus - AVP78042.1
23. Bat coronavirus - QHR63300.1
24. EPI_ISL_402131|bat/Yunnan/RaTG13/2013|...
25. EPI_ISL_402125|Wuhan-Hu-1/2019|MN908...

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 (<u>RRA</u>R), 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.

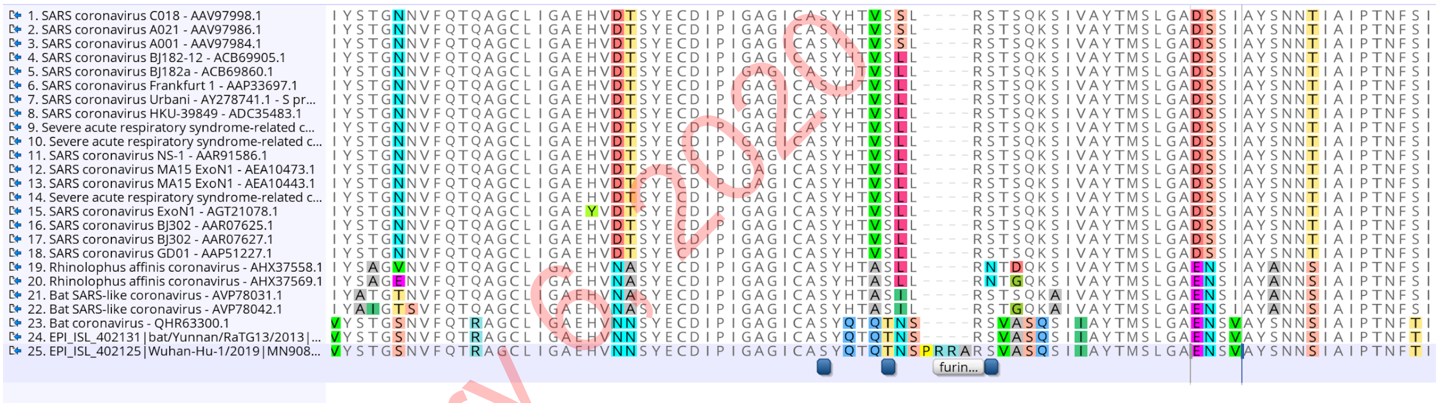

1. SARS coronavirus C018 - AAV97998.1
2. SARS coronavirus A021 - AAV97986.1
3. SARS coronavirus A001 - AAV97984.1
4. SARS coronavirus BJ182-12 - ACB69905.1
5. SARS coronavirus BJ182a - ACB69860.1
6. SARS coronavirus Frankfurt 1 - AAP33697.1
7. SARS coronavirus Urbani - AY278741.1 - S pr...
8. SARS coronavirus HKU-39849 - ADC35483.1
9. Severe acute respiratory syndrome-related c...
10. Severe acute respiratory syndrome-related c...
11. SARS coronavirus NS-1 - AAR91586.1
12. SARS coronavirus MA15 ExoN1 - AEA10473.1
13. SARS coronavirus MA15 ExoN1 - AEA10443.1
14. Severe acute respiratory syndrome-related c...
15. SARS coronavirus ExoN1 - AGT21078.1
16. SARS coronavirus BJ302 - AAR07625.1
17. SARS coronavirus BJ302 - AAR07627.1
18. SARS coronavirus GD01 - AAP51227.1
19. Rhinolophus affinis coronavirus - AHX37558.1
20. Rhinolophus affinis coronavirus - AHX37569.1
21. Bat SARS-like coronavirus - AVP78031.1
22. Bat SARS-like coronavirus - AVP78042.1
23. Bat coronavirus - QHR63300.1
24. EPI_ISL_402131|bat/Yunnan/RaTG13/2013|...
25. EPI_ISL_402125|Wuhan-Hu-1/2019|MN908...
furin...

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.

Evolution of 2019-nCoV

As described in the beginning, we believe deliberate engineering can be ruled out with a high degree of confidence as the data is inconsistent with this scenario. In addition, if engineering would have been

February 3:09am

performed, one would also 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, including those recently posited by various conspiracy theories, based on a 2015 paper in *Nature Medicine*[10].

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.

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.,[11–14]). 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.

February 6, 2020 3:09am

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.

February 6, 2020 3:09am

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. Ge, X.-Y. *et al.* Isolation and characterization of a bat SARS-like coronavirus that uses the ACE2 receptor. *Nature* **503**, 535–538 (2013).

12. 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).

13. 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).

February 6, 2020 3:09am

14. 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).

February 6, 2020 3:09am

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. As rumours have been circulating about this virus being engineered or otherwise created with intent, we wish to make it clear that our analyses show that such scenarios are largely incompatible with the data.

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 closely related viruses, including 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.

First Distributed Version

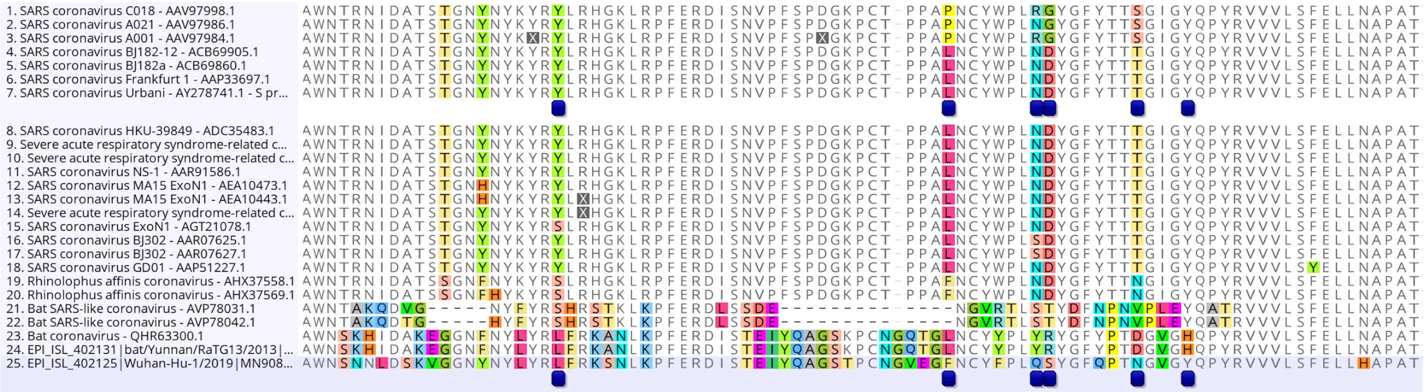

1. SARS coronavirus C018 - AAV97998.1
2. SARS coronavirus A021 - AAV97986.1
3. SARS coronavirus A001 - AAV97984.1
4. SARS coronavirus BJ182-12 - ACB69905.1
5. SARS coronavirus BJ182a - ACB69860.1
6. SARS coronavirus Frankfurt 1 - AAP33697.1
7. SARS coronavirus Urbani - AY278741.1 - S pr...
8. SARS coronavirus HKU-39849 - ADC35483.1
9. Severe acute respiratory syndrome-related c...
10. Severe acute respiratory syndrome-related c...
11. SARS coronavirus NS-1 - AAR91586.1
12. SARS coronavirus MA15 ExoN1 - AEA10473.1
13. SARS coronavirus MA15 ExoN1 - AEA10443.1
14. Severe acute respiratory syndrome-related c...
15. SARS coronavirus ExoN1 - AGT21078.1
16. SARS coronavirus BJ302 - AAR07625.1
17. SARS coronavirus BJ302 - AAR07627.1
18. SARS coronavirus GD01 - AAP51227.1
19. Rhinolophus affinis coronavirus - AHX37558.1
20. Rhinolophus affinis coronavirus - AHX37569.1
21. Bat SARS-like coronavirus - AVP78031.1
22. Bat SARS-like coronavirus - AVP78042.1
23. Bat coronavirus - QHR63300.1
24. EPI_ISL_402131|bat/Yunnan/RaTG13/2013|...
25. EPI_ISL_402125|Wuhan-Hu-1/2019|MN908...

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.

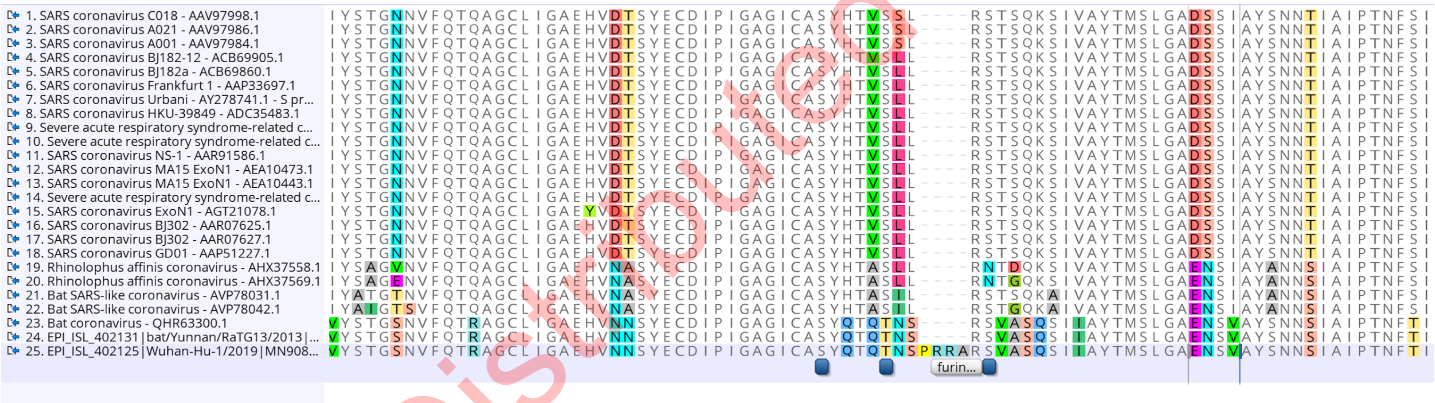

1. SARS coronavirus C018 - AAV97998.1
2. SARS coronavirus A021 - AAV97986.1
3. SARS coronavirus A001 - AAV97984.1
4. SARS coronavirus BJ182-12 - ACB69905.1
5. SARS coronavirus BJ182a - ACB69860.1
6. SARS coronavirus Frankfurt 1 - AAP33697.1
7. SARS coronavirus Urbani - AY278741.1 - S pr...
8. SARS coronavirus HKU-39849 - ADC35483.1
9. Severe acute respiratory syndrome-related c...
10. Severe acute respiratory syndrome-related c...
11. SARS coronavirus NS-1 - AAR91586.1
12. SARS coronavirus MA15 ExoN1 - AEA10473.1
13. SARS coronavirus MA15 ExoN1 - AEA10443.1
14. Severe acute respiratory syndrome-related c...
15. SARS coronavirus ExoN1 - AGT21078.1
16. SARS coronavirus BJ302 - AAR07625.1
17. SARS coronavirus BJ302 - AAR07627.1
18. SARS coronavirus GD01 - AAP51227.1
19. Rhinolophus affinis coronavirus - AHX37558.1
20. Rhinolophus affinis coronavirus - AHX37569.1
21. Bat SARS-like coronavirus - AVP78031.1
22. Bat SARS-like coronavirus - AVP78042.1
23. Bat coronavirus - QHR63300.1
24. EPI_ISL_402131|bat/Yunnan/RaTG13/2013|...
25. EPI_ISL_402125|Wuhan-Hu-1/2019|MN908...
furin...

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.

Evolution of 2019-nCoV

As described in the beginning, we believe deliberate engineering can be ruled out with a high degree of confidence as the data is inconsistent with this scenario. In addition, if engineering would have been

First Distributed Version

performed, one would also 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, including those recently posited by various conspiracy theories, based on a 2015 paper in *Nature Medicine*[10].

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.

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.,[11–14]). 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.

First Distributed Version

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.

First Distributed Version

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. Ge, X.-Y. *et al.* Isolation and characterization of a bat SARS-like coronavirus that uses the ACE2 receptor. *Nature* **503**, 535–538 (2013).
12. 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).
13. 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).

First Distributed Version

14. 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).

First Distributed Version

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.

Second Distributed Version

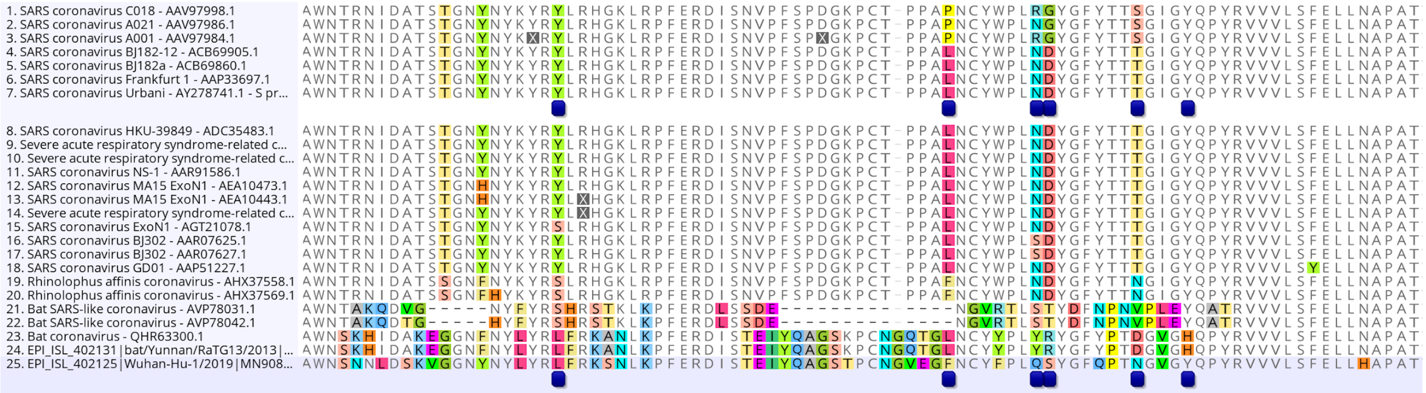

1. SARS coronavirus C018 - AAV97998.1
2. SARS coronavirus A021 - AAV97986.1
3. SARS coronavirus A001 - AAV97984.1
4. SARS coronavirus BJ182-12 - ACB69905.1
5. SARS coronavirus BJ182a - ACB69860.1
6. SARS coronavirus Frankfurt 1 - AAP33697.1
7. SARS coronavirus Urbani - AY278741.1 - S pr...
8. SARS coronavirus HKU-39849 - ADC35483.1
9. Severe acute respiratory syndrome-related c...
10. Severe acute respiratory syndrome-related c...
11. SARS coronavirus NS-1 - AAR91586.1
12. SARS coronavirus MA15 ExoN1 - AEA10473.1
13. SARS coronavirus MA15 ExoN1 - AEA10443.1
14. Severe acute respiratory syndrome-related c...
15. SARS coronavirus ExoN1 - AGT21078.1
16. SARS coronavirus BJ302 - AAR07625.1
17. SARS coronavirus BJ302 - AAR07627.1
18. SARS coronavirus GD01 - AAP51227.1
19. Rhinolophus affinis coronavirus - AHX37558.1
20. Rhinolophus affinis coronavirus - AHX37569.1
21. Bat SARS-like coronavirus - AVP78031.1
22. Bat SARS-like coronavirus - AVP78042.1
23. Bat coronavirus - QHR63300.1
24. EPI_ISL_402131|bat/Yunnan/RaTG13/2013|...
25. EPI_ISL_402125|Wuhan-Hu-1/2019|MN908...

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.

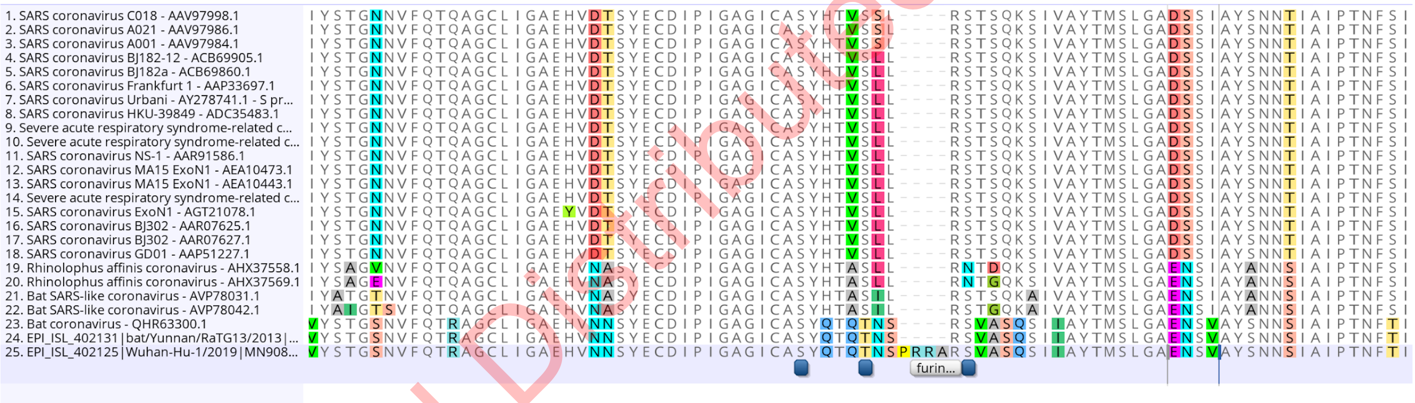

1. SARS coronavirus C018 - AAV97998.1
2. SARS coronavirus A021 - AAV97986.1
3. SARS coronavirus A001 - AAV97984.1
4. SARS coronavirus BJ182-12 - ACB69905.1
5. SARS coronavirus BJ182a - ACB69860.1
6. SARS coronavirus Frankfurt 1 - AAP33697.1
7. SARS coronavirus Urbani - AY278741.1 - S pr...
8. SARS coronavirus HKU-39849 - ADC35483.1
9. Severe acute respiratory syndrome-related c...
10. Severe acute respiratory syndrome-related c...
11. SARS coronavirus NS-1 - AAR91586.1
12. SARS coronavirus MA15 ExoN1 - AEA10473.1
13. SARS coronavirus MA15 ExoN1 - AEA10443.1
14. Severe acute respiratory syndrome-related c...
15. SARS coronavirus ExoN1 - AGT21078.1
16. SARS coronavirus BJ302 - AAR07625.1
17. SARS coronavirus BJ302 - AAR07627.1
18. SARS coronavirus GD01 - AAP51227.1
19. Rhinolophus affinis coronavirus - AHX37558.1
20. Rhinolophus affinis coronavirus - AHX37569.1
21. Bat SARS-like coronavirus - AVP78031.1
22. Bat SARS-like coronavirus - AVP78042.1
23. Bat coronavirus - QHR63300.1
24. EPI_ISL_402131|bat/Yunnan/RaTG13/2013|...
25. EPI_ISL_402125|Wuhan-Hu-1/2019|MN908...
furin...

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

Second Draft Version

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.

Second Distributed Version

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.

Second Distributed Version

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).

Second Distributed Version

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).

Second Distributed Version

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.

February 8, 2020 5:39 AM

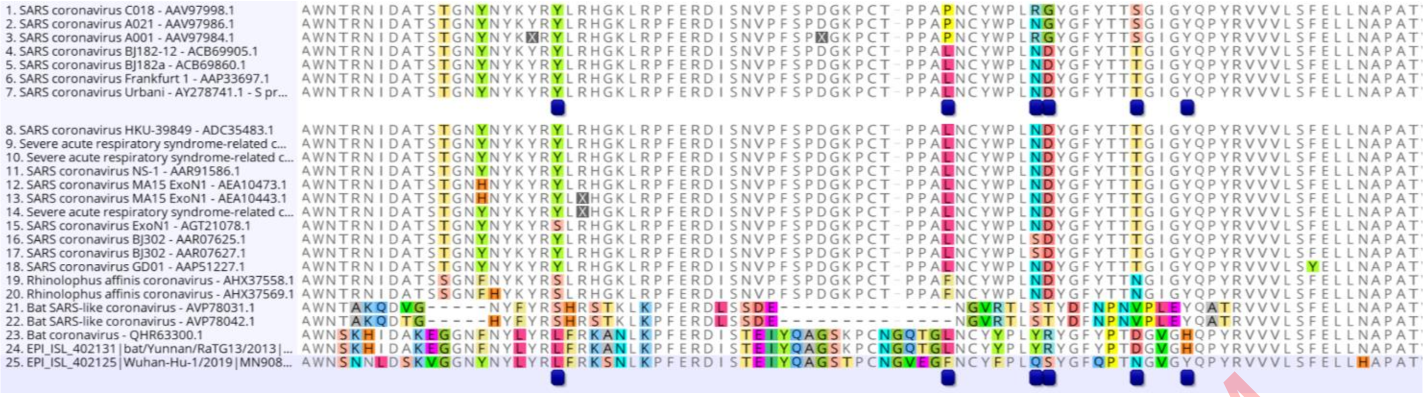

1. SARS coronavirus C018 - AAV97998.1
2. SARS coronavirus A021 - AAV97986.1
3. SARS coronavirus A001 - AAV97984.1
4. SARS coronavirus BJ182-12 - ACB69905.1
5. SARS coronavirus BJ182a - ACB69860.1
6. SARS coronavirus Frankfurt 1 - AAP33697.1
7. SARS coronavirus Urbani - AY278741.1 - S pr...
8. SARS coronavirus HKU-39849 - ADC35483.1
9. Severe acute respiratory syndrome-related c...
10. Severe acute respiratory syndrome-related c...
11. SARS coronavirus NS-1 - AAR91586.1
12. SARS coronavirus MA15 ExoN1 - AEA10473.1
13. SARS coronavirus MA15 ExoN1 - AEA10443.1
14. Severe acute respiratory syndrome-related c...
15. SARS coronavirus ExoN1 - AGT21078.1
16. SARS coronavirus BJ302 - AAR07625.1
17. SARS coronavirus BJ302 - AAR07627.1
18. SARS coronavirus GD01 - AAP51227.1
19. Rhinolophus affinis coronavirus - AHX37558.1
20. Rhinolophus affinis coronavirus - AHX37569.1
21. Bat SARS-like coronavirus - AVP78031.1
22. Bat SARS-like coronavirus - AVP78042.1
23. Bat coronavirus - QHR63300.1
24. EPI_ISL_402131|bat/Yunnan/RaTG13/2013|...
25. EPI_ISL_402125|Wuhan-Hu-1/2019|MN908...

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.

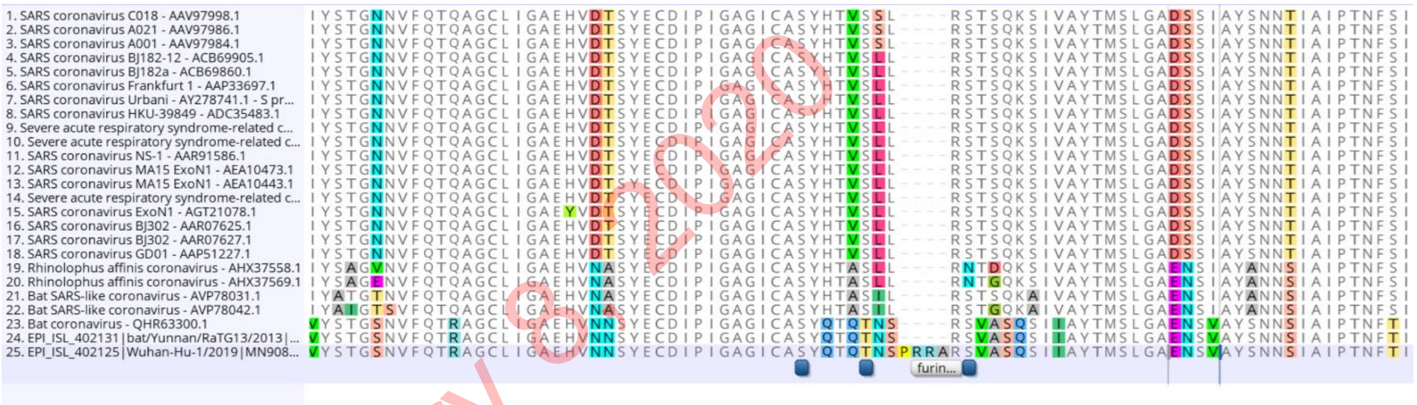

1. SARS coronavirus C018 - AAV97998.1
2. SARS coronavirus A021 - AAV97986.1
3. SARS coronavirus A001 - AAV97984.1
4. SARS coronavirus BJ182-12 - ACB69905.1
5. SARS coronavirus BJ182a - ACB69860.1
6. SARS coronavirus Frankfurt 1 - AAP33697.1
7. SARS coronavirus Urbani - AY278741.1 - S pr...
8. SARS coronavirus HKU-39849 - ADC35483.1
9. Severe acute respiratory syndrome-related c...
10. Severe acute respiratory syndrome-related c...
11. SARS coronavirus NS-1 - AAR91586.1
12. SARS coronavirus MA15 ExoN1 - AEA10473.1
13. SARS coronavirus MA15 ExoN1 - AEA10443.1
14. Severe acute respiratory syndrome-related c...
15. SARS coronavirus ExoN1 - AGT21078.1
16. SARS coronavirus BJ302 - AAR07625.1
17. SARS coronavirus BJ302 - AAR07627.1
18. SARS coronavirus GD01 - AAP51227.1
19. Rhinolophus affinis coronavirus - AHX37558.1
20. Rhinolophus affinis coronavirus - AHX37569.1
21. Bat SARS-like coronavirus - AVP78031.1
22. Bat SARS-like coronavirus - AVP78042.1
23. Bat coronavirus - QHR63300.1
24. EPI_ISL_402131|bat/Yunnan/RaTG13/2013|...
25. EPI_ISL_402125|Wuhan-Hu-1/2019|MN908...
furin...

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

February 8, 2020 5:39 AM

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.

February 8, 2020 5:39 AM

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.

February 8, 2020 5:39 AM

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).

February 8, 2020 5:39 AM

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).

February 8, 2020 5:39 AM

The proximal origin of SARS-CoV-2

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.

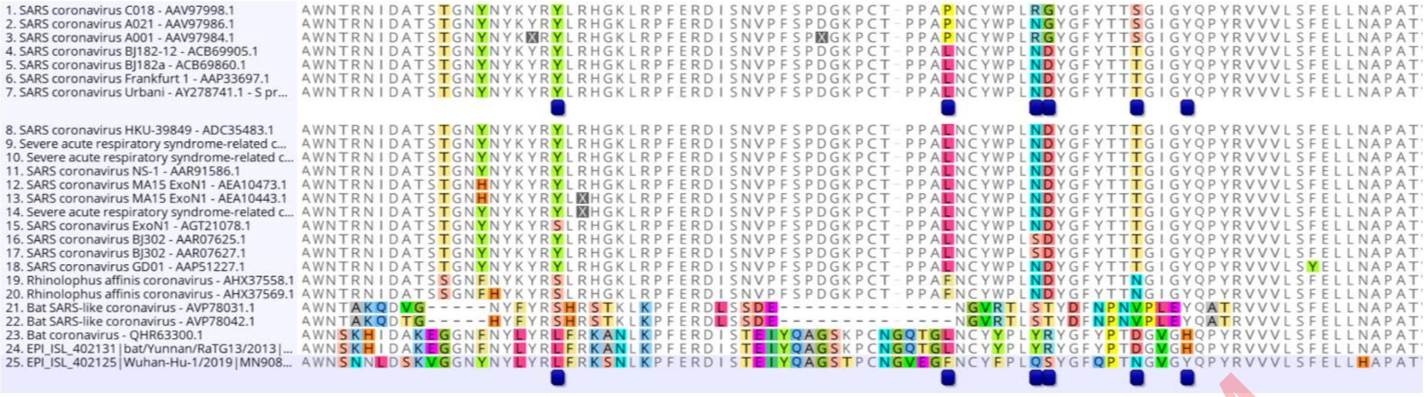
Mutations in the receptor binding domain of 2019-nCoV

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.

February 8, 2020, 5_45 AM



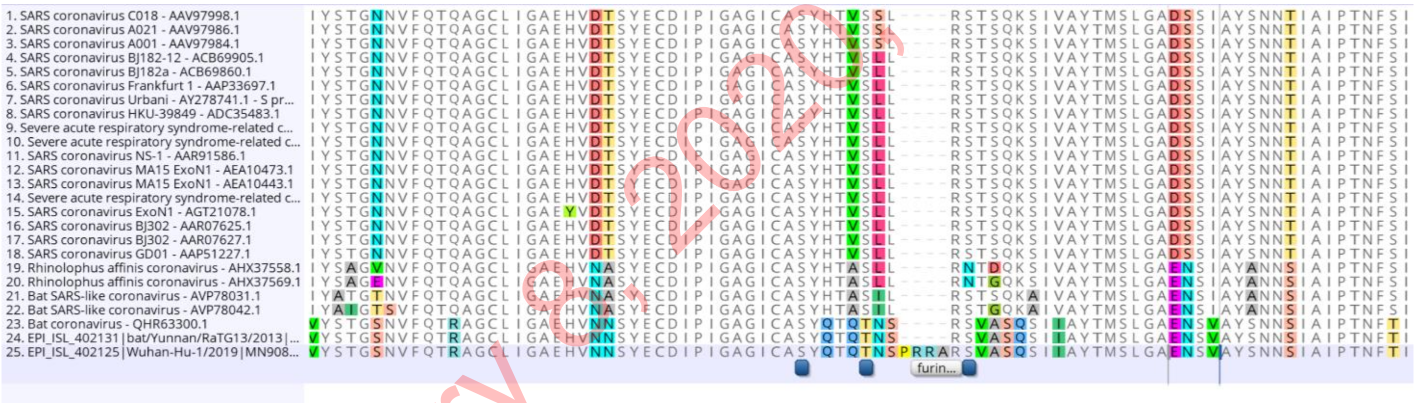
```
1. SARS coronavirus C018 - AAV97998.1                 AWNTRNIDATSTGNYNYKYRYLRHGKLRPFERDISNVPFSPDGKPCT-PPAPNCYWPLRGYGFYTTSGIGYQPYRVVVLSFELLNAPAT
2. SARS coronavirus A021 - AAV97986.1                 AWNTRNIDATSTGNYNYKYRYLRHGKLRPFERDISNVPFSPDGKPCT-PPAPNCYWPLNGYGFYTTSGIGYQPYRVVVLSFELLNAPAT
3. SARS coronavirus A001 - AAV97984.1                 AWNTRNIDATSTGNYNYKXRYLRHGKLRPFERDISNVPFSPXGKPCT-PPAPNCYWPLRGYGFYTTSGIGYQPYRVVVLSFELLNAPAT
4. SARS coronavirus BJ182-12 - ACB69905.1             AWNTRNIDATSTGNYNYKYRYLRHGKLRPFERDISNVPFSPDGKPCT-PPALNCYWPLNDYGFYTTTGIGYQPYRVVVLSFELLNAPAT
5. SARS coronavirus BJ182a - ACB69860.1               AWNTRNIDATSTGNYNYKYRYLRHGKLRPFERDISNVPFSPDGKPCT-PPALNCYWPLNDYGFYTTTGIGYQPYRVVVLSFELLNAPAT
6. SARS coronavirus Frankfurt 1 - AAP33697.1          AWNTRNIDATSTGNYNYKYRYLRHGKLRPFERDISNVPFSPDGKPCT-PPALNCYWPLNDYGFYTTTGIGYQPYRVVVLSFELLNAPAT
7. SARS coronavirus Urbani - AY278741.1 - S pr...     AWNTRNIDATSTGNYNYKYRYLRHGKLRPFERDISNVPFSPDGKPCT-PPALNCYWPLNDYGFYTTTGIGYQPYRVVVLSFELLNAPAT

8. SARS coronavirus HKU-39849 - ADC35483.1            AWNTRNIDATSTGNYNYKYRYLRHGKLRPFERDISNVPFSPDGKPCT-PPALNCYWPLNDYGFYTTTGIGYQPYRVVVLSFELLNAPAT
9. Severe acute respiratory syndrome-related c...     AWNTRNIDATSTGNYNYKYRYLRHGKLRPFERDISNVPFSPDGKPCT-PPALNCYWPLNDYGFYTTTGIGYQPYRVVVLSFELLNAPAT
10. Severe acute respiratory syndrome-related c...    AWNTRNIDATSTGNYNYKYRYLRHGKLRPFERDISNVPFSPDGKPCT-PPALNCYWPLNDYGFYTTTGIGYQPYRVVVLSFELLNAPAT
11. SARS coronavirus NS-1 - AAR91586.1                AWNTRNIDATSTGNYNYKYRYLRHGKLRPFERDISNVPFSPDGKPCT-PPALNCYWPLNDYGFYTTTGIGYQPYRVVVLSFELLNAPAT
12. SARS coronavirus MA15 ExoN1 - AEA10473.1          AWNTRNIDATSTGNHNYKYRYLRHGKLRPFERDISNVPFSPDGKPCT-PPALNCYWPLNDYGFYTTTGIGYQPYRVVVLSFELLNAPAT
13. SARS coronavirus MA15 ExoN1 - AEA10443.1          AWNTRNIDATSTGNHNYKYRYLXHGKLRPFERDISNVPFSPDGKPCT-PPALNCYWPLNDYGFYTTTGIGYQPYRVVVLSFELLNAPAT
14. Severe acute respiratory syndrome-related c...    AWNTRNIDATSTGNYNYKYRYLXHGKLRPFERDISNVPFSPDGKPCT-PPALNCYWPLNDYGFYTTTGIGYQPYRVVVLSFELLNAPAT
15. SARS coronavirus ExoN1 - AGT21078.1               AWNTRNIDATSTGNYNYKYRSLRHGKLRPFERDISNVPFSPDGKPCT-PPALNCYWPLNDYGFYTTTGIGYQPYRVVVLSFELLNAPAT
16. SARS coronavirus BJ302 - AAR07625.1               AWNTRNIDATSTGNYNYKYRYLRHGKLRPFERDISNVPFSPDGKPCT-PPALNCYWPLSDYGFYTTTGIGYQPYRVVVLSFELLNAPAT
17. SARS coronavirus BJ302 - AAR07627.1               AWNTRNIDATSTGNYNYKYRYLRHGKLRPFERDISNVPFSPDGKPCT-PPALNCYWPLSDYGFYTTTGIGYQPYRVVVLSFELLNAPAT
18. SARS coronavirus GD01 - AAP51227.1                AWNTRNIDATSTGNYNYKYRYLRHGKLRPFERDISNVPFSPDGKPCT-PPALNCYWPLNDYGFYTTTGIGYQPYRVVVLSYELLNAPAT
19. Rhinolophus affinis coronavirus - AHX37558.1      AWNTRNIDATSSGNFNYKYRSLRHGKLRPFERDISNVPFSPDGKPCT-PPAFNCYWPLNDYGFYTTNGIGYQPYRVVVLSFELLNAPAT
20. Rhinolophus affinis coronavirus - AHX37569.1      AWNTRNIDATSSGNFHYKYRSLRHGKLRPFERDISNVPFSPDGKPCT-PPAFNCYWPLNDYGFYTTNGIGYQPYRVVVLSFELLNAPAT
21. Bat SARS-like coronavirus - AVP78031.1            AWNTAKQDVG-----NYFYRSHRSTKLKPFERDLSSDE--------------NGVRTLSTYDFNPNVPLEYQATRVVVLSFELLNAPAT
22. Bat SARS-like coronavirus - AVP78042.1            AWNTAKQDTG-----HYFYRSHRSTKLKPFERDLSSDE--------------NGVRTLSTYDFNPNVPLEYQATRVVVLSFELLNAPAT
23. Bat coronavirus - QHR63300.1                      AWNSKHIDAKEGGNFNYLYRLFRKANLKPFERDISTEIYQAGSKPCNGQTGLNCYYPLYRYGFYPTDGVGHQPYRVVVLSFELLNAPAT
24. EPI_ISL_402131|bat/Yunnan/RaTG13/2013|...         AWNSKHIDAKEGGNFNYLYRLFRKANLKPFERDISTEIYQAGSKPCNGQTGLNCYYPLYRYGFYPTDGVGHQPYRVVVLSFELLNAPAT
25. EPI_ISL_402125|Wuhan-Hu-1/2019|MN908...           AWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVEGFNCYFPLQSYGFQPTNGVGYQPYRVVVLSFELLHAPAT
```

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.



```
1. SARS coronavirus C018 - AAV97998.1                 IYSTGNNVFQTQAGCLIGAEHVDTSYECDIPIGAGICASYHTVSSL----RSTSQKSIVAYTMSLGADSSIAYSNNTIAIPTNFSI
2. SARS coronavirus A021 - AAV97986.1                 IYSTGNNVFQTQAGCLIGAEHVDTSYECDIPIGAGICASYHTVSSL----RSTSQKSIVAYTMSLGADSSIAYSNNTIAIPTNFSI
3. SARS coronavirus A001 - AAV97984.1                 IYSTGNNVFQTQAGCLIGAEHVDTSYECDIPIGAGICASYHTVSSL----RSTSQKSIVAYTMSLGADSSIAYSNNTIAIPTNFSI
4. SARS coronavirus BJ182-12 - ACB69905.1             IYSTGNNVFQTQAGCLIGAEHVDTSYECDIPIGAGICASYHTVSLL----RSTSQKSIVAYTMSLGADSSIAYSNNTIAIPTNFSI
5. SARS coronavirus BJ182a - ACB69860.1               IYSTGNNVFQTQAGCLIGAEHVDTSYECDIPIGAGICASYHTVSLL----RSTSQKSIVAYTMSLGADSSIAYSNNTIAIPTNFSI
6. SARS coronavirus Frankfurt 1 - AAP33697.1          IYSTGNNVFQTQAGCLIGAEHVDTSYECDIPIGAGICASYHTVSLL----RSTSQKSIVAYTMSLGADSSIAYSNNTIAIPTNFSI
7. SARS coronavirus Urbani - AY278741.1 - S pr...     IYSTGNNVFQTQAGCLIGAEHVDTSYECDIPIGAGICASYHTVSLL----RSTSQKSIVAYTMSLGADSSIAYSNNTIAIPTNFSI
8. SARS coronavirus HKU-39849 - ADC35483.1            IYSTGNNVFQTQAGCLIGAEHVDTSYECDIPIGAGICASYHTVSLL----RSTSQKSIVAYTMSLGADSSIAYSNNTIAIPTNFSI
9. Severe acute respiratory syndrome-related c...     IYSTGNNVFQTQAGCLIGAEHVDTSYECDIPIGAGICASYHTVSLL----RSTSQKSIVAYTMSLGADSSIAYSNNTIAIPTNFSI
10. Severe acute respiratory syndrome-related c...    IYSTGNNVFQTQAGCLIGAEHVDTSYECDIPIGAGICASYHTVSLL----RSTSQKSIVAYTMSLGADSSIAYSNNTIAIPTNFSI
11. SARS coronavirus NS-1 - AAR91586.1                IYSTGNNVFQTQAGCLIGAEHVDTSYECDIPIGAGICASYHTVSLL----RSTSQKSIVAYTMSLGADSSIAYSNNTIAIPTNFSI
12. SARS coronavirus MA15 ExoN1 - AEA10473.1          IYSTGNNVFQTQAGCLIGAEHVDTSYECDIPIGAGICASYHTVSLL----RSTSQKSIVAYTMSLGADSSIAYSNNTIAIPTNFSI
13. SARS coronavirus MA15 ExoN1 - AEA10443.1          IYSTGNNVFQTQAGCLIGAEHVDTSYECDIPIGAGICASYHTVSLL----RSTSQKSIVAYTMSLGADSSIAYSNNTIAIPTNFSI
14. Severe acute respiratory syndrome-related c...    IYSTGNNVFQTQAGCLIGAEHVDTSYECDIPIGAGICASYHTVSLL----RSTSQKSIVAYTMSLGADSSIAYSNNTIAIPTNFSI
15. SARS coronavirus ExoN1 - AGT21078.1               IYSTGNNVFQTQAGCLIGAEYVDTSYECDIPIGAGICASYHTVSLL----RSTSQKSIVAYTMSLGADSSIAYSNNTIAIPTNFSI
16. SARS coronavirus BJ302 - AAR07625.1               IYSTGNNVFQTQAGCLIGAEHVDTSYECDIPIGAGICASYHTVSLL----RSTSQKSIVAYTMSLGADSSIAYSNNTIAIPTNFSI
17. SARS coronavirus BJ302 - AAR07627.1               IYSTGNNVFQTQAGCLIGAEHVDTSYECDIPIGAGICASYHTVSLL----RSTSQKSIVAYTMSLGADSSIAYSNNTIAIPTNFSI
18. SARS coronavirus GD01 - AAP51227.1                IYSTGNNVFQTQAGCLIGAEHVDTSYECDIPIGAGICASYHTVSLL----RSTSQKSIVAYTMSLGADSSIAYSNNTIAIPTNFSI
19. Rhinolophus affinis coronavirus - AHX37558.1      IYSAGVNVFQTQAGCLIGAEHVNASYECDIPIGAGICASYHTASLL----RNTDQKSIVAYTMSLGAENSIAYANNSIAIPTNFSI
20. Rhinolophus affinis coronavirus - AHX37569.1      IYSAGENVFQTQAGCLIGAEHVNASYECDIPIGAGICASYHTASLL----RNTGQKSIVAYTMSLGAENSIAYANNSIAIPTNFSI
21. Bat SARS-like coronavirus - AVP78031.1            IYATGTNVFQTQAGCLIGAEHVNASYECDIPIGAGICASYHTASIL----RSTSQKAIVAYTMSLGAENSIAYANNSIAIPTNFSI
22. Bat SARS-like coronavirus - AVP78042.1            IYAIGTSVFQTQAGCLIGAEHVNASYECDIPIGAGICASYHTASIL----RSTGQKAIVAYTMSLGAENSIAYANNSIAIPTNFSI
23. Bat coronavirus - QHR63300.1                      VYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICASYQTQTNS----RSVASQSIIAYTMSLGAENSVAYSNNSIAIPTNFTI
24. EPI_ISL_402131|bat/Yunnan/RaTG13/2013|...         VYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICASYQTQTNS----RSVASQSIIAYTMSLGAENSVAYSNNSIAIPTNFTI
25. EPI_ISL_402125|Wuhan-Hu-1/2019|MN908...           VYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICASYQTQTNSPRRARSVASQSIIAYTMSLGAENSVAYSNNSIAIPTNFTI
```

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

February 8, 2020, 5:45 AM

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.

February 8, 2020, 5:45 AM

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.

February 8, 2020, 5:45 AM

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).

February 8, 2020, 5_45 AM

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).

February 8, 2020, 5_45 AM

The proximal origin of 2019-nCoV

Since the first reports of a novel pneumonia in Wuhan city, Hubei province, China there has been considerable uncertainty and discussion over the possible origin of the causative virus, 2019-nCoV. Herein, we review what can be deduced about the origin of this virus from the comparative analysis of available genome sequence data. In particular, we describe notable features in the 2019-nCoV genome, outline mechanisms for how the virus may have acquired them, and discuss scenarios by which these features could have arisen. Importantly, this analysis clearly demonstrates that 2019-nCoV is not a laboratory construct or experimentally manipulated virus.

Genomic comparisons identified two features of the 2019-nCoV genome of note: (i) based on structural modeling and early biochemical experiments, 2019-nCoV appears to be optimized for binding to the human ACE2 receptor; (ii) thighly 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.

February 10, 2020 8:54 PM

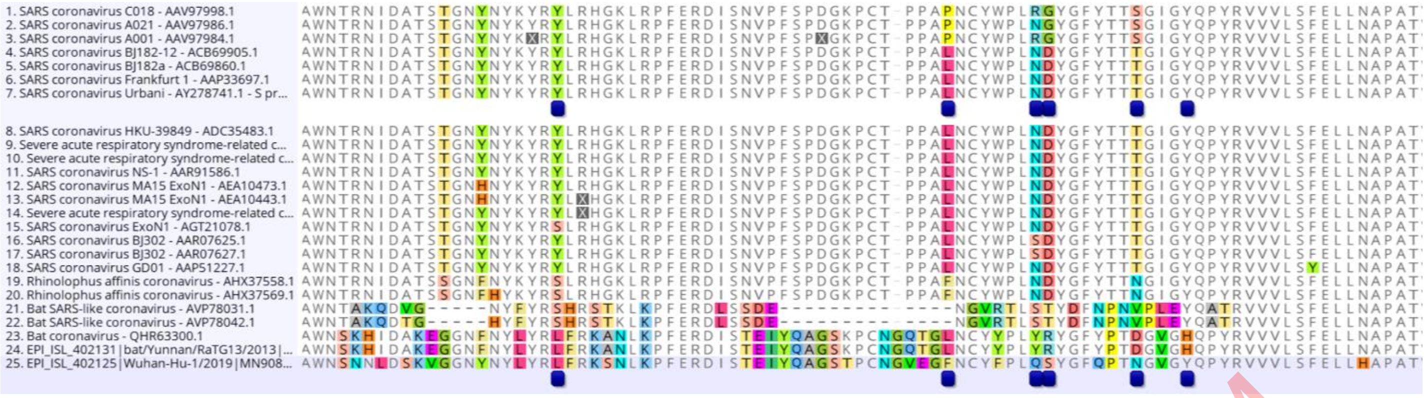

1. SARS coronavirus C018 - AAV97998.1
2. SARS coronavirus A021 - AAV97986.1
3. SARS coronavirus A001 - AAV97984.1
4. SARS coronavirus BJ182-12 - ACB69905.1
5. SARS coronavirus BJ182a - ACB69860.1
6. SARS coronavirus Frankfurt 1 - AAP33697.1
7. SARS coronavirus Urbani - AY278741.1 - S pr...
8. SARS coronavirus HKU-39849 - ADC35483.1
9. Severe acute respiratory syndrome-related c...
10. Severe acute respiratory syndrome-related c...
11. SARS coronavirus NS-1 - AAR91586.1
12. SARS coronavirus MA15 ExoN1 - AEA10473.1
13. SARS coronavirus MA15 ExoN1 - AEA10443.1
14. Severe acute respiratory syndrome-related c...
15. SARS coronavirus ExoN1 - AGT21078.1
16. SARS coronavirus BJ302 - AAR07625.1
17. SARS coronavirus BJ302 - AAR07627.1
18. SARS coronavirus GD01 - AAP51227.1
19. Rhinolophus affinis coronavirus - AHX37558.1
20. Rhinolophus affinis coronavirus - AHX37569.1
21. Bat SARS-like coronavirus - AVP78031.1
22. Bat SARS-like coronavirus - AVP78042.1
23. Bat coronavirus - QHR63300.1
24. EPI_ISL_402131|bat/Yunnan/RaTG13/2013|...
25. EPI_ISL_402125|Wuhan-Hu-1/2019|MN908...

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

The second notable 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.

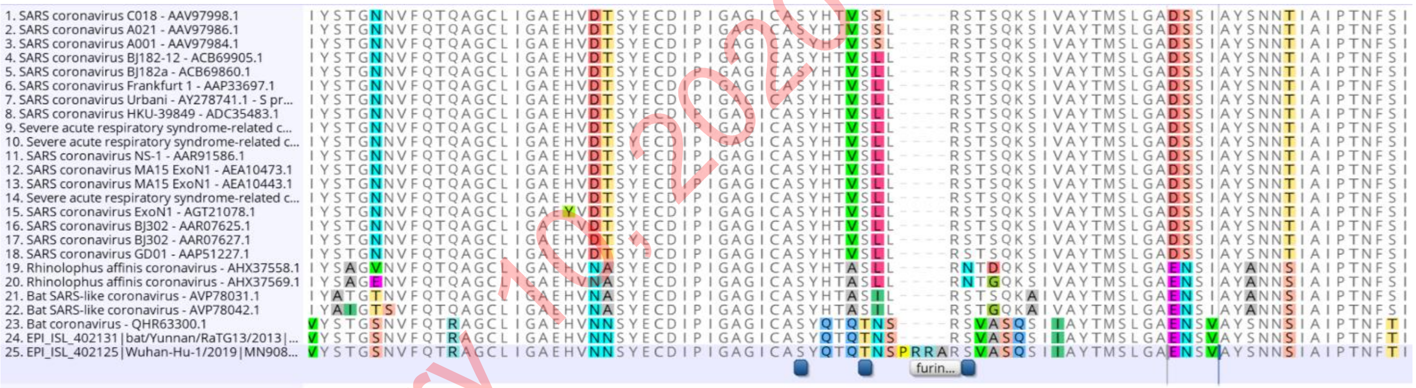

1. SARS coronavirus C018 - AAV97998.1
2. SARS coronavirus A021 - AAV97986.1
3. SARS coronavirus A001 - AAV97984.1
4. SARS coronavirus BJ182-12 - ACB69905.1
5. SARS coronavirus BJ182a - ACB69860.1
6. SARS coronavirus Frankfurt 1 - AAP33697.1
7. SARS coronavirus Urbani - AY278741.1 - S pr...
8. SARS coronavirus HKU-39849 - ADC35483.1
9. Severe acute respiratory syndrome-related c...
10. Severe acute respiratory syndrome-related c...
11. SARS coronavirus NS-1 - AAR91586.1
12. SARS coronavirus MA15 ExoN1 - AEA10473.1
13. SARS coronavirus MA15 ExoN1 - AEA10443.1
14. Severe acute respiratory syndrome-related c...
15. SARS coronavirus ExoN1 - AGT21078.1
16. SARS coronavirus BJ302 - AAR07625.1
17. SARS coronavirus BJ302 - AAR07627.1
18. SARS coronavirus GD01 - AAP51227.1
19. Rhinolophus affinis coronavirus - AHX37558.1
20. Rhinolophus affinis coronavirus - AHX37569.1
21. Bat SARS-like coronavirus - AVP78031.1
22. Bat SARS-like coronavirus - AVP78042.1
23. Bat coronavirus - QHR63300.1
24. EPI_ISL_402131|bat/Yunnan/RaTG13/2013|...
25. EPI_ISL_402125|Wuhan-Hu-1/2019|MN908...
furin...

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.

February 10, 2020 8:54 PM

Theories of 2019-nCoV origins

As noted above 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 available 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 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 therefore theoretically 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.

Selection in an animal host

Given the similarity of 2019-nCoV to bat SARS-like CoVs, particularly RaTG13, it is 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 probable 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. Although analyses are ongoing, the pangolin viruses described to date 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

February 10, 2020 8:54 PM

sarbecoviruses before, it is unclear what conditions would be required for it to be acquired in the lineage leading to 2019-nCoV.

Conclusions

The genomic features described here, which may in part explain the infectiousness and transmissibility of 2019-nCoV in humans, could have arisen through a process of adaptive evolution prior to the start of the outbreak. Although we can readily dismiss the idea that 2019-nCoV is an experimentally manipulated virus, it is impossible to prove or disprove other theories of its origin, 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 revealing virus origins. In addition, it would be helpful to obtain additional genetic and functional data about the virus, including 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, would also be informative. 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.

February 10, 2020 8:54 PM

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).

February 10, 2020 8:54 PM

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).

February 10, 2020 8:54 PM

The proximal origin of 2019-nCoV

Since the first reports of a novel pneumonia in Wuhan city, Hubei province, China there has been considerable uncertainty and discussion over the possible origin of the causative virus, 2019-nCoV. Herein, we review what can be deduced about the origin of this virus from the comparative analysis of available genome sequence data. In particular, we describe notable features in the 2019-nCoV genome, outline mechanisms for how the virus may have acquired them, and discuss scenarios by which these features could have arisen. Importantly, this analysis clearly demonstrates that 2019-nCoV is not a laboratory construct or experimentally manipulated virus.

Genomic comparisons identified two features of the 2019-nCoV genome of note: (i) based on structural modeling and early biochemical experiments, 2019-nCoV appears to be optimized for binding to the human ACE2 receptor; (ii) thighly 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.

February 11, 2020 7:17 AM

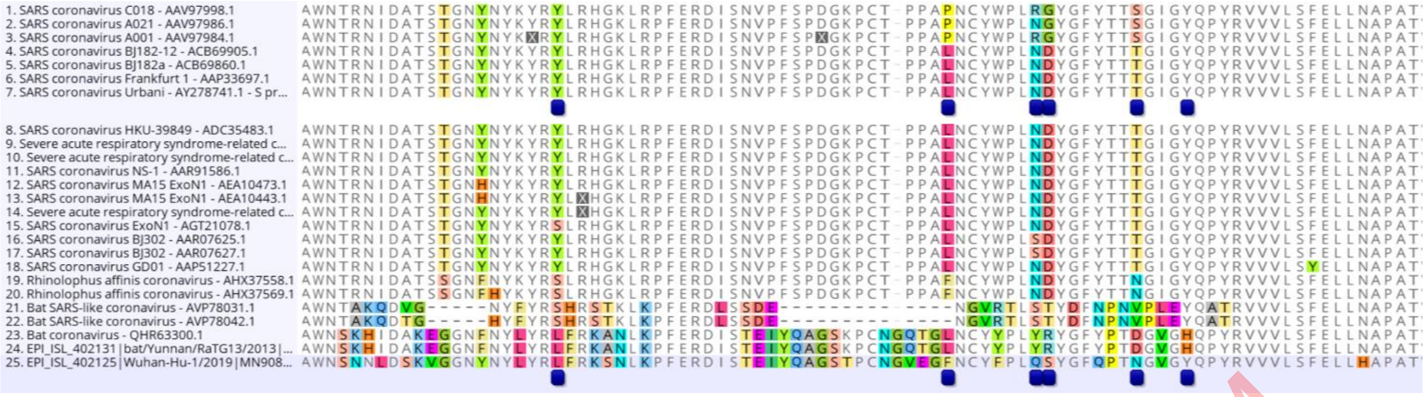

1. SARS coronavirus C018 - AAV97998.1
2. SARS coronavirus A021 - AAV97986.1
3. SARS coronavirus A001 - AAV97984.1
4. SARS coronavirus BJ182-12 - ACB69905.1
5. SARS coronavirus BJ182a - ACB69860.1
6. SARS coronavirus Frankfurt 1 - AAP33697.1
7. SARS coronavirus Urbani - AY278741.1 - S pr...
8. SARS coronavirus HKU-39849 - ADC35483.1
9. Severe acute respiratory syndrome-related c...
10. Severe acute respiratory syndrome-related c...
11. SARS coronavirus NS-1 - AAR91586.1
12. SARS coronavirus MA15 ExoN1 - AEA10473.1
13. SARS coronavirus MA15 ExoN1 - AEA10443.1
14. Severe acute respiratory syndrome-related c...
15. SARS coronavirus ExoN1 - AGT21078.1
16. SARS coronavirus BJ302 - AAR07625.1
17. SARS coronavirus BJ302 - AAR07627.1
18. SARS coronavirus GD01 - AAP51227.1
19. Rhinolophus affinis coronavirus - AHX37558.1
20. Rhinolophus affinis coronavirus - AHX37569.1
21. Bat SARS-like coronavirus - AVP78031.1
22. Bat SARS-like coronavirus - AVP78042.1
23. Bat coronavirus - QHR63300.1
24. EPI_ISL_402131|bat/Yunnan/RaTG13/2013|...
25. EPI_ISL_402125|Wuhan-Hu-1/2019|MN908...

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

The second notable 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.

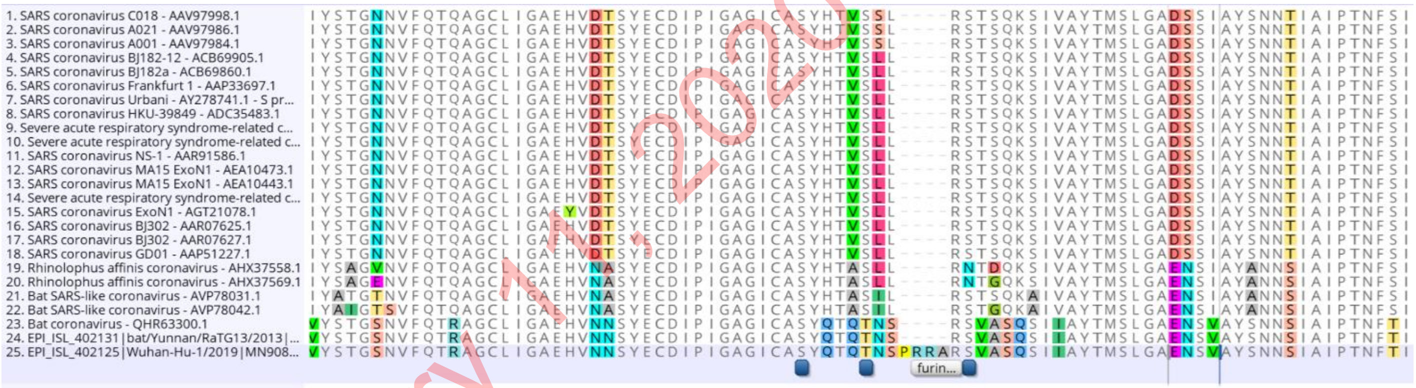

1. SARS coronavirus C018 - AAV97998.1
2. SARS coronavirus A021 - AAV97986.1
3. SARS coronavirus A001 - AAV97984.1
4. SARS coronavirus BJ182-12 - ACB69905.1
5. SARS coronavirus BJ182a - ACB69860.1
6. SARS coronavirus Frankfurt 1 - AAP33697.1
7. SARS coronavirus Urbani - AY278741.1 - S pr...
8. SARS coronavirus HKU-39849 - ADC35483.1
9. Severe acute respiratory syndrome-related c...
10. Severe acute respiratory syndrome-related c...
11. SARS coronavirus NS-1 - AAR91586.1
12. SARS coronavirus MA15 ExoN1 - AEA10473.1
13. SARS coronavirus MA15 ExoN1 - AEA10443.1
14. Severe acute respiratory syndrome-related c...
15. SARS coronavirus ExoN1 - AGT21078.1
16. SARS coronavirus BJ302 - AAR07625.1
17. SARS coronavirus BJ302 - AAR07627.1
18. SARS coronavirus GD01 - AAP51227.1
19. Rhinolophus affinis coronavirus - AHX37558.1
20. Rhinolophus affinis coronavirus - AHX37569.1
21. Bat SARS-like coronavirus - AVP78031.1
22. Bat SARS-like coronavirus - AVP78042.1
23. Bat coronavirus - QHR63300.1
24. EPI_ISL_402131|bat/Yunnan/RaTG13/2013|...
25. EPI_ISL_402125|Wuhan-Hu-1/2019|MN908...
furin...

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.

February 7, 2020 7:17 AM

Theories of 2019-nCoV origins

As noted above 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 available 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 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 therefore theoretically 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.

Selection in an animal host

Given the similarity of 2019-nCoV to bat SARS-like CoVs, particularly RaTG13, it is 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 probable 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. Although analyses are ongoing, the pangolin viruses described to date 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

February 11, 2020 7:17 AM

sarbecoviruses before, it is unclear what conditions would be required for it to be acquired in the lineage leading to 2019-nCoV.

Conclusions

The genomic features described here, which may in part explain the infectiousness and transmissibility of 2019-nCoV in humans, could have arisen through a process of adaptive evolution prior to the start of the outbreak. Although we can readily dismiss the idea that 2019-nCoV is an experimentally manipulated virus, it is impossible to prove or disprove other theories of its origin, 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 revealing virus origins. In addition, it would be helpful to obtain additional genetic and functional data about the virus, including 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, would also be informative. 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.

February 11, 2020 7:17 AM

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).

February 11, 2020 7:17 AM

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).

February 11, 2020 7:17 AM

The proximal origin of SARS-CoV-2

Since the first reports of a novel pneumonia (COVID-19) in Wuhan city, Hubei province, China there has been considerable uncertainty and discussion over the possible origin of the causative virus. Herein, we review what can be deduced about the origin of this virus, SARS-CoV-2, from the comparative analysis of available genome sequence data. In particular, we describe notable features in the SARS-CoV-2 genome, outline mechanisms for how the virus may have acquired them, and discuss scenarios by which these features could have arisen. Importantly, this analysis provides evidence that SARS-CoV-2 is not a laboratory construct or experimentally manipulated virus.

Genomic comparisons identified two features of the SARS-CoV-2 genome of note: (i) based on structural modeling and early biochemical experiments, SARS-CoV-2 appears to be optimized for binding to the human ACE2 receptor; (ii) the highly variable spike protein of SARS-CoV-2 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 SARS-CoV-2

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, SARS-CoV-2 displays a similar level of diversity as predicted from previous studies, including to its most closely related virus, a SARS-like CoV isolated from bats (RaTG13) to which it is ~96% identical.

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 Urbani strain of SARS-CoV, they are Y442, L472, N479, D480, T487, and Y491. The corresponding residues in SARS-CoV-2 are L455, F486, Q493, S494, N501, and Y505. Five out of six of these residues are mutated in SARS-CoV-2 compared to the most closely related virus, RaTG13 (**Figure 1**). Based on modeling[1] and biochemical experiments[2,3], SARS-CoV-2 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, SARS-CoV-2 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 SARS-CoV-2 corresponds to L472 in the SARS-CoV Urbani strain. In cell culture experiments, the leucine at position 472 is mutated to phenylalanine (L472F)[4]; this mutation is 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 SARS-CoV-2 may be capable of binding the human ACE2 receptor with high affinity, the interaction is not predicted to be optimal[1]. Additionally, several of the key residues in the RBD of SARS-CoV-2 are different from those previously described as optimal for human ACE2 receptor binding[5]. This is strong evidence that SARS-CoV-2 is *not* the product of genetic engineering.

February 12, 2020 6:38PM

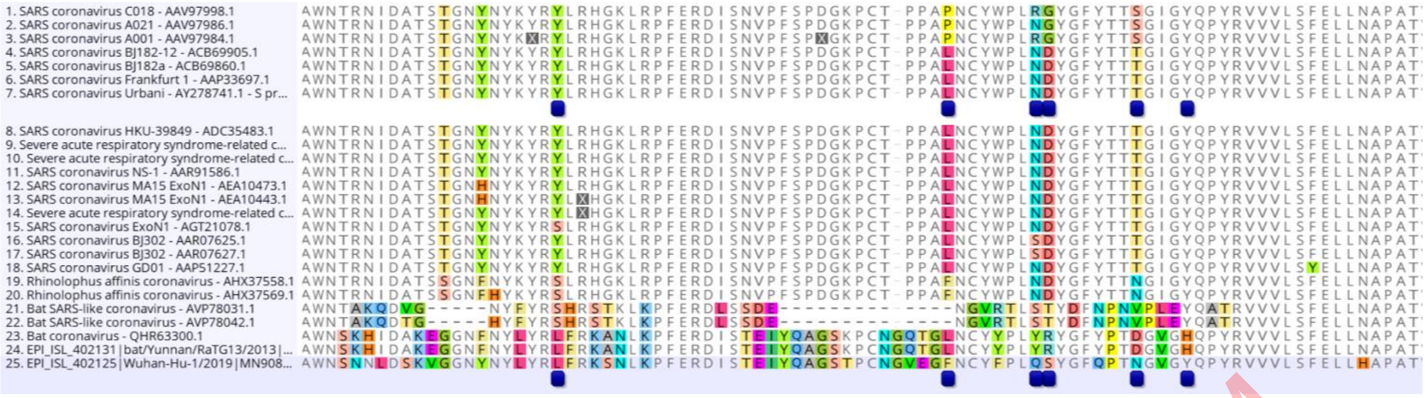

1. SARS coronavirus C018 - AAV97998.1
2. SARS coronavirus A021 - AAV97986.1
3. SARS coronavirus A001 - AAV97984.1
4. SARS coronavirus BJ182-12 - ACB69905.1
5. SARS coronavirus BJ182a - ACB69860.1
6. SARS coronavirus Frankfurt 1 - AAP33697.1
7. SARS coronavirus Urbani - AY278741.1 - S pr...
8. SARS coronavirus HKU-39849 - ADC35483.1
9. Severe acute respiratory syndrome-related c...
10. Severe acute respiratory syndrome-related c...
11. SARS coronavirus NS-1 - AAR91586.1
12. SARS coronavirus MA15 ExoN1 - AEA10473.1
13. SARS coronavirus MA15 ExoN1 - AEA10443.1
14. Severe acute respiratory syndrome-related c...
15. SARS coronavirus ExoN1 - AGT21078.1
16. SARS coronavirus BJ302 - AAR07625.1
17. SARS coronavirus BJ302 - AAR07627.1
18. SARS coronavirus GD01 - AAP51227.1
19. Rhinolophus affinis coronavirus - AHX37558.1
20. Rhinolophus affinis coronavirus - AHX37569.1
21. Bat SARS-like coronavirus - AVP78031.1
22. Bat SARS-like coronavirus - AVP78042.1
23. Bat coronavirus - QHR63300.1
24. EPI_ISL_402131|bat/Yunnan/RaTG13/2013|...
25. EPI_ISL_402125|Wuhan-Hu-1/2019|MN908...

Figure 1 | Mutations in contact residues of the SARS-CoV-2 spike protein. The spike protein of SARS-CoV-2 (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 SARS-CoV-2 and the SARS-CoV Urbani strain.

Furin cleavage site and O-linked glycans

The second notable feature of SARS-CoV-2 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; thus, the fully inserted sequence is PRRA (**Figure 2**). This proline is predicted to create three flanking O-linked glycans at S673, T678, and S686. A furin site has not previously been observed in the lineage B betacoronaviruses and is a unique feature of SARS-CoV-2. Some human betacoronaviruses, including HCoV-HKU1 (lineage A) have furin cleavage sites (typically RRKR), although not in such an optimal position.

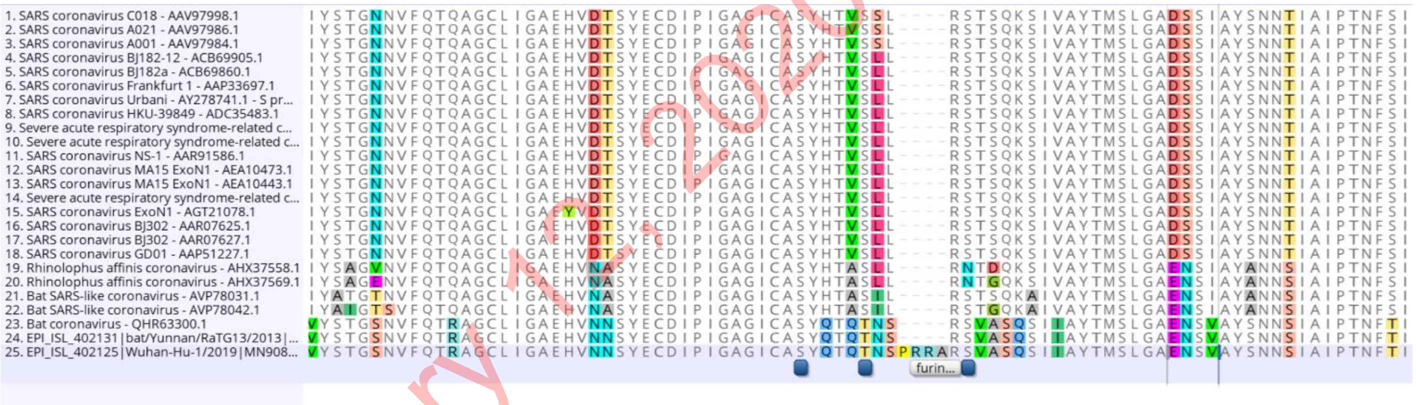

1. SARS coronavirus C018 - AAV97998.1
2. SARS coronavirus A021 - AAV97986.1
3. SARS coronavirus A001 - AAV97984.1
4. SARS coronavirus BJ182-12 - ACB69905.1
5. SARS coronavirus BJ182a - ACB69860.1
6. SARS coronavirus Frankfurt 1 - AAP33697.1
7. SARS coronavirus Urbani - AY278741.1 - S pr...
8. SARS coronavirus HKU-39849 - ADC35483.1
9. Severe acute respiratory syndrome-related c...
10. Severe acute respiratory syndrome-related c...
11. SARS coronavirus NS-1 - AAR91586.1
12. SARS coronavirus MA15 ExoN1 - AEA10473.1
13. SARS coronavirus MA15 ExoN1 - AEA10443.1
14. Severe acute respiratory syndrome-related c...
15. SARS coronavirus ExoN1 - AGT21078.1
16. SARS coronavirus BJ302 - AAR07625.1
17. SARS coronavirus BJ302 - AAR07627.1
18. SARS coronavirus GD01 - AAP51227.1
19. Rhinolophus affinis coronavirus - AHX37558.1
20. Rhinolophus affinis coronavirus - AHX37569.1
21. Bat SARS-like coronavirus - AVP78031.1
22. Bat SARS-like coronavirus - AVP78042.1
23. Bat coronavirus - QHR63300.1
24. EPI_ISL_402131|bat/Yunnan/RaTG13/2013|...
25. EPI_ISL_402125|Wuhan-Hu-1/2019|MN908...
furin...

Figure 2 | Acquisition of furin cleavage site and O-linked glycans. The spike protein of SARS-CoV-2 (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 SARS-CoV-2 and not previously seen in this group of viruses.

While the functional consequence of the furin cleavage site in SARS-CoV-2 is unknown, 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 conditions 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 SARS-CoV-2 spike protein.

Theories of SARS-CoV-2 origins

We believe it is unlikely that SARS-CoV-2 emerged through laboratory manipulation of an existing SARS-related coronavirus. As noted above, the RBD of SARS-CoV-2 is not optimized for human ACE2 receptor binding. Furthermore, if genetic manipulation had been performed, one would expect that a researcher

February 17, 2020 6:38PM

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 SARS-CoV-2 is not derived from any previously used virus backbone, for example those described in a 2015 paper in *Nature Medicine*[10]. Instead, we propose three scenarios that plausibly explain the origin of SARS-CoV-2: (1) natural selection in humans, (2) selection during passage in culture, and (3) natural selection in an animal host.

Adaptation to humans

Adaptations to humans as a host may have arose after the virus jumped from a non-human host, during the early stages of the epidemic. However, all known genome sequences include the RBD associated with human ACE2 receptor binding. Timing of the most recent common ancestor estimates point to the emergence of SARS-CoV-2 in early December. This would afford little opportunity for adaptation to occur. Importantly, however, we cannot exclude the possibility of an earlier hidden epidemic of sufficient magnitude and duration for the adaptations to occur. This issue could be addressed with retrospective serological studies.

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]). In theory it is possible that SARS-CoV-2 acquired the RBD mutations and furin cleavage site during adaptation to passage in cell culture, as has been observed in studies with SARS-CoV[4]. However, it is less clear how the O-linked glycans - if functional - would have been acquired in such a manner, as these typically suggest the involvement of an immune system, that is not present *in vitro*.

Selection in an animal host

Given the similarity of SARS-CoV-2 to bat SARS-like CoVs, particularly RaTG13, it is likely that bats serve as the long-term 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 probable that there was an intermediate host for SARS-CoV-2. Provisional analyses indicate that Malayan pangolins (*Manis javanica*) illegally imported into Guangdong province contain CoVs that are similar to SARS-CoV-2[11]. Although the bat virus RaTG13 remains the closest relative to SARS-CoV-2 across the whole genome, the Malayan pangolin CoVs are identical to SARS-CoV-2 at all six key RBD residues. However, those Pangolin viruses described to date do not carry the furin cleavage site insertion. For a precursor virus to acquire the furin cleavage site and mutations in the spike proteins that appear to be suitable for human ACE2 receptor binding, an animal host would likely 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.

Conclusions

The genomic features described here, which may in part explain the infectiousness and transmissibility of SARS-CoV-2 in humans, could have arisen through a process of adaptive evolution prior to the start of the outbreak. Although we can readily dismiss the idea that SARS-CoV-2 is an experimentally manipulated virus, it is impossible to prove or disprove other theories of its origin, 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 revealing virus origins. In addition, it would be helpful to obtain additional genetic and functional data about the virus, including 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 SARS-CoV-2, as well as sequencing of very early cases including those not connected to the market, would also be informative. Even in the light of such data, however, it is not

February 12, 2020 6:38PM

guaranteed that data can be obtained to conclusively prove all aspects of the initial emergence of SARS-CoV-2.

February 12, 2020 6:38PM

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 SARS-CoV-2. *bioRxiv* 2020.01.22.915660 (2020) doi:10.1101/2020.01.22.915660.

3. Hoffmann, M. *et al.* The novel coronavirus 2019 (SARS-CoV-2) 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).

February 12, 2020 6:38PM

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).

February 12, 2020 6:38PM

The proximal origin of SARS-CoV-2

Since the first reports of a novel pneumonia (COVID-19) in Wuhan city, Hubei province, China there has been considerable uncertainty and discussion over the possible origin of the causative virus, SARS-CoV-2. Herein, we review what can be deduced about the origin SARS-CoV-2 from the comparative analysis of available genome sequence data. In particular, we offer a perspective of notable features in the SARS-CoV-2 genome, outline mechanisms for how the virus may have acquired them, and discuss scenarios by which these features could have arisen. Importantly, this analysis provides evidence that SARS-CoV-2 is not a laboratory construct or experimentally manipulated virus.

Genomic comparisons of both alpha- and betacoronaviruses (family *Coronaviridae*) identify two notable features of the SARS-CoV-2 genome: (i) based on structural modeling and early biochemical experiments, SARS-CoV-2 appears to be optimized for binding to the human ACE2 receptor; (ii) the highly variable spike protein of SARS-CoV-2 has a furin cleavage site at the S1 and S2 boundary via the insertion of twelve 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 SARS-CoV-2

The receptor binding domain (RBD) in the spike protein of SARS-CoV and SARS-related coronaviruses is the most variable part of the virus genome. Six residues in the RBD appear to be critical for binding to the human ACE2 receptor and determining host range[1]. Using coordinates based on the Urbani strain of SARS-CoV, they are Y442, L472, N479, D480, T487, and Y491(Wan et al. 2020). The corresponding residues in SARS-CoV-2 are L455, F486, Q493, S494, N501, and Y505. Five out of six of these residues are mutated in SARS-CoV-2 compared to its most closely related virus, RaTG13 sampled from a bat, to which it is ~96% identical (Wu et al. 2020). (**Figure 1**). Based on modeling[1] and biochemical experiments[2,3], SARS-CoV-2 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(Wan et al. 2020). In contrast, SARS-CoV-2 may bind less efficiently to ACE2 in other species associated with SARS-like viruses, including rodents and civets[1].

A phenylalanine at F486 in SARS-CoV-2 corresponds to L472 in the SARS-CoV Urbani strain. In cell culture experiments, the leucine at position 472 is mutated to phenylalanine (L472F)[4]; this mutation is 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 SARS-CoV-2 may be capable of binding the human ACE2 receptor with high affinity, the interaction is not predicted to be optimal[1]. Additionally, several of the key residues in the RBD of SARS-CoV-2 are different from those previously described as optimal for human ACE2 receptor binding[5]. This is strong evidence that SARS-CoV-2 is *not* the product of genetic engineering.

February 13, 2020 8:10 PM

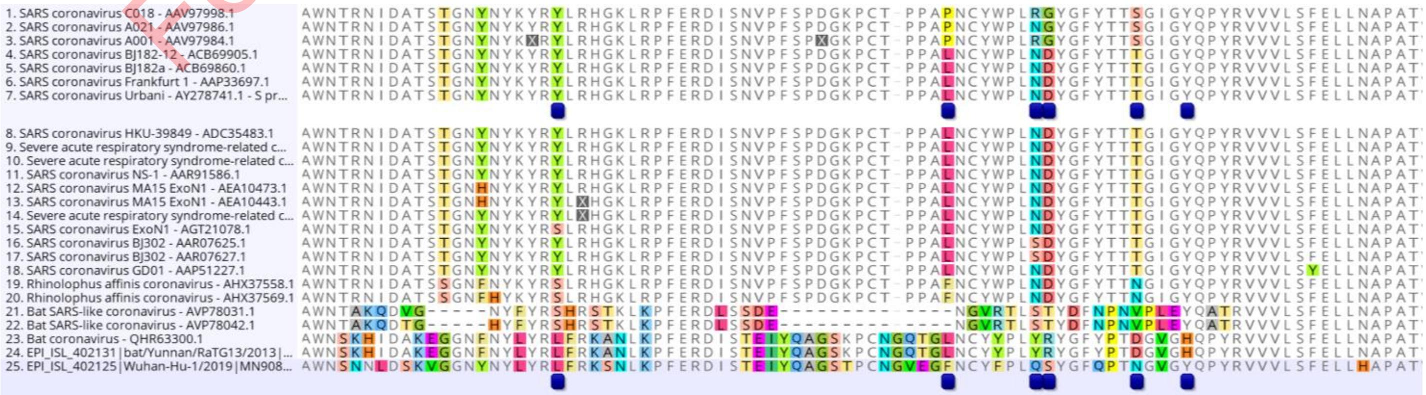

1. SARS coronavirus C018 - AAV97998.1
2. SARS coronavirus A021 - AAV97986.1
3. SARS coronavirus A001 - AAV97984.1
4. SARS coronavirus BJ182-12 - ACB69905.1
5. SARS coronavirus BJ182a - ACB69860.1
6. SARS coronavirus Frankfurt 1 - AAP33697.1
7. SARS coronavirus Urbani - AY278741.1 - S pr...
8. SARS coronavirus HKU-39849 - ADC35483.1
9. Severe acute respiratory syndrome-related c...
10. Severe acute respiratory syndrome-related c...
11. SARS coronavirus NS-1 - AAR91586.1
12. SARS coronavirus MA15 ExoN1 - AEA10473.1
13. SARS coronavirus MA15 ExoN1 - AEA10443.1
14. Severe acute respiratory syndrome-related c...
15. SARS coronavirus ExoN1 - AGT21078.1
16. SARS coronavirus BJ302 - AAR07625.1
17. SARS coronavirus BJ302 - AAR07627.1
18. SARS coronavirus GD01 - AAP51227.1
19. Rhinolophus affinis coronavirus - AHX37558.1
20. Rhinolophus affinis coronavirus - AHX37569.1
21. Bat SARS-like coronavirus - AVP78031.1
22. Bat SARS-like coronavirus - AVP78042.1
23. Bat coronavirus - QHR63300.1
24. EPI_ISL_402131|bat/Yunnan/RaTG13/2013|...
25. EPI_ISL_402125|Wuhan-Hu-1/2019|MN908...

Figure 1 | Mutations in contact residues of the SARS-CoV-2 spike protein. The spike protein of SARS-CoV-2 (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 SARS-CoV-2 and the SARS-CoV Urbani strain.

Furin cleavage site and O-linked glycans

The second notable feature of SARS-CoV-2 is a predicted polybasic (furin) cleavage site (RRAR) in the spike protein at the junction of S1 and S2, the two subunits of the spike (**Figure 2**)(Gallaher 2020; Coutard et al. 2020). In addition to two basic arginines and an alanine at the cleavage site, a leading proline is also inserted; thus, the fully inserted sequence is PRRA (**Figure 2**). This proline is predicted to create three flanking O-linked glycans at S673, T678, and S686. A polybasic cleavage site has not previously been observed in the lineage B betacoronaviruses and is a unique feature of SARS-CoV-2. Some human betacoronaviruses, including HCoV-HKU1 (lineage A), have polybasic cleavage sites, as well as predicted O-linked glycans near the S1/S2 cleavage site.

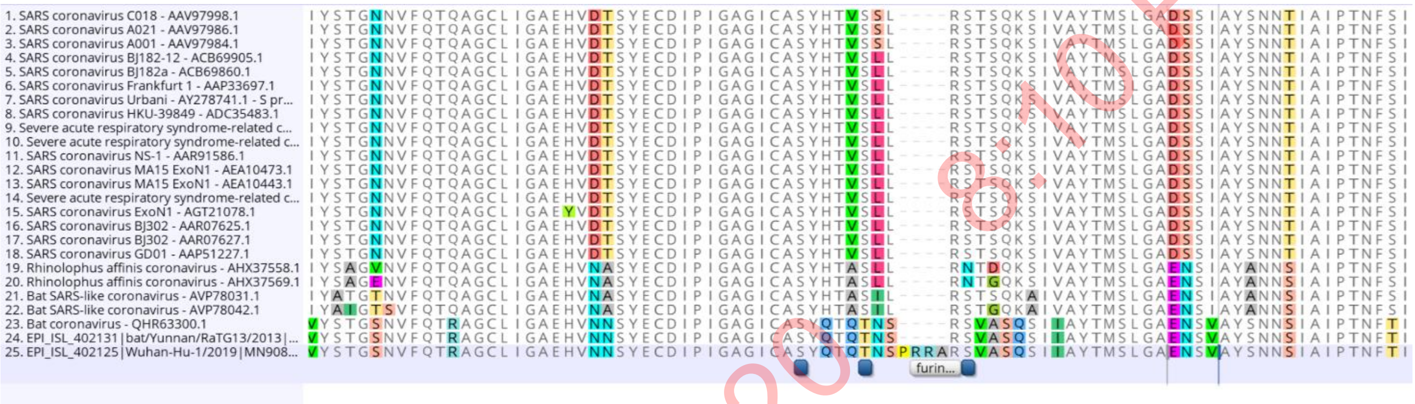

1. SARS coronavirus C018 - AAV97998.1
2. SARS coronavirus A021 - AAV97986.1
3. SARS coronavirus A001 - AAV97984.1
4. SARS coronavirus BJ182-12 - ACB69905.1
5. SARS coronavirus BJ182a - ACB69860.1
6. SARS coronavirus Frankfurt 1 - AAP33697.1
7. SARS coronavirus Urbani - AY278741.1 - S pr...
8. SARS coronavirus HKU-39849 - ADC35483.1
9. Severe acute respiratory syndrome-related c...
10. Severe acute respiratory syndrome-related c...
11. SARS coronavirus NS-1 - AAR91586.1
12. SARS coronavirus MA15 ExoN1 - AEA10473.1
13. SARS coronavirus MA15 ExoN1 - AEA10443.1
14. Severe acute respiratory syndrome-related c...
15. SARS coronavirus ExoN1 - AGT21078.1
16. SARS coronavirus BJ302 - AAR07625.1
17. SARS coronavirus BJ302 - AAR07627.1
18. SARS coronavirus GD01 - AAP51227.1
19. Rhinolophus affinis coronavirus - AHX37558.1
20. Rhinolophus affinis coronavirus - AHX37569.1
21. Bat SARS-like coronavirus - AVP78031.1
22. Bat SARS-like coronavirus - AVP78042.1
23. Bat coronavirus - QHR63300.1
24. EPI_ISL_402131|bat/Yunnan/RaTG13/2013|...
25. EPI_ISL_402125|Wuhan-Hu-1/2019|MN908...
furin...

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

While the functional consequence of the polybasic cleavage site in SARS-CoV-2 is unknown, experiments with SARS-CoV have shown that engineering a polybasic cleavage site at the S1/S2 junction enhances cell–cell fusion but does not affect virus entry[6]. Polybasic cleavage sites allow effective cleavage by furin and other proteases, and can be acquired at the junction of the two subunits of the hemagglutinin (HA) of avian influenza viruses in conditions selecting for rapid virus replication and transmission (e.g., highly dense chicken populations). HA serves a similar function in cell-cell fusion and viral entry as the coronavirus spike. Acquisition of a polybasic cleavage site in HA, by insertion or recombination, converts low pathogenicity avian influenza viruses to a highly pathogenic avian influenza viruses[7–9]. The acquisition of polybasic cleavage sites by influenza virus HA has also been observed after repeated forced passage in cell culture or through animals (refs). 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 SARS-CoV-2 spike protein. Biochemical analyses or structural studies are required to determine whether or not the predicted O-linked glycan sites are utilized.

February 13, 2020 3:10 PM

Theories of SARS-CoV-2 origins

It is unlikely that SARS-CoV-2 emerged through laboratory manipulation of an existing SARS-related coronavirus. As noted above, the RBD of SARS-CoV-2 is not optimized for human ACE2 receptor binding. Furthermore, if genetic manipulation had 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 SARS-CoV-2 is not derived from any previously used virus backbone, for example those described in a 2015 study[10]. Instead, we propose three scenarios that plausibly explain the origin of SARS-CoV-2: (1) natural selection in humans, (2) selection during passage in culture, and (3) natural selection in an animal host.

Selection in an animal host

Given the similarity of SARS-CoV-2 to bat SARS-like CoVs, particularly RaTG13, it is plausible that bats serve as reservoir hosts for the SARS-CoV-2. It is important, however, to consider that previous human outbreaks caused by betacoronaviruses have involved direct human exposures to animals other than bats, including civets (SARS) and camels (MERS) that carry viruses that are highly genetically similar to SARS-CoV-1 or MERS-CoV, respectively. These observations suggest that civets and perhaps other animals are reservoirs for SARS-CoV-1 and that camels are the reservoirs for MERS-CoV. In contrast, bat coronaviruses with closely related genomes to SARS-CoV-1 or MERS-CoV have not yet been characterised so that they are unlikely to be proximal hosts, although this clearly needs to be confirmed with additional sampling.

By analogy to SARS-CoV-1 and MERS-CoV, viruses closely related to SARS-Cov-2 may be circulating in one or more animal species. Provisional analyses indicate that Malayan pangolins (*Manis javanica*) illegally imported into Guangdong province contain a CoV that is similar to SARS-CoV-2 (Wong et al. 2020). Although the bat virus RaTG13 remains the closest relative to SARS-CoV-2 across the whole genome, the Malayan pangolin CoV is identical to SARS-CoV-2 at all six key RBD residues (**Figure 1**). However, no pangolin CoV has yet been identified that has sufficient genetically similar to the SARS-CoV-2 across its entire genome to be consistent with the hypothesis that the emergence of SARS-CoV-2 was a result of direct infection by a pangolin CoV. In addition, the pangolin CoV described to date does not carry a polybasic furin cleavage site insertion. For a precursor virus to acquire the polybasic cleavage site and mutations in the spike proteins that appear to be suitable for human ACE2 receptor binding, an animal host would likely 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. Further characterization of CoVs in pangolins and other animals that may harbor SARS-CoV-like viruses should be an urgent public health priority.

Adaptation to humans

SARS-CoV-2 may have emerged from a less pathogenic progenitor that has been circulating in humans for an extended period of time. Most human infections of SARS-CoV-2 do not appear to lead to disease requiring hospitalization and it is possible that a SARS-CoV-2 progenitor with lower pathogenic potential circulated undetected in humans prior to the current epidemic. Other human CoVs, including HKU1, OC43, HL63 and E229, caused human infections for decades before they were recognized and characterized (ref).

Estimates of the timing of the most recent common ancestor (tMRCA) using currently available genome sequence data point to the emergence of SARS-CoV-2 in late November, early December 2019, compatible with the earliest retrospectively confirmed cases (Huang et al. 2020). This recent tMCRA date would seem to afford little opportunity for human adaptation, arguing against the circulation of SARS-CoV-2 in humans for an extended time period. However, all known genome sequences of SARS-CoV-2 have a RBD that is seemingly well, although not optimally, adapted for associated with human ACE2 receptor binding. A variable in this scenario is whether or not the animal CoV contained the polybasic cleavage site and how long this would take to acquire by adaptive evolution. Acquisition of a polybasic cleavage site in HA influenza virus has been observed to occur by insertion or recombination events. The human adaptation scenario requires that a specific insertion or recombination occurred to allow emergence of SARS-CoV-2. If there was a relative recent acquisition of the furin site this could be consistent with the current tMCRA estimate.

We cannot exclude the possibility of an earlier hidden epidemic of sufficient magnitude and duration for the adaptations of an animal CoV to humans to occur. Metagenomic studies of banked serum samples could provide important information, but given the relative short period of viremia it may be impossible to

February 13, 2020 8:10 PM

detect of low level SARS-CoV-2 circulation in historical sample. Retrospective serological studies potentially could be informative and a few such studies have already been conducted [refs]. One such study found that hunters involved in animal importation had a XX% seropositivity to betacoronaviruses, while another found that X% residents of a village in Southern China were seropositive to these viruses. Interestingly, this prior study found that 200 residents of Wuhan did not show betacornavirus seroreactivity. Critically, however, these studies could not have distinguished whether the positive serological responses were due to a prior infection with SARS-CoV-1 or -2 or another betacoronavirus. Further retrospective serological studies should be conducted to determine the extent of prior human exposure to betacoronaviruses in different geographic areas, with attention directed to the development of serological assays that distinguish among multiple betacoronaviruses.

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 in multiple laboratories across the world[12–15]. Because we have taken an unbiased approach to attempt to discern the origin of SARS-CoV-2, we considered the possibility that SARS-Cov-2 originated in a laboratory and inadvertently infected a laboratory worker who transmitted the infection. We consider this scenario to be extremely unlikely. However, after the emergence of SARS-CoV-1 several instances of laboratory acquisition of this virus by laboratory personnel working under BSL-2 containment have been documented and thus we cannot eliminate this possibility beyond doubt (ref). In theory, it is possible that SARS-CoV-2 acquired the observed RBD mutations site during adaptation to passage in cell culture, as has been observed in studies with SARS-CoV[4] as well as MERS-CoV (ref). However, acquisition of the polybasic cleavage site or O-linked glycans - if functional - argues strongly against this scenario. Acquisition of polybasic cleavage sites has been observed after passage of low pathogenicity avian influenza virus in either cell culture or animals. However, the generation of these sites required lengthy forced passage in cells or animals. Furthermore, the generation of SARS-CoV-2 by cell culture or animal passage would have required prior isolation of a progenitor virus with a very genetic high similarity. Subsequent generation of a polybasic cleavage site would have then required an intense program of passage in cell culture or animals with ACE-2 receptor similar to humans (e.g. ferrets). It is also questionable whether generation of the O-linked glycan would have occurred cell culture passage, as such mutations typically suggest the involvement of an immune system, that is not present *in vitro*.

Conclusions

The genomic features described here, which may in part explain the infectiousness and transmissibility of SARS-CoV-2 in humans, could have arisen through a process of adaptive evolution prior to the start of the outbreak. Although current evidence does not support the idea that SARS-CoV-2 is a laboratory construct, it is impossible to prove or disprove other theories of its origin, 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 revealing virus origins. In addition, it would be helpful to obtain additional genetic and functional data about the virus, including experimental studies of receptor binding and the role of the polybasic cleavage site and predicted O-linked glycans. The identification of a potential intermediate host of SARS-CoV-2, as well as sequencing of very early cases including those not connected to the market, would also be informative.

February 13, 2020 8:10 PM

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 SARS-CoV-2. *bioRxiv* 2020.01.22.915660 (2020) doi:10.1101/2020.01.22.915660.
3. Hoffmann, M. *et al.* The novel coronavirus 2019 (SARS-CoV-2) 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).

February 13, 2020 8:10 PM

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).

February 13, 2020 8:10 PM

The proximal origin of SARS-CoV-2

Since the first reports of a novel pneumonia (COVID-19) in Wuhan city, Hubei province, China there has been considerable discussion and uncertainty over the origin of the causative virus, SARS-CoV-2. Infections with SARS-CoV-2 are now widespread in China, with cases in every province. As of 14 February 2020, 64,473 cases have been confirmed, with 1,384 deaths attributed to the virus. These official case numbers are likely a large underestimate because of limited reporting of mild and asymptomatic cases, and the virus is clearly capable of efficient human-to-human transmission. Based on the possibility of spread to countries with weaker healthcare systems, the World Health Organization declared the COVID-9 outbreak a Public Health Emergency of International Concern (PHEIC). There are currently no vaccines nor specific treatments for this disease.

SARS-CoV-2 is the seventh member of the *Coronaviridae* that can infect humans, including SARS CoV-1, MERS, and four viruses (HKU1, NL63, OC43 and E229) that cause generally mild respiratory symptoms. Herein, we review what can be deduced about the origin and early evolution of SARS-CoV-2 from the comparative analysis of available genome sequence data. In particular, we offer a perspective of notable features in the SARS-CoV-2 genome, outline mechanisms for how the virus may have acquired them, and discuss scenarios by which these features could have arisen. Importantly, this analysis provides evidence that SARS-CoV-2 is not a laboratory construct nor an experimentally manipulated virus.

The genomic comparisons of both alpha- and betacoronaviruses (family *Coronaviridae*) described below identify two notable features of the SARS-CoV-2 genome: (i) based on structural modeling and early biochemical experiments, SARS-CoV-2 appears to be optimized for binding to the human ACE2 receptor; (ii) the highly variable spike protein of SARS-CoV-2 has a furin cleavage site at the S1 and S2 boundary via the insertion of twelve 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 SARS-CoV-2

The receptor binding domain (RBD) in the spike protein of SARS-CoV and SARS-related coronaviruses is the most variable part of the virus genome. Six residues in the RBD appear to be critical for binding to the human ACE2 receptor and determining host range[1]. Using coordinates based on the Urbani strain of SARS-CoV, they are Y442, L472, N479, D480, T487, and Y491(Wan et al. 2020). The corresponding residues in SARS-CoV-2 are L455, F486, Q493, S494, N501, and Y505. Five out of six of these residues are mutated in SARS-CoV-2 compared to its most closely related virus, RaTG13 sampled from a *Rhinolophus affinis* bat, to which it is ~96% identical (Wu et al. 2020). (**Figure 1**). Based on modeling[1] and biochemical experiments[2,3], SARS-CoV-2 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(Wan et al. 2020). In contrast, SARS-CoV-2 may bind less efficiently to ACE2 in other species associated with SARS-like viruses, including rodents and civets[1].

A phenylalanine at F486 in SARS-CoV-2 corresponds to L472 in the SARS-CoV Urbani strain. In cell culture experiments the leucine at position 472 is mutated to phenylalanine (L472F)[4]; this mutation is 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 SARS-CoV-2 may be capable of binding the human ACE2 receptor with high affinity, the interaction is not predicted to be optimal[1]. Additionally, several of the key residues in the RBD of SARS-CoV-2 are different from those previously described as optimal for human ACE2 receptor binding[5]. This is strong evidence that SARS-CoV-2 is *not* the product of genetic engineering.

February 14, 2020, 8:44 PM

Furin cleavage site and O-linked glycans

The second notable feature of SARS-CoV-2 is a predicted polybasic (furin) cleavage site (RRAR) in the spike protein at the junction of S1 and S2, the two subunits of the spike (**Figure 2**)(Coutard et al. 2020). In addition to two basic arginines and an alanine at the cleavage site, a leading proline is also inserted; thus, the fully inserted sequence is PRRA (**Figure 2**). This proline is predicted to create three flanking O-linked glycans at S673, T678, and S686. A polybasic cleavage site has not previously been observed in the related lineage B betacoronaviruses and is a unique feature of SARS-CoV-2. Some human betacoronaviruses, including HCoV-HKU1 (lineage A), have polybasic cleavage sites, as well as predicted O-linked glycans near the S1/S2 cleavage site.

While the functional consequence of the polybasic cleavage site in SARS-CoV-2 is unknown, experiments with SARS-CoV have shown that engineering such a site at the S1/S2 junction enhances cell–cell fusion but does not affect virus entry[6]. Polybasic cleavage sites allow effective cleavage by furin and other proteases, and can be acquired at the junction of the two subunits of the hemagglutinin (HA) of avian influenza viruses in conditions that select for rapid virus replication and transmission (e.g., highly dense chicken populations). HA serves a similar function in cell-cell fusion and viral entry as the coronavirus spike. Acquisition of a polybasic cleavage site in HA, by insertion or recombination, converts low pathogenicity avian influenza viruses to a highly pathogenic avian influenza viruses[7–9]. The acquisition of polybasic cleavage sites by influenza virus HA has also been observed after repeated forced passage in cell culture or through animals (refs). 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 SARS-CoV-2 spike protein. Biochemical analyses or structural studies are required to determine whether or not the predicted O-linked glycan sites are utilized.

Theories of SARS-CoV-2 origins

It is unlikely that SARS-CoV-2 emerged through laboratory manipulation of an existing SARS-related coronavirus. As noted above, the RBD of SARS-CoV-2 is not optimized for human ACE2 receptor binding. Furthermore, if genetic manipulation had 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 SARS-CoV-2 is not derived from any previously used virus backbone, for example those described in a key 2015 study[10]. Instead, we propose two scenarios that can plausibly explain the origin of SARS-CoV-2: (1) natural selection in a non-human animal host prior to zoonotic transfer, (2) natural selection in humans following zoonotic transfer. We also briefly discuss whether selection during passage in culture could have given rise to the same observed features.

Selection in an animal host

Given the similarity of SARS-CoV-2 to bat SARS-like CoVs, particularly RaTG13, it is plausible that bats serve as reservoir hosts for the SARS-CoV-2. Indeed, as many of the early cases were linked to the Huanan seafood market in Wuhan, it is possible that any zoonotic source was present at this location. It is important, however, to consider that previous human outbreaks caused by betacoronaviruses have involved direct human exposures to animals other than bats, including civets (SARS) and camels (MERS) that carry viruses that are genetically very similar to SARS-CoV-1 or MERS-CoV, respectively. This suggests that civets and perhaps other animals are reservoirs for SARS-CoV-1 and that camels are the reservoirs for MERS-CoV. In contrast, bat coronaviruses with closely related genomes to SARS-CoV-1 or MERS-CoV have not yet been characterised so that they are unlikely to be proximal hosts, although this clearly needs to be confirmed with additional sampling.

By analogy to SARS-CoV-1 and MERS-CoV, viruses closely related to SARS-Cov-2 may be circulating in one or more animal species. Provisional analyses indicate that Malayan pangolins (*Manis javanica*) illegally

February 14, 2020, 8:44 PM

imported into Guangdong province contain a CoV that is similar to SARS-CoV-2 (Wong et al. 2020). Although the bat virus RaTG13 remains the closest relative to SARS-CoV-2 across the whole genome, the Malayan pangolin CoV is identical to SARS-CoV-2 at all six key RBD residues (**Figure 1**). However, no pangolin CoV has yet been identified that has sufficient genetically similar to the SARS-CoV-2 across its entire genome to support direct human infection by a pangolin CoV. In addition, the pangolin CoV described to date does not carry a furin cleavage site insertion. For a precursor virus to acquire the polybasic cleavage site and mutations in the spike proteins that appear to be suitable for human ACE2 receptor binding, an animal host would likely 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. Further characterization of CoVs in pangolins and other animals that may harbor SARS-CoV-like viruses should be a public health priority.

Cryptic adaptation to humans

The second scenario is that the virus jumped from a non-human animal to humans with the genomic features described above acquired through adaptation to human infection and human-to-human transmission. We would surmise that once these adaptations were acquired (either together or in series) it would enable the outbreak to take off, producing a sufficiently large and unusual cluster of pneumonia cases to trigger the surveillance system that ultimately detected it.

All SARS-CoV-2 genomes sequenced so far have the well adapted RBD and the polybasic cleavage site, and are thus derived from a common ancestor that had these features. The presence of an RBD in pangolins that is very similar to the one in SARS-CoV-2 means that this was likely already present in the virus that jumped to humans, even if we don't yet have the exact non-human progenitor virus. This leaves the furin cleavage site insertion to occur during human-to-human transmission. Following the example of the HA gene of influenza A virus, this requires a specific insertion or recombination event to enable the emergence of SARS-CoV-2 as an epidemic pathogen.

Estimates of the timing of the most recent common ancestor (tMRCA) of SARS-CoV-2 using currently available genome sequence data point to virus emergence in late November to early December 2019, compatible with the earliest retrospectively confirmed cases (Huang et al. 2020). So, this scenario presumes a phase of unrecognised transmission in humans between the initial zoonotic transfer event and the acquisition of the furin cleavage site. Sufficient opportunity could occur if there had been many prior zoonotic events producing short chains of human-to-human transmission (so-called 'stuttering chains') over an extended period. This is essentially the situation for MERS-CoV in the Arabian Peninsula where all the human cases are the result of repeated jumps of the virus from dromedary camels, producing single infections or short chains of transmission which eventually resolve. To date, after 2499 cases over 8 years, no human adaptation has emerged that has allowed the virus to take hold in the human population.

We cannot exclude the cryptic spread of SARS-CoV-2 of sufficient magnitude and duration an animal CoV to adapt to humans. Metagenomic studies of banked serum samples could provide important information, but given the relatively short period of viremia it may be impossible to detect low level SARS-CoV-2 circulation in historical samples. Retrospective serological studies potentially could be informative and a few such studies have already been conducted. One such study found that traders involved in animal importation had a 13% seropositivity to coronaviruses (PMID: 14561956), while another found that 3% residents of a village in Southern China were seropositive to these viruses (PMID: 14561956, PMID: 29500691). Interestingly, 200 residents of Wuhan did not show coronavirus seroreactivity. Critically, however, these studies could not have distinguished whether positive serological responses were due to a prior infection with SARS-CoV-1 or -2. Further retrospective serological studies should be conducted to determine the

February 14, 2020, 8:44 PM

extent of prior human exposure to betacoronaviruses in different geographic areas, particularly using assays that can distinguish among multiple betacoronaviruses.

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 in multiple laboratories across the world[12–15]. Because we have taken an unbiased approach to attempt to discern the origin of SARS-CoV-2, we necessarily considered the possibility that SARS-Cov-2 originated in a laboratory and inadvertently infected a laboratory worker who transmitted the infection. While we consider this scenario to be highly unlikely, several instances of laboratory acquisition of SARS-CoV-1 by laboratory personnel working under BSL-2 containment were documented and thus we cannot eliminate this possibility beyond doubt (ref). In theory, it is possible that SARS-CoV-2 acquired the observed RBD mutations site during adaptation to passage in cell culture, as has been observed in studies with SARS-CoV[4] as well as MERS-CoV (PMID: 30110630). Importantly, however, the acquisition of the furin cleavage site or O-linked glycans - if functional - argues strongly against this scenario. Acquisition of polybasic cleavage sites has been observed after passage of low pathogenicity avian influenza virus in either cell culture or animals. However, the generation of these sites required lengthy forced passage in cells or animals. Furthermore, the generation of SARS-CoV-2 by cell culture or animal passage would have required prior isolation of a progenitor virus with a very genetic high similarity. Subsequent generation of a furin cleavage site would have then required an intense program of passage in cell culture or animals with ACE-2 receptor similar to humans (e.g. ferrets). It is also questionable whether generation of the O-linked glycan would have occurred on cell culture passage, as such mutations typically suggest the involvement of an immune system, that is not present *in vitro*.

Conclusions

In the midst of the global COVID-19 public health emergency it is reasonable to wonder why the origins of the epidemic matter. We believe that a detailed understanding of how an animal virus jumped species boundaries to infect humans so productively will undoubtedly help in the prevention of future zoonotic events. For example, if SARS-CoV-2 pre-adapted in another animal species then we are at risk of future reemergence events, even if the current epidemic is controlled. In contrast, if the adaptive process we describe occured in humans then even if we have repeated zoonotic transfers they are unlikely to take off unless the same series of mutations occur. In addition, identifying the closest animal relatives of SARS-CoV-2 will greatly assist studies of virus function. For example, the availability of the RaTG13 bat sequence facilitated the comparative genomic analysis performed, helping to reveal the key mutations in the RBD as well as the furin cleavage site insertion.

The genomic features described here may in part explain the infectiousness and transmissibility of SARS-CoV-2 in humans. Although current evidence does not support the idea that SARS-CoV-2 is a laboratory construct, it is impossible to prove or disprove the other theories of its origin outlined above, 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 revealing virus origins. In addition, it would be helpful to obtain additional genetic and functional data about the virus, including experimental studies of receptor binding and the role of the polybasic cleavage site and predicted O-linked glycans. The identification of a potential intermediate host of SARS-CoV-2, as well as sequencing of very early cases including those not connected to the fish market, would similarly be highly informative.

February 14, 2020, 8:44 PM

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 SARS-CoV-2. *bioRxiv* 2020.01.22.915660 (2020) doi:10.1101/2020.01.22.915660.
3. Hoffmann, M. *et al.* The novel coronavirus 2019 (SARS-CoV-2) 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).

February 14, 2020, 8:44 PM

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).

February 14, 2020, 8:44 PM

Figures

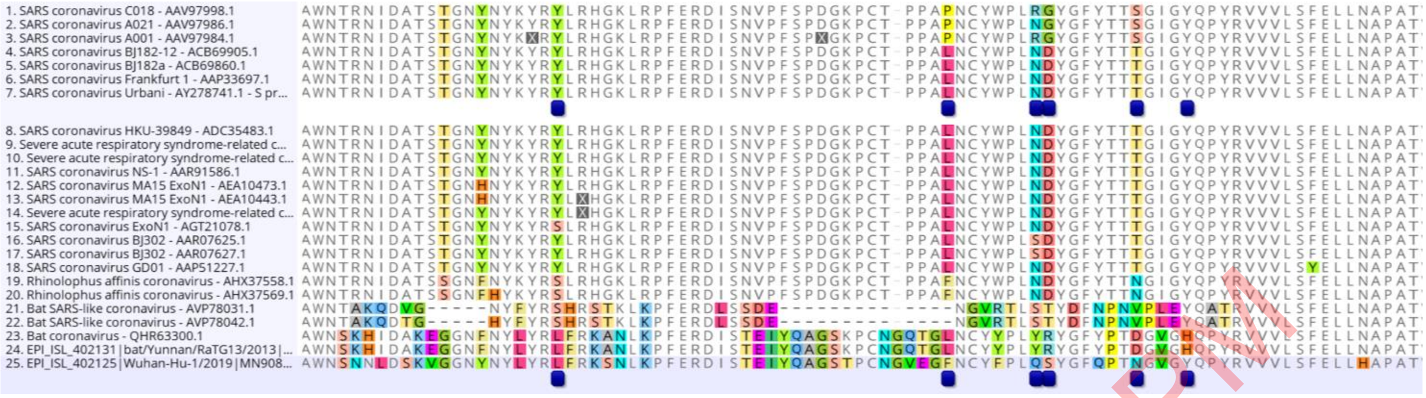

1. SARS coronavirus C018 - AAV97998.1
2. SARS coronavirus A021 - AAV97986.1
3. SARS coronavirus A001 - AAV97984.1
4. SARS coronavirus BJ182-12 - ACB69905.1
5. SARS coronavirus BJ182a - ACB69860.1
6. SARS coronavirus Frankfurt 1 - AAP33697.1
7. SARS coronavirus Urbani - AY278741.1 - S pr...
8. SARS coronavirus HKU-39849 - ADC35483.1
9. Severe acute respiratory syndrome-related c...
10. Severe acute respiratory syndrome-related c...
11. SARS coronavirus NS-1 - AAR91586.1
12. SARS coronavirus MA15 ExoN1 - AEA10473.1
13. SARS coronavirus MA15 ExoN1 - AEA10443.1
14. Severe acute respiratory syndrome-related c...
15. SARS coronavirus ExoN1 - AGT21078.1
16. SARS coronavirus BJ302 - AAR07625.1
17. SARS coronavirus BJ302 - AAR07627.1
18. SARS coronavirus GD01 - AAP51227.1
19. Rhinolophus affinis coronavirus - AHX37558.1
20. Rhinolophus affinis coronavirus - AHX37569.1
21. Bat SARS-like coronavirus - AVP78031.1
22. Bat SARS-like coronavirus - AVP78042.1
23. Bat coronavirus - QHR63300.1
24. EPI_ISL_402131|bat/Yunnan/RaTG13/2013|...
25. EPI_ISL_402125|Wuhan-Hu-1/2019|MN908...

Figure 1 | Mutations in contact residues of the SARS-CoV-2 spike protein. The spike protein of SARS-CoV-2 (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 SARS-CoV-2 and the SARS-CoV Urbani strain.

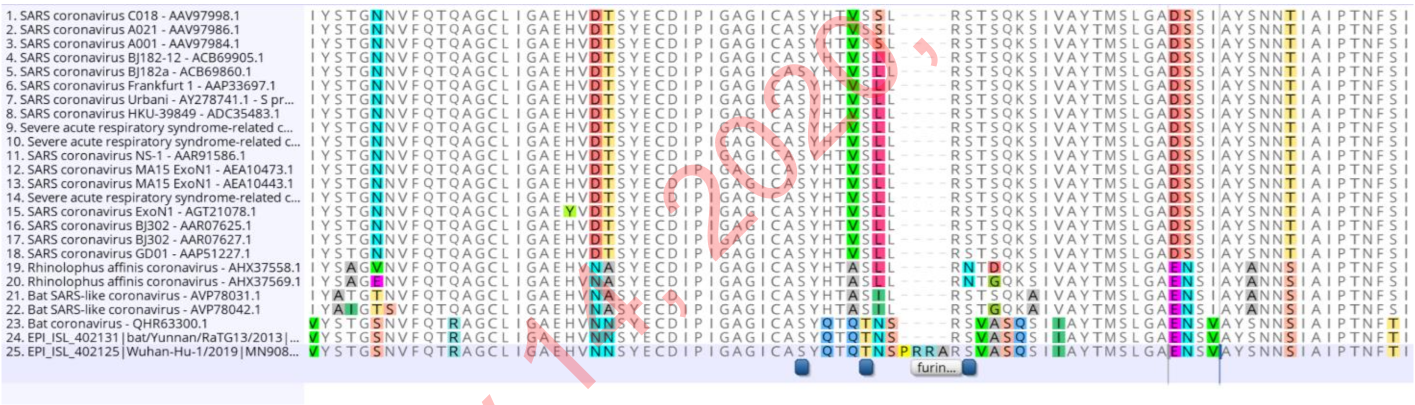

1. SARS coronavirus C018 - AAV97998.1
2. SARS coronavirus A021 - AAV97986.1
3. SARS coronavirus A001 - AAV97984.1
4. SARS coronavirus BJ182-12 - ACB69905.1
5. SARS coronavirus BJ182a - ACB69860.1
6. SARS coronavirus Frankfurt 1 - AAP33697.1
7. SARS coronavirus Urbani - AY278741.1 - S pr...
8. SARS coronavirus HKU-39849 - ADC35483.1
9. Severe acute respiratory syndrome-related c...
10. Severe acute respiratory syndrome-related c...
11. SARS coronavirus NS-1 - AAR91586.1
12. SARS coronavirus MA15 ExoN1 - AEA10473.1
13. SARS coronavirus MA15 ExoN1 - AEA10443.1
14. Severe acute respiratory syndrome-related c...
15. SARS coronavirus ExoN1 - AGT21078.1
16. SARS coronavirus BJ302 - AAR07625.1
17. SARS coronavirus BJ302 - AAR07627.1
18. SARS coronavirus GD01 - AAP51227.1
19. Rhinolophus affinis coronavirus - AHX37558.1
20. Rhinolophus affinis coronavirus - AHX37569.1
21. Bat SARS-like coronavirus - AVP78031.1
22. Bat SARS-like coronavirus - AVP78042.1
23. Bat coronavirus - QHR63300.1
24. EPI_ISL_402131|bat/Yunnan/RaTG13/2013|...
25. EPI_ISL_402125|Wuhan-Hu-1/2019|MN908...
furin...

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

Hanging Text (not sure whether to include)

But this produces a paradox because without the adaptations it is not clear that there would need to be sufficient human infection and transmission to provide the opportunity for the virus to adapt. On the other hand if the virus was already readily transmissible prior to the acquisition of the furin cleavage site but was simply less pathogenic, and thus just less likely to be detected, then we would expect it to be widely disseminated.

February 1, 2020, 8:44 PM

The proximal origin of SARS-CoV-2

Since the first reports of a novel pneumonia (COVID-19) in Wuhan city, Hubei province, China there has been considerable discussion and uncertainty over the origin of the causative virus, SARS-CoV-2. Infections with SARS-CoV-2 are now widespread in China, with cases in every province. As of 14 February 2020, 64,473 cases have been confirmed, with 1,384 deaths attributed to the virus. These official case numbers are likely a large underestimate because of limited reporting of mild and asymptomatic cases, and the virus is clearly capable of efficient human-to-human transmission. Based on the possibility of spread to countries with weaker healthcare systems, the World Health Organization has declared the COVID-19 outbreak a Public Health Emergency of International Concern (PHEIC). There are currently neither vaccines nor specific treatments for this disease.

SARS-CoV-2 is the seventh member of the *Coronaviridae* that is known to infect humans. Three of these viruses, SARS CoV-1, MERS, and SARS-CoV-2, can cause severe disease; four, HKU1, NL63, OC43 and 229E, are associated with mild respiratory symptoms. Herein, we review what can be deduced about the origin and early evolution of SARS-CoV-2 from the comparative analysis of available genome sequence data. In particular, we offer a perspective of notable features in the SARS-CoV-2 genome, outline mechanisms for how the virus may have acquired them, and discuss scenarios by which these features could have arisen. Importantly, this analysis provides evidence that SARS-CoV-2 is not a laboratory construct nor an experimentally manipulated virus.

The genomic comparisons of both alpha- and betacoronaviruses (family *Coronaviridae*) described below identify two notable features of the SARS-CoV-2 genome: (i) based on structural modeling and early biochemical experiments, SARS-CoV-2 appears to be optimized for binding to the human ACE2 receptor; (ii) the highly variable spike protein of SARS-CoV-2 has a polybasic (furin) cleavage site at the S1 and S2 boundary via the insertion of twelve nucleotides. Additionally, this event also led to the acquisition of three predicted O-linked glycans around the polybasic cleavage site.

Mutations in the receptor binding domain of SARS-CoV-2

The receptor binding domain (RBD) in the spike protein of SARS-CoV and SARS-related coronaviruses is the most variable part of the virus genome. Six residues in the RBD appear to be critical for binding to the human ACE2 receptor and determining host range[1]. Using coordinates based on the Urbani strain of SARS-CoV, they are Y442, L472, N479, D480, T487, and Y491[1]. The corresponding residues in SARS-CoV-2 are L455, F486, Q493, S494, N501, and Y505. Five out of six of these residues are mutated in SARS-CoV-2 compared to its most closely related virus, RaTG13 sampled from a *Rhinolophus affinis* bat, to which it is ~96% identical[2] (**Figure 1a**). Based on modeling[1] and biochemical experiments[3,4], SARS-CoV-2 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[1]. In contrast, SARS-CoV-2 may bind less efficiently to ACE2 in other species associated with SARS-like viruses, including rodents and civets[1].

A phenylalanine at F486 in SARS-CoV-2 corresponds to L472 in the SARS-CoV Urbani strain. In cell culture experiments the leucine at position 472 is mutated to phenylalanine (L472F)[5]; this mutation is predicted to be optimal for binding of the SARS-CoV RBD to the human ACE2 receptor[6]. However, a phenylalanine in this position is also present in several SARS-like CoVs from bats (**Figure 1a**). While these analyses suggest that SARS-CoV-2 may be capable of binding the human ACE2 receptor with high affinity, the interaction is not predicted to be optimal[1]. Additionally, several of the key residues in the RBD of SARS-CoV-2 are different from those previously described as optimal for human ACE2 receptor binding[6]. This is strong evidence that SARS-CoV-2 is *not* the product of genetic engineering.

February 15, 2020 4:52 PM

Polybasic cleavage site and O-linked glycans

The second notable feature of SARS-CoV-2 is a predicted polybasic cleavage site (RRAR) in the spike protein at the junction of S1 and S2, the two subunits of the spike (**Figure 1b**)[7,8]. In addition to two basic arginines and an alanine at the cleavage site, a leading proline is also inserted; thus, the fully inserted sequence is PRRA (**Figure 1b**). The strong turn created by the proline insertion is predicted to result in the addition of O-linked glycans to serine-673, threonine-678, and serine-686 that flank the polybasic cleavage site. A polybasic cleavage site has not previously been observed in the related lineage B betacoronaviruses and is a unique feature of SARS-CoV-2. Some human betacoronaviruses, including HCoV-HKU1 (lineage A), have polybasic cleavage sites, as well as predicted O-linked glycans near the S1/S2 cleavage site.

While the functional consequence of the polybasic cleavage site in SARS-CoV-2 is unknown, experiments with SARS-CoV have shown that engineering such a site at the S1/S2 junction enhances cell–cell fusion but does not affect virus entry[9]. Polybasic cleavage sites allow effective cleavage by furin and other proteases, and can be acquired at the junction of the two subunits of the hemagglutinin (HA) of avian influenza viruses in conditions that select for rapid virus replication and transmission (e.g., highly dense chicken populations). HA serves a similar function in cell-cell fusion and viral entry as the coronavirus spike. Acquisition of a polybasic cleavage site in HA, by insertion or recombination, converts low pathogenicity avian influenza viruses to highly pathogenic avian influenza viruses[10–12]. The acquisition of polybasic cleavage sites by influenza virus HA has also been observed after repeated forced passage in cell culture or through animals[13,14]. An avirulent isolate of Newcastle Disease virus became highly pathogenic during serial passage in chickens by incremental acquisition of a polybasic cleavage site at the junction of its fusion protein subunits[15]. A potential function of the three predicted O-linked glycans is less clear, but could create a "mucin-like domain" that would shield potential epitopes or key residues on the SARS-CoV-2 spike protein. Biochemical analyses or structural studies are required to determine whether or not the predicted O-linked glycan sites are utilized.

Theories of SARS-CoV-2 origins

It is unlikely that SARS-CoV-2 emerged through laboratory manipulation of an existing SARS-related coronavirus. As noted above, the RBD of SARS-CoV-2 is not optimized for human ACE2 receptor binding. Furthermore, if genetic manipulation had been performed, one would expect that a researcher would have used one of the several reverse genetic systems available for betacoronaviruses. However, this is not the case as the genetic data shows that SARS-CoV-2 is not derived from any previously used virus backbone[16]. Instead, we propose two scenarios that can plausibly explain the origin of SARS-CoV-2: (1) natural selection in a non-human animal host prior to zoonotic transfer, (2) natural selection in humans following zoonotic transfer. We also discuss whether selection during passage in culture could have given rise to the same observed features.

Selection in an animal host

Given the similarity of SARS-CoV-2 to bat SARS-like CoVs, particularly RaTG13, it is plausible that bats serve as reservoir hosts for the SARS-CoV-2. Indeed, as many of the early cases were linked to the Huanan seafood and wildlife market in Wuhan, it is possible that a bat source was present at this location. It is important, however, to consider that previous human outbreaks caused by betacoronaviruses have involved direct human exposures to animals other than bats, including civets (SARS) and camels (MERS) that carry viruses that are genetically very similar to SARS-CoV-1 or MERS-CoV, respectively. By analogy to SARS-CoV-1 and MERS-CoV, viruses closely related to SARS-Cov-2 may be circulating in one or more animal species. Provisional analyses indicate that Malayan pangolins (*Manis javanica*) illegally imported into Guangdong province contain a CoV that is similar to SARS-CoV-2 [17]. Although the bat virus RaTG13 remains the closest relative to SARS-CoV-2 across the whole genome, the Malayan pangolin CoV is identical to SARS-CoV-2 at all six key RBD residues (**Figure 1**). However, no pangolin CoV has yet been identified that is

February 15, 2020 4:52 PM

sufficiently similar to the SARS-CoV-2 across its entire genome to support direct human infection by a pangolin CoV. In addition, the pangolin CoV described to date does not carry a polybasic cleavage site insertion. For a precursor virus to acquire the polybasic cleavage site and mutations in the spike proteins that appear to be suitable for human ACE2 receptor binding, an animal host would likely 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. Further characterization of CoVs in pangolins and other animals that may harbor SARS-CoV-like viruses should be a public health priority.

Cryptic adaptation to humans

The second scenario is that a progenitor to SARS-CoV-2 jumped from a non-human animal to humans with the genomic features described above then acquired through adaptation during human-to-human transmission. We surmise that once these adaptations were acquired (either together or in series) it would enable the outbreak to take off, producing a sufficiently large and unusual cluster of pneumonia cases to trigger the surveillance system that ultimately detected it.

All SARS-CoV-2 genomes sequenced so far have the well adapted RBD and the polybasic cleavage site, and are thus derived from a common ancestor that had these features. The presence of an RBD in pangolins that is very similar to the one in SARS-CoV-2 means that this was likely already present in the virus that jumped to humans, even if we don't yet have the exact non-human progenitor virus. This leaves the polybasic cleavage site insertion to occur during human-to-human transmission. Following the example of the HA gene of influenza A virus, this requires a specific insertion or recombination event to enable the emergence of SARS-CoV-2 as an epidemic pathogen.

Estimates of the timing of the most recent common ancestor (tMRCA) of SARS-CoV-2 using currently available genome sequence data point to virus emergence in late November to early December 2019[18,19], compatible with the earliest retrospectively confirmed cases[20]. So, this scenario presumes a phase of unrecognised transmission in humans between the initial zoonotic transfer event and the acquisition of the polybasic cleavage site. Sufficient opportunity could occur if there had been many prior zoonotic events producing short chains of human-to-human transmission (so-called 'stuttering chains') over an extended period. This is essentially the situation for MERS-CoV in the Arabian Peninsula where all the human cases are the result of repeated jumps of the virus from dromedary camels, producing single infections or short chains of transmission that eventually resolve. To date, after 2,499 cases over 8 years, no human adaptation has emerged that has allowed the virus to take hold in the human population.

How could we test whether cryptic spread of SARS-CoV-2 enabled adaptation to humans? Metagenomic studies of banked serum samples could provide important information, but given the relatively short period of viremia it may be impossible to detect low level SARS-CoV-2 circulation in historical samples. Retrospective serological studies potentially could be informative and a few such studies have already been conducted. One such study found that traders involved in animal importation had a 13% seropositivity to coronaviruses[21], while another found that 3% residents of a village in Southern China were seropositive to these viruses[22]. Interestingly, 200 residents of Wuhan did not show coronavirus seroreactivity. Critically, however, these studies could not have distinguished whether positive serological responses were due to a prior infection with SARS-CoV-1 or -2. Further retrospective serological studies should be conducted to determine the extent of prior human exposure to betacoronaviruses in different geographic areas, particularly using assays that can distinguish among multiple betacoronaviruses.

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 in multiple laboratories across the world[23–26]. There are also documented instances of laboratory acquisition of SARS-CoV-1 by laboratory personnel working under

February 15, 2020 4:52 PM

BSL-2 containment[27,28]. We must consider, therefore, the possibility of a deliberate or inadvertent release of SARS-CoV-2. In theory, it is possible that SARS-CoV-2 acquired the observed RBD mutations site during adaptation to passage in cell culture, as has been observed in studies with SARS-CoV[5] as well as MERS-CoV[29]. However, the acquisition of the polybasic cleavage site or O-linked glycans - if functional - argues against this scenario. New polybasic cleavage sites have only been observed after prolonged passaging of low pathogenicity avian influenza virus in cell culture or animals. Furthermore, the generation of SARS-CoV-2 by cell culture or animal passage would have required prior isolation of a progenitor virus with a very high genetic similarity. Subsequent generation of a polybasic cleavage site would have then required an intense program of passage in cell culture or animals with ACE-2 receptor similar to humans (e.g. ferrets). It is also questionable whether generation of the O-linked glycans would have occurred on cell culture passage, as such mutations typically suggest the involvement of an immune system, that is not present *in vitro*.

Conclusions

In the midst of the global COVID-19 public health emergency it is reasonable to wonder why the origins of the epidemic matter. A detailed understanding of how an animal virus jumped species boundaries to infect humans so productively will undoubtedly help in the prevention of future zoonotic events. For example, if SARS-CoV-2 pre-adapted in another animal species then we are at risk of future reemergence events, even if the current epidemic is controlled. In contrast, if the adaptive process we describe occurred in humans then even if we have repeated zoonotic transfers they are unlikely to take off unless the same series of mutations occur. In addition, identifying the closest animal relatives of SARS-CoV-2 will greatly assist studies of virus function. Indeed, the availability of the RaTG13 bat sequence facilitated the comparative genomic analysis performed, helping to reveal the key mutations in the RBD as well as the polybasic cleavage site insertion.

The genomic features described here may in part explain the infectiousness and transmissibility of SARS-CoV-2 in humans. Although genomic evidence does not support the idea that SARS-CoV-2 is a laboratory construct, it is currently impossible to prove or disprove the other theories of its origin outlined above, 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 revealing virus origins. In addition, it would be helpful to obtain additional genetic and functional data about the virus, including experimental studies of receptor binding and the role of the polybasic cleavage site and predicted O-linked glycans. The identification of a potential intermediate host of SARS-CoV-2, as well as sequencing of very early cases including those not connected to the market, would similarly be highly informative.

February 15, 2020 4:52 PM

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. Wu, F. *et al.* A new coronavirus associated with human respiratory disease in China. *Nature* (2020) doi:10.1038/s41586-020-2008-3.
3. 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.
4. 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.
5. 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).
6. Cui, J., Li, F. & Shi, Z.-L. Origin and evolution of pathogenic coronaviruses. *Nat. Rev. Microbiol.* **17**, 181–192 (2019).
7. Gallaher, W. Analysis of Wuhan Coronavirus: Deja Vu. *Virological* http://virological.org/t/analysis-of-wuhan-coronavirus-deja-vu/357 (2020).
8. Coutard, B. *et al.* The spike glycoprotein of the new coronavirus 2019-nCoV contains a furin-like cleavage site absent in CoV of the same clade. *Antiviral Res.* 104742 (2020).
9. 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).
10. 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).

February 15, 2020 4:52 PM

11. Alexander, D. J. & Brown, I. H. History of highly pathogenic avian influenza. *Rev. Sci. Tech.* **28**, 19–38 (2009).

12. 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).

13. Ito, T. *et al.* Generation of a highly pathogenic avian influenza A virus from an avirulent field isolate by passaging in chickens. *J. Virol.* **75**, 4439–4443 (2001).

14. Li, S. Q., Orlich, M. & Rott, R. Generation of seal influenza virus variants pathogenic for chickens, because of hemagglutinin cleavage site changes. *J. Virol.* **64**, 3297–3303 (1990).

15. Shengqing, Y. *et al.* Generation of velogenic Newcastle disease viruses from a nonpathogenic waterfowl isolate by passaging in chickens. *Virology* **301**, 206–211 (2002).

16. Menachery, V. D. *et al.* A SARS-like cluster of circulating bat coronaviruses shows potential for human emergence. *Nat. Med.* **21**, 1508–1513 (2015).

17. Wong, M. C., Javornik Cregeen, S. J., Ajami, N. J. & Petrosino, J. F. Evidence of recombination in coronaviruses implicating pangolin origins of nCoV-2019. *bioRxiv* 2020.02.07.939207 (2020) doi:10.1101/2020.02.07.939207.

18. Phylodynamic Analysis | 90 genomes | 12 Feb 2020 – Novel 2019 coronavirus / nCoV-2019 Genomic Epidemiology. http://virological.org/t/phylodynamic-analysis-90-genomes-12-feb-2020/356 (2020).

19. Phylodynamic estimation of incidence and prevalence of novel coronavirus (nCoV) infections through time – Novel 2019 coronavirus / nCoV-2019 Genomic Epidemiology. http://virological.org/t/phylodynamic-estimation-of-incidence-and-prevalence-of-novel-coronavirus-ncov-infections-through-time/391 (2020).

20. Huang, C. *et al.* Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet* (2020) doi:10.1016/S0140-6736(20)30183-5.

21. Centers for Disease Control and Prevention (CDC). Prevalence of IgG antibody to SARS-associated

February 15, 2020 4:52 PM

coronavirus in animal traders--Guangdong Province, China, 2003. *MMWR Morb. Mortal. Wkly. Rep.* **52**, 986–987 (2003).

22. Wang, N. *et al.* Serological Evidence of Bat SARS-Related Coronavirus Infection in Humans, China. *Virol. Sin.* **33**, 104–107 (2018).
23. Ge, X.-Y. *et al.* Isolation and characterization of a bat SARS-like coronavirus that uses the ACE2 receptor. *Nature* **503**, 535–538 (2013).
24. 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).
25. 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).
26. 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).
27. Lim, P. L. *et al.* Laboratory-acquired severe acute respiratory syndrome. *N. Engl. J. Med.* **350**, 1740–1745 (2004).
28. Senior, K. Recent Singapore SARS case a laboratory accident. *Lancet Infect. Dis.* **3**, 679 (2003).
29. Letko, M. *et al.* Adaptive Evolution of MERS-CoV to Species Variation in DPP4. *Cell Rep.* **24**, 1730–1737 (2018).

February 15, 2020 4:52 PM

Figures

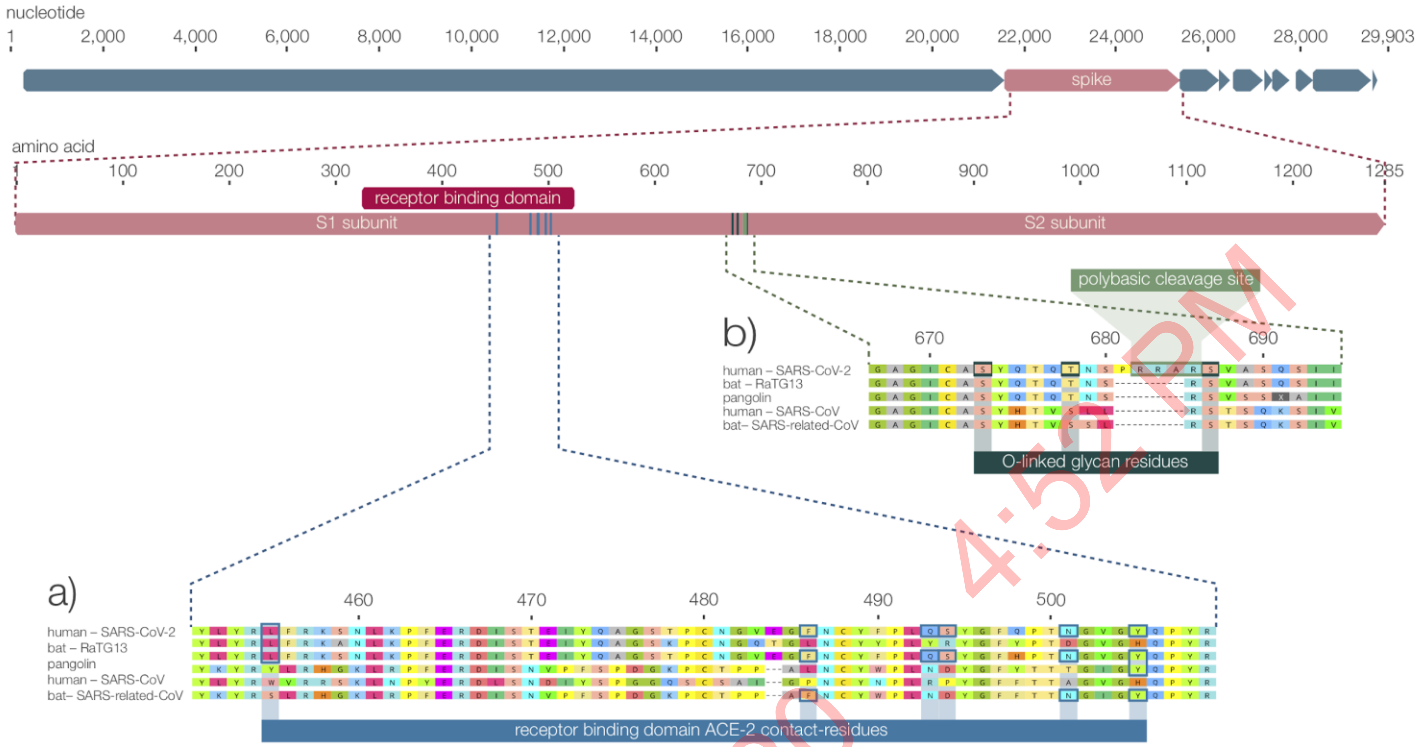


Figure 1 | a) Mutations in contact residues of the SARS-CoV-2 spike protein. The spike protein of SARS-CoV-2 (bottom) was aligned against the most closely related SARS and SARS-like CoVs. Key residues in the spike protein that make contact to the ACE2 receptor have been marked with blue boxes in both SARS-CoV-2 and the SARS-CoV Urbani strain. **b) Acquisition of polybasic cleavage site and O-linked glycans.** The spike protein of SARS-CoV-2 (bottom) was aligned against the most closely related SARS and SARS-like CoVs. The polybasic cleavage site is marked in grey with the three adjacent predicted O-linked glycans in blue. Both the polybasic cleavage site and O-linked glycans are unique to SARS-CoV-2 and not previously seen in lineage B betacoronaviruses.

Hanging Text (not sure whether to include)

But this produces a paradox because without the adaptations it is not clear that there would need to be sufficient human infection and transmission to provide the opportunity for the virus to adapt. On the other hand if the virus was already readily transmissible prior to the acquisition of the polybasic cleavage site but was simply less pathogenic, and thus just less likely to be detected, then we would expect it to be widely disseminated.

February 15, 2020 4:52 PM

The Proximal Origin of SARS-CoV-2

Kristian G. Andersen[1,2*], Andrew Rambaut[3], W. Ian Lipkin[4], Edward C. Holmes[5] & Robert F. Garry[6,7]

[1]Department of Immunology and Microbiology, The Scripps Research Institute, La Jolla, CA 92037, USA.

[2]Scripps Research Translational Institute, La Jolla, CA 92037, USA.

[3]Institute of Evolutionary Biology, University of Edinburgh, Edinburgh, UK.

[4]Center for Infection and Immunity, Mailman School of Public Health of Columbia University, New York, New York, USA.

[5]Marie Bashir Institute for Infectious Diseases and Biosecurity, School of Life and Environmental Sciences and School of Medical Sciences, The University of Sydney, Sydney, Australia.

[6]Tulane University, School of Medicine, Department of Microbiology and Immunology, New Orleans, LA, USA.

[7]Zalgen Labs, LCC, Germantown, MD, USA.

*Corresponding author:

Kristian G. Andersen

Department of Immunology and Microbiology,

The Scripps Research Institute,

La Jolla, CA 92037, USA.

February 17, 2020 2:32 PM

Since the first reports of a novel pneumonia (COVID-19) in Wuhan city, Hubei province, China there has been considerable discussion and uncertainty over the origin of the causative virus, SARS-CoV-2. Infections with SARS-CoV-2 are now widespread in China, with cases in every province. As of 14 February 2020, 64,473 such cases have been confirmed, with 1,384 deaths attributed to the virus. These official case numbers are likely an underestimate because of limited reporting of mild and asymptomatic cases, and the virus is clearly capable of efficient human-to-human transmission. Based on the possibility of spread to countries with weaker healthcare systems, the World Health Organization has declared the COVID-19 outbreak a Public Health Emergency of International Concern (PHEIC). There are currently neither vaccines nor specific treatments for this disease.

SARS-CoV-2 is the seventh member of the *Coronaviridae* known to infect humans. Three of these viruses, SARS CoV-1, MERS CoV, and SARS-CoV-2, can cause severe disease; four, HKU1, NL63, OC43 and 229E, are associated with mild respiratory symptoms. Herein, we review what can be deduced about the origin and early evolution of SARS-CoV-2 from the comparative analysis of available genome sequence data. In particular, we offer a perspective on the notable features in the SARS-CoV-2 genome and discuss scenarios by which these features could have arisen. Importantly, this analysis provides evidence that SARS-CoV-2 is not a laboratory construct nor a purposefully manipulated virus.

The genomic comparison of both alpha- and betacoronaviruses (family *Coronaviridae*) described below identifies two notable features of the SARS-CoV-2 genome: (i) based on structural modeling and early biochemical experiments, SARS-CoV-2 appears to be optimized for binding to the human ACE2 receptor; (ii) the highly variable spike (S) protein of SARS-CoV-2 has a polybasic (furin) cleavage site at the S1 and S2 boundary via the insertion of twelve nucleotides. Additionally, this event led to the acquisition of three predicted O-linked glycans around the polybasic cleavage site.

Mutations in the receptor binding domain of SARS-CoV-2

The receptor binding domain (RBD) in the spike protein of SARS-CoV and SARS-related coronaviruses is the most variable part of the virus genome. Six residues in the RBD appear to be critical for binding to the human ACE2 receptor and determining host range[1]. Using coordinates based on the Urbani strain of SARS-CoV, they are Y442, L472, N479, D480, T487, and Y491[1]. The corresponding residues in SARS-CoV-2 are L455, F486, Q493, S494, N501, and Y505. Five of these six residues are mutated in SARS-CoV-2 compared to its most closely related virus, RaTG13 sampled from a *Rhinolophus affinis* bat, to which it is ~96% identical[2] (**Figure 1a**). Based on modeling[1] and biochemical experiments[3,4], SARS-CoV-2 seems to have an RBD that may bind with high affinity to ACE2 from human, non-human primate, ferret, pig, and cat, as well as other species with high receptor homology[1]. In contrast, SARS-CoV-2 may bind less efficiently to ACE2 in other species associated with SARS-like viruses, including rodents and civets[1].

The phenylalanine (F) at residue 486 in the SARS-CoV-2 S protein corresponds to L472 in the SARS-CoV Urbani strain. Notably, in SARS-CoV cell culture experiments the L472 mutates to phenylalanine (L472F)[5], which is predicted to be optimal for binding of the SARS-CoV RBD to the human ACE2 receptor[6]. However, a phenylalanine in this position is also present in several SARS-like CoVs from bats (**Figure 1a**). While these analyses suggest that SARS-CoV-2 may be capable of binding the human ACE2 receptor with high affinity, the interaction is not predicted to be optimal[1]. Additionally, several of the key residues in the RBD of SARS-CoV-2 are different to those previously described as optimal for human ACE2 receptor binding[6]. In contrast to these computational predictions, recent binding studies indicate that SARS-CoV-2 binds with high affinity to human ACE2[7]. Thus the SARS-CoV-2 spike appears to be the result of selection on human or human-like ACE2 permitting another optimal binding solution to arise. This is strong evidence that SARS-CoV-2 is *not* the product of genetic engineering.

Polybasic cleavage site and O-linked glycans

The second notable feature of SARS-CoV-2 is a predicted polybasic cleavage site (RRAR) in the spike protein at the junction of S1 and S2, the two subunits of the spike protein (**Figure 1b**)[8,9]. In addition to two basic arginines and an alanine at the cleavage site, a leading proline is also inserted; thus, the fully inserted sequence is PRRA (**Figure 1b**). The strong turn created by the proline insertion is predicted to result in the addition of O-linked glycans to S673, T678, and S686 that flank the polybasic cleavage site. A polybasic cleavage site has not previously been observed in related lineage B betacoronaviruses and is a unique feature of SARS-CoV-2. Some human betacoronaviruses, including HCoV-HKU1 (lineage A), have polybasic cleavage sites, as well as predicted O-linked glycans near the S1/S2 cleavage site.

While the functional consequence of the polybasic cleavage site in SARS-CoV-2 is unknown, experiments with SARS-CoV have shown that engineering such a site at the S1/S2 junction enhances cell–cell fusion but does not affect virus entry[10]. Polybasic cleavage sites allow effective cleavage by furin and other proteases, and can be acquired at the junction of the two subunits of the haemagglutinin (HA) protein of avian influenza viruses in conditions that select for rapid virus replication and transmission (e.g.,. highly dense chicken populations). HA serves a similar function in cell-cell fusion and viral entry as the coronavirus S protein. Acquisition of a polybasic cleavage site in HA, by either insertion or recombination, converts low pathogenicity avian influenza viruses into highly pathogenic forms[11–13]. The acquisition of polybasic

February 17, 2020 2:32 PM

cleavage sites by the influenza virus HA has also been observed after repeated forced passage in cell culture or through animals[14,15]. Similarly, an avirulent isolate of Newcastle Disease virus became highly pathogenic during serial passage in chickens by incremental acquisition of a polybasic cleavage site at the junction of its fusion protein subunits[16]. The potential function of the three predicted O-linked glycans is less clear, but they could create a "mucin-like domain" that would shield potential epitopes or key residues on the SARS-CoV-2 spike protein. Biochemical analyses or structural studies are required to determine whether or not the predicted O-linked glycan sites are utilized.

Theories of SARS-CoV-2 origins

It is improbable that SARS-CoV-2 emerged through laboratory manipulation of an existing SARS-related coronavirus. As noted above, the RBD of SARS-CoV-2 is optimized for human ACE2 receptor binding with an efficient binding solution different to that which would have been predicted. Further, if genetic manipulation had been performed, one would expect that one of the several reverse genetic systems available for betacoronaviruses would have been used. However, this is not the case as the genetic data shows that SARS-CoV-2 is not derived from any previously used virus backbone[17]. Instead, we propose two scenarios that can plausibly explain the origin of SARS-CoV-2: (i) natural selection in a non-human animal host prior to zoonotic transfer, and (ii) natural selection in humans following zoonotic transfer. We also discuss whether selection during passage in culture could have given rise to the same observed features.

Selection in an animal host

As many of the early cases of COVID-19 were linked to the Huanan seafood and wildlife market in Wuhan, it is possible that an animal source was present at this location. Given the similarity of SARS-CoV-2 to bat SARS-like CoVs, particularly RaTG13, it is plausible that bats serve as reservoir hosts for SARS-CoV-2. It is important, however, to note that previous outbreaks of betacoronaviruses in humans involved direct exposure to animals other than bats, including civets (SARS) and camels (MERS), that carry viruses that are genetically very similar to SARS-CoV or MERS-CoV, respectively. By analogy, viruses closely related to SARS-Cov-2 may be circulating in one or more animal species. Initial analyses indicate that Malayan pangolins (*Manis javanica*) illegally imported into Guangdong province contain a CoV that is similar to SARS-CoV-2[18]. Although the bat virus RaTG13 remains the closest relative to SARS-CoV-2 across the whole genome, the Malayan pangolin CoV is identical to SARS-CoV-2 at all six key RBD residues (**Figure 1**). However, no pangolin CoV has yet been identified that is sufficiently similar to SARS-CoV-2 across its entire genome to support direct human infection. In addition, the pangolin CoV does not carry a polybasic cleavage site insertion. For a precursor virus to acquire the polybasic cleavage site and mutations in the spike protein suitable for human ACE2 receptor binding, an animal host would likely have to have a high population density – to allow natural selection to proceed efficiently – and an ACE2 gene that is similar to the human orthologue. Further characterization of CoVs in pangolins and other animals that may harbour SARS-CoV-like viruses should be a public health priority.

Cryptic adaptation to humans

It is also possible that a progenitor to SARS-CoV-2 jumped from a non-human animal to humans, with the genomic features described above acquired through adaptation during subsequent human-to-human transmission. We surmise that once these adaptations were acquired (either together or in series) it would enable the outbreak to take-off, producing a sufficiently large and unusual cluster of pneumonia cases to trigger the surveillance system that ultimately detected it.

All SARS-CoV-2 genomes sequenced so far have the well adapted RBD and the polybasic cleavage site and are thus derived from a common ancestor that had these features. The presence of an RBD in pangolins that is very similar to the one in SARS-CoV-2 means that this was likely already present in the virus that jumped to humans, even if we don't yet have the exact non-human progenitor virus. This leaves the

February 17, 2020 2:32 PM

polybasic cleavage site insertion to occur during human-to-human transmission. Following the example of the influenza A virus HA gene, a specific insertion or recombination event is required to enable the emergence of SARS-CoV-2 as an epidemic pathogen.

Estimates of the timing of the most recent common ancestor (tMRCA) of SARS-CoV-2 using currently available genome sequence data point to virus emergence in late November to early December 2019[19,20], compatible with the earliest retrospectively confirmed cases[21]. Hence, this scenario presumes a period of unrecognised transmission in humans between the initial zoonotic transfer event and the acquisition of the polybasic cleavage site. Sufficient opportunity could occur if there had been many prior zoonotic events producing short chains of human-to-human transmission (so-called 'stuttering chains') over an extended period. This is essentially the situation for MERS-CoV in the Arabian Peninsula where all the human cases are the result of repeated jumps of the virus from dromedary camels, producing single infections or short chains of transmission that eventually resolve. To date, after 2,499 cases over 8 years, no human adaptation has emerged that has allowed MERS-CoV to take hold in the human population.

How could we test whether cryptic spread of SARS-CoV-2 enabled human adaptation? Metagenomic studies of banked serum samples could provide important information, but given the relatively short period of viremia it may be impossible to detect low level SARS-CoV-2 circulation in historical samples. Retrospective serological studies potentially could be informative and a few such studies have already been conducted. One found that animal importation traders had a 13% seropositivity to coronaviruses[22], while another noted that 3% residents of a village in Southern China were seropositive to these viruses[23]. Interestingly, 200 residents of Wuhan did not show coronavirus seroreactivity. Critically, however, these studies could not have distinguished whether positive serological responses were due to a prior infection with SARS-CoV or SARS-CoV-2. Further retrospective serological studies should be conducted to determine the extent of prior human exposure to betacoronaviruses in different geographic areas, particularly using assays that can distinguish among multiple betacoronaviruses.

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 in multiple laboratories across the world[24–27]. There are also documented instances of the laboratory acquisition of SARS-CoV by laboratory personnel working under BSL-2 containment[28,29]. We must therefore consider the possibility of a deliberate or inadvertent release of SARS-CoV-2. In theory, it is possible that SARS-CoV-2 acquired the observed RBD mutations site during adaptation to passage in cell culture, as has been observed in studies with SARS-CoV[5] as well as MERS-CoV[30]. However, the acquisition of the polybasic cleavage site or O-linked glycans - if functional - argues against this scenario. New polybasic cleavage sites have only been observed after prolonged passaging of low pathogenicity avian influenza virus in cell culture or animals. Furthermore, the generation of SARS-CoV-2 by cell culture or animal passage would have required prior isolation of a progenitor virus with a very high genetic similarity. Subsequent generation of a polybasic cleavage site would have then required an intense program of passage in cell culture or animals with ACE-2 receptor similar to humans (e.g. ferrets). It is also questionable whether generation of the O-linked glycans would have occurred on cell culture passage, as such mutations typically suggest the involvement of an immune system, that is not present *in vitro*.

Conclusions

In the midst of the global COVID-19 public health emergency it is reasonable to wonder why the origins of the epidemic matter. A detailed understanding of how an animal virus jumped species boundaries to infect humans so productively will help in the prevention of future zoonotic events. For example, if SARS-CoV-2 pre-adapted in another animal species then we are at risk of future re-emergence events even if the current

February 17, 2020 2:32 PM

epidemic is controlled. In contrast, if the adaptive process we describe occurred in humans, then even if we have repeated zoonotic transfers they are unlikely to take-off unless the same series of mutations occurs. In addition, identifying the closest animal relatives of SARS-CoV-2 will greatly assist studies of virus function. Indeed, the availability of the RaTG13 bat sequence facilitated the comparative genomic analysis performed here, helping to reveal the key mutations in the RBD as well as the polybasic cleavage site insertion.

The genomic features described here may in part explain the infectiousness and transmissibility of SARS-CoV-2 in humans. Although genomic evidence does not support the idea that SARS-CoV-2 is a laboratory construct, it is currently impossible to prove or disprove the other theories of its origin described here, and it is unclear whether future data 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 revealing virus origins. In addition, it would be helpful to obtain more genetic and functional data about the virus, including experimental studies of receptor binding and the role of the polybasic cleavage site and predicted O-linked glycans. The identification of a potential intermediate host of SARS-CoV-2, as well as the sequencing of very early cases including those not connected to the Wuhan market, would similarly be highly informative. Irrespective of how SARS-CoV-2 originated, the ongoing surveillance of pneumonia in humans and other animals is clearly of utmost importance.

Acknowledgements

We thank all those who have contributed SARS-CoV-2 genome sequences to the GISAID database (https://www.gisaid.org/) and contributed analyses and ideas to Virological.org (http://virological.org/). We thank M. Farzan for discussions. We thank the Wellcome Trust for supporting this work. ECH is supported by an ARC Australian Laureate Fellowship (FL170100022). KGA is a Pew Biomedical Scholar and is supported by NIH NIAID grant U19AI135995. AR is supported by the Wellcome Trust (Collaborators Award 206298/Z/17/Z – ARTIC network) and the European Research Council (grant agreement no. 725422 – ReservoirDOCS). RFG is supported by NIH grants U19AI135995, U54 HG007480 and U19AI142790.

Figure Legends

Figure 1 | a) Mutations in contact residues of the SARS-CoV-2 spike protein. The spike protein of SARS-CoV-2 (top) was aligned against the most closely related SARS-like CoVs and SARS-CoV-1. Key residues in the spike protein that make contact to the ACE2 receptor are marked with blue boxes in both SARS-CoV-2 and the SARS-CoV Urbani strain. **b) Acquisition of polybasic cleavage site and O-linked glycans**. The polybasic cleavage site is marked in grey with the three adjacent predicted O-linked glycans in blue. Both the polybasic cleavage site and O-linked glycans are unique to SARS-CoV-2 and not previously seen in lineage B betacoronaviruses. Sequences shown are from NCBI GenBank, accession numbers MN908947, MN996532, AY278741, KY417146 & MK211376. The pangolin coronavirus sequences are a consensus generated from SRR10168377 and SRR10168378 (NCBI BioProject PRJNA573298).

February 17, 2020 2:32 PM

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. Wu, F. *et al.* A new coronavirus associated with human respiratory disease in China. *Nature* (2020) doi:10.1038/s41586-020-2008-3.

3. 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.

4. 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.

5. 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).

6. Cui, J., Li, F. & Shi, Z.-L. Origin and evolution of pathogenic coronaviruses. *Nat. Rev. Microbiol.* **17**, 181–192 (2019).

7. Wrapp, D. *et al.* Cryo-EM Structure of the 2019-nCoV Spike in the Prefusion Conformation. *bioRxiv* 2020.02.11.944462 (2020) doi:10.1101/2020.02.11.944462.

8. Gallaher, W. Analysis of Wuhan Coronavirus: Deja Vu. *Virological* http://virological.org/t/analysis-of-wuhan-coronavirus-deja-vu/357 (2020).

9. Coutard, B. *et al.* The spike glycoprotein of the new coronavirus 2019-nCoV contains a furin-like cleavage site absent in CoV of the same clade. *Antiviral Res.* 104742 (2020).

10. 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).

February 17, 2020 2:32 PM

11. 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).

12. Alexander, D. J. & Brown, I. H. History of highly pathogenic avian influenza. *Rev. Sci. Tech.* **28**, 19–38 (2009).

13. 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).

14. Ito, T. *et al.* Generation of a highly pathogenic avian influenza A virus from an avirulent field isolate by passaging in chickens. *J. Virol.* **75**, 4439–4443 (2001).

15. Li, S. Q., Orlich, M. & Rott, R. Generation of seal influenza virus variants pathogenic for chickens, because of hemagglutinin cleavage site changes. *J. Virol.* **64**, 3297–3303 (1990).

16. Shengqing, Y. *et al.* Generation of velogenic Newcastle disease viruses from a nonpathogenic waterfowl isolate by passaging in chickens. *Virology* **301**, 206–211 (2002).

17. Menachery, V. D. *et al.* A SARS-like cluster of circulating bat coronaviruses shows potential for human emergence. *Nat. Med.* **21**, 1508–1513 (2015).

18. Wong, M. C., Javornik Cregeen, S. J., Ajami, N. J. & Petrosino, J. F. Evidence of recombination in coronaviruses implicating pangolin origins of nCoV-2019. *bioRxiv* 2020.02.07.939207 (2020) doi:10.1101/2020.02.07.939207.

19. Phylodynamic Analysis | 90 genomes | 12 Feb 2020 – Novel 2019 coronavirus / nCoV-2019 Genomic Epidemiology. http://virological.org/t/phylodynamic-analysis-90-genomes-12-feb-2020/356 (2020).

20. Phylodynamic estimation of incidence and prevalence of novel coronavirus (nCoV) infections through time – Novel 2019 coronavirus / nCoV-2019 Genomic Epidemiology. http://virological.org/t/phylodynamic-estimation-of-incidence-and-prevalence-of-novel-coronavirus-ncov-infections-through-time/391 (2020).

21. Huang, C. *et al.* Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China.

February 17, 2020 2:32 PM

Lancet (2020) doi:10.1016/S0140-6736(20)30183-5.

22. Centers for Disease Control and Prevention (CDC). Prevalence of IgG antibody to SARS-associated coronavirus in animal traders--Guangdong Province, China, 2003. *MMWR Morb. Mortal. Wkly. Rep.* **52**, 986–987 (2003).

23. Wang, N. *et al.* Serological Evidence of Bat SARS-Related Coronavirus Infection in Humans, China. *Virol. Sin.* **33**, 104–107 (2018).

24. Ge, X.-Y. *et al.* Isolation and characterization of a bat SARS-like coronavirus that uses the ACE2 receptor. *Nature* **503**, 535–538 (2013).

25. 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).

26. 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).

27. 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).

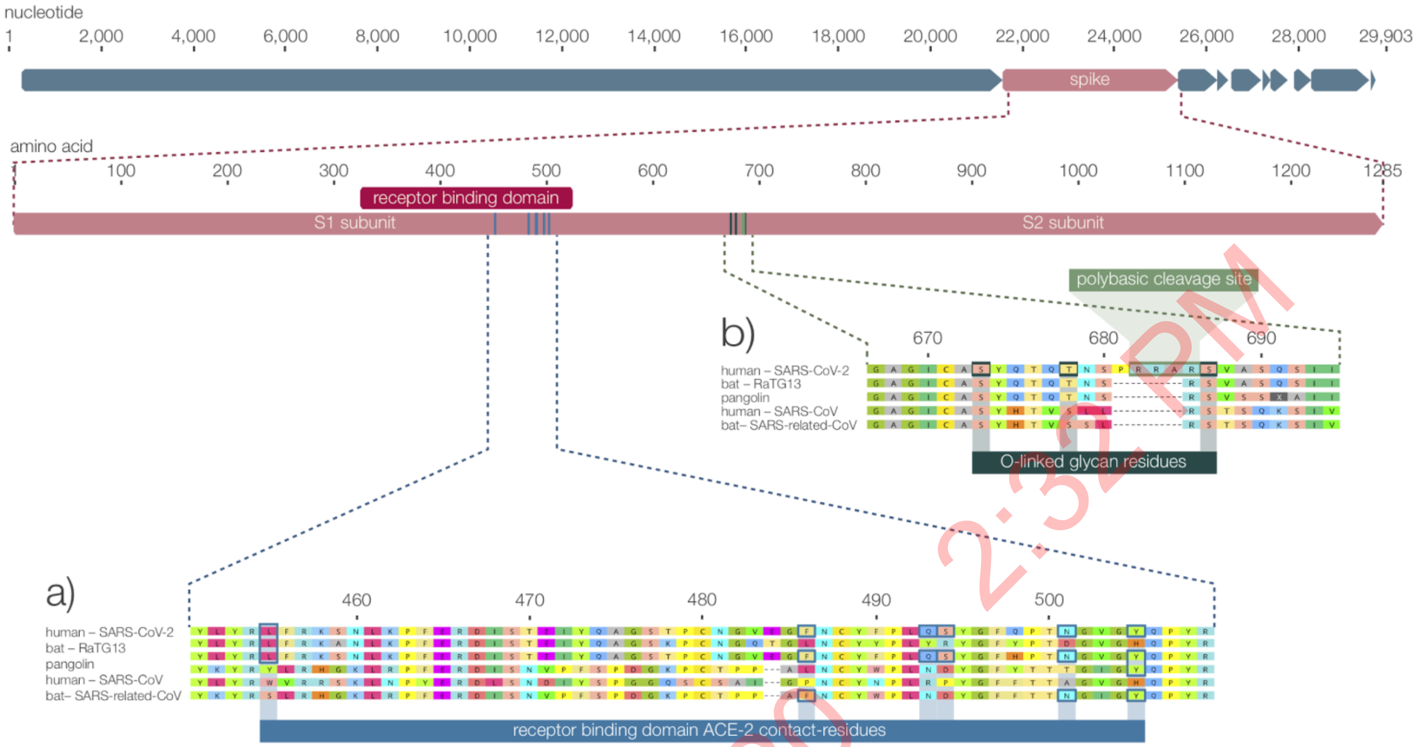
28. Lim, P. L. *et al.* Laboratory-acquired severe acute respiratory syndrome. *N. Engl. J. Med.* **350**, 1740–1745 (2004).

29. Senior, K. Recent Singapore SARS case a laboratory accident. *Lancet Infect. Dis.* **3**, 679 (2003).

30. Letko, M. *et al.* Adaptive Evolution of MERS-CoV to Species Variation in DPP4. *Cell Rep.* **24**, 1730–1737 (2018).

February 17, 2020 2:32 PM

Figure 1.


nucleotide
1
2,000
4,000
6,000
8,000
10,000
12,000
14,000
16,000
18,000
20,000
22,000
24,000
26,000
28,000
29,903
spike
amino acid
1
100
200
300
400
500
600
700
800
900
1000
1100
1200
1285
receptor binding domain
S1 subunit
S2 subunit
polybasic cleavage site
b)
670
680
690
human – SARS-CoV-2
bat – RaTG13
pangolin
human – SARS-CoV
bat– SARS-related-CoV
G A G I C A S Y Q T Q T N S P R R A R S V A S Q S I I
G A G I C A S Y Q T Q T N S ---- R S V A S Q S I I
G A G I C A S Y Q T Q T N S ---- R S V S S X A I I
G A G I C A S Y H T V S L L ---- R S T S Q K S I V
G A G I C A S Y H T V S S L ---- R S T S Q K S I V
O-linked glycan residues
a)
460
470
480
490
500
human – SARS-CoV-2
bat – RaTG13
pangolin
human – SARS-CoV
bat– SARS-related-CoV
Y L Y R L F R K S N L K P F E R D I S T E I Y Q A G S T P C N G V E G F N C Y F P L Q S Y G F Q P T N G V G Y Q P Y R
Y L Y R L F R K A N L K P F E R D I S T E I Y Q A G S K P C N G Q T G L N C Y Y P L Y R Y G F Y P T D G V G H Q P Y R
Y L Y R L F R K S N L K P F E R D I S T E I Y Q A G S T P C N G V E G F N C Y F P L Q S Y G F H P T N G V G Y Q P Y R
Y K Y R Y L R H G K L R P F E R D I S N V P F S P D G K P C T P P --- A L N C Y W P L N D Y G F Y T T T G I G Y Q P Y R
Y L Y R W V R R S K L N P Y E R D L S N D I Y S P G G Q S C S A I --- G P N C Y N P L R P Y G F F T T A G V G H Q P Y R
Y K Y R S L R H G K L R P F E R D I S N V P F S P D G K P C T P P --- A F N C Y W P L N D Y G F F T T N G I G Y Q P Y R
receptor binding domain ACE-2 contact-residues

February 17, 2020 2:32 PM

Referee #1 (Remarks to the Author):

Anderson presented a timely manuscript to share their points of view about the origin of SARS-CoV-2. There are several rumors about the origin of this virus. However, these "hypotheses" are entirely based on very limited, if any, scientific evidences.

This reviewer sees most of the arguments raised by the authors are valid and convincing. However, the authors might want to consider these minor suggestions:

1. The sections for the RBD and cleavage site of Spike protein basically have summarized the existing findings from other recent publications. The authors might want to spell out that these two sections are review summaries. In addition, the author can present these two sections in a more condense format and save some space for something else (also see points 6 and 7 below)

[RESPONSE]

2. Fig. 1. This figure has 6 aligned sequences, but with only 5 sequence titles. The order of these titles are also not correct.

[RESPONSE]

3. Lines 170 -174. It is correct that no adaptive mutation has been found in the spike of MERS-CoV. Deletions in other ORF regions, however, were detected in some human MERS-CoV viruses (PMID: 26981770). In addition, the 29nt deletion of human SARS-CoV (PMID: 12958366) was suggested to have effects on host adaptation. The authors should also consider these findings. It is premature to say that this would not happen in SARS-CoV-2.

[RESPONSE]

4. Line 194. The accident at Singapore occurred in a BSL3, not BSL2, containment.

[RESPONSE]

5. Line 194. Laboratory escapes of SARS occurred in Singapore, China and Taiwan (PMID: 16830004).

[RESPONSE]

6. There are two recent reports about coronaviruses in pangolins (https://www.biorxiv.org/content/10.1101/2020.02.13.945485v1.full.pdf; https://www.biorxiv.org/content/10.1101/2020.02.08.939660v2.full.pdf). The authors might want to comments on these.

[RESPONSE]

7. Optional: Can the authors share their views on the possibility of having a lab escape of a natural coronavirus? This is also one of the hypotheses that have been extensively discussed. The reviewer understand that this is entirely a different topic, but any insights are welcomed.

[RESPONSE]

Nature Rebuttal February, 20, 2020 10:52 AM

Referee #2 (Remarks to the Author):

This is a perspective discussing evidence against a hypothetical lab origin of SARS-CoV-2. The paper addresses suboptimal composition of ACE2-binding sites in the RBD, 3 predicted O-linked glycosylation sites and a furin cleavage site in the glycoprotein that was speculated upon before.

The paper is itself interesting, but unnecessarily speculative. It’s not clear why the authors do not refute a hypothetical lab origin in their coming publication on the ancestors of SARS-CoV-2 in bats and pangolins. The tree showing diverse pangolin viruses has kindly been made available by some of the authors in GISAID. Once the authors publish their new pangolin sequences, a lab origin will be extremely unlikely. It is not clear why the authors rush with a speculative perspective if their central hypothesis can be supported by their own data. Please explain.

[RESPONSE]

Another critical aspect of this text is the complete lack of referencing to a potential debate on a hypothetical lab origin. Who said this, why is this considered a problem? There are indeed a few apparently uninformed statements claiming the virus may be a Chinese bioweapon, but is this really problematic on a larger scale? The central reason for issuing this text must be exhaustively referenced and discussed.

[RESPONSE]

The authors state that a predicted polybasic cleavage sites is unique to SARS-CoV-2 in SARS viruses. Who knows how many out of thousands undiscovered bat ancestors also acquired such a motif, the sampling bias in descriptions of remote bat viruses is dramatic. This should be discussed. Also state clearly that this site is only predicted so far and that experimental evidence for its biological function and its potential impact on pathogenesis are required.

[RESPONSE]

We agreed that it is possible that a bat CoV

The predicted O-linked glycosylation sites are mysterious. What do the authors imply with those sites? In silico prediction of O-linked glycosylation sites is not robust and whether these sites indeed exist requires experimental validation. Even if those sites exist, why are they relevant? This is not addressed at all. If the authors assume these sites constitute part of a glycan shield, they should say so and weigh their assumption carefully.

[RESPONSE]

We agree that the O-linked glycosylation sites require experimental validation not only in SARS-CoV-2 but in other CoVs that have similar predicted sites. This was already stated explicitly, but in revision we have stated this even more directly. Yes, we consider that these sites if utilized may be part of the glycan shield and have added this important point to the text.

Finally, the main argument against a hypothetical lab origin seems the required reconstruction of a backbone of a bat virus of unknown pathogenesis. It does not seem feasible that any scientist would disembark on such an uncertain endeavor. This difficulties of coronavirus reverse genetics should be stated clearly.

[RESPONSE]

Nature Rebuttal February 20, 2020 10:52 AM

Referee #1 (Remarks to the Author):

Anderson presented a timely manuscript to share their points of view about the origin of SARS-CoV-2. There are several rumors about the origin of this virus. However, these "hypotheses" are entirely based on very limited, if any, scientific evidences.

This reviewer sees most of the arguments raised by the authors are valid and convincing. However, the authors might want to consider these minor suggestions:

1. The sections for the RBD and cleavage site of Spike protein basically have summarized the existing findings from other recent publications. The authors might want to spell out that these two sections are review summaries. In addition, the author can present these two sections in a more condense format and save some space for something else (also see points 6 and 7 below)

[RESPONSE]

We have edited these sections to be more concise.

2. Fig. 1. This figure has 6 aligned sequences, but with only 5 sequence titles. The order of these titles are also not correct.

[RESPONSE]

The sequence titles have been corrected.

3. Lines 170 -174. It is correct that no adaptive mutation has been found in the spike of MERS-CoV. Deletions in other ORF regions, however, were detected in some human MERS-CoV viruses (PMID: 26981770). In addition, the 29nt deletion of human SARS-CoV (PMID: 12958366) was suggested to have effects on host adaptation. The authors should also consider these findings. It is premature to say that this would not happen in SARS-CoV-2.

[RESPONSE]

We thank the reviewer for pointing out these relevant references. We have included these important points in the revised text.

4. Line 194. The accident at Singapore occurred in a BSL3, not BSL2, containment.

[RESPONSE]

We have corrected this point in revision.

5. Line 194. Laboratory escapes of SARS occurred in Singapore, China and Taiwan (PMID: 16830004).

[RESPONSE]

We have added this reference and the point about lab escapes of SARS-CoV-1 in Singapore, China and Taiwan.

6. There are two recent reports about coronaviruses in pangolins (https://www.biorxiv.org/content/10.1101/2020.02.13.945485v1.full.pdf; https://www.biorxiv.org/content/10.1101/2020.02.08.939660v2.full.pdf). The authors might want to comments on these.

[RESPONSE]

Nature Rebuttal February 20, 2020 2:44pm

In response to both reviewers we have clarified and expanded our discussion of the newly available CoV sequences from pangolins.

7. Optional: Can the authors share their views on the possibility of having a lab escape of a natural coronavirus? This is also one of the hypotheses that have been extensively discussed. The reviewer understand that this is entirely a different topic, but any insights are welcomed.

[RESPONSE]

Escape of a natural CoV (SARS-CoV-2 or a close progenitor) from a lab could not be distinguished from an animal-to-human transfer in another environment. Given the limited numbers of labs compared to the frequent opportunities for animal-to-human transfer the latter is much less likely than the former.

Nature Rebuttal February 20, 2020 2:44pm

Referee #2 (Remarks to the Author):

This is a perspective discussing evidence against a hypothetical lab origin of SARS-CoV-2. The paper addresses suboptimal composition of ACE2-binding sites in the RBD, 3 predicted O-linked glycosylation sites and a furin cleavage site in the glycoprotein that was speculated upon before.

The paper is itself interesting, but unnecessarily speculative. It's not clear why the authors do not refute a hypothetical lab origin in their coming publication on the ancestors of SARS-CoV-2 in bats and pangolins. The tree showing diverse pangolin viruses has kindly been made available by some of the authors in GISAID. Once the authors publish their new pangolin sequences, a lab origin will be extremely unlikely. It is not clear why the authors rush with a speculative perspective if their central hypothesis can be supported by their own data. Please explain.

[RESPONSE]

Our manuscript is written to explore the potential natural origin of SARS-CoV-2, and thus some speculation is necessary. Unfortunately, the newly available pangolin sequences do not elucidate the origin of SARS-CoV-2 or refute a lab origin. The reviewer is wrong about this point. Had this been the point we would have included it. We have been analyzing the pangolin CoV sequences, and it is increasingly unlikely that they serve as the intermediate host. While pangolins harbor SARS-like CoVs, it is unlikely that they have a direct connection to the COVID-19 epidemic via any of the scenarios outlined in our manuscript.

Another critical aspect of this text is the complete lack of referencing to a potential debate on a hypothetical lab origin. Who said this, why is this considered a problem? There are indeed a few apparently uninformed statements claiming the virus may be a Chinese bioweapon, but is this really problematic on a larger scale? The central reason for issuing this text must be exhaustively referenced and discussed.

[RESPONSE]

The possibility that SARS-CoV-2 originated as an engineered bioweapon has had widespread discussion in the press and on social media and is problematic on a large scale. Agroup of public health scientists recently wrote a letter to The Lancet in which they "strongly condemn conspiracy theories suggesting that COVID-19 does not have a natural origin." We now reference this publication. While our manuscript briefly discusses evidence against the bioweapon scenario, it was written to explore the proximal origins of SARS-CoV-2.

The authors state that a predicted polybasic cleavage sites is unique to SARS-CoV-2 in SARS viruses. Who knows how many out of thousands undiscovered bat ancestors also acquired such a motif, the sampling bias in descriptions of remote bat viruses is dramatic. This should be discussed. Also state clearly that this site is only predicted so far and that experimental evidence for its biological function and its potential impact on pathogenesis are required.

[RESPONSE]

We agreed that the diversity in bat CoVs is undersampled and that it is possible that a bat CoVprogenitor could be discovered that contains a polybasic site. This is explicitly discussed in revision. While other CoVs have polybasic sites that are in fact utilized, we did refer to the site as a predicted cleavage site. We are more explicit in revision that this will require experimental verification. We also state that it will be necessary to test the effect of the polybasic site on pathogenesis, which will require the establishment of an animal model.

The predicted O-linked glycosylation sites are mysterious. What do the authors imply with those sites? In silico prediction of O-linked glycosylation sites is not robust and whether these sites indeed exist requires

Nature Rebuttal February 20, 2020 2:44pm

experimental validation. Even if those sites exist, why are they relevant? This is not addressed at all. If the authors assume these sites constitute part of a glycan shield, they should say so and weigh their assumption carefully.

[RESPONSE]

Although not previously described for CoV proteins, numerous other viral proteins have mucin-like domains that are involved in immune evasion. The revised text adds relevant references and is more explicit about the potential relevance of the predicted O-linked glycan sites. As stated previously, we consider that these sites if utilized may be part of the glycan shield and have made this point directly in the revised text. We also agree that the predicted O-linked glycosylation sites require experimental validation not only in SARS-CoV-2 but in other CoVs that have similar predicted sites. This was also already stated explicitly, but in revision we have stated this even more directly.

Finally, the main argument against a hypothetical lab origin seems the required reconstruction of a backbone of a bat virus of unknown pathogenesis. It does not seem feasible that any scientist would disembark on such an uncertain endeavor. This difficulties of coronavirus reverse genetics should be stated clearly.

[RESPONSE]

The reviewer's restatement of our argument is correct. We reiterate that the purpose of our manuscript was not to refute the conspiracy theory that SARS-CoV-2 was bioengineered. We added an additional statement about the difficulties of CoV reverse engineering.

Editor's comments:

While the Perspective is interesting and timely one of our referees raised concerns (also emphasised to the editors) about whether such a piece would feed or quash the conspiracy theories.

[RESPONSE]

We weren't setting out to quash conspiracy theories but to examine the evidence for and against a number of possible origins of the virus.

But more importantly this reviewer feels, and we agree, that the Perspective would quickly become outdated when more scientific data are published (for example on potential reservoir hosts).

[RESPONSE]

We believe our piece would remain relevant if more data becomes available because it would potentially confirm which of these scenarios is correct and our paper sets out what evidence would be needed. We would also say that there is a very real possibility that no clear reservoir host will be found (the pangolins are as yet not clearly in the frame - as we explain in our revised section on these).

Nature Rebuttal February 20, 2020 2:44pm

The Proximal Origin of SARS-CoV-2

Kristian G. Andersen[1,2*], Andrew Rambaut[3], W. Ian Lipkin[4], Edward C. Holmes[5] & Robert F. Garry[6,7]

[1]Department of Immunology and Microbiology, The Scripps Research Institute, La Jolla, CA 92037, USA.

[2]Scripps Research Translational Institute, La Jolla, CA 92037, USA.

[3]Institute of Evolutionary Biology, University of Edinburgh, Edinburgh, UK.

[4]Center for Infection and Immunity, Mailman School of Public Health of Columbia University, New York, New York, USA.

[5]Marie Bashir Institute for Infectious Diseases and Biosecurity, School of Life and Environmental Sciences and School of Medical Sciences, The University of Sydney, Sydney, Australia.

[6]Tulane University, School of Medicine, Department of Microbiology and Immunology, New Orleans, LA, USA.

[7]Zalgen Labs, LCC, Germantown, MD, USA.

*Corresponding author:

Kristian G. Andersen

Department of Immunology and Microbiology,

The Scripps Research Institute,

La Jolla, CA 92037, USA.

Since the first reports of a novel pneumonia (COVID-19) in Wuhan city, Hubei province, China there has been considerable discussion and uncertainty over the origin of the causative virus, SARS-CoV-2. Infections with SARS-CoV-2 are now widespread in China, with cases in every province. As of 14 February 2020, 64,473 such cases have been confirmed, with 1,384 deaths attributed to the virus. These official case numbers are likely an underestimate because of limited reporting of mild and asymptomatic cases, and the virus is clearly capable of efficient human-to-human transmission. Based on the possibility of spread to countries with weaker healthcare systems, the World Health Organization has declared the COVID-19 outbreak a Public Health Emergency of International Concern (PHEIC). There are currently neither vaccines nor specific treatments for this disease.

SARS-CoV-2 is the seventh member of the *Coronaviridae* known to infect humans. Three of these viruses, SARS CoV-1, MERS CoV, and SARS-CoV-2, can cause severe disease; four, HKU1, NL63, OC43 and 229E, are associated with mild respiratory symptoms. Herein, we review what can be deduced about the origin and early evolution of SARS-CoV-2 from the comparative analysis of available genome sequence data. In particular, we offer a perspective on the notable features in the SARS-CoV-2 genome and discuss scenarios by which these features could have arisen. Importantly, this analysis provides evidence that SARS-CoV-2 is not a laboratory construct nor a purposefully manipulated virus.

February 21, 2020 2:49 PM

The genomic comparison of both alpha- and betacoronaviruses (family *Coronaviridae*) described below identifies two notable features of the SARS-CoV-2 genome: (i) based on structural modeling and early biochemical experiments, SARS-CoV-2 appears to be optimized for binding to the human ACE2 receptor; (ii) the highly variable spike (S) protein of SARS-CoV-2 has a polybasic (furin) cleavage site at the S1 and S2 boundary via the insertion of twelve nucleotides. Additionally, this event led to the acquisition of three predicted O-linked glycans around the polybasic cleavage site.

Mutations in the receptor binding domain of SARS-CoV-2

The receptor binding domain (RBD) in the spike protein of SARS-CoV and SARS-related coronaviruses is the most variable part of the virus genome. Six residues in the RBD appear to be critical for binding to the human ACE2 receptor and determining host range[1]. Using coordinates based on the Urbani strain of SARS-CoV, they are Y442, L472, N479, D480, T487, and Y491[1]. The corresponding residues in SARS-CoV-2 are L455, F486, Q493, S494, N501, and Y505. Five of these six residues are mutated in SARS-CoV-2 compared to its most closely related virus, RaTG13 sampled from a *Rhinolophus affinis* bat, to which it is ~96% identical[2] (**Figure 1a**). Based on modeling[1] and biochemical experiments[3,4], SARS-CoV-2 seems to have an RBD that may bind with high affinity to ACE2 from human, non-human primate, ferret, pig, and cat, as well as other species with high receptor homology[1]. In contrast, SARS-CoV-2 may bind less efficiently to ACE2 in other species associated with SARS-like viruses, including rodents and civets[1].

The phenylalanine (F) at residue 486 in the SARS-CoV-2 S protein corresponds to L472 in the SARS-CoV Urbani strain. Notably, in SARS-CoV cell culture experiments the L472 mutates to phenylalanine (L472F)[5], which is predicted to be optimal for binding of the SARS-CoV RBD to the human ACE2 receptor[6]. However, a phenylalanine in this position is also present in several SARS-like CoVs from bats (**Figure 1a**). While these analyses suggest that SARS-CoV-2 may be capable of binding the human ACE2 receptor with high affinity, the interaction is not predicted to be optimal[1]. Additionally, several of the key residues in the RBD of SARS-CoV-2 are different to those previously described as optimal for human ACE2 receptor binding[6]. In contrast to these computational predictions, recent binding studies indicate that SARS-CoV-2 binds with high affinity to human ACE2[7]. Thus the SARS-CoV-2 spike appears to be the result of selection on human or human-like ACE2 permitting another optimal binding solution to arise. This is strong evidence that SARS-CoV-2 is *not* the product of genetic engineering.

Polybasic cleavage site and O-linked glycans

The second notable feature of SARS-CoV-2 is a predicted polybasic cleavage site (RRAR) in the spike protein at the junction of S1 and S2, the two subunits of the spike protein (**Figure 1b**)[8,9]. In addition to two basic arginines and an alanine at the cleavage site, a leading proline is also inserted; thus, the fully inserted sequence is PRRA (**Figure 1b**). The strong turn created by the proline insertion is predicted to result in the addition of O-linked glycans to S673, T678, and S686 that flank the polybasic cleavage site. A polybasic cleavage site has not previously been observed in related lineage B betacoronaviruses and is a unique feature of SARS-CoV-2. Some human betacoronaviruses, including HCoV-HKU1 (lineage A), have polybasic cleavage sites, as well as predicted O-linked glycans near the S1/S2 cleavage site.

While the functional consequence of the polybasic cleavage site in SARS-CoV-2 is unknown, experiments with SARS-CoV have shown that engineering such a site at the S1/S2 junction enhances cell–cell fusion but does not affect virus entry[10]. Polybasic cleavage sites allow effective cleavage by furin and other proteases, and can be acquired at the junction of the two subunits of the haemagglutinin (HA) protein of avian influenza viruses in conditions that select for rapid virus replication and transmission (e.g.,. highly dense chicken populations). HA serves a similar function in cell-cell fusion and viral entry as the coronavirus S protein. Acquisition of a polybasic cleavage site in HA, by either insertion or recombination, converts low pathogenicity avian influenza viruses into highly pathogenic forms[11–13]. The acquisition of polybasic

February 21, 2020 2:49 PM

cleavage sites by the influenza virus HA has also been observed after repeated forced passage in cell culture or through animals[14,15]. Similarly, an avirulent isolate of Newcastle Disease virus became highly pathogenic during serial passage in chickens by incremental acquisition of a polybasic cleavage site at the junction of its fusion protein subunits[16]. The potential function of the three predicted O-linked glycans is less clear, but they could create a “mucin-like domain” that would shield potential epitopes or key residues on the SARS-CoV-2 spike protein(Bagdonaite and Wandall 2018; Tran et al. 2014). Although the algorithms for prediction of O-linked glycosylation are robust(Steentoft et al. 2013), biochemical analyses or structural studies are required to determine whether or not the predicted O-linked glycan sites are utilized.

Theories of SARS-CoV-2 origins

It is improbable that SARS-CoV-2 emerged through laboratory manipulation of an existing SARS-related coronavirus. As noted above, the RBD of SARS-CoV-2 is optimized for human ACE2 receptor binding with an efficient binding solution different to that which would have been predicted. Further, if genetic manipulation had been performed, one would expect that one of the several reverse genetic systems available for betacoronaviruses would have been used. However, this is not the case as the genetic data shows that SARS-CoV-2 is not derived from any previously used virus backbone[17]. Instead, we propose two scenarios that can plausibly explain the origin of SARS-CoV-2: (i) natural selection in a non-human animal host prior to zoonotic transfer, and (ii) natural selection in humans following zoonotic transfer. We also discuss whether selection during passage in culture could have given rise to the same observed features.

Selection in an animal host

As many of the early cases of COVID-19 were linked to the Huanan seafood and wildlife market in Wuhan, it is possible that an animal source was present at this location. Given the similarity of SARS-CoV-2 to bat SARS-like CoVs, particularly RaTG13, it is plausible that bats serve as reservoir hosts for SARS-CoV-2. It is important, however, to note that previous outbreaks of betacoronaviruses in humans involved direct exposure to animals other than bats, including civets (SARS) and camels (MERS), that carry viruses that are genetically very similar to SARS-CoV or MERS-CoV, respectively. By analogy, viruses closely related to SARS-Cov-2 may be circulating in one or more animal species. Initial analyses indicate that Malayan pangolins (*Manis javanica*) illegally imported into Guangdong province contain a CoV that is similar to SARS-CoV-2[(Wong et al. 2020; Phylodynamic Analysis | 90 genomes | ...)]. Although the bat virus RaTG13 remains the closest relative to SARS-CoV-2 across the whole genome, the Malayan pangolin CoV is identical to SARS-CoV-2 at all six key RBD residues (**Figure 1**). However, no pangolin CoV has yet been identified that is sufficiently similar to SARS-CoV-2 across its entire genome to support direct human infection. In addition, the pangolin CoV does not carry a polybasic cleavage site insertion. For a precursor virus to acquire the polybasic cleavage site and mutations in the spike protein suitable for human ACE2 receptor binding, an animal host would likely have to have a high population density – to allow natural selection to proceed efficiently – and an ACE2 gene that is similar to the human orthologue. Further characterization of CoVs in pangolins and other animals that may harbour SARS-CoV-like viruses should be a public health priority.

Cryptic adaptation to humans

It is also possible that a progenitor to SARS-CoV-2 jumped from a non-human animal to humans, with the genomic features described above acquired through adaptation during subsequent human-to-human transmission. We surmise that once these adaptations were acquired (either together or in series) it would enable the outbreak to take-off, producing a sufficiently large and unusual cluster of pneumonia cases to trigger the surveillance system that ultimately detected it.

All SARS-CoV-2 genomes sequenced so far have the well adapted RBD and the polybasic cleavage site and are thus derived from a common ancestor that had these features. The presence of an RBD in pangolins that is very similar to the one in SARS-CoV-2 means that this was likely already present in the virus that

February 21, 2020 2:49 PM

jumped to humans, even if we don't yet have the exact non-human progenitor virus. This leaves the polybasic cleavage site insertion to occur during human-to-human transmission. Following the example of the influenza A virus HA gene, a specific insertion or recombination event is required to enable the emergence of SARS-CoV-2 as an epidemic pathogen.

Estimates of the timing of the most recent common ancestor (tMRCA) of SARS-CoV-2 using currently available genome sequence data point to virus emergence in late November to early December 2019[19,20], compatible with the earliest retrospectively confirmed cases[21]. Hence, this scenario presumes a period of unrecognised transmission in humans between the initial zoonotic transfer event and the acquisition of the polybasic cleavage site. Sufficient opportunity could occur if there had been many prior zoonotic events producing short chains of human-to-human transmission (so-called 'stuttering chains') over an extended period. This is essentially the situation for MERS-CoV in the Arabian Peninsula where all the human cases are the result of repeated jumps of the virus from dromedary camels, producing single infections or short chains of transmission that eventually resolve. To date, after 2,499 cases over 8 years, no human adaptation has emerged that has allowed MERS-CoV to take hold in the human population.

How could we test whether cryptic spread of SARS-CoV-2 enabled human adaptation? Metagenomic studies of banked serum samples could provide important information, but given the relatively short period of viremia it may be impossible to detect low level SARS-CoV-2 circulation in historical samples. Retrospective serological studies potentially could be informative and a few such studies have already been conducted. One found that animal importation traders had a 13% seropositivity to coronaviruses[22], while another noted that 3% residents of a village in Southern China were seropositive to these viruses[23]. Interestingly, 200 residents of Wuhan did not show coronavirus seroreactivity. Critically, however, these studies could not have distinguished whether positive serological responses were due to a prior infection with SARS-CoV or SARS-CoV-2. Further retrospective serological studies should be conducted to determine the extent of prior human exposure to betacoronaviruses in different geographic areas, particularly using assays that can distinguish among multiple betacoronaviruses.

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 in multiple laboratories across the world[24–27]. There are also documented instances of the laboratory acquisition of SARS-CoV by laboratory personnel working under BSL-2 containment[28,29]. We must therefore consider the possibility of a deliberate or inadvertent release of SARS-CoV-2. In theory, it is possible that SARS-CoV-2 acquired the observed RBD mutations site during adaptation to passage in cell culture, as has been observed in studies with SARS-CoV[5] as well as MERS-CoV[30]. However, the acquisition of the polybasic cleavage site or O-linked glycans - if functional - argues against this scenario. New polybasic cleavage sites have only been observed after prolonged passaging of low pathogenicity avian influenza virus in cell culture or animals. Furthermore, the generation of SARS-CoV-2 by cell culture or animal passage would have required prior isolation of a progenitor virus with a very high genetic similarity. Subsequent generation of a polybasic cleavage site would have then required an intense program of passage in cell culture or animals with ACE-2 receptor similar to humans (e.g. ferrets). It is also questionable whether generation of the O-linked glycans would have occurred on cell culture passage, as such mutations typically suggest the involvement of an immune system, that is not present *in vitro*.

Conclusions

In the midst of the global COVID-19 public health emergency it is reasonable to wonder why the origins of the epidemic matter. A detailed understanding of how an animal virus jumped species boundaries to infect humans so productively will help in the prevention of future zoonotic events. For example, if SARS-CoV-2

February 21, 2020 2:49 PM

pre-adapted in another animal species then we are at risk of future re-emergence events even if the current epidemic is controlled. In contrast, if the adaptive process we describe occurred in humans, then even if we have repeated zoonotic transfers they are unlikely to take-off unless the same series of mutations occurs. In addition, identifying the closest animal relatives of SARS-CoV-2 will greatly assist studies of virus function. Indeed, the availability of the RaTG13 bat sequence facilitated the comparative genomic analysis performed here, helping to reveal the key mutations in the RBD as well as the polybasic cleavage site insertion.

The genomic features described here may in part explain the infectiousness and transmissibility of SARS-CoV-2 in humans. Although genomic evidence does not support the idea that SARS-CoV-2 is a laboratory construct, it is currently impossible to prove or disprove the other theories of its origin described here, and it is unclear whether future data 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 revealing virus origins. In addition, it would be helpful to obtain more genetic and functional data about the virus, including experimental studies of receptor binding and the role of the polybasic cleavage site and predicted O-linked glycans. The identification of a potential intermediate host of SARS-CoV-2, as well as the sequencing of very early cases including those not connected to the Wuhan market, would similarly be highly informative. Irrespective of how SARS-CoV-2 originated, the ongoing surveillance of pneumonia in humans and other animals is clearly of utmost importance.

Acknowledgements

We thank all those who have contributed SARS-CoV-2 genome sequences to the GISAID database (https://www.gisaid.org/) and contributed analyses and ideas to Virological.org (http://virological.org/). We thank M. Farzan for discussions. We thank the Wellcome Trust for supporting this work. ECH is supported by an ARC Australian Laureate Fellowship (FL170100022). KGA is a Pew Biomedical Scholar and is supported by NIH NIAID grant U19AI135995. AR is supported by the Wellcome Trust (Collaborators Award 206298/Z/17/Z – ARTIC network) and the European Research Council (grant agreement no. 725422 – ReservoirDOCS). RFG is supported by NIH grants U19AI135995, U54 HG007480 and U19AI142790.

Figure Legends

Figure 1 | a) Mutations in contact residues of the SARS-CoV-2 spike protein. The spike protein of SARS-CoV-2 (top) was aligned against the most closely related SARS-like CoVs and SARS-CoV-1. Key residues in the spike protein that make contact to the ACE2 receptor are marked with blue boxes in both SARS-CoV-2 and the SARS-CoV Urbani strain. **b) Acquisition of polybasic cleavage site and O-linked glycans**. The polybasic cleavage site is marked in grey with the three adjacent predicted O-linked glycans in blue. Both the polybasic cleavage site and O-linked glycans are unique to SARS-CoV-2 and not previously seen in lineage B betacoronaviruses. Sequences shown are from NCBI GenBank, accession numbers MN908947, MN996532, AY278741, KY417146 & MK211376. The pangolin coronavirus sequences are a consensus generated from SRR10168377 and SRR10168378 (NCBI BioProject PRJNA573298)(Phylodynamic Analysis | 90 genomes | ...; Wong et al. 2020).

February 21, 2020 2:49 PM

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. Wu, F. *et al.* A new coronavirus associated with human respiratory disease in China. *Nature* (2020) doi:10.1038/s41586-020-2008-3.

3. 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.

4. 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.

5. 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).

6. Cui, J., Li, F. & Shi, Z.-L. Origin and evolution of pathogenic coronaviruses. *Nat. Rev. Microbiol.* **17**, 181–192 (2019).

7. Wrapp, D. *et al.* Cryo-EM Structure of the 2019-nCoV Spike in the Prefusion Conformation. *bioRxiv* 2020.02.11.944462 (2020) doi:10.1101/2020.02.11.944462.

8. Gallaher, W. Analysis of Wuhan Coronavirus: Deja Vu. *Virological* http://virological.org/t/analysis-of-wuhan-coronavirus-deja-vu/357 (2020).

9. Coutard, B. *et al.* The spike glycoprotein of the new coronavirus 2019-nCoV contains a furin-like cleavage site absent in CoV of the same clade. *Antiviral Res.* 104742 (2020).

10. 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).

February 21, 2020 2:49 PM

11. 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).

12. Alexander, D. J. & Brown, I. H. History of highly pathogenic avian influenza. *Rev. Sci. Tech.* **28**, 19–38 (2009).

13. 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).

14. Ito, T. *et al.* Generation of a highly pathogenic avian influenza A virus from an avirulent field isolate by passaging in chickens. *J. Virol.* **75**, 4439–4443 (2001).

15. Li, S. Q., Orlich, M. & Rott, R. Generation of seal influenza virus variants pathogenic for chickens, because of hemagglutinin cleavage site changes. *J. Virol.* **64**, 3297–3303 (1990).

16. Shengqing, Y. *et al.* Generation of velogenic Newcastle disease viruses from a nonpathogenic waterfowl isolate by passaging in chickens. *Virology* **301**, 206–211 (2002).

17. Menachery, V. D. *et al.* A SARS-like cluster of circulating bat coronaviruses shows potential for human emergence. *Nat. Med.* **21**, 1508–1513 (2015).

18. Wong, M. C., Javornik Cregeen, S. J., Ajami, N. J. & Petrosino, J. F. Evidence of recombination in coronaviruses implicating pangolin origins of nCoV-2019. *bioRxiv* 2020.02.07.939207 (2020) doi:10.1101/2020.02.07.939207.

19. Phylodynamic Analysis | 90 genomes | 12 Feb 2020 – Novel 2019 coronavirus / nCoV-2019 Genomic Epidemiology. http://virological.org/t/phylodynamic-analysis-90-genomes-12-feb-2020/356 (2020).

20. Phylodynamic estimation of incidence and prevalence of novel coronavirus (nCoV) infections through time – Novel 2019 coronavirus / nCoV-2019 Genomic Epidemiology. http://virological.org/t/phylodynamic-estimation-of-incidence-and-prevalence-of-novel-coronavirus-ncov-infections-through-time/391 (2020).

21. Huang, C. *et al.* Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China.

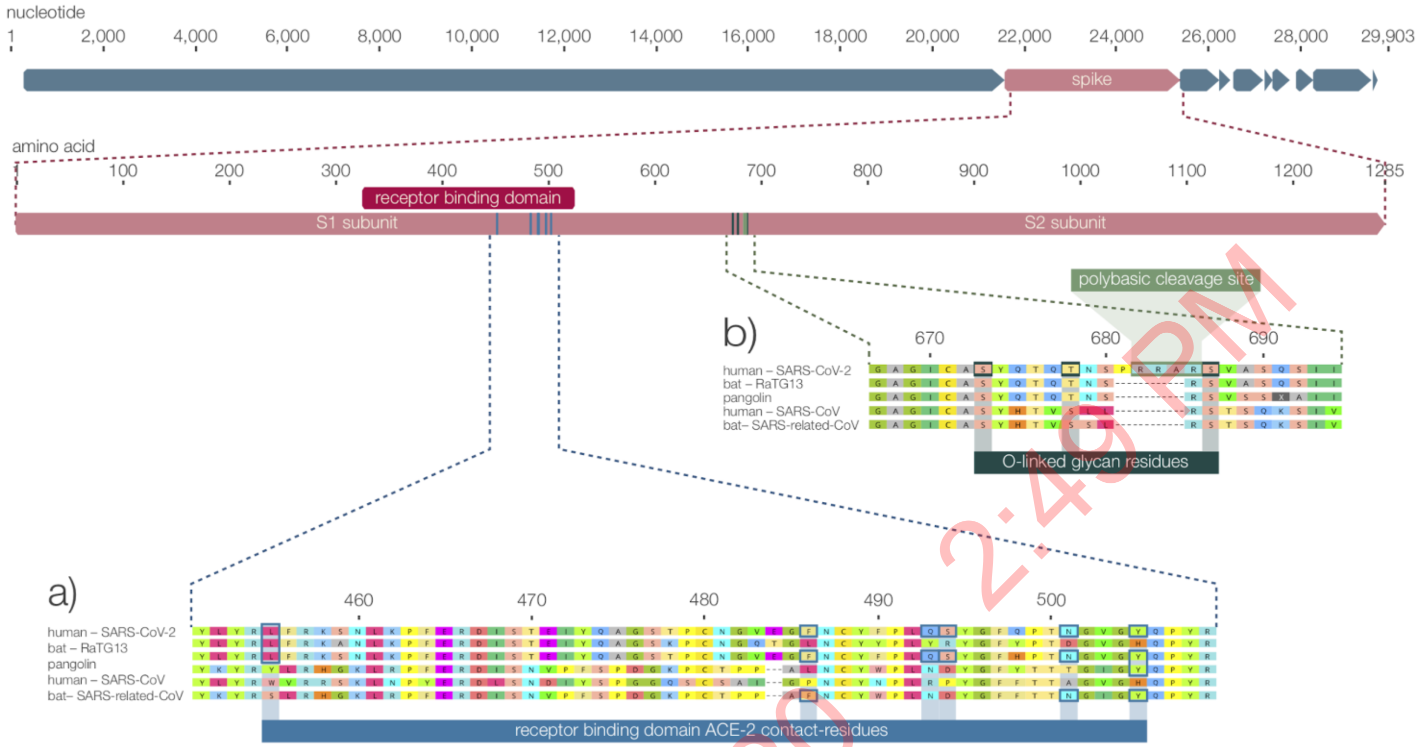
February 21, 2020 2:49 PM

Lancet (2020) doi:10.1016/S0140-6736(20)30183-5.

22. Centers for Disease Control and Prevention (CDC). Prevalence of IgG antibody to SARS-associated coronavirus in animal traders--Guangdong Province, China, 2003. *MMWR Morb. Mortal. Wkly. Rep.* **52**, 986–987 (2003).
23. Wang, N. *et al.* Serological Evidence of Bat SARS-Related Coronavirus Infection in Humans, China. *Virol. Sin.* **33**, 104–107 (2018).
24. Ge, X.-Y. *et al.* Isolation and characterization of a bat SARS-like coronavirus that uses the ACE2 receptor. *Nature* **503**, 535–538 (2013).
25. 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).
26. 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).
27. 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).
28. Lim, P. L. *et al.* Laboratory-acquired severe acute respiratory syndrome. *N. Engl. J. Med.* **350**, 1740–1745 (2004).
29. Senior, K. Recent Singapore SARS case a laboratory accident. *Lancet Infect. Dis.* **3**, 679 (2003).
30. Letko, M. *et al.* Adaptive Evolution of MERS-CoV to Species Variation in DPP4. *Cell Rep.* **24**, 1730–1737 (2018).

February 21, 2020 2:49 PM

Figure 1.


nucleotide
1
2,000
4,000
6,000
8,000
10,000
12,000
14,000
16,000
18,000
20,000
22,000
24,000
26,000
28,000
29,903
spike
amino acid
1
100
200
300
400
500
600
700
800
900
1000
1100
1200
1285
receptor binding domain
S1 subunit
S2 subunit
polybasic cleavage site
b)
670
680
690
human – SARS-CoV-2
bat – RaTG13
pangolin
human – SARS-CoV
bat– SARS-related-CoV
G A G I C A S Y Q T Q T N S P R R A R S V A S Q S I I
G A G I C A S Y Q T Q T N S - - - - R S V A S Q S I I
G A G I C A S Y Q T Q T N S - - - - R S V S S X A I I
G A G I C A S Y H T V S L L - - - - R S T S Q K S I V
G A G I C A S Y H T V S S L - - - - R S T S Q K S I V
O-linked glycan residues
a)
460
470
480
490
500
human – SARS-CoV-2
bat – RaTG13
pangolin
human – SARS-CoV
bat– SARS-related-CoV
Y L Y R L F R K S N L K P F E R D I S T E I Y Q A G S T P C N G V E G F N C Y F P L Q S Y G F Q P T N G V G Y Q P Y R
Y L Y R L F R K A N L K P F E R D I S T E I Y Q A G S K P C N G Q T G L N C Y Y P L Y R Y G F Y P T D G V G H Q P Y R
Y L Y R L F R K S N L K P F E R D I S T E I Y Q A G S T P C N G V E G F N C Y F P L Q S Y G F H P T N G V G Y Q P Y R
Y K Y R Y L R H G K L R P F E R D I S N V P F S P D G K P C T P P - - - A L N C Y W P L N D Y G F Y T T T G I G Y Q P Y R
Y L Y R W V R R S K L N P Y E R D L S N D I Y S P G G Q S C S A I - - - G P N C Y N P L R P Y G F F T T A G V G H Q P Y R
Y K Y R S L R H G K L R P F E R D I S N V P F S P D G K P C T P P - - - A F N C Y W P L N D Y G F F T T N G I G Y Q P Y R
receptor binding domain ACE-2 contact-residues

February 21, 2020 2:49 PM

Referee #1:

Anderson presented a timely manuscript to share their points of view about the origin of SARS-CoV-2. There are several rumors about the origin of this virus. However, these "hypotheses" are entirely based on very limited, if any, scientific evidences. This reviewer sees most of the arguments raised by the authors are valid and convincing. However, the authors might want to consider these minor suggestions:

1. The sections for the RBD and cleavage site of Spike protein basically have summarized the existing findings from other recent publications. The authors might want to spell out that these two sections are review summaries. In addition, the author can present these two sections in a more condense format and save some space for something else (also see points 6 and 7 below)

We have edited these sections to be more concise. However, we think that it is important that the features are described in detail as they form the basis of some of the discussion that follows.

2. Fig. 1. This figure has 6 aligned sequences, but with only 5 sequence titles. The order of these titles are also not correct.

We have corrected these mistakes and updated the figure and titles.

3. Lines 170 -174. It is correct that no adaptive mutation has been found in the spike of MERS-CoV. Deletions in other ORF regions, however, were detected in some human MERS-CoV viruses (PMID: 26981770). In addition, the 29nt deletion of human SARS-CoV (PMID: 12958366) was suggested to have effects on host adaptation. The authors should also consider these findings. It is premature to say that this would not happen in SARS-CoV-2.

We have revised the text to incorporate this information as we agree that they are important for context.

4. Line 194. The accident at Singapore occurred in a BSL3, not BSL2, containment.

We have corrected this.

5. Line 194. Laboratory escapes of SARS occurred in Singapore, China and Taiwan (PMID: 16830004).

We have added this reference as well as the point about laboratory escapes of SARS-CoV in Singapore, China and Taiwan.

6. There are two recent reports about coronaviruses in pangolins. The authors might want to comment on these.

We have included these references as well as several others that have investigated pangolin CoV. In response to Reviewer 2 below as well, we should, however, point out that these additional pangolin CoV sequences do not further clarify the different scenarios we discuss in our manuscript. There is nothing in these reports that changes our statements regarding the role of pangolins.

7. Optional: Can the authors share their views on the possibility of having a lab escape of a natural coronavirus? This is also one of the hypotheses that have been extensively discussed. The reviewer understand that this is entirely a different topic, but any insights are welcomed.

Nature rebuttal February 21, 2020 8:37 PM

Escape of a natural CoV (SARS-CoV-2 or a close progenitor) from a lab could not be distinguished from an animal-to-human transfer in another environment. Given the limited numbers of labs compared to the frequent opportunities for animal-to-human transfer, it is obvious that the latter is much less likely than the former.

Nature rebuttal   February 21, 2020   8:37 PM

Referee #2:

This is a perspective discussing evidence against a hypothetical lab origin of SARS-CoV-2. The paper addresses suboptimal composition of ACE2-binding sites in the RBD, 3 predicted O-linked glycosylation sites and a furin cleavage site in the glycoprotein that was speculated upon before.

The paper is itself interesting, but unnecessarily speculative. It's not clear why the authors do not refute a hypothetical lab origin in their coming publication on the ancestors of SARS-CoV-2 in bats and pangolins. The tree showing diverse pangolin viruses has kindly been made available by some of the authors in GISAID. Once the authors publish their new pangolin sequences, a lab origin will be extremely unlikely. It is not clear why the authors rush with a speculative perspective if their central hypothesis can be supported by their own data. Please explain.

Our manuscript is written to explore the potential origin of SARS-CoV-2. We do not believe it is speculative: rather, it simply takes the available data and proposes a series of hypotheses that explain how these data may have arisen. We try to do this in a logical, balanced and unbiased manner: this is critically important because it gives our work credibility. It is opinion, but science.

Unfortunately, the newly available pangolin sequences do not elucidate the origin of SARS-CoV-2 or refute a lab origin. Hence, the reviewer is incorrect on this point. To clarify, while the RBD from the Guangdong pangolin CoVs is the closest to that found SARS-CoV-2, they are more divergent in the remainder of the viral genome (for which the bat virus RaTG13 is still the closest) and do not possess the polybasic cleavage site insertion. Hence, there is no evidence on present data that the pangolin CoVs are directly related to the COVID-19 epidemic.

Another critical aspect of this text is the complete lack of referencing to a potential debate on a hypothetical lab origin. Who said this, why is this considered a problem? There are indeed a few apparently uninformed statements claiming the virus may be a Chinese bioweapon, but is this really problematic on a larger scale? The central reason for issuing this text must be exhaustively referenced and discussed.

The possibility that SARS-CoV-2 originated as an engineered bioweapon has had widespread discussion in the press and on social media (particularly in China) and is hence problematic on a very large scale. This particular topic was also the reason for a recent request from the White House: https://bit.ly/2HMndCi. A group of public health scientists recently wrote a letter to *The Lancet* in which they *"strongly condemn conspiracy theories suggesting that COVID-19 does not have a natural origin."* We now reference this publication (the publication itself also references our Virological post).

While our manuscript briefly discusses evidence against the bioweapon scenario, it was written to explore valid scientific theories about the proximal origins of SARS-CoV-2. Importantly, it appears that the reviewer is considering bioweapons, engineering, and lab accidents as one and the same theory lumped in under a "lab origin". This is incorrect. As we very clearly state in the manuscript, there are far more subtle scenarios that need to be considered carefully and scientifically: for example, of accidental infections in a lab while culturing SARS-like CoVs. Since accidental infections and other lab 'escapes' happen frequently across the world (and as we mention, happened multiple times with SARS-CoV-1 following the SARS epidemic) we firmly believe that this discussion is of major importance and must be had. In particular, the culturing of SARS-like CoVs from animals is typically performed under BSL-2 and has been ongoing for years. Dismissing this potential accidental scenario out of hand - or considering it in the same category as conspiracy theories about bioweapons and deliberate engineering (that, as we outline, are clearly wrong)

Nature rebuttal February 21, 2020 8:37 PM

- would be irresponsible. Our manuscript will only serve its purpose if it considers all possible hypotheses equally. Any perceived bias will undermine its credibility.

We have modified the text to more clearly state the scenarios we are considering.

The authors state that a predicted polybasic cleavage sites is unique to SARS-CoV-2 in SARS viruses. Who knows how many out of thousands undiscovered bat ancestors also acquired such a motif, the sampling bias in descriptions of remote bat viruses is dramatic. This should be discussed. Also state clearly that this site is only predicted so far and that experimental evidence for its biological function and its potential impact on pathogenesis are required.

We agree that the diversity in bat CoVs is undersampled and that it is possible that a bat CoV progenitor could be discovered that contains a polybasic site. We say nothing against this in the manuscript. However, given the diversity of SARS-like CoVs already sampled from bats, pangolins, and many other animals - none of which possess this insertion - it is obviously reasonable to hypothesize that it may have been gained in the lineage leading to SARS-CoV-2. This is explicitly discussed in our revision. While other CoVs utilize polybasic sites, we did refer to the site as a predicted cleavage site. We are more explicit in our revision that this will require experimental verification, although it is important to keep in mind that the sequences that define polybasic (furin) cleavage sites have been defined with high precision. It is very likely given the exact location at the S1/S2 junction and the fact that the sequence (RRAR) conforms precisely to an optimal cleavage site (RRXR) for furin or furin-like endoproteases that this site is utilized. We also state that it will be necessary to test the effect of the polybasic site on pathogenesis, which will require the establishment of an animal model.

The predicted O-linked glycosylation sites are mysterious. What do the authors imply with those sites? In silico prediction of O-linked glycosylation sites is not robust and whether these sites indeed exist requires experimental validation. Even if those sites exist, why are they relevant? This is not addressed at all. If the authors assume these sites constitute part of a glycan shield, they should say so and weigh their assumption carefully.

Although not previously described for CoV proteins, numerous other viral proteins have mucin-like domains that are involved in immune evasion. The revised text adds relevant references and is more explicit about the potential relevance of the predicted O-linked glycan sites. As stated previously, we consider that these sites - if indeed utilized - may be part of the glycan shield and have made this point directly in the revised text. We also agree that the predicted O-linked glycosylation sites require experimental validation not only in SARS-CoV-2 but in other CoVs that have similar predicted sites. This was also already stated explicitly, but in our revision we have stated this even more directly.

Finally, the main argument against a hypothetical lab origin seems the required reconstruction of a backbone of a bat virus of unknown pathogenesis. It does not seem feasible that any scientist would disembark on such an uncertain endeavor. This difficulties of coronavirus reverse genetics should be stated clearly.

The reviewer's restatement of our argument is correct. We reiterate that the purpose of our manuscript was not to refute the conspiracy theory that SARS-CoV-2 was bioengineered. Rather, it was carefully designed to be a balanced and unbiased assessment of the available data. We added an additional statement about the difficulties of CoV reverse engineering. However, the statement that no scientist would embark on such an endeavor is a subjective one with no supporting evidence.

Nature rebuttal February 21, 2020 8:37 PM

Editor's comments:

While the Perspective is interesting and timely one of our referees raised concerns (also emphasised to the editors) about whether such a piece would feed or quash the conspiracy theories.

Critically, the main purpose of our manuscript was not to quash conspiracy theories. Rather, our aim was to carefully examine in a balanced and unbiased manner the evidence for and against a number of possible probable scenarios for the proximal origins of SARS-CoV-2.

But more importantly this reviewer feels, and we agree, that the Perspective would quickly become outdated when more scientific data are published (for example on potential reservoir hosts).

We believe our piece would remain relevant if more data becomes available because it would potentially confirm which of these scenarios is correct. Most importantly, our manuscript sets out what evidence would be needed to test the hypotheses outlined and will serve as an important starting point for guiding future research. In addition, we make it clear that there is a very real possibility that no definitive intermediate host will ever be found.

Nature rebuttal February 21, 2020 8:37 PM

The Proximal Origin of SARS-CoV-2

Kristian G. Andersen[1,2*], Andrew Rambaut[3], W. Ian Lipkin[4], Edward C. Holmes[5] & Robert F. Garry[6,7]

[1]Department of Immunology and Microbiology, The Scripps Research Institute, La Jolla, CA 92037, USA.

[2]Scripps Research Translational Institute, La Jolla, CA 92037, USA.

[3]Institute of Evolutionary Biology, University of Edinburgh, Edinburgh, UK.

[4]Center for Infection and Immunity, Mailman School of Public Health of Columbia University, New York, New York, USA.

[5]Marie Bashir Institute for Infectious Diseases and Biosecurity, School of Life and Environmental Sciences and School of Medical Sciences, The University of Sydney, Sydney, Australia.

[6]Tulane University, School of Medicine, Department of Microbiology and Immunology, New Orleans, LA, USA.

[7]Zalgen Labs, LCC, Germantown, MD, USA.

*Corresponding author:

Kristian G. Andersen

Department of Immunology and Microbiology,

The Scripps Research Institute,

La Jolla, CA 92037, USA.

February 22, 2020 2:17 AM

Since the first reports of a novel pneumonia (COVID-19) in Wuhan city, Hubei province, China there has been considerable discussion and uncertainty over the origin of the causative virus, SARS-CoV-2. Infections with SARS-CoV-2 are now widespread in China, with cases in every province. As of 14 February 2020, 64,473 such cases have been confirmed, with 1,384 deaths attributed to the virus. These official case numbers are likely an underestimate because of limited reporting of mild and asymptomatic cases, and the virus is clearly capable of efficient human-to-human transmission. Based on the possibility of spread to countries with weaker healthcare systems, the World Health Organization has declared the COVID-19 outbreak a Public Health Emergency of International Concern (PHEIC). There are currently neither vaccines nor specific treatments for this disease.

SARS-CoV-2 is the seventh member of the *Coronaviridae* known to infect humans. Three of these viruses, SARS CoV-1, MERS CoV, and SARS-CoV-2, can cause severe disease; four, HKU1, NL63, OC43 and 229E, are associated with mild respiratory symptoms. Herein, we review what can be deduced about the origin and early evolution of SARS-CoV-2 from the comparative analysis of available genome sequence data. In particular, we offer a perspective on the notable features in the SARS-CoV-2 genome and discuss scenarios by which these features could have arisen. Importantly, this analysis provides evidence that SARS-CoV-2 is not a laboratory construct nor a purposefully manipulated virus.

The genomic comparison of both alpha- and betacoronaviruses (family *Coronaviridae*) described below identifies two notable features of the SARS-CoV-2 genome: (i) based on structural modeling and early biochemical experiments, SARS-CoV-2 appears to be optimized for binding to the human ACE2 receptor; (ii) the highly variable spike (S) protein of SARS-CoV-2 has a polybasic (furin) cleavage site at the S1 and S2 boundary via the insertion of twelve nucleotides. Additionally, this event led to the acquisition of three predicted O-linked glycans around the polybasic cleavage site.

Mutations in the receptor binding domain of SARS-CoV-2

The receptor binding domain (RBD) in the spike protein of SARS-CoV and SARS-related coronaviruses is the most variable part of the virus genome. Six residues in the RBD appear to be critical for binding to the human ACE2 receptor and determining host range[1]. Using coordinates based on the Urbani strain of SARS-CoV, they are Y442, L472, N479, D480, T487, and Y491[1]. The corresponding residues in SARS-CoV-2 are L455, F486, Q493, S494, N501, and Y505. Five of these six residues are mutated in SARS-CoV-2 compared to its most closely related virus, RaTG13 sampled from a *Rhinolophus affinis* bat, to which it is ~96% identical[2] (**Figure 1a**). Based on modeling[1] and biochemical experiments[3,4], SARS-CoV-2 seems to have an RBD that may bind with high affinity to ACE2 from human, non-human primate, ferret, pig, and cat, as well as other species with high receptor homology[1]. In contrast, SARS-CoV-2 may bind less efficiently to ACE2 in other species associated with SARS-like viruses, including rodents and civets[1].

The phenylalanine (F) at residue 486 in the SARS-CoV-2 S protein corresponds to L472 in the SARS-CoV Urbani strain. Notably, in SARS-CoV cell culture experiments the L472 mutates to phenylalanine (L472F)[5], which is predicted to be optimal for binding of the SARS-CoV RBD to the human ACE2 receptor[6]. However, a phenylalanine in this position is also present in several SARS-like CoVs from bats (**Figure 1a**). While these analyses suggest that SARS-CoV-2 may be capable of binding the human ACE2 receptor with high affinity, the interaction is not predicted to be optimal[1]. Additionally, several of the key residues in the RBD of SARS-CoV-2 are different to those previously described as optimal for human ACE2 receptor binding[6]. In contrast to these computational predictions, recent binding studies indicate that SARS-CoV-2 binds with high affinity to human ACE2[7]. Thus the SARS-CoV-2 spike appears to be the result of selection on human or human-like ACE2 permitting another optimal binding solution to arise. This is strong evidence that SARS-CoV-2 is *not* the product of genetic engineering.

Polybasic cleavage site and O-linked glycans

The second notable feature of SARS-CoV-2 is a predicted polybasic cleavage site (RRAR) in the spike protein at the junction of S1 and S2, the two subunits of the spike protein (**Figure 1b**)[8,9]. In addition to two basic arginines and an alanine at the cleavage site, a leading proline is also inserted; thus, the fully inserted sequence is PRRA (**Figure 1b**). The strong turn created by the proline insertion is predicted to result in the addition of O-linked glycans to S673, T678, and S686 that flank the polybasic cleavage site. A polybasic cleavage site has not previously been observed in related lineage B betacoronaviruses and is a unique feature of SARS-CoV-2. Some human betacoronaviruses, including HCoV-HKU1 (lineage A), have polybasic cleavage sites, as well as predicted O-linked glycans near the S1/S2 cleavage site.

While the functional consequence of the polybasic cleavage site in SARS-CoV-2 is unknown, experiments with SARS-CoV have shown that engineering such a site at the S1/S2 junction enhances cell–cell fusion but does not affect virus entry[10]. Polybasic cleavage sites allow effective cleavage by furin and other proteases, and can be acquired at the junction of the two subunits of the haemagglutinin (HA) protein of avian influenza viruses in conditions that select for rapid virus replication and transmission (e.g.,. highly dense chicken populations). HA serves a similar function in cell-cell fusion and viral entry as the coronavirus S protein. Acquisition of a polybasic cleavage site in HA, by either insertion or recombination, converts low pathogenicity avian influenza viruses into highly pathogenic forms[11–13]. The acquisition of polybasic

February 22, 2020 2:17 AM

cleavage sites by the influenza virus HA has also been observed after repeated forced passage in cell culture or through animals[14,15]. Similarly, an avirulent isolate of Newcastle Disease virus became highly pathogenic during serial passage in chickens by incremental acquisition of a polybasic cleavage site at the junction of its fusion protein subunits[16]. The potential function of the three predicted O-linked glycans is less clear, but they could create a "mucin-like domain" that would shield potential epitopes or key residues on the SARS-CoV-2 spike protein(Bagdonaite and Wandall 2018; Tran et al. 2014). Although the algorithms for prediction of O-linked glycosylation are robust(Steentoft et al. 2013), biochemical analyses or structural studies are required to determine whether or not the predicted O-linked glycan sites are utilized.

Theories of SARS-CoV-2 origins

It is improbable that SARS-CoV-2 emerged through laboratory manipulation of an existing SARS-related coronavirus. As noted above, the RBD of SARS-CoV-2 is optimized for human ACE2 receptor binding with an efficient binding solution different to that which would have been predicted. Further, if genetic manipulation had been performed, one would expect that one of the several reverse genetic systems available for betacoronaviruses would have been used. However, this is not the case as the genetic data shows that SARS-CoV-2 is not derived from any previously used virus backbone[17]. Instead, we propose two scenarios that can plausibly explain the origin of SARS-CoV-2: (i) natural selection in a non-human animal host prior to zoonotic transfer, and (ii) natural selection in humans following zoonotic transfer. We also discuss whether selection during passage in culture could have given rise to the same observed features.

Selection in an animal host

As many of the early cases of COVID-19 were linked to the Huanan seafood and wildlife market in Wuhan, it is possible that an animal source was present at this location. Given the similarity of SARS-CoV-2 to bat SARS-like CoVs, particularly RaTG13, it is plausible that bats serve as reservoir hosts for SARS-CoV-2. It is important, however, to note that previous outbreaks of betacoronaviruses in humans involved direct exposure to animals other than bats, including civets (SARS) and camels (MERS), that carry viruses that are genetically very similar to SARS-CoV or MERS-CoV, respectively. By analogy, viruses closely related to SARS-Cov-2 may be circulating in one or more animal species. Initial analyses indicate that Malayan pangolins (*Manis javanica*) illegally imported into Guangdong province contain a CoV that is similar to SARS-CoV-2[(Wong et al. 2020; Phylodynamic Analysis | 90 genomes | ...)]. Although the bat virus RaTG13 remains the closest relative to SARS-CoV-2 across the whole genome, the Malayan pangolin CoV is identical to SARS-CoV-2 at all six key RBD residues (**Figure 1**). However, no pangolin CoV has yet been identified that is sufficiently similar to SARS-CoV-2 across its entire genome to support direct human infection. In addition, the pangolin CoV does not carry a polybasic cleavage site insertion. For a precursor virus to acquire the polybasic cleavage site and mutations in the spike protein suitable for human ACE2 receptor binding, an animal host would likely have to have a high population density – to allow natural selection to proceed efficiently – and an ACE2 gene that is similar to the human orthologue. Further characterization of CoVs in pangolins and other animals that may harbour SARS-CoV-like viruses should be a public health priority.

Cryptic adaptation to humans

It is also possible that a progenitor to SARS-CoV-2 jumped from a non-human animal to humans, with the genomic features described above acquired through adaptation during subsequent human-to-human transmission. We surmise that once these adaptations were acquired (either together or in series) it would enable the outbreak to take-off, producing a sufficiently large and unusual cluster of pneumonia cases to trigger the surveillance system that ultimately detected it.

All SARS-CoV-2 genomes sequenced so far have the well adapted RBD and the polybasic cleavage site and are thus derived from a common ancestor that had these features. The presence of an RBD in pangolins that is very similar to the one in SARS-CoV-2 means that this was likely already present in the virus that

February 22, 2020 2:17 AM

jumped to humans, even if we don't yet have the exact non-human progenitor virus. This leaves the polybasic cleavage site insertion to occur during human-to-human transmission. Following the example of the influenza A virus HA gene, a specific insertion or recombination event is required to enable the emergence of SARS-CoV-2 as an epidemic pathogen.

Estimates of the timing of the most recent common ancestor (tMRCA) of SARS-CoV-2 using currently available genome sequence data point to virus emergence in late November to early December 2019[19,20], compatible with the earliest retrospectively confirmed cases[21]. Hence, this scenario presumes a period of unrecognised transmission in humans between the initial zoonotic transfer event and the acquisition of the polybasic cleavage site. Sufficient opportunity could occur if there had been many prior zoonotic events producing short chains of human-to-human transmission (so-called 'stuttering chains') over an extended period. This is essentially the situation for MERS-CoV in the Arabian Peninsula where all the human cases are the result of repeated jumps of the virus from dromedary camels, producing single infections or short chains of transmission that eventually resolve. To date, after 2,499 cases over 8 years, no human adaptation has emerged that has allowed MERS-CoV to take hold in the human population.

How could we test whether cryptic spread of SARS-CoV-2 enabled human adaptation? Metagenomic studies of banked serum samples could provide important information, but given the relatively short period of viremia it may be impossible to detect low level SARS-CoV-2 circulation in historical samples. Retrospective serological studies potentially could be informative and a few such studies have already been conducted. One found that animal importation traders had a 13% seropositivity to coronaviruses[22], while another noted that 3% residents of a village in Southern China were seropositive to these viruses[23]. Interestingly, 200 residents of Wuhan did not show coronavirus seroreactivity. Critically, however, these studies could not have distinguished whether positive serological responses were due to a prior infection with SARS-CoV or SARS-CoV-2. Further retrospective serological studies should be conducted to determine the extent of prior human exposure to betacoronaviruses in different geographic areas, particularly using assays that can distinguish among multiple betacoronaviruses.

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 in multiple laboratories across the world[24–27]. There are also documented instances of the laboratory acquisition of SARS-CoV by laboratory personnel working under BSL-2 containment[28,29]. We must therefore consider the possibility of a deliberate or inadvertent release of SARS-CoV-2. In theory, it is possible that SARS-CoV-2 acquired the observed RBD mutations site during adaptation to passage in cell culture, as has been observed in studies with SARS-CoV[5] as well as MERS-CoV[30]. However, the acquisition of the polybasic cleavage site or O-linked glycans - if functional - argues against this scenario. New polybasic cleavage sites have only been observed after prolonged passaging of low pathogenicity avian influenza virus in cell culture or animals. Furthermore, the generation of SARS-CoV-2 by cell culture or animal passage would have required prior isolation of a progenitor virus with a very high genetic similarity. Subsequent generation of a polybasic cleavage site would have then required an intense program of passage in cell culture or animals with ACE-2 receptor similar to humans (e.g. ferrets). It is also questionable whether generation of the O-linked glycans would have occurred on cell culture passage, as such mutations typically suggest the involvement of an immune system, that is not present *in vitro*.

Conclusions

In the midst of the global COVID-19 public health emergency it is reasonable to wonder why the origins of the epidemic matter. A detailed understanding of how an animal virus jumped species boundaries to infect humans so productively will help in the prevention of future zoonotic events. For example, if SARS-CoV-2

February 22, 2020 2:17 AM

pre-adapted in another animal species then we are at risk of future re-emergence events even if the current epidemic is controlled. In contrast, if the adaptive process we describe occurred in humans, then even if we have repeated zoonotic transfers they are unlikely to take-off unless the same series of mutations occurs. In addition, identifying the closest animal relatives of SARS-CoV-2 will greatly assist studies of virus function. Indeed, the availability of the RaTG13 bat sequence facilitated the comparative genomic analysis performed here, helping to reveal the key mutations in the RBD as well as the polybasic cleavage site insertion.

The genomic features described here may in part explain the infectiousness and transmissibility of SARS-CoV-2 in humans. Although genomic evidence does not support the idea that SARS-CoV-2 is a laboratory construct, it is currently impossible to prove or disprove the other theories of its origin described here, and it is unclear whether future data 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 revealing virus origins. In addition, it would be helpful to obtain more genetic and functional data about the virus, including experimental studies of receptor binding and the role of the polybasic cleavage site and predicted O-linked glycans. The identification of a potential intermediate host of SARS-CoV-2, as well as the sequencing of very early cases including those not connected to the Wuhan market, would similarly be highly informative. Irrespective of how SARS-CoV-2 originated, the ongoing surveillance of pneumonia in humans and other animals is clearly of utmost importance.

Acknowledgements

We thank all those who have contributed SARS-CoV-2 genome sequences to the GISAID database (https://www.gisaid.org/) and contributed analyses and ideas to Virological.org (http://virological.org/). We thank M. Farzan for discussions. We thank the Wellcome Trust for supporting this work. ECH is supported by an ARC Australian Laureate Fellowship (FL170100022). KGA is a Pew Biomedical Scholar and is supported by NIH NIAID grant U19AI135995. AR is supported by the Wellcome Trust (Collaborators Award 206298/Z/17/Z – ARTIC network) and the European Research Council (grant agreement no. 725422 – ReservoirDOCS). RFG is supported by NIH grants U19AI135995, U54 HG007480 and U19AI142790.

Figure Legends

Figure 1 | a) Mutations in contact residues of the SARS-CoV-2 spike protein. The spike protein of SARS-CoV-2 (top) was aligned against the most closely related SARS-like CoVs and SARS-CoV-1. Key residues in the spike protein that make contact to the ACE2 receptor are marked with blue boxes in both SARS-CoV-2 and the SARS-CoV Urbani strain. **b) Acquisition of polybasic cleavage site and O-linked glycans**. Both the polybasic cleavage site and the three adjacent predicted O-linked glycans are unique to SARS-CoV-2 and not previously seen in lineage B betacoronaviruses. Sequences shown are from NCBI GenBank, accession numbers MN908947, MN996532, AY278741, KY417146 & MK211376. The pangolin coronavirus sequences are a consensus generated from SRR10168377 and SRR10168378 (NCBI BioProject PRJNA573298)(Wong et al. 2020; Liu et al. 2019).

February 22, 2020 2:17 AM

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. Wu, F. *et al.* A new coronavirus associated with human respiratory disease in China. *Nature* (2020) doi:10.1038/s41586-020-2008-3.
3. 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.
4. 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.
5. 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).
6. Cui, J., Li, F. & Shi, Z.-L. Origin and evolution of pathogenic coronaviruses. *Nat. Rev. Microbiol.* **17**, 181–192 (2019).
7. Wrapp, D. *et al.* Cryo-EM Structure of the 2019-nCoV Spike in the Prefusion Conformation. *bioRxiv* 2020.02.11.944462 (2020) doi:10.1101/2020.02.11.944462.
8. Gallaher, W. Analysis of Wuhan Coronavirus: Deja Vu. *Virological* http://virological.org/t/analysis-of-wuhan-coronavirus-deja-vu/357 (2020).
9. Coutard, B. *et al.* The spike glycoprotein of the new coronavirus 2019-nCoV contains a furin-like cleavage site absent in CoV of the same clade. *Antiviral Res.* 104742 (2020).
10. 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).

February 22, 2020 2:17 AM

11. 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).

12. Alexander, D. J. & Brown, I. H. History of highly pathogenic avian influenza. *Rev. Sci. Tech.* **28**, 19–38 (2009).

13. 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).

14. Ito, T. *et al.* Generation of a highly pathogenic avian influenza A virus from an avirulent field isolate by passaging in chickens. *J. Virol.* **75**, 4439–4443 (2001).

15. Li, S. Q., Orlich, M. & Rott, R. Generation of seal influenza virus variants pathogenic for chickens, because of hemagglutinin cleavage site changes. *J. Virol.* **64**, 3297–3303 (1990).

16. Shengqing, Y. *et al.* Generation of velogenic Newcastle disease viruses from a nonpathogenic waterfowl isolate by passaging in chickens. *Virology* **301**, 206–211 (2002).

17. Menachery, V. D. *et al.* A SARS-like cluster of circulating bat coronaviruses shows potential for human emergence. *Nat. Med.* **21**, 1508–1513 (2015).

18. Wong, M. C., Javornik Cregeen, S. J., Ajami, N. J. & Petrosino, J. F. Evidence of recombination in coronaviruses implicating pangolin origins of nCoV-2019. *bioRxiv* 2020.02.07.939207 (2020) doi:10.1101/2020.02.07.939207.

19. Phylodynamic Analysis | 90 genomes | 12 Feb 2020 – Novel 2019 coronavirus / nCoV-2019 Genomic Epidemiology. http://virological.org/t/phylodynamic-analysis-90-genomes-12-feb-2020/356 (2020).

20. Phylodynamic estimation of incidence and prevalence of novel coronavirus (nCoV) infections through time – Novel 2019 coronavirus / nCoV-2019 Genomic Epidemiology. http://virological.org/t/phylodynamic-estimation-of-incidence-and-prevalence-of-novel-coronavirus-ncov-infections-through-time/391 (2020).

21. Huang, C. *et al.* Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China.

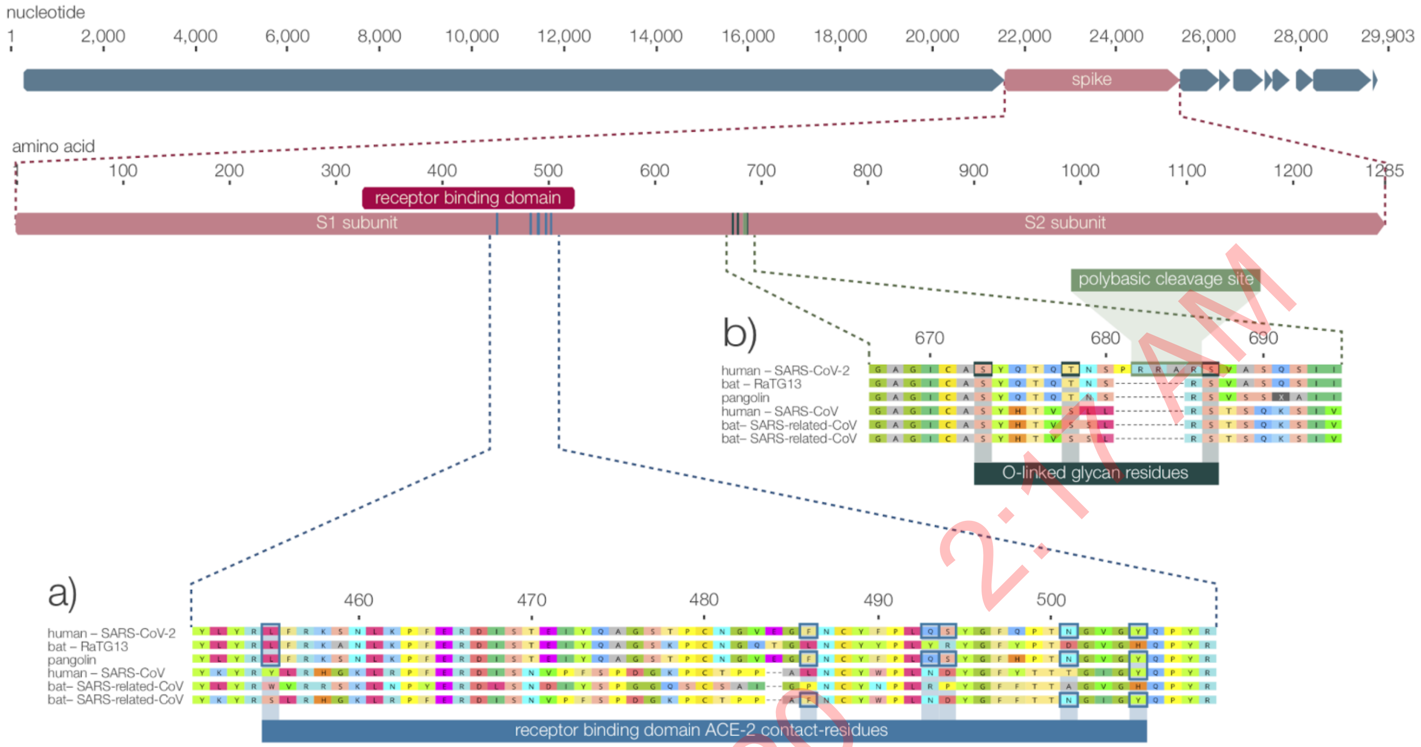
February 22, 2020 2:17 AM

Lancet (2020) doi:10.1016/S0140-6736(20)30183-5.

22. Centers for Disease Control and Prevention (CDC). Prevalence of IgG antibody to SARS-associated coronavirus in animal traders--Guangdong Province, China, 2003. *MMWR Morb. Mortal. Wkly. Rep.* **52**, 986–987 (2003).
23. Wang, N. *et al.* Serological Evidence of Bat SARS-Related Coronavirus Infection in Humans, China. *Virol. Sin.* **33**, 104–107 (2018).
24. Ge, X.-Y. *et al.* Isolation and characterization of a bat SARS-like coronavirus that uses the ACE2 receptor. *Nature* **503**, 535–538 (2013).
25. 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).
26. 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).
27. 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).
28. Lim, P. L. *et al.* Laboratory-acquired severe acute respiratory syndrome. *N. Engl. J. Med.* **350**, 1740–1745 (2004).
29. Senior, K. Recent Singapore SARS case a laboratory accident. *Lancet Infect. Dis.* **3**, 679 (2003).
30. Letko, M. *et al.* Adaptive Evolution of MERS-CoV to Species Variation in DPP4. *Cell Rep.* **24**, 1730–1737 (2018).

February 22, 2020 2:17 AM

Figure 1.


nucleotide
1
2,000
4,000
6,000
8,000
10,000
12,000
14,000
16,000
18,000
20,000
22,000
24,000
26,000
28,000
29,903
spike
amino acid
100
200
300
400
500
600
700
800
900
1000
1100
1200
1285
receptor binding domain
S1 subunit
S2 subunit
polybasic cleavage site
b)
670
680
690
human – SARS-CoV-2
bat – RaTG13
pangolin
human – SARS-CoV
bat– SARS-related-CoV
bat– SARS-related-CoV
GAGICASYQTQTNSPRRARSVASQSII
GAGICASYQTQTNS---RSVASQSII
GAGICASYQTQTNS---RSVSSXAII
GAGICASYHTVSLL---RSTSQKSIV
GAGICASYHTVSSL---RSTSQKSIV
GAGICASYHTVSSL---RSTSQKSIV
O-linked glycan residues
a)
460
470
480
490
500
human – SARS-CoV-2
bat – RaTG13
pangolin
human – SARS-CoV
bat– SARS-related-CoV
bat– SARS-related-CoV
YLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVEGFNCYFPLQSYGFQPTNGVGYQPYR
YLYRLFRKANLKPFERDISTEIYQAGSKPCNGQTGLNCYYPLYRYGFYPTDGVGHQPYR
YLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVEGFNCYFPLQSYGFHPTNGVGYQPYR
YKYRYLRHGKLRPFERDISNVPFSPDGKPCTPP---ALNCYWPLNDYGFYTTTGIGYQPYR
YLYRWVRRSKLNPYERDLSNDIYSPGGQSCSAI---GPNCYNPLRPYGFFTTAGVGHQPYR
YKYRSLRHGKLRPFERDISNVPFSPDGKPCTPP---AFNCYWPLNDYGFFTTNGIGYQPYR
receptor binding domain ACE-2 contact-residues

February 22, 2020 2:17 AM

Referee #1:

Anderson presented a timely manuscript to share their points of view about the origin of SARS-CoV-2. There are several rumors about the origin of this virus. However, these "hypotheses" are entirely based on very limited, if any, scientific evidences. This reviewer sees most of the arguments raised by the authors are valid and convincing. However, the authors might want to consider these minor suggestions:

1. The sections for the RBD and cleavage site of Spike protein basically have summarized the existing findings from other recent publications. The authors might want to spell out that these two sections are review summaries. In addition, the author can present these two sections in a more condense format and save some space for something else (also see points 6 and 7 below)

We have edited these sections to be more concise. However, we think that it is important that the key genomic features are described in detail as they form the basis of the discussion that follows.

2. Fig. 1. This figure has 6 aligned sequences, but with only 5 sequence titles. The order of these titles are also not correct.

We have corrected these mistakes and updated the figure and titles.

3. Lines 170 -174. It is correct that no adaptive mutation has been found in the spike of MERS-CoV. Deletions in other ORF regions, however, were detected in some human MERS-CoV viruses (PMID: 26981770). In addition, the 29nt deletion of human SARS-CoV (PMID: 12958366) was suggested to have effects on host adaptation. The authors should also consider these findings. It is premature to say that this would not happen in SARS-CoV-2.

We have revised the text to incorporate this information as we agree that they are important for context.

4. Line 194. The accident at Singapore occurred in a BSL3, not BSL2, containment.

We have corrected this.

5. Line 194. Laboratory escapes of SARS occurred in Singapore, China and Taiwan (PMID: 16830004).

We have added this reference as well as the point about laboratory escapes of SARS-CoV in Singapore, China and Taiwan.

6. There are two recent reports about coronaviruses in pangolins. The authors might want to comment on these.

We have included these references as well as several others that have investigated pangolin CoV. In addition, and in response to Reviewer 2 (see below) we should point out that these additional pangolin CoV sequences do not further clarify the different scenarios discussed in our manuscript. There is nothing in these reports that changes our statements regarding the role of pangolins.

7. Optional: Can the authors share their views on the possibility of having a lab escape of a natural coronavirus? This is also one of the hypotheses that have been extensively discussed. The reviewer understand that this is entirely a different topic, but any insights are welcomed.

Nature Rebuttal February 22, 2020 4:04PM

Escape of a natural CoV (SARS-CoV-2 or a close progenitor) from a lab could not be distinguished from an animal-to-human transfer in another environment. Given the limited numbers of labs doing work of this kind compared to the frequent opportunities for animal-to-human transfer, it is obvious that the latter is much less likely than the former.

Nature Rebuttal February 22, 2020 4:04PM

Referee #2:

This is a perspective discussing evidence against a hypothetical lab origin of SARS-CoV-2. The paper addresses suboptimal composition of ACE2-binding sites in the RBD, 3 predicted O-linked glycosylation sites and a furin cleavage site in the glycoprotein that was speculated upon before.

The paper is itself interesting, but unnecessarily speculative. It's not clear why the authors do not refute a hypothetical lab origin in their coming publication on the ancestors of SARS-CoV-2 in bats and pangolins. The tree showing diverse pangolin viruses has kindly been made available by some of the authors in GISAID. Once the authors publish their new pangolin sequences, a lab origin will be extremely unlikely. It is not clear why the authors rush with a speculative perspective if their central hypothesis can be supported by their own data. Please explain.

Our manuscript is written to explore the potential origin of SARS-CoV-2. We do not believe it is speculative: rather, it simply takes the available data and proposes a series of hypotheses that explain how these data may have arisen. We try to do this in a logical, balanced and unbiased manner: this is critically important because it gives our work credibility. It is science, not opinion.

Unfortunately, the newly available pangolin sequences do not elucidate the origin of SARS-CoV-2 or refute a lab origin. Hence, the reviewer is incorrect on this point. To clarify, while the RBD from the Guangdong pangolin CoVs is the closest to that found SARS-CoV-2, they are more divergent in the remainder of the viral genome (for which the bat virus RaTG13 is still the closest) and do not possess the polybasic cleavage site insertion. Hence, there is no evidence on present data that the pangolin CoVs are directly related to the COVID-19 epidemic.

Another critical aspect of this text is the complete lack of referencing to a potential debate on a hypothetical lab origin. Who said this, why is this considered a problem? There are indeed a few apparently uninformed statements claiming the virus may be a Chinese bioweapon, but is this really problematic on a larger scale? The central reason for issuing this text must be exhaustively referenced and discussed.

The possibility that SARS-CoV-2 originated as an engineered bioweapon has had widespread discussion in the press and on social media (particularly in China) and is hence problematic on a very large scale. This particular topic was also the reason for a recent request from the White House: https://bit.ly/2HMndCi. A group of public health scientists recently wrote a letter to *The Lancet* in which they *"strongly condemn conspiracy theories suggesting that COVID-19 does not have a natural origin."* We now reference this publication (the publication itself also references our pre-print of this manuscript).

It appears that the reviewer is considering bioweapons, engineering, and lab accidents as one and the same theory lumped in under a "lab origin". This is inappropriate and will likely only increase confusion. We do briefly discuss the notion of deliberate engineering of the virus and explain how the evidence is strongly against this explanation. Importantly, we do not discuss the concept of 'bioweapons' per se as there are many, legitimate, research uses of bio-engineering of viruses.

As we very clearly state in the manuscript, there are far more subtle scenarios that need to be considered carefully and scientifically: for example, of accidental infections in a lab while culturing SARS-like CoVs. Since accidental infections and other lab 'escapes' happen frequently across the world (and as we mention, happened multiple times with SARS-CoV-1 following the SARS epidemic) we firmly believe that this

Nature Rebuttal February 22, 2020 4:04PM

discussion is of major importance and must be had. In particular, the culturing of SARS-like CoVs from animals is typically performed under BSL-2 and has been ongoing for years.

Dismissing this potential accidental scenario out of hand - or considering it in the same category as conspiracy theories about bioweapons and deliberate engineering (that, as we outline, are clearly wrong) - would be irresponsible. Our manuscript will only serve its purpose if it considers all possible hypotheses equally. Any perceived bias will undermine its credibility.

We have modified the text to more clearly state the scenarios we are considering.

The authors state that a predicted polybasic cleavage sites is unique to SARS-CoV-2 in SARS viruses. Who knows how many out of thousands undiscovered bat ancestors also acquired such a motif, the sampling bias in descriptions of remote bat viruses is dramatic. This should be discussed. Also state clearly that this site is only predicted so far and that experimental evidence for its biological function and its potential impact on pathogenesis are required.

We agree that the diversity in bat CoVs is undersampled and that it is possible that a bat CoV progenitor could be discovered that contains a polybasic site. We say nothing against this in the manuscript. However, given the diversity of SARS-like CoVs already sampled from bats, pangolins, and many other animals - none of which possess this insertion - it is obviously reasonable to hypothesize that it may have been gained in the lineage leading to SARS-CoV-2. This is explicitly discussed in our revision.

While other CoVs utilize polybasic sites, we did refer to the site as a predicted cleavage site. We are more explicit in our revision that this will require experimental verification, although it is important to note that the sequences that define polybasic (furin) cleavage sites have been defined with high precision. It is very likely given the exact location at the S1/S2 junction and the fact that the sequence (RRAR) conforms to an optimal cleavage site (RRXR) for furin or furin-like endoproteases that this site is utilized. We also state that it will be necessary to test the effect of the polybasic site on pathogenesis, which will require the establishment of an animal model.

The predicted O-linked glycosylation sites are mysterious. What do the authors imply with those sites? In silico prediction of O-linked glycosylation sites is not robust and whether these sites indeed exist requires experimental validation. Even if those sites exist, why are they relevant? This is not addressed at all. If the authors assume these sites constitute part of a glycan shield, they should say so and weigh their assumption carefully.

Although not previously described for CoV proteins, numerous other viral proteins have mucin-like domains that are involved in immune evasion. The revised text adds relevant references and is more explicit about the potential relevance of the predicted O-linked glycan sites. As stated previously, we consider that these sites - if indeed utilized - may be part of the glycan shield and have made this point directly in the revised text. We also agree that the predicted O-linked glycosylation sites require experimental validation not only in SARS-CoV-2 but in other CoVs that have similar predicted sites. This was also already stated explicitly, but in our revision we have stated this even more directly.

Finally, the main argument against a hypothetical lab origin seems the required reconstruction of a backbone of a bat virus of unknown pathogenesis. It does not seem feasible that any scientist would disembark on such an uncertain endeavor. This difficulties of coronavirus reverse genetics should be stated clearly.

Nature Rebuttal February 22, 2020 4:04PM

The reviewer’s restatement of our argument is correct. We reiterate that the purpose of our manuscript was not to refute the conspiracy theory that SARS-CoV-2 was bioengineered. Rather, it was carefully designed to be a balanced and unbiased assessment of the available data. We added an additional statement about the difficulties of CoV reverse engineering. However, the statement that no scientist would embark on such an endeavor is a subjective one with no supporting evidence.

Nature Rebuttal February 22, 2020 4:04PM

Editor's comments:

While the Perspective is interesting and timely one of our referees raised concerns (also emphasised to the editors) about whether such a piece would feed or quash the conspiracy theories.

Critically, the purpose of our manuscript was not to quash conspiracy theories. Rather, our aim was to carefully examine in a balanced and unbiased manner the evidence for and against a number of possible probable scenarios for the proximal origins of SARS-CoV-2.

But more importantly this reviewer feels, and we agree, that the Perspective would quickly become outdated when more scientific data are published (for example on potential reservoir hosts).

Of course, it is likely that more scientific data will swing the balance in favor of one hypothesis over another. However, the same can be said of many of the papers published on COVID-19: as we learn more about the virus and the disease so previous publications may be quickly revised. We contend, however, our piece will remain relevant even if more data becomes available because these data would potentially confirm which of these scenarios is correct. Most importantly, our manuscript sets out what evidence is needed to test the hypotheses outlined and will therefore serve as an important starting point for guiding future research. In addition, we make it clear that there is a very real possibility that no definitive intermediate host will ever be found.

Nature Rebuttal   February 22, 2020   4:04PM

The Proximal Origin of nCoV-19 or HCoV-19

Kristian G. Andersen[1,2*], Andrew Rambaut[3], W. Ian Lipkin[4], Edward C. Holmes[5] & Robert F. Garry[6,7]

[1]Department of Immunology and Microbiology, The Scripps Research Institute, La Jolla, CA 92037, USA.

[2]Scripps Research Translational Institute, La Jolla, CA 92037, USA.

[3]Institute of Evolutionary Biology, University of Edinburgh, Edinburgh, UK.

[4]Center for Infection and Immunity, Mailman School of Public Health of Columbia University, New York, New York, USA.

[5]Marie Bashir Institute for Infectious Diseases and Biosecurity, School of Life and Environmental Sciences and School of Medical Sciences, The University of Sydney, Sydney, Australia.

[6]Tulane University, School of Medicine, Department of Microbiology and Immunology, New Orleans, LA, USA.

[7]Zalgen Labs, LCC, Germantown, MD, USA.

*Corresponding author:

Kristian G. Andersen

Department of Immunology and Microbiology,

The Scripps Research Institute,

La Jolla, CA 92037, USA.

February 26, 2020 4:12 PM

Since the first reports of a novel pneumonia (COVID-19) in Wuhan city, Hubei province, China there has been considerable discussion and uncertainty over the origin of the causative virus, SARS-CoV-2. Infections with SARS-CoV-2 are now widespread in China, with cases in every province. As of 14 February 2020, 64,473 such cases have been confirmed, with 1,384 deaths attributed to the virus. These official case numbers are likely an underestimate because of limited reporting of mild and asymptomatic cases, and the virus is clearly capable of efficient human-to-human transmission. Based on the possibility of spread to countries with weaker healthcare systems, the World Health Organization has declared the COVID-19 outbreak a Public Health Emergency of International Concern (PHEIC). There are currently neither vaccines nor specific treatments for this disease.

SARS-CoV-2 is the seventh member of the *Coronaviridae* known to infect humans. Three of these viruses, SARS CoV-1, MERS CoV, and SARS-CoV-2, can cause severe disease; four, HKU1, NL63, OC43 and 229E, are associated with mild respiratory symptoms. Herein, we review what can be deduced about the origin and early evolution of SARS-CoV-2 from the comparative analysis of available genome sequence data. In particular, we offer a perspective on the notable features in the SARS-CoV-2 genome and discuss scenarios by which these features could have arisen. Importantly, this analysis provides evidence that SARS-CoV-2 is not a laboratory construct nor a purposefully manipulated virus.

The genomic comparison of both alpha- and betacoronaviruses (family *Coronaviridae*) described below identifies two notable features of the SARS-CoV-2 genome: (i) based on structural modeling and early biochemical experiments, SARS-CoV-2 appears to be optimized for binding to the human ACE2 receptor; (ii) the highly variable spike (S) protein of SARS-CoV-2 has a polybasic (furin) cleavage site at the S1 and S2 boundary via the insertion of twelve nucleotides. Additionally, this event led to the acquisition of three predicted O-linked glycans around the polybasic cleavage site.

Mutations in the receptor binding domain of SARS-CoV-2

The receptor binding domain (RBD) in the spike protein of SARS-CoV and SARS-related coronaviruses is the most variable part of the virus genome. Six residues in the RBD appear to be critical for binding to the human ACE2 receptor and determining host range[1]. Using coordinates based on the Urbani strain of SARS-CoV, they are Y442, L472, N479, D480, T487, and Y491[1]. The corresponding residues in SARS-CoV-2 are L455, F486, Q493, S494, N501, and Y505. Five of these six residues are mutated in SARS-CoV-2 compared to its most closely related virus, RaTG13 sampled from a *Rhinolophus affinis* bat, to which it is ~96% identical[2] (**Figure 1a**). Based on modeling[1] and biochemical experiments[3,4], SARS-CoV-2 seems to have an RBD that may bind with high affinity to ACE2 from human, non-human primate, ferret, pig, and cat, as well as other species with high receptor homology[1]. In contrast, SARS-CoV-2 may bind less efficiently to ACE2 in other species associated with SARS-like viruses, including rodents and civets[1].

The phenylalanine (F) at residue 486 in the SARS-CoV-2 S protein corresponds to L472 in the SARS-CoV Urbani strain. Notably, in SARS-CoV cell culture experiments the L472 mutates to phenylalanine (L472F)[5], which is predicted to be optimal for binding of the SARS-CoV RBD to the human ACE2 receptor[6]. However, a phenylalanine in this position is also present in several SARS-like CoVs from bats (**Figure 1a**). While these analyses suggest that SARS-CoV-2 may be capable of binding the human ACE2 receptor with high affinity, the interaction is not predicted to be optimal[1]. Additionally, several of the key residues in the RBD of SARS-CoV-2 are different to those previously described as optimal for human ACE2 receptor binding[6]. In contrast to these computational predictions, recent binding studies indicate that SARS-CoV-2 binds with high affinity to human ACE2[7]. Thus the SARS-CoV-2 spike appears to be the result of selection on human or human-like ACE2 permitting another optimal binding solution to arise. This is strong evidence that SARS-CoV-2 is *not* the product of genetic engineering.

Polybasic cleavage site and O-linked glycans

The second notable feature of SARS-CoV-2 is a predicted polybasic cleavage site (RRAR) in the spike protein at the junction of S1 and S2, the two subunits of the spike protein (**Figure 1b**)[8,9]. In addition to two basic arginines and an alanine at the cleavage site, a leading proline is also inserted; thus, the fully inserted sequence is PRRA (**Figure 1b**). The strong turn created by the proline insertion is predicted to result in the addition of O-linked glycans to S673, T678, and S686 that flank the polybasic cleavage site. A polybasic cleavage site has not previously been observed in related lineage B betacoronaviruses and is a unique feature of SARS-CoV-2. Some human betacoronaviruses, including HCoV-HKU1 (lineage A), have polybasic cleavage sites, as well as predicted O-linked glycans near the S1/S2 cleavage site.

While the functional consequence of the polybasic cleavage site in SARS-CoV-2 is unknown, experiments with SARS-CoV have shown that engineering such a site at the S1/S2 junction enhances cell–cell fusion but does not affect virus entry[10]. Polybasic cleavage sites allow effective cleavage by furin and other proteases, and can be acquired at the junction of the two subunits of the haemagglutinin (HA) protein of avian influenza viruses in conditions that select for rapid virus replication and transmission (e.g.,. highly dense chicken populations). HA serves a similar function in cell-cell fusion and viral entry as the coronavirus S protein. Acquisition of a polybasic cleavage site in HA, by either insertion or recombination, converts low pathogenicity avian influenza viruses into highly pathogenic forms[11–13]. The acquisition of polybasic

February 26, 2020 4:12 PM

cleavage sites by the influenza virus HA has also been observed after repeated forced passage in cell culture or through animals[14,15]. Similarly, an avirulent isolate of Newcastle Disease virus became highly pathogenic during serial passage in chickens by incremental acquisition of a polybasic cleavage site at the junction of its fusion protein subunits[16]. The potential function of the three predicted O-linked glycans is less clear, but they could create a “mucin-like domain” that would shield potential epitopes or key residues on the SARS-CoV-2 spike protein[17,18]. Although the algorithms for prediction of O-linked glycosylation are robust[19], biochemical analyses or structural studies are required to determine whether or not the predicted O-linked glycan sites are utilized.

Theories of SARS-CoV-2 origins

It is improbable that SARS-CoV-2 emerged through laboratory manipulation of an existing SARS-related coronavirus. As noted above, the RBD of SARS-CoV-2 is optimized for human ACE2 receptor binding with an efficient binding solution different to that which would have been predicted. Further, if genetic manipulation had been performed, one would expect that one of the several reverse genetic systems available for betacoronaviruses would have been used. However, this is not the case as the genetic data shows that SARS-CoV-2 is not derived from any previously used virus backbone[20]. Instead, we propose two scenarios that can plausibly explain the origin of SARS-CoV-2: (i) natural selection in a non-human animal host prior to zoonotic transfer, and (ii) natural selection in humans following zoonotic transfer. We also discuss whether selection during passage in culture could have given rise to the same observed features.

Selection in an animal host

As many of the early cases of COVID-19 were linked to the Huanan seafood and wildlife market in Wuhan, it is possible that an animal source was present at this location. Given the similarity of SARS-CoV-2 to bat SARS-like CoVs, particularly RaTG13, it is plausible that bats serve as reservoir hosts for SARS-CoV-2. It is important, however, to note that previous outbreaks of betacoronaviruses in humans involved direct exposure to animals other than bats, including civets (SARS) and camels (MERS), that carry viruses that are genetically very similar to SARS-CoV or MERS-CoV, respectively. By analogy, viruses closely related to SARS-Cov-2 may be circulating in one or more animal species. Initial analyses indicate that Malayan pangolins (*Manis javanica*) illegally imported into Guangdong province contain a CoV that is similar to SARS-CoV-2[21,22]. Although the bat virus RaTG13 remains the closest relative to SARS-CoV-2 across the whole genome, the Malayan pangolin CoV is identical to SARS-CoV-2 at all six key RBD residues (**Figure 1**). However, no pangolin CoV has yet been identified that is sufficiently similar to SARS-CoV-2 across its entire genome to support direct human infection. In addition, the pangolin CoV does not carry a polybasic cleavage site insertion. For a precursor virus to acquire the polybasic cleavage site and mutations in the spike protein suitable for human ACE2 receptor binding, an animal host would likely have to have a high population density – to allow natural selection to proceed efficiently – and an ACE2 gene that is similar to the human orthologue. Further characterization of CoVs in pangolins and other animals that may harbour SARS-CoV-like viruses should be a public health priority.

Cryptic adaptation to humans

It is also possible that a progenitor to SARS-CoV-2 jumped from a non-human animal to humans, with the genomic features described above acquired through adaptation during subsequent human-to-human transmission. We surmise that once these adaptations were acquired (either together or in series) it would enable the outbreak to take-off, producing a sufficiently large and unusual cluster of pneumonia cases to trigger the surveillance system that ultimately detected it.

All SARS-CoV-2 genomes sequenced so far have the well adapted RBD and the polybasic cleavage site and are thus derived from a common ancestor that had these features. The presence of an RBD in pangolins that is very similar to the one in SARS-CoV-2 means that this was likely already present in the virus that

February 26, 2020 4:12 PM

jumped to humans, even if we don't yet have the exact non-human progenitor virus. This leaves the polybasic cleavage site insertion to occur during human-to-human transmission. Following the example of the influenza A virus HA gene, a specific insertion or recombination event is required to enable the emergence of SARS-CoV-2 as an epidemic pathogen.

Estimates of the timing of the most recent common ancestor (tMRCA) of SARS-CoV-2 using currently available genome sequence data point to virus emergence in late November to early December 2019[23,24], compatible with the earliest retrospectively confirmed cases[25]. Hence, this scenario presumes a period of unrecognised transmission in humans between the initial zoonotic transfer event and the acquisition of the polybasic cleavage site. Sufficient opportunity could occur if there had been many prior zoonotic events producing short chains of human-to-human transmission (so-called 'stuttering chains') over an extended period. This is essentially the situation for MERS-CoV in the Arabian Peninsula where all the human cases are the result of repeated jumps of the virus from dromedary camels, producing single infections or short chains of transmission that eventually resolve. To date, after 2,499 cases over 8 years, no human adaptation has emerged that has allowed MERS-CoV to take hold in the human population.

How could we test whether cryptic spread of SARS-CoV-2 enabled human adaptation? Metagenomic studies of banked serum samples could provide important information, but given the relatively short period of viremia it may be impossible to detect low level SARS-CoV-2 circulation in historical samples. Retrospective serological studies potentially could be informative and a few such studies have already been conducted. One found that animal importation traders had a 13% seropositivity to coronaviruses[26], while another noted that 3% residents of a village in Southern China were seropositive to these viruses[27]. Interestingly, 200 residents of Wuhan did not show coronavirus seroreactivity. Critically, however, these studies could not have distinguished whether positive serological responses were due to a prior infection with SARS-CoV or SARS-CoV-2. Further retrospective serological studies should be conducted to determine the extent of prior human exposure to betacoronaviruses in different geographic areas, particularly using assays that can distinguish among multiple betacoronaviruses.

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 in multiple laboratories across the world[28–31]. There are also documented instances of the laboratory acquisition of SARS-CoV by laboratory personnel working under BSL-2 containment[32,33]. We must therefore consider the possibility of a deliberate or inadvertent release of SARS-CoV-2. In theory, it is possible that SARS-CoV-2 acquired the observed RBD mutations site during adaptation to passage in cell culture, as has been observed in studies with SARS-CoV[5] as well as MERS-CoV[34]. However, the acquisition of the polybasic cleavage site or O-linked glycans - if functional - argues against this scenario. New polybasic cleavage sites have only been observed after prolonged passaging of low pathogenicity avian influenza virus in cell culture or animals. Furthermore, the generation of SARS-CoV-2 by cell culture or animal passage would have required prior isolation of a progenitor virus with a very high genetic similarity. Subsequent generation of a polybasic cleavage site would have then required an intense program of passage in cell culture or animals with ACE-2 receptor similar to humans (e.g. ferrets). It is also questionable whether generation of the O-linked glycans would have occurred on cell culture passage, as such mutations typically suggest the involvement of an immune system, that is not present *in vitro*.

Conclusions

In the midst of the global COVID-19 public health emergency it is reasonable to wonder why the origins of the epidemic matter. A detailed understanding of how an animal virus jumped species boundaries to infect humans so productively will help in the prevention of future zoonotic events. For example, if SARS-CoV-2

February 26, 2020 4:12 PM

pre-adapted in another animal species then we are at risk of future re-emergence events even if the current epidemic is controlled. In contrast, if the adaptive process we describe occurred in humans, then even if we have repeated zoonotic transfers they are unlikely to take-off unless the same series of mutations occurs. In addition, identifying the closest animal relatives of SARS-CoV-2 will greatly assist studies of virus function. Indeed, the availability of the RaTG13 bat sequence facilitated the comparative genomic analysis performed here, helping to reveal the key mutations in the RBD as well as the polybasic cleavage site insertion.

The genomic features described here may in part explain the infectiousness and transmissibility of SARS-CoV-2 in humans. Although genomic evidence does not support the idea that SARS-CoV-2 is a laboratory construct, it is currently impossible to prove or disprove the other theories of its origin described here, and it is unclear whether future data 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 revealing virus origins. In addition, it would be helpful to obtain more genetic and functional data about the virus, including experimental studies of receptor binding and the role of the polybasic cleavage site and predicted O-linked glycans. The identification of a potential intermediate host of SARS-CoV-2, as well as the sequencing of very early cases including those not connected to the Wuhan market, would similarly be highly informative. Irrespective of how SARS-CoV-2 originated, the ongoing surveillance of pneumonia in humans and other animals is clearly of utmost importance.

Acknowledgements

We thank all those who have contributed SARS-CoV-2 genome sequences to the GISAID database (https://www.gisaid.org/) and contributed analyses and ideas to Virological.org (http://virological.org/). We thank M. Farzan for discussions. We thank the Wellcome Trust for supporting this work. ECH is supported by an ARC Australian Laureate Fellowship (FL170100022). KGA is a Pew Biomedical Scholar and is supported by NIH NIAID grant U19AI135995. AR is supported by the Wellcome Trust (Collaborators Award 206298/Z/17/Z – ARTIC network) and the European Research Council (grant agreement no. 725422 – ReservoirDOCS). RFG is supported by NIH grants U19AI135995, U54 HG007480 and U19AI142790.

Figure Legends

Figure 1 | a) Mutations in contact residues of the SARS-CoV-2 spike protein. The spike protein of SARS-CoV-2 (top) was aligned against the most closely related SARS-like CoVs and SARS-CoV-1. Key residues in the spike protein that make contact to the ACE2 receptor are marked with blue boxes in both SARS-CoV-2 and the SARS-CoV Urbani strain. **b) Acquisition of polybasic cleavage site and O-linked glycans**. Both the polybasic cleavage site and the three adjacent predicted O-linked glycans are unique to SARS-CoV-2 and not previously seen in lineage B betacoronaviruses. Sequences shown are from NCBI GenBank, accession numbers MN908947, MN996532, AY278741, KY417146 & MK211376. The pangolin coronavirus sequences are a consensus generated from SRR10168377 and SRR10168378 (NCBI BioProject PRJNA573298)[21,35].

Figure 2 | a) Recombination in spike protein.

February 26, 2020 4:12 PM

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. Wu, F. *et al.* A new coronavirus associated with human respiratory disease in China. *Nature* (2020) doi:10.1038/s41586-020-2008-3.

3. 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.

4. 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.

5. 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).

6. Cui, J., Li, F. & Shi, Z.-L. Origin and evolution of pathogenic coronaviruses. *Nat. Rev. Microbiol.* **17**, 181–192 (2019).

7. Wrapp, D. *et al.* Cryo-EM Structure of the 2019-nCoV Spike in the Prefusion Conformation. *bioRxiv* 2020.02.11.944462 (2020) doi:10.1101/2020.02.11.944462.

8. Gallaher, W. Analysis of Wuhan Coronavirus: Deja Vu. *Virological* http://virological.org/t/analysis-of-wuhan-coronavirus-deja-vu/357 (2020).

9. Coutard, B. *et al.* The spike glycoprotein of the new coronavirus 2019-nCoV contains a furin-like cleavage site absent in CoV of the same clade. *Antiviral Res.* 104742 (2020).

10. 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).

February 26, 2020 4:12 PM

11. 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).

12. Alexander, D. J. & Brown, I. H. History of highly pathogenic avian influenza. *Rev. Sci. Tech.* **28**, 19–38 (2009).

13. 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).

14. Ito, T. *et al.* Generation of a highly pathogenic avian influenza A virus from an avirulent field isolate by passaging in chickens. *J. Virol.* **75**, 4439–4443 (2001).

15. Li, S. Q., Orlich, M. & Rott, R. Generation of seal influenza virus variants pathogenic for chickens, because of hemagglutinin cleavage site changes. *J. Virol.* **64**, 3297–3303 (1990).

16. Shengqing, Y. *et al.* Generation of velogenic Newcastle disease viruses from a nonpathogenic waterfowl isolate by passaging in chickens. *Virology* **301**, 206–211 (2002).

17. Bagdonaite, I. & Wandall, H. H. Global aspects of viral glycosylation. *Glycobiology* **28**, 443–467 (2018).

18. Tran, E. E. H. *et al.* Spatial localization of the Ebola virus glycoprotein mucin-like domain determined by cryo-electron tomography. *J. Virol.* **88**, 10958–10962 (2014).

19. Steentoft, C. *et al.* Precision mapping of the human O-GalNAc glycoproteome through SimpleCell technology. *EMBO J.* **32**, 1478–1488 (2013).

20. Menachery, V. D. *et al.* A SARS-like cluster of circulating bat coronaviruses shows potential for human emergence. *Nat. Med.* **21**, 1508–1513 (2015).

21. Wong, M. C., Javornik Cregeen, S. J., Ajami, N. J. & Petrosino, J. F. Evidence of recombination in coronaviruses implicating pangolin origins of nCoV-2019. *bioRxiv* 2020.02.07.939207 (2020) doi:10.1101/2020.02.07.939207.

22. Lam, T. T.-Y. *et al.* Identification of 2019-nCoV related coronaviruses in Malayan pangolins in southern China. doi:10.1101/2020.02.13.945485.

February 26, 2020 4:12 PM

23. Phylodynamic Analysis | 90 genomes | 12 Feb 2020 – Novel 2019 coronavirus / nCoV-2019 Genomic Epidemiology. http://virological.org/t/phylodynamic-analysis-90-genomes-12-feb-2020/356 (2020).

24. Phylodynamic estimation of incidence and prevalence of novel coronavirus (nCoV) infections through time – Novel 2019 coronavirus / nCoV-2019 Genomic Epidemiology. http://virological.org/t/phylodynamic-estimation-of-incidence-and-prevalence-of-novel-coronavirus-ncov-infections-through-time/391 (2020).

25. Huang, C. *et al.* Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet* (2020) doi:10.1016/S0140-6736(20)30183-5.

26. Centers for Disease Control and Prevention (CDC). Prevalence of IgG antibody to SARS-associated coronavirus in animal traders--Guangdong Province, China, 2003. *MMWR Morb. Mortal. Wkly. Rep.* **52**, 986–987 (2003).

27. Wang, N. *et al.* Serological Evidence of Bat SARS-Related Coronavirus Infection in Humans, China. *Virol. Sin.* **33**, 104–107 (2018).

28. Ge, X.-Y. *et al.* Isolation and characterization of a bat SARS-like coronavirus that uses the ACE2 receptor. *Nature* **503**, 535–538 (2013).

29. 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).

30. 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).

31. 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).

32. Lim, P. L. *et al.* Laboratory-acquired severe acute respiratory syndrome. *N. Engl. J. Med.* **350**, 1740–1745 (2004).

February 26, 2020 4:12 PM

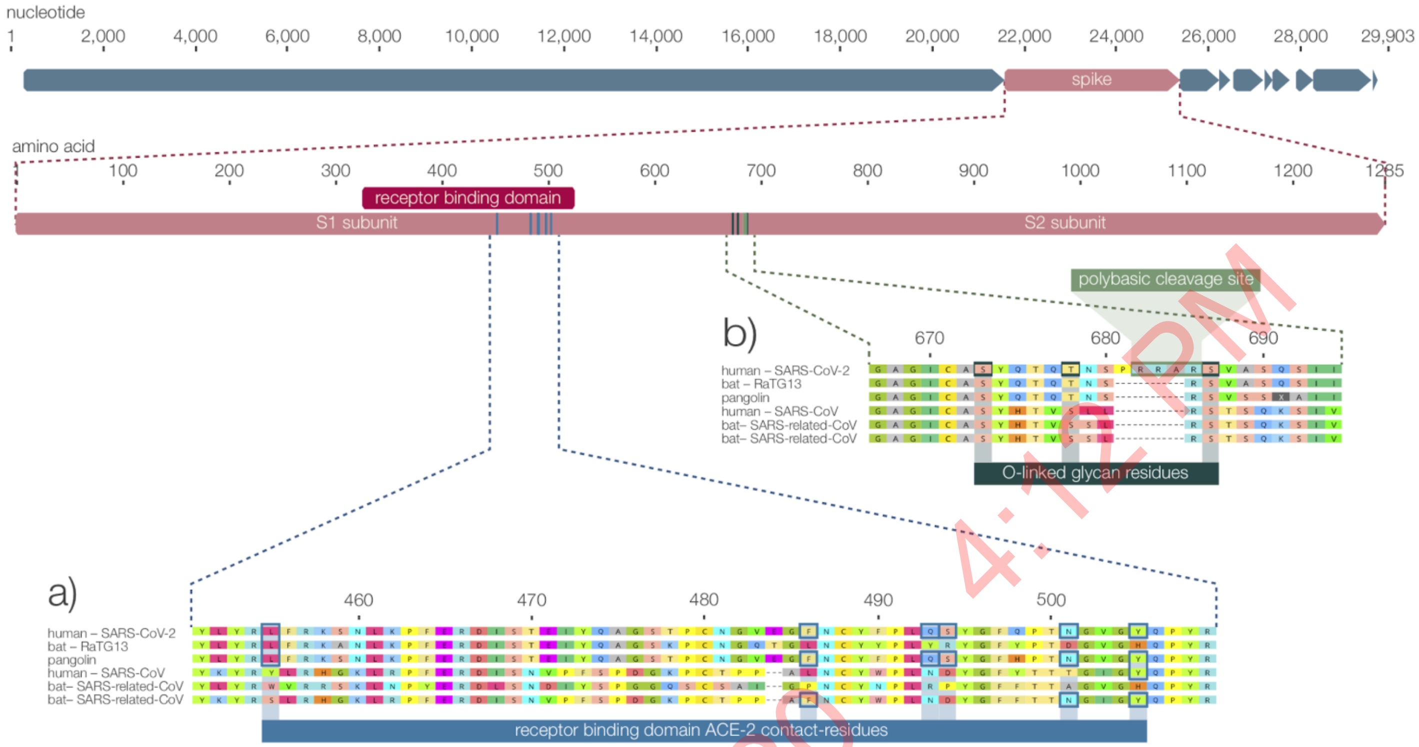
33. Senior, K. Recent Singapore SARS case a laboratory accident. *Lancet Infect. Dis.* **3**, 679 (2003).

34. Letko, M. *et al.* Adaptive Evolution of MERS-CoV to Species Variation in DPP4. *Cell Rep.* **24**, 1730–1737 (2018).

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

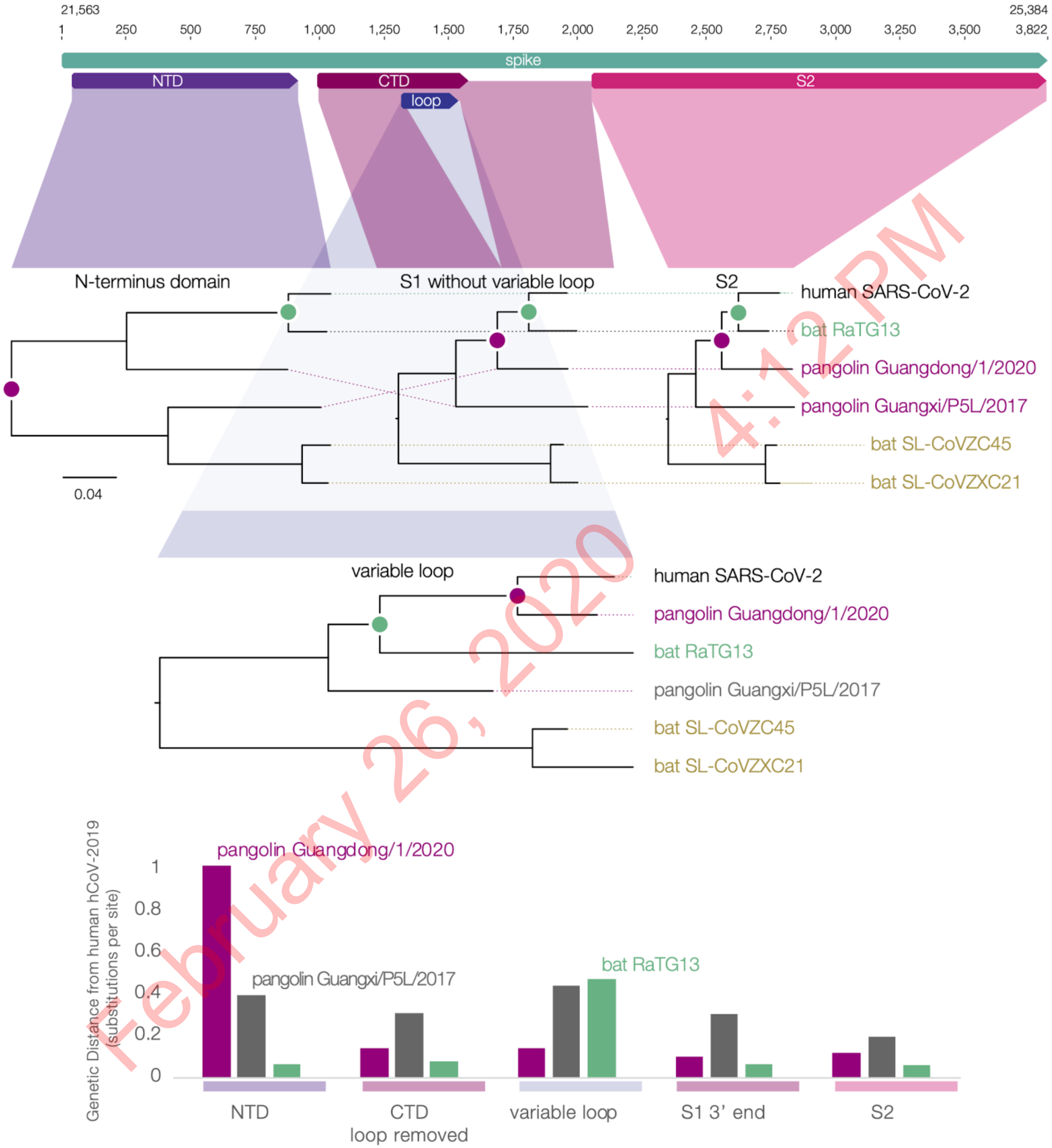
February 26, 2020 4:12 PM

Figure 1.


nucleotide
1
2,000
4,000
6,000
8,000
10,000
12,000
14,000
16,000
18,000
20,000
22,000
24,000
26,000
28,000
29,903
spike
amino acid
100
200
300
400
500
600
700
800
900
1000
1100
1200
1285
receptor binding domain
S1 subunit
S2 subunit
polybasic cleavage site
b)
670
680
690
human – SARS-CoV-2
bat – RaTG13
pangolin
human – SARS-CoV
bat– SARS-related-CoV
bat– SARS-related-CoV
G A G I C A S Y Q T Q T N S P R R A R S V A S Q S I I
G A G I C A S Y Q T Q T N S - - - - R S V A S Q S I I
G A G I C A S Y Q T Q T N S - - - - R S V S S X A I I
G A G I C A S Y H T V S L L - - - - R S T S Q K S I V
G A G I C A S Y H T V S S L - - - - R S T S Q K S I V
G A G I C A S Y H T V S S L - - - - R S T S Q K S I V
O-linked glycan residues
a)
460
470
480
490
500
human – SARS-CoV-2
bat – RaTG13
pangolin
human – SARS-CoV
bat– SARS-related-CoV
bat– SARS-related-CoV
Y L Y R L F R K S N L K P F E R D I S T E I Y Q A G S T P C N G V E G F N C Y F P L Q S Y G F Q P T N G V G Y Q P Y R
Y L Y R L F R K A N L K P F E R D I S T E I Y Q A G S K P C N G Q T G L N C Y Y P L Y R Y G F Y P T D G V G H Q P Y R
Y L Y R L F R K S N L K P F E R D I S T E I Y Q A G S T P C N G V E G F N C Y F P L Q S Y G F H P T N G V G Y Q P Y R
Y K Y R Y L R H G K L R P F E R D I S N V P F S P D G K P C T P P - - - A L N C Y W P L N D Y G F Y T T T G I G Y Q P Y R
Y L Y R W V R R S K L N P Y E R D L S N D I Y S P G G Q S C S A I - - - G P N C Y N P L R P Y G F F T T A G V G H Q P Y R
Y K Y R S L R H G K L R P F E R D I S N V P F S P D G K P C T P P - - - A F N C Y W P L N D Y G F F T T N G I G Y Q P Y R
receptor binding domain ACE-2 contact-residues

February 26, 2020 4:12 PM

Figure 2.



February 26, 2020 4:12 PM

The Proximal Origin of HCoV-19

Kristian G. Andersen[1,2*], Andrew Rambaut[3], W. Ian Lipkin[4], Edward C. Holmes[5] & Robert F. Garry[6,7]

[1]Department of Immunology and Microbiology, The Scripps Research Institute, La Jolla, CA 92037, USA.

[2]Scripps Research Translational Institute, La Jolla, CA 92037, USA.

[3]Institute of Evolutionary Biology, University of Edinburgh, Edinburgh, UK.

[4]Center for Infection and Immunity, Mailman School of Public Health of Columbia University, New York, New York, USA.

[5]Marie Bashir Institute for Infectious Diseases and Biosecurity, School of Life and Environmental Sciences and School of Medical Sciences, The University of Sydney, Sydney, Australia.

[6]Tulane University, School of Medicine, Department of Microbiology and Immunology, New Orleans, LA, USA.

[7]Zalgen Labs, LCC, Germantown, MD, USA.

*Corresponding author:

Kristian G. Andersen

Department of Immunology and Microbiology,

The Scripps Research Institute,

La Jolla, CA 92037, USA.

February 27, 2020 4:12 PM

Since the first reports of a novel pneumonia (COVID-19) in Wuhan city, Hubei province, China there has been considerable discussion and uncertainty over the origin of the causative virus, HCoV-19 (also referred to as SARS-CoV-2). Infections with HCoV-19 are now widespread across China, with cases in every province. As of 27 February 2020, 82,588 such cases across 50 countries have been confirmed, with 2,814 deaths attributed to the virus. These official case numbers are likely an underestimate because of limited reporting of mild and asymptomatic cases, and the virus is clearly capable of efficient human-to-human transmission. Based on the possibility of spread to countries with weaker healthcare systems, the World Health Organization has declared the COVID-19 outbreak a Public Health Emergency of International Concern (PHEIC). There are currently neither vaccines nor specific treatments for this disease.

HCoV-19 is the seventh member of the *Coronaviridae* known to infect humans. Three of these viruses, SARS CoV-1, MERS CoV, and HCoV-19 can cause severe disease; four, HKU1, NL63, OC43 and 229E, are associated with mild respiratory symptoms. Herein, we review what can be deduced about the origin and early evolution of HCoV-19 from the comparative analysis of available genome sequence data. In particular, we offer a perspective on the notable features in the HCoV-19 genome and discuss scenarios by which these features could have arisen. Importantly, this analysis provides evidence that HCoV-19 is not a laboratory construct nor a purposefully manipulated virus.

The genomic comparison of both alpha- and betacoronaviruses (family *Coronaviridae*) described below identifies two notable features of the HCoV-19 genome: (i) based on structural modeling and early biochemical experiments, HCoV-19 appears to be optimized for binding to the human ACE2 receptor; (ii) the highly variable spike (S) protein of HCoV-19 has a polybasic (furin) cleavage site at the S1 and S2 boundary via the insertion of twelve nucleotides. Additionally, this event led to the acquisition of three predicted O-linked glycans around the polybasic cleavage site.

Mutations in the receptor binding domain of HCoV-19

The sequences encoding the receptor binding domain (RBD) in the spike protein of SARS-CoV and SARS-related coronaviruses are the most variable part of the virus genome. Six amino acids in the RBD appear to be critical for binding to the human ACE2 receptor and determining host range[1]. Five of these six residues differ in the RBDs of HCoV-19 and SARS-CoV-1 (Figure 1a). Based on modeling[1] and biochemical experiments[3,4], HCoV-19 seems to have an RBD that may bind with high affinity to ACE2 from human, non-human primate, ferret, pig, and cat, as well as other species with high receptor homology[1]. In contrast, HCoV-19 may bind less efficiently to ACE2 in other species associated with SARS-like viruses, including rodents and civets[1]. While these analyses suggest that HCoV-19 may be capable of binding the human ACE2 receptor with high affinity, computational analyses predict that the interaction is not optimal[1]. The phenylalanine (F) at residue 486 in the HCoV-19 S protein corresponds to L472 in the SARS-CoV Urbani strain. Notably, in SARS-CoV cell culture passage experiments the L472 mutates to phenylalanine (L472F)[5], which is predicted to be optimal for binding of the SARS-CoV RBD to the human ACE2 receptor[6]. Additionally, several of the key residues in the RBD of HCoV-19 are different from those previously described as optimal for human ACE2 receptor binding[6]. In contrast to these computational predictions, recent binding studies indicate that HCoV-19 binds with high affinity to human ACE2[7]. Thus, the HCoV-19 spike appears to be the result of selection on a human or human-like ACE2 permitting another optimal binding solution to arise. This is strong evidence that HCoV-19 is *not* the product of genetic engineering, a widely propagated conspiracy theory PMID: 32087122 .

Polybasic cleavage site and O-linked glycans

The second notable feature of HCoV-19 is a predicted polybasic cleavage site (RRAR) in the spike protein at the junction of S1 and S2, the two subunits of the spike protein (Figure 1b)[8,9]. In addition to two basic arginines and an alanine at the cleavage site, a leading proline is also inserted; thus, the fully inserted sequence is PRRA (Figure 1b). The strong turn created by the proline insertion is predicted to result in the addition of O-linked glycans to S673, T678, and S686 that flank the polybasic cleavage site. A polybasic cleavage site has not previously been observed in related "lineage B" betacoronaviruses and is a unique feature of HCoV-19. Some human betacoronaviruses, including HCoV-HKU1 (lineage A), have polybasic cleavage sites, as well as predicted O-linked glycans near the S1/S2 cleavage site.

The functional consequence of the predicted polybasic cleavage site in HCoV-19 is unknown. It will be important to verify that this cleavage site is utilized and to determine what impact, if any, the site has on transmissibility and pathogenesis in yet to be established animal models. Experiments with SARS-CoV have shown that engineering such a site at the S1/S2 junction enhances cell–cell fusion but does not affect virus entry[10]. Polybasic cleavage sites allow effective cleavage by furin and other proteases, and can be acquired at the junction of the two subunits of the haemagglutinin (HA) protein of avian influenza viruses in conditions that select for rapid virus replication and transmission (e.g., highly dense chicken populations). HA serves a similar function in cell-cell fusion and viral entry as the coronavirus S protein. Acquisition of a polybasic cleavage site in HA, by either insertion or recombination, converts low pathogenicity avian

February 27, 2020 4:12 PM

influenza viruses into highly pathogenic forms[11–13]. The acquisition of polybasic cleavage sites by the influenza virus HA has also been observed after repeated forced passage in cell culture or through animals[14,15]. Similarly, an avirulent isolate of Newcastle Disease virus became highly pathogenic during serial passage in chickens by incremental acquisition of a polybasic cleavage site at the junction of its fusion protein subunits[16].

The potential function of the three predicted O-linked glycans is less clear, but they could create a "mucin-like domain" that would shield potential epitopes or key residues on the HCoV-19 spike protein[17,18]. Several viruses employ mucin-like domains as part of a glycan shield that is involved in immune evasion (PMID: 29579213) Although the algorithms for prediction of O-linked glycosylation are robust[19], biochemical analyses or structural studies are required to determine whether or not the predicted O-linked glycan sites are utilized.

Theories of HCoV-19 origins

It is improbable that HCoV-19 emerged through laboratory manipulation of an existing SARS-related coronavirus. As noted above, the RBD of HCoV-19 is optimized for human ACE2 receptor binding with an efficient binding solution different to that which would have been predicted. Further, if genetic manipulation had been performed, one would expect that one of the several reverse genetic systems available for betacoronaviruses would have been used. However, this is not the case as the genetic data shows that HCoV-19 is not derived from any previously used virus backbone[20]. Instead, we propose two scenarios that can plausibly explain the origin of HCoV-19: (i) natural selection in a non-human animal host prior to zoonotic transfer, and (ii) natural selection in humans following zoonotic transfer. We also discuss whether selection during passage in culture could have given rise to the same observed features.

Selection in an animal host

As many of the early cases of COVID-19 were linked to the Huanan seafood and wildlife market in Wuhan, it is possible that an animal source was present at this location. Given the similarity of HCoV-19 to bat SARS-like CoVs, particularly RaTG13, it is plausible that bats serve as reservoir hosts for its progenitor. Although RaTG13 sampled from a *Rhinolophus affinis* bats is ~96% identical overall to HCoV-192, its S protein possesses two deletions and distinct sequences in the RBD suggesting that it may use an entirely different cellular receptor. It is also important to note that previous outbreaks of betacoronaviruses in humans involved direct exposure to animals other than bats, including civets (SARS) and camels (MERS), that carry viruses that are genetically very similar to SARS-CoV or MERS-CoV, respectively. By analogy, viruses closely related to HCoV-19 may be circulating in one or more animal species. Malayan pangolins (*Manis javanica*) illegally imported into Guangdong province contain CoVs that are similar to HCoV-19[1,22]. Although the bat virus RaTG13 remains the closest relative to HCoV-19 across the whole genome, a Malayan pangolin CoV exhibits strong similarity to HCoV-19 in the RBD, including all six key RBD residues (Figure 1). However, no CoV from a pangolin, nor any other animal species, has yet been identified that is sufficiently similar to HCoV-19 across its entire genome to support direct human infection. Similarly, although there is likely to be a history of complex recombination events in these viruses, including in the RBD and other domains of the S protein, none of the available bat or pangolin betacoronaviruses are sufficiently similar to HCoV-19 to have directly generated this virus by recombination.

Neither the bat nor pangolin betacoronaviruses sampled to date carry polybasic cleavage sites. However, as the diversity of CoVs in bats and other species is hugely undersampled, it is possible that an animal betacoronavirus will eventually be identified with a polybasic cleavage site. Mutations, including point mutations, insertions and deletions can occur near the S1/S2 junction of CoVs [MHV FECov ref PMID: 23763835 PMID: 19553314 PMID: 9782269] suggesting that the polybasic site could arise by a natural evolutionary process. For a precursor virus to acquire both the polybasic cleavage site and mutations in the spike protein suitable

February 27, 2020 4:12 PM

for human ACE2 receptor binding, an animal host would likely have to have a high population density – to allow natural selection to proceed efficiently – and an ACE2 gene that is similar to the human orthologue. Further characterization of CoVs in animals that may harbour SARS-CoV-like viruses should be a public health priority.

Cryptic adaptation to humans

It is also possible that a progenitor to HCoV-19 jumped from a non-human animal to humans, with the genomic features described above acquired through adaptation during subsequent human-to-human transmission. We surmise that once these adaptations were acquired (either together or in series) it would enable the outbreak to take-off, producing a sufficiently large and unusual cluster of pneumonia cases to trigger the surveillance system that ultimately detected it.

All HCoV-19 genomes sequenced so far have the well adapted RBD and the polybasic cleavage site and are thus derived from a common ancestor that had these features. The presence of an RBD in pangolins that is very similar to the one in HCoV-19 means that this was likely already present in the virus that jumped to humans, even if we don't yet have the exact non-human progenitor virus. This leaves the polybasic cleavage site insertion to occur during human-to-human transmission. Following the example of the influenza A virus HA gene, a specific insertion or recombination event is required to enable the emergence of HCoV-19 as an epidemic pathogen.

Estimates of the timing of the most recent common ancestor (tMRCA) of HCoV-19 using currently available genome sequence data point to virus emergence in late November to early December 2019[23,24], compatible with the earliest retrospectively confirmed cases[25]. Hence, this scenario presumes a period of unrecognised transmission in humans between the initial zoonotic transfer event and the acquisition of the polybasic cleavage site. Sufficient opportunity could occur if there had been many prior zoonotic events producing short chains of human-to-human transmission (so-called 'stuttering chains') over an extended period. This is essentially the situation for MERS-CoV in the Arabian Peninsula where all the human cases are the result of repeated jumps of the virus from dromedary camels, producing single infections or short chains of transmission that eventually resolve. To date, after 2,499 cases over 8 years, no human adaptation has emerged that has allowed MERS-CoV to take hold in the human population.

How could we test whether cryptic spread of HCoV-19 enabled human adaptation? Metagenomic studies of banked serum samples could provide important information, but given the relatively short period of viraemia it may be impossible to detect low level HCoV-19 circulation in historical samples. Retrospective serological studies potentially could be informative and a few such studies have already been conducted. One found that animal importation traders had a 13% seropositivity to coronaviruses[26], while another noted that 3% residents of a village in Southern China were seropositive to these viruses[27]. Interestingly, 200 residents of Wuhan did not show coronavirus seroreactivity. Critically, however, these studies could not have distinguished whether positive serological responses were due to a prior infection with SARS-CoV or HCoV-19. Further retrospective serological studies should be conducted to determine the extent of prior human exposure to betacoronaviruses in different geographic areas, particularly using assays that can distinguish among multiple betacoronaviruses.

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 in multiple laboratories across the world[28–31]. There are also documented instances of the acquisition of SARS-CoV by laboratory personnel working under BSL-2 or BSL-3 containment[32,33] (PMID: 16830004). While reverse engineering is not a trivial task, methods to add insertions, deletions or otherwise modify large CoV genomes are well established [refs]. We must therefore consider the possibility of a deliberate or inadvertent release of HCoV-19. In theory, it is possible that HCoV-

February 27, 2020 4:12 PM

19 acquired the observed RBD mutations site during adaptation to passage in cell culture, as has been observed in studies with SARS-CoV[5] as well as MERS-CoV[34]. However, the acquisition of the polybasic cleavage site or O-linked glycans - if functional - argues against this scenario. New polybasic cleavage sites have only been observed after prolonged passaging of low pathogenicity avian influenza virus in cell culture or animals. Furthermore, the generation of HCoV-19 by cell culture or animal passage would have required prior isolation of a progenitor virus with a very high genetic similarity. Subsequent generation of a polybasic cleavage site would have then required an intense program of passage in cell culture or animals with ACE-2 receptor similar to humans (e.g. ferrets). It is also questionable whether generation of the O-linked glycans would have occurred on cell culture passage, as such mutations typically suggest the involvement of an immune system that is not present *in vitro*.

Conclusions

In the midst of the global COVID-19 public health emergency it is reasonable to wonder why the origins of the epidemic matter. A detailed understanding of how an animal virus jumped species boundaries to infect humans so productively will help in the prevention of future zoonotic events. For example, if HCoV-19 pre-adapted in another animal species then we are at risk of future re-emergence events even if the current epidemic is controlled. In contrast, if the adaptive process we describe occurred in humans, then even if we have repeated zoonotic transfers they are unlikely to take-off unless the same series of mutations occurs. In addition, identifying the closest animal relatives of HCoV-19 will greatly assist studies of virus function. Indeed, the availability of the RaTG13 bat sequence facilitated the comparative genomic analysis performed here, helping to reveal the key mutations in the RBD as well as the polybasic cleavage site insertion.

The genomic features described here may in part explain the infectiousness and transmissibility of HCoV-19 in humans. Although genomic evidence does not support the idea that HCoV-19 is a laboratory construct, it is currently impossible to prove or disprove the other theories of its origin described here. While it is unclear whether future data will help resolve this issue, it is likely that more scientific data will swing the balance in favor of one hypothesis over another. Identifying the immediate non-human animal source and obtaining virus sequences from it would be the most definitive way of revealing virus origins. In addition, it would be helpful to obtain more genetic and functional data about the virus, including experimental studies of receptor binding and the role of the polybasic cleavage site and predicted O-linked glycans. The identification of a potential intermediate host of HCoV-19, as well as the sequencing of very early cases including those not connected to the Wuhan market, would similarly be highly informative. Irrespective of how HCoV-19 originated, the ongoing surveillance of pneumonia in humans and other animals is clearly of utmost importance.

Acknowledgements

We thank all those who have contributed HCoV-19 genome sequences to the GISAID database (https://www.gisaid.org/) and contributed analyses and ideas to Virological.org (http://virological.org/). We thank M. Farzan for discussions. We thank the Wellcome Trust for supporting this work. ECH is supported by an ARC Australian Laureate Fellowship (FL170100022). KGA is a Pew Biomedical Scholar and is supported by NIH NIAID grant U19AI135995. AR is supported by the Wellcome Trust (Collaborators Award 206298/Z/17/Z – ARTIC network) and the European Research Council (grant agreement no. 725422 – ReservoirDOCS). RFG is supported by NIH grants U19AI135995, U54 HG007480 and U19AI142790.

Figure Legends

Figure 1. (a) Mutations in contact residues of the HCoV-19 spike protein. The spike protein of HCoV-19 (top) was aligned against the most closely related SARS-like CoVs and SARS-CoV-1. Key residues in the spike protein that make contact to the ACE2 receptor are marked with blue boxes in both HCoV-19 and the SARS-

February 27, 2020 4:12 PM

CoV Urbani strain. (b) Acquisition of polybasic cleavage site and O-linked glycans. Both the polybasic cleavage site and the three adjacent predicted O-linked glycans are unique to HCoV-19 and not previously seen in lineage B betacoronaviruses. Sequences shown are from NCBI GenBank, accession numbers MN908947, MN996532, AY278741, KY417146 and MK211376. The pangolin coronavirus sequences are a consensus generated from SRR10168377 and SRR10168378 (NCBI BioProject PRJNA573298)[21,35].

February 27, 2020 4:12 PM

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. Wu, F. *et al.* A new coronavirus associated with human respiratory disease in China. *Nature* (2020) doi:10.1038/s41586-020-2008-3.

3. 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.

4. 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.

5. 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).

6. Cui, J., Li, F. & Shi, Z.-L. Origin and evolution of pathogenic coronaviruses. *Nat. Rev. Microbiol.* **17**, 181–192 (2019).

7. Wrapp, D. *et al.* Cryo-EM Structure of the 2019-nCoV Spike in the Prefusion Conformation. *bioRxiv* 2020.02.11.944462 (2020) doi:10.1101/2020.02.11.944462.

8. Gallaher, W. Analysis of Wuhan Coronavirus: Deja Vu. *Virological* http://virological.org/t/analysis-of-wuhan-coronavirus-deja-vu/357 (2020).

9. Coutard, B. *et al.* The spike glycoprotein of the new coronavirus 2019-nCoV contains a furin-like cleavage site absent in CoV of the same clade. *Antiviral Res.* 104742 (2020).

10. 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).

February 27, 2020 4:12 PM

11. 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).

12. Alexander, D. J. & Brown, I. H. History of highly pathogenic avian influenza. *Rev. Sci. Tech.* **28**, 19–38 (2009).

13. 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).

14. Ito, T. *et al.* Generation of a highly pathogenic avian influenza A virus from an avirulent field isolate by passaging in chickens. *J. Virol.* **75**, 4439–4443 (2001).

15. Li, S. Q., Orlich, M. & Rott, R. Generation of seal influenza virus variants pathogenic for chickens, because of hemagglutinin cleavage site changes. *J. Virol.* **64**, 3297–3303 (1990).

16. Shengqing, Y. *et al.* Generation of velogenic Newcastle disease viruses from a nonpathogenic waterfowl isolate by passaging in chickens. *Virology* **301**, 206–211 (2002).

17. Bagdonaite, I. & Wandall, H. H. Global aspects of viral glycosylation. *Glycobiology* **28**, 443–467 (2018).

18. Tran, E. E. H. *et al.* Spatial localization of the Ebola virus glycoprotein mucin-like domain determined by cryo-electron tomography. *J. Virol.* **88**, 10958–10962 (2014).

19. Steentoft, C. *et al.* Precision mapping of the human O-GalNAc glycoproteome through SimpleCell technology. *EMBO J.* **32**, 1478–1488 (2013).

20. Menachery, V. D. *et al.* A SARS-like cluster of circulating bat coronaviruses shows potential for human emergence. *Nat. Med.* **21**, 1508–1513 (2015).

21. Wong, M. C., Javornik Cregeen, S. J., Ajami, N. J. & Petrosino, J. F. Evidence of recombination in coronaviruses implicating pangolin origins of nCoV-2019. *bioRxiv* 2020.02.07.939207 (2020) doi:10.1101/2020.02.07.939207.

22. Lam, T. T.-Y. *et al.* Identification of 2019-nCoV related coronaviruses in Malayan pangolins in southern China. doi:10.1101/2020.02.13.945485.

February 27, 2020 4:12 PM

23. Phylodynamic Analysis | 90 genomes | 12 Feb 2020 – Novel 2019 coronavirus / nCoV-2019 Genomic Epidemiology. http://virological.org/t/phylodynamic-analysis-90-genomes-12-feb-2020/356 (2020).

24. Phylodynamic estimation of incidence and prevalence of novel coronavirus (nCoV) infections through time – Novel 2019 coronavirus / nCoV-2019 Genomic Epidemiology. http://virological.org/t/phylodynamic-estimation-of-incidence-and-prevalence-of-novel-coronavirus-ncov-infections-through-time/391 (2020).

25. Huang, C. *et al.* Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet* (2020) doi:10.1016/S0140-6736(20)30183-5.

26. Centers for Disease Control and Prevention (CDC). Prevalence of IgG antibody to SARS-associated coronavirus in animal traders--Guangdong Province, China, 2003. *MMWR Morb. Mortal. Wkly. Rep.* **52**, 986–987 (2003).

27. Wang, N. *et al.* Serological Evidence of Bat SARS-Related Coronavirus Infection in Humans, China. *Virol. Sin.* **33**, 104–107 (2018).

28. Ge, X.-Y. *et al.* Isolation and characterization of a bat SARS-like coronavirus that uses the ACE2 receptor. *Nature* **503**, 535–538 (2013).

29. 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).

30. 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).

31. 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).

32. Lim, P. L. *et al.* Laboratory-acquired severe acute respiratory syndrome. *N. Engl. J. Med.* **350**, 1740–1745 (2004).

February 27, 2020 4:12 PM

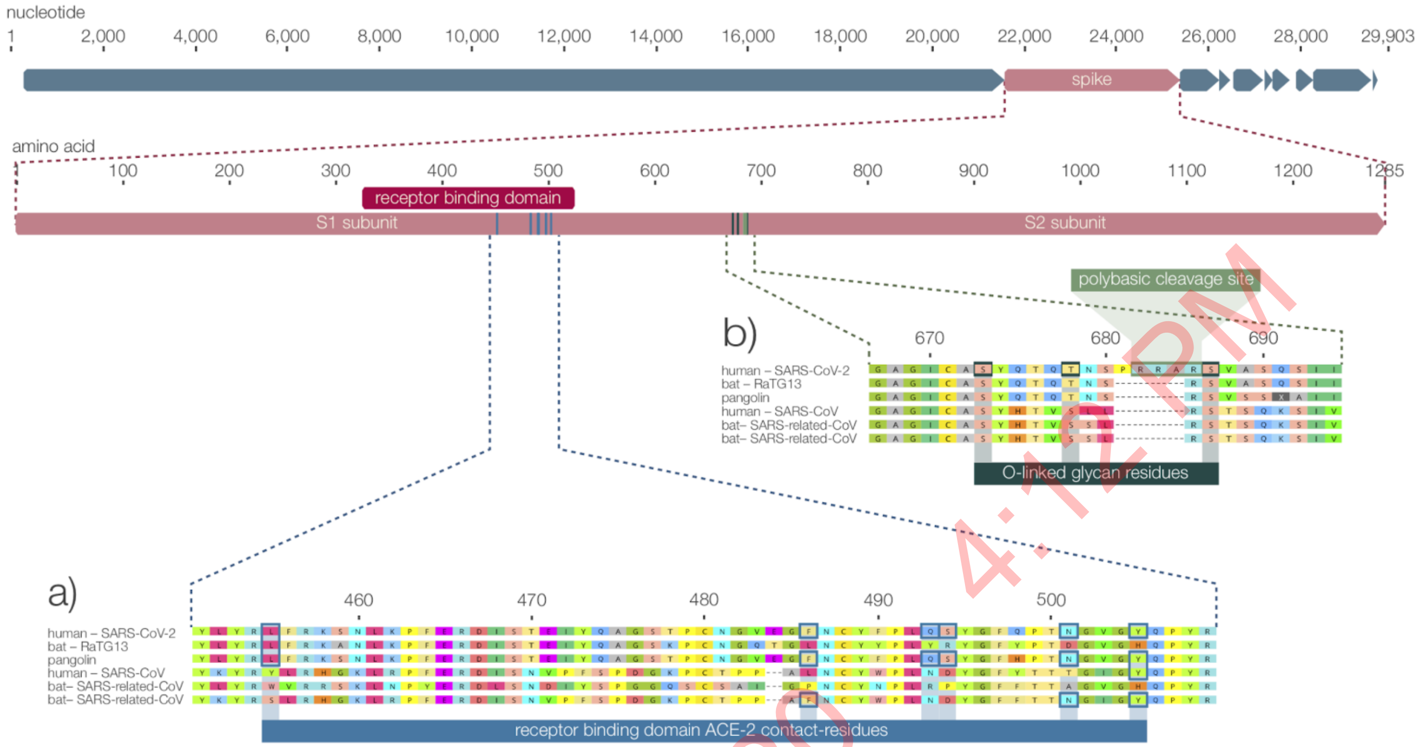
33. Senior, K. Recent Singapore SARS case a laboratory accident. *Lancet Infect. Dis.* **3**, 679 (2003).

34. Letko, M. *et al.* Adaptive Evolution of MERS-CoV to Species Variation in DPP4. *Cell Rep.* **24**, 1730–1737 (2018).

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

February 27, 2020 4:12 PM

Figure 1.



February 27, 2020 4:12 PM

Referee #1:

Anderson presented a timely manuscript to share their points of view about the origin of SARS-CoV-2. There are several rumors about the origin of this virus. However, these “hypotheses” are entirely based on very limited, if any, scientific evidences. This reviewer sees most of the arguments raised by the authors are valid and convincing. However, the authors might want to consider these minor suggestions:

1. The sections for the RBD and cleavage site of Spike protein basically have summarized the existing findings from other recent publications. The authors might want to spell out that these two sections are review summaries. In addition, the author can present these two sections in a more condense format and save some space for something else (also see points 6 and 7 below)

We have edited these sections to be more concise. However, we think that it is important that the key genomic features are described in detail as they form the basis of the discussion that follows.

2. Fig. 1. This figure has 6 aligned sequences, but with only 5 sequence titles. The order of these titles are also not correct.

We have corrected these mistakes and updated the figure and titles.

3. Lines 170 -174. It is correct that no adaptive mutation has been found in the spike of MERS-CoV. Deletions in other ORF regions, however, were detected in some human MERS-CoV viruses (PMID: 26981770). In addition, the 29nt deletion of human SARS-CoV (PMID: 12958366) was suggested to have effects on host adaptation. The authors should also consider these findings. It is premature to say that this would not happen in SARS-CoV-2.

We have revised the text to incorporate this information as we agree that they are important for context.

4. Line 194. The accident at Singapore occurred in a BSL3, not BSL2, containment.

We have corrected this.

5. Line 194. Laboratory escapes of SARS occurred in Singapore, China and Taiwan (PMID: 16830004).

We have added this reference as well as the point about laboratory escapes of SARS-CoV in Singapore, China and Taiwan.

6. There are two recent reports about coronaviruses in pangolins. The authors might want to comment on these.

We have included these references as well as several others that have investigated pangolin CoV. In addition, and in response to Reviewer 2 (see below) we should point out that these additional pangolin CoV sequences do not further clarify the different scenarios discussed in our manuscript. There is nothing in these reports that changes our statements regarding a potential role of pangolins.

7. Optional: Can the authors share their views on the possibility of having a lab escape of a natural coronavirus? This is also one of the hypotheses that have been extensively discussed. The reviewer understand that this is entirely a different topic, but any insights are welcomed.

Nature Rebuttal February 28, 2020 6:46PM

Escape of a natural CoV (SARS-CoV-2 or a close progenitor) from a lab could not be distinguished from an animal-to-human transfer in another environment. Given the limited numbers of labs doing work of this kind compared to the frequent opportunities for animal-to-human transfer, we believe the latter is much less likely than the former.

Nature Rebuttal February 28, 2020 6:46PM

Referee #2:

This is a perspective discussing evidence against a hypothetical lab origin of SARS-CoV-2. The paper addresses suboptimal composition of ACE2-binding sites in the RBD, 3 predicted O-linked glycosylation sites and a furin cleavage site in the glycoprotein that was speculated upon before.

The paper is itself interesting, but unnecessarily speculative. It's not clear why the authors do not refute a hypothetical lab origin in their coming publication on the ancestors of SARS-CoV-2 in bats and pangolins. The tree showing diverse pangolin viruses has kindly been made available by some of the authors in GISAID. Once the authors publish their new pangolin sequences, a lab origin will be extremely unlikely. It is not clear why the authors rush with a speculative perspective if their central hypothesis can be supported by their own data. Please explain.

Our manuscript is written to explore the potential origin of SARS-CoV-2. We do not believe it is speculative: rather, it simply takes the available data and proposes a series of plausible hypotheses for how these data may have arisen. We do this in a logical, balanced and unbiased manner; this is critically important because it gives our work credibility. It is science, not opinion.

Unfortunately, the newly available pangolin sequences do not elucidate the origin of SARS-CoV-2 or refute a lab origin. Hence, the reviewer is incorrect on this point. To clarify, while the RBD from the Guangdong pangolin CoVs is the closest to that found SARS-CoV-2, they are more divergent in the remainder of the viral genome (for which the bat virus RaTG13 is still the closest) and do not possess the polybasic cleavage site insertion. Hence, there is no evidence on present data that the pangolin CoVs are directly related to the COVID-19 epidemic.

Another critical aspect of this text is the complete lack of referencing to a potential debate on a hypothetical lab origin. Who said this, why is this considered a problem? There are indeed a few apparently uninformed statements claiming the virus may be a Chinese bioweapon, but is this really problematic on a larger scale? The central reason for issuing this text must be exhaustively referenced and discussed.

The possibility that SARS-CoV-2 originated as an engineered bioweapon has had widespread discussion in the press and on social media (particularly in China) and is hence problematic on a very large scale. This particular topic was also the reason for a recent request from the White House: https://bit.ly/2HMndCi. A group of public health scientists recently wrote a letter to *The Lancet* in which they *"strongly condemn conspiracy theories suggesting that COVID-19 does not have a natural origin."* We now reference this publication (the publication itself also references our pre-print of this manuscript).

It appears that the reviewer is considering bioweapons, engineering, and lab accidents as one and the same theory lumped in under a "lab origin". This is inappropriate and will likely only increase confusion. We do briefly discuss the notion of deliberate engineering of the virus and explain how the evidence is strongly against this explanation. Importantly, we do not discuss the concept of 'bioweapons' per se as there are many, legitimate, research uses of bio-engineering of viruses.

As we very clearly state in the manuscript, there are far more subtle scenarios that need to be considered carefully and scientifically: for example, of accidental infections in a lab while culturing SARS-like CoVs. Since accidental infections and other lab 'escapes' happen frequently across the world (and as we mention, happened multiple times with SARS-CoV following the SARS epidemic) we firmly believe that this discussion

Nature Rebuttal February 28, 2020 6:46PM

is of major importance and must be had. In particular, the culturing of SARS-like CoVs from animals is typically performed under BSL-2 and has been ongoing for years.

Dismissing this potential accidental scenario out of hand - or considering it in the same category as conspiracy theories about bioweapons and deliberate engineering (that, as we outline, are clearly wrong) - would be irresponsible. Our manuscript will only serve its purpose if it considers all possible hypotheses equally. Any perceived bias will undermine its credibility.

We have modified the text to more clearly state the scenarios we are considering.

The authors state that a predicted polybasic cleavage sites is unique to SARS-CoV-2 in SARS viruses. Who knows how many out of thousands undiscovered bat ancestors also acquired such a motif, the sampling bias in descriptions of remote bat viruses is dramatic. This should be discussed. Also state clearly that this site is only predicted so far and that experimental evidence for its biological function and its potential impact on pathogenesis are required.

We agree that the diversity in bat CoVs is highly undersampled and that it is possible that a bat CoV progenitor could be discovered that contains a polybasic site. We say nothing against this in the manuscript. However, given the diversity of SARS-like CoVs already sampled from bats, pangolins, and many other animals - none of which possess this insertion - it is obviously reasonable to hypothesize that it may have been gained in the lineage leading to SARS-CoV-2. This is explicitly discussed in our revision.

While other CoVs utilize polybasic sites, we did refer to the site as a predicted cleavage site. Studies have since been performed showing that this site is indeed functional, which we now reference (Walls et al., DOI: 10.1016/j.cell.2020.02.058). We also state that it will be necessary to test the effect of the polybasic site on pathogenesis, which will require the establishment of an animal model.

The predicted O-linked glycosylation sites are mysterious. What do the authors imply with those sites? *In silico* prediction of O-linked glycosylation sites is not robust and whether these sites indeed exist requires experimental validation. Even if those sites exist, why are they relevant? This is not addressed at all. If the authors assume these sites constitute part of a glycan shield, they should say so and weigh their assumption carefully.

Although not previously described for CoV proteins, numerous other viral proteins have mucin-like domains that are involved in immune evasion. The revised text adds relevant references and is more explicit about the potential relevance of the predicted O-linked glycan sites. As stated previously, we consider that these sites - if indeed utilized - may be part of the glycan shield and have made this point directly in the revised text. We also agree that the predicted O-linked glycosylation sites require experimental validation not only in SARS-CoV-2 but in other CoVs that have similar predicted sites. This was also already stated explicitly, but in our revision we have stated this even more directly.

Finally, the main argument against a hypothetical lab origin seems the required reconstruction of a backbone of a bat virus of unknown pathogenesis. It does not seem feasible that any scientist would disembark on such an uncertain endeavor. This difficulties of coronavirus reverse genetics should be stated clearly.

The reviewer’s restatement of our argument is correct. We reiterate that the purpose of our manuscript was not to refute the conspiracy theory that SARS-CoV-2 was bioengineered. Rather, it was carefully designed to be a balanced and unbiased assessment of the available data. The statement that no scientist would embark on such an endeavor is a subjective one with no supporting evidence - in fact, scientists

Nature Rebuttal February 28, 2020 6:46PM

have already created a reverse genetics system for SARS-CoV-2, which was completed in two weeks (Thao et al., bioRxiv 10.1101/2020.02.21.959817).

Nature Rebuttal February 28, 2020 6:46PM

Editor’s comments:

While the Perspective is interesting and timely one of our referees raised concerns (also emphasised to the editors) about whether such a piece would feed or quash the conspiracy theories.

Critically, the purpose of our manuscript was not to quash conspiracy theories. Rather, our aim was to carefully examine in a balanced and unbiased manner the evidence for and against a number of possible probable scenarios for the proximal origins of SARS-CoV-2.

But more importantly this reviewer feels, and we agree, that the Perspective would quickly become outdated when more scientific data are published (for example on potential reservoir hosts).

Of course, it is likely that more scientific data will swing the balance in favor of one hypothesis over another. However, the same can be said of many of the papers published on COVID-19: as we learn more about the virus and the disease so previous publications may be quickly revised. We contend, however, our piece will remain relevant even if more data becomes available because these data would potentially confirm which of these scenarios is correct. Most importantly, our manuscript sets out what evidence is needed to test the hypotheses outlined and will therefore serve as an important starting point for guiding future research. In addition, we make it clear that there is a very real possibility that no definitive intermediate host will ever be found.

Nature Rebuttal February 28, 2020 6:46PM

The Proximal Origin of HCoV-19

Kristian G. Andersen[1,2*], Andrew Rambaut[3], W. Ian Lipkin[4], Edward C. Holmes[5] & Robert F. Garry[6,7]

[1]Department of Immunology and Microbiology, The Scripps Research Institute, La Jolla, CA 92037, USA.

[2]Scripps Research Translational Institute, La Jolla, CA 92037, USA.

[3]Institute of Evolutionary Biology, University of Edinburgh, Edinburgh, UK.

[4]Center for Infection and Immunity, Mailman School of Public Health of Columbia University, New York, New York, USA.

[5]Marie Bashir Institute for Infectious Diseases and Biosecurity, School of Life and Environmental Sciences and School of Medical Sciences, The University of Sydney, Sydney, Australia.

[6]Tulane University, School of Medicine, Department of Microbiology and Immunology, New Orleans, LA, USA.

[7]Zalgen Labs, LCC, Germantown, MD, USA.

*Corresponding author:

Kristian G. Andersen

Department of Immunology and Microbiology,

The Scripps Research Institute,

La Jolla, CA 92037, USA.

February 28, 2020 8:41 PM

Since the first reports of a novel pneumonia (coronavirus disease 2019; COVID-19) in Wuhan city, Hubei province, China(Zhou et al. 2020; Wu et al. 2020) there has been considerable discussion and uncertainty over the origin of the causative virus, human coronavirus 2019 (HCoV-19(Jiang et al. 2020); also referred to as SARS-CoV-2(Gorbalenya 2020)). Infections with HCoV-19 are now widespread across the world and as of 28 February 2020, 85,176 COVID-19 cases have been confirmed in 57 countries, with 2,919 deaths attributed to the virus(Dong et al. 2020). These official numbers likely represent an underestimate because of limited reporting of mild and asymptomatic cases, and the virus is clearly capable of efficient human-to-human transmission. Based on the possibility of spread to countries with weaker healthcare systems, the World Health Organization has declared the COVID-19 epidemic a Public Health Emergency of International Concern (PHEIC)(Statement on the second meeting of th...). There are currently neither vaccines nor specific treatments for this disease.

HCoV-19 is the seventh member of the *Coronaviridae* known to infect humans. Three of these viruses, SARS-CoV, MERS-CoV, and HCoV-19 can cause severe disease; four, HKU1, NL63, OC43 and 229E, are associated with mild respiratory symptoms(Corman et al. 2018). Herein, we review what can be deduced about the origin and early evolution of HCoV-19 from the comparative analysis of available genome sequence data.

In particular, we offer a perspective on the notable features in the HCoV-19 genome and discuss scenarios by which these features could have arisen. Importantly, our analysis provides strong evidence that HCoV-19 is not a laboratory construct nor a purposefully manipulated virus.

Notable features of the HCoV-19 genome

Our genomic comparison of both alpha- and betacoronaviruses (family *Coronaviridae*) described below identifies two notable features of the HCoV-19 genome: (*i*) based on structural studies(Wan et al. 2020; Wrapp et al. 2020; Walls et al. 2020) and early biochemical experiments(Zhou et al. 2020; Letko et al. 2020; Wrapp et al. 2020; Hoffmann et al. 2020), HCoV-19 appears to be optimized for binding to the human ACE2 receptor; (*ii*) the highly variable spike (S) protein of HCoV-19 has a functional polybasic (furin) cleavage site at the S1 and S2 boundary via the insertion of twelve nucleotides(Gallaher 2020; Coutard et al. 2020; Walls et al. 2020). Additionally, this event led to the acquisition of three predicted O-linked glycans around the polybasic cleavage site.

Mutations in the receptor binding domain of HCoV-19

The the receptor binding domain (RBD) in the spike protein of HCoV-19 and SARS-related coronaviruses is the most variable part of the virus genome(Zhou et al. 2020; Wu et al. 2020). Six amino acids in the RBD have been shown to be critical for binding to the human ACE2 receptor and determining the host range of SARS-like viruses(Wan et al. 2020). Using coordinates based on the Urbani strain of SARS-CoV, they are Y442, L472, N479, D480, T487, and Y4911. The corresponding residues in HCoV-19 are L455, F486, Q493, S494, N501, and Y505(Wan et al. 2020). Five of these six residues differ in the RBDs of HCoV-19 and SARS-CoV (**Fig. 1a**). Based on structural studies(Wan et al. 2020; Wrapp et al. 2020; Walls et al. 2020) and biochemical experiments(Zhou et al. 2020; Letko et al. 2020; Wrapp et al. 2020; Hoffmann et al. 2020), HCoV-19 seems to have an RBD that may bind with high affinity to ACE2 from human, non-human primate, ferret, pig, and cat, as well as other species with high receptor homology(Wan et al. 2020). In contrast, HCoV-19 may bind less efficiently to ACE2 in other species associated with SARS-like viruses, including rodents and civets(Wan et al. 2020; Zhang et al. 2020).

While these analyses suggest that HCoV-19 may be capable of binding the human ACE2 receptor with high affinity, computational analyses predict that the interaction is not ideal(Wan et al. 2020) and the RBD sequence is different from those previously shown in SARS-CoV to be optimal for receptor binding(Sheahan et al. 2008). Thus, the optimized binding of the HCoV-19 spike protein to the human ACE2 receptor is most likely the result of selection on a human or human-like ACE2 permitting another optimal binding solution to arise. This is strong evidence that HCoV-19 is *not* the product of a purposefully manipulated virus, a widely propagated conspiracy theory(Calisher et al. 2020).

Polybasic cleavage site and O-linked glycans

The second notable feature of HCoV-19 is a polybasic furin cleavage site (RRAR) in the S protein at the junction of S1 and S2, the two subunits of the spike (**Fig. 1b**)(Gallaher 2020; Coutard et al. 2020). Polybasic cleavage sites allow effective cleavage by furin and other proteases and play important roles in determining virus infectivity and host range(Nao et al. 2017). In addition to three basic arginines and an alanine at the cleavage site, a leading proline is also inserted around this site in HCoV-19; thus, the fully inserted sequence is PRRA (**Fig. 1b**). The strong turn created by the proline insertion is predicted to result in the addition of O-linked glycans to S673, T678, and S686 that flank the polybasic cleavage site. A polybasic cleavage site has not previously been observed in related “lineage B” betacoronaviruses and is a unique feature of HCoV-19(Gallaher 2020; Coutard et al. 2020). Some human betacoronaviruses, including HKU1, have polybasic cleavage sites, however, as well as predicted O-linked glycans near the S1/S2 cleavage site(Chan et al. 2008).

February 28, 2020 8:41 PM

The functional consequence of the furin cleavage site in HCoV-19 is unknown and it will be important to determine what impact, if any, the site has on transmissibility and pathogenesis in animal models(Bao et al. 2020). Experiments with SARS-CoV have shown that insertion of a furin cleavage site at the S1/S2 junction enhances cell–cell fusion, but does not affect virus entry(Follis et al. 2006). In addition, efficient cleavage of the MERS-CoV spike protein has been shown to enable MERS-like CoVs from bats to infect human cells(Menachery et al. 2019). In avian influenza viruses, polybasic cleavage sites can be acquired at the S1/S2 junction of the haemagglutinin (HA) protein under conditions that select for rapid virus replication and transmission, such as when the virus is replicating in highly dense chicken populations(Longping et al. 2014; Alexander and Brown 2009; Luczo et al. 2018). HA serves a similar function in cell-cell fusion and viral entry as the coronavirus spike protein. Acquisition of a polybasic cleavage site in HA, by either insertion or recombination, converts low pathogenicity avian influenza viruses into highly pathogenic forms(Longping et al. 2014; Alexander and Brown 2009; Luczo et al. 2018). The acquisition of polybasic cleavage sites by the influenza virus HA has also been observed after repeated forced passage in cell culture or through animals(Ito et al. 2001; Li et al. 1990). Similarly, an avirulent isolate of Newcastle Disease virus became highly pathogenic during serial passage in chickens by incremental acquisition of a polybasic cleavage site at the junction of its fusion protein subunits(Shengqing et al. 2002).

The potential function of the three predicted O-linked glycans is less clear, but they could create a "mucin-like domain" that would shield potential epitopes or key residues on the HCoV-19 spike protein(Bagdonaite and Wandall 2018; Tran et al. 2014). Several viruses employ mucin-like domains as part of a glycan shield that is involved in immune evasion (Bagdonaite and Wandall 2018). Although the algorithms for prediction of O-linked glycosylation are robust(Steentoft et al. 2013), biochemical analyses or structural studies are required to determine whether or not the predicted O-linked glycan sites are utilized.

Theories of HCoV-19 origins

It is improbable that HCoV-19 emerged through laboratory manipulation or engineering of a related SARS-like coronavirus. As noted above, the RBD of HCoV-19 is optimized for human ACE2 receptor binding with an efficient solution that is different from those previously predicted(Sheahan et al. 2008; Wan et al. 2020). Further, if genetic manipulation had been performed, one would expect that one of the several reverse genetic systems available for betacoronaviruses would have been used(Cui et al. 2019). This is not the case, however, as the genetic data irrefutably shows that HCoV-19 is not derived from any previously used virus backbone(Almazán et al. 2014). Instead, we propose two scenarios that can plausibly explain the origin of HCoV-19: (*i*) natural selection in a non-human animal host prior to zoonotic transfer, and (*ii*) natural selection in humans following zoonotic transfer. We also discuss whether selection during passage in culture could have given rise to the same observed features and conclude that such a scenario is unlikely.

Selection in an animal host

As many of the early cases of COVID-19 were linked to the Huanan seafood and wildlife market in Wuhan(Zhou et al. 2020; Wu et al. 2020), it is possible that an animal source was present at this location. Given the similarity of HCoV-19 to bat SARS-like CoVs(Wu et al. 2020), particularly RaTG13(Zhou et al. 2020), it is plausible that bats serve as reservoir hosts for its progenitor. Although RaTG13 sampled from a *Rhinolophus affinis* bat is ~96% identical overall to HCoV-19(Zhou et al. 2020), its S protein possesses distinct sequences in the RBD suggesting that it may not bind efficiently to the human ACE2 receptor (**Fig. 1a**)(Wan et al. 2020). It is also important to note that previous outbreaks of betacoronaviruses in humans involved direct exposure to animals other than bats, including civets and camels, which carry viruses that are genetically very similar to SARS-CoV(Ge et al. 2013) or MERS-CoV(Dudas et al. 2018), respectively. By analogy, viruses closely related to HCoV-19 may be circulating in one or more animal species.

February 28, 2020 8:41 PM

Malayan pangolins (*Manis javanica*) illegally imported into Guangdong province contain CoVs that are similar to HCoV-19(Wong et al. 2020; Lam et al. ; Xiao et al. 2020; Zhang et al. 2020). Although the bat virus RaTG13 remains the closest relative to HCoV-19 across the whole genome, a Malayan pangolin CoV exhibits strong similarity to HCoV-19 in the RBD, including all six key RBD residues (**Fig.** 1) (Zhang et al. 2020; Xiao et al. 2020). However, no CoV from a pangolin, nor any other animal species, has yet been identified that is sufficiently similar to HCoV-19 across its entire genome to support direct human infection. Similarly, although there is likely a history of complex recombination events in these viruses, including in the RBD and other domains of the S protein, none of the available bat or pangolin CoVs are sufficiently similar to HCoV-19 to have directly generated it by recombination.

Neither the bat nor pangolin betacoronaviruses sampled to date carry polybasic cleavage sites. However, as the diversity of CoVs in bats and other species is hugely undersampled, it is possible that an animal betacoronavirus will eventually be identified with a polybasic cleavage site. Mutations, including point mutations, insertions and deletions, can occur near the S1/S2 junction of CoVs(Licitra et al. 2013; Yamada and Liu 2009; Yamada et al. 1998; Lamers et al. 2016; Guan et al. 2003) suggesting that the polybasic site could arise by a natural evolutionary process. For a precursor virus to acquire both the polybasic cleavage site and mutations in the spike protein suitable for human ACE2 receptor binding, an animal host would likely have to have a high population density – to allow natural selection to proceed efficiently – and an ACE2 gene that is similar to the human orthologue. Further characterization of CoVs in animals that may harbour SARS-CoV-like viruses should be a public health priority.

Cryptic adaptation to humans

It is also possible that a progenitor to HCoV-19 jumped from a non-human animal to humans, with the genomic features described above acquired through adaptation during subsequent human-to-human transmission. We surmise that once these adaptations were acquired (either together or in series) it would enable the epidemic to take-off, producing a sufficiently large and unusual cluster of pneumonia cases to trigger the surveillance system that ultimately detected it.

All HCoV-19 genomes sequenced so far have the well adapted RBD and the polybasic cleavage site and are thus derived from a common ancestor that had these features. The presence of an RBD in pangolins that is very similar to the one in HCoV-19 means that this was likely already present in the virus that jumped to humans, even if we don't yet have the exact non-human progenitor virus. This leaves the polybasic cleavage site insertion to occur during human-to-human transmission. Following the example of the influenza A virus HA gene, a specific insertion or recombination event is required to enable the emergence of HCoV-19 as an epidemic pathogen.

Estimates of the timing of the most recent common ancestor (tMRCA) of HCoV-19 using currently available genome sequence data point to virus emergence in late November to early December 2019(Phylodynamic Analysis | 90 genomes | ...; Phylodynamic estimation of incidence ...), compatible with the earliest retrospectively confirmed cases(Huang et al. 2020). Hence, this scenario presumes a period of unrecognised transmission in humans between the initial zoonotic transfer event and the acquisition of the polybasic cleavage site. Sufficient opportunity could occur if there had been many prior zoonotic events producing short chains of human-to-human transmission (so-called 'stuttering chains') over an extended period. This is essentially the situation for MERS-CoV in the Arabian Peninsula where all the human cases are the result of repeated jumps of the virus from dromedary camels, producing single infections or short chains of transmission that eventually resolve. To date, after 2,499 cases over 8 years, no human adaptation has emerged that has allowed MERS-CoV to take hold in the human population.

How could we test whether cryptic spread of HCoV-19 enabled human adaptation? Metagenomic studies of banked serum samples could provide important information, but given the relatively short period of

February 28, 2020 8:41 PM

viraemia it may be impossible to detect low level HCoV-19 circulation in historical samples. Retrospective serological studies potentially could be informative and a few such studies have already been conducted. One found that animal importation traders had a 13% seropositivity to coronaviruses(Centers for Disease Control and Preve...), while another noted that 3% residents of a village in Southern China were seropositive to these viruses(Wang et al. 2018). Interestingly, 200 residents of Wuhan did not show coronavirus seroreactivity. Critically, however, these studies could not have distinguished whether positive serological responses were due to a prior infection with SARS-CoV or HCoV-19. Further retrospective serological studies should be conducted to determine the extent of prior human exposure to betacoronaviruses in different geographic areas, particularly using assays that can distinguish among multiple betacoronaviruses.

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 in multiple laboratories across the world(Ge et al. 2013; Hu et al. 2017; Zeng et al. 2016; Yang et al. 2015). There are also documented instances of the acquisition of SARS-CoV by laboratory personnel working under BSL-2 or BSL-3 containment(Lim et al. 2004; Senior 2003; Lim et al. 2006). While reverse engineering is not a trivial task, methods to add insertions, deletions or otherwise modify large CoV genomes are well established(Almazán et al. 2014). We must therefore consider the possibility of a deliberate or inadvertent release of HCoV-19. In theory, it is possible that HCoV-19 acquired the observed RBD mutations site during adaptation to passage in cell culture, as has been observed in studies with SARS-CoV(Sheahan et al. 2008) as well as MERS-CoV(Letko et al. 2018). However, the acquisition of the polybasic cleavage site or O-linked glycans - if functional - argues against this scenario. New polybasic cleavage sites have only been observed after prolonged passaging of low pathogenicity avian influenza virus in cell culture or animals. Furthermore, the generation of HCoV-19 by cell culture or animal passage would have required prior isolation of a progenitor virus with a very high genetic similarity. Subsequent generation of a polybasic cleavage site would have then required an intense program of passage in cell culture or animals with ACE-2 receptor similar to humans (e.g. ferrets). It is also questionable whether generation of the O-linked glycans would have occurred on cell culture passage, as such mutations typically suggest the involvement of an immune system that is not present *in vitro*.

Conclusions

In the midst of the global COVID-19 public health emergency it is reasonable to wonder why the origins of the epidemic matter. A detailed understanding of how an animal virus jumped species boundaries to infect humans so productively will help in the prevention of future zoonotic events. For example, if HCoV-19 pre-adapted in another animal species then we are at risk of future re-emergence events even if the current epidemic is controlled. In contrast, if the adaptive process we describe occurred in humans, then even if we have repeated zoonotic transfers they are unlikely to take-off unless the same series of mutations occurs. In addition, identifying the closest animal relatives of HCoV-19 will greatly assist studies of virus function. Indeed, the availability of the RaTG13 bat sequence facilitated the comparative genomic analysis performed here, helping to reveal the key mutations in the RBD as well as the polybasic cleavage site insertion.

The genomic features described here may in part explain the infectiousness and transmissibility of HCoV-19 in humans. Although genomic evidence does not support the idea that HCoV-19 is a laboratory construct, it is currently impossible to prove or disprove the other theories of its origin described here. While it is unclear whether future data will help resolve this issue, it is likely that more scientific data will swing the balance in favor of one hypothesis over another. Identifying the immediate non-human animal source and obtaining virus sequences from it would be the most definitive way of revealing virus origins. In addition, it would be helpful to obtain more genetic and functional data about the virus, including

February 28, 2020 8:41 PM

experimental studies of receptor binding and the role of the polybasic cleavage site and predicted O-linked glycans. The identification of a potential intermediate host of HCoV-19, as well as the sequencing of very early cases including those not connected to the Wuhan market, would similarly be highly informative. Irrespective of how HCoV-19 originated, the ongoing surveillance of pneumonia in humans and other animals is clearly of utmost importance.

Acknowledgements

We thank all those who have contributed HCoV-19 genome sequences to the GISAID database (https://www.gisaid.org/) and contributed analyses and ideas to Virological.org (http://virological.org/). We thank M. Farzan for discussions. We thank the Wellcome Trust for supporting this work. ECH is supported by an ARC Australian Laureate Fellowship (FL170100022). KGA is a Pew Biomedical Scholar and is supported by NIH NIAID grant U19AI135995. AR is supported by the Wellcome Trust (Collaborators Award 206298/Z/17/Z – ARTIC network) and the European Research Council (grant agreement no. 725422 – ReservoirDOCS). RFG is supported by NIH grants U19AI135995, U54 HG007480 and U19AI142790.

February 28, 2020 8:41 PM

Figure Legends

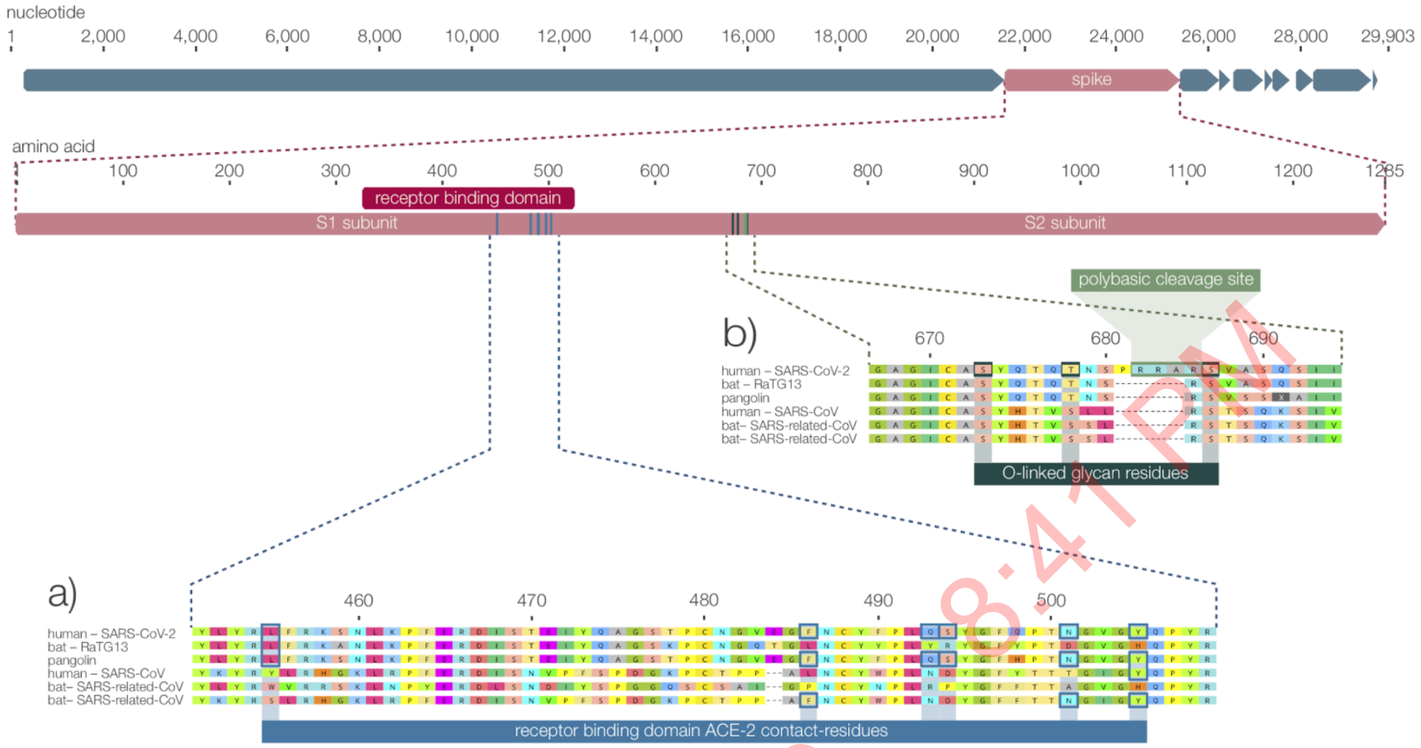


Figure 1. (a) Mutations in contact residues of the HCoV-19 spike protein. The spike protein of HCoV-19 (top) was aligned against the most closely related SARS-like CoVs and SARS-CoV. Key residues in the spike protein that make contact to the ACE2 receptor are marked with blue boxes in both HCoV-19 and the SARS-CoV Urbani strain. (b) Acquisition of polybasic cleavage site and O-linked glycans. Both the polybasic cleavage site and the three adjacent predicted O-linked glycans are unique to HCoV-19 and not previously seen in lineage B betacoronaviruses. Sequences shown are from NCBI GenBank, accession numbers MN908947, MN996532, AY278741, KY417146 and MK211376. The pangolin coronavirus sequences are a consensus generated from SRR10168377 and SRR10168378 (NCBI BioProject PRJNA573298)(Wong et al. 2020; Liu et al. 2019).

February 28, 2020 8:41 PM

References

February 28, 2020 8:41 PM

Referee #1:

Anderson presented a timely manuscript to share their points of view about the origin of SARS-CoV-2. There are several rumors about the origin of this virus. However, these "hypotheses" are entirely based on very limited, if any, scientific evidence. This reviewer sees most of the arguments raised by the authors are valid and convincing. However, the authors might want to consider these minor suggestions:

1. The sections for the RBD and cleavage site of Spike protein basically have summarized the existing findings from other recent publications. The authors might want to spell out that these two sections are review summaries. In addition, the author can present these two sections in a more condense format and save some space for something else (also see points 6 and 7 below)

We have edited these sections to be more concise. However, we think that it is important that the key genomic features are described in detail as they form the basis of the discussion that follows.

2. Fig. 1. This figure has 6 aligned sequences, but with only 5 sequence titles. The order of these titles are also not correct.

We have corrected these mistakes and updated the figure and titles.

3. Lines 170 -174. It is correct that no adaptive mutation has been found in the spike of MERS-CoV. Deletions in other ORF regions, however, were detected in some human MERS-CoV viruses (PMID: 26981770). In addition, the 29nt deletion of human SARS-CoV (PMID: 12958366) was suggested to have effects on host adaptation. The authors should also consider these findings. It is premature to say that this would not happen in SARS-CoV-2.

We have revised the text to incorporate this information as we agree that they are important for context.

4. Line 194. The accident at Singapore occurred in a BSL3, not BSL2, containment.

We have corrected this.

5. Line 194. Laboratory escapes of SARS occurred in Singapore, China and Taiwan (PMID: 16830004).

We have added this reference as well as the point about laboratory escapes of SARS-CoV in Singapore, China and Taiwan.

6. There are two recent reports about coronaviruses in pangolins. The authors might want to comment on these.

We have included these references as well as several others that have investigated pangolin CoV. In addition, and in response to Reviewer 2 (see below) we should point out that these additional pangolin CoV sequences do not further clarify the different scenarios discussed in our manuscript. There is nothing in these reports that changes our statements regarding a potential role of pangolins.

7. Optional: Can the authors share their views on the possibility of having a lab escape of a natural coronavirus? This is also one of the hypotheses that have been extensively discussed. The reviewer understands that this is entirely a different topic, but any insights are welcomed.

Nature rebuttal February 29, 2020 1:47 PM

The referee makes a good point but the exact nature of the zoonotic event (whether in a seafood market, an outlying rural area, or, indeed, a lab) cannot be meaningfully addressed here. We would prefer not to speculate.

Referee #2:

This is a perspective discussing evidence against a hypothetical lab origin of SARS-CoV-2. The paper addresses suboptimal composition of ACE2-binding sites in the RBD, 3 predicted O-linked glycosylation sites and a furin cleavage site in the glycoprotein that was speculated upon before.

The paper is itself interesting, but unnecessarily speculative. It's not clear why the authors do not refute a hypothetical lab origin in their coming publication on the ancestors of SARS-CoV-2 in bats and pangolins. The tree showing diverse pangolin viruses has kindly been made available by some of the authors in GISAID. Once the authors publish their new pangolin sequences, a lab origin will be extremely unlikely. It is not clear why the authors rush with a speculative perspective if their central hypothesis can be supported by their own data. Please explain.

Our manuscript is written to explore the potential origin of SARS-CoV-2. We do not believe it is speculative: rather, it simply takes the available data and proposes a series of plausible hypotheses for how these data may have arisen. We do this in a logical, balanced and unbiased manner; this is critically important because it gives our work credibility. It is science, not opinion.

Unfortunately, the newly available pangolin sequences do not elucidate the origin of SARS-CoV-2 or refute a lab origin. Hence, the reviewer is incorrect on this point. To clarify, while the RBD from the Guangdong pangolin CoVs is the closest to that found SARS-CoV-2, they are more divergent in the remainder of the viral genome (for which the bat virus RaTG13 is still the closest) and do not possess the polybasic cleavage site insertion. Hence, there is no evidence on present data that the pangolin CoVs are directly related to the COVID-19 epidemic.

Another critical aspect of this text is the complete lack of referencing to a potential debate on a hypothetical lab origin. Who said this, why is this considered a problem? There are indeed a few apparently uninformed statements claiming the virus may be a Chinese bioweapon, but is this really problematic on a larger scale? The central reason for issuing this text must be exhaustively referenced and discussed.

The possibility that SARS-CoV-2 originated as an engineered bioweapon has had widespread discussion in the press and on social media (particularly in China) and is hence problematic on a very large scale. This particular topic was also the reason for a recent request from the White House: https://bit.ly/2HMndCi. A group of public health scientists recently wrote a letter to *The Lancet* in which they *"strongly condemn conspiracy theories suggesting that COVID-19 does not have a natural origin."* We now reference this publication (the publication itself also references our pre-print of this manuscript).

It appears that the reviewer is considering bioweapons, engineering, and lab accidents as one and the same theory lumped in under a "lab origin". This is inappropriate and will likely only increase confusion. We do briefly discuss the notion of deliberate engineering of the virus and explain how the evidence is strongly against this explanation. Importantly, we do not discuss the concept of 'bioweapons' per se as there are many, legitimate, research uses of bio-engineering of viruses and do not want to equate the two in public perception.

Nature rebuttal February 29, 2020 1:47 PM

As we very clearly state in the manuscript, there are far more subtle scenarios that need to be considered carefully and scientifically: for example, of accidental infections in a lab while culturing SARS-like CoVs. Since accidental infections and other lab 'escapes' happen frequently across the world (and as we mention, happened multiple times with SARS-CoV following the SARS epidemic) we firmly believe that this discussion is of major importance and must be had. In particular, the culturing of SARS-like CoVs from animals is typically performed under BSL-2 and has been ongoing for years.

Dismissing this potential accidental scenario out of hand - or considering it in the same category as conspiracy theories about bioweapons and deliberate engineering (that, as we outline, are clearly wrong) - would be irresponsible. Our manuscript will only serve its purpose if it considers all possible hypotheses equally. Any perceived bias or omission will undermine its credibility.

We have modified the text to more clearly state the scenarios we are considering.

The authors state that a predicted polybasic cleavage site is unique to SARS-CoV-2 in SARS viruses. Who knows how many out of thousands undiscovered bat ancestors also acquired such a motif, the sampling bias in descriptions of remote bat viruses is dramatic. This should be discussed. Also state clearly that this site is only predicted so far and that experimental evidence for its biological function and its potential impact on pathogenesis are required.

We agree that the diversity in bat CoVs is highly undersampled and that it is possible that a bat CoV progenitor could be discovered that contains a polybasic site. We say nothing against this in the manuscript. However, given the diversity of SARS-like CoVs already sampled from bats, pangolins, and many other animals - none of which possess this insertion - it is obviously reasonable to hypothesize that it may have been gained in the lineage leading to SARS-CoV-2. This is explicitly discussed in our revision.

While other CoVs utilize polybasic sites, we did refer to the site as a predicted cleavage site. Studies have since been performed showing that this site is indeed functional, which we now reference (Walls et al., DOI: 10.1016/j.cell.2020.02.058). We also state that it will be necessary to test the effect of the polybasic site on pathogenesis, which will require the establishment of an animal model.

The predicted O-linked glycosylation sites are mysterious. What do the authors imply with those sites? *In silico* prediction of O-linked glycosylation sites is not robust and whether these sites indeed exist requires experimental validation. Even if those sites exist, why are they relevant? This is not addressed at all. If the authors assume these sites constitute part of a glycan shield, they should say so and weigh their assumption carefully.

Although not previously described for CoV proteins, numerous other viral proteins have mucin-like domains that are involved in immune evasion. The revised text adds relevant references and is more explicit about the potential relevance of the predicted O-linked glycan sites. As stated previously, we consider that these sites - if indeed utilized - may be part of the glycan shield and have made this point directly in the revised text. We also agree that the predicted O-linked glycosylation sites require experimental validation not only in SARS-CoV-2 but in other CoVs that have similar predicted sites. This was also already stated explicitly, but in our revision we have stated this even more directly.

Finally, the main argument against a hypothetical lab origin seems the required reconstruction of a backbone of a bat virus of unknown pathogenesis. It does not seem feasible that any scientist would disembark on such an uncertain endeavor. This difficulties of coronavirus reverse genetics should be stated clearly.

Nature rebuttal February 29, 2020 1:47 PM

The reviewer's restatement of our argument is correct. We reiterate that the purpose of our manuscript was not to refute the conspiracy theory that SARS-CoV-2 was bioengineered. Rather, it was carefully designed to be a balanced and unbiased assessment of the available data. The statement that no scientist would embark on such an endeavor is a subjective one with no supporting evidence - in fact, scientists have already created a reverse genetics system for SARS-CoV-2, which was completed in two weeks (Thao et al., bioRxiv 10.1101/2020.02.21.959817).

Nature rebuttal February 29, 2020 1:47 PM

Editor’s comments:

While the Perspective is interesting and timely one of our referees raised concerns (also emphasised to the editors) about whether such a piece would feed or quash the conspiracy theories.

Critically, the purpose of our manuscript was not to quash conspiracy theories. Rather, our aim was to carefully examine in a balanced and unbiased manner the evidence for and against a number of possible probable scenarios for the proximal origins of SARS-CoV-2.

But more importantly this reviewer feels, and we agree, that the Perspective would quickly become outdated when more scientific data are published (for example on potential reservoir hosts).

Of course, it is likely that more scientific data will swing the balance in favor of one hypothesis over another. However, the same can be said of many of the papers published on COVID-19: as we learn more about the virus and the disease so previous publications may be quickly revised. We contend, however, our piece will remain relevant even if more data becomes available because these data would potentially confirm which of these scenarios is correct. Most importantly, our manuscript sets out what evidence is needed to test the hypotheses outlined and will therefore serve as an important starting point for guiding future research.

Nature rebuttal February 29, 2020 1:47 PM

The Proximal Origin of HCoV-19

Kristian G. Andersen[1,2*], Andrew Rambaut[3], W. Ian Lipkin[4], Edward C. Holmes[5] & Robert F. Garry[6,7]

[1]Department of Immunology and Microbiology, The Scripps Research Institute, La Jolla, CA 92037, USA.
[2]Scripps Research Translational Institute, La Jolla, CA 92037, USA.
[3]Institute of Evolutionary Biology, University of Edinburgh, Edinburgh, UK.
[4]Center for Infection and Immunity, Mailman School of Public Health of Columbia University, New York, New York, USA.
[5]Marie Bashir Institute for Infectious Diseases and Biosecurity, School of Life and Environmental Sciences and School of Medical Sciences, The University of Sydney, Sydney, Australia.
[6]Tulane University, School of Medicine, Department of Microbiology and Immunology, New Orleans, LA, USA.
[7]Zalgen Labs, LCC, Germantown, MD, USA.

*Corresponding author:
Kristian G. Andersen
Department of Immunology and Microbiology,
The Scripps Research Institute,
La Jolla, CA 92037, USA.

Since the first reports of novel pneumonia (COVID-19) in Wuhan, Hubei province, China[1,2] there has been considerable discussion on the origin of the causative virus, human coronavirus 2019 (HCoV-19[3]; also referred to as SARS-CoV-2[4]). Infections with HCoV-19 are now widespread, and as of 29 February 2020, 86,012 cases have been confirmed in 57 countries, with 2,941 deaths[5], although these are likely an underestimate with limited reporting of mild and asymptomatic cases.

HCoV-19 is the seventh coronavirus known to infect humans. Three of these viruses, SARS-CoV, MERS-CoV, and HCoV-19 can cause severe disease; four, HKU1, NL63, OC43 and 229E, are associated with mild respiratory symptoms[7]. Herein, we review what can be deduced about the origin of HCoV-19 from the comparative analysis of genome sequence data. In particular, we offer a perspective on the notable features in the HCoV-19 genome and discuss scenarios by which they could have arisen. Our analysis provides strong evidence that HCoV-19 is not a laboratory construct nor a purposefully manipulated virus.

Notable features of the HCoV-19 genome

Our genomic comparison of alpha- and betacoronaviruses (family *Coronaviridae*) identifies two notable features of the HCoV-19 genome: (*i*) based on structural studies[8–10] and biochemical experiments[1,9,11,12], HCoV-19 appears optimized for binding to the human ACE2 receptor; (*ii*) the highly variable spike (S) protein of HCoV-19 has a functional polybasic (furin) cleavage site at the S1/S2 boundary through the insertion of twelve nucleotides[10,13,14]. Additionally, this led to the predicted acquisition of three O-linked glycans around the polybasic cleavage site.

Mutations in the receptor binding domain of HCoV-19

The receptor binding domain (RBD) in the spike protein is the most variable part of the coronavirus genome[1,2]. Six RBD amino acids have been shown to be critical for binding to ACE2 receptors and determining the host range of SARS-like viruses[8]. Using coordinates based SARS-CoV, they are Y442, L472, N479, D480, T487, and Y4911 corresponding to L455, F486, Q493, S494, N501, and Y505 in HCoV-19[8]. Five of these six residues differ between HCoV-19 and SARS-CoV (**Fig. 1a**). Based on structural studies[8–10] and biochemical experiments[1,9,11,12], HCoV-19 seems to have an RBD that binds with high affinity to ACE2 from human, non-human primate, ferret, pig, and cat, and some other species with high receptor homology[8].

March 3, 2020 6:09 PM

While these analyses suggest that HCoV-19 may be capable of binding human ACE2 with high affinity, computational analyses predict that the interaction is not ideal[8] and the RBD sequence is different from those shown in SARS-CoV to be optimal for receptor binding[16]. Thus, the optimized binding of HCoV-19 spike protein to human ACE2 is most likely the result of natural selection on a human or human-like ACE2 permitting another optimal binding solution to arise. This is strong evidence that HCoV-19 is *not* the product of purposeful manipulation.

Polybasic furin cleavage site and O-linked glycans

The second notable feature of HCoV-19 is a polybasic furin cleavage site (RRAR) at the junction of S1 and S2, the two subunits of the spike (**Fig. 1b**)[13,14]. This allows effective cleavage by furin and other proteases and plays an important role in determining virus infectivity and host range[18]. In addition, a leading proline is also inserted at this site in HCoV-19; thus, the inserted sequence is PRRA (**Fig. 1b**). The strong turn created by the proline is predicted to result in the addition of O-linked glycans to S673, T678, and S686 flanking the cleavage site that are unique to HCoV-19 (**Fig. 1b**). Polybasic cleavage sites have not been observed in related "lineage B" betacoronaviruses, although other human betacoronaviruses, including HKU1 (lineage A), have them and predicted O-linked glycans near the S1/S2 cleavage site[19]. Given the level of genetic variation in the S protein it is likely that HCoV-19-like viruses with partial or full polybasic sites will be discovered in other species.

The functional consequence of the furin cleavage site in HCoV-19 is unknown and it will be important to determine what impact the feature has on transmissibility and pathogenesis in animal models[21]. Experiments with SARS-CoV have shown that insertion of a furin cleavage site at the S1/S2 junction enhances cell–cell fusion without affecting virus entry[22]. In addition, efficient cleavage of the MERS-CoV spike enables MERS-like coronaviruses from bats to infect human cells[23]. In avian influenza viruses, rapid virus replication and transmission in highly dense chicken populations selects for the acquisition of polybasic cleavage sites in the haemagglutinin (HA) protein , [24–26]. HA serves a similar function in cell-cell fusion and viral entry as the coronavirus spike protein. Acquisition polybasic cleavage sites in HA, by insertion or recombination, converts low pathogenicity avian influenza viruses into highly pathogenic forms[24–26]. The acquisition of polybasic cleavage sites by HA has also been observed after repeated passage in cell culture or through animals[27,28].

The function of the predicted O-linked glycans is less clear, but they could create a "mucin-like domain" shielding potential epitopes or key residues on the HCoV-19 spike protein[30,31]. Several viruses employ mucin-like domains as part of a glycan shield involved in immune evasion[30]. Although prediction of O-linked glycosylation is robust[32], biochemical analyses or structural studies are required to determine whether or not these sites in HCoV-19 are utilized.

Theories of HCoV-19 origins

It is improbable that HCoV-19 emerged through laboratory manipulation or engineering of a related SARS-like coronavirus. As noted above, the RBD of HCoV-19 is optimized for human ACE2 binding with an efficient solution different from those previously predicted[8,16]. Further, had genetic manipulation had been performed, one of the several reverse genetic systems available for betacoronaviruses would likely have been used[20]. However, the genetic data irrefutably show that HCoV-19 is not derived from any previously used virus backbone[33]. Instead, we propose two scenarios that can plausibly explain the origin of HCoV-19: (*i*) natural selection in a non-human animal host prior to zoonotic transfer, and (***ii***) natural selection in humans following zoonotic transfer. We also discuss whether selection during passage in culture could have given rise to the same observed features.

March 3, 2020 6:09 PM

Natural selection in an animal host prior to zoonotic transfer

As many of the early cases of COVID-19 were linked to the Huanan seafood and wildlife market in Wuhan[1,2], it is possible that an animal source was present at this location. Given the similarity of HCoV-19 to bat SARS-like coronaviruses[2], it is likely that bats serve as reservoir hosts for its progenitor. Although RaTG13, sampled from a *Rhinolophus affinis* bat, is ~96% identical overall to HCoV-19[1], its spike diverges in the RBD suggesting that it may not bind efficiently to the human ACE2 receptor (**Fig. 1a**)[8].

Malayan pangolins (*Manis javanica*) illegally imported into Guangdong province contain coronaviruses similar to HCoV-19[15,37–39]. Although the RaTG13 bat virus remains the closest relative to HCoV-19 across the whole genome[1], some pangolin coronaviruses exhibit strong similarity to HCoV-19 in the RBD, including all six key RBD residues (**Fig. 1**)[15,39]. This clearly shows the HCoV-19 spike protein optimized for binding to human-like ACE2 occurs in nature and is the result of natural selection. Similarly, neither the bat nor pangolin betacoronaviruses sampled to date carry polybasic cleavage sites. Although a non-human animal coronavirus, sufficiently similar to HCoV-19 across its entire genome that it could have served as the direct progenitor of the virus, has yet to be identified, the diversity of coronaviruses in bats and other species is massively undersampled. Mutations, including point mutations, insertions and deletions, can occur near the S1/S2 junction of coronaviruses[34,40–43] suggesting that the polybasic site could arise by a natural evolutionary process. For a precursor virus to acquire both the polybasic cleavage site and mutations in the spike protein suitable for human ACE2 receptor binding, an animal host would likely have to have a high population density – to allow natural selection to proceed efficiently – and an ACE2 gene that is similar to the human orthologue.

Natural selection in humans following zoonotic transfer

It is possible that a progenitor to HCoV-19 jumped into humans, acquiring the genomic features described above through adaptation during (undetected) human-to-human transmission. Once acquired, these adaptations would enable the epidemic to take off, producing a sufficiently large and unusual cluster of pneumonia cases to trigger the surveillance system that ultimately detected it[1,2].

All HCoV-19 genomes sequenced so far have the genomic features derived above and are thus derived from a common ancestor that had them too. The presence in pangolins of an RBD very similar to that in HCoV-19 means we can infer this was also likely in the virus that jumped to humans. This leaves the polybasic cleavage site insertion to occur during human-to-human transmission.

Estimates of the timing of the most recent common ancestor of HCoV-19 using currently available genome sequence data point to virus emergence in late November to early December 2019[44–46], compatible with the earliest retrospectively confirmed cases[47]. Hence, this scenario presumes a period of unrecognised transmission in humans between the initial zoonotic transfer event and the acquisition of the polybasic cleavage site. Sufficient opportunity could occur if there had been many prior zoonotic events producing short chains of human-to-human transmission over an extended period. This is essentially the situation for MERS-CoV in the Arabian Peninsula where all the human cases are the result of repeated jumps of the virus from dromedary camels, producing single infections or short transmission chainsthat eventually resolve, with no adaptation to sustained human transmission[48].

Metagenomic studies of banked human samples could provide important information on whether this cryptic spread has occurred, although given the relatively short period of viremia it may be impossible to detect low level HCoV-19 circulation in historical samples. Retrospective serological studies could also be informative and a few such studies have been conducted. One found that animal importation traders had a 13% seropositivity to coronaviruses[49], while another noted that 3% residents of a village in Southern China were seropositive to SARS-like coronaviruses[50]. Critically, however, these studies could not have distinguished whether positive serological responses were due to prior infections with SARS-CoV, HCoV-19,

March 3, 2020 6:09 PM

or other SARS-like coronaviruses. Further serological studies should be conducted to determine the extent of prior human exposure to HCoV-19.

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 in laboratories across the world[51–54]. There are also several documented instances of laboratory escapes of SARS-CoV[55–57]. We must therefore examine the possibility of a inadvertent laboratory release of HCoV-19.

In theory, it is possible that HCoV-19 acquired RBD mutations (**Fig. 1a**) during adaptation to passage in cell culture, as has been observed in studies with SARS-CoV[16] and MERS-CoV[58]. The finding of SARS-like coronaviruses from pangolins with near-identical RBDs, however, provides a much stronger and parsimonious explanation for how HCoV-19 acquired these via recombination or mutation[20].

The acquisition of both the polybasic cleavage site and predicted O-linked glycans also argues against any type of culture-based scenario. New polybasic cleavage sites have only been observed after prolonged passage of low pathogenicity avian influenza virus *in vitro* or *in vivo*[24,26–28]. Furthermore, a hypothetical generation of HCoV-19 by cell culture or animal passage would have required prior isolation of a progenitor virus with a very high genetic similarity, which has not been described. Subsequent generation of a polybasic cleavage site would have then required repeated passage in cell culture or animals with ACE2 receptors similar to humans (e.g. ferrets), but such work has also not previously been described. Finally, the generation of the predicted O-linked glycans is also unlikely to have occured due to cell culture passage, as such features suggest the involvement of an immune system[30].

Conclusions

In the midst of the global COVID-19 public health emergency it is reasonable to wonder why the origins of the epidemic matter. A detailed understanding of how an animal virus jumped species boundaries to infect humans so productively will help in the prevention of future zoonotic events. For example, if HCoV-19 pre-adapted in another animal species then we are at risk of future re-emergence events. In contrast, if the adaptive process we describe occurred in humans, then even if we have repeated zoonotic transfers they are unlikely to take-off without the same series of mutations. In addition, identifying the closest animal relatives of HCoV-19 will greatly assist studies of virus function. Indeed, the availability of the RaTG13 bat sequence helped reveal the key mutations in the RBD as well as the polybasic cleavage site insertion.

The genomic features described here may in part explain the infectiousness and transmissibility of HCoV-19 in humans. Although the evidence shows that HCoV-19 is not a purposefully manipulated virus, it is currently impossible to prove or disprove the other theories of its origin described here. However, since we observe all notable HCoV-19 features - including the optimized RBD and furin cleavage site - in related coronaviruses in nature, we do not believe that selection during passage, or any other type of laboratory-based scenario, is necessary.

More scientific data could swing the balance of evidence to favor one hypothesis over another. Obtaining virus sequences from any immediate non-human animal source would be the most definitive way of revealing virus origins. For example, a future observation of an intermediate or fully formed polybasic cleavage site in an HCoV-19 related virus from animals would lend very strong support to the natural selection hypotheses. It would also be helpful to obtain more genetic and functional data about HCoV-19, including experimental studies of receptor binding and the role of the polybasic cleavage site and predicted O-linked glycans. The identification of a potential intermediate host of HCoV-19, as well as the sequencing of very early cases would similarly be highly informative. Irrespective of the exact mechanisms of how

March 3, 2020 6:09 PM

HCoV-19 originated via natural selection, the ongoing surveillance of pneumonia in humans and other animals is clearly of utmost importance.

Acknowledgements

We thank all those who have contributed HCoV-19 genome sequences to the GISAID database (https://www.gisaid.org/) and contributed analyses and ideas to Virological.org (http://virological.org/). We thank M. Farzan for discussions. We thank the Wellcome Trust for supporting this work. KGA is a Pew Biomedical Scholar and is supported by NIH NIAID grant U19AI135995. AR is supported by the Wellcome Trust (Collaborators Award 206298/Z/17/Z – ARTIC network) and the European Research Council (grant agreement no. 725422 – ReservoirDOCS). ECH is supported by an ARC Australian Laureate Fellowship (FL170100022). RFG is supported by NIH grants U19AI135995, U54 HG007480 and U19AI142790.

March 3, 2020 6:09 PM

Figure Legends

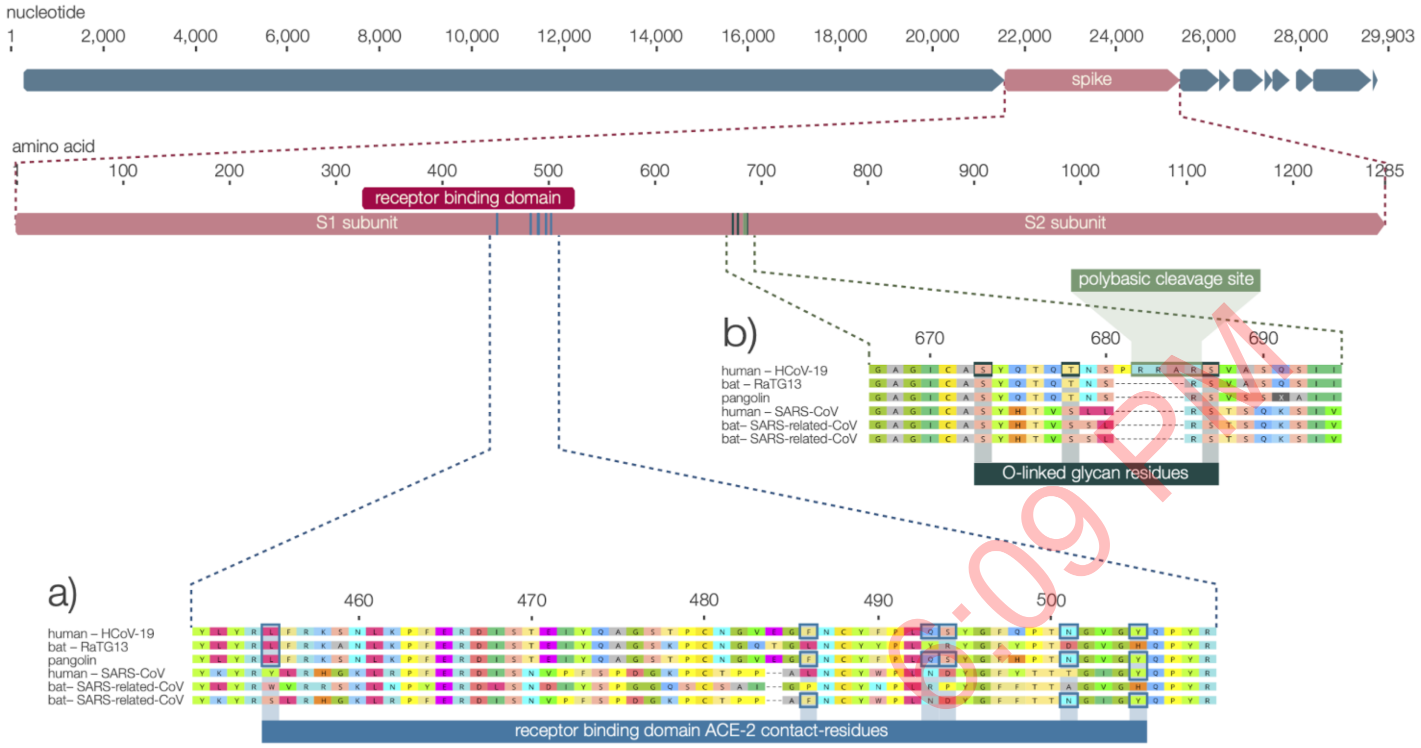


Figure 1. (**a**) Mutations in contact residues of the HCoV-19 spike protein. The spike protein of HCoV-19 (top) was aligned against the most closely related SARS-like CoVs and SARS-CoV. Key residues in the spike protein that make contact to the ACE2 receptor are marked with blue boxes in both HCoV-19 and the SARS-CoV Urbani strain. (**b**) Acquisition of polybasic cleavage site and O-linked glycans. Both the polybasic cleavage site and the three adjacent predicted O-linked glycans are unique to HCoV-19 and not previously seen in lineage B betacoronaviruses. Sequences shown are from NCBI GenBank, accession numbers MN908947, MN996532, AY278741, KY417146 and MK211376. The pangolin coronavirus sequences are a consensus generated from SRR10168377 and SRR10168378 (NCBI BioProject PRJNA573298)[37,59].

March 3, 2020 6:09 PM

References

1. Zhou, P. *et al.* A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* (2020) doi:10.1038/s41586-020-2012-7.
2. Wu, F. *et al.* A new coronavirus associated with human respiratory disease in China. *Nature* (2020) doi:10.1038/s41586-020-2008-3.
3. Jiang, S. *et al.* A distinct name is needed for the new coronavirus. *Lancet* (2020) doi:10.1016/S0140-6736(20)30419-0.
4. Gorbalenya, A. E. Severe acute respiratory syndrome-related coronavirus–The species and its viruses, a statement of the Coronavirus Study Group. *BioRxiv* (2020).
5. Dong, E., Du, H. & Gardner, L. An interactive web-based dashboard to track COVID-19 in real time. *Lancet Infect. Dis.* (2020) doi:10.1016/S1473-3099(20)30120-1.
6. Statement on the second meeting of the International Health Regulations (2005) Emergency Committee regarding the outbreak of novel coronavirus (2019-nCoV). https://www.who.int/news-room/detail/30-01-2020-statement-on-the-second-meeting-of-the-international-health-regulations-(2005)-emergency-committee-regarding-the-outbreak-of-novel-coronavirus-(2019-ncov).
7. Corman, V. M., Muth, D., Niemeyer, D. & Drosten, C. Hosts and Sources of Endemic Human Coronaviruses. *Adv. Virus Res.* **100**, 163–188 (2018).
8. 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.
9. Wrapp, D. *et al.* Cryo-EM Structure of the 2019-nCoV Spike in the Prefusion Conformation. *bioRxiv* 2020.02.11.944462 (2020) doi:10.1101/2020.02.11.944462.
10. Walls, A. C. *et al.* Structure, function and antigenicity of the SARS-CoV-2 spike glycoprotein. *bioRxiv* 2020.02.19.956581 (2020) doi:10.1101/2020.02.19.956581.

March 3, 2020 6:09 PM

11. Letko, M., Marzi, A. & Munster, V. Functional assessment of cell entry and receptor usage for SARS-CoV-2 and other lineage B betacoronaviruses. *Nat Microbiol* (2020) doi:10.1038/s41564-020-0688-y.
12. 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.
13. Gallaher, W. Analysis of Wuhan Coronavirus: Deja Vu. *Virological* http://virological.org/t/analysis-of-wuhan-coronavirus-deja-vu/357 (2020).
14. Coutard, B. *et al.* The spike glycoprotein of the new coronavirus 2019-nCoV contains a furin-like cleavage site absent in CoV of the same clade. *Antiviral Res.* 104742 (2020).
15. Zhang, T., Wu, Q. & Zhang, Z. Pangolin homology associated with 2019-nCoV. *bioRxiv* 2020.02.19.950253 (2020) doi:10.1101/2020.02.19.950253.
16. 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).
17. Calisher, C. *et al.* Statement in support of the scientists, public health professionals, and medical professionals of China combatting COVID-19. *Lancet* (2020) doi:10.1016/S0140-6736(20)30418-9.
18. Nao, N. *et al.* Genetic Predisposition To Acquire a Polybasic Cleavage Site for Highly Pathogenic Avian Influenza Virus Hemagglutinin. *MBio* **8**, (2017).
19. Chan, C.-M. *et al.* Spike protein, S, of human coronavirus HKU1: role in viral life cycle and application in antibody detection. *Exp. Biol. Med.* **233**, 1527–1536 (2008).
20. Cui, J., Li, F. & Shi, Z.-L. Origin and evolution of pathogenic coronaviruses. *Nat. Rev. Microbiol.* **17**, 181–192 (2019).
21. Bao, L. *et al.* The Pathogenicity of SARS-CoV-2 in hACE2 Transgenic Mice. *bioRxiv* 2020.02.07.939389 (2020) doi:10.1101/2020.02.07.939389.
22. Follis, K. E., York, J. & Nunberg, J. H. Furin cleavage of the SARS coronavirus spike glycoprotein

March 3, 2020 6:09 PM

enhances cell-cell fusion but does not affect virion entry. *Virology* **350**, 358–369 (2006).

23. Menachery, V. D. *et al.* Trypsin treatment unlocks barrier for zoonotic bat coronaviruses infection. *J. Virol.* (2019) doi:10.1128/JVI.01774-19.
24. 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).
25. Alexander, D. J. & Brown, I. H. History of highly pathogenic avian influenza. *Rev. Sci. Tech.* **28**, 19–38 (2009).
26. 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).
27. Ito, T. *et al.* Generation of a highly pathogenic avian influenza A virus from an avirulent field isolate by passaging in chickens. *J. Virol.* **75**, 4439–4443 (2001).
28. Li, S. Q., Orlich, M. & Rott, R. Generation of seal influenza virus variants pathogenic for chickens, because of hemagglutinin cleavage site changes. *J. Virol.* **64**, 3297–3303 (1990).
29. Shengqing, Y. *et al.* Generation of velogenic Newcastle disease viruses from a nonpathogenic waterfowl isolate by passaging in chickens. *Virology* **301**, 206–211 (2002).
30. Bagdonaite, I. & Wandall, H. H. Global aspects of viral glycosylation. *Glycobiology* **28**, 443–467 (2018).
31. Tran, E. E. H. *et al.* Spatial localization of the Ebola virus glycoprotein mucin-like domain determined by cryo-electron tomography. *J. Virol.* **88**, 10958–10962 (2014).
32. Steentoft, C. *et al.* Precision mapping of the human O-GalNAc glycoproteome through SimpleCell technology. *EMBO J.* **32**, 1478–1488 (2013).
33. Almazán, F. *et al.* Coronavirus reverse genetic systems: infectious clones and replicons. *Virus Res.* **189**, 262–270 (2014).
34. Guan, Y. *et al.* Isolation and characterization of viruses related to the SARS coronavirus from animals in southern China. *Science* **302**, 276–278 (2003).

March 3, 2020 6:09 PM

35. Haagmans, B. L. *et al.* Middle East respiratory syndrome coronavirus in dromedary camels: an outbreak investigation. *Lancet Infect. Dis.* **14**, 140–145 (2014).

36. Azhar, E. I. *et al.* Evidence for camel-to-human transmission of MERS coronavirus. *N. Engl. J. Med.* **370**, 2499–2505 (2014).

37. Wong, M. C., Javornik Cregeen, S. J., Ajami, N. J. & Petrosino, J. F. Evidence of recombination in coronaviruses implicating pangolin origins of nCoV-2019. *bioRxiv* 2020.02.07.939207 (2020) doi:10.1101/2020.02.07.939207.

38. Lam, T. T.-Y. *et al.* Identification of 2019-nCoV related coronaviruses in Malayan pangolins in southern China. *bioRxiv* (2020) doi:10.1101/2020.02.13.945485.

39. Xiao, K. *et al.* Isolation and Characterization of 2019-nCoV-like Coronavirus from Malayan Pangolins. *bioRxiv* 2020.02.17.951335 (2020) doi:10.1101/2020.02.17.951335.

40. Licitra, B. N. *et al.* Mutation in spike protein cleavage site and pathogenesis of feline coronavirus. *Emerg. Infect. Dis.* **19**, 1066–1073 (2013).

41. Yamada, Y. & Liu, D. X. Proteolytic activation of the spike protein at a novel RRRR/S motif is implicated in furin-dependent entry, syncytium formation, and infectivity of coronavirus infectious bronchitis virus in cultured cells. *J. Virol.* **83**, 8744–8758 (2009).

42. Yamada, Y. K., Takimoto, K., Yabe, M. & Taguchi, F. Requirement of proteolytic cleavage of the murine coronavirus MHV-2 spike protein for fusion activity. *Adv. Exp. Med. Biol.* **440**, 89–93 (1998).

43. Lamers, M. M. *et al.* Deletion Variants of Middle East Respiratory Syndrome Coronavirus from Humans, Jordan, 2015. *Emerg. Infect. Dis.* **22**, 716–719 (2016).

44. Phylodynamic Analysis | 90 genomes | 12 Feb 2020 – Novel 2019 coronavirus / nCoV-2019 Genomic Epidemiology. http://virological.org/t/phylodynamic-analysis-90-genomes-12-feb-2020/356 (2020).

45. Phylodynamic estimation of incidence and prevalence of novel coronavirus (nCoV) infections through time – Novel 2019 coronavirus / nCoV-2019 Genomic Epidemiology.

March 3, 2020 6:09 PM

http://virological.org/t/phylodynamic-estimation-of-incidence-and-prevalence-of-novel-coronavirus-ncov-infections-through-time/391 (2020).

46. Clock and TMRCA based on 27 genomes. *Virological* http://virological.org/t/clock-and-tmrca-based-on-27-genomes/347 (2020).

47. Huang, C. *et al.* Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet* (2020) doi:10.1016/S0140-6736(20)30183-5.

48. Dudas, G., Carvalho, L. M., Rambaut, A. & Bedford, T. MERS-CoV spillover at the camel-human interface. *Elife* **7**, (2018).

49. Centers for Disease Control and Prevention (CDC). Prevalence of IgG antibody to SARS-associated coronavirus in animal traders--Guangdong Province, China, 2003. *MMWR Morb. Mortal. Wkly. Rep.* **52**, 986–987 (2003).

50. Wang, N. *et al.* Serological Evidence of Bat SARS-Related Coronavirus Infection in Humans, China. *Virol. Sin.* **33**, 104–107 (2018).

51. Ge, X.-Y. *et al.* Isolation and characterization of a bat SARS-like coronavirus that uses the ACE2 receptor. *Nature* **503**, 535–538 (2013).

52. 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).

53. 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).

54. 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).

55. Lim, P. L. *et al.* Laboratory-acquired severe acute respiratory syndrome. *N. Engl. J. Med.* **350**, 1740–1745 (2004).

March 3, 2020 6:09 PM

56. Senior, K. Recent Singapore SARS case a laboratory accident. *Lancet Infect. Dis.* **3**, 679 (2003).

57. Lim, W., Ng, K.-C. & Tsang, D. N. C. Laboratory containment of SARS virus. *Ann. Acad. Med. Singapore* **35**, 354–360 (2006).

58. Letko, M. *et al.* Adaptive Evolution of MERS-CoV to Species Variation in DPP4. *Cell Rep.* **24**, 1730–1737 (2018).

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

March 3, 2020 6:09 PM

The Proximal Origin of SARS-CoV-2

Kristian G. Andersen[1,2*], Andrew Rambaut[3], W. Ian Lipkin[4], Edward C. Holmes[5] & Robert F. Garry[6,7]

[1]Department of Immunology and Microbiology, The Scripps Research Institute, La Jolla, CA 92037, USA.
[2]Scripps Research Translational Institute, La Jolla, CA 92037, USA.
[3]Institute of Evolutionary Biology, University of Edinburgh, Edinburgh, UK.
[4]Center for Infection and Immunity, Mailman School of Public Health of Columbia University, New York, New York, USA.
[5]Marie Bashir Institute for Infectious Diseases and Biosecurity, School of Life and Environmental Sciences and School of Medical Sciences, The University of Sydney, Sydney, Australia.
[6]Tulane University, School of Medicine, Department of Microbiology and Immunology, New Orleans, LA, USA.
[7]Zalgen Labs, LCC, Germantown, MD, USA.

*Corresponding author: andersen@scripps.edu

March 4, 2020 6:50 PM

TO THE EDITOR - Since the first reports of novel pneumonia (COVID-19) in Wuhan, Hubei province, China[1,2] there has been considerable discussion on the origin of the causative virus SARS-CoV-2[3] (also referred to as HCoV-19[4]. Infections with SARS-CoV-2 are now widespread, and as of 29 February 2020, 86,012 cases have been confirmed in more than 60 countries, with 2,941 deaths[5].

SARS-CoV-2 is the seventh coronavirus known to infect humans. SARS-CoV, MERS-CoV, and SARS-CoV-2 can cause severe disease, whereas HKU1, NL63, OC43 and 229E, are associated with mild symptoms[6]. Herein, we review what can be deduced about the origin of SARS-CoV-2 from the comparative analysis of genomic data. We offer a perspective on the notable features in the SARS-CoV-2 genome and discuss scenarios by which they could have arisen. Our analyses clearly show that SARS-CoV-2 is not a laboratory construct nor a purposefully manipulated virus.

Notable features of the SARS-CoV-2 genome

Our comparison of alpha- and betacoronaviruses identifies two notable genomic features of SARS-CoV-2: (*i*) based on structural studies[7–9] and biochemical experiments[1,9,10], SARS-CoV-2 appears optimized for binding to the human ACE2 receptor; (*ii*) the spike (S) protein of SARS-CoV-2 has a functional polybasic (furin) cleavage site at the S1/S2 boundary through the insertion of twelve nucleotides[8]. Additionally, this led to the predicted acquisition of three O-linked glycans around the site.

1. Mutations in the receptor binding domain of SARS-CoV-2

The receptor binding domain (RBD) in the spike protein is the most variable part of the coronavirus genome[1,2]. Six RBD amino acids have been shown to be critical for binding to ACE2 receptors and determining the host range of SARS-like viruses[7]. Using coordinates based on SARS-CoV, they are Y442, L472, N479, D480, T487, and Y4911 corresponding to L455, F486, Q493, S494, N501, and Y505 in SARS-CoV-2[7]. Five of these six residues differ between SARS-CoV-2 and SARS-CoV (**Fig. 1a**). Based on structural studies[7–9] and biochemical experiments[1,9,10], SARS-CoV-2 seems to have an RBD that binds with high affinity to ACE2 from human, ferret, cat, and other species with high receptor homology[7].

While these analyses suggest that SARS-CoV-2 may bind human ACE2 with high affinity, computational analyses predict that the interaction is not ideal[7] and the RBD sequence is different from those shown in SARS-CoV to be optimal for receptor binding[7,11]. Thus, the high affinity binding of the SARS-CoV-2 spike protein to human ACE2 is most likely the result of natural selection on a human or human-like ACE2 permitting another optimal binding solution to arise. This is strong evidence that SARS-CoV-2 is *not* the product of purposeful manipulation.

2. Polybasic furin cleavage site and O-linked glycans

The second notable feature of SARS-CoV-2 is a polybasic cleavage site (RRAR) at the S1/S2 junction, the two subunits of the spike (**Fig. 1b**)[8]. This allows effective cleavage by furin and other proteases and plays a role in determining virus infectivity and host range[12]. In addition, a leading proline is also inserted at this site in SARS-CoV-2; thus, the inserted sequence is PRRA (**Fig. 1b**). The turn created by the proline is predicted to result in the addition of O-linked glycans to S673, T678, and S686 flanking the cleavage site and are unique to SARS-CoV-2 (**Fig. 1b**). Polybasic cleavage sites have not been observed in related "lineage B" betacoronaviruses, although other human betacoronaviruses, including HKU1 (lineage A), have them and predicted O-linked glycans[13]. Given the level of genetic variation in the spike it is likely that SARS-CoV-2-like viruses with partial or full polybasic cleavage sites will be discovered in other species.

The functional consequence of the polybasic cleavage site in SARS-CoV-2 is unknown and it will be important to determine its impact on transmissibility and pathogenesis in animal models. Experiments with SARS-CoV have shown that insertion of a furin cleavage site at the S1/S2 junction enhances cell–cell fusion without affecting virus entry[14]. In addition, efficient cleavage of the MERS-CoV spike enables MERS-like coronaviruses from bats to infect human cells[15]. In avian influenza viruses, rapid replication and transmission in highly dense chicken populations selects for the acquisition of polybasic cleavage sites in the haemagglutinin (HA) protein[16], which serves a similar function as the coronavirus spike protein. Acquisition of polybasic cleavage sites in HA, by insertion or recombination, converts low pathogenicity avian influenza viruses into highly pathogenic forms[16]. The acquisition of polybasic cleavage sites by HA has also been observed after repeated passage in cell culture or through animals[17].

The function of the predicted O-linked glycans is unclear, but they could create a "mucin-like domain" shielding epitopes or key residues on the SARS-CoV-2 spike protein[18]. Several viruses employ mucin-like domains as glycan shields involved in immune evasion[18]. Although prediction of O-linked glycosylation is robust, experimental studies are required to determine if these sites are utilized in SARS-CoV-2.

Theories of SARS-CoV-2 origins

It is improbable that SARS-CoV-2 emerged through laboratory manipulation of a related SARS-like coronavirus. As noted above, the RBD of SARS-CoV-2 is optimized for human ACE2 binding with an efficient solution different from those previously predicted[7,11]. Further, had genetic manipulation had been performed, one of the several reverse genetic systems available for betacoronaviruses would likely have been used[19]. However, the genetic data irrefutably show that SARS-CoV-2 is not derived from any previously used virus backbone[20]. Instead, we propose two scenarios that can plausibly explain the origin of SARS-CoV-2: (*i*) natural selection in an animal host prior to zoonotic transfer, and (*ii*) natural selection in humans following zoonotic transfer. We also discuss whether selection during passage could have given rise to SARS-CoV-2.

1. Natural selection in an animal host prior to zoonotic transfer

As many early cases of COVID-19 were linked to the Huanan market in Wuhan[1,2], it is possible that an animal source was present at this location. Given the similarity of SARS-CoV-2 to bat SARS-like coronaviruses[2], it is likely that bats serve as reservoir hosts for its progenitor. Although RaTG13, sampled from a *Rhinolophus affinis* bat[1], is ~96% identical overall to SARS-CoV-2, its spike diverges in the RBD suggesting that it may not bind efficiently to the human ACE2 receptor (**Fig. 1a**)[7].

Malayan pangolins (*Manis javanica*) illegally imported into Guangdong province contain coronaviruses similar to SARS-CoV-2[21]. Although the RaTG13 bat virus remains the closest relative to SARS-CoV-2 across the genome[1], some pangolin coronaviruses exhibit strong similarity to SARS-CoV-2 in the RBD, including all

March 4, 2020 6:50 PM

six key RBD residues (**Fig. 1**)[21]. This clearly shows that the SARS-CoV-2 spike protein optimized for binding to human-like ACE2 is the result of natural selection.

Neither the bat nor pangolin betacoronaviruses sampled to date have polybasic cleavage sites. Although no animal coronavirus has been identified that is sufficiently similar to have served as the direct SARS-CoV-2 progenitor, the diversity of coronaviruses in bats and other species is massively undersampled. Mutations, insertions and deletions, can occur near the S1/S2 junction of coronaviruses[22] showing that the polybasic cleavage site can arise by a natural evolutionary process. For a precursor virus to acquire both the polybasic cleavage site and mutations in the spike protein suitable for human ACE2 receptor binding, an animal host would likely have to have a high population density – to allow natural selection to proceed efficiently – and an ACE2 gene that is similar to the human orthologue.

2. Natural selection in humans following zoonotic transfer

It is possible that a progenitor to SARS-CoV-2 jumped into humans, acquiring the genomic features described above through adaptation during undetected human-to-human transmission. Once acquired, these adaptations would enable the epidemic to take off, producing a sufficiently large cluster of cases to trigger the surveillance system that detected it[1,2].

All SARS-CoV-2 genomes sequenced so far have the genomic features derived above and are thus derived from a common ancestor that had them too. The presence in pangolins of an RBD very similar to that in SARS-CoV-2 means we can infer this was also likely in the virus that jumped to humans. This leaves the polybasic cleavage site insertion to occur during human-to-human transmission.

Estimates of the timing of the most recent common ancestor of SARS-CoV-2 using current sequence data point to virus emergence in late November to early December 2019[23], compatible with the earliest retrospectively confirmed cases[24]. Hence, this scenario presumes a period of unrecognised transmission in humans between the initial zoonotic event and the acquisition of the polybasic cleavage site. Sufficient opportunity could occur if there had been many prior zoonotic events producing short chains of human-to-human transmission over an extended period. This is essentially the situation for MERS-CoV where all human cases are the result of repeated jumps of the virus from dromedary camels, producing single infections or short transmission chains that eventually resolve, with no adaptation to sustained transmission[25].

Studies of banked human samples could provide information on whether such cryptic spread has occurred. Retrospective serological studies could also be informative and a few such studies have been conducted showing low-level exposures to SARS-like coronaviruses in certain areas of China [26]. Critically, however, these studies could not have distinguished whether exposures were due to prior infections with SARS-CoV, SARS-CoV-2, or other SARS-like coronaviruses. Further serological studies should be conducted to determine the extent of prior human exposure to SARS-CoV-2.

3. 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 in laboratories across the world[27] and there are documented instances of laboratory escapes of SARS-CoV[28]. We must therefore examine the possibility of a inadvertent laboratory release of SARS-CoV-2.

In theory, it is possible that SARS-CoV-2 acquired RBD mutations (**Fig. 1a**) during adaptation to passage in cell culture, as has been observed in studies with SARS-CoV[11]. The finding of SARS-like coronaviruses from pangolins with near-identical RBDs, however, provides a much stronger and parsimonious explanation for how SARS-CoV-2 acquired these via recombination or mutation[19].

March 4, 2020 6:50 PM

The acquisition of both the polybasic cleavage site and predicted O-linked glycans also argues against culture-based scenarios. New polybasic cleavage sites have only been observed after prolonged passage of low pathogenicity avian influenza virus *in vitro* or *in vivo*[17]. Furthermore, a hypothetical generation of SARS-CoV-2 by cell culture or animal passage would have required prior isolation of a progenitor virus with very high genetic similarity, which has not been described. Subsequent generation of a polybasic cleavage site would have then required repeated passage in cell culture or animals with ACE2 receptors similar to humans, but such work has also not previously been described. Finally, the generation of the predicted O-linked glycans is also unlikely to have occured due to cell culture passage, as such features suggest the involvement of an immune system[18].

Conclusions

In the midst of the global COVID-19 public health emergency it is reasonable to wonder why the origins of the epidemic matter. A detailed understanding of how an animal virus jumped species boundaries to infect humans so productively will help in the prevention of future zoonotic events. For example, if SARS-CoV-2 pre-adapted in another animal species then we are at risk of future re-emergence events. In contrast, if the adaptive process occurred in humans, then even if we have repeated zoonotic transfers they are unlikely to take off without the same series of mutations. In addition, identifying the closest animal relatives of SARS-CoV-2 will greatly assist studies of virus function. Indeed, the availability of the RaTG13 bat sequence helped reveal key RBD mutations and the polybasic cleavage site.

The genomic features described here may in part explain the infectiousness and transmissibility of SARS-CoV-2 in humans. Although the evidence shows that SARS-CoV-2 is not a purposefully manipulated virus, it is currently impossible to prove or disprove the other theories of its origin described here. However, since we observe all notable SARS-CoV-2 features - including the optimized RBD and polybasic cleavage site - in related coronaviruses in nature, we do not believe that any type of laboratory-based scenario is plausible.

More scientific data could swing the balance of evidence to favor one hypothesis over another. Obtaining related virus sequences from animal sources would be the most definitive way of revealing virus origins. For example, a future observation of an intermediate or fully formed polybasic cleavage site in an SARS-CoV-2-like virus from animals would lend even further support to the natural selection hypotheses. It would also be helpful to obtain more genetic and functional data about SARS-CoV-2, including animal studies. The identification of a potential intermediate host of SARS-CoV-2, as well as the sequencing of very early cases would similarly be highly informative. Irrespective of the exact mechanisms of how SARS-CoV-2 originated via natural selection, the ongoing surveillance of pneumonia in humans and other animals is clearly of utmost importance.

Acknowledgements

We thank all those who have contributed sequences to the GISAID database (https://www.gisaid.org/) and analyses to Virological.org (http://virological.org/). We thank M. Farzan for discussions. We thank the Wellcome Trust for support. KGA is a Pew Biomedical Scholar and is supported by NIH grant U19AI135995. AR is supported by the Wellcome Trust (Collaborators Award 206298/Z/17/Z – ARTIC network) and the European Research Council (grant agreement no. 725422 – ReservoirDOCS). ECH is supported by an ARC Australian Laureate Fellowship (FL170100022). RFG is supported by NIH grants U19AI135995, U54 HG007480 and U19AI142790.

Competing Interests

RFG is co-founder of Zalgen Labs, a biotechnology company developing countermeasures to emerging viruses. None of the other authors declare any conflicts of interest.

March 4, 2020 6:50 PM

Figure Legends

Figure 1. (**a**) Mutations in contact residues of the SARS-CoV-2 spike protein. The spike protein of SARS-CoV-2 (top) was aligned against the most closely related SARS-like CoVs and SARS-CoV. Key residues in the spike protein that make contact to the ACE2 receptor are marked with blue boxes in both SARS-CoV-2 and the SARS-CoV Urbani strain. (**b**) Acquisition of polybasic cleavage site and O-linked glycans. Both the polybasic cleavage site and the three adjacent predicted O-linked glycans are unique to SARS-CoV-2 and not previously seen in lineage B betacoronaviruses. Sequences shown are from NCBI GenBank, accession numbers MN908947, MN996532, AY278741, KY417146 and MK211376. The pangolin coronavirus sequences are a consensus generated from SRR10168377 and SRR10168378 (NCBI BioProject PRJNA573298)[29,30].

March 4, 2020 6:50 PM

References

1. Zhou, P. *et al. Nature* (2020) doi:10.1038/s41586-020-2012-7.

2. Wu, F. *et al. Nature* (2020) doi:10.1038/s41586-020-2008-3.

3. Gorbalenya, A. E. *et al. bioRxiv* 2020.02.07.937862 (2020) doi:10.1101/2020.02.07.937862.

4. Jiang, S. *et al. Lancet* (2020) doi:10.1016/S0140-6736(20)30419-0.

5. Dong, E., Du, H. & Gardner, L. *Lancet Infect. Dis.* (2020) doi:10.1016/S1473-3099(20)30120-1.

6. Corman, V. M., Muth, D., Niemeyer, D. & Drosten, C. *Adv. Virus Res.* **100**, 163–188 (2018).

7. Wan, Y., Shang, J., Graham, R., Baric, R. S. & Li, F. *J. Virol.* (2020) doi:10.1128/JVI.00127-20.

8. Walls, A. C. *et al. bioRxiv* 2020.02.19.956581 (2020) doi:10.1101/2020.02.19.956581.

9. Wrapp, D. *et al. Science* (2020) doi:10.1126/science.abb2507.

10. Letko, M., Marzi, A. & Munster, V. *Nat Microbiol* (2020) doi:10.1038/s41564-020-0688-y.

11. Sheahan, T. *et al. J. Virol.* **82**, 2274–2285 (2008).

12. Nao, N. *et al. MBio* **8**, (2017).

13. Chan, C.-M. *et al. Exp. Biol. Med.* **233**, 1527–1536 (2008).

14. Follis, K. E., York, J. & Nunberg, J. H. *Virology* **350**, 358–369 (2006).

15. Menachery, V. D. *et al. J. Virol.* (2019) doi:10.1128/JVI.01774-19.

16. Alexander, D. J. & Brown, I. H. *Rev. Sci. Tech.* **28**, 19–38 (2009).

17. Ito, T. *et al. J. Virol.* **75**, 4439–4443 (2001).

18. Bagdonaite, I. & Wandall, H. H. *Glycobiology* **28**, 443–467 (2018).

19. Cui, J., Li, F. & Shi, Z.-L. *Nat. Rev. Microbiol.* **17**, 181–192 (2019).

20. Almazán, F. *et al. Virus Res.* **189**, 262–270 (2014).

21. Zhang, T., Wu, Q. & Zhang, Z. *bioRxiv* 2020.02.19.950253 (2020) doi:10.1101/2020.02.19.950253.

22. Yamada, Y. & Liu, D. X. *J. Virol.* **83**, 8744–8758 (2009).

23. http://virological.org/t/phylodynamic-analysis-90-genomes-12-feb-2020/356 (2020).

March 4, 2020 6:50 PM

24. Huang, C. *et al. Lancet* (2020) doi:10.1016/S0140-6736(20)30183-5.

25. Dudas, G., Carvalho, L. M., Rambaut, A. & Bedford, T. *Elife* **7**, (2018).

26. Wang, N. *et al. Virol. Sin.* **33**, 104–107 (2018).

27. Ge, X.-Y. *et al. Nature* **503**, 535–538 (2013).

28. Lim, P. L. *et al. N. Engl. J. Med.* **350**, 1740–1745 (2004).

29. Wong, M. C., Javornik Cregeen, S. J., Ajami, N. J. & Petrosino, J. F. *bioRxiv* 2020.02.07.939207 (2020) doi:10.1101/2020.02.07.939207.

30. Liu, P., Chen, W. & Chen, J.-P. *Viruses* vol. 11 979 (2019).

March 4, 2020 6:50 PM

The Proximal Origin of SARS-CoV-2

Kristian G. Andersen[1,2*], Andrew Rambaut[3], W. Ian Lipkin[4], Edward C. Holmes[5] & Robert F. Garry[6,7]

[1]Department of Immunology and Microbiology, The Scripps Research Institute, La Jolla, CA 92037, USA.
[2]Scripps Research Translational Institute, La Jolla, CA 92037, USA.
[3]Institute of Evolutionary Biology, University of Edinburgh, Edinburgh, UK.
[4]Center for Infection and Immunity, Mailman School of Public Health of Columbia University, New York, New York, USA.
[5]Marie Bashir Institute for Infectious Diseases and Biosecurity, School of Life and Environmental Sciences and School of Medical Sciences, The University of Sydney, Sydney, Australia.
[6]Tulane University, School of Medicine, Department of Microbiology and Immunology, New Orleans, LA, USA.
[7]Zalgen Labs, LCC, Germantown, MD, USA.

*Corresponding author: andersen@scripps.edu

TO THE EDITOR - Since the first reports of novel pneumonia (COVID-19) in Wuhan, Hubei province, China[1,2] there has been considerable discussion on the origin of the causative virus SARS-CoV-2[3] (also referred to as HCoV-19)[4]. Infections with SARS-CoV-2 are now widespread, and as of 29 February 2020, 86,012 cases have been confirmed in more than 60 countries, with 2,941 deaths[5].

SARS-CoV-2 is the seventh coronavirus known to infect humans. SARS-CoV, MERS-CoV, and SARS-CoV-2 can cause severe disease, whereas HKU1, NL63, OC43 and 229E, are associated with mild symptoms[6]. Herein, we review what can be deduced about the origin of SARS-CoV-2 from the comparative analysis of genomic data. We offer a perspective on the notable features in the SARS-CoV-2 genome and discuss scenarios by which they could have arisen. Our analyses clearly show that SARS-CoV-2 is not a laboratory construct nor a purposefully manipulated virus.

Notable features of the SARS-CoV-2 genome

Our comparison of alpha- and betacoronaviruses identifies two notable genomic features of SARS-CoV-2: (*i*) based on structural studies[7–9] and biochemical experiments[1,9,10], SARS-CoV-2 appears optimized for binding to the human ACE2 receptor; (*ii*) the spike (S) protein of SARS-CoV-2 has a functional polybasic (furin) cleavage site at the S1/S2 boundary through the insertion of twelve nucleotides[8]. Additionally, this led to the predicted acquisition of three O-linked glycans around the site.

1. Mutations in the receptor binding domain of SARS-CoV-2

The receptor binding domain (RBD) in the spike protein is the most variable part of the coronavirus genome[1,2]. Six RBD amino acids have been shown to be critical for binding to ACE2 receptors and determining the host range of SARS-like viruses[7]. Using coordinates based on SARS-CoV, they are Y442, L472, N479, D480, T487, and Y4911 corresponding to L455, F486, Q493, S494, N501, and Y505 in SARS-CoV-2[7]. Five of these six residues differ between SARS-CoV-2 and SARS-CoV (**Fig. 1a**). Based on structural studies[7–9] and biochemical experiments[1,9,10], SARS-CoV-2 seems to have an RBD that binds with high affinity to ACE2 from human, ferret, cat, and other species with high receptor homology[7].

While these analyses suggest that SARS-CoV-2 may bind human ACE2 with high affinity, computational analyses predict that the interaction is not ideal[7] and the RBD sequence is different from those shown in SARS-CoV to be optimal for receptor binding[7,11]. Thus, the high affinity binding of the SARS-CoV-2 spike protein to human ACE2 is most likely the result of natural selection on a human or human-like ACE2 permitting another optimal binding solution to arise. This is strong evidence that SARS-CoV-2 is *not* the product of purposeful manipulation.

March 5, 2020 4:48 AM

2. Polybasic furin cleavage site and O-linked glycans

The second notable feature of SARS-CoV-2 is a polybasic cleavage site (RRAR) at the S1/S2 junction, the two subunits of the spike (**Fig. 1b**)[8]. This allows effective cleavage by furin and other proteases and plays a role in determining virus infectivity and host range[12]. In addition, a leading proline is also inserted at this site in SARS-CoV-2; thus, the inserted sequence is PRRA (**Fig. 1b**). The turn created by the proline is predicted to result in the addition of O-linked glycans to S673, T678, and S686 flanking the cleavage site and are unique to SARS-CoV-2 (**Fig. 1b**). Polybasic cleavage sites have not been observed in related "lineage B" betacoronaviruses, although other human betacoronaviruses, including HKU1 (lineage A), have them and predicted O-linked glycans[13]. Given the level of genetic variation in the spike it is likely that SARS-CoV-2-like viruses with partial or full polybasic cleavage sites will be discovered in other species.

The functional consequence of the polybasic cleavage site in SARS-CoV-2 is unknown and it will be important to determine its impact on transmissibility and pathogenesis in animal models. Experiments with SARS-CoV have shown that insertion of a furin cleavage site at the S1/S2 junction enhances cell–cell fusion without affecting virus entry[14]. In addition, efficient cleavage of the MERS-CoV spike enables MERS-like coronaviruses from bats to infect human cells[15]. In avian influenza viruses, rapid replication and transmission in highly dense chicken populations selects for the acquisition of polybasic cleavage sites in the haemagglutinin (HA) protein[16], which serves a similar function as the coronavirus spike protein. Acquisition of polybasic cleavage sites in HA, by insertion or recombination, converts low pathogenicity avian influenza viruses into highly pathogenic forms[16]. The acquisition of polybasic cleavage sites by HA has also been observed after repeated passage in cell culture or through animals[17].

The function of the predicted O-linked glycans is unclear, but they could create a "mucin-like domain" shielding epitopes or key residues on the SARS-CoV-2 spike protein[18]. Several viruses employ mucin-like domains as glycan shields involved in immune evasion[18]. Although prediction of O-linked glycosylation is robust, experimental studies are required to determine if these sites are utilized in SARS-CoV-2.

Theories of SARS-CoV-2 origins

It is improbable that SARS-CoV-2 emerged through laboratory manipulation of a related SARS-like coronavirus. As noted above, the RBD of SARS-CoV-2 is optimized for human ACE2 binding with an efficient solution different from those previously predicted[7,11]. Further, had genetic manipulation had been performed, one of the several reverse genetic systems available for betacoronaviruses would likely have been used[19]. However, the genetic data irrefutably show that SARS-CoV-2 is not derived from any previously used virus backbone[20]. Instead, we propose two scenarios that can plausibly explain the origin of SARS-CoV-2: (*i*) natural selection in an animal host prior to zoonotic transfer, and (*ii*) natural selection in humans following zoonotic transfer. We also discuss whether selection during passage could have given rise to SARS-CoV-2.

1. Natural selection in an animal host prior to zoonotic transfer

As many early cases of COVID-19 were linked to the Huanan market in Wuhan[1,2], it is possible that an animal source was present at this location. Given the similarity of SARS-CoV-2 to bat SARS-like coronaviruses[2], it is likely that bats serve as reservoir hosts for its progenitor. Although RaTG13, sampled from a *Rhinolophus affinis* bat[1], is ~96% identical overall to SARS-CoV-2, its spike diverges in the RBD suggesting that it may not bind efficiently to the human ACE2 receptor (**Fig. 1a**)[7].

Malayan pangolins (*Manis javanica*) illegally imported into Guangdong province contain coronaviruses similar to SARS-CoV-2[21]. Although the RaTG13 bat virus remains the closest relative to SARS-CoV-2 across the genome[1], some pangolin coronaviruses exhibit strong similarity to SARS-CoV-2 in the RBD, including all

March 5, 2020 4:48 AM

six key RBD residues (**Fig. 1**)[21]. This clearly shows that the SARS-CoV-2 spike protein optimized for binding to human-like ACE2 is the result of natural selection.

Neither the bat nor pangolin betacoronaviruses sampled to date have polybasic cleavage sites. Although no animal coronavirus has been identified that is sufficiently similar to have served as the direct SARS-CoV-2 progenitor, the diversity of coronaviruses in bats and other species is massively undersampled. Mutations, insertions and deletions, can occur near the S1/S2 junction of coronaviruses[22] showing that the polybasic cleavage site can arise by a natural evolutionary process. For a precursor virus to acquire both the polybasic cleavage site and mutations in the spike protein suitable for human ACE2 receptor binding, an animal host would likely have to have a high population density – to allow natural selection to proceed efficiently – and an ACE2 gene that is similar to the human orthologue.

2. Natural selection in humans following zoonotic transfer

It is possible that a progenitor to SARS-CoV-2 jumped into humans, acquiring the genomic features described above through adaptation during undetected human-to-human transmission. Once acquired, these adaptations would enable the epidemic to take off, producing a sufficiently large cluster of cases to trigger the surveillance system that detected it[1,2].

All SARS-CoV-2 genomes sequenced so far have the genomic features derived above and are thus derived from a common ancestor that had them too. The presence in pangolins of an RBD very similar to that in SARS-CoV-2 means we can infer this was also likely in the virus that jumped to humans. This leaves the polybasic cleavage site insertion to occur during human-to-human transmission.

Estimates of the timing of the most recent common ancestor of SARS-CoV-2 using current sequence data point to virus emergence in late November to early December 2019[23], compatible with the earliest retrospectively confirmed cases[24]. Hence, this scenario presumes a period of unrecognised transmission in humans between the initial zoonotic event and the acquisition of the polybasic cleavage site. Sufficient opportunity could occur if there had been many prior zoonotic events producing short chains of human-to-human transmission over an extended period. This is essentially the situation for MERS-CoV where all human cases are the result of repeated jumps of the virus from dromedary camels, producing single infections or short transmission chains that eventually resolve, with no adaptation to sustained transmission[25].

Studies of banked human samples could provide information on whether such cryptic spread has occurred. Retrospective serological studies could also be informative and a few such studies have been conducted showing low-level exposures to SARS-like coronaviruses in certain areas of China [26]. Critically, however, these studies could not have distinguished whether exposures were due to prior infections with SARS-CoV, SARS-CoV-2, or other SARS-like coronaviruses. Further serological studies should be conducted to determine the extent of prior human exposure to SARS-CoV-2.

3. 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 in laboratories across the world[27] and there are documented instances of laboratory escapes of SARS-CoV[28]. We must therefore examine the possibility of a inadvertent laboratory release of SARS-CoV-2.

In theory, it is possible that SARS-CoV-2 acquired RBD mutations (**Fig. 1a**) during adaptation to passage in cell culture, as has been observed in studies with SARS-CoV[11]. The finding of SARS-like coronaviruses from pangolins with near-identical RBDs, however, provides a much stronger and parsimonious explanation for how SARS-CoV-2 acquired these via recombination or mutation[19].

March 5, 2020 4:48 AM

The acquisition of both the polybasic cleavage site and predicted O-linked glycans also argues against culture-based scenarios. New polybasic cleavage sites have only been observed after prolonged passage of low pathogenicity avian influenza virus *in vitro* or *in vivo*[17]. Furthermore, a hypothetical generation of SARS-CoV-2 by cell culture or animal passage would have required prior isolation of a progenitor virus with very high genetic similarity, which has not been described. Subsequent generation of a polybasic cleavage site would have then required repeated passage in cell culture or animals with ACE2 receptors similar to humans, but such work has also not previously been described. Finally, the generation of the predicted O-linked glycans is also unlikely to have occured due to cell culture passage, as such features suggest the involvement of an immune system[18].

Conclusions

In the midst of the global COVID-19 public health emergency it is reasonable to wonder why the origins of the epidemic matter. A detailed understanding of how an animal virus jumped species boundaries to infect humans so productively will help in the prevention of future zoonotic events. For example, if SARS-CoV-2 pre-adapted in another animal species then we are at risk of future re-emergence events. In contrast, if the adaptive process occurred in humans, then even if we have repeated zoonotic transfers they are unlikely to take off without the same series of mutations. In addition, identifying the closest animal relatives of SARS-CoV-2 will greatly assist studies of virus function. Indeed, the availability of the RaTG13 bat sequence helped reveal key RBD mutations and the polybasic cleavage site.

The genomic features described here may in part explain the infectiousness and transmissibility of SARS-CoV-2 in humans. Although the evidence shows that SARS-CoV-2 is not a purposefully manipulated virus, it is currently impossible to prove or disprove the other theories of its origin described here. However, since we observe all notable SARS-CoV-2 features - including the optimized RBD and polybasic cleavage site - in related coronaviruses in nature, we do not believe that any type of laboratory-based scenario is plausible.

More scientific data could swing the balance of evidence to favor one hypothesis over another. Obtaining related virus sequences from animal sources would be the most definitive way of revealing virus origins. For example, a future observation of an intermediate or fully formed polybasic cleavage site in an SARS-CoV-2-like virus from animals would lend even further support to the natural selection hypotheses. It would also be helpful to obtain more genetic and functional data about SARS-CoV-2, including animal studies. The identification of a potential intermediate host of SARS-CoV-2, as well as the sequencing of very early cases would similarly be highly informative. Irrespective of the exact mechanisms of how SARS-CoV-2 originated via natural selection, the ongoing surveillance of pneumonia in humans and other animals is clearly of utmost importance.

March 5, 2020 4:48 AM

Acknowledgements

We thank all those who have contributed sequences to the GISAID database (https://www.gisaid.org/) and analyses to Virological.org (http://virological.org/). We thank M. Farzan for discussions. We thank the Wellcome Trust for support. KGA is a Pew Biomedical Scholar and is supported by NIH grant U19AI135995. AR is supported by the Wellcome Trust (Collaborators Award 206298/Z/17/Z – ARTIC network) and the European Research Council (grant agreement no. 725422 – ReservoirDOCS). ECH is supported by an ARC Australian Laureate Fellowship (FL170100022). RFG is supported by NIH grants U19AI135995, U54 HG007480 and U19AI142790.

Competing Interests

RFG is co-founder of Zalgen Labs, a biotechnology company developing countermeasures to emerging viruses. None of the other authors declare any conflicts of interest.

Figure Legends

Figure 1. (**a**) Mutations in contact residues of the SARS-CoV-2 spike protein. The spike protein of SARS-CoV-2 (top) was aligned against the most closely related SARS-like CoVs and SARS-CoV. Key residues in the spike protein that make contact to the ACE2 receptor are marked with blue boxes in both SARS-CoV-2 and the SARS-CoV Urbani strain. (**b**) Acquisition of polybasic cleavage site and O-linked glycans. Both the polybasic cleavage site and the three adjacent predicted O-linked glycans are unique to SARS-CoV-2 and not previously seen in lineage B betacoronaviruses. Sequences shown are from NCBI GenBank, accession numbers MN908947, MN996532, AY278741, KY417146 and MK211376. The pangolin coronavirus sequences are a consensus generated from SRR10168377 and SRR10168378 (NCBI BioProject PRJNA573298)[29,30].

March 5, 2020 4:48 AM

References

1. Zhou, P. *et al. Nature* (2020) doi:10.1038/s41586-020-2012-7.
2. Wu, F. *et al. Nature* (2020) doi:10.1038/s41586-020-2008-3.
3. Gorbalenya, A. E. *et al. bioRxiv* 2020.02.07.937862 (2020) doi:10.1101/2020.02.07.937862.
4. Jiang, S. *et al. Lancet* (2020) doi:10.1016/S0140-6736(20)30419-0.
5. Dong, E., Du, H. & Gardner, L. *Lancet Infect. Dis.* (2020) doi:10.1016/S1473-3099(20)30120-1.
6. Corman, V. M., Muth, D., Niemeyer, D. & Drosten, C. *Adv. Virus Res.* **100**, 163–188 (2018).
7. Wan, Y., Shang, J., Graham, R., Baric, R. S. & Li, F. *J. Virol.* (2020) doi:10.1128/JVI.00127-20.
8. Walls, A. C. *et al. bioRxiv* 2020.02.19.956581 (2020) doi:10.1101/2020.02.19.956581.
9. Wrapp, D. *et al. Science* (2020) doi:10.1126/science.abb2507.
10. Letko, M., Marzi, A. & Munster, V. *Nat Microbiol* (2020) doi:10.1038/s41564-020-0688-y.
11. Sheahan, T. *et al. J. Virol.* **82**, 2274–2285 (2008).
12. Nao, N. *et al. MBio* **8**, (2017).
13. Chan, C.-M. *et al. Exp. Biol. Med.* **233**, 1527–1536 (2008).
14. Follis, K. E., York, J. & Nunberg, J. H. *Virology* **350**, 358–369 (2006).
15. Menachery, V. D. *et al. J. Virol.* (2019) doi:10.1128/JVI.01774-19.
16. Alexander, D. J. & Brown, I. H. *Rev. Sci. Tech.* **28**, 19–38 (2009).
17. Ito, T. *et al. J. Virol.* **75**, 4439–4443 (2001).
18. Bagdonaite, I. & Wandall, H. H. *Glycobiology* **28**, 443–467 (2018).
19. Cui, J., Li, F. & Shi, Z.-L. *Nat. Rev. Microbiol.* **17**, 181–192 (2019).
20. Almazán, F. *et al. Virus Res.* **189**, 262–270 (2014).
21. Zhang, T., Wu, Q. & Zhang, Z. *bioRxiv* 2020.02.19.950253 (2020) doi:10.1101/2020.02.19.950253.
22. Yamada, Y. & Liu, D. X. *J. Virol.* **83**, 8744–8758 (2009).
23. http://virological.org/t/phylodynamic-analysis-90-genomes-12-feb-2020/356 (2020).

March 5, 2020 4:48 AM

24. Huang, C. *et al. Lancet* (2020) doi:10.1016/S0140-6736(20)30183-5.

25. Dudas, G., Carvalho, L. M., Rambaut, A. & Bedford, T. *Elife* **7**, (2018).

26. Wang, N. *et al. Virol. Sin.* **33**, 104–107 (2018).

27. Ge, X.-Y. *et al. Nature* **503**, 535–538 (2013).

28. Lim, P. L. *et al. N. Engl. J. Med.* **350**, 1740–1745 (2004).

29. Wong, M. C., Javornik Cregeen, S. J., Ajami, N. J. & Petrosino, J. F. *bioRxiv* 2020.02.07.939207 (2020) doi:10.1101/2020.02.07.939207.

30. Liu, P., Chen, W. & Chen, J.-P. *Viruses* vol. 11 979 (2019).

March 5, 2020 4:48 AM

The Proximal Origin of SARS-CoV-2

Kristian G. Andersen[1,2*], Andrew Rambaut[3], W. Ian Lipkin[4], Edward C. Holmes[5] & Robert F. Garry[6,7]

[1]Department of Immunology and Microbiology, The Scripps Research Institute, La Jolla, CA 92037, USA.

[2]Scripps Research Translational Institute, La Jolla, CA 92037, USA.

[3]Institute of Evolutionary Biology, University of Edinburgh, Edinburgh, UK.

[4]Center for Infection and Immunity, Mailman School of Public Health of Columbia University, New York, New York, USA.

[5]Marie Bashir Institute for Infectious Diseases and Biosecurity, School of Life and Environmental Sciences and School of Medical Sciences, The University of Sydney, Sydney, Australia.

[6]Tulane University, School of Medicine, Department of Microbiology and Immunology, New Orleans, LA, USA.

[7]Zalgen Labs, LCC, Germantown, MD, USA.

*Corresponding author:

Kristian G. Andersen

Department of Immunology and Microbiology,

The Scripps Research Institute,

La Jolla, CA 92037, USA.

version before revisions for resubmission

Since the first reports of a novel pneumonia (COVID-19) in Wuhan city, Hubei province, China there has been considerable discussion and uncertainty over the origin of the causative virus, SARS-CoV-2. Infections with SARS-CoV-2 are now widespread in China, with cases in every province. As of 14 February 2020, 64,473 such cases have been confirmed, with 1,384 deaths attributed to the virus. These official case numbers are likely an underestimate because of limited reporting of mild and asymptomatic cases, and the virus is clearly capable of efficient human-to-human transmission. Based on the possibility of spread to countries with weaker healthcare systems, the World Health Organization has declared the COVID-19 outbreak a Public Health Emergency of International Concern (PHEIC). There are currently neither vaccines nor specific treatments for this disease.

SARS-CoV-2 is the seventh member of the *Coronaviridae* known to infect humans. Three of these viruses, SARS CoV-1, MERS CoV, and SARS-CoV-2, can cause severe disease; four, HKU1, NL63, OC43 and 229E, are associated with mild respiratory symptoms. Herein, we review what can be deduced about the origin and early evolution of SARS-CoV-2 from the comparative analysis of available genome sequence data. In particular, we offer a perspective on the notable features in the SARS-CoV-2 genome and discuss scenarios by which these features could have arisen. Importantly, this analysis provides evidence that SARS-CoV-2 is not a laboratory construct nor a purposefully manipulated virus.

The genomic comparison of both alpha- and betacoronaviruses (family *Coronaviridae*) described below identifies two notable features of the SARS-CoV-2 genome: (i) based on structural modeling and early biochemical experiments, SARS-CoV-2 appears to be optimized for binding to the human ACE2 receptor; (ii) the highly variable spike (S) protein of SARS-CoV-2 has a polybasic (furin) cleavage site at the S1 and S2 boundary via the insertion of twelve nucleotides. Additionally, this event led to the acquisition of three predicted O-linked glycans around the polybasic cleavage site.

Mutations in the receptor binding domain of SARS-CoV-2

The receptor binding domain (RBD) in the spike protein of SARS-CoV and SARS-related coronaviruses is the most variable part of the virus genome. Six residues in the RBD appear to be critical for binding to the human ACE2 receptor and determining host range[1]. Using coordinates based on the Urbani strain of SARS-CoV, they are Y442, L472, N479, D480, T487, and Y491[1]. The corresponding residues in SARS-CoV-2 are L455, F486, Q493, S494, N501, and Y505. Five of these six residues are mutated in SARS-CoV-2 compared to its most closely related virus, RaTG13 sampled from a *Rhinolophus affinis* bat, to which it is ~96% identical[2] (**Figure 1a**). Based on modeling[1] and biochemical experiments[3,4], SARS-CoV-2 seems to have an RBD that may bind with high affinity to ACE2 from human, non-human primate, ferret, pig, and cat, as well as other species with high receptor homology[1]. In contrast, SARS-CoV-2 may bind less efficiently to ACE2 in other species associated with SARS-like viruses, including rodents and civets[1].

The phenylalanine (F) at residue 486 in the SARS-CoV-2 S protein corresponds to L472 in the SARS-CoV Urbani strain. Notably, in SARS-CoV cell culture experiments the L472 mutates to phenylalanine (L472F)[5], which is predicted to be optimal for binding of the SARS-CoV RBD to the human ACE2 receptor[6]. However, a phenylalanine in this position is also present in several SARS-like CoVs from bats (**Figure 1a**). While these analyses suggest that SARS-CoV-2 may be capable of binding the human ACE2 receptor with high affinity, the interaction is not predicted to be optimal[1]. Additionally, several of the key residues in the RBD of SARS-CoV-2 are different to those previously described as optimal for human ACE2 receptor binding[6]. In contrast to these computational predictions, recent binding studies indicate that SARS-CoV-2 binds with high affinity to human ACE2[7]. Thus the SARS-CoV-2 spike appears to be the result of selection on human or human-like ACE2 permitting another optimal binding solution to arise. This is strong evidence that SARS-CoV-2 is *not* the product of genetic engineering.

Polybasic cleavage site and O-linked glycans

The second notable feature of SARS-CoV-2 is a predicted polybasic cleavage site (RRAR) in the spike protein at the junction of S1 and S2, the two subunits of the spike protein (**Figure 1b**)[8,9]. In addition to two basic arginines and an alanine at the cleavage site, a leading proline is also inserted; thus, the fully inserted sequence is PRRA (**Figure 1b**). The strong turn created by the proline insertion is predicted to result in the addition of O-linked glycans to S673, T678, and S686 that flank the polybasic cleavage site. A polybasic cleavage site has not previously been observed in related lineage B betacoronaviruses and is a unique feature of SARS-CoV-2. Some human betacoronaviruses, including HCoV-HKU1 (lineage A), have polybasic cleavage sites, as well as predicted O-linked glycans near the S1/S2 cleavage site.

While the functional consequence of the polybasic cleavage site in SARS-CoV-2 is unknown, experiments with SARS-CoV have shown that engineering such a site at the S1/S2 junction enhances cell–cell fusion but does not affect virus entry[10]. Polybasic cleavage sites allow effective cleavage by furin and other proteases, and can be acquired at the junction of the two subunits of the haemagglutinin (HA) protein of avian influenza viruses in conditions that select for rapid virus replication and transmission (e.g.,. highly dense chicken populations). HA serves a similar function in cell-cell fusion and viral entry as the coronavirus S protein. Acquisition of a polybasic cleavage site in HA, by either insertion or recombination, converts low pathogenicity avian influenza viruses into highly pathogenic forms[11–13]. The acquisition of polybasic

Version before revisions for resubmission

cleavage sites by the influenza virus HA has also been observed after repeated forced passage in cell culture or through animals[14,15]. Similarly, an avirulent isolate of Newcastle Disease virus became highly pathogenic during serial passage in chickens by incremental acquisition of a polybasic cleavage site at the junction of its fusion protein subunits[16]. The potential function of the three predicted O-linked glycans is less clear, but they could create a "mucin-like domain" that would shield potential epitopes or key residues on the SARS-CoV-2 spike protein[17,18]. Although the algorithms for prediction of O-linked glycosylation are robust[19], biochemical analyses or structural studies are required to determine whether or not the predicted O-linked glycan sites are utilized.

Theories of SARS-CoV-2 origins

It is improbable that SARS-CoV-2 emerged through laboratory manipulation of an existing SARS-related coronavirus. As noted above, the RBD of SARS-CoV-2 is optimized for human ACE2 receptor binding with an efficient binding solution different to that which would have been predicted. Further, if genetic manipulation had been performed, one would expect that one of the several reverse genetic systems available for betacoronaviruses would have been used. However, this is not the case as the genetic data shows that SARS-CoV-2 is not derived from any previously used virus backbone[20]. Instead, we propose two scenarios that can plausibly explain the origin of SARS-CoV-2: (i) natural selection in a non-human animal host prior to zoonotic transfer, and (ii) natural selection in humans following zoonotic transfer. We also discuss whether selection during passage in culture could have given rise to the same observed features.

Selection in an animal host

As many of the early cases of COVID-19 were linked to the Huanan seafood and wildlife market in Wuhan, it is possible that an animal source was present at this location. Given the similarity of SARS-CoV-2 to bat SARS-like CoVs, particularly RaTG13, it is plausible that bats serve as reservoir hosts for SARS-CoV-2. It is important, however, to note that previous outbreaks of betacoronaviruses in humans involved direct exposure to animals other than bats, including civets (SARS) and camels (MERS), that carry viruses that are genetically very similar to SARS-CoV or MERS-CoV, respectively. By analogy, viruses closely related to SARS-Cov-2 may be circulating in one or more animal species. Initial analyses indicate that Malayan pangolins (*Manis javanica*) illegally imported into Guangdong province contain a CoV that is similar to SARS-CoV-2[21,22]. Although the bat virus RaTG13 remains the closest relative to SARS-CoV-2 across the whole genome, the Malayan pangolin CoV is identical to SARS-CoV-2 at all six key RBD residues (**Figure 1**). However, no pangolin CoV has yet been identified that is sufficiently similar to SARS-CoV-2 across its entire genome to support direct human infection. In addition, the pangolin CoV does not carry a polybasic cleavage site insertion. For a precursor virus to acquire the polybasic cleavage site and mutations in the spike protein suitable for human ACE2 receptor binding, an animal host would likely have to have a high population density – to allow natural selection to proceed efficiently – and an ACE2 gene that is similar to the human orthologue. Further characterization of CoVs in pangolins and other animals that may harbour SARS-CoV-like viruses should be a public health priority.

Cryptic adaptation to humans

It is also possible that a progenitor to SARS-CoV-2 jumped from a non-human animal to humans, with the genomic features described above acquired through adaptation during subsequent human-to-human transmission. We surmise that once these adaptations were acquired (either together or in series) it would enable the outbreak to take-off, producing a sufficiently large and unusual cluster of pneumonia cases to trigger the surveillance system that ultimately detected it.

All SARS-CoV-2 genomes sequenced so far have the well adapted RBD and the polybasic cleavage site and are thus derived from a common ancestor that had these features. The presence of an RBD in pangolins that is very similar to the one in SARS-CoV-2 means that this was likely already present in the virus that

Version before revisions for resubmission

jumped to humans, even if we don't yet have the exact non-human progenitor virus. This leaves the polybasic cleavage site insertion to occur during human-to-human transmission. Following the example of the influenza A virus HA gene, a specific insertion or recombination event is required to enable the emergence of SARS-CoV-2 as an epidemic pathogen.

Estimates of the timing of the most recent common ancestor (tMRCA) of SARS-CoV-2 using currently available genome sequence data point to virus emergence in late November to early December 2019[23,24], compatible with the earliest retrospectively confirmed cases[25]. Hence, this scenario presumes a period of unrecognised transmission in humans between the initial zoonotic transfer event and the acquisition of the polybasic cleavage site. Sufficient opportunity could occur if there had been many prior zoonotic events producing short chains of human-to-human transmission (so-called 'stuttering chains') over an extended period. This is essentially the situation for MERS-CoV in the Arabian Peninsula where all the human cases are the result of repeated jumps of the virus from dromedary camels, producing single infections or short chains of transmission that eventually resolve. To date, after 2,499 cases over 8 years, no human adaptation has emerged that has allowed MERS-CoV to take hold in the human population.

How could we test whether cryptic spread of SARS-CoV-2 enabled human adaptation? Metagenomic studies of banked serum samples could provide important information, but given the relatively short period of viremia it may be impossible to detect low level SARS-CoV-2 circulation in historical samples. Retrospective serological studies potentially could be informative and a few such studies have already been conducted. One found that animal importation traders had a 13% seropositivity to coronaviruses[26], while another noted that 3% residents of a village in Southern China were seropositive to these viruses[27]. Interestingly, 200 residents of Wuhan did not show coronavirus seroreactivity. Critically, however, these studies could not have distinguished whether positive serological responses were due to a prior infection with SARS-CoV or SARS-CoV-2. Further retrospective serological studies should be conducted to determine the extent of prior human exposure to betacoronaviruses in different geographic areas, particularly using assays that can distinguish among multiple betacoronaviruses.

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 in multiple laboratories across the world[28–31]. There are also documented instances of the laboratory acquisition of SARS-CoV by laboratory personnel working under BSL-2 containment[32,33]. We must therefore consider the possibility of a deliberate or inadvertent release of SARS-CoV-2. In theory, it is possible that SARS-CoV-2 acquired the observed RBD mutations site during adaptation to passage in cell culture, as has been observed in studies with SARS-CoV[5] as well as MERS-CoV[34]. However, the acquisition of the polybasic cleavage site or O-linked glycans - if functional - argues against this scenario. New polybasic cleavage sites have only been observed after prolonged passaging of low pathogenicity avian influenza virus in cell culture or animals. Furthermore, the generation of SARS-CoV-2 by cell culture or animal passage would have required prior isolation of a progenitor virus with a very high genetic similarity. Subsequent generation of a polybasic cleavage site would have then required an intense program of passage in cell culture or animals with ACE-2 receptor similar to humans (e.g. ferrets). It is also questionable whether generation of the O-linked glycans would have occurred on cell culture passage, as such mutations typically suggest the involvement of an immune system, that is not present *in vitro*.

Conclusions

In the midst of the global COVID-19 public health emergency it is reasonable to wonder why the origins of the epidemic matter. A detailed understanding of how an animal virus jumped species boundaries to infect humans so productively will help in the prevention of future zoonotic events. For example, if SARS-CoV-2

Version before revisions for resubmission

pre-adapted in another animal species then we are at risk of future re-emergence events even if the current epidemic is controlled. In contrast, if the adaptive process we describe occurred in humans, then even if we have repeated zoonotic transfers they are unlikely to take-off unless the same series of mutations occurs. In addition, identifying the closest animal relatives of SARS-CoV-2 will greatly assist studies of virus function. Indeed, the availability of the RaTG13 bat sequence facilitated the comparative genomic analysis performed here, helping to reveal the key mutations in the RBD as well as the polybasic cleavage site insertion.

The genomic features described here may in part explain the infectiousness and transmissibility of SARS-CoV-2 in humans. Although genomic evidence does not support the idea that SARS-CoV-2 is a laboratory construct, it is currently impossible to prove or disprove the other theories of its origin described here, and it is unclear whether future data 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 revealing virus origins. In addition, it would be helpful to obtain more genetic and functional data about the virus, including experimental studies of receptor binding and the role of the polybasic cleavage site and predicted O-linked glycans. The identification of a potential intermediate host of SARS-CoV-2, as well as the sequencing of very early cases including those not connected to the Wuhan market, would similarly be highly informative. Irrespective of how SARS-CoV-2 originated, the ongoing surveillance of pneumonia in humans and other animals is clearly of utmost importance.

Acknowledgements

We thank all those who have contributed SARS-CoV-2 genome sequences to the GISAID database (https://www.gisaid.org/) and contributed analyses and ideas to Virological.org (http://virological.org/). We thank M. Farzan for discussions. We thank the Wellcome Trust for supporting this work. ECH is supported by an ARC Australian Laureate Fellowship (FL170100022). KGA is a Pew Biomedical Scholar and is supported by NIH NIAID grant U19AI135995. AR is supported by the Wellcome Trust (Collaborators Award 206298/Z/17/Z – ARTIC network) and the European Research Council (grant agreement no. 725422 – ReservoirDOCS). RFG is supported by NIH grants U19AI135995, U54 HG007480 and U19AI142790.

Figure Legends

Figure 1 | a) Mutations in contact residues of the SARS-CoV-2 spike protein. The spike protein of SARS-CoV-2 (top) was aligned against the most closely related SARS-like CoVs and SARS-CoV-1. Key residues in the spike protein that make contact to the ACE2 receptor are marked with blue boxes in both SARS-CoV-2 and the SARS-CoV Urbani strain. **b) Acquisition of polybasic cleavage site and O-linked glycans**. Both the polybasic cleavage site and the three adjacent predicted O-linked glycans are unique to SARS-CoV-2 and not previously seen in lineage B betacoronaviruses. Sequences shown are from NCBI GenBank, accession numbers MN908947, MN996532, AY278741, KY417146 & MK211376. The pangolin coronavirus sequences are a consensus generated from SRR10168377 and SRR10168378 (NCBI BioProject PRJNA573298)[21,35].

Version before revisions for resubmission

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. Wu, F. *et al.* A new coronavirus associated with human respiratory disease in China. *Nature* (2020) doi:10.1038/s41586-020-2008-3.

3. 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.

4. 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.

5. 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).

6. Cui, J., Li, F. & Shi, Z.-L. Origin and evolution of pathogenic coronaviruses. *Nat. Rev. Microbiol.* **17**, 181–192 (2019).

7. Wrapp, D. *et al.* Cryo-EM Structure of the 2019-nCoV Spike in the Prefusion Conformation. *bioRxiv* 2020.02.11.944462 (2020) doi:10.1101/2020.02.11.944462.

8. Gallaher, W. Analysis of Wuhan Coronavirus: Deja Vu. *Virological* http://virological.org/t/analysis-of-wuhan-coronavirus-deja-vu/357 (2020).

9. Coutard, B. *et al.* The spike glycoprotein of the new coronavirus 2019-nCoV contains a furin-like cleavage site absent in CoV of the same clade. *Antiviral Res.* 104742 (2020).

10. 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).

Version before revisions for resubmission

11. 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).
12. Alexander, D. J. & Brown, I. H. History of highly pathogenic avian influenza. *Rev. Sci. Tech.* **28**, 19–38 (2009).
13. 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).
14. Ito, T. *et al.* Generation of a highly pathogenic avian influenza A virus from an avirulent field isolate by passaging in chickens. *J. Virol.* **75**, 4439–4443 (2001).
15. Li, S. Q., Orlich, M. & Rott, R. Generation of seal influenza virus variants pathogenic for chickens, because of hemagglutinin cleavage site changes. *J. Virol.* **64**, 3297–3303 (1990).
16. Shengqing, Y. *et al.* Generation of velogenic Newcastle disease viruses from a nonpathogenic waterfowl isolate by passaging in chickens. *Virology* **301**, 206–211 (2002).
17. Bagdonaite, I. & Wandall, H. H. Global aspects of viral glycosylation. *Glycobiology* **28**, 443–467 (2018).
18. Tran, E. E. H. *et al.* Spatial localization of the Ebola virus glycoprotein mucin-like domain determined by cryo-electron tomography. *J. Virol.* **88**, 10958–10962 (2014).
19. Steentoft, C. *et al.* Precision mapping of the human O-GalNAc glycoproteome through SimpleCell technology. *EMBO J.* **32**, 1478–1488 (2013).
20. Menachery, V. D. *et al.* A SARS-like cluster of circulating bat coronaviruses shows potential for human emergence. *Nat. Med.* **21**, 1508–1513 (2015).
21. Wong, M. C., Javornik Cregeen, S. J., Ajami, N. J. & Petrosino, J. F. Evidence of recombination in coronaviruses implicating pangolin origins of nCoV-2019. *bioRxiv* 2020.02.07.939207 (2020) doi:10.1101/2020.02.07.939207.
22. Lam, T. T.-Y. *et al.* Identification of 2019-nCoV related coronaviruses in Malayan pangolins in southern China. doi:10.1101/2020.02.13.945485.

Version before revisions for resubmission

23. Phylodynamic Analysis | 90 genomes | 12 Feb 2020 – Novel 2019 coronavirus / nCoV-2019 Genomic Epidemiology. http://virological.org/t/phylodynamic-analysis-90-genomes-12-feb-2020/356 (2020).

24. Phylodynamic estimation of incidence and prevalence of novel coronavirus (nCoV) infections through time – Novel 2019 coronavirus / nCoV-2019 Genomic Epidemiology. http://virological.org/t/phylodynamic-estimation-of-incidence-and-prevalence-of-novel-coronavirus-ncov-infections-through-time/391 (2020).

25. Huang, C. *et al.* Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet* (2020) doi:10.1016/S0140-6736(20)30183-5.

26. Centers for Disease Control and Prevention (CDC). Prevalence of IgG antibody to SARS-associated coronavirus in animal traders--Guangdong Province, China, 2003. *MMWR Morb. Mortal. Wkly. Rep.* **52**, 986–987 (2003).

27. Wang, N. *et al.* Serological Evidence of Bat SARS-Related Coronavirus Infection in Humans, China. *Virol. Sin.* **33**, 104–107 (2018).

28. Ge, X.-Y. *et al.* Isolation and characterization of a bat SARS-like coronavirus that uses the ACE2 receptor. *Nature* **503**, 535–538 (2013).

29. 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).

30. 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).

31. 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).

32. Lim, P. L. *et al.* Laboratory-acquired severe acute respiratory syndrome. *N. Engl. J. Med.* **350**, 1740–1745 (2004).

Version before revisions for resubmission

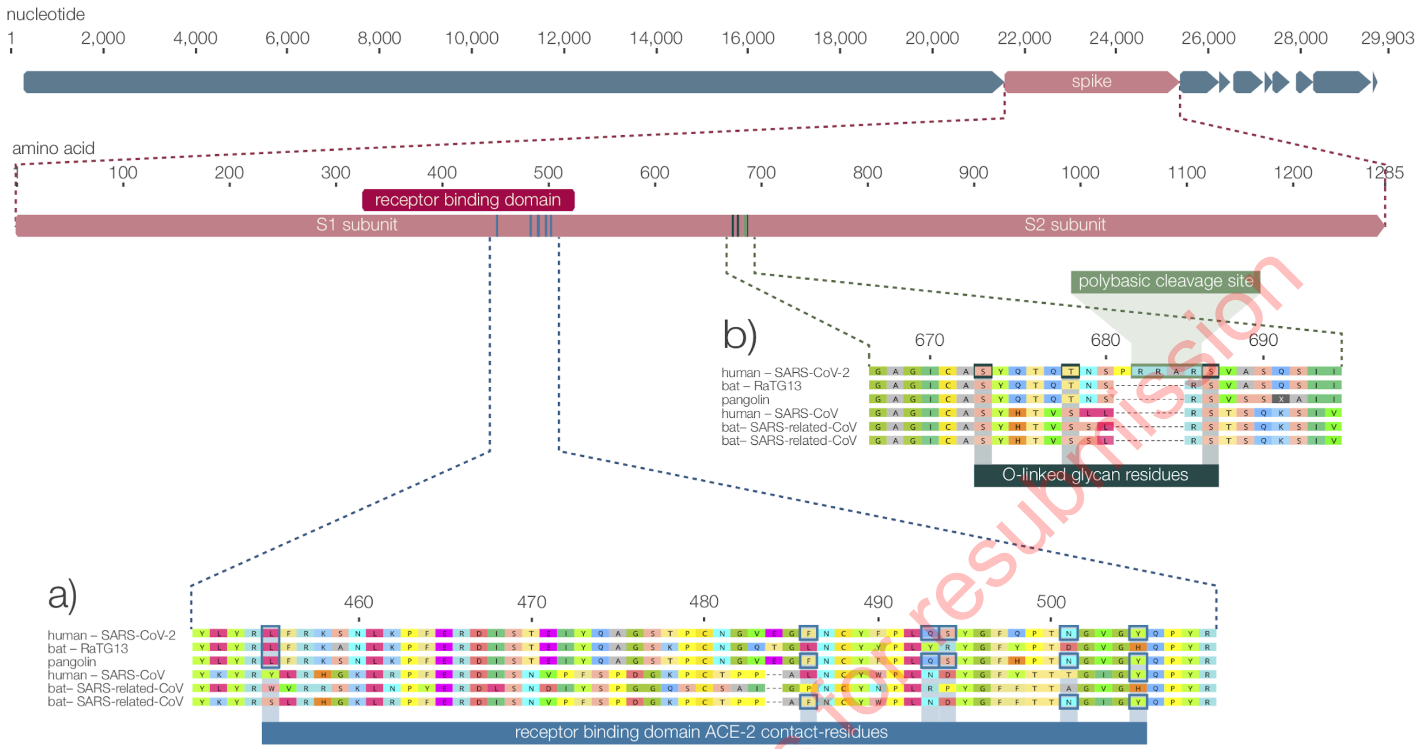
33. Senior, K. Recent Singapore SARS case a laboratory accident. *Lancet Infect. Dis.* **3**, 679 (2003).

34. Letko, M. *et al.* Adaptive Evolution of MERS-CoV to Species Variation in DPP4. *Cell Rep.* **24**, 1730–1737 (2018).

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

Version before revisions for resubmission

Figure 1.


nucleotide
1
2,000
4,000
6,000
8,000
10,000
12,000
14,000
16,000
18,000
20,000
22,000
24,000
26,000
28,000
29,903
spike
amino acid
100
200
300
400
500
600
700
800
900
1000
1100
1200
1285
receptor binding domain
S1 subunit
S2 subunit
polybasic cleavage site
b)
670
680
690
human – SARS-CoV-2 G A G I C A S Y Q T Q T N S P R R A R S V A S Q S I I
bat – RaTG13 G A G I C A S Y Q T Q T N S ---- R S V A S Q S I I
pangolin G A G I C A S Y Q T Q T N S ---- R S V S S X A I I
human – SARS-CoV G A G I C A S Y H T V S L L ---- R S T S Q K S I V
bat– SARS-related-CoV G A G I C A S Y H T V S S L ---- R S T S Q K S I V
bat– SARS-related-CoV G A G I C A S Y H T V S S L ---- R S T S Q K S I V
O-linked glycan residues
a)
460
470
480
490
500
human – SARS-CoV-2 Y L Y R L F R K S N L K P F E R D I S T E I Y Q A G S T P C N G V E G F N C Y F P L Q S Y G F Q P T N G V G Y Q P Y R
bat – RaTG13 Y L Y R L F R K A N L K P F E R D I S T E I Y Q A G S K P C N G Q T G L N C Y Y P L Y R Y G F Y P T D G V G H Q P Y R
pangolin Y L Y R L F R K S N L K P F E R D I S T E I Y Q A G S T P C N G V E G F N C Y F P L Q S Y G F H P T N G V G Y Q P Y R
human – SARS-CoV Y K Y R Y L R H G K L R P F E R D I S N V P F S P D G K P C T P P --- A L N C Y W P L N D Y G F Y T T T G I G Y Q P Y R
bat– SARS-related-CoV Y L Y R W V R R S K L N P Y E R D L S N D I Y S P G G Q S C S A I --- G P N C Y N P L R P Y G F F T T A G V G H Q P Y R
bat– SARS-related-CoV Y K Y R S L R H G K L R P F E R D I S N V P F S P D G K P C T P P --- A F N C Y W P L N D Y G F F T T N G I G Y Q P Y R
receptor binding domain ACE-2 contact-residues

Version before revisions for resubmission

The Proximal Origin of HCoV-19

Kristian G. Andersen[1,2*], Andrew Rambaut[3], W. Ian Lipkin[4], Edward C. Holmes[5] & Robert F. Garry[6,7]

[1]Department of Immunology and Microbiology, The Scripps Research Institute, La Jolla, CA 92037, USA.
[2]Scripps Research Translational Institute, La Jolla, CA 92037, USA.
[3]Institute of Evolutionary Biology, University of Edinburgh, Edinburgh, UK.
[4]Center for Infection and Immunity, Mailman School of Public Health of Columbia University, New York, New York, USA.
[5]Marie Bashir Institute for Infectious Diseases and Biosecurity, School of Life and Environmental Sciences and School of Medical Sciences, The University of Sydney, Sydney, Australia.
[6]Tulane University, School of Medicine, Department of Microbiology and Immunology, New Orleans, LA, USA.
[7]Zalgen Labs, LCC, Germantown, MD, USA.

*Corresponding author:
Kristian G. Andersen
Department of Immunology and Microbiology,
The Scripps Research Institute,
La Jolla, CA 92037, USA.

Since the first reports of a novel pneumonia (coronavirus disease 2019; COVID-19) in Wuhan city, Hubei province, China[1,2] there has been considerable discussion and uncertainty over the origin of the causative virus, human coronavirus 2019 (HCoV-19[3]; also referred to as SARS-CoV-2[4]). Infections with HCoV-19 are now widespread across the world and as of 29 February 2020, 86,012 COVID-19 cases have been confirmed in 57 countries, with 2,941 deaths attributed to the virus[5]. These numbers likely represent an underestimate because of limited reporting of mild and asymptomatic cases, and the virus is clearly capable of efficient human-to-human transmission. Based on the possibility of spread to countries with weaker healthcare systems, the World Health Organization has declared the COVID-19 epidemic a Public Health Emergency of International Concern (PHEIC)[6].

HCoV-19 is the seventh member of the *Coronaviridae* known to infect humans. Three of these viruses, SARS-CoV, MERS-CoV, and HCoV-19 can cause severe disease; four, HKU1, NL63, OC43 and 229E, are associated with mild respiratory symptoms[7]. Herein, we review what can be deduced about the origin and early evolution of HCoV-19 from the comparative analysis of available genome sequence data. In particular, we offer a perspective on the notable features in the HCoV-19 genome and discuss scenarios by which these features could have arisen. Importantly, our analysis provides strong evidence that HCoV-19 is not a laboratory construct nor a purposefully manipulated virus.

Notable features of the HCoV-19 genome

Our genomic comparison of both alpha- and betacoronaviruses (family *Coronaviridae*) described below identifies two notable features of the HCoV-19 genome: (*i*) based on structural studies[8–10] and early biochemical experiments[1,9,11,12], HCoV-19 appears to be optimized for binding to the human ACE2 receptor; (*ii*) the highly variable spike (S) protein of HCoV-19 has a functional polybasic (furin) cleavage site at the S1 and S2 boundary via the insertion of twelve nucleotides[10,13,14]. Additionally, this event led to the acquisition of three predicted O-linked glycans around the polybasic cleavage site.

Mutations in the receptor binding domain of HCoV-19

The receptor binding domain (RBD) in the spike protein of HCoV-19 and SARS-related coronaviruses is the most variable part of the virus genome[1,2]. Six amino acids in the RBD have been shown to be critical for binding to the human ACE2 receptor and determining the host range of SARS-like viruses[8]. Using coordinates based on the Urbani strain of SARS-CoV, they are Y442, L472, N479, D480, T487, and Y4911. The corresponding residues in HCoV-19 are L455, F486, Q493, S494, N501, and Y505[8]. Five of these six

Before Cutting for NatMed

residues differ in the RBDs of HCoV-19 and SARS-CoV (**Fig. 1a**). Based on structural studies[8–10] and biochemical experiments[1,9,11,12], HCoV-19 seems to have an RBD that bind with high affinity to ACE2 from human, non-human primate, ferret, pig, and cat, as well as other species with high receptor homology[8]. In contrast, HCoV-19 may bind less efficiently to ACE2 in other species associated with SARS-like viruses, including rodents and civets[8,15].

While these analyses suggest that HCoV-19 may be capable of binding the human ACE2 receptor with high affinity, computational analyses predict that the interaction is not ideal[8] and the RBD sequence is different from those previously shown in SARS-CoV to be optimal for receptor binding[16]. Thus, the optimized binding of the HCoV-19 spike protein to the human ACE2 receptor is most likely the result of natural selection on a human or human-like ACE2 permitting another optimal binding solution to arise. This is strong evidence that HCoV-19 is *not* the product of a purposefully manipulated virus, a widely propagated baseless conspiracy theory[17].

Polybasic furin cleavage site and O-linked glycans

The second notable feature of HCoV-19 is a polybasic furin cleavage site (RRAR) in the S protein at the junction of S1 and S2, the two subunits of the spike (**Fig. 1b**)[13,14]. Polybasic cleavage sites allow effective cleavage by furin and other proteases and play important roles in determining virus infectivity and host range[18]. In addition to two basic arginines and an alanine at the cleavage site, a leading proline is also inserted around this site in HCoV-19; thus, the fully inserted sequence is PRRA (**Fig. 1b**). The strong turn created by the proline insertion is predicted to result in the addition of O-linked glycans to S673, T678, and S686 that flank the polybasic cleavage site. While these three residues are conserved in related viruses from bats and pangolins, they are not predicted to have O-linked glycans, because of the lack of the leading proline (**Fig. 1b**). A polybasic cleavage site has not previously been observed in related "lineage B" betacoronaviruses, however, other human betacoronaviruses, including HKU1 (lineage A), have polybasic cleavage sites, as well as predicted O-linked glycans near the S1/S2 cleavage site[19]. While the furin cleavage site is a unique feature of HCoV-19 (**Fig. 1b**), given how common genetic variation is in the S protein of betacoronaviruses[20], we expect that bats or intermediate hosts harboring HCoV-19-like viruses with partial or full polybasic sites will be identified as part of the flurry of research ongoing in response to the COVID-19 epidemic.

The functional consequence of the furin cleavage site in HCoV-19 is unknown and it will be important to determine what impact, if any, the site has on transmissibility and pathogenesis in animal models[21]. Experiments with SARS-CoV have shown that insertion of a furin cleavage site at the S1/S2 junction enhances cell–cell fusion, but does not affect virus entry[22]. In addition, efficient cleavage of the MERS-CoV spike protein has been shown to enable MERS-like coronaviruses from bats to infect human cells[23]. In avian influenza viruses, polybasic cleavage sites can be acquired at the S1/S2 junction of the haemagglutinin (HA) protein under conditions that select for rapid virus replication and transmission, such as when the virus is replicating in highly dense chicken populations[24–26]. HA serves a similar function in cell-cell fusion and viral entry as the coronavirus spike protein. Acquisition of a polybasic cleavage site in HA, by either insertion or recombination, converts low pathogenicity avian influenza viruses into highly pathogenic forms[24–26]. The acquisition of polybasic cleavage sites by the influenza virus HA has also been observed after repeated passage in cell culture or through animals[27,28]. Similarly, an avirulent isolate of Newcastle Disease virus became highly pathogenic during serial passage in chickens by incremental acquisition of a polybasic cleavage site at the junction of its fusion protein subunits[29].

The potential function of the three predicted O-linked glycans is less clear, but they could create a "mucin-like domain" that would shield potential epitopes or key residues on the HCoV-19 spike protein[30,31]. Several viruses employ mucin-like domains as part of a glycan shield that is involved in immune evasion[30]. Although

Before Cutting for NatMed

the algorithms for prediction of O-linked glycosylation are robust[32], biochemical analyses or structural studies are required to determine whether or not the predicted O-linked glycan sites in HCoV-19 are utilized.

Theories of HCoV-19 origins

It is improbable that HCoV-19 emerged through laboratory manipulation or engineering of a related SARS-like coronavirus. As noted above, the RBD of HCoV-19 is optimized for human ACE2 receptor binding with an efficient solution that is different from those previously predicted[8,16]. Further, if genetic manipulation had been performed, one would expect that one of the several reverse genetic systems available for betacoronaviruses would have been used[20]. This is not the case, however, as the genetic data irrefutably show that HCoV-19 is not derived from any previously used virus backbone[33]. Instead, we propose two scenarios that can plausibly explain the origin of HCoV-19: (***i***) natural selection in a non-human animal host prior to zoonotic transfer, and (***ii***) natural selection in humans following zoonotic transfer. We also discuss whether selection during passage in culture could have given rise to the same observed features and conclude that such a scenario is unlikely.

Natural selection in an animal host prior to zoonotic transfer

As many of the early cases of COVID-19 were linked to the Huanan seafood and wildlife market in Wuhan[1,2], it is possible that an animal source was present at this location. Given the similarity of HCoV-19 to bat SARS-like coronaviruses[2], particularly RaTG13[1], it is highly plausible that bats serve as reservoir hosts for its progenitor. Although RaTG13 sampled from a *Rhinolophus affinis* bat is ~96% identical overall to HCoV-19[1], its S protein possesses distinct sequences in the RBD suggesting that it may not bind efficiently to the human ACE2 receptor (**Fig. 1a**)[8]. It is also important to note that previous human outbreaks of betacoronaviruses involved direct exposure to animals other than bats, including civets and camels, which carry viruses that are genetically very similar to SARS-CoV[34] or MERS-CoV[35,36], respectively. By analogy, viruses closely related to HCoV-19 may be circulating in one or more animal species.

Malayan pangolins (*Manis javanica*) illegally imported into Guangdong province contain coronaviruses that are similar to HCoV-19[15,37–39]. Although the RaTG13 bat virus remains the closest relative to HCoV-19 across the whole genome[1], a Malayan pangolin coronavirus exhibits strong similarity to HCoV-19 in the RBD, including all six key RBD residues (**Fig. 1**)[15,39]. This finding clearly shows that the optimized binding of the HCoV-19 spike protein to human ACE2 can occur in nature and is the result of natural selection. However, no coronavirus from a pangolin, nor any other animal species, has yet been identified that is sufficiently similar to HCoV-19 across its entire genome that it could have served as the direct progenitor of the virus. Similarly, although there is likely a history of complex recombination events in these viruses[20], including in the RBD and other domains of the S protein, none of the available bat or pangolin coronaviruses are sufficiently similar to HCoV-19 to have directly generated it by recombination.

Neither the bat nor pangolin betacoronaviruses sampled to date carry polybasic cleavage sites. However, as the diversity of coronaviruses in bats and other species is massively undersampled, it is possible that an animal betacoronavirus will eventually be identified with a polybasic cleavage site. Mutations, including point mutations, insertions and deletions, can occur near the S1/S2 junction of coronaviruses[34,40–43] suggesting that the polybasic site could arise by a natural evolutionary process. For a precursor virus to acquire both the polybasic cleavage site and mutations in the spike protein suitable for human ACE2 receptor binding, an animal host would likely have to have a high population density – to allow natural selection to proceed efficiently – and an ACE2 gene that is similar to the human orthologue. Further characterization of viruses in animals that may harbour SARS-like coronaviruses should be a public health priority.

Before Cutting for NatMed

Natural selection in humans following zoonotic transfer

It is also possible that a progenitor to HCoV-19 jumped from an animal to humans, with the genomic features described above acquired through adaptation during subsequent (undetected) human-to-human transmission. We surmise that once these adaptations were acquired (either together or in series) it would enable the epidemic to take off, producing a sufficiently large and unusual cluster of pneumonia cases to trigger the surveillance system that ultimately detected it[1,2].

All HCoV-19 genomes sequenced so far have the well adapted RBD and the polybasic cleavage site and are thus derived from a common ancestor that had these features. The presence of an RBD in pangolins that is very similar to the one in HCoV-19 means that this was likely already present in the virus that jumped to humans, even if we do not yet have the exact non-human progenitor virus. This leaves the polybasic cleavage site insertion to occur during human-to-human transmission. Following the example of the influenza A virus HA gene, a specific insertion or recombination event is required to enable the emergence of HCoV-19 as an epidemic pathogen.

Estimates of the timing of the most recent common ancestor (tMRCA) of HCoV-19 using currently available genome sequence data point to virus emergence in late November to early December 2019[44–46], compatible with the earliest retrospectively confirmed cases[47]. Hence, this scenario presumes a period of unrecognised transmission in humans between the initial zoonotic transfer event and the acquisition of the polybasic cleavage site. Sufficient opportunity could occur if there had been many prior zoonotic events producing short chains of human-to-human transmission (so-called 'stuttering chains') over an extended period. This is essentially the situation for MERS-CoV in the Arabian Peninsula where all the human cases are the result of repeated jumps of the virus from dromedary camels, producing single infections or short chains of transmission that eventually resolve[48]. To date, after 2,499 cases over eight years, no human adaptation has emerged that has allowed MERS-CoV to take hold in the human population.

How could we test whether cryptic spread of HCoV-19 enabled human adaptation? Metagenomic studies of banked human samples could provide important information, but given the relatively short period of viremia it may be impossible to detect low level HCoV-19 circulation in historical samples. Retrospective serological studies could also be informative and a few such studies have already been conducted. One found that animal importation traders had a 13% seropositivity to coronaviruses[49], while another noted that 3% residents of a village in Southern China were seropositive to SARS-like coronaviruses[50]. Critically, however, these studies could not have distinguished whether positive serological responses were due to prior infections with SARS-CoV, HCoV-19, or other SARS-like coronaviruses. Further serological studies should be conducted to determine the extent of prior human exposure to HCoV-19 and other betacoronaviruses before, during, and after the COVID-19 epidemic.

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 in laboratories across the world[51–54]. There are also several documented instances of laboratory escapes of SARS-CoV[55–57]. We must therefore examine the possibility of a inadvertent laboratory release of HCoV-19.

In theory, it is possible that HCoV-19 acquired RBD mutations (**Fig. 1a**) during adaptation to passage in cell culture, as has been observed in studies with SARS-CoV[16] and MERS-CoV[58]. The finding of SARS-like coronaviruses from pangolins with near-identical RBDs, however, provides a much stronger and parsimonious explanation for how HCoV-19 acquired these via recombination or mutation[20].

The acquisition of both the polybasic cleavage site and predicted O-linked glycans also argues against any type of culture-based scenario. New polybasic cleavage sites have only been observed after prolonged

Before Cutting for NatMed

passage of low pathogenicity avian influenza virus *in vitro* or *in vivo*[24,26–28]. Furthermore, a hypothetical generation of HCoV-19 by cell culture or animal passage would have required prior isolation of a progenitor virus with a very high genetic similarity, which has not been described. Subsequent generation of a polybasic cleavage site would have then required repeated passage in cell culture or animals with ACE2 receptors similar to humans (e.g. ferrets), but that type of work has also not previously been described. Finally, the generation of the predicted O-linked glycans is also unlikely to have occured due to cell culture passage, as such features suggest the involvement of an immune system[30], which is obviously not present *in vitro*.

Conclusions

In the midst of the global COVID-19 public health emergency it is reasonable to wonder why the origins of the epidemic matter. A detailed understanding of how an animal virus jumped species boundaries to infect humans so productively will help in the prevention of future zoonotic events. For example, if HCoV-19 pre-adapted in another animal species then we are at risk of future re-emergence events even if the current epidemic is controlled. In contrast, if the adaptive process we describe occurred in humans, then even if we have repeated zoonotic transfers they are unlikely to take-off unless the same series of mutations occurs. In addition, identifying the closest animal relatives of HCoV-19 will greatly assist studies of virus function. Indeed, the availability of the RaTG13 bat sequence facilitated the comparative genomic analysis performed here, helping to reveal the key mutations in the RBD as well as the polybasic cleavage site insertion.

The genomic features described here may in part explain the infectiousness and transmissibility of HCoV-19 in humans. Although the evidence shows that HCoV-19 is not a laboratory construct nor a purposefully manipulated virus, it is currently impossible to prove or disprove the other theories of its origin described here. However, since we observe all notable HCoV-19 features - including the optimized RBD and furin cleavage site - in highly related coronaviruses in nature, we do not believe that selection during passage, or any other type of laboratory-based scenario, is plausible.

While it is unclear whether future data will fully resolve this issue, it is likely that more scientific data will swing the balance of evidence to favor one hypothesis over another. Identifying the immediate non-human animal source and obtaining virus sequences from it would be the most definitive way of revealing virus origins. For example, a future observation of an intermediate or fully formed polybasic cleavage site in an HCoV-19 related virus from animals would lend even stronger support to our natural selection hypotheses described above. Given the immense amount of research ongoing in response to the COVID-19 epidemic, we are hopeful that such data may be obtained in the near future. In addition, it would be helpful to obtain more genetic and functional data about HCoV-19, including experimental studies of receptor binding and the role of the polybasic cleavage site and predicted O-linked glycans. The identification of a potential intermediate host of HCoV-19, as well as the sequencing of very early cases, including those not connected to the Wuhan market, would similarly be highly informative. Irrespective of the exact mechanisms of how HCoV-19 originated via natural selection, the ongoing surveillance of pneumonia in humans and other animals is clearly of utmost importance.

Acknowledgements

We thank all those who have contributed HCoV-19 genome sequences to the GISAID database (https://www.gisaid.org/) and contributed analyses and ideas to Virological.org (http://virological.org/). We thank M. Farzan for discussions. We thank the Wellcome Trust for supporting this work. KGA is a Pew Biomedical Scholar and is supported by NIH NIAID grant U19AI135995. AR is supported by the Wellcome Trust (Collaborators Award 206298/Z/17/Z – ARTIC network) and the European Research Council (grant

Before Cutting for NatMed

agreement no. 725422 – ReservoirDOCS). ECH is supported by an ARC Australian Laureate Fellowship (FL170100022). RFG is supported by NIH grants U19AI135995, U54 HG007480 and U19AI142790.

Before Cutting for NatMed

Figure Legends

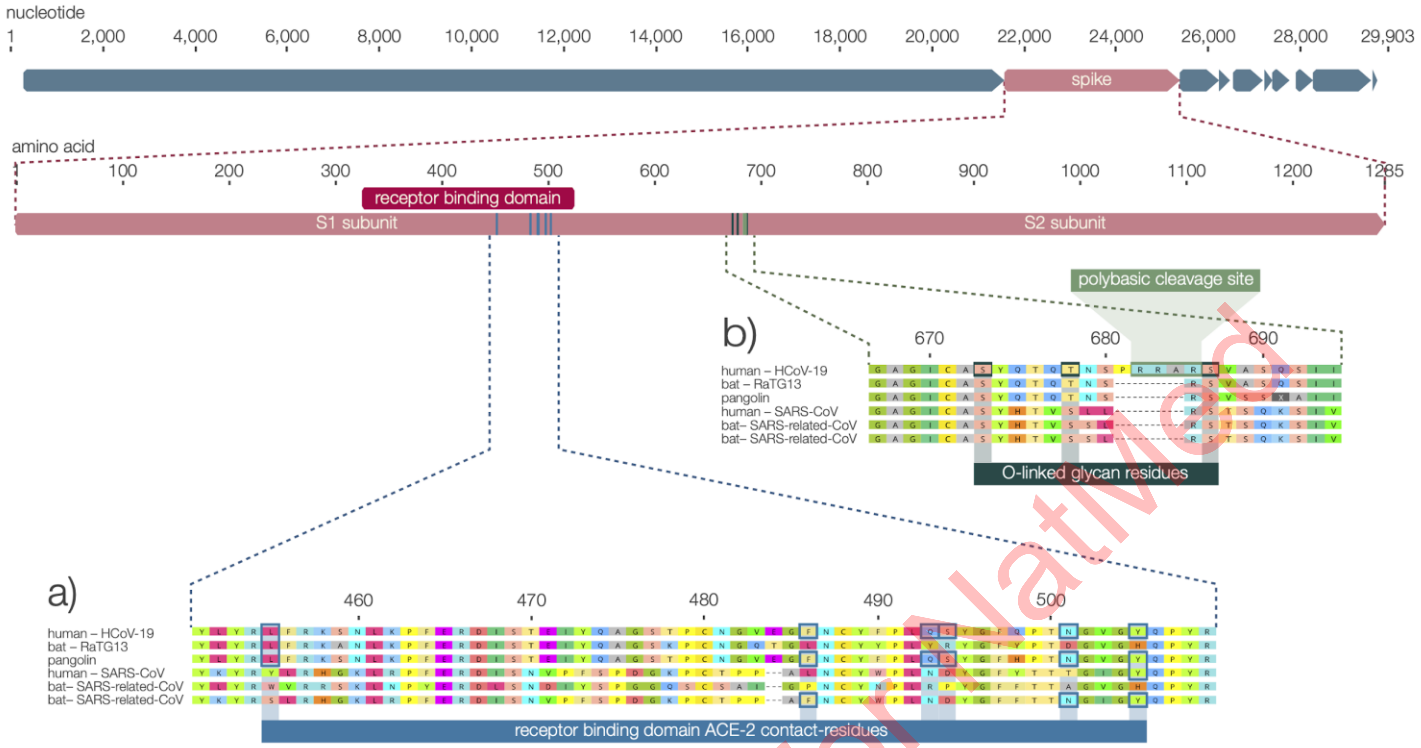


Figure 1. (**a**) Mutations in contact residues of the HCoV-19 spike protein. The spike protein of HCoV-19 (top) was aligned against the most closely related SARS-like CoVs and SARS-CoV. Key residues in the spike protein that make contact to the ACE2 receptor are marked with blue boxes in both HCoV-19 and the SARS-CoV Urbani strain. (**b**) Acquisition of polybasic cleavage site and O-linked glycans. Both the polybasic cleavage site and the three adjacent predicted O-linked glycans are unique to HCoV-19 and not previously seen in lineage B betacoronaviruses. Sequences shown are from NCBI GenBank, accession numbers MN908947, MN996532, AY278741, KY417146 and MK211376. The pangolin coronavirus sequences are a consensus generated from SRR10168377 and SRR10168378 (NCBI BioProject PRJNA573298)[37,59].

Before Cutting for NatMed

References

1. Zhou, P. *et al.* A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* (2020) doi:10.1038/s41586-020-2012-7.
2. Wu, F. *et al.* A new coronavirus associated with human respiratory disease in China. *Nature* (2020) doi:10.1038/s41586-020-2008-3.
3. Jiang, S. *et al.* A distinct name is needed for the new coronavirus. *Lancet* (2020) doi:10.1016/S0140-6736(20)30419-0.
4. Gorbalenya, A. E. Severe acute respiratory syndrome-related coronavirus–The species and its viruses, a statement of the Coronavirus Study Group. *BioRxiv* (2020).
5. Dong, E., Du, H. & Gardner, L. An interactive web-based dashboard to track COVID-19 in real time. *Lancet Infect. Dis.* (2020) doi:10.1016/S1473-3099(20)30120-1.
6. Statement on the second meeting of the International Health Regulations (2005) Emergency Committee regarding the outbreak of novel coronavirus (2019-nCoV). https://www.who.int/news-room/detail/30-01-2020-statement-on-the-second-meeting-of-the-international-health-regulations-(2005)-emergency-committee-regarding-the-outbreak-of-novel-coronavirus-(2019-ncov).
7. Corman, V. M., Muth, D., Niemeyer, D. & Drosten, C. Hosts and Sources of Endemic Human Coronaviruses. *Adv. Virus Res.* **100**, 163–188 (2018).
8. 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.
9. Wrapp, D. *et al.* Cryo-EM Structure of the 2019-nCoV Spike in the Prefusion Conformation. *bioRxiv* 2020.02.11.944462 (2020) doi:10.1101/2020.02.11.944462.
10. Walls, A. C. *et al.* Structure, function and antigenicity of the SARS-CoV-2 spike glycoprotein. *bioRxiv* 2020.02.19.956581 (2020) doi:10.1101/2020.02.19.956581.

Before Cutting for NatMed

11. Letko, M., Marzi, A. & Munster, V. Functional assessment of cell entry and receptor usage for SARS-CoV-2 and other lineage B betacoronaviruses. *Nat Microbiol* (2020) doi:10.1038/s41564-020-0688-y.

12. 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.

13. Gallaher, W. Analysis of Wuhan Coronavirus: Deja Vu. *Virological* http://virological.org/t/analysis-of-wuhan-coronavirus-deja-vu/357 (2020).

14. Coutard, B. *et al.* The spike glycoprotein of the new coronavirus 2019-nCoV contains a furin-like cleavage site absent in CoV of the same clade. *Antiviral Res.* 104742 (2020).

15. Zhang, T., Wu, Q. & Zhang, Z. Pangolin homology associated with 2019-nCoV. *bioRxiv* 2020.02.19.950253 (2020) doi:10.1101/2020.02.19.950253.

16. 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).

17. Calisher, C. *et al.* Statement in support of the scientists, public health professionals, and medical professionals of China combatting COVID-19. *Lancet* (2020) doi:10.1016/S0140-6736(20)30418-9.

18. Nao, N. *et al.* Genetic Predisposition To Acquire a Polybasic Cleavage Site for Highly Pathogenic Avian Influenza Virus Hemagglutinin. *MBio* **8**, (2017).

19. Chan, C.-M. *et al.* Spike protein, S, of human coronavirus HKU1: role in viral life cycle and application in antibody detection. *Exp. Biol. Med.* **233**, 1527–1536 (2008).

20. Cui, J., Li, F. & Shi, Z.-L. Origin and evolution of pathogenic coronaviruses. *Nat. Rev. Microbiol.* **17**, 181–192 (2019).

21. Bao, L. *et al.* The Pathogenicity of SARS-CoV-2 in hACE2 Transgenic Mice. *bioRxiv* 2020.02.07.939389 (2020) doi:10.1101/2020.02.07.939389.

22. Follis, K. E., York, J. & Nunberg, J. H. Furin cleavage of the SARS coronavirus spike glycoprotein

Before Cutting for NatMed

enhances cell-cell fusion but does not affect virion entry. *Virology* **350**, 358–369 (2006).

23. Menachery, V. D. *et al.* Trypsin treatment unlocks barrier for zoonotic bat coronaviruses infection. *J. Virol.* (2019) doi:10.1128/JVI.01774-19.
24. 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).
25. Alexander, D. J. & Brown, I. H. History of highly pathogenic avian influenza. *Rev. Sci. Tech.* **28**, 19–38 (2009).
26. 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).
27. Ito, T. *et al.* Generation of a highly pathogenic avian influenza A virus from an avirulent field isolate by passaging in chickens. *J. Virol.* **75**, 4439–4443 (2001).
28. Li, S. Q., Orlich, M. & Rott, R. Generation of seal influenza virus variants pathogenic for chickens, because of hemagglutinin cleavage site changes. *J. Virol.* **64**, 3297–3303 (1990).
29. Shengqing, Y. *et al.* Generation of velogenic Newcastle disease viruses from a nonpathogenic waterfowl isolate by passaging in chickens. *Virology* **301**, 206–211 (2002).
30. Bagdonaite, I. & Wandall, H. H. Global aspects of viral glycosylation. *Glycobiology* **28**, 443–467 (2018).
31. Tran, E. E. H. *et al.* Spatial localization of the Ebola virus glycoprotein mucin-like domain determined by cryo-electron tomography. *J. Virol.* **88**, 10958–10962 (2014).
32. Steentoft, C. *et al.* Precision mapping of the human O-GalNAc glycoproteome through SimpleCell technology. *EMBO J.* **32**, 1478–1488 (2013).
33. Almazán, F. *et al.* Coronavirus reverse genetic systems: infectious clones and replicons. *Virus Res.* **189**, 262–270 (2014).
34. Guan, Y. *et al.* Isolation and characterization of viruses related to the SARS coronavirus from animals in southern China. *Science* **302**, 276–278 (2003).

Before Cutting for NatMed

35. Haagmans, B. L. *et al.* Middle East respiratory syndrome coronavirus in dromedary camels: an outbreak investigation. *Lancet Infect. Dis.* **14**, 140–145 (2014).

36. Azhar, E. I. *et al.* Evidence for camel-to-human transmission of MERS coronavirus. *N. Engl. J. Med.* **370**, 2499–2505 (2014).

37. Wong, M. C., Javornik Cregeen, S. J., Ajami, N. J. & Petrosino, J. F. Evidence of recombination in coronaviruses implicating pangolin origins of nCoV-2019. *bioRxiv* 2020.02.07.939207 (2020) doi:10.1101/2020.02.07.939207.

38. Lam, T. T.-Y. *et al.* Identification of 2019-nCoV related coronaviruses in Malayan pangolins in southern China. *bioRxiv* (2020) doi:10.1101/2020.02.13.945485.

39. Xiao, K. *et al.* Isolation and Characterization of 2019-nCoV-like Coronavirus from Malayan Pangolins. *bioRxiv* 2020.02.17.951335 (2020) doi:10.1101/2020.02.17.951335.

40. Licitra, B. N. *et al.* Mutation in spike protein cleavage site and pathogenesis of feline coronavirus. *Emerg. Infect. Dis.* **19**, 1066–1073 (2013).

41. Yamada, Y. & Liu, D. X. Proteolytic activation of the spike protein at a novel RRRR/S motif is implicated in furin-dependent entry, syncytium formation, and infectivity of coronavirus infectious bronchitis virus in cultured cells. *J. Virol.* **83**, 8744–8758 (2009).

42. Yamada, Y. K., Takimoto, K., Yabe, M. & Taguchi, F. Requirement of proteolytic cleavage of the murine coronavirus MHV-2 spike protein for fusion activity. *Adv. Exp. Med. Biol.* **440**, 89–93 (1998).

43. Lamers, M. M. *et al.* Deletion Variants of Middle East Respiratory Syndrome Coronavirus from Humans, Jordan, 2015. *Emerg. Infect. Dis.* **22**, 716–719 (2016).

44. Phylodynamic Analysis | 90 genomes | 12 Feb 2020 – Novel 2019 coronavirus / nCoV-2019 Genomic Epidemiology. http://virological.org/t/phylodynamic-analysis-90-genomes-12-feb-2020/356 (2020).

45. Phylodynamic estimation of incidence and prevalence of novel coronavirus (nCoV) infections through time – Novel 2019 coronavirus / nCoV-2019 Genomic Epidemiology.

Before Cutting for NatMed

http://virological.org/t/phylodynamic-estimation-of-incidence-and-prevalence-of-novel-coronavirus-ncov-infections-through-time/391 (2020).

46. Clock and TMRCA based on 27 genomes. *Virological* http://virological.org/t/clock-and-tmrca-based-on-27-genomes/347 (2020).

47. Huang, C. *et al.* Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet* (2020) doi:10.1016/S0140-6736(20)30183-5.

48. Dudas, G., Carvalho, L. M., Rambaut, A. & Bedford, T. MERS-CoV spillover at the camel-human interface. *Elife* **7**, (2018).

49. Centers for Disease Control and Prevention (CDC). Prevalence of IgG antibody to SARS-associated coronavirus in animal traders--Guangdong Province, China, 2003. *MMWR Morb. Mortal. Wkly. Rep.* **52**, 986–987 (2003).

50. Wang, N. *et al.* Serological Evidence of Bat SARS-Related Coronavirus Infection in Humans, China. *Virol. Sin.* **33**, 104–107 (2018).

51. Ge, X.-Y. *et al.* Isolation and characterization of a bat SARS-like coronavirus that uses the ACE2 receptor. *Nature* **503**, 535–538 (2013).

52. 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).

53. 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).

54. 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).

55. Lim, P. L. *et al.* Laboratory-acquired severe acute respiratory syndrome. *N. Engl. J. Med.* **350**, 1740–1745 (2004).

Before Cutting for NatMed

56. Senior, K. Recent Singapore SARS case a laboratory accident. *Lancet Infect. Dis.* **3**, 679 (2003).

57. Lim, W., Ng, K.-C. & Tsang, D. N. C. Laboratory containment of SARS virus. *Ann. Acad. Med. Singapore* **35**, 354–360 (2006).

58. Letko, M. *et al.* Adaptive Evolution of MERS-CoV to Species Variation in DPP4. *Cell Rep.* **24**, 1730–1737 (2018).

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

Before Cutting for NatMed