Insertions in human SARS-CoV-2 samples shed light on the origin of the furin cleavage site.

Guy Gadboit

Contents

List of Figures

Abstract

Since the start of the COVID-19 pandemic several million SARS-CoV-2 sequences have been uploaded to GISAID. Some of these sequences contain insertions relative to Wuhan-Hu-1, and some of those insertions match sequences in species of bacteria in the predominant human oral microbiota including with significant sequence identity outside the insertion itself (p=0.00013, p=0.08 for $\geq$3bp and $\geq$6bp of adjacent sequence identity respectively). The furin cleavage site insertion also matches a species of bacterium in this set with 6bp of sequence identity adjacent to the insertion ($E=3\times10^{-4}$). This implies that the particular nucleotides encoding the furin cleavage site have a natural origin not an artificial one, and may also be a clue to how the pandemic started.

1 Introduction

The most notable difference between the genome of SARS-CoV-2 and that of its close relatives is a 12-nucleotide insertion of the sequence CTCCTCGGCGGG at position 23601, which results in a furin cleavage site between the S1 and S2 domains of the spike protein[7].

It is thought to be an important determinant of transmissibility[7] and may have been critical to helping the virus start a pandemic.

It is heavily implicated in concerns that the virus may have had an artificial origin due to a documented research interest in this subject[4] that can be linked to work that was likely taking place in Wuhan, where the pandemic is thought to have started[28].

However, insertions into coronaviruses like this do happen naturally, and many have occurred since the virus has been circulating in the human population[14][15]. We are able to observe this due to the unprecedented number of viral sequences from human samples that have been gathered during the pandemic.

This gives us a reference point for comparison. Although we are missing crucial early data about the very first SARS-CoV-2 sequences, and do not have any without the FCS insert, there are aspects of the later insertions that can nevertheless give us some important clues.

```
NC_045512.2     TTATCAGACTCAGACTAATTCTCCTCGGCGGGCACGTAGTGTAGCTAGTCAATCCATCAT    23640
OR233328.1      TTATCAGACTCAAACTAATT----------CACGTAGTGTAACCAGTCAATCCATTAT    23640
OR233322.1      TTATCAGACTCAAACTAATT----------CACGTAGTGTAACCAGTCAATCCATTAT    23640
OR233323.1      TTATCAGACTCAAACTAATT----------CACGTAGTGTAACCAGTCAATCCATTAT    23640
OR233324.1      TTATCAGACTCAAACTAATT----------CACGTAGTGTAACCAGTCAATCCATTAT    23640
MZ937003.2      TTATCAGACTCAAACTAATT----------CACGTAGTGTGGCCAGTCAATCCATTAT    23640
MZ937001.1      TTATCAGACTCAAACTAATT----------CACGTAGTGTGGCCAGTCAATCCATTAT    23640
MZ937000.1      TTATCAGACTCAAACTAATT----------CACGTAGTGTGGCCAGTCAATCCATTAT    23640
MN996532.2      TTATCAGACTCAAACTAATT----------CACGTAGTGTGGCCAGTCAATCTATTAT    23640
MZ081381.1      CTACCATGCGGCTTCTATAT----------TACGTAGTACAAGTCAGAAAGCTATTGT    23640
MW703458.1      CTACCATACGGCTCCTATAT----------TACGTAGTACAAGCCAGAAGGCTATTGT    23640
OK017806.1      CTACCATATGGCTCCTATAT----------TACGTAGTACAAGCCAGAAGGCTATTGT    23640
OK287355.1      CTACCACACGGCTTCTATAT----------TACGTAGTACAAGCCAGAAGGCTATTGT    23640
OR233302.1      TTATCAGACTCAAACTAATT----------CACGTAGTGTGGCCAGTCAATCTATTAT    23640
OP963576.1      TTATCAGACTCAAACTAATT----------CACGTAGTGTGGCCAGTCAATCTATTAT    23640
OQ297732.1      CTACCACACAGCTTCTATAT----------TACGTAGTACAAGCCASAAAGCYATTGT    23640
OQ297706.1      CTACCACACGGCATCTATAT----------TACGTAGTACAAGCCAGAAAGCCATTGT    23640
OK017804.1      CTACCACACGGCTTCTATAT----------TACGTAGTACAAGCCAGAAAGCTATTGT    23640
OK017803.1      CTACCACACGGCTTCTATAT----------TACGTAGTACAAGCCAGAAAGCTATTGT    23640
OK017805.1      CTACCACACGGCTTCTATAT----------TACGTAGTACAAGCCAGAAAGCTATTGT    23640
MG772934.1      CTACCATACGGCTTCTATAT----------TACGTAGTACAGGCCAGAAAGCTATTGT    23640
MG772933.1      CTACCATACGGCTTCTATAT----------TACGCAGTACAAGCCAGAAAGCTATTGT    23640
OQ297708.1      TTATCAGACTCAAACTAATT----------CACGTAGTGTTTCAAGTCAAGCTATTAT    23640
OQ297700.1      TTATCAGACTCAAACTAATT----------CACGTAGTGTTTCAAGTCAAGCTATTAT    23640
OQ297707.1      TTATCAGACTCAAACTAATT----------CACGTAGTGTTTCAAGTCAAGCTATTAT    23640
MT121216.1      TTATCAGACTCAAACTAATT----------CACGTAGTGTTTCAAGTCAAGCTATTAT    23640
MT084071.1      TTATCAGACTCAAACTAATT----------CACGTAGTGTTTCAAGTCNAGCTATTAT    23640
MW532698.1      TTACCATTCCATGTCATCAT----------TTCGTAGTGTCAACCAGCGTTCAATCAT    23640
MT072864.1      TTACCATTCCATGTCATCAT----------TTCGTAGTGTCAACCAGCGTTCAATCAT    23640
MT040336.1      TTACCATTCCATGTCATCAT----------TTCGTAGTGTCAACCAGCGTTCAATCAT    23640
MT040335.1      TTACCATTCCATGTCATCAT----------TTCGTAGTGTCAACCAGCGTTCAATCAT    23640
MT040334.1      TTACCATTCCATGTCATCAT----------TGCGTAGTGTCAACCAGCGTTCAATCAT    23640
MT040333.1      TTACCATTCCATGTCATCAT----------TGCGTAGTGTCAACCAGCGTTCAATCAT    23640
MT072865.1      TTACCATTCCATGTCATCAT----------TGCGTAGTGTCAACCAGCGTTCAATCAT    23640
LC663959.1      ATATCACACACCATCCACAT----------TACGTAGCGCAAACAATAAGAGAATTGT    23640
LC663958.1      ATATCACACACCATCTACGC----------TACGTAGCGCAAACAATAAGAGAATTGT    23640
LC663793.1      ATATCACACACCATCCACAT----------TACGTAGCGTAAACAATAAGAGAATTGT    23640
LC556375.1      ATATCACACGCCATCTATGC----------TACGTAGCGCAAACAATAAGAGAATTGT    23640
```

Figure 1: Alignment of SARS-CoV-2 furin cleavage site

2 Where do insertions come from?

For an insertion to happen, a minimum requirement is that the sequence being inserted has to be present, as RNA, in the same place where viral replication is happening, which, in the case of coronaviruses, is inside host cells, either within double-membraned vesicles or in the "convoluted membrane" near them[13].

Insertions are rare events, in part because the conditions in which source RNA is present in the location where viral RNA is being copied are rare. Some insertions have been traced to the host genome[11], and others to other circulating endemic coronaviruses (HCoVs)[14]. Others are from distal parts of the SARS-CoV-2 genome itself[15].

But those sources leave many insertions unaccounted for. SARS-CoV-2 enters the body and first starts infecting cells at mucosal surfaces which are inhabited by a diverse community of trillions of bacteria[23].

A clue that under some circumstances bacteria might be contributing short genetic sequences is to look at dinucleotide composition [24]. The following graphs show the dinucleotide composition of the insertions from human sequences in GISAID (filtered as described below) together with that of some potential sources. Note that the CG content of the insertions is higher than that in either the human genome or that of other HCoVs, consistent with some of their having come from bacteria like the three *actinomyces* species on the right-hand graphs.

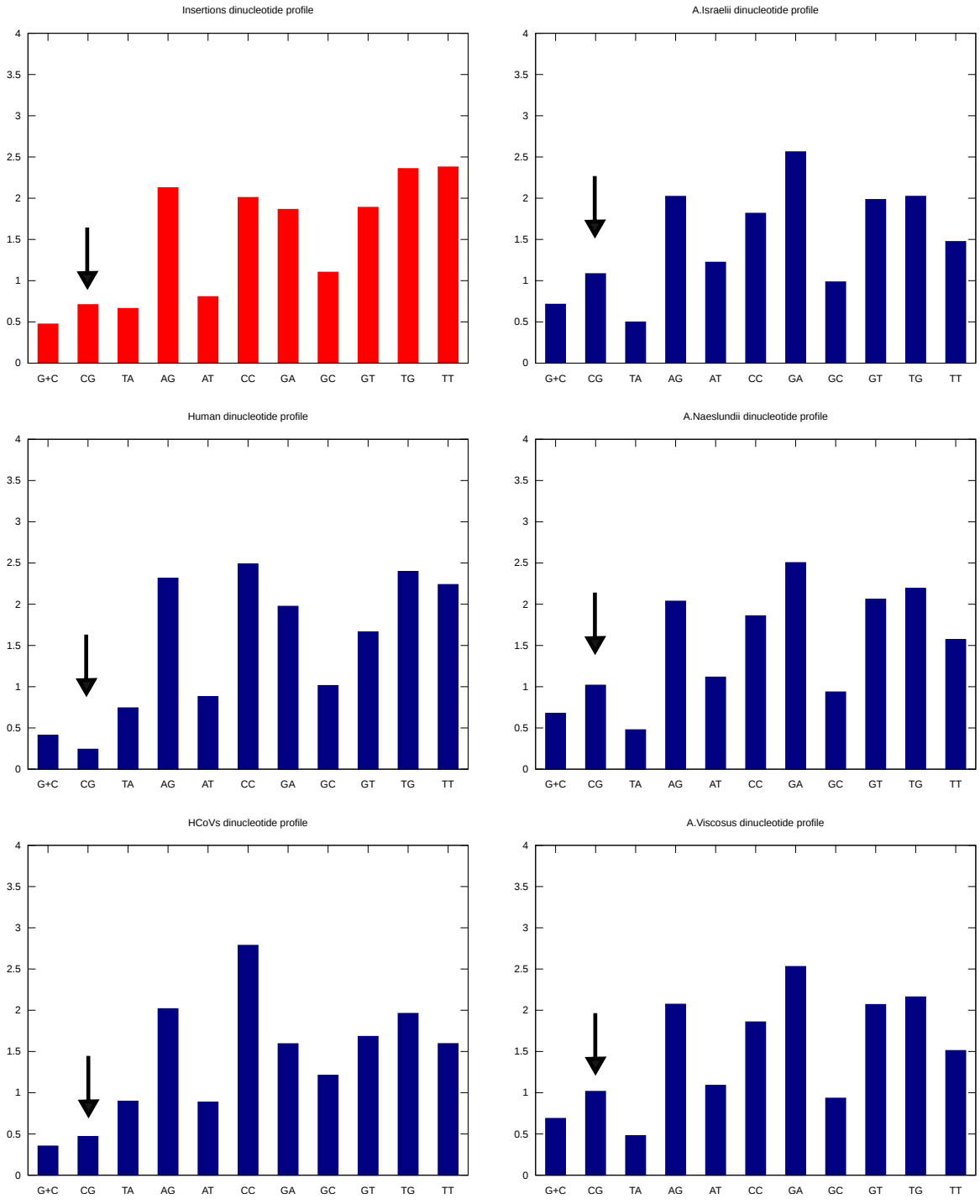

Figure 2: Dinucleotide compositions of insertions and potential sources

This is a clue, but we shall see in later sections stronger statistical associations between the insertions and bacteria in the human oral microbiota, adding weight to this hypothesis.

It is not known what the exact mechanism is by which bacterial RNA might end up in the same place at the same time as SARS-CoV-2 replication. One possibility is that the bacteria themselves get inside the cell. *Tannerella forsythia* and *treponema denticola* for example invade epithelial cells[16][20], which may also increase their susceptibility to SARS-CoV-2 infection[17][18][19].

When bacteria replicate inside cells they use RNA primers of about 12 nucleotides in length[21] (which actually seems a very "popular" size for these insertions, including the furin cleavage site itself). Although you would expect the bacterial cell membrane to isolate it from any viral replication that might be occurring in the same cell, there may be situations in which the bacterium has recently been lysed.

SARS-CoV-2 can also sometimes infect phagocytic immune cells[25], which might already contain bacteria or material from them. Normally that should be contained within the lysosome, but there may be circumstances in which that containment breaks down.

Another possibility is that SARS-CoV-2 is actually infecting bacteria[22] although I consider this less likely as coronaviruses are highly adapted to replicate inside eukaryotic cells[13].

3 Matches in gram-positive actinomyces

Out of the over 700[1] species of oral bacteria, we are going to focus on three of them: *actinomyces naeslundii*, *actinomyces viscosus* and *actinomyces israelii*.

A number of insertions from GISAID match in these species with between 3 and 6 nucleotides of sequence identity outside the match. Here is an example from *actinomyces naeslundii*:

```
Wuhan-Hu-1          ACGCCTATTAATTTAGTGCGCAGTGTGAAGAATGATCTCCCTCAGGGTTTTTC
NZ_CP066049.1       GGCGTGCCCTCCCAGGTGCGCAGTGTGAAGAACAGGTCGACGCGCTGCTCCA-
```

Figure 3: Insertion at 22204

The pattern highlighted in red is the insertion into SARS-CoV-2, shown here in context, in other words, inserted at position 22204. The other genome aligned with it is that of *actinomyces naeslundii*, shown as its reverse complement in this example, because the insertion matches on the negative strand. The pattern highlighted in blue shows that there are five nucleotides of sequence identity on one side of the insertion. These nucleotides shown in blue are not part of the insertion itself, but they match in both genomes outside it, at the place where they align.

The exact mechanism by which these insertions happen is not known[11]. It is likely to involve template-switching events, similar to those involved in recombination, which may be made more likely by adjacent sequence similarity[12], as seen in this example. Further analysis below has found a significantly elevated frequency of adjacent sequence similarity between SARS-CoV-2 and the hypothesized origin of some of these insertions.

A widely used program called BLAST[2] generates a value called an *E-value* which is supposed to be an estimate of how often we would expect to find a particular sequence in a genome, or set of genomes, by chance. In this case the E-value of the 12 nucleotides of the insertion itself in this genome is 0.31. This is not highly significant. You would not expect it to be as such a short pattern is reasonably likely to match in a 3152123-nucleotide-long bacterial genome. However, the E-value for all 17 nucleotides, including the 5 nucleotides before the insert, is much more significant, at 0.000244.

Both E-values are included in the table of examples below: that for the insertion itself, and for the insertion plus the sequence identity outside it. Note that an E-value is always a function of two things: what you are looking for, and where you are looking for it. The smaller the needle, or the larger the haystack, the more times you will expect to find it (i.e. the higher an E you will get) "just by chance".

Also included in the table are the number of insertions found at each position. The higher the number the more likely this is to be a real insertion and not some kind of sequencing artefact. This is because it's likely that the SARS-CoV-2 genome will only "tolerate" insertions in particular locations. It is also possible that particular locations are more susceptible to insertions due to

RNA structure. But whatever the reasons are, sequencing artefacts would be unlikely to cluster by location, because something that wasn't a real insertion wouldn't be expected to have any association with location.

The number of sequences in GISAID in which each insertion appeared is also included. The more sequences an insertion appears in, the more likely it is to be a real insertion and not an artefact.

Species	Insertion	Entire Match	Number of insertions here	Number of sequences	Position	Number of matching nts ahead	Number of matching nts behind	E (insertion only)	E (entire match)
naesl.	CACCGACTGCCC	CACCGACTGCCCAGCCCA	1	1	4118	0	6	0.31	0.000244
naesl.	TGGCGGCGGCGG	CGAGCTTGGCGGCGGCGG	288	3	22204	6	0	0.31	0.000244
naesl.	CAGTGTGAAGAA	CAGTGTGAAGAACAGGT	288	1	22204	0	5	0.31	0.000803
israel.	GTCCTGGCTGGG	ATCATGTCCTGGCTGGG	36	1	27300	0	5	0.39	0.001
israel.	GTCCACGGTCTG	GTCCACGGTCTGCCTC	4	5	23603	4	0	0.39	0.003
visc.	GCGTGCCGTGCA	AGCAGCGTGCCGTGCA	261	1	18	4	0	0.34	0.003
visc.	CACGGGCGGCGG	TGCGCACGGGCGGCGG	288	2	22204	0	4	0.34	0.003
israel.	TCCTGCCGCTGT	GTGTTCCTGCCGCTGT	100	1	21608	0	4	0.39	0.003
naesl.	CGGGCGACGCGG	TCGGGCGACGCGGAGT	85	2	22198	3	1	0.31	0.003
naesl.	GCACACCGTGCG	TGAGCACACCGTGCG	1	1	20382	3	0	0.31	0.008
israel.	GCACCAGCACCT	CTGGCACCAGCACCT	4	2	15827	0	3	0.39	0.009
israel.	TCCGAGCCCACG	TCCGAGCCCACGCAG	1	1	14314	3	0	0.39	0.009
visc.	GGGCTGGGCGGG	GCGGGGCTGGGCGGG	288	1	22204	0	3	0.34	0.008
visc.	CCGCTCGCGTCG	TCGCCGCTCGCGTCG	29	2	26184	3	0	0.34	0.008
israel.	AGCGCGGCTGGG	GTCAGCGCGGCTGGGT	2	1	26043	3	1	0.39	0.003
visc.	GGAGCGCGCGAG	GCGGGAGCGCGCGAG	288	1	22204	0	3	0.34	0.008
israel.	GCTGAGGAACGC	CGGGCTGAGGAACGC	2	1	27240	3	0	0.39	0.009
naesl.	AGTCTTCGCTGG	AAGTCTTCGCTGGAGA	2	19	6425	1	3	0.31	0.003
visc.	CTGGCGCGCGTC	GACCTGGCGCGCGTCT	472	1	17	3	1	0.34	0.003
visc.	CAGTGCGGCTTG	GACCAGTGCGGCTTGCC	1	1	6628	3	2	0.34	0.000889
naesl.	TGGCTGAGCCGG	TGATGGCTGAGCCGG	3	1	25057	0	3	0.31	0.008
visc.	CTCTCGCTGCGC	CAGCTCTCGCTGCGC	72	1	24	0	3	0.34	0.008
naesl.	GGCGGCGGCGAG	GATGGCGGCGGCGAG	1	2	11581	3	0	0.31	0.008
naesl.	CCACGGTCTGGT	CCACGGTCTGGTTCG	4	1	23605	3	0	0.31	0.008
israel.	GCTCAGCGTCAG	ACCGCTCAGCGTCAG	2	2	610	0	3	0.39	0.009
naesl.	GTTCATGCCGCT	GTGGTTCATGCCGCT	42	6	21607	0	3	0.31	0.008
israel.	GCGCCTCCGCCC	TGCGCCTCCGCCCTGT	261	1	18	1	3	0.39	0.003
israel.	TCACGCCGCTGT	TGTTCACGCCGCTGT	100	16	21608	0	3	0.39	0.009
naesl.	CCCGGGAGCCGG	ATTCCCGGGAGCCGG	18	3	22193	0	3	0.31	0.008

4 Quantifying the significance of the adjacent sequence identity

Since these insertions are mostly short (12 nucleotides) many of them will match in other genomes, which may or may not be possible alternative candidates for where they originated. We can look at how much sequence identity we find *outside and adjacent to* the insertions in these matches in these other genomes, and use this as a reference to estimate how unusual the level of sequence identity found in the bacterial matches is.

Two genomes were chosen for comparison: that of atlantic cod (*gadus morhua* NC_044048.1), because this is surely not a source for insertions into human sequences in GISAID, and the human genome (*homo sapiens* NC_000001.11) which probably is a source for some of them.

The complete set of 90771 insertions were first filtered according to the following criteria:

1. Minimum length 12 nucleotides (12 is the length we are interested in, any shorter will add noise by increasing the number of random matches).
2. Maximum length 24 nucleotides (we're interested in shorter insertions).
3. Position less than 29870 ("insertions" right at the end are more likely to be artefacts).
4. Not a mononucleotide repeat (these will match in genomes like cod or human with high sequence identity outside).
5. Does not contain a dinucleotide repeat of 5 or more repeats (same reason we excluded mononucleotide repeats).
6. Length is a multiple of 3 (inserts that aren't a multiple of 3 will break the coding sequence downstream of them and are therefore more likely to be artefacts).
7. Is not an exact match in Wuhan-Hu-1 (we're looking for other sources besides Wuhan-Hu-1).
8. Appears in two or more GISAID sequences (artefacts are less likely to do this).
9. Shares its location with two or more other inserts that appear in two or more GISAID sequences (artefacts are much less likely to do this).

After filtering, there were 2727 insertions, and 144044 matches in cod, and 235614 in human (these numbers are much bigger than 2727 because the same insert often matches in multiple places). The following frequencies of different amounts of sequence identity either side of the matches were found:

Min number of matching nucleotides outside the insertion	**Occurrences in cod**	**Occurrences in human**
3	1312/144044	7174/235614
4	404/144044	2380/235614
5	123/144044	629/235614
6	41/144044	184/235614

An additional "control" called "randomized self" was constructed as follows: for each of the insertions that matched the filters described above, the nucleotides in it were replaced with a random string of nucleotides the same length (made by picking randomly from the distribution of the four bases found in all of the insertions matching the filters). This was then treated as if it had been an insertion in the same location as the original. Those randomized insertions were matched up in each of the *actinomyces* genomes and the frequencies of 3, 4, 5 and 6 nucleotide-long sequence identities outside the "insertions" counted.

The rationale here is that we are trying to measure whether the content of the insertion reveals anything about its origin. By randomizing the content, but keeping everything else the same, we can get a measure of how often we would expect to find rates of 3, 4, 5, and 6 nucleotide sequence identity outside the insertion under the null hypothesis that the insertion did not originate in the genome we're looking at.

Because those rates are quite low, the randomization process was repeated 500 times for each genome. Those expected frequencies were used to calculate the odds ratios and p-values in the "Comparison with randomized self" columns in the table below.

There were significant results for all three *actinomyces* species, with the highest significance found for *a.naeslundii*.

Minimum number of matching nucleotides outside the insertion	**Occurrences in *a.naeslundii***	**Comparison with cod**	**Comparison with human**	**Comparison with randomized self**
3	15/160	OR=11.25/p=3.4×10^{11}	OR=3.29/p=0.00013	OR=3.30/p=0.00013
4	4/160	OR=9.12/p=0.0012	OR=2.51/p=0.08	OR=3.27/p=0.038
5	1/160	OR=7.36/p=0.13	OR=2.35/p=0.35	OR=3.00/p=0.29
6	1/160	OR=22.09/p=0.046	OR=8.05/p=0.12	OR=13.38/p=0.076

Minimum number of matching nucleotides outside the insertion	**Occurrences in *a.israelii***	**Comparison with cod**	**Comparison with human**	**Comparison with randomized self**
3	10/193	OR=5.94/p=1.4×10^{5}	OR=1.74/p=0.073	OR=1.64/p=0.097
4	2/193	OR=3.72/p=0.1	OR=1.03/p=0.58	OR=1.24/p=0.48
5	0/193	OR=0.00/p=1	OR=0.00/p=1	OR=0.00/p=1
6	0/193	OR=0.00/p=1	OR=0.00/p=1	OR=0.00/p=1

Minimum number of matching nucleotides outside the insertion	**Occurrences in *a.viscosus***	**Comparison with cod**	**Comparison with human**	**Comparison with randomized self**
3	6/191	OR=3.53/p=0.0086	OR=1.03/p=0.53	OR=1.02/p=0.54
4	3/191	OR=5.67/p=0.017	OR=1.56/p=0.3	OR=2.04/p=0.19
5	1/191	OR=6.16/p=0.15	OR=1.97/p=0.4	OR=2.39/p=0.34
6	0/191	OR=0.00/p=1	OR=0.00/p=1	OR=0.00/p=1

The odds ratios above are presented graphically below, together with additional results (not listed in the tables) for *treponema denticola*, *porphyromonas gingivalis*, *tannerella forsythia* and *aggregatibacter actinomycetemcomitans*. In each graph the odds ratio of 1.0 is marked with a red line. The higher the odds ratio the more adjacent sequence identity we found in the places the insertions matched between that species of bacterium and that reference.

It's interesting that the more "alien" genome of cod provides the lowest adjacent sequence identity. We might expect this compared to human because it's likely some insertions actually are coming from the human genome. But randomized insertions in bacteria still score generally higher than the actual insertions in cod. This is likely due to more "compatibility" between codon usage and/or dinucleotide bias in the bacterial genomes and that of SARS-CoV-2.

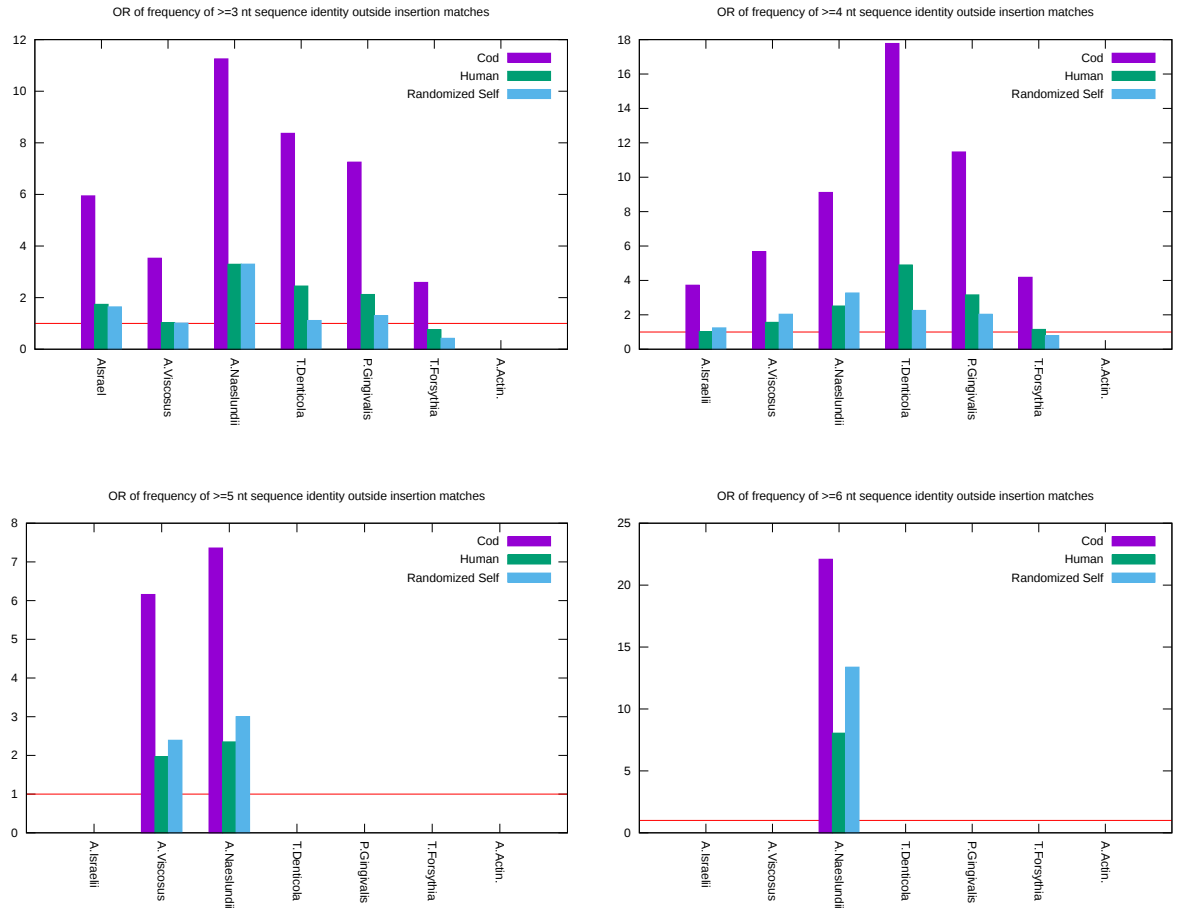

Figure 4: ORs for comparisons with cod, human and randomized self

These results indicate with some confidence that *a.naeslundii* (or something closely related to it) is a source of the some of the insertions in the human sequences found in GISAID since the start of the pandemic. It looks as though *a.israelii* and *a.viscosus* may also have contributed some, although these results have lower significance. However, if *a.naeslundii* can contribute insertions, it is likely that the other two can as well, even if this has happened too infrequently so far to be detectable with high confidence.

5 The furin cleavage site insertion

The FCS is a 12 nucleotide insertion of the sequence CTCCTCGGCGGG at position 23601, very similar to those in the table above, that has six contiguous nucleotides of sequence identity with *actinomyces israelii* on one side of the site.

```
Wuhan-Hu-1          TTATCAGACTCAGACTAATTCTCCTCGGCGGGCACGTAGTGTAGCTAGTCAA
NZ_CP124548.1       TCGTCCGGCCCCTCCTCGACCTCCTCGGCGGGCACGTACTGGCCGCGCTCGT
```

Figure 5: Furin cleavage site insertion

The E-value for the insertion in *actinomyces israelii* together with the six matching nucleotides is 3×10^{-4}. The additional matching "TG" separated from the six-nucleotide adjacent identity (CACGTA) has not been used in this analysis as in the insertions and controls I am only counting strictly adjacent identity.

This is a significant match, and the full 18 nucleotide sequence is not found anywhere in the human genome [1]. The best match in human is 14 of the 18, with an E value of 21. Note however that you can expect to find *any* 18 nucleotide sequence about 2.74% of the time in a genome as large as the human one, and you will find these 18 nucleotides in lots of other places if you BLAST against all of the genomes known to science (but with high E-values, because the "haystack" is so big). The key point here is that we have a three-way assocation. Other insertions in actual human sequences, and the additional matching nucleotides outside them, show a highly significant association with these bacteria. That independent observation leads us to a much smaller haystack.

It is hard to estimate an absolute confidence level for the proposition that the FCS actually did originate in *actinomyces israelii*. But we can use the odds ratio estimated above as a Bayes factor for a weaker hypothesis (which does not rely on E values at all). We have 6 nucleotides of sequence identity after the FCS insertion, which is ~10 times more likely under the hypothesis that it came from the same source as multiple other insertions in human sequences since the start of the pandemic than that it came from somewhere else and the adjacent sequence identity is just a coincidence.

Informally we can express this as: if all those insertions that look just like the FCS came from these (or similar) bacteria, then why not the FCS itself? Suddenly it doesn't look very unusual at all any more.

6 Possible origin of high frequency of CTCCTCGGCGGG in *actinomyces*

The furin cleavage site pattern (CTCCTCGGCGGG) appears 9 times in *a.naeslundii*, 19 times in *a.israelii* and 10 times in *a.viscosus*. The BLAST E-values are 0.31, 0.39 and 0.34. In other words, it appears 29 times, 48 times and 29 times more often than we might expect in each of these three genomes respectively, and far more frequently than it does in the genomes of, for example, human, or other mammals.

This fact by itself doesn't mean it originated in any of these genomes (the reasons for thinking it might have have already been discussed). But it's worth investigating, and it might help us to understand whether, if the origin is not *a.israelii*, it might still be more likely to be a related bacterial source than something else.

[1] or in those of raccoon dog (*nyctereutes procyonoides*), bat (*myotis davidii*), or pangolin (*manis javanica*)

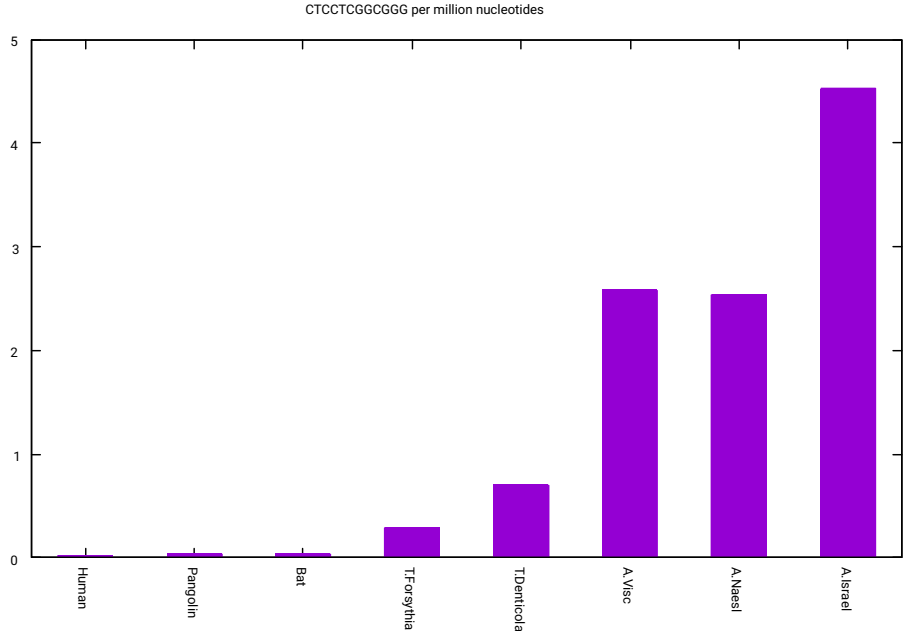

Figure 6: Frequency of CTCCTCGGCGGG in genomes of interest

All 19 occurrences of the pattern in *a.naeslundii* are frame-aligned and code for either the amino acid sequence Leu-Leu-Gly-Gly or, more often, Pro-Ala-Glu-Glu, since the pattern more often appears on the negative strand (or in a reverse complement coding sequence on the positive strand). As the alignment below shows, there is generally little sequence identity outside the pattern itself.

```
---CGAGCGCTGCGCGGCGACCTCCTCGGCGGGACGCTGTGGGCGGTAGAC--
GGGCGCCTGGCTCGCCC---CCTCCTCGGCGGGGGCCTCGCTGGCCGCCTC--
---CAGCAGGTTGATCTGCCGCTCCTCGGCGGGGAT--CCTCTGGGCGGGCGT
---GTCCGGCCCCTCCTCGACCTCCTCGGCGGGCACGTACTGGCCGCGCTC--
GCACTGCTCGCCACCGT---GCTCCTCGGCGGGGCCCTGACGCTGAGCGCC--
GCGCGAGGCGCGAGCTT---GCTCCTCGGCGGGCCCTGCGGCGCTCATGCC--
---GGCGAAGAGCCCGGCGAGCTCCTCGGCGGGAGCCGCGGGGGCCCAGCG--
---GTCCTGATGCGCCTGAGTCTCCTCGGCGGGCCCGGCACCCTCGGGGAG--
---CTCCTTCCACGCCTCGTACTCCTCGGCGGGCATGAGGTAGGCGCGGCC--
---CAGGGGGCGCGCGCGCACCTCCTCGGCGGGGGCGTCCACCGGGTCGGC--
GGCCATAATGCGCGAGCGC---TCCTCGGCGGGCAGCAGGTACCAGTCGTA--
---TCCGACGACGTGCGCGACCTCCTCGGCGGGCTCGTCGCGGAGGGGCTT--
---AGACGTCATCCATGGTGCCTCCTCGGCGGGGTCCCGACCCATCATTGC--
---CTTTCCCGCGACCTCGACCTCCTCGGCGGGGACGACGACGGTGACGAT--
GCCCGGGTGGGCCGCGG---GCTCCTCGGCGGGCTCGCTCGCCGATGTTTC--
---GTCGGGGCGGAAGGCGGCCTCCTCGGCGGGCAGGTCGCCCCAGCGCAG--
---GATGTCGCCCGCGTGGGCCTCCTCGGCGGGCCTGCGGTCCAGGCCCTC--
---GACGAAGGTGCCGTTCTTCTCCTCGGCGGGGGCGACCGGCAGGACGAC--
```

Figure 7: Alignment of occurrences of CTCCTCGGCGGG in *a.israelii*

Location	Strand	Amino acid
1821627	+	Pro-Ala-Glu-Glu
2341829	+	Pro-Ala-Glu-Glu
2405018	+	Leu-Leu-Gly-Gly
2751737	+	Pro-Ala-Glu-Glu
3383987	+	Pro-Ala-Glu-Glu
3582148	+	Pro-Ala-Glu-Glu
3634022	+	Pro-Ala-Glu-Glu
3810985	+	Leu-Leu-Gly-Gly
3927580	+	Pro-Ala-Glu-Glu
3884975	-	Pro-Ala-Glu-Glu
2529147	-	Pro-Ala-Glu-Glu
2030306	-	Pro-Ala-Glu-Glu
1903246	-	Pro-Ala-Glu-Glu
929102	-	Pro-Ala-Glu-Glu
857634	-	Pro-Ala-Glu-Glu
424543	-	Pro-Ala-Glu-Glu
291519	-	Pro-Ala-Glu-Glu

The two facts that all the occurrences of the pattern are frame-aligned, and that there is little to no sequence identity between them outside the pattern, make it much more likely that the high frequency is a consequence of codon usage bias than because the pattern is, for example, part of a transposable element.

This means that although we have a significant match for the pattern and adjacent sequence identity in *a.naeslundii*, it is quite likely that equally good matches would appear in other species of bacteria with similar codon usage bias.

We know that *a.naeslundii* is present in an environment were we also find SARS-CoV-2—the human mouth—but we can't rule out that similar bacteria don't occupy another niche where we might find the virus in another animal.

However the hypothesis that SARS-CoV-2 acquired the FCS in a human host after a zoonotic infection from a bat is certainly a highly plausible one.

7 Comparison with alternative FCS encodings

Some variants of the lab leak hypothesis contend that the FCS might be an *artificial* insert in the sense that the particular choice of nucleotides encoding it was not made by nature, but by an engineer sitting at a computer. It has even been suggested that they were human codon-optimized [3].

If the FCS was chosen like that it would be a coincidence if it showed the relationship to other natural insertions and to these *actinomyces* species that it does.

We can quantify this by running the same set of tests on alternative ways that 12 nucleotides could have been inserted to encode the same protein at this location. For example, instead of CTCCTCGGCGGG, the engineer might have chosen to insert CACCTCGGAGGG at the same location instead. The protein would then have still been Ser-Pro-Arg-Arg-Ala at that point. Altogether there are 575 other ways of encoding a 12 nucleotide insert to get the same FCS. It's likely that some of them will match up in our *actinomyces* genomes, but will they show the same sequence identity outside the places where they match?

Out of the 576 ways of encoding the same FCS, only one matches *actinomyces israelii* with higher adjacent sequence identity than the actual furin cleavage site insert. 157 of the 576 are equally as or more "human codon optimized" than CTCCTCGGCGGG, using a simple scoring system in which each codon, aligned as they are in SARS-CoV-2, gets a score equal to its inverse rank in the human codon frequency table. In other words, the least frequent codon for a particular amino acid scores 0, the next least frequent 1, etc.

None of those that are equally or more codon-optimized has as much adjacent sequence identity in *a.israelii* as the actual insert.

In other words: P(the observed FCS encoding)|(origin in *a.israelii*) > P(the observed FCS encoding)|(origin not in *a.israelii*) by a factor of 575. P(the observed FCS encoding)|(lab origin with human codon optimization) > P(the observed FCS encoding)|(not human codon optimized) by a factor of 3.7 (576/157). This gives us an overall Bayes factor in favour of the *a.israelii* hypothesis compared to the lab human-codon optimization hypothesis of 156.

This table shows some of the alternative encodings of the FCS that rank as equally or more human-codon optimized, sorted by total adjacent sequence identity between SARS-CoV-2 and *a.israelii* either side of the insertion. The complete table is in the supplement.

FCS encoding	**Total adjacent sequence identity of best match in *a.israelii***	**Human codon optimization score**
CTCCTCGGCGGG (actual FCS insertion)	6	8
CTCCCCGGCGGG	5	9
CTCCCCGCCGGG	5	8
CTCCAAGACGGG	5	9
CGCCCCGCAGGG	5	9
CCCCAAGGAGGG	5	9
CTCCCAGAAGAG	5	13
CTCCCCGGCGCG	4	8
CTCCAAGGAGGG	4	9
CTCCGAGGAGGG	4	8
CACCCCGGCGCG	4	8
CACCACGGAGGG	4	8
CGCCCCGGAGGG	4	10
CGCCCCGGCGCG	4	8
CGCCAAGGCGGG	4	8
CGCCCAGGCGGG	4	10
CGCCCCGCCGGG	4	8
CACCCCGCCGGG	4	8
etc... 140 rows deleted (see supplement)		

8 Frequency of CGGCGG in insertions

The use of the CGG codon for Arginine is very rare in coronaviruses, but in the furin cleavage site there are two of them. This has been used to make a case for an artificial origin[3], and also for a host insertion from human[9].

We can get a better sense for how unusual it is now that we have so much data on other insertions by counting how often it occurs *in them* compared to alternatives that would also be capable of coding for Arginine-Arginine if aligned.

In all insertions between 6 and 200 nucleotides in length from the latest set downloaded on 2024-09-11, there are 104 occurrences of the pattern CGGCGG out of a total of 4081 6-nucleotide strings that could also code for Arg-Arg, which is a rate of 2.6%, only just under the 1/36 (or 2.8%) that you would expect if all possible synonyms for Arg-Arg were used evenly.

If we filter the insertions by the same criteria used for the analysis of matches above, which was intended to reduce noise by excluding insertions more likely to be artefacts, we find an even higher rate: 16 out of 339, which is 4.72%. The reason for this relatively high frequency of CGGCGG is likely to be explained by the high GC content and/or codon usage bias of insertion sources.

A complete list of these insertions is included in the supplementary material.

9 Association with FCSes in feline coronaviruses

It has been argued that the leading Proline in the FCS may strengthen the case for a lab origin due to an interest in feline coronaviruses.

While that is an interesting idea, the presence of a single amino acid cannot be considered strong evidence for anything: there are only 20 amino acids to choose from and not all of them will be functional in that location. Perhaps only a handful are.

10 Association with Human ENaC α subunit

It has been argued[26] that the match between the amino acid sequence RRARSVAS in SARS-CoV-2, the initial RRA part of which is contributed by the FCS insert, and an identical protein pattern in a furin cleavage site in the human genome suggests a deliberate insertion.

This is an alternative hypothesis to the feline coronavirus one, and I believe incompatible with it. But it has a similar problem. The E-value given by BLAST[2] for finding RRARSVAS in the human genome is about 0.3. So whatever you might think of the case made for[26] or against[27] this feature's assocation with engineering, the likelihood of finding it by chance in the human genome is not low, and that it should have a similar function where it does appear is not particularly surprising.

11 Contamination

Contamination of the samples which yielded the sequences in GISAID with these bacteria is not just a possibility. It is a near certainty. But the assembly software should have removed any inserts that did not show considerable read depth with the surrounding genome. We are looking at short inserts, much shorter than the usual read length, so this ought to be minimized. We should gain some additional confidence from the filtering described above. However a limitation of this work is that I have not been able to investigate any raw reads directly.

12 Methods

The insert data was obtained from the website https://cov-spectrum.org, setting the filters to World/All Times, waiting for it to load, and copying out all the insertions. After basic reformatting that data is included in the supplement.

Alignments were done with Clustal Omega 1.2.4 and E-values calculated with BLAST 2.16.0. All the other code used for the searching, matching and calculations is available here but I've tried to include everything you need to check the results as data files in the supplement.

13 Discussion

SARS-CoV-2 is the only known sarbecovirus to have a furin cleavage site[4]. This is known to have been an area of active research interest. Proposals to insert furin cleavage sites artificially into SARS-related coronaviruses have been published[3], and even if that specific project may not have been funded, it is hard to argue that such work may not have been taking place at the Wuhan Institute of Virology around the time that SARS-CoV-2 was first observed in Wuhan.

But it's important to temper this evidence with the consideration that the reason FCS insertion is an area of research interest in the first place is because it's something that can and does happen naturally[6]. If the virus had some completely arbitrary feature that was associated only with research, and not with nature, that would be different. The FCS is likely to confer an evolutionary advantage[7] and that fact alone makes its presence consistent with a natural origin.

The finding that the FCS has a significant association with a possible insertion source (bacteria in the human oral microbiota) that is shared by other insertions in human sequences that have happened since the pandemic started lends support to the idea that it also shares their origin. In other words: that it is a natural insertion that may have occurred in a human host. Under some lab origin scenarios, this association has to be attributed to coincidence. But it is worth listing some of the possibilities and discussing them in turn.

A: Zoonosis from bat to human, then FCS acquisition. Research has found 2.7% seropositivity for SARS-related coronaviruses in people living in close proximity to the habitats of the bats who have these viruses [8] . Although relatively rare, and usually with little consequence, if one of these infections resulted in the acquisition of the FCS, it could start a pandemic. The current finding adds support for this hypothesis.

B: FCS acquisition in an intermediate host. Another possibility is that the virus jumped from a bat to another host (for example a raccoon dog) where it acquired the FCS. We have found here an association between the FCS insert and other inserts acquired in humans, but we were only able to infer this because we have several million SARS-CoV-2 sequences from human hosts in GISAID. In contrast, we have 0 sequences from raccoon dogs (or any other putative intermediate host).

C: FCS acquisition in a bat. It is possible that the FCS was acquired in a bat, although out of at least 1535 known sarbecoviruses[10], most of them in bats, an FCS has so far not been found. But it's difficult to know how often we should have expected to find it. And, just as for scenario B above, we don't have the millions of individual sequences from these hosts we would need to see whether the FCS shared any characteristics with other insertions that we have for humans.

D: FCS acquisition in a lab. The current finding makes insertion of a "designed" FCS (according to such criteria as copying feline or human FCSes or codon-optimizing for humans) less likely. However other lab scenarios are possible: the FCS might have been copied into a bat virus from a related virus in which it had been found naturally (perhaps in another host, for example a Pangolin) or it might just be that a natural bat virus with an FCS was found in samples that had been collected from bat caves and either cultured or recovered with reverse genetics. Certainly if such a virus had been found, it would have been at the top of the "interesting" list and therefore quite likely to have been recovered in this way.

Our findings here lend support to Scenario A over the others and also tend to equate Scenarios B to D. It has been previously argued that the apparent artifice of the FCS: its unusual (for a sarbecovirus) codons, or perhaps the leading proline, imply engineering. But as we have seen, the codons are not unusual for an *insert* into a human-circulating coronavirus, and the particular insert has a high association with a natural origin.

I therefore don't think there is a case for an artificial origin based on any characteristics of the FCS insert. Everything about it looks more natural than artificial. But there may still be a case just based on its existence.

14 Conclusion

That it's better not to jump one. I have made several hypotheses here, which are given varying degrees of support by the findings:

- That bacteria contribute insertions.
- That insertions are "similarity-assisted" (i.e. that adjacent sequence identity helps them to happen).
- That the FCS insert may be a similarity-assisted insertion from *a.israelii* or something closely related to it.
- That the pandemic started after a zoonotic virus acquired the FCS in a human host.

15 References

[1] Kah Yan How, Keang Peng Song, and Kok Gan Chan Porphyromonas gingivalis: An Overview of Periodontopathic Pathogen below the Gum Line.

[2] C. Camacho, G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer, and T.L. Madden BLAST+: architecture and applications 2009 BMC Bioinformatics, 10, 421

[3] S. Quay A Bayesian analysis concludes beyond a reasonable doubt that SARS-CoV-2 is not a natural zoonosis but instead is laboratory derived

[4] Daoyu Zhang, Gilles Demaneuf, Billy Bostickson, and Monali Rahalkar DRASTIC - An Analysis of Project DEFUSE

[5] Yujia Alina Chan and Shing Hei Zhan The Emergence of the Spike Furin Cleavage Site in SARS-CoV-2

[6] Yiran Wu and Suwen Zhao Furin cleavage sites naturally occur in coronaviruses

[7] Thomas Peacock et al. The furin cleavage site in the SARS-CoV-2 spike protein is required for transmission in ferrets

[8] Ning Wang et al. Serological Evidence of Bat SARS-Related Coronavirus Infection in Humans, China

[9] Antonio Romeu Probable human origin of the SARS-CoV-2 polybasic furin cleavage motif

[10] Bo Liu et al. A comprehensive dataset of animal-associated sarbecoviruses

[11] Yiyan Yang, Keith Dufault-Thompson, Rafaela Salgado Fontenele, and Xiaofang Jiang Putative Host-Derived Insertions in the Genomes of Circulating SARS-CoV-2 Variants

[12] Heather L. Wells, Cassandra M. Bonavita, Isamara Navarrete-Macias, Blake Vilchez, Angela L. Rasmussen, and Simon J. Anthony The coronavirus recombination pathway

[13] SARS-Coronavirus Replication Is Supported by a Reticulovesicular Network of Modified Endoplasmic Reticulum Kèvin Knoops, Marjolein Kikkert, Sjoerd H. E. van den Worm, Jessika C. Zevenhoven-Dobbe, Yvonne van der Meer, Abraham J. Koster, A. Mieke Mommaas1, and Eric J. Snijder

[14] Omicron variant of SARS-CoV-2 harbors a unique insertion mutation of putative viral or human genomic origin. A Venkatakrishnan, P Anand, PJ Lenehan, R Suratekar, B Raghunathan, MJ Niesen, and V Soundararajan

[15] Putative host origins of RNA insertions in SARS-CoV-2 genomes Thomas P. Peacock and David L. V. Bauer

[16] The intriguing strategies of Tannerella forsythia's host interaction Christina Schäffer and Oleh Andrukhov

[17] Covid-19 and SARS-CoV-2 infection in periodontology: A narrative review Agnieszka Drozdzik

[18] SARS-CoV-2 infection causes periodontal fibrotic pathogenesis through deregulating mitochondrial beta-oxidation Yan Gao, Wai Ling Kok, Vikram Sharma, Charlotte Sara Illsley, Sally Hanks, Christopher Tredwin, and Bing Hu

[19] Periodontal tissues are targets for Sars-Cov-2: a post-mortem study Bruno Fernandes Matuck, Marisa Dolhnikoff, Gilvan V. A. Maia, and Daniel Isaac Sendy et al.

[20] The fate of Treponema denticola within human gingival epithelial cells J. Shin and Y. Choi

[21] Principles and Concepts of DNA Replication in Bacteria, Archaea, and Eukarya Michael O'Donnell, Lance Langston, and Bruce Stillman

[22] Could SARS-CoV-2 Have Bacteriophage Behavior or Induce the Activity of Other Bacteriophages? Carlo Brogna and Barbara Brogna et al.

[23] Commensal Bacteria: An Emerging Player in Defense Against Respiratory Pathogens Rabia Khan, Fernanda Cristina Petersen, and Sudhanshu Shekhar

[24] Compositional differences within and between eukaryotic genomes Samuel Karlin and Jan Mrázek

[25] Inflammasome activation in infected macrophages drives COVID-19 pathology Esen Sefik, Rihao Qu, and Caroline Junqueira et al.

[26] A call for an independent inquiry into the origin of the SARS-CoV-2 virus Neil L. Harrison and Jeffrey D. Sachs

[27] SARS-CoV-2 furin cleavage site was not engineered Robert F.Garry

[28] The emergence, genomic diversity and global spread of SARS-CoV-2 Juan Li, Shengjie Lai, George F. Gao, and Weifeng Shi