Ubiquitous genomic fragment in human 2019-nCoV viruses in the spike-protein, also encoding a novel 87 aa protein, completely missing in all other coronaviruses

Sandeep Chakraborty,

Abstract

The origins of the highly virulent coronavirus isolated from Wuhan (Hubei, China) are uncertain, as are the reasons for its highly virulent nature (human-to-human transmission before the onset of symptoms). Here, 29 genomes of 2019-nCoV in GISAID reveals a genomic fragment which is present in all 2019-nCoV genomes, (and also in the recent Nanopore sequencing data from a family [1]), and absent in other species. The only entry in GISAID from bats (BatCoV-RaTG13) is a mystery (it does not have any publications linked to it), but is very close to human 2019-nCoV. Mutations in the viral genome need to translate in changes in protein sequences (and function) in order modulate its virulence. This genomic fragment is in the N-terminal of the spike-protein (98-228), a known-epitope region and implicated in viral entry into host cells. Interestingly, this region also encodes a novel 87 aa protein, with a shifted open-reading frame (a phenomenon common in viruses). The genomic fragment will help in faster diagnosis (excluding all other coronaviruses), while the protein information will aid in vaccine or inhibitor design. Note, there are no other fragments which have this property - present in nCov and absent in others. Coincidentally, amino acids '17-240 were deleted from the N-terminal domain of the TGEV Spike gene' using CRISPR, an experiment carried out in Wuhan [2].

Introduction

Coronaviruses are enveloped RNA viruses [3]. Out of the six coronavirus species that are known to cause human disease, four (HKU1, 229E, OC43 and NL63) cause mild symptoms, while SARS-CoV and MERS-CoV have caused disease outbreaks in China and the Middle-east [4]. The sequence (NC_045512.2 [5]) of the extremely virulent novel coronavirus (2019-nCoV) isolated from the city of Wuhan in China shows that it forms a separate clade within the subgenus sarbecovirus [5,6]. It is closely related to the single bat entry (BatCoV-RaTG13) in GISAID, the origin of which remains unclear.

In the current work, a fragment is shown that is present in all human 2019-nCoV, slightly mutated in the bat BatCoV-RaTG13, and completely missing in all other species. This genomic region is within the spike-protein, which is implicated in host entry and virulence. This region also encodes a putative 87 aa protein (frame-shifted within the spike-protein - a common phenomenon in viruses [7]),

Results and discussion

29 nCov genomes from GISAID

GISAID has currently 29 genomes of 2019-nCoV. One of this is BatCoV-RaTG13 from bats, whose origin is the missing link to other coronaviruses.

The fragment from human 2019-nCoV (Accid:NC_045512.2 [5]) (21852-22427, SI:nCoVFULLSLICE.fa) can be split into two with lesser homology elsewhere. These are 21883-22056 (SI:nCovSLICE1.fa,n=174) and 22167-22347 (SI:nCovSLICE2.fa,n=181), and are present in all 29 2019-nCoV genomes, slightly mutated in the BatCoV-RaTG13, and absent in other species. This region is at end of the 'middle region (10,901-22,830)' which grouped-nCoV and RaTG13 in a separate lineage from the sarbecovirus branch [6].

Nterminal of the spike protein is highly mutated:

Mutations in the genome need to translate into changes in protein sequences (and function) in order modulate its virulence (Table 1). nCoVFULLSLICE is in the N-terminal (98-228 of the 1273 spike protein). In BatCoV-RaTG13, this has only one amino-acid changed. However, it has many differences (including insertions) in other species (Fig 1).;

Note, ORF1ab which falls in the 'middle segment' [6] has 96% homology in the protein (Table 1). So, genomic variability here does not translate into as much protein changes. On the other hand, ORF8 which is at the end has only 59% homology.

Introduction of a new protein?

The changes in this region of the spike protein introduces a 87 novel aa protein (SI:novelprotein.fa), which is frame-shifted (Table 1). Gene overlaps (nucleotides coding for multiple proteins by being read in different reading frames) are common in viruses [7]. If this indeed is a protein, then it would greatly aid the development of a vaccine or an inhibitor.

Table 1: **Novel proteins in nCov (Accid:NC_045512.2 [5] encoded by nCovSLICE1** The genomic variation in viruses need to be looked at protein-level changes. And there are just a handful of proteins. A recent pre-print found that "a middle segment" has more variation from known Covs. Here, the end (27888 - 28256) which encodes ORF8 seems to have diverged the most (59%).

Protein	ORF in nCov	% in Batcov	% in others
ORF1a	254 - 13480	98	81
ORF1ab	13450 - 21552	99	96
NOVEL	21894 - 22196	93	—
spike glycoprotein	21503 - 25381	97	75
ORF3a	25348 - 26217	98	71
membrane glycoprotein	26460 - 27188	99	87
ORF7a	27388 - 27756	98	88
ORF8	27888 - 28256	95	59
ORF9a (**)	28260 - 28574	93	75
nucleocapsid phosphoprotein	28232 - 29530	99	90

nCov sequences from a family of six

nCoV from 'a family of six patients who traveled to Wuhan from Shenzhen between Dec 29, 2019 and Jan 4, 2020' was sequenced' [1]. The genomes (MN938384 and MN975262) are not yet available on ncbi. However, three samples sequenced using Nanopore (PRJNA601630) are available. The analysis of these reads also confirm the presence of these fragments (note Nanopore sequencing has a 10% error rate - so one needs to sequence deeper).

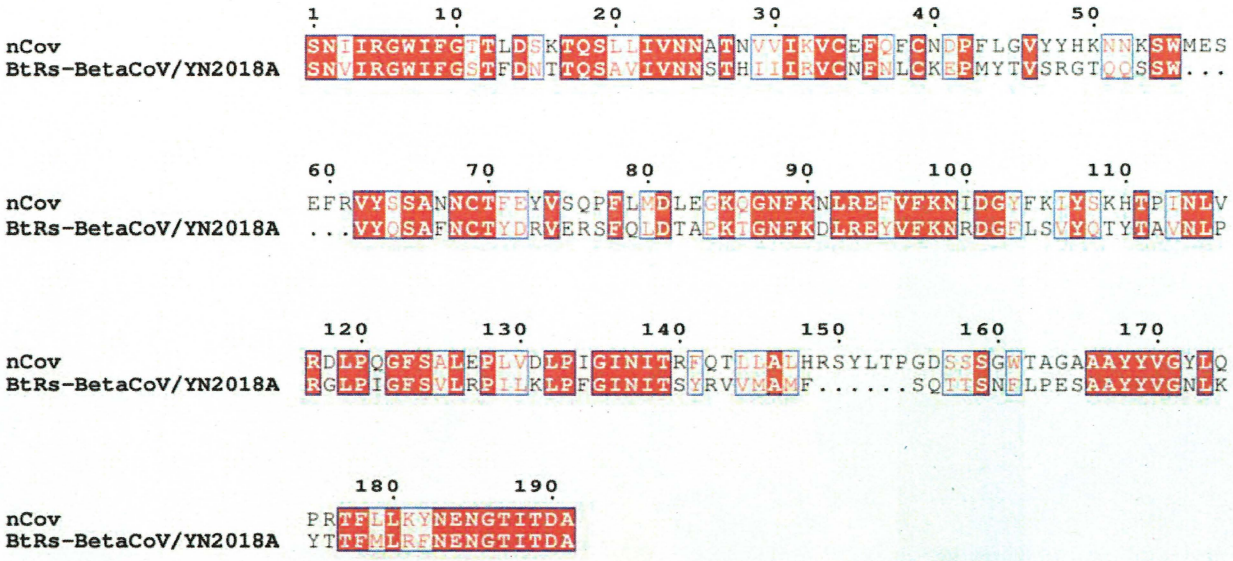

Figure 1: **MSA of the spike protein within the genomic fragment unique to nCov**: Identities=92/191 (48%), Gaps = 12/191. The spike-protein (1273 aa long) is a multifunctional enzyme that mediates host cell entry, and mutations in the protein has been implicated in virulence.

Discussion:

Here, two genomic fragments (proximal to each other) is identified which is unique to 2019-nCoV. Note, there are no other fragments which have this property - present in 2019-nCoV and absent in others. This has two important implications.

First, it mutates the N-terminal of the spike protein. The coronavirus spike protein mediates viral entry into host cells, among many other functions [8]. The 300 aa N-terminal was the epitope-binding site of the murine coronavirus [12]. Mutations in the protein has been implicated in virulence in many instances [9]. Specifically, N-terminal mutations in mice-specific coronaviruses extended the host range [10]. Such N-terminal deletions were being done for a while [11]. Interestingly, aa '17-240 were deleted from the N-terminal domain of the TGEV Spike gene' using CRISPR, an experiment carried out in Wuhan [2].

Secondly, it encodes a putative 87 aa protein, with a shifted open-reading frame, which is again missing in all other coronavirus species. The Lancet study noted that all 'respiratory samples were negative on two point-of-care multiplex PCR systems for 18 respiratory viral and four bacterial targets.' [1]. These two fragments could be added to this panel. Taken together, this aids faster diagnosis of exact 2019-nCoV, and gives a plausible explanation for the high virulence - simultaneously suggesting targets for inhibition/vaccine development.

Materials and methods

A small script splits the genome into sliding kmers of 300 nucleotides, each of which was aligned to a collection of coronavirus genomes (SI:Others.fa) which excluded the genomes from GISAID. Kmers that did not align to any of the genomes were then selected - it was found there were only 2 such fragments, and nothing else which satisfied this criteria.

MAFFT (v7.123b) [13] was used for doing the multiple sequence alignment (MSA). MSA figures were generated using the ENDscript server [14].

Competing interests

No competing interests were disclosed.

References

1. Chan JFW, Yuan S, Kok KH, To KKW, Chu H, et al. (2020) A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster. The Lancet .

2. Wang G, Liang R, Liu Z, Shen Z, Shi J, et al. (2019) The n-terminal domain of spike protein is not the enteric tropism determinant for transmissible gastroenteritis virus in piglets. Viruses 11: 313.

3. Fehr AR, Perlman S (2015) Coronaviruses: an overview of their replication and pathogenesis. In: Coronaviruses, Springer. pp. 1–23.

4. Lu G, Wang Q, Gao GF (2015) Bat-to-human: spike features determining 'host jump'of coronaviruses sars-cov, mers-cov, and beyond. Trends in microbiology 23: 468–478.

5. Zhu N, Zhang D, Wang W, Li X, Yang B, et al. (2020) A novel coronavirus from patients with pneumonia in china, 2019. New England Journal of Medicine .

6. Paraskevis D, Kostaki EG, Magiorkinis G, Panayiotakopoulos G, Tsiodras S (2020) Full-genome evolutionary analysis of the novel corona virus (2019-nCoV) rejects the hypothesis of emergence as a result of a recent recombination event. bioRxiv .

7. Chirico N, Vianelli A, Belshaw R (2010) Why genes overlap in viruses. Proceedings of the Royal Society B: Biological Sciences 277: 3809–3817.

8. Li F (2016) Structure, function, and evolution of coronavirus spike proteins. Annual review of virology 3: 237–261.

9. Chang HW, Egberink HF, Halpin R, Spiro DJ, Rottier PJ (2012) Spike protein fusion peptide and feline coronavirus virulence. Emerging infectious diseases 18: 1089.

10. Schickli JH, Thackray LB, Sawicki SG, Holmes KV (2004) The n-terminal region of the murine coronavirus spike glycoprotein is associated with the extended host range of viruses from persistently infected murine cells. Journal of virology 78: 9073–9083.

11. Hou Y, Lin CM, Yokoyama M, Yount BL, Marthaler D, et al. (2017) Deletion of a 197-amino-acid region in the n-terminal domain of spike protein attenuates porcine epidemic diarrhea virus in piglets. Journal of virology 91: e00227–17.

12. Kubo H, Yamada YK, Taguchi F (1994) Localization of neutralizing epitopes and the receptor-binding site within the amino-terminal 330 amino acids of the murine coronavirus spike protein. Journal of Virology 68: 5403–5410.

13. Katoh K, Standley DM (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol 30: 772–780.

14. Robert X, Gouet P (2014) Deciphering key features in protein structures with the new endscript server. Nucleic acids research 42: W320–W324.