A Palindromic RNA Sequence as Common Breakpoint Contributor to Copy-choice Recombination in SARS-CoV-2

William R. Gallaher (profbillg1901@gmail.com)
LSU School of Medicine, New Orleans, LA, USA https://orcid.org/0000-0002-5270-7468

Short Report

Keywords: SARS-CoV-2, coronavirus, COVID-19, recombination

Posted Date: July 22nd, 2020

DOI: https://doi.org/10.21203/rs.3.rs-46379/v1

Version of Record: A version of this preprint was published at Archives of Virology on July 31st, 2020. See the published version at https://doi.org/10.1007/s00705-020-04750-z.

Abstract

Much remains unknown concerning the origin of the novel pandemic Coronavirus that has raged across the globe since emerging in Wuhan of Hubei province, near the center of the People's Republic of China in December of 2019. All current strains of Coronaviridae have arisen by a combination of incremental adaptive mutations, against the backdrop of many recombinational events throughout the past, rendering each a unique mosaic of RNA sequence from diverse sources. The consensus among virologists is that the base sequence of the novel coronavirus, designated SARS-CoV-2, was derived from a common ancestor of a bat coronavirus, represented by the strain RaTG13 isolated in Yunnan province in 2013. Into that ancestral genetic background, several recombination events have since occurred from other divergent bat-derived coronaviruses, resulting in localized discordance between the two. One such event left SARS-CoV-2 with a receptor binding domain (RBD) capable of binding the human ACE-2 receptor lacking in RaTG13, and a second event uniquely added to SARS-CoV-2 a site specific for furin, capable of efficient endoproteolytic cleavage and activation of the spike glycoprotein responsible for virus entry and cell fusion. This paper demonstrates by bioinformatic analysis that such recombinational events are facilitated by short oligonucleotide "breakpoint sequences", similar to CAGAC, that direct recombination naturally to certain positions in the genome at the boundaries between blocks of RNA code and potentially RNA structure. This "breakpoint sequence hypothesis" provides a natural explanation for the biogenesis of SARS-CoV-2 over time and in the wild.

Main Text

By mid-December of 2019, hospitals in the area of Wuhan, Hubei Province, China, became aware of an outbreak of novel pneumonia (1-3) that was not due to either influenza virus or a reoccurrence of the severe acute respiratory syndrome coronavirus (SARS-CoV) of 2002-2003 (4-6). The early history of what developed into a global pandemic of epic proportions was summarized in early January, indicating that community spread was likely before the center of the outbreak erupted in association with the Huanan Seafood Market in central Wuhan (7,8).

Upon isolation, the causative agent was identified as a novel coronavirus of the viral genus *Betacoronaviridae*. The viral genome was rapidly sequenced by several groups, and reported out of Shanghai in early January 2020 (2). The initial analysis showed that the novel virus was most similar (89%) to a bat coronavirus, isolate SL-CoVZC45, but only 78% to SARS-CoV of 2003. Ultimately, the sequence was classified in the sub- genus Sarbecovirus and officially named Severe Acute Respiratory Syndrome Coronavirus, Hu-1, abbreviated SARS-CoV-2 (Genbank NC_045512.2) (9).

Coronaviruses comprise a very large and diverse family of single-stranded RNA viruses with positive polarity (10,11), with the longest RNA genome known, in the case of SARS-CoV-2, 29,903 bases long. The genome is divided into two overall regions. The 5' region, roughly 2/3 of the genome, is 21,555 bases long. It is translated from the entire genome serving as an enormous mRNA, into a polyprotein (Orf1ab) subdivided into two sections by a ribosomal frameshift just past midway. From the first section, three nonstructural and multifunctional proteins, nsp1, nsp2 and nsp3, are cleaved at sites recognized by an autocatalytic papain-like protease (PL-pro) encoded within nsp3. From the second section, eleven additional proteins, nsp4 through 16, are cleaved by a second autocatalytic protease 3CL-pro that cuts just past a glutamine (Q), as the uniformly last amino acid in each protein.

The second genomic region codes for nine proteins, including the four major structural proteins of the virus, in gene order, the spike glycoprotein (S), which is cleaved into two subunits S1 and S2 (12,13); the membrane protein E (E); the matrix glycoprotein (M); and the nucleocapsid protein (N). Each of the nine proteins is expressed from a nested set of progressively shorter mRNAs that begin with a leader RNA derived from the 5' end of the genome, linked to a transcriptional regulatory sequence (ACGAAC) repeated at the beginning of each gene (14). All mRNAs derived from the genome have a common untranslated 3'OH terminus (nt29674-29903). Coronaviruses are very complex RNA viruses, encoding a total of 23 different proteins, many interactive with one another in multiprotein aggregates, and some of which having multiple structural and functional domains.

Here I present evidence that the RNA genome of SARS-CoV-2 is organized into structural and fucntional blocks of RNA information that are demarcaed by short RNA breakpoint sequences that promote recombination at specific nonrandom locations within the viral genome.

Once the SARS-CoV-2 sequence became available, public and private communication among virologists, molecular biologists and phylogeneticists exploded, much of it playing out on the blog site *virological.org* or other internet means before any formal publications were processed, in an unusually free and open atmosphere of communication.

In that spirit, on January 24, scientists at the Wuhan Institute of Virology (WIV) published the sequence of a viral isolate obtained from the bat species *Rhinolopus affinis* in 2013, designated Bat_RaTG13, which was 96% identical to SARS-CoV-2, with an even higher degree of amino acid identity (2). The existence of such a close relative to the pandemic virus at a research lab a short distance (12km) from the center of the outbreak set off a firestorm of commentary that SARS-CoV-2 was derived from laboratory manipulation, or at least, from accidental release of a laboratory isolate. Added to that was the observation that SARS-CoV-2 had an apparent insert of a four amino acid site that created a site recognized by the endoproteolytic host enzyme, furin. Furin-specific activation of the Spike glycoprotein of coronaviruses is common in that peptide region, but lacking in SARS-CoV of 2003 or other members of the *Betacoronavidae* closely related to SARS-CoV-2 (15). Investigators at WIV had previously participated in "gain of function" experiments, to probe how bat coronaviruses from the wild could acquire functions permitting human infection (16,17). The combination of circumstances produced a high level of public suspicion that SARS-CoV-2 was the product of such a "gain of function" experiment, with a furin site deliberately inserted into the Bat_RaTG13 framework.

Several scientists, with great urgency, posted analyses on *virological.org* that showed, either at the RNA or protein level, that there was no molecular evidence supporting either laboratory manipulation or that SARS-CoV-2 could have been derived from any known virus in any laboratory inventory. My analysis grounded in RNA sequence concerning Bat_RaTG13 was posted shortly before midnight Feb 6 (18). A wider analysis, also ruling out origin from laboratory isolates derived from pangolins, grounded in peptide sequence, was posted 10 days later, on February 16, and later published (19).

Closer examination of the recombinant regions among these viruses, to be reported here, led to the identification of a short oligonucleotide sequence at the recombinant boundaries within SARS-CoV-2 that may provide a unifying hypothesis concerning some, but not all, of the recombinational events known to have occurred in SARS-CoV-2.

It has been widely known for decades that coronaviruses are profligate in engaging in recombination, through copy-choice errors involving jumping from one template to another (20-22). I shall focus on key and relatively recent recombination events, and not attempt to comprehensively explain all such events in the past. Other breakpoint sequences may well be present in the genome, and I have made no attempt to comprehensively identify them.

Conjecture on the geographic origin of SARS-CoV-2, i.e. the when, where and by what pathway, is beyond the scope of this paper. Indeed, the origin of SARS-CoV-2 is unknown as of this writing, and perhaps ultimately unknowable.

Further work by others has generated a general picture of the structure of the SARS-CoV-2 genome based on breaks in homology in comparison to other known coronaviruses. As determined by Li et al (23), the vast majority of the RNA sequence of SARS-CoV-2 is derived from a relatively recent common ancestor with Bat_RaTG13. Similarity plots (SimPlot) detected a number of sudden changes in the percent similarity between SARS-CoV-2 and the base RaTG13-like sequence, accompanied by a change in similarity of RaTG13 to other viruses. Where quantitatively significant, such a sudden change can signal a breakpoint for recombination. Two such breakpoints were observed, on either side of an apparent recombination involving an ancestor of RaTG13, an ancestor of SARS-CoV-2 and an ancestor of a betacoronavirus isolated in 2019 from a confiscated Malayan pangolin, namely Pan_SL-CoV_GD/P1L (24). This recombination event left SARS-CoV-2 with a receptor binding domain (RBD) capable of binding the human ACE2 receptor, and RaTG13 with an RBD incapable of do ing so.

In a parallel study, Boni et al (25) identified a much larger number of breakpoints between blocks of RNA, and a more complex evolutionary history for both RaTG13 and SARS-CoV-2 within a subclade of similar viruses. Indeed there are multiple localized regions of sequence discordance between the two viruses consistent with a number of recombination events. In their view, against the backdrop of other closely related bat and pangolin cornaviruses, the ACE2 receptor binding region is ancestral to SARS-CoV-2, and was lost by RaTG13 in the recombination event, as the most parsimonious explanation of the evolutionary pattern.

For the purpose of my analysis, the direction of donation of the ACE2 receptor binding function is not relevant, simply that a clear area of recombination involving these otherwise closely related viruses from the same subclade has been identified by both Li et al (23) and Boni et al (25).

My present analysis extended to the entire genomes of five viral isolates, all available to the general public. Four were acquired from GENBANK, the reference sequence for SARS-CoV-2 (NC_045512.2), for SARS-CoV of 2003 (NC_004718.3), for Bat_RaTG13 (MN996532) and for Bat Coronavirus HKU9-1 (EF065513). The fifth, isolated from pangolin, Pan_SL-CoV_GD/P1L, deposited in the GISAID database as Pangolin.Guangdong-1- 2019.EPI_ISL_410721.2019 (24). While the genomes themselves are RNA, reverse-transcribed DNA is the primary data source, so DNA code will be used here. Sequence queries, analysis and alignment was performed using the BLAST utility, especially the BLAST2 version, available online from the National Center for Biotechnology Information (NCBI) https://blast.ncbi.nlm.nih.gov/Blast.cgi . Additional alignments and graphic renditions were performed using the program CLUSTAL X 2.1 (26). All alignments were checked visually and a very few minor adjustments made that did not affect the results, but removed specious gaps. For RNA structure prediction the online utility at the University of Vienna, Austria, was used (http://rna.tbi.univie.ac.at/forna/). For estimates of a time for the most recent common ancestor to two sequences (TMRCA), the determinations of Boni et al (25) were used, i.e. 1946-1980 between SARS-CoV-2 and its closest relative, RaTG13.

Three months after my original post contrasting the RNA sequences of RaTG13 and SARS-CoV-2 in the vicinity of the furin insert, I revisited that analysis in greater detail.

While the S protein is exposed to host selection and the immune response and thereby more variable than other parts of the genome, these factors do not apply to base changes that leave the peptide sequence unchanged. Nevertheless, further analysis of a 100% identical peptide sequence of 519 amino acids in S glycoprotein, between RaTG13 and SARS-CoV-2, shows that a 5.1% divergence is consistent. Even such a long region of amino acid identity, when examined at the RNA level, may conceal an underlying divergence that spans decades. This level of divergence renders impossible the assertion that SARS-CoV-2 isolated in 2019 was derived from RaTG13 isolated in 2013.

The furin insert is remarkable because it within this 519 amino acid region otherwise identical between RaTG13 and SARS-CoV-2. Coutard et al. (15) had elegantly described the length and sequence polymorphism that is common in the immediate vicinity of the furin site, but a mechanism of how it occurred at this specific location was lacking. It was then that I noticed a tandem duplication of sequence immediately before the furin insert of SARS-CoV-2, namely CAGACTCAGACT (Fig. 1a). Such sequence duplication occurs when the viral polymerase complex stops and starts, a so-called stutter, and may insert either extra bases or produce a tandem duplication. When the polymerase complex comes upon a new region of highly base-paired stem-loop structure, its processivity is impeded as helicase unwinds and “melts” the structure ahead (27). Finding of the tandem duplication raised the possibility that the S1/S2 borderline may be such a transition from one block of genetic information to the next.

There are only two other locations where a significant length polymorphism occurs in comparing the RaTG13 and SARS-CoV-2 sequences, that are otherwise colinear along the entire genome. The first is the insertion of three nucleotides and a single amino acid at nt3332, within the Orf1a polyprotein. The other is a similar insertion of a single amino acid at the beginning of the M protein. In each of these instances, the RNA sequence immediately preceding the insert was 3327CAGAC and 26527CAGAT.

I then further examined the recombination site for the pangolin-like RBD peptide sequence highly similar to that in Pan_SL-CoV_GD/P1L, but also comparing the S1 RNA sequence from RaTG13 (Fig 1b). Its divergence at the RNA level is high with respect to both RaTG13 and the virus isolated from pangolin, from the latter overwhelmingly in synonymous mutations, at 10.4%. This clearly indicates that the recombinant sequence was derived from a virus that only shares a quite distant ancestor (TMRCA range of 1945 to 1981) with the virus present in pangolins in 2019. Most likely, divergence that long ago occurred in bats. Nevertheless, just prior to the upstream recombination breakpoint is again found 22839CAGAT, conserved in both SARS-CoV-2 and RaTG13.

There are three iterationsof CAGAC/T within a span of 91 nucleotides, between 22752-22843, of which only the third was noted in Fig 1b. Three of the three in the cluster are conserved between SARS-Cov-2, and RaTG13, while two of three are conserved in Pan_SL-CoV_GD/P1L, Pan_CoV/GXP2V, SARS-CoV of 2003, and Bat_CoV/YN2018B, a much wider phylogenetic range of betacoronaviruses (not shown).

In further examining the 3327CAGAC location in Orf1a, I noted that there was an unusual degree of divergence between the RaTG13 and SARS-CoV-2 sequences 5' to that breakpoint, an extraordinary 9.1% in RNA and 16 amino acid differences (18.1%), over a region of just 255 nucleotides (Fig 2). At the beginning of this disparity, at the transition between a CT rich region and the AG rich region encoding the acidic-rich region in nsp3, lies the sequence 3045CAGAT. This region would appear to define another recombination site between an ancestor of RaTG13 and yet another distant coronavirus sequence. In this case, assessment by the BLAST utility indicates that the source virus remains unsampled. It is noteworthy, however, that the boundaries of the recombination site are CAGAT and CAGAC.

Li et al.(23) also demonstrated an unusually high degree of similarity (>99%) in the 3'OH untranslated region among not only RaTG13, SARS-CoV-2 and Pan_SL-CoV_GD/P1L, but also SARS-CoV of 2003 (94.5% vs. 78% genome-wide). In other coronaviruses, such as Middle Eastern Respiratory Syndrome (MERS) virus this region can be much less conserved. Close examination of the RNA sequence at the beginning of the 3'OH region demonstrates that SARS-CoV-2, RaTG13 and Pan_SL-CoV_GD/P1L share a common identical sequence that is replete with palindromic sequences (Fig.3a) This may imply that the 3'OH region is “contagious” among bat coronaviruses, maintaining a very high degree of similarity by frequent recombination. Indeed, both CAGAC and CAGAT are found at the 5'end of the 3'OH sequence.

The origin of the furin site RNA sequence has been a complete mystery, even though similar amino acid motifs are found in many coronaviruses. A BLAST search for the actual insert sequence, CTCCTCGGCGGG, identifies a sequence derived from Bat coronavirus HKU-9, from a region in S protein downstream of the S1/S2 junction, with identity at the last 10 nucleotides. Looking further at the HKU9 sequence, CAGAC lies directly prior to the identity (Fig.3b).

The insert itself is also replete with short palindromes, CTCCTC and GGCGG. Overall, there are four such palindromes, comprising 21 out of a span of just 27 nucleotides of the single stranded RNA, an unusual feature. (In DNA parlance, a palindrome is a sequence that reads the same way on each of both strands. Here the term is used in its original sense of a series of letters that read the same in both directions on a single strand.) Common deletions have been shown to frequently occur, after infection of VeroE6 cells by SARS-CoV-2, within only a few passages, repeatedly either CAGACTCAGACTAAT or AATTCTCCTCGGCGGGCACGT, but neither both at once, nor a wider overlap of the two than the three underlined nucleotides (30). This demonstrates that these sequences define repeatedly used breakpoints for recombination with biological relevance.

The SARS-CoV-2 and HKU-9 sequences can be aligned as 18 identities in a span of 28 nucleotides of HKU-9, with minimal gaps. While not perfect, this does support the hypothesis that the entire furin insert may have been derived by recombination from an as yet unsampled coronavirus, mediated by the presence of CAGAC as the lead nucleotide in both progenitor viruses, with other palindromic oligonucleotides as contributors. A recent bat coronavirus isolate from Yunnan (28) also has an apparent insert at the S1/S2 border, PAA rather than PRRA, supporting the concept of a natural insertion at this site. However, this insert is far too dissimilar from the RNA sequence encoding S protein of SARS-CoV-2 to be at all closely related in origin, and does not contribute an endoproteolytic site for furin..

All of these recombination events, in Orf1a, in the RBD and the furin site itself, have most likely occurred since SARS-CoV-2 and RaTG13 diverged from their most recent common ancestor (MRCA). As estimated above, this would give a wide range for the time period during which these recombination events may have taken place, from the last half of the 20th century to the present.

The palindromic CAGAC or its variant CAGAT appear to be common, while far from universal, boundary elements, “breakpoint sequences”, to all of these recombination events identified above. In single-stranded RNA viruses, recombination does not commonly take place by breakage and rejoining of two complete strands, but rather by the polymerase slowing or stopping during transcription and jumping to another template strand, “copy choice”, during a mixed infection. This can take place in either direction of synthesis, even though the vast majority of replications are of the positive strand from the negative template. It is perhaps significant that the breakpoint sequences in the examples within SARS-CoV-2 are more uniformly at the 5’ end of the recombinant or insert. This would imply that the events occurred during positive strand synthesis from the negative strand template. Since the actual breakpoint sequence would be on the template being read 3’ to 5’, this would imply that the complement of CAGAC, namely 3’-GTCTG-5’, is what is really recognized by the polymerase complex. It is much more convenient, however, to define the site on the positive genomic strand.

The recombination events delineated here are notably defined at either end for the borders of functional regions of the genetic information – the acidic rich region of nsp3, the RBD, the S1/S2 border. They are nonrandomably distributed, often in clusters, 88 on the positive strand, 46 on the minus strand, but sparse between nt3331 and nt18855 (not shown) that Boni et al (25) identify as a breakpoint-free region.

Therefore, CAGAC and CAGAT may define points in the RNA genome that mark the boundaries between discreet blocks of RNA code information, corresponding to the block pattern of breakpoints observed by Boni et al (25). If so, then recombination events enhancing evolutionary adaptation of coronaviruses over vast distances in time would involve whole blocks of RNA information, and not randomly disrupt the RNA code in the middle of such blocks. Domains can be exchanged *en bloc*, a superior outcome to creating hybrid blocks of information that may not function at the protein level.

Echoes of such an organization of the RNA code into blocks, marked by discreet “breakpoint sequences”, can be found in the Orf1b region of the coronavirus genome where the polyprotein is broken down into component nsp proteins at recognition sites for 3CL-pro protease (14). For instance, the RNA boundaries for 3CL-pro itself are CAGAG and CAAAG. For helicase, they are CAGGC and CAAGC. The reader will note that these sequences preserve 3 or 4 out of 5 of the nucleotide sequences CAGAC and CAGAT discussed above, in part because the codons for Q, the uniformly terminal amino acid for each nsp, are CAG and CAA. The boundaries between nsp proteins, at 11 sites, have been ensconced in the peptide sequence for the entire known history of coronaviruses, likely extending many millions of years (29). Yet a portion of the underlying RNA breakpoint sequences CAGAC and CAGAT are still recognizable as that in the ancestral organization of the RNA code.

The hypothesis being put forward here is that the RNA genomeof members of the *Betacoronaviridae*, including SARS-CoV-2, is organized into blocks of information bounded by short “breakpoint sequences”. A subset of these is identified here in multiple sites as CAGAC and CAGAT. It has been previously suggested that the TRS, ACGAAC, already active in mRNA splicing, may also serve as a recombiantion signal (22). These blocks correspond to functional domains in the encoded proteins, and recombination is directed to these breakpoints by slowing the processivity of RNA polymerase complexes at breakpoints, enhancing their jumping the track to an alternate template. A series of such events, entirely natural and using a means of genetic exchange frequently employed by coronaviruses, accounts entirely for the biogenesis of SARS-CoV-2 in the wild.

One correlate to this hypothesis recognizes that “single-stranded RNA” is something of a misnomer. A 30kb genome cannot exist as an incredibly long rope of RNA. Rather, it is routinely organized into highly base- paired “double stranded” regions that fold back onto themselves in potentially complex stem-loop and pseudoknot structures. Virus-encoded helicase is essential precisely because these structures would otherwise be an absolute impediment to RNA duplication, as well as recombination. Despite RNA structure prediction programs, we have a poor understanding of these structures, but a correlate of the breakpoint hypothesis would be that “breakpoint sequences” separate structural regions of RNA as they do functional regions of RNA and protein. It has also been suggested that single-stranded interstices or loops may be preferential targets of RNAse mediated recombination by a breakage and repair mechanism (see 22) An illustration of a possible RNA structure for the RBD insert is shown in Supplementary Fig.S1. While such proposed structures are hypothetical and speculative, this example shows the extraordinary potential of base-pairing and double stranded structure that lies hidden in single-stranded RNA code. When recombination preferentially occurs at breakpoints between such structures, the overall structural integrity of the RNA genome would therefore be preserved.

The identification of CAGAC and CAGAT as such “breakpoint sequences” may be only one of many similar signals in RNA code. We already know of the TRS (ACGAAC) for transcription (14), and start and stop codons for translation, and the longer stem-loop structures that regulate the molecular biology of many viruses. There may be many more.

I would end with a warning. Karst limestone formations extend widely across southern China, from Yunnan province, through Guangxi province and into westernmost Guangdong province. Portions of such wilderness are designated a UNESCO World Heritage Site (https://whc.unesco.org/en/list/1248). The limestone is riddled with caves carved by water flow that are occupied by large colonies of bats. Several colonies of both insectivorous microbats and fruit macrobats can occupy the same caves. Bats and bat viruses of innumerable variety cohabit this ecological environment, providing ample and ongoing opportunities for mixed infection of bats by a variety of bat coronaviruses (10,11,31-33) . Likewise, gat guano, mined for fertilizer, is a viral soup with many contributing sources. The proximal source SARS-CoV-2, and particularly of the critical sequences in the RBD or furin-sensntive site, remains to be identified.

As these recombination processes are endless, natural and ongoing, the possibilities for emergence of pandemic viruses from these natural virological parks of Southeast Asia are also endless. As the emergence of the 1918 pandemic influenza from western Kansas (34,35), Hantavirus from the American Southwest (36), Ebola from Africa (37), and Zika from Africa via South America (38) prove, emergence can occur from every wilderness area on planet Earth. Every human virus represents an emergence from an animal source. In the past they have occurred only occasionally, but in the 21st century, emergence of pandemic virus has just changed the world as we have known it.

Declarations

Acknowledgments: This analysis was produced without grant or institutional support, as the author is long retired from his university post, a Professor Emeritus since 2005. I would like to thank Profs. Edward Holmes, Andrew Rambaut and Timothy Skern for their encouragement and support. I am deeply indebted, as always, to my longtime collaborator and friend, Prof. Robert Garry, for his commentary and offers of help. I want to thank my son, MSgt Andrew D. Gallaher, USMC, for his commentary and assistance in preparing the manuscript.

Competing Interests: None

Supplement Material

Supplementary Figure S1. Predicted RNA structure of the RBD recombinant region of SARS-CoV-2. A. The linear sequence of SARS-CoV-2 RNA is shown for the RBD recombinant region, beginning with the CAGAT “breakpoint sequence” at nt23590 and extending for 272 nucleotides. B. Predicted base-paired RNA structure for the RBD rrecombinant region using the utility hosted by the Institute for Theoretical Chemistry at the University of Vienna, Austria, http://rna.tbi.univie.ac.at/forna/ . The model structure consists of a highly base-paired core with five stem-loop branches. For most of the structure, shown in green, confidence is high.

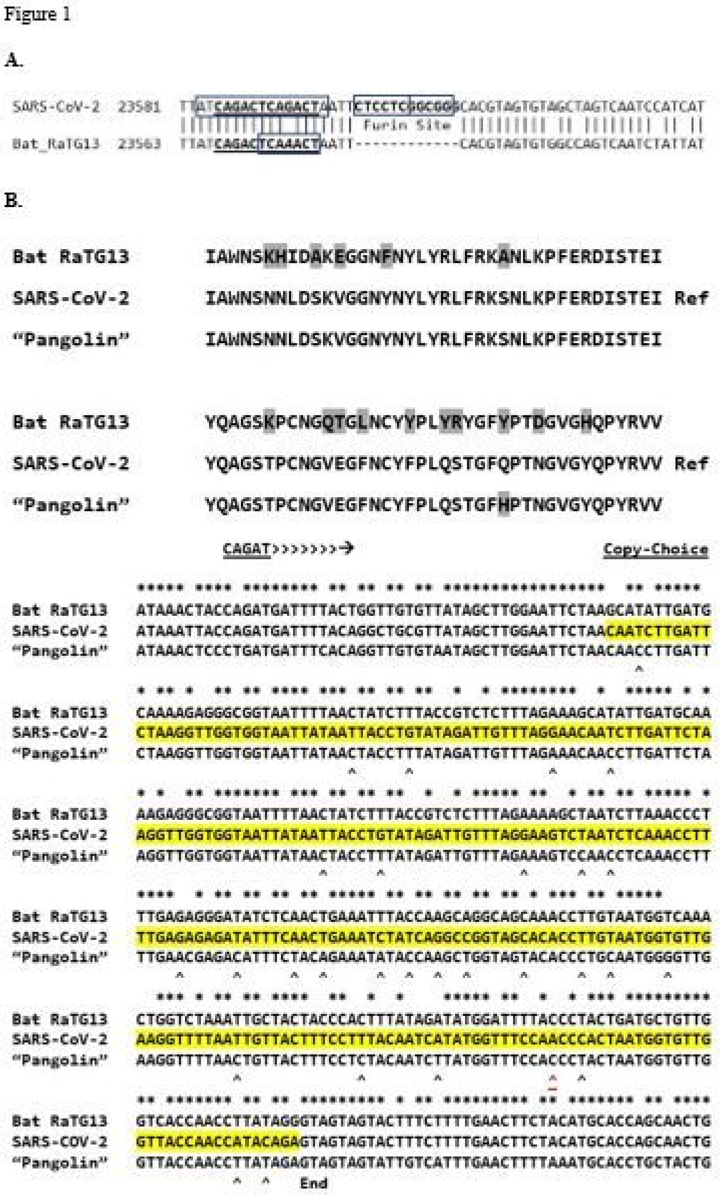
References

1. Zhu N et al (2020 A Novel Coronavirus from Patients with Pneumonia in China, 2019. New Engl J Med 382, 727-733
2. Zhou P et al (2020) A pneumonia outbreak associated with a new coronavirus of probable bat Nature 579, 270-273
3. Wu F et al (2020) A new coronavirus associated with human respiratory disease in China. Nature 579, 265- 269 (2020).
4. Drosten C et al. (2003) Identification of a novel coronavirus in patients with severe acute respiratory syndrome. N Engl J Med 348, 1967–1976.
5. Ksiazek TG et al (2003) A Novel Coronavirus Associated with Severe Acute Respiratory Syndrome, N Engl J Med 348:1953-1966
6. Peiris JSM et al (2003) Coronavirus as a possible cause of severe acute respiratory syndrome. Lancet 361:1319-1325
7. Li Q et al (2020) Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus-Infected Pneumonia. N Engl J Med 382:1199-1207
8. Zhao S et al (2020) Preliminary estimation of the basic reproduction number of novel coronavirus (2019- nCoV) in China, from 2019 to 2020: A data-driven analysis in the early phase of the outbreak. Int J Infect Dis 2020, 92:214-217
9. Gorbalenya AE et al (2020) The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. Nature Microbiology 5:536-544
10. Cui J, Li F, Shi ZL (2019) Origin and evolution of pathogenic coronaviruses. Nat Rev Microbiol 17:181- 192
11. Banerjee A et al (2019) Bats and Coronaviruses. Viruses 11, 41 (2019).
12. Li F (2026) Structure, Function, and Evolution of Coronavirus Spike Proteins. Annu Rev Virol 3, 237-261
13. Walls AC et al (2020) Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein. Cell 181:281-292
14. Thiel V et al (2003) Mechanisms and enzymes involved in SARS coronavirus genome expression. J Gen Virol 84:2305-2315

15. Coutard B et al (2020) The spike glycoprotein of the new coronavirus 2019-nCoV contains a furin-like cleavage site absent in CoV of the same clade. Antiviral Res 176, 104742
16. Menachery VD et al (2015) A SARS-like cluster of circulating bat coronaviruses shows potential for human emergence. Nature medicine 21, 1508-1513
17. Menachery VD et al (2016) SARS-like WIV1-CoV poised for human emergence. Proc Nat Acad Sci USA 113:3048-3053
18. Gallaher WR (2020) Tackling rumors of a suspicious origin of nCoV19 http://virological.org/t/tackling-rumors-of-a-suspicious-origin-of-ncov2019/384/17.
19. Andersen KG et al (2020) The proximal origin of SARS-CoV-2. Nat Med 26:450-452
20. Lai MM et al (1985) Recombination between nonsegmented RNA genomes of murine coronaviruses. J Virol 56:449–456
21. Makino S et al (1986) High-frequency RNA recombination of murine coronaviruses. J Virol 57:729-737
22. Graham RL, Baric RS (2010) Recombination, reservoirs, and the modular spike: mechanisms of coronavirus cross-species transmission. J Virol 84:3134 3146
23. Li X et al (2020) Emergence of SARS-CoV-2 through Recombination and Strong Purifying Selection. Sci Adv 29 May 2020: eabb9153 DOI: 1126/sciadv.abb9153
24. Lam TTY et al (2020) Identifying SARS-CoV-2 related coronaviruses in Malayan pangolins. Nature 10.1038/s41586-020-2169-0. doi:10.1038/s41586-020-2169-0
25. Boni MF et al (2020) Evolutionary origins of the SARS-CoV-2 sarbecovirus lineage responsible for the COVID-19 pandemic. bioRxiv 2020.03.30.015008; doi: https://doi.org/10.1101/2020.03.30.015008
26. Larkin MA et al (2007) Clustal W and Clustal X version 2.0. Bioinformatics 23, 2947-2948
27. Adedeji AO et al (2012) Mechanism of nucleic acid unwinding by SARS-CoV helicase. PLoS 7:e36521. doi:10.1371/journal.pone.0036521
28. Zhou H et al (2020) A novel bat coronavirus reveals natural insertions at the S1/S2 cleavage site of the Spike protein and a possible recombinant origin of HCoV-19. bioRxiv 2020.03.02.974139; doi: https://doi.org/10.1101/2020.03.02.974139.
29. Wertheim JO et al (2013) A case for the ancient origin of coronaviruses. J Virol 87:7039-45.
30. Liu Z et al (2020) Identification of common deletions in the spike protein of SARS-CoV-2. J Virol doi:10.1128/JVI.00790-20
31. Lin X-D et al (2017) Extensive diversity of coronaviruses in bats from China. Virology 507:1-10

32. Guan Y et al (2003) Isolation and characterization of viruses related to the SARS coronavirus from animals in southern China. Science 302:276-278
33. Hu B et al (2017) Discovery of a rich gene pool of bat SARS-related coronaviruses provides new insights into the origin of SARS coronavirus. PLoS Pathogens 13, e1006698
34. Barry, J (2004). The site of origin of the 1918 influenza pandemic and its public health Journal of Translational Medicine. 2. 3. 10.1186/1479-5876-2-3.
35. Taubenberger JK, Morens DM. (2006) 1918 influenza: the mother of all pandemics. Emerg Infect Dis 12:15-22.
36. Nichol ST et al (1993) Genetic identification of a hantavirus associated with an outbreak of acute respiratory illness. Science. 262:914-917
37. Goba A et al (2016) An Outbreak of Ebola Virus Disease in the Lassa Fever Zone. J Infect Dis 214(suppl 3):S110-S121
38. Kindhauser MK et al (2016) Zika: the origin and spread of a mosquito-borne virus. Bull World Health Organ. 94:675-686C

Figures



```
Figure 1

A.

SARS-CoV-2  23581  TTATCAGACTCAGACTAATTCTCCTCGGCGGGCACGTAGTGTAGCTAGTCAATCCATCAT
                   ||||||||||||||||||||| Furin Site |||||||||| || |||||||| || ||
Bat_RaTG13  23563  TTATCAGACTCAAACTAATT------------CACGTAGTGTGGCCAGTCAATCTATTAT

B.

Bat RaTG13      IAWNSKHIDAKEGGNFNYLYRLFRKANLKPFERDISTEI
SARS-CoV-2      IAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEI Ref
"Pangolin"      IAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEI


Bat RaTG13      YQAGSKPCNGQTGLNCYYPLYRYGFYPTDGVGHQPYRVV
SARS-CoV-2      YQAGSTPCNGVEGFNCYFPLQSTGFQPTNGVGYQPYRVV Ref
"Pangolin"      YQAGSTPCNGVEGFNCYFPLQSTGFHPTNGVGYQPYRVV

                  CAGAT>>>>>>>→                           Copy-Choice

            ***** **** ******** ** ** ** ** *****************  ** *****
Bat RaTG13  ATAAACTACCAGATGATTTTACTGGTTGTGTTATAGCTTGGAATTCTAAGCATATTGATG
SARS-CoV-2  ATAAATTACCAGATGATTTTACAGGCTGCGTTATAGCTTGGAATTCTAACAATCTTGATT
"Pangolin"  ATAAACTCCCTGATGATTTCACAGGTTGTGTAATAGCTTGGAATTCTAACAACCTTGATT
                                                                ^
            * ** *  ** ** **** *** ** ** **  *  * ********  *  ***** * *
Bat RaTG13  CAAAAGAGGGCGGTAATTTTAACTATCTTTACCGTCTCTTTAGAAAGCATATTGATGCAA
SARS-CoV-2  CTAAGGTTGGTGGTAATTATAATTACCTGTATAGATTGTTTAGGAACAATCTTGATTCTA
"Pangolin"  CTAAGGTTGGTGGTAATTATAACTACCTTTATAGATTGTTTAGAAACAACCTTGATTCTA
                                  ^     ^              ^     ^
            * *  ** ******* *** ** ** **  *  * ***** **  * ** ** ***** *
Bat RaTG13  AAGAGGGCGGTAATTTTAACTATCTTTACCGTCTCTTTAGAAAAGCTAATCTTAAACCCT
SARS-CoV-2  AGGTTGGTGGTAATTATAATTACCTGTATAGATTGTTTAGGAAGTCTAATCTCAAACCTT
"Pangolin"  AGGTTGGTGGTAATTATAACTACCTTTATAGATTGTTTAGAAAGTCCAACCTCAAACCTT
                               ^     ^              ^     ^  ^
            ****  * ** ** ** ** ***** ** ** ** ** ** * *** ** *****
Bat RaTG13  TTGAGAGGGATATCTCAACTGAAATTTACCAAGCAGGCAGCAAACCTTGTAATGGTCAAA
SARS-CoV-2  TTGAGAGAGATATTTCAACTGAAATCTATCAGGCCGGTAGCACACCTTGTAATGGTGTTG
"Pangolin"  TTGAACGAGACATTTCTACAGAAATATACCAAGCTGGTAGTACACCCTGCAATGGGGTTG
                ^     ^     ^  ^     ^  ^  ^  ^     ^     ^  ^     ^
              *** * ** ** ****  **  *  *    ***** **  *  * *** *********
Bat RaTG13  CTGGTCTAAATTGCTACTACCCACTTTATAGATATGGATTTTACCCTACTGATGCTGTTG
SARS-CoV-2  AAGGTTTTAATTGTTACTTTCCTTTACAATCATATGGTTTCCAACCCACTAATGGTGTTG
"Pangolin"  AAGGTTTTAACTGTTACTTTCCTCTACAATCTTATGGTTTCCACCCTACTAATGGTGTTG
                      ^            ^       ^           ^  ^
            ** ******* ** ** ********* * ** ********* ** ******* ** ****
Bat RaTG13  GTCACCAACCTTATAGGGTAGTAGTACTTTCTTTTGAACTTCTACATGCACCAGCAACTG
SARS-COV-2  GTTACCAACCATACAGAGTAGTAGTACTTTCTTTTGAACTTCTACATGCACCAGCAACTG
"Pangolin"  GTTACCAACCTTATAGAGTAGTAGTATTGTCATTTGAACTTTTAAATGCACCTGCTACTG
                      ^  ^   End
```

Figure 1

Alignments within and at the boundaries of recombinant sequences in the Spike (S) protein. A. Alignment beginning at nt23581 of SARS-Cov-2 and the corresponding nt23563 of RaTG13 through the RNA code of the furin site insert. Underlining shows the tandem repeat in RNA of SARS-CoV-2, i.e. CAGACT. Boxes highlight the several palindromic sequences in the SARS-CoV-2 (CAGAC, CAGAC, the outer ATCAGACTCAGACTA, as well as two within the insert itself, CTCCTC and GGCGG. A box also highlights the palindromic TCAAACT created by mutation of the tandem repeat in RaTG13. B. Alignment of both the peptide (in standard single letter code) and RNA sequences in the region of the receptor binding domain (RBD) of RaTG13, SARS-CoV-2 and BetaCoV/pan/P1L/2019, beginning at nt22838 of SARS-CoV-2. Amino acid differences between SARS-CoV-2 and the other two viruses are highlighted in grey shading. In the RNA alignment, the recombinant sequence is highlighted in yellow shading. The two non-synonymous mutations between SARS-CoV-2 and BetaCoV/pan/P1L/2019 are noted by red arrows. Synonymous wobble base mutations between the two are indicated by blacks arrows, 28 of 268, a divergence of 10.4%, indicative of decades of divergence between the pangolin sequence and its ancestor that recombined with the ancestor of RaTG13 to yield the sequence of SARS-CoV-2.

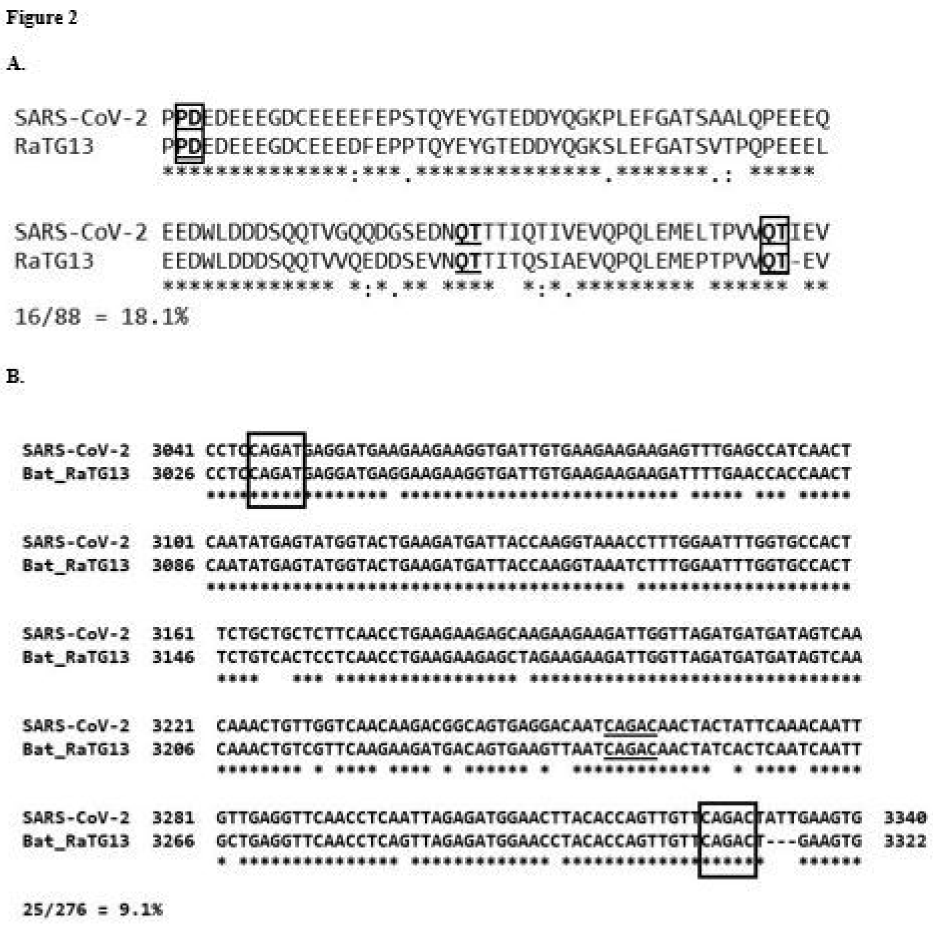


Figure 2

A.

```
SARS-CoV-2 PPDEDEEEGDCEEEEFEPSTQYEYGTEDDYQGKPLEFGATSAALQPEEEQ
RaTG13     PPDEDEEEGDCEEEDFEPPTQYEYGTEDDYQGKSLEFGATSVTPQPEEEL
           **************:***.**************.*******.: *****

SARS-CoV-2 EEDWLDDDSQQTVGQQDGSEDNQTTTIQTIVEVQPQLEMELTPVVQTIEV
RaTG13     EEDWLDDDSQQTVVQEDDSEVNQTTITQSIAEVQPQLEMEPTPVVQT-EV
           ************* *:*.** ****  *:*.********* ****** **
16/88 = 18.1%
```

B.

```
SARS-CoV-2  3041 CCTCCAGATGAGGATGAAGAAGAAGGTGATTGTGAAGAAGAAGAGTTTGAGCCATCAACT
Bat_RaTG13  3026 CCTCCAGATGAGGATGAGGAAGAAGGTGATTGTGAAGAAGAAGATTTTGAACCACCAACT
                 ***************** ************************** ***** *** *****

SARS-CoV-2  3101 CAATATGAGTATGGTACTGAAGATGATTACCAAGGTAAACCTTTGGAATTTGGTGCCACT
Bat_RaTG13  3086 CAATATGAGTATGGTACTGAAGATGATTACCAAGGTAAATCTTTGGAATTTGGTGCCACT
                 *************************************** ********************

SARS-CoV-2  3161  TCTGCTGCTCTTCAACCTGAAGAAGAGCAAGAAGAAGATTGGTTAGATGATGATAGTCAA
Bat_RaTG13  3146  TCTGTCACTCCTCAACCTGAAGAAGAGCTAGAAGAAGATTGGTTAGATGATGATAGTCAA
                  ****   *** ***************** *******************************

SARS-CoV-2  3221  CAAACTGTTGGTCAACAAGACGGCAGTGAGGACAATCAGACAACTACTATTCAAACAATT
Bat_RaTG13  3206  CAAACTGTCGTTCAAGAAGATGACAGTGAAGTTAATCAGACAACTATCACTCAATCAATT
                  ******** * **** **** * ****** *  *************  * **** *****

SARS-CoV-2  3281  GTTGAGGTTCAACCTCAATTAGAGATGGAACTTACACCAGTTGTTCAGACTATTGAAGTG  3340
Bat_RaTG13  3266  GCTGAGGTTCAACCTCAGTTAGAGATGGAACCTACACCAGTTGTTCAGACT---GAAGTG  3322
                  * *************** ************* *******************   ******

25/276 = 9.1%
```

Figure 2

Alignment of amino acid and RNA sequence in the putative recombinant region within nsp3 of Orf1a. A. Alignment of SARS-CoV-2 and RaTG13 in standard single-letter amino acid code for nsp3, beginning just prior to the acidic-rich region. The boxed PD indicates the location of the CAGAT "breakpoint sequence", in this case out of frame with the nsp3 reading frame. The boxed QT indicates the site of the CAGAC "breakpoint sequence", in frame with nsp3, just prior to a single amino acid insertion in SARS-CoV-2 relative to RaTG13. The overall difference in amino acid sequence is 18.1% B. The corresponding alignment in nsp3 for RNA sequences of SARS-Co-V-2 beginning at nt3041, and the corresponding nt3026 of RaTG13. The "breakpoint sequences CAGAC and CAGAT are underlined or boxed. A divergence of 9.1% is observed, commensurate with decades of divergence between RaTG13 and the source of the recombinant sequence, as yet unsampled per BLAST analysis.

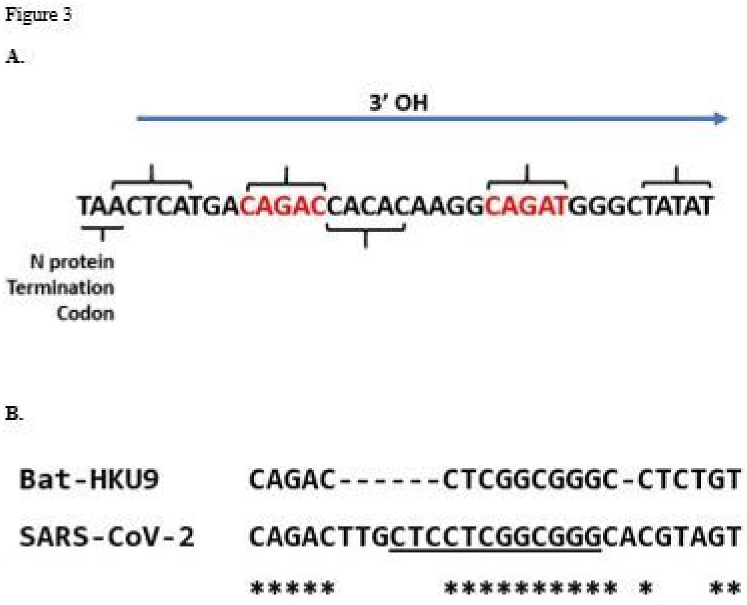


Figure 3
A.
3' OH
TAACTCATGACAGACCACACAAGGCAGATGGGCTATAT
N protein
Termination
Codon
B.
Bat-HKU9 CAGAC------CTCGGCGGGC-CTCTGT
SARS-CoV-2 CAGACTTGCTCCTCGGCGGGCACGTAGT
***** ********** * **

Figure 3

RNA sequences of the common 5’ terminus of the 3’OH region and a putative source of the furin insert RNA. A. RNA sequence of the common 5’ terminus of the 3’OH region of SARS-Co2, RaTG13 and BetaCoV/pan/P1L/2019. RNA sequence begins with the termination codon for the nucleocapsid (N) protein TAA, and continues through a region replete with five non-overlapping pentanucleotide palindromic sequences, indicated by brackets, including one each of the “breakpoint sequences”, CAGAC and CAGAT, highlighted in red type. B. Alignment of SARS-CoV-2 in the furin insert region with a candidate source of the insert sequence, Bat-HKU9-1. Alignment begins at nt23601 of SARS-CoV, paired with nt24190 of Bat-HKU9-1 corresponding to a peptide region 29 amino acids prior to the beginning of the heptad repeat 1 region of SARS-CoV-2, 206 amino acids from the beginning of S2. The RNA code for the insert in underlined, with 21 of 28 nculeotides identical, and two gaps.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- GALLAHERFigS1ALL.JPG