The murky origins of the coronavirus SARS-CoV-2, the causative agent of the COVID-19 pandemic

P. Balaram

'*...whenever a new and startling fact is brought to light in science, people first say, "it is not true," then that "it is contrary to religion," and lastly, "that everybody knew it before."* '

– Louis Aggassiz

'*I suppose the process of acceptance will pass through the usual four stages: 1. This is worthless nonsense, 2. This is an interesting, but perverse, point of view, 3. This is true, but quite unimportant, 4. I always said so.*'

– J. B. S. Haldane

The coronavirus with its spherical surface decorated with innumerable spikes may well become the defining image of our times. No image produced by scientists has ever before been so readily recognized by millions and possibly billions of people around the world. The ability of an invisible particle of about 100 nm diameter to bring the world to its knees reminds us that nature can trump the dominant forces that govern the world today, politics and religion. Where did this fearsome pathogen emerge from in late December 2019? Virologists readily accepted the conventional explanation of zoonotic transfer, where animal reservoirs of pathogenic viruses, bats being the usual culprits, occasionally breach species barriers by jumping to intermediate hosts before infecting humans. In previous coronaviral outbreaks, 2003 (SARS1) and 2007 (MERS), the intermediate hosts had been identified, civets in the former and camels in the latter. In early 2020, when scientists by the hundreds turned their focus to coronaviruses, the Wuhan wet market in China, with its collection of esoteric live animals, appeared the place to look for the intermediate animal hosts. None has been found so far. A way out to explain this failure is to argue that this was a direct case of transmission from bats to humans, although it remains a conjecture with no evidence. An even more unlikely explanation is that the virus appeared through a contaminated cold food chain. Finally, there is the possibility (however improbable) that the virus was engineered in a Wuhan laboratory and an accident, not uncommon even in high safety laboratories, allowed direct human infection and subsequent human–human transmission. This last scenario was, of course, the favourite amongst those fond of conspiracy theories, most notably the former US President Donald Trump, who famously and publicly christened the SARS-CoV-2 as the 'Chinese virus'. Even a joint G-7 statement was scrapped when its members refused to endorse the term 'Wuhan virus'. The American scientific establishment predictably closed ranks and high profile groups of scientists published letters in major journals arguing that the virus was clearly the handiwork of nature, through the glacially slow processes of evolution by random mutational processes and natural selection. Both evolutionary processes and laboratory genetic manipulation can leave subtle imprints on the sequences of proteins involved in mediating the transition of a bat virus into a virulent human pathogen. Early on in the pandemic in March–April 2020 scientists of little repute, but connoisseurs of sequences, began to point out that the sequence of the SARS-CoV-2 glycoprotein may hold the key to the mystery of the remarkable virulence of the causative agent of COVID-19. This commentary makes no effort to examine the vast literature on SARS-CoV-2 but is stimulated by the sudden spurt of interest in the American scientific community, which resulted in a second letter to *Science*, 14 May 2021, which notes: '*We must take hypotheses about both natural and laboratory spillovers seriously until we have sufficient data. A proper investigation should be transparent, objective, data-driven*' (doi:10.1126/science.abj0016). Curiously, these authors, even while citing CNN, make no reference to an earlier letter in *The Lancet* that appeared in February 2020, which was emphatic in its dismissal of 'conspiracy theories', by citing high scientific authorities: '……*sharing of data on this outbreak is now being threatened by rumours and misinformation around its origins. We stand together to strongly condemn conspiracy theories suggesting that COVID-19 does not have a natural origin. Scientists from multiple countries have published and analysed genomes of the causative agent, severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), and they overwhelmingly conclude that this coronavirus originated in wildlife, as have so many other emerging pathogens. This is further supported by a letter from the presidents of the US National Academies of Science, Engineering, and Medicine and by the scientific communities they represent. Conspiracy theories do nothing but create fear, rumours, and prejudice that jeopardise our global collaboration in the fight against this virus. We support the call from the Director-General of WHO to promote scientific evidence and unity over misinformation and conjecture*' (https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(20)30418-9/fulltext). The cast of authors in this letter bears careful scrutiny. Their contributions in reopening discussions, sponsored by the US National Academies on dangerous 'gain of function' researches, which led to the US National Institutes of Health, lifting an earlier ban on funding for such research, merit attention. It is specific undisclosed conflicts of interest of one of the authors, now reputed to be the motivating force behind the letter, which has resulted in the current exploding interest in the possible laboratory origins of the SARS-CoV-2 virus. A second letter in April 2020 by other prominent researchers in *Nature Medicine* weighed in decisively: '*Our analyses clearly show that SARS-CoV-2 is not a laboratory construct or a purposefully manipulated virus*' (https://doi.org/10.1038/s41591-020-0820-9). A careful reading of the evidence, marshalled by these authors, favouring natural evolution of the virus, suggests that they were overstating their case. A year ago the scientific establishment in the US, which invariably dominates scientific discourse, seemed to be united in its views, a clear example of the scientific orthodoxy closing ranks in the face of new and unusual observations beginning to emerge from inspections of the SARS-CoV-2 genome and more specifically,

the unusual features of the spike protein amino acid sequence. What triggered the early Internet conspiracy theories, some of which were apparently postings based on sequence analysis, and what is responsible for the remarkable volte-face by the scientific establishment in the US?

Early rumblings: the end of the beginning

One of the early credible signs that all may not be well with the natural origins hypothesis came from a posting on *bioRxiv*, the open access repository for biology, by a postdoctoral researcher from the Broad Institute at Harvard-MIT, USA, Alina Chan, who raised substantive concerns. These were, of course, quickly criticized by prominent scientists, including Peter Daszak of the EcoHealth Foundation, USA, whose association with the Wuhan Institute of Virology was not publicly known at that time. A *Boston Magazine* article in September 2020, provides a fascinating account of a young, female researcher raising uncomfortable questions, that are magisterially dismissed by senior established researchers (https://www.bostonmagazine.com/news/2020/09/09/alina-chan-broad-institute-coronavirus/). Chan then retreated to the safer and apparently more democratic medium, of Twitter, describing herself as a 'scientist turned detective'. Her tweetorials, which make interesting reading, as she worked tirelessly to dismantle the increasingly flimsy arguments in favour of zoonotic transfer, are based on sequences reported from Wuhan (https://twitter.com/ayjchan/status/1391753059504738308). The formidable scientific establishment exerts far greater control over the most prestigious scientific journals than commonly imagined. Heretics are easily banished from sight, consigned to the dark corners of the internet, the modern equivalent of the stake on which Giordano Bruno was burnt. The wheel appears to have turned the full circle. Chan is now a co-author of the latest letter to *Science*, which argues that the laboratory origin of the SARS-CoV-2 virus must be taken seriously. She is in the midst of a distinguished group of authors, all leaders in their fields, from the best of institutions. If there is a lesson here for senior, established scientists it is this: It is always advisable to listen carefully to younger and committed colleagues, who are often more familiar with the mind-numbing details of analysing data, than those who work at higher levels of the scientific stratosphere. The wall of resistance truly crumbled with the appearance of the deeply investigated article by Nicholas Wade, a veteran of science journalism, in the *Bulletin of Atomic Scientists*. Wade's highly readable and detailed account sets the cat among the pigeons with a telling commentary on the consequences of 'virologists *omerta*' and the fact that '*science reporters, unlike political reporters, have little innate skepticism of their sources' motives; most see their role largely as purveying the wisdom of scientists to the unwashed masses. So when their sources won't help, these journalists are at a loss*'. One quotation from Wade's article merits reproduction: ' "*When I first saw the furin cleavage site in the viral sequence, with its arginine codons, I said to my wife it was the smoking gun for the origin of the virus,*" *said David Baltimore, an eminent virologist and former president of CalTech.* "*These features make a powerful challenge to the idea of a natural origin for SARS2,*" *he said*' (https://thebulletin.org/2021/05/the-origin-of-coviddid-people-or-nature-open-pandoras-box-at-wuhan/). Baltimore, the 1975 Nobel laureate, one of the high priests of molecular biology and the co-discoverer of the enzyme reverse transcriptase, central to the RT-PCR diagnostic for SARS-CoV-2, is not a voice to be easily dismissed. One can only wonder, why did he not voice his suspicions earlier.

On the trail of Baltimore's smoking gun

'*What one man can invent another can discover.*'

– Arthur Conan Doyle,
The Return of Sherlock Holmes

The principal actor in the hypothetical drama of the laboratory engineered virus is the protein that constitutes the spiky projections on the pathogen surface, the spike glycoprotein or more simply, the spike protein (Figure 1). The immunologist Peter Medawar in his dictionary of biology, *From Aristotle to Zoos*, defined a virus, with an anonymous quote, as '*a piece of bad news wrapped up in a protein*'. The bad news is, of course, the genetic material, ribonucleic acid (RNA), in the case of coronaviruses that needs to be delivered to the interior of a human cell (in the case of SARS-CoV-2). Viruses which inhabit the shadowy no-man's land between chemistry and biology, are quintessential parasites that subvert the biochemical machinery of a host cell in order to reproduce. It is the spike protein (Figure 1) that guides the virus to a specific receptor protein site (the angiotensin-converting enzyme, ACE 2) on the surface membrane of human host cells. Docking to a host cell is necessary, but insufficient to gain entry. Another step which involves the mediation of a host cell enzyme, furin, is needed to break the spike protein, approximately at its centre, to set in motion a complex set of events that lead to membrane fusion and viral entry. Furin, a protease, is an enzyme that breaks specific peptide bonds, that link the over 1200 amino acids together in the long polymeric backbone of the spike protein. This biochemical scissor then cleaves the long spike protein into two segments in preparation for viral entry. This is a key step in infection; the more efficient the cleavage, the more effectively will the virus breach the protective barriers of its reluctant host. The genome sequence and by extension the spike protein sequence of the SARS-CoV-2 virus isolated from a patient, was made available from the Wuhan laboratory in January 2020. Almost immediately, an unusual feature became visible. The new virus contained a furin cleavage site that appeared almost optimal for furin cleavage, clearly different from the sequence of the earlier SARS-CoV-1 virus, which caused the 2003 disease outbreak. Where did this new feature emerge from? There was certainly a dramatic 'gain of function', an enhanced ability to infect that became evident as the pandemic exploded worldwide. Could the often obscure and slow processes of evolutionary change be responsible for the emergence of this new and most virulent pathogen? Or is a human hand visible in the fashioning of the virus? On the surface, these are not easy questions to answer, even by those steeped in virology and genome sequence analysis, with the limited data publicly available. But the smoking gun may still be found, hidden in the sequences of the spike proteins from bats and humans. Figure 2 shows a comparison of the furin cleavage site sequences

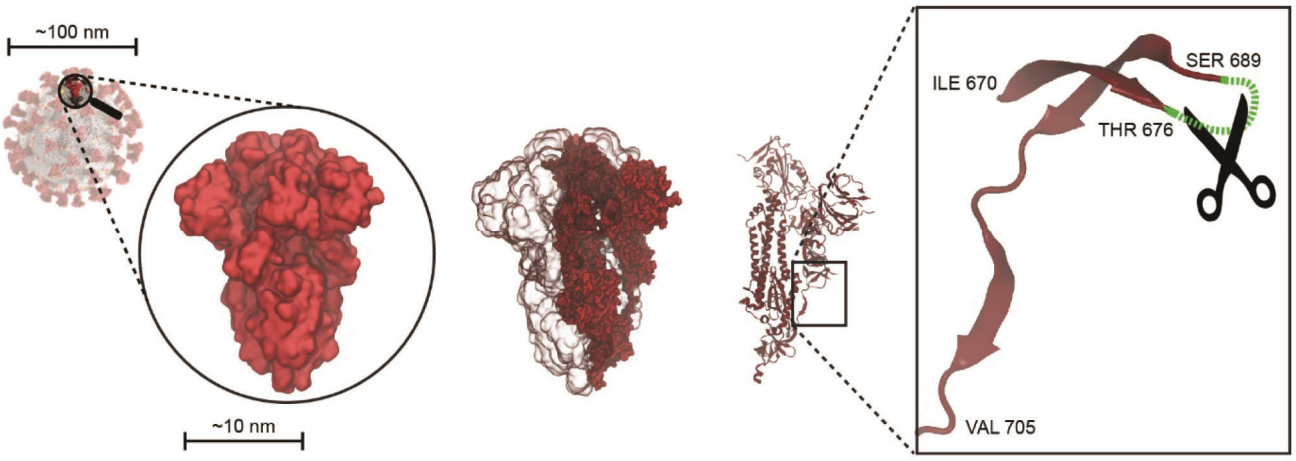

Figure 1. Image of the coronavirus (CDC, Atlanta, USA) with expansion of the trimeric spike, highlighting one protein molecule. (far right) atomic level structure of the protruding part of the trimeric spike protein, with an expanded view of the long projecting loop harbouring the furin cleavage site. The furin site is missing, presumably disordered (green loop), in the determined structure (Protein Data Bank Code:6VXX)

Coronavirus/seq no(SARS CoV2)	680											690					695			
Bat RmYN01(AGC74176.1)	C	A	S	Y	H	T	A	S	-	L	L	-	-	-	-	R	N	T	G	Q
Bat RaTG13(QHR63300.2)	C	A	S	Y	Q	T	Q	T	N	S	-	-	-	-	-	R	S	V	A	Q
SARS-CoV2(CAD0240766.1)	C	A	S	Y	Q	T	Q	T	N	S	-	P	R	R	A	R	S	V	A	Q
SARS-CoV1(P59594)	C	A	S	Y	H	T	V	S	-	L	L	-	-	-	-	R	S	T	S	Q
MERS(K9N5Q8)	C	A	L	P	D	T	P	S	T	L	T	P	R	S	V	R	S	V	P	G
Human 229E(P15423)	C	A	D	G	S	I	I	A	V	Q	-	P	R	N	V	-	S	Y	D	S
Human NL63(Q6Q1S2)	C	A	D	G	S	L	I	P	V	R	-	P	R	N	-	S	S	-	D	N
Human HKU1(Q0ZME7)	C	I	D	Y	A	L	P	S	S	-	-	R	R	K	R	R	G	I	S	S
Human OC43(P36334)	C	V	D	Y	S	K	N	-	-	-	-	R	R	S	R	G	A	I	T	T

Figure 2. Comparison of the spike protein segment containing the furin cleavage site across viruses specific for bat and human hosts. The top two rows are bat sequences. The middle three are the agents of severe human disease. The last four rows are sequences from the coronaviruses generally causing relatively mild respiratory infections. The blue coloured rows highlight the identity of the segments flanking the furin cleavage site in the virus responsible for COVID-19 and a bat virus. Residues conserved at the furin cleavage site (690–695) are highlighted in yellow. The NCBI accession numbers for the sequences are indicated in parentheses.

of bat coronaviruses closely related to SARS-CoV-2, the three infectious coronaviruses which have caused disease in the 21st century and the much less dangerous coronaviruses, isolated from humans suffering from relatively mild respiratory infections. Interestingly, the sequences of the bat RaTG13 and the current pathogen, SARS-CoV-2 are practically identical across a length of over 1200 letters (amino acids), with the offending PRRAR segment sticking out like a sore thumb. The question forbidden by the scientific establishment throughout 2020, was whether this bait for attracting furin cleavage was deliberately engineered into a bat template, thereby 'humanizing' the bat virus. Experiments using 'humanized' mouse models, where the mouse now carries an engineered human ACE receptor, thus permitting infection by a virus specific for human host, have been done. It is then a small step to test the effects of engineering a new cleavage site and thus produce more 'infectious' viruses. It is such '*gain of function*' experiments that have been so widely debated in the United States, leading to even the lifting of a ban on funding for such experiments.

How would one design a better furin site to enhance human infection by a bat virus? The clues may lie in Figure 2.

The relatively mild coronaviruses isolated from humans decades ago, Dorothy Hamre's original isolate 229E (https://science.thewire.in/the-sciences/finding-dorothy-hamre-the-first-person-to-isolate-a-strain-of-a-coronavirus/) and NL-63 contain the doublet sequence PR. The other two previously studied human pathogens contain the more highly basic segments RRKRR (HKU1) and RRSRG (OC43). Is the furin site absolutely necessary for viral infectivity? No. Alternative cleavage pathways, much less efficient than furin, can substitute and promote, albeit poorly, virus entry into the host cell. Furin does the job much better. Furin homes in on the basic segments rich in the residue arginine (R). This inviting cleavage site projects outwards as a long unstructured loop, hanging out as a bait to attract furin. Indeed, in the crystal structure of SARS-CoV-2 (Figure 1), this magnet for the protease furin is invisible, disordered and disobeys the dictates of local symmetry. What has been outlined above is a 'thought experiment', creating a possible obvious rationale for choosing a PRRAR segment for insertion into a bat sequence, in order to 'humanize' it. Such manipulations must, of course, be done at the level of the gene by altering the sequences of nucleotides, a process

cct	*cgg*	*cgg*	*gca*	*cgt*
P	**R**	**R**	**A**	**R**

Figure 3. Codons used for the furin cleavage site in SARS-CoV-2.

which has often captivated the public imagination, loosely described as 'genetic engineering'. It is there that we must search for Baltimore's 'smoking gun'. Figure 3 shows the sequence of the nucleic acid bases corresponding to the PRRAR segment of SARS-CoV-2.

The RR amino acid sequence is coded by the triplet codons ***cggcgg***. For the uninitiated, three letters of the nucleic acid alphabet (which contains only 4 letters) translate into single letters of the protein alphabet (which contains 20 letters). Arginine (R) is coded for by as many as six triplet codons. The frequency of occurrence (%) of these codons across the eight natural viruses in Figure 2 (omitting SARS-CoV-2) is: ***agg***16.1, ***aga***30.6, ***cga***6.6, ***cgt***32.8, ***cgg***3.5, ***cgc***10.4. It is clear that the most infrequent codon used in nature is ***cgg***. Yet this is found to code for the contiguous RR segment in the causative agent of the ongoing pandemic. Is this Baltimore's smoking gun? If it is then the barrel appears hot enough to singe the hand. Even as this commentary is being written, an article authored by groups at the Imperial College, London and the University of Sheffield has appeared with the provocative title, '*The SARS-CoV-2 variants associated with infections in India, B.1.617, show enhanced spike cleavage by furin*' (https://doi.org/10.1101/2021.05.28.446163). The variant B.1.617, now widely spread in India and elsewhere, carries a single mutation at the furin cleavage site. The PRRAR bait for furin is now mutated to RRRAR, a mutational change that requires only a flip of a single letter in the coding triplets. Whatever be the origin of the furin site, deliberate design or an accident of biology, random mutations by successive passages through human hosts, followed by natural selection for the phenotype of enhanced cleavage, may have indeed been at play. The article goes on to use small peptide substrates mimicking the viral cleavage sites to argue that this single mutation may indeed contribute substantially to transmissibility of the new viral strain. If this scenario is indeed true, then nature can quickly improve on human constructs given the rapid viral replication time scales. For the mutant watchers inspecting the growing database of SARS-CoV-2 sequences, this raises the issue of which site is likely to enhance viral infectivity as a consequence of mutation, spike protein receptor binding domain (RBD) or the furin cleavage site.

Aficionados of biomolecular recognition will recognize that the step of virus binding to its host-membrane receptor, is determined by the physics of interatomic interactions, non-covalent in nature, a process determined by diffusion and collision. The cleavage step goes beyond recognition of the cleavage site on the spike protein by the enzyme furin. This step necessarily requires chemistry to break covalent bonds. Loosely, physics is faster than chemistry. One would imagine that speeding up the slower step, 'rate determining' in the parlance of chemistry, would be an effective approach to enhancing viral entry into cells.

The beginning of the end

'*Survival ... is an infinite capacity for suspicion.*'

– George Smiley in John Le Carre's *Tinker, Soldier, Tailor, Spy*

More flags have been raised on the sequence data on bat coronaviruses deposited by the Wuhan laboratory. The abstract of a very recent analysis (*A reconstructed historical Aetiology of the SARS coronavirus-2 spike* by B. Sorensen *et al.*, slated to appear in the *Quarterly Review of Biophysics Discovery*), concludes: '*Henceforth, those who maintain the zoonotic transfer hypothesis need to explain precisely why our simpler model of laboratory manipulation is wrong, before asserting that their evidence is persuasive*' (https://www.dailymail.co.uk/news/article-9629563/Chinese-scientists-created-COVID-19-lab-tried-cover-tracks-new-study-claims.html). The various hypotheses to explain the origins of the coronavirus SARS-CoV-2 still lack definitive evidence to establish their veracity beyond doubt. That evidence is unlikely to be forthcoming in the near future, hidden as it is (or even destroyed) behind the impenetrable wall of secrecy that surrounds coronavirus research in China. However, it is becoming increasingly likely that the gain of function research on pathogenic viruses, so vigorously defended in the US by influential sections of the scientific establishment, may have contributed significantly to Chinese efforts in this area. Thus far, the major scientific journals, which act as gatekeepers for the credibility of the scientific literature have refrained from weighing in on the controversy surrounding the origins of the coronavirus. Their own credibility has been strained by their uncritical publication of correspondence last year, declaring that a natural origin for the virus was almost a foregone conclusion. Has a '*prima facie*' case, a phrase beloved by our hyperactive investigative agencies, been established for the laboratory origin of the coronavirus? Sifting through the available evidence should challenge the best of science detectives, but we might well remember the immortal words of Sherlock Holmes: '*… when you have eliminated the impossible, whatever remains, however improbable, must be the truth.*'

Note added in proof: In this fast-moving detective story some of the key clues have been unearthed by those who work in the shadows of high-profile science. Readers may like to see the following: Rahalkar, M. C. and Bahulikar, R. C., *Front. Public Health*, 20 October 2020, https://doi.org/10.3389/fpubh.2020.581569; a team of largely anonymous internet detectives https://drasticresearch.org/ and those who marshal data, but whose analyses are consigned to what are dismissively termed as 'low impact' journals, Segreto, R. *et al.*, *Environmental Chemistry Letters*, https://doi.org/10.1007/s10311-021-01211-0.

ACKNOWLEDGEMENT. I am grateful to M. Vijayasarathy and Sahil Lall for their help in drawing the Figures and for many helpful discussions.

P. Balaram is at the National Centre for Biological Sciences, Bengaluru 560 065, India.
e-mail: pb@iisc.ac.in