Investigations into the origin of SARS-CoV-2: an update

Anirban Mitra*

Department of Biotechnology, Institute of Genetic Engineering, Badu 700 128, India

Since January 2020, scientists have been using both experimental and bioinformatic approaches to study the key molecular features of SARS-CoV-2, the causative agent of COVID-19. These studies have established that the genome of this virus is overall similar to that of viruses found in bats. However, there are genomic stretches which show strong similarity with viruses identified from other animals. The rapid developments of this subject have provided insights into how this novel virus has evolved from a number of progenitors and gained attributes that have made it a formidable pathogen. This review presents the salient features of these peer-reviewed findings and how the scientific evidence contradicts the 'conspiracy theories' floating around.

Keywords: ACE2 receptor, betacoronavirus, COVID, pangolins, SARS-CoV-2, spike protein.

SINCE the beginning of 2020, perhaps no question has rocked humanity more than 'where did this novel coronavirus come from?'. It has been asked again and again whenever people meet and exchange opinions and information, and a recurring answer has been, 'this must have been genetically-engineered…' or 'accidentally released from a lab…'. This 'belief' – that SARS-CoV-2 is a dark product of human innovation – is a strong one, strengthened by formal political statements, millions of informal gatherings and echo-chambers of social media. What is noticeable however, is that professional biologists have largely stayed away from these heated debates. Instead, they have done what they do – pile up scientific results to search for the answer. A wealth of scientific literature has delved into the molecular features of the causative agent of COVID-19. The question that scientists, working in several labs across the world, have asked is 'what does scientific investigation tell us about the origin and evolution of SARS-CoV-2?.' Several research articles have been published so far. This review presents their findings and summarizes the overall scientific understanding of the subject.

It is, of course, well-established that all diseases and epidemics known till date – malaria, TB, cholera, typhoid, smallpox, dengue, chicken pox, AIDS – have been caused by natural pathogens. The causative agents of these maladies are all products of biological evolution. None of them were manufactured by human hands. Another noteworthy point is that a significant number of pathogens that cause disease today have actually jumped from animals to humans. Sixty per cent of known infectious diseases and 75% of emerging pathogens are of zoonotic origin[1,2]. There is not anything surprising about this; most pathogens have fine-tuned their lives with their hosts. But, once in a while, one among them gains access to a new species. Prominent examples of such zoonosis are influenza viruses, which have jumped from birds, and the human immunodeficiency virus (HIV) which originated in African apes before evolving to enter a related species – *Homo sapiens*[3]. The best known recent example is the severe acute respiratory syndrome (SARS) coronavirus (referred to as SARS-classic in this article) which crossed over from bats and palm civets to humans and caused the pandemic of 2003 (ref. 4). All biologists are aware of these fundamentals. Hence, regardless of what social media tells us, virologists and epidemiologists were certain that the starting point for their research would be the vast natural world of Virosphere. Indeed, that is where the salient answers come from.

A new pathogen

In early January 2020, when several patients turned up in large numbers in Wuhan's (China) hospitals with symptoms like dry cough, fever and pneumonia that progressed to respiratory distress and fatal alveolar damage[5], Chinese scientists noted that many (but not all) patients had links to the city's wet and seafood market[6]. They also recalled that, in addition to the symptoms and a possible animal connection, the new disease had emerged in winter – similar to the 2002–2003 SARS pandemic[5]. Furthermore, antibiotics were ineffective, indicating that the causative agent was not bacteria. This necessitated that virologists scan the patient samples for coronaviruses.

The initial reverse transcription polymerase chain reaction (RT-PCR) tests searched for any coronavirus and five out of seven patient samples gave positive results. Subsequent metagenomic analysis and allied experiments identified a viral genome that was 29,891 nt (nucleotide) long. Initially found in one patient's sample, the same viral sequence (99.9% identical to each other) was soon

*e-mail: informanirbanmitra@gmail.com

observed in all the positive samples[5]. The next obvious step was bioinformatic analysis of the viral genome, to understand its important molecular features and identify its closest relatives. Comparative analysis of viral sequences (in sync with experimental studies) allows us to group viruses into families and genera. And, in this case, the new virus was identified to be a coronavirus – a member of the family Coronaviridae and it had 79.6% sequence identity with the genome of the SARS-classic virus[5]. They tentatively named it novel coronavirus 2019 (nCoV-2019), till the International Committee on Taxonomy of Viruses gave it the formal name SARS-CoV-2 (ref. 7).

Coronaviruses – once ignored, now worrisome

Coronaviruses are not particularly new. First identified in 1968 (ref. 8), several have been discovered over the decades. They are now grouped into four genera – Alpha, Beta, Gamma and Delta. The alphacoronavirus and betacoronaviruses infect mammals, while the gammacoronavirus and deltacoronaviruses are mainly avian pathogens[9]. For years, coronaviruses had largely been restricted to small paragraphs in textbooks of microbiology; ~15% of common colds are caused by coronaviruses. Two alphacoronaviruses NL63 and 229E, and two betacoronaviruses OC43 and HKU1 are causative agents of mild respiratory illnesses. But, medical and public consciousness about them rocketed up with the emergence of two highly pathogenic betacoronaviruses: SARS-classic of 2002–2003 and the Middle East respiratory syndrome (MERS) virus of 2012 (refs 9, 10). Both pathogens killed a big percentage of the people they infected and, notably, both had entered humans from animals before acquiring the ability to spread from one person to another. The ancestors of both viruses had been residents in bats and then reached humans via an intermediate host – civets for SARS-classic virus and dromedary camels for MERS virus[11,12]. The recurring pattern of viral entry, coupled with the knowledge that unbridled deforestation and illegal wildlife trade was bringing human populations and wildlife (and viruses resident in them) in dangerously close proximity, set the alarm bells ringing. Identifying the animal hosts of SARS-CoV-2 went beyond academic interest. It was essential to understand which animals could be significant 'viral depots' in the future. Economic and political policies would have to be re-wired accordingly. By the end of the last decade, scientists had already cautioned that other pathogenic coronaviruses could also come the way SARS-classic and MERS had evolved, and now the spread of SARS-CoV-2 seemed to validate their concern[13].

Sequence analysis showed that SARS-CoV-2 genome has 14 Open Reading Frames (ORFs). Six of these – ORF1a, ORF1b, S, E, M and N – are found in all coronaviruses. Together, they encode for 27 proteins. This is because the first two ORFs – named ORF1a and ORF1ab – encode proteins pp1a and pp1ab, from which a total of 15 non-structural proteins (nsps) are carved out. In addition, the genome houses four structural genes (*S*, *E*, *M* and *N*) and 8 accessory genes[9]. The functions of some of the genes have been partially deduced, and their roles in the spread of COVID-19 are now being studied. It had already been clear that the first SARS-classic virus (of 2003 pandemic) and SARS-CoV-2 (of COVID-19) were related, but not too closely. So, a similarity of 79% meant 21% difference and it was impossible that this much difference would be there if SARS-CoV-2 had directly originated from SARS-classic. There had to be other viruses.

Genome analysis and the first indicators

Phylogenetic tree construction with several already known coronavirus genomes confirmed that SARS-CoV-2 genome belonged to the betacoronavirus group (like SARS-classic, MERS and the several SARS-like bat CoVs). But, it was more closely related to the SARS-like CoVs from bats compared to the two human pathogens. Also, among the bat viruses, SARS-CoV-2 showed maximum identity with a virus that had been sampled from *Rhinolophus affinis* bats in the Chinese province of Yunnan in 2013. Its name was CoV-RaTG13 (Figure 1). At the whole genome level, the sequence identity between the two viruses is 96.2% (refs 5, 10), which certainly indicated close relatedness; this is a good evidence that the causative agent of COVID-19 has evolved from a bat virus closely related to CoV-RaTG13. But, it was necessary to see whether this strong similarity was uniformly seen across all the genes of RaTG13 and SARS-CoV-2 or not. Of particular interest was the spike (*S*) gene that codes for spike glycoprotein which projects out from the envelope of the virion.

The spike protein is essential for virus entry into the host cell. Detailed structural studies on the SARS-classic virus over the last 15 years have shown that its spike protein contains a receptor-binding domain (RBD) that interacts with the ACE2 (angiotensin-converting enzyme 2) membrane protein of human cells, initiating viral fusion and entry. More specifically, RBD consists of a core structure from which a smaller receptor-binding motif (RBM) projects out[4]. It is the amino acid residues of the RBM that bind to specific 'partner' residues on the ACE2 receptor. To give the oft-used analogy of biochemistry, spike protein is the molecular key and ACE2 receptor is the molecular lock. All SARS viruses carry the spike protein, but subtle variations (mutations) in the nucleotide sequence have resulted in spikes where one or more critical amino acid residues could be different. This is important because experiments have also established that while

ACE2 membrane protein is present in all mammalian hosts, which SARS virus will successfully infect which host depends mainly on the affinity between the RBD and RBM of a viral spike protein and the host cell's ACE2. For example, the SARS-classic virus has six critical residues that bind to human ACE2 (refs 4, 5, 10). Thus, the question was whether the spike genes of RaTG13, SARS-classic virus and SARS-CoV-2 identical or dissimilar?

The analysis showed that the spike protein of SARS-CoV-2 was rather different (i.e. divergent) compared to other betacoronaviruses; sequence identity was less than 75%. Even for the closely related RaTG13, sequence identity of *S* gene dropped significantly to 93.1% (Figure 2). This indicated that RaTG13 would not infect human cells and this prediction has been recently demonstrated experimentally[5,14,15]. And, when compared to the spike protein of the SARS-classic virus, only 76% of the amino acid residues of the two spike proteins were identical. The two sequences diverged further in the ACE2-binding region of *S* gene – only 50% for the RBM! Six residues of SARS-classic's spike were critical for binding to human ACE2 and five of them did not match with SARS-CoV-2 (refs 4, 14, 15).

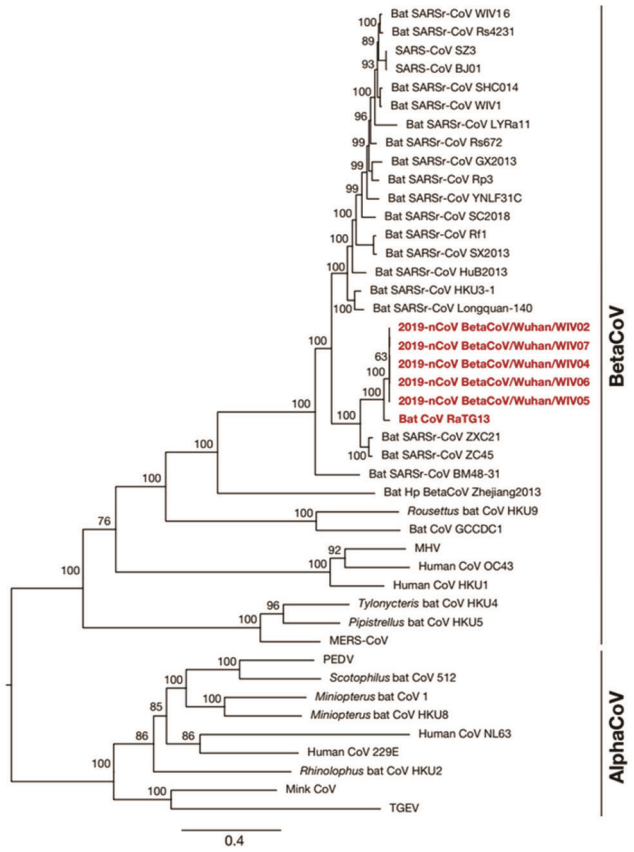

Figure 1. Phylogenetic tree based on complete genome sequences of several alphacoronaviruses (AlphaCoV) and betacoronaviruses (BetaCoV) shows the close relationship between SARS-CoV-2 identified in Wuhan and the RaTG13 virus isolated from *Rhinolophus affinis* bats. (Slightly adapted from the open-access article, Zhou *et al.*[5].)

Despite this difference, SARS-CoV-2 used ACE2 protein to enter host cells. Experiments showed that when ACE2 protein from humans, horseshoe bats, civets, mice and pigs was expressed on the membrane of human HeLa cells and then these cells were incubated with SARS-CoV-2 virions, the virus could enter all these cells except the ones that expressed the murine homolog of ACE2 (ref. 5). Notably, HeLa cells that did not express any ACE2 protein were not infected by the virus. Other experiments also confirmed that the spike protein of the new virus could effectively use human ACE2 (ref. 16). Subsequent experiments demonstrated that it binds to human ACE2 with higher affinity (Kd of SARS-classic RBM for human ACE2 is 31 nM and Kd of SARS-CoV-2 RBM for human ACE2 is 4.7 nM)[14]. But its key residues were different from SARS-classic, other known SARS-CoVs from bats or even RaTG13. Where had SARS-CoV-2 acquired this *S* gene come from?

The uniqueness of the new spike protein did not end here. Like other spikes, this too is made of two subunits, S1 and S2. But, the sequence analysis also showed that it contained a decisive sequence of four amino acids (RRAR – three arginines and one alanine) at the S1–S2 junction. This was recognized as the site which the protease furin (present in many human tissues) could cleave[10,14]. This cleavage is important for activating the spike protein. Cleaving a specific peptide bond removes a part from these large precursors and this, in turn, results in the formation of a smaller but active protein that can carry out catalysis. The identification of a distinct furin-recognition sequence (called the polybasic cleavage site)

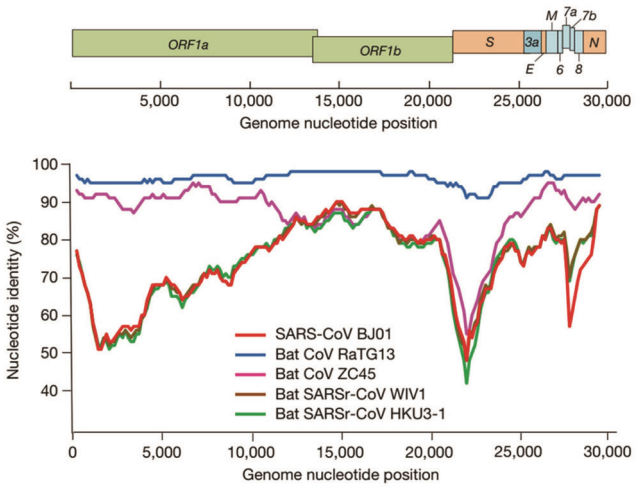

Figure 2. (Top) Genome organization of SARS-CoV-2 shows the positions of various genes. The *S* gene encoding the spike protein occurs around 22,000–25,000 nucleotides. (Bottom) Similarity plot based on the genome sequence of SARS-CoV-2. The *X*-axis shows the nucleotide sequences while *Y*-axis shows per cent identity when SARS-CoV-2 genome sequence is compared to the genomes of SARS-CoV and four coronaviruses from bats, including virus RaTG13. RaTG13's close similarity is shown by the flat blue line. (Slightly adapted from the open-access article, Zhou *et al.*[5].)

in SARS-CoV-2's spike protein immediately alarmed virologists. Polybasic cleavage sites had earlier been observed in influenza viruses – the hemagglutinin proteins of virulent strains have such sites, while the less-virulent strains do not carry them. In deadly flu viruses, the presence of such sites on hemagglutinin proteins facilitates swift activation (because furin is abundantly present in tissues of the respiratory tract). And once activated, the hemagglutinin protein docks onto receptor proteins on host cells – the first step of infection[10]. Furin would do a similar swift activation for SARS-CoV-2's spike protein and it was clear that this was one of the reasons why the novel virus was so infectious. But such a polybasic cleavage site had never been seen in SARS-classic or in any bat coronavirus related to SARS-CoV-2. The major question was from where had the new virus acquired this polybasic sequence that was contributing to its pathogenicity?

An 'unfortunate' animal provides evidence

If a new sequence is not already there in the databases, it must be out there in nature! It is true that bats are the reservoirs for several pathogens, but the last two pandemic-causing coronaviruses had not sprung directly from bats to man; civets and dromedary camels were the intermediate species for SARS-classic and MERS viruses respectively. Not only that, the differences even with the closest-known virus RaTG13 as well as the distinct differences in the RBD of the spike protein indicated SARS-CoV-2 had evolved, at least partly, in another species. Which one? One easy way to find out would have been to go back to the wet market in Wuhan and examine the animals being sold there. But, the market had been shut down and sanitized soon after the first cases emerged. Scientists would have to look elsewhere.

Since betacoronaviruses infect several mammals, it would be logical to look into animal specimens – tissues of animals from forests of China and East Asia. But which mammal? Several were being sold at the Wuhan market. Or it could be another source, maybe an animal farm. Logical guesswork and actual search gave scientists the (probable) answer rather fast – it was the pangolin. Easily recognized by their large scales, pangolins have the unfortunate tag of being the most poached mammals both due to their meat as well the use of their keratinized scales in Chinese traditional medicine[17,18]. As a result, several pangolin species have now been listed as critically endangered by the International Union for Conservation of Nature (IUCN), and customs officials and forest rangers across East Asia regularly retrieve pangolins from poachers and smugglers. It was from these pangolins that virologists had their break.

In fact, the first report of coronaviruses infecting pangolins was reported in October 2019 itself, two months before COVID-19 hit the world[16]. In March 2019, the Wildlife Rescue Center of the Chinese province of Guangdong had rescued 21 sick Malayan pangolins (*Manis javanica*). Sixteen of these animals died soon after. Autopsy showed their swollen lungs contained a frothy liquid and virologists identified SARS-like coronaviruses as one of the likely pathogens. Then, in March–April 2020, three research articles got published[18–20]. One of them presented a detailed analysis of the findings of October 2019. The other two papers identified viral genomes found in the tissues of more pangolins recovered in the Chinese provinces of Guangxi and Guangdong. The titles of the three papers summed up their results – 'Probable pangolin origin of SARS-CoV-2 associated with the COVID-19 outbreak', 'Identifying SARS-CoV-2 related coronaviruses in Malayan pangolins' and 'Isolation of SARS-CoV-2 related coronavirus from Malayan pangolins'.

In one study[18], 17 out of 25 Malayan pangolins retrieved from poachers during March–August 2019 showed presence of such coronaviruses in the lungs. Notably, the pangolins fell sick, showed respiratory distress, alveolar damage and died (usually, a virus does not cause disease in its natural reservoir host species (bat), but the intermediate hosts might show symptoms of infection). Initial BLAST search for SARS-like coronaviruses in mammalian and avian database turned up positive results from viral metagenomic sequences identified in pangolins[18]. The lung of an infected pangolin was homogenized and the supernatant was added to Vero cells in tissue culture. Within 72 h, the cells showed cytopathogenic effects typical of viral infection. Electron microscopy also identified the proliferation of coronaviruses. The 29.8 kb viral genome (called Pangolin-CoV) showed >90% sequence similarity to both RaTG13 and SARS-CoV-2. Moreover, at the protein level, the *S*, *E*, *M* and *N* genes of SARS-CoV-2 and the Pangolin-CoV showed 90.7%, 100%, 98.6% and 97.8% amino acid identity respectively. But the most remarkable observation was that the RBDs of spike proteins from Pangolin-CoV and SARS-CoV-2 differed by only one amino acid, i.e. they were practically identical[18]. Furthermore, phylogenetic analysis showed that for most of the length of the *S* gene, the SARS-CoV-2 and RaTG13 genes were closely related, while the *S* gene from Pangolin-CoV was a more distant relative. But when it came to the RBD sequence, it was the Pangolin sequence that closely matched with SARS-CoV-2; RaTG13 sequence became distant (Figure 3). This was a good indication that SARS-CoV-2 had originated by the recombination of a Pangolin-CoV-like virus with a RaTG13-like virus. The main genetic backbone of SARS-CoV-2 had probably come from the bat virus, but a pangolin virus had 'donated' the RBD to it. Natural selection had done the rest.

The second study[19] also found a virus in the dead pangolins, and this genome was undeniably related to

both SARS-CoV-2 as well as bat viruses RaTG13, ZXC21 and ZC45. But again, the clincher was the strong conservation between the RBDs – the amino acids which were known to bind to human ACE2 were all identical, only one non-essential residue differed. The third group's findings[20] were from pangolins recovered by the Guangxi Customs, and these viruses shared 85.5–92.4% sequence similarity to SARS-CoV-2. The convergence of evidence indicated that all these could not be due to chance; the parsimonious explanation was recombination between various betacoronaviruses from pangolins and bats. It fitted in with what was already known, i.e. coronaviruses indulged in extensive recombination, with even small genomic subregions having originated in different ancestral viruses[15,16]. It is this constant 'cut and paste' that provides the fodder for natural selection.

Most importantly, scientists identified recombination 'signals' in the viral genome sequence[15]. By using bioinformatics tools that detected recombination signals embedded within genome sequences they observed that, while SARS-CoV-2 did show highest pan-genomic similarity with RaTG13, there were two 'recombination breakspots' – (i) At beginning of *ORF1a* gene and (ii) before and after the DNA that codes for the RBM of the spike protein. The latter was a zone which was distinctly similar to the viruses isolated from pangolins. In sync with the other results, the best explanation for this scenario was that a RaTG13-like bat coronavirus acquired a human ACE2-binding RBM from a pangolin coronavirus by recombination and thus gained the ability to infect human cells.

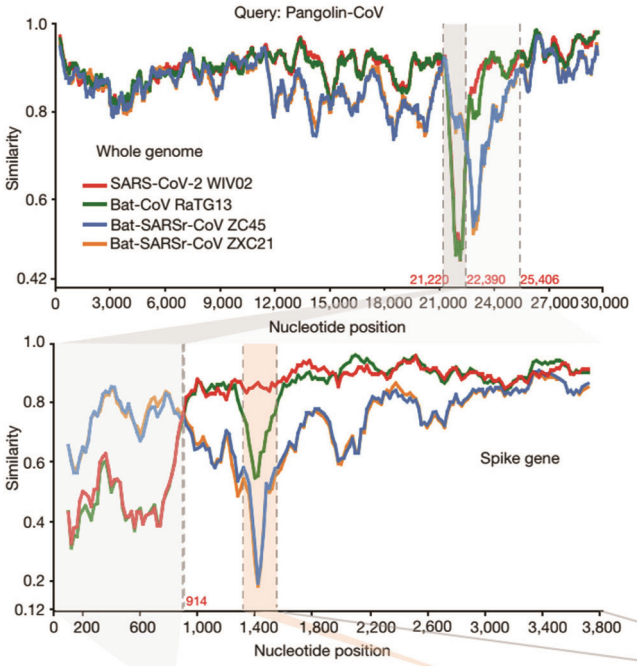

Figure 3. Similarity plot based on the genome sequence of coronavirus isolated from pangolins. The *X*-axis shows the nucleotide sequences while *Y*-axis shows per cent identity when pangolin-CoV genome is compared to the genomes of SARS-CoV-2 and three coronaviruses from bats, including virus RaTG13. (Top) The results show that the pangolin-CoV is highly similar to SARS-CoV-2 and RaTG13. (Bottom) Zooming on the *S* gene sequences also shows that SARS-CoV-2 (red line) is almost identical to the pangolin-CoV around the sequence coding for the RBD while the other genomes differ significantly. (Reproduced with author's permission from Xiao *et al.*[18].)

Yet-to-be-answered questions

How viruses from bats and pangolins got together, of course, remains unknown. But, given that they are both nocturnal, eat insects and share the same ecological spaces in East Asia[18], maybe half-eaten 'meals' and faeces of bats got mixed with pangolin food sources. On the other hand, wet markets (and the networks/farms that supply them) are places where animals of several species are closely packed together, potentially increasing the chances of viruses spilling over to new hosts. After all, viruses are using hosts to evolve and spread. However, there are gaps in this scenario – pangolins are solitary animals and viral evolution would be better in species with dense populations. So, were other animals involved as intermediate or amplifying hosts? A leading coronavirus expert speculates that raccoons, thousands of which are reared for the Chinese fur industry, are worth checking[21]. This scenario gets more complicated because there is some doubt whether the earliest cases came out of the Wuhan wet market or not[6].

Typical of scientific quest, there are other unanswered questions. No bat or pangolin virus closely related to SARS-CoV-2 has so far been shown to contain the functional furin site. So, where was this critical component acquired from? Scientists are still on the lookout for that, but a strong lead emerged in June 2020. An article titled 'A novel bat coronavirus closely related to SARS-CoV-2 contains natural insertions at the S1/S2 cleavage site of the spike protein' presented the viruses found from 227 bats from the Chinese Province of Yunnan[22]. Among them was the coronavirus RmYN02, a close genomic relative of SARS-CoV-2. The two viruses showed 93.3% nucleotide sequence identity at whole-genome level, but this identity dipped to 71.8% identity in the *S* gene. Notably, like RaTG13, only one of the six amino acid residues that helped SARS-CoV-2's RBM to hook onto human ACE2 matched. However, three amino acids – proline–alanine–alanine – were present at the *S*1–*S*2 junction. Although different from the furin site, this was a clear demonstration that such insertions were present in natural coronaviruses and recombination can transfer them from one virus to another. If many more viral genomes are searched, it is quite likely that the progenitor who 'gifted' that sequence to the pandemic-pathogen will be found.

Moreover, it is unknown whether the virus gained its complete ability to infect humans while in another

species? Or did it jump over to humans rather inefficiently and then fine-tuned its proteins by selection while spreading from one person to another?[10] Were Pangolins involved or was there a different host species that lived in big herds? For example, a recent publication has shown that viruses closely related to SARS-CoV-2 have been circulating in horseshoe bats for many decades[23]. A conclusion of this article was that the RBD of SARS-CoV-2 came from an as-yet-unidentified bat virus and not from pangolins. Another major study published in September 2020 had probed deep into how the spike of SARS-CoV-2 interacted to various extents with ACE2 orthologs from fourteen mammalian species. Phylogenetic study, structural modelling and infection of 293T cells transiently expressing ACE2 orthologs, showed that the human and rhesus monkey receptors were most efficient for viral entry, while ACE2 from rabbit, pangolin and dog were also good targets for SARS-CoV-2 (ref. 24). Another article submitted in bioRxiv claimed that the 5′UTR of the SARS-CoV-2 genome had high similarity with the 5′UTR of the coronaviruses isolated from the Guangdong pangolins[25]. Hardly twelve months into the situation, it is no wonder that the research landscape is still young and very dynamic. Nevertheless, an unprecedented speed of research in this field will provide these answers sooner or later.

How can one be sure all these viral features are products of natural selection, and not some sinister laboratory working on biological weapons?

The huge scientific evidence presented above, in hardly six months of systematic peer-reviewed research from different groups is the biggest proof for this question. If evidence was scant, alternative scenarios would have to be investigated. But when the data builds up well, and in harmony with our accumulated knowledge of modern biology, there can be no place for semi-literate conspiracies. However, to further nullify 'cut n paste' theories, some points should be sufficient.

Some of the earlier studies of the viral genome have revealed no proof of human manipulation[2,10]. The viral sequences have been easily accessible since January 2020; scores of experts have scanned them, and yet no one has reported any signatures of genetic manipulation. Rather, one pre-accepted manuscript that claimed such manipulation had to be quickly retracted when other scientists pointed out glaring mistakes in their analysis[26].

Is there a possibility that there was a scientific cover-up? Most unlikely, as that would demand a gigantic clandestine international collaboration across universities and institutes and for no strong reason. Rather, if such a thing happened, sooner than later, it would be scientifically exposed and all those implicated in such 'data fudging' would be blacklisted by peers and journals for their entire careers. No sincere professional scientist would do this. Overall, the international scientific community is sure that SARS-CoV-2 is a natural virus and serious academic circles are not even talking about such conspiracy theories anymore. That is now exclusively the domain of politicians and social media 'experts'.

Among the big scientific reasons is the evidence in the viral genome itself. It is evident that its backbone comes from a virus closely related to the bat viruses RaTG13 and RmYN02. But none of these viruses are known to cause any disease. If SARS-CoV-2 was really a genetically engineered bioweapon, why would unknown harmless viruses be used for making its main structure? If weapon making is the objective, then most likely a well-known (or less-known) pathogenic virus would have been used[2,10]. To give an analogy, if you had to make a sword, would you use steel or cardboard? The question is whether it is possible at all to make a deadly bioweapon using a non-pathogenic virus. In theory, the answer is an easy 'yes', but in reality there is no such guarantee that such 'reverse genetics' would be successful even after years after years efforts. Would any clandestine bioweapon project waste time and resources on such shaky things? The answer is an obvious no.

The protein that was not predicted

However, the finest evidence that SARS-CoV-2 is a natural virus comes from its biggest weapon – the RBM of the spike protein. As presented earlier, six residues in its RBM are critical for binding to human ACE2 receptor. This is the same with the RBM of the spike protein of SARS-classic virus, although five of the six amino acid residues are different. Unlike SARS-CoV-2, the molecular structure of RBDs of SARS-classic and other SARS coronaviruses isolated from different species have been extensively studied, both *in silico* and by experiments, for more than a decade[4]. The amino acid residues that determine their binding with ACE2 orthologs from corresponding host species have been investigated. It is now known that five residues at positions 442, 472, 479, 480 and 487 of the spike protein majorly determine its affinity for human ACE2. These residues are Tyr442, Leu472, Asn479, Asp480 and Thr487 for SARS-classic, but they vary between the various SARS viruses. When a RBM containing the five residues that bound optimally to human ACE2 was designed in the laboratory[4,27], it bound with super-affinity and the corresponding spike protein was super-efficient in mediating viral entry into host cells. It was expected that if SARS-CoV-2 was genetically engineered for high infectivity, its residues would be identical to the optimized high-affinity RBM. The next obvious question was – Did these residues in SARS-CoV-2 match with the optimized/designed RBM? The answer is 'No'.

Table 1. Comparing the critical residues in the receptor binding motifs (RBM) of SARS-Classic virus, the new SARS-CoV-2 and the optimized RBM designed in lab[27]

Key residues	SARS-classic RBM	SARS-CoV-2 RBM	Optimized/ designed RBM	Comparison with SARS-CoV-2 residue
1	Tyr442	Leu455	Phe442	Better, but not optimal
2	Leu472	Phe486	Phe472	Optimal
3	Asn479	Gln493	Asn479	Not optimal
4	Asp480	Ser494	Asp480	Not optimal
5	Thr487	Asn501	Thr487	Not optimal
6 (outside RBM)	Val404	Lys417	x	Better
7 (outside RBM)	Pro462	Ala475	x	Better

Table 1 clearly shows that of the five critical RBM residues of SARS-CoV-2, only one of them matches with the optimized/designed RBM and thus, the RBM of SARS-CoV-2 has a suboptimal binding efficiency. It is impossible to believe that a covert bioweapons project would deliberately design a partly-efficient weapon. Rather, the parsimonious scientific explanation was that the recombination and natural selection generated a RBM in SARS-CoV-2 that was different from both the RBM of SARS-classic and the lab-engineered RBM. Moreover, there were two residues outside RBM which provided additional grip to human ACE2, and this was something the optimized/designed RBM had missed[14]. Thus, it is clear that SARS-CoV-2's RBM is not a man-made construct. Rather, the spike proteins of SARS-classic, SARS-CoV-2 and optimized/designed RBM are three different solutions to the same problem, i.e. binding to human ACE2. This is an example of convergent evolution.

The spike protein of SARS-CoV-2 provides further evidence about its natural origin. It is a glycoprotein and has carbohydrate molecules covalently attached to certain residues. Which residues? Immediately preceding the furin site (RRAR) is a proline residue (Pro681). Pro681, which is present only in SARS-CoV-2 and RmYN02, facilitates the attachment of oligosaccharides to serine and threonine residues nearby (O-linked glycans). The function of these moieties is not yet clear, but they probably serve as 'mucin shields' that camouflage and protect the protein from attacks by the host's immune system[10]. An easiest explanation for the development of this mucin shield was that the virus (or its progenitor) was resident in a mammal with a functional immune system. It was in that host that such viral variants arose. They could fool the host's immune system and hence they got selected. Is it not possible to do this in the lab? No, it is practically impossible to generate this in cell cultures in the absence of an active immune system[10]. Thus, the presence of the mucin shield in the spike protein of SARS-CoV-2 is another good evidence that the pathogen is NOT a lab product. It is a product of biological evolution.

Supplementary lines of evidence including the Ka/Ks ratio[10,15], also prove that this virus is, like all other known pathogens, a natural virus. Moreover, the reputation of the Wuhan Institute of Virology, one of the most prestigious institutions where researchers from several countries work, argues against such conspiracy theories.

Concluding remarks

Do humans not have any role at all in this pandemic? Of course, we have. Although SARS-CoV-2 is quite unlikely to be creation of an evil human mind, it is undeniable that deforestation and illegal wildlife trade[18–20] – both examples of unbridled mega-scale human profiteering – are prime drivers for this viral spillover. The warnings and advice of scientists who have investigated the origins of SARS-CoV-2: 'the simplest and most cost-effective way to reduce the risk of future outbreaks is to limit our exposure to animal pathogens as much as possible[6]' and 'international co-operation and stricter regulations against illegal wildlife trade and consumption of game meat should be implemented. They can offer stronger protection of endangered animals as well as the prevention of major outbreaks caused by SARSr-CoV' – must be taken with utmost seriousness[18]. Along these related lines, a recent review speculated whether 'the sampled pangolins could have also been exposed to CoVs by other animal species or humans along the wildlife trade route[28]'. It is undeniable that COVID-19 and its aftermath is the product of human misadventures. It is also a time for collective introspection. We have to learn our lesson because the next 'reminder' is likely to be a more dangerous one.

Update: During the time this manuscript was being reviewed, the WHO sent an international team of experts to Wuhan in January 2021. Amidst political mistrust and allegations, they collaborated with Chinese experts, visited several sites associated with the pandemic and scanned through records. The detailed report of this 'first phase' of investigation will soon be published. However, the team have already summarized their main conclusions at a formal press conference. The lab-leak hypothesis is 'extremely unlikely'. There is a possibility that refrigerated/frozen animal food carried the virus into Wuhan. But, the WHO mission's leader Peter Ben Embarek has elaborated on the most likely scenario[29] – 'Some traders

at the Huanan market were trading in farmed wild animals like badgers, bamboo rats, rabbits, crocodiles and many others. Several of these animals are known to be susceptible to SARS viruses. Some of them come from farms in provinces where coronaviruses have been isolated from bats Guangdong Guanxi, Yunnan. Potentially, some of these animals were infected at those farms and then brought the virus into the market.' Thus, the WHO team's findings are in sync with the data presented in this review. Additionally, more viruses closely related to SARS-CoV-2 have been discovered in bats across South Asia. However, the intermediate species' are yet to be conclusively identified and animals like minks, raccoon dogs and foxes, all reared in many Chinese farms for their fur, will be studied in greater detail.

It is not surprising that more studies have unearthed facets of this subject in the last few months. The WHO team's final report has received criticism as well as defence. More studies have discovered coronavirus-harbouring bat populations in parts of southern and eastern Asia. The effect of forest fragmentation, concentrated livestock production, increased human encroachment into wildlife habitat has been quantified. Viral genome sequences, first submitted then deleted from databases, have been recovered. The origin of SARS-CoV-2 is indeed a field whose dynamics has few parallels.

1. Belay, E. D. *et al*., Zoonotic disease programs for enhancing global health security. *Emerg. Infect Dis*., 2017, **23**(Suppl. 1): S65–S70.
2. Stevens, C., 2020; https://leelabvirus.host/COVID19/origins-part1; https://leelabvirus.host/COVID19/origins-part2; https://leelabvirus.host/COVID19/origins-part3 (a lucid summary by the Benhur Lee Lab, Icahn School of Medicine at Mount Sinai, New York, USA).
3. Zimmer, C., *A Planet of Viruses*, The University of Chicago Press, USA, 2012, pp. 64–65.
4. Wan, Y. *et al*., Receptor recognition by the novel coronavirus from Wuhan: an analysis based on decade-long structural studies of SARS coronavirus. *J. Virol*., 2020, **94**, e00127-20.
5. Zhou, P. *et al*., A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature*, 2020, **579**, 270–273.
6. Zhang, Y. and Holmes, E., A genomic perspective on the origin and emergence of SARS-CoV-2. *Cell*, 2020, **181**, 1–5.
7. Coronaviridae Study Group of the International Committee on Taxonomy of Viruses. The species severe acute respiratory syndrome related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. *Nat. Microbiol*., 2020, **5**, 536–544.
8. Balaram, P., *Outlook*, 22 June 2020, pp. 92–93; https://magazine.outloookindia.com/story/india-news-why-does-science-come-last/303337
9. Wu, A. *et al*., Genome composition and divergence of the novel coronavirus (2019-nCoV) originating in China. *Cell Host Microbe*, 2020, **27**, 325–328.
10. Andersen, K. G. *et al*., The proximal origin of SARS-CoV-2. *Nat. Med.*, 2020, **26**, 450–452.
11. Guan, Y. *et al*., Isolation and characterization of viruses related to the SARS coronavirus from animals in southern China. *Science*, 2003, **302**, 276–278.
12. Drosten, C., Kellam, P. and Memish, Z. A., Evidence for camel-to-human transmission of MERS coronavirus. *New Engl. J. Med*., 2014, **371**, 1359–1360.
13. Cheng, V. C. C. *et al*., Severe acute respiratory syndrome coronavirus as an agent of emerging and reemerging infection. *Clin. Microb. Rev.*, 2007, **20**, 660–694.
14. Lan, J. *et al*., Structure of the SARS-CoV-2 spike receptor binding domain bound to the ACE2 receptor. *Nature*, 2020, **581**, 215–220.
15. Li, X. *et al*., Emergence of SARS-CoV-2 through recombination and strong purifying selection. *Sci. Adv.*, 2020, **6**, eabb9153.
16. Letko, M. *et al*., Functional assessment of cell entry and receptor usage for SARS-CoV-2 and other lineage B betacoronaviruses. *Nat. Microbiol*., 2020, **5**, 562–569.
17. Liu, P., Chen, W. and Chen, J.-P., Viral metagenomics revealed sendai virus and coronavirus infection of Malayan Pangolins (*Manis javanica*). *Viruses*, 2019, **11**, 979.
18. Xiao, K. *et al*., Isolation of SARS-CoV-2-related coronavirus from Malayan pangolins. *Nature*, 2020, **583**, 286–289.
19. Zhang, T. *et al*., Probable pangolin origin of SARS-CoV-2 associated with the COVID-19 outbreak. *Curr. Biol*., 2020, **30**, 1–6.
20. Lam, T. *et al*., Identifying SARS-CoV-2-related coronaviruses in Malayan pangolins. *Nature*, 2020, **583**, 282–285.
21. Kupferschmidt, K., How the pandemic made this virologist an unlikely cult figure. *Science*, 2020; https://www.sciencemag.org/news/2020/04/how-pandemic-made-virologist-unlikely-cult-figure
22. Zhou, H. *et al*., A novel bat coronavirus closely related to SARS-CoV-2 contains natural insertions at the S1/S2 cleavage site of the spike protein. *Curr. Biol*., 2020, **30**, 2196–2203.
23. Boni, M. F. *et al*., Evolutionary origins of the SARS-CoV-2 sarbecovirus lineage responsible for the COVID-19 pandemic. *Nat. Microbiol*., 2020, **5**, 1408–1417.
24. Zhao, X. *et al*., Broad and differential animal angiotensin-converting enzyme 2 receptor usage by SARS-CoV-2. *J. Virol*., 2020, **94**, e00940-20.
25. Afrasiabi, A., Evidence for ZAP-independent CpG reduction in SARS-CoV-2 genome, and pangolin coronavirus origin of 5′UTR. bioRxiv, 2020; https://www.biorxiv.org/content/10.1101/2020.10.23.351353v1.
26. Pradhan, P. *et al*., Uncanny similarity of unique inserts in the 2019-nCoV spike protein to HIV-1 gp120 and Gag. bioRxiv (withdrawn), 2020.
27. Wu, K. *et al*., Mechanisms of host receptor adaptation by severe acute respiratory syndrome coronavirus. *J. Biol. Chem.*, 2012, 2012, **287**, 8904–8911.
28. Banerjee, A. *et al*., Unraveling the zoonotic origin and transmission of SARS-CoV-2. *Trends Eco. Evol*., 2020; https://doi.org/10.1016/j.tree.2020.12.002.
29. Kupferschmidt, K., 'Politics was always in the room'. WHO mission chief reflects on China trip seeking COVID-19's origin. *Science*, 2021; https://www.sciencemag.org/news/2021/02/politics-was-always-room-who-mission-chief-reflects-china-trip-seeking-covid-19-s

Received 29 October 2020; revised accepted 9 April 2021

doi: 10.18520/cs/v121/i1/77-84