Tracing the origins of SARS-COV-2 in coronavirus phylogenies: a review

Erwan Sallard, José Halloy, Didier Casane, Etienne Decroly, Jacques van Helden

Environmental Chemistry Letters (2021) 19:769–785
https://doi.org/10.1007/s10311-020-01151-1

REVIEW

Tracing the origins of SARS-COV-2 in coronavirus phylogenies: a review

Erwan Sallard[1] · José Halloy[2] · Didier Casane[3,4] · Etienne Decroly[5] · Jacques van Helden[6,7]

Received: 11 November 2020 / Accepted: 26 November 2020 / Published online: 4 February 2021

Abstract

SARS-CoV-2 is a new human coronavirus (CoV), which emerged in China in late 2019 and is responsible for the global COVID-19 pandemic that caused more than 97 million infections and 2 million deaths in 12 months. Understanding the origin of this virus is an important issue, and it is necessary to determine the mechanisms of viral dissemination in order to contain future epidemics. Based on phylogenetic inferences, sequence analysis and structure–function relationships of coronavirus proteins, informed by the knowledge currently available on the virus, we discuss the different scenarios on the origin—natural or synthetic—of the virus. The data currently available are not sufficient to firmly assert whether SARS-CoV2 results from a zoonotic emergence or from an accidental escape of a laboratory strain. This question needs to be solved because it has important consequences on the risk/benefit balance of our interactions with ecosystems, on intensive breeding of wild and domestic animals, on some laboratory practices and on scientific policy and biosafety regulations. Regardless of COVID-19 origin, studying the evolution of the molecular mechanisms involved in the emergence of pandemic viruses is essential to develop therapeutic and vaccine strategies and to prevent future zoonoses. This article is a translation and update of a French article published in Médecine/Sciences, August/September 2020 (https://doi.org/10.1051/medsci/2020123).

Keywords SARS-CoV-2 · Coronavirus · Covid-19 · Pandemic · Bioinformatics · Virology · Phylogeny · Genome analysis · Gain of function · Furin · Zoonosis · Biosafety · Spike protein

Etienne Decroly and Jacques van Helden contributed equally to the article.

✉ Jacques van Helden
Jacques.van-Helden@univ-amu.fr

Etienne Decroly
etienne.decroly@univ-amu.fr

1 École Normale Supérieure de Paris, 45 rue d'Ulm, 75005 Paris, France

2 Université de Paris, CNRS, LIED UMR 8236, 85 bd Saint-Germain, 75006 Paris, France

3 Université Paris-Saclay, CNRS, IRD, UMR Évolution, Génomes, Comportement et Écologie, 91198 Gif-sur-Yvette, France

4 Université de Paris, UFR Sciences du Vivant, 75013 Paris, France

5 Aix-Marseille Univ, CNRS, UMR 7257, AFMB, Case 925, 163 Avenue de Luminy, 13288 Marseille Cedex 09, France

6 CNRS, Institut Français de Bioinformatique, IFB-core, UMS 3601, Evry, France

7 Aix-Marseille Univ, INSERM, Lab. Theory and Approaches of Genome Complexity (TAGC), Marseille, France

Introduction

SARS-CoV-2 is the third human coronavirus (CoV) responsible for severe respiratory syndrome that emerged in the last 20 years, the two previous ones being SARS-CoV in 2002 (Drosten et al. 2003) and MERS-CoV in 2012 (Zaki et al. 2012). SARS-CoV-2, which causes the COVID-19 disease in humans, spread in early 2019 leding to a pandemie. By January 20, 2021, more than 97 million infections had been reported with at least 1.4 million deaths. The etiological agent of COVID-19 was rapidly identified at the beginning of the pandemic, and by January 26, 2020, 10 viral genomes had been sequenced (Lu et al. 2020). Sequence comparisons revealed a 99.98% pairwise identity between those genomes, which is characteristic of a recent emergence.

When the first SARS-CoV-2 isolates were sequenced, the closest coronaviruses available in databases were bat-SL-CoVZXC21 and bat-SL-CoVZC45 strains, isolated in 2015 and 2017 from bats in the Zhoushan region of eastern China, and whose genomes showed 88% identity with SARS-CoV-2 (Lu et al. 2020). The SARS-CoV-2 genome sequence is

more distant from SARS-CoV (79% identity) and MERS-CoV (50% identity), the viruses responsible for the previous human epidemics. Researchers concluded that SARS-CoV-2 is a new infectious agent belonging to the SARS-CoV family, able of human-to-human transmission, and whose animal reservoir is a bat (Zhou et al. 2020a; Lu et al. 2020).

Based on phylogenetic inferences, sequence analysis and structure–function relationships of coronavirus proteins, informed by the knowledge currently available on the SARS-CoV-2 virus, we present our re-analysis of the available data and discuss the different scenarios evoked to account for the origin of this coronavirus. Addressing this question is important not only to understand the causes of the pandemic, but also because the actual events at the origin of the virus should be taken into account for decision-making about science policy.

This article is the English translation and update of a French article published in médecines/sciences (https://doi.org/10.1051/medsci/2020123) on July 10, 2020. Since our study included a complete re-analysis by ourselves of the genomic and peptidic sequences, this English translation contains an additional section "Materials and methods." We also added the ferret in Fig. 4, made some minor revisions, added a short conclusion at the end of each paragraph and discussed a few key articles on the subject that were published after our initial publication.

Evolutionary origin of the new virus

The zoonotic origin of CoVs is well documented. This family of viruses infects more than 500 species of chiropterans (a mammalian order consisting of more than 1200 species of bats) which represent an important reservoir for CoV evolution, allowing the recombination of viral genomes in animals co-infected by different strains (Hu et al. 2017; Luk et al. 2019; Menachery et al. 2015). It is generally accepted that zoonotic transmission of CoVs to humans occurs through an intermediate host species, in which viruses better adapted to human receptors can be selected, thereby facilitating the species barrier crossing (Cui et al. 2019). Vectors of zoonotic transmission can be identified by examining the phylogenetic relationships between new viruses and viruses isolated from animal species living in the regions of CoV emergence.

Figure 1a, which presents the phylogenetic tree produced from full-genome alignments of different CoVs, shows the close proximity (99% genome identity) between the coronaviruses responsible for the two previous epidemics and the respective strains isolated from the last intermediate hosts before transmission to humans: civets for SARS-CoV in 2003 (Fig. 1b) (Guan et al. 2003; Song et al. 2005), and camels for MERS-CoV (Fig. 1c) (Sabir et al. 2016). In the latter case, several zoonotic transmission (from animal hosts to humans) have been demonstrated.

Although no epidemic related to direct bat-to-human transmission has been identified to date, experimental studies have shown that more than 60 chiropteran CoVs are capable of infecting cultured human cells (Luis et al. 2013; Menachery et al. 2015). The identification, in 2017, of viral isolates very similar to SARS-CoV in bats raises the issue of a possible direct transmission from chiropterans to humans, which could result from mutations in the receptor-binding domain (RBD) of the viral spike protein having enabled its entry into the host cell (Hu et al. 2017).

In summary, mechanisms of viral emergence and spreading have to be elucidated, and molecular phylogeny can contribute to provide clues about the possible paths of transmissions from bats to humans.

SARS-CoV-2: From Yunnan to Wuhan?

The origin of SARS-CoV-2 is a matter of debate. Bioinformatic studies revealed that it has a 96.2% identity with a CoV genome (RaTG13) reconstructed from feces and anal samples of *Rhinolophus affinis* bats. Interestingly, these samples were collected in 2013, but the full-genome sequence was only published in early February 2020 (Zhou et al. 2020a). Unfortunately, the precise location of the sample collection is documented neither in the original article nor in the sequence databases. However, when the current analysis was led (April 2020), we found an exact match between RaTG13 and a 370 nucleotide fragment published in 2016 (KP876546), encoding a BtCoV/4991 polymerase domain, which had been sequenced from isolates collected from a mine shaft in Yunnan Province following the death of 3 miners from an atypical pneumonia (Ge et al. 2016; Wu et al. 2014; Rissanen et al. 2017). We thus inferred that this strain resulted from this mineshaft. In the meantime, concerns have been raised by the scientific community about this lack of information concerning the source of the RaTG13 strain, and a recent addendum to the publication confirmed our deduction (Zhou et al. 2020b).

More recently, a metagenome (RmYN02) was assembled from feces samples of 11 bats of the species *Rhinolophus malayanus*, collected in 2019 in Yunnan province. This sequence has 97.2% identity with the first two-thirds of the SARS-CoV-2 (ORF 1ab) genome. However, on the remaining third of the genome it diverges quite strongly, especially at the level of the S1 protein and ORF 8 (Fig. 2) (Zhou et al. 2020c).

Thus, even though viral strain sequencing enabled the identification of several viruses related to SARS-CoV-2, the genetic distance is still too high to consider them as the proximal ancestors.

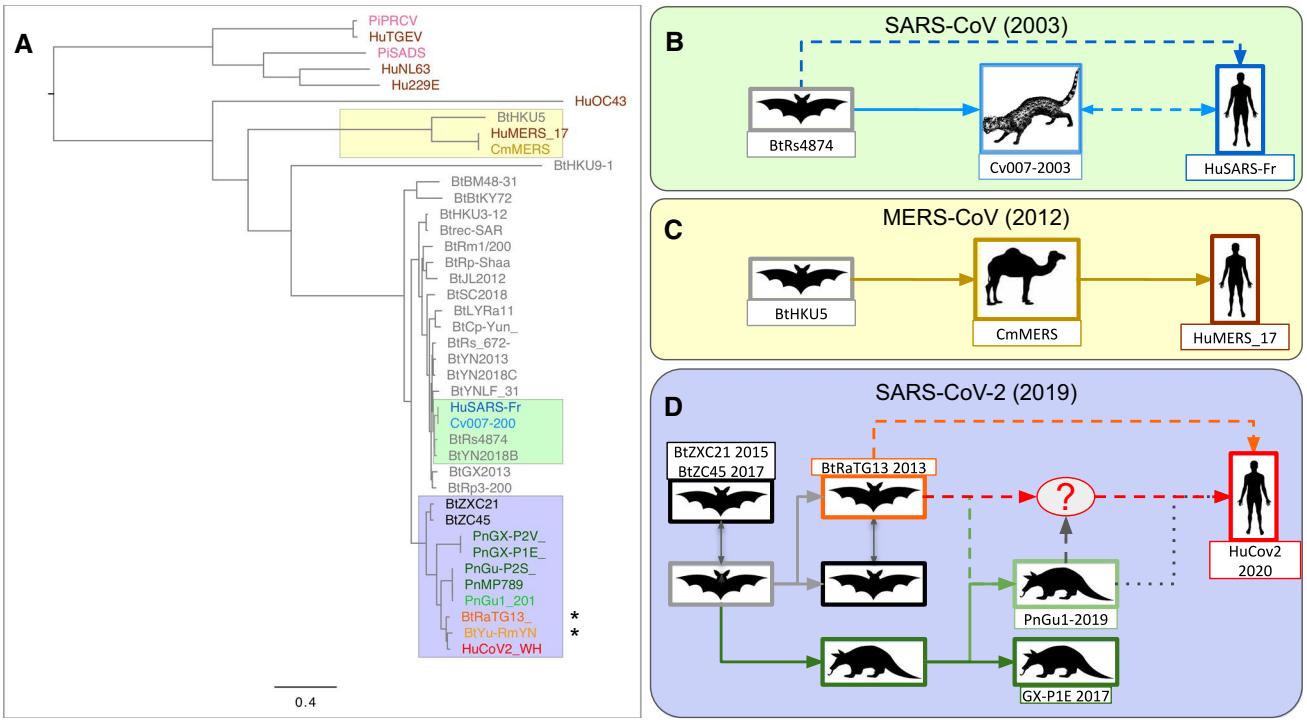


Fig. 1 Phylogeny and emergence of coronaviruses. **a** Tree inferred from complete coronavirus genomes, based on multiple alignment (clustalw) followed by maximum likelihood inference (PhyML). Genomes assembled from metagenomic data are marked with a star. The prefixes of virus names indicate the host species: Bt (bat), Hu (human), Pn (pangolin), Cv (civet), Cm (camel), Pi (pig). Note that the distances between HuCoV2 and the closest viral strains (BrY-uRmYN02, BtRaTG13) are higher than for SARS-CoV (human–civet) or MERS-CoV (human–camel). **b–d** Hypotheses of transmission from the animal reservoir (bats) to humans, based on the molecular phylogeny of viral genomes. **b** For the SARS-CoV pandemic of 2003, the civet has been proposed as intermediate host. Direct bat-to-human transmission is also under consideration. **c** Pandemic MERS-CoV of 2012, with the camel as an intermediate host. Several direct transmission events have been documented. **d** COVID-19 pandemic. Several scenarios are proposed about the last host before transmission to humans. Distances between HuCoV2 and the closest viral strains are found to be greater than for SARS-CoV (human–civet) or MERS-CoV (human–camel)

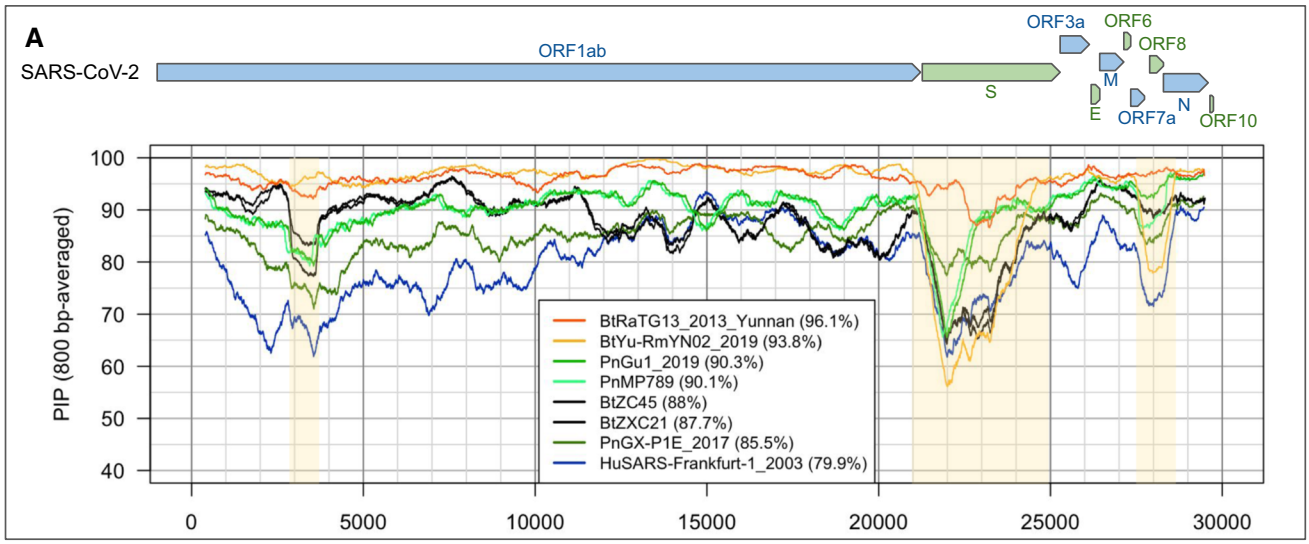
An evolutionary history by fragments

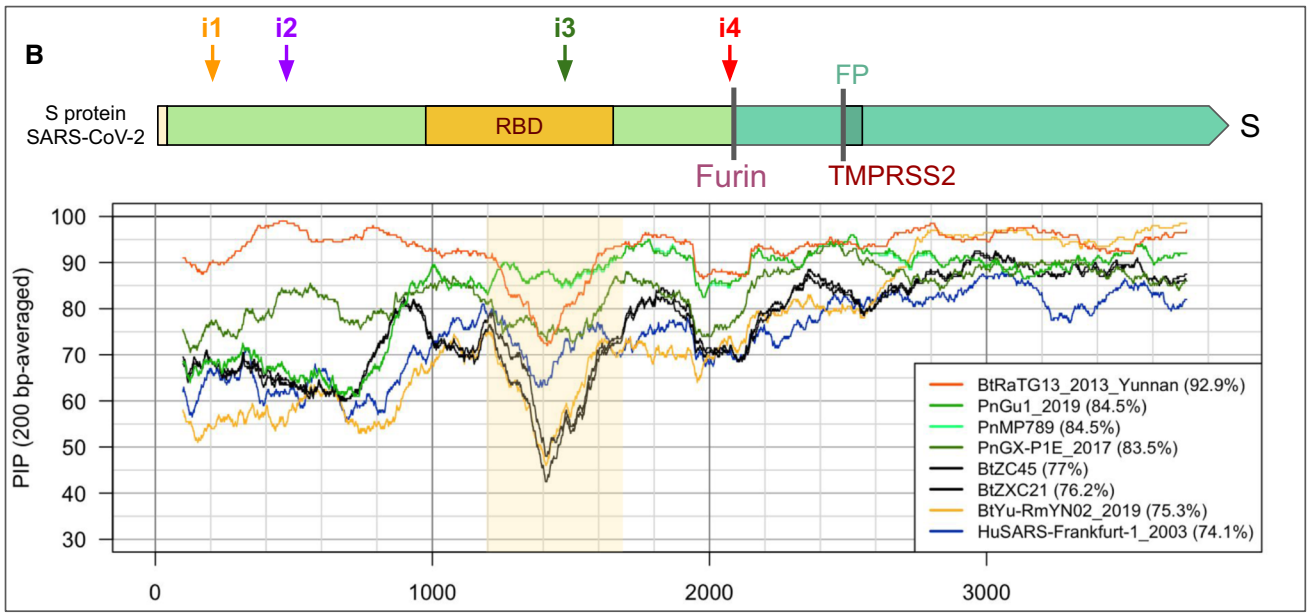
The length of CoV genomes is about 30,000 nucleotides, which is exceptionally long for an RNA virus. (By comparison, the length of AIDS and Ebola virus genomes is about 10,000 and 19,000 nucleotides, respectively) CoVs are able to maintain such a long genome thanks to a replication error correction system unique in the world of RNA viruses that ensures a proofreading mechanism limiting the mutation rate (Eckerle et al. 2010; Ferron et al. 2018; Casane et al. 2019). The first two-thirds of the genome corresponds to a single gene, ORF1ab, coding for a polyprotein precursor, which is then cleaved into 16 proteins forming the replication/transcription complex. The last third contains 9 genes coding for proteins produced from subgenomic RNAs synthesized by viral polymerase (Fig. 2a).

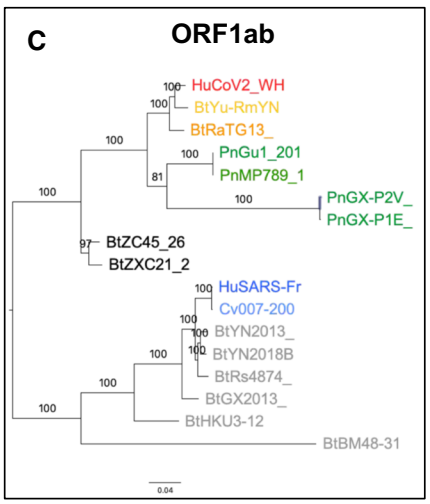
The CoV viral polymerase, in addition to its canonical RdRp activity, is able to jump between different RNA strands during replication (template switching), a property that probably plays a key role in the recombination capacity of CoVs, and promotes their evolution and host change. Genomic recombination is frequent in chiropteran CoVs (Hu et al. 2017) and is thought to have played a role in the emergence of SARS-CoV in 2002 (Graham and Baric 2010). It is believed that the SARS-CoV-2 genome is a "mosaic" genome composed of pieces of at least two preexisting CoVs.

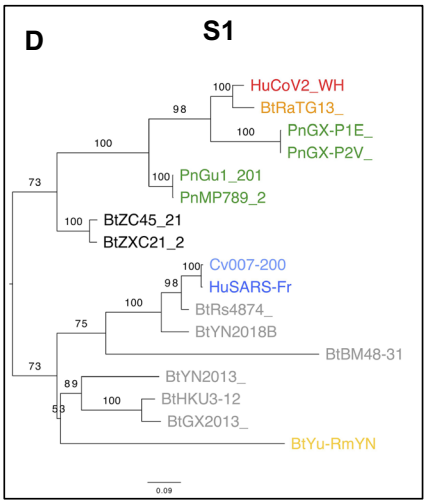
A recombinant genome can be detected on the Percent of Identical Positions (PIP) profiles, which are obtained by aligning different genomes to a reference genome. A recombination site is evidenced by the fact that the profiles of two different strains cross each other. In the genomic PIP profiles comparing SARS-CoV-2 with other genetically related viruses (Fig. 2a, b), recombinations appear in multiple regions, for example, 2,900–3,800, 21,000–24,000, 27,500–28,500 (highlighted with a yellow background).

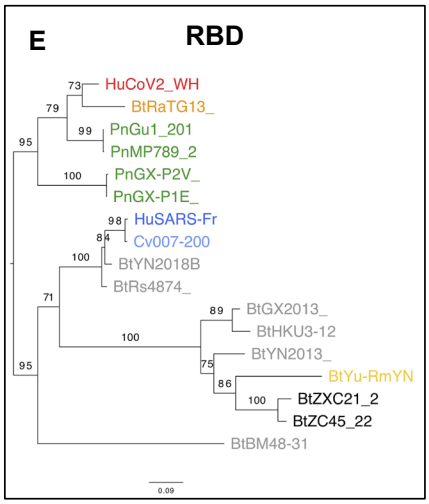
This mosaicism biases genome-based phylogenies because the inferred tree is a combination of the different evolutionary histories of the recombinant fragments. A phylogenetic


A
SARS-CoV-2
ORF1ab
S
ORF3a
ORF6
ORF8
M
E
ORF7a
N
ORF10
PIP (800 bp-averaged)
BtRaTG13_2013_Yunnan (96.1%)
BtYu-RmYN02_2019 (93.8%)
PnGu1_2019 (90.3%)
PnMP789 (90.1%)
BtZC45 (88%)
BtZXC21 (87.7%)
PnGX-P1E_2017 (85.5%)
HuSARS-Frankfurt-1_2003 (79.9%)


B
i1
i2
i3
i4
FP
S protein
SARS-CoV-2
RBD
S
Furin
TMPRSS2
PIP (200 bp-averaged)
BtRaTG13_2013_Yunnan (92.9%)
PnGu1_2019 (84.5%)
PnMP789 (84.5%)
PnGX-P1E_2017 (83.5%)
BtZC45 (77%)
BtZXC21 (76.2%)
BtYu-RmYN02_2019 (75.3%)
HuSARS-Frankfurt-1_2003 (74.1%)


C
ORF1ab
HuCoV2_WH
BtYu-RmYN
BtRaTG13_
PnGu1_201
PnMP789_1
PnGX-P2V_
PnGX-P1E_
BtZC45_26
BtZXC21_2
HuSARS-Fr
Cv007-200
BtYN2013_
BtYN2018B
BtRs4874_
BtGX2013_
BtHKU3-12
BtBM48-31


D
S1
HuCoV2_WH
BtRaTG13_
PnGX-P1E_
PnGX-P2V_
PnGu1_201
PnMP789_2
BtZC45_21
BtZXC21_2
Cv007-200
HuSARS-Fr
BtRs4874_
BtYN2018B
BtBM48-31
BtYN2013_
BtHKU3-12
BtGX2013_
BtYu-RmYN


E
RBD
HuCoV2_WH
BtRaTG13_
PnGu1_201
PnMP789_2
PnGX-P2V_
PnGX-P1E_
HuSARS-Fr
Cv007-200
BtYN2018B
BtRs4874_
BtGX2013_
BtHKU3-12
BtYN2013_
BtYu-RmYN
BtZXC21_2
BtZC45_22
BtBM48-31

◂**Fig. 2** Profiles of Percent Identical Positions (PIP) between SARS-CoV-2 and other coronavirus genomic sequences. **a** Genome-wide PIP profile (with sliding windows of 800 base pairs). **b** PIP profile along the S gene (200 bp sliding windows). **c–e** Impact of recombinations on the topology of phylogenetic trees inferred from different genomic regions: ORF1ab (**c**), S1 (**d**) and RBD (**e**)

inference should thus be made for each recombinant region separately, as illustrated in Fig. 2c–e, where we have inferred evolutionary trees, respectively, from the genomic sequences of the ORF1ab polyprotein (Fig. 2c), the S1 subunit (Fig. 2d) and the receptor-binding domain (RBD) (Fig. 2e). There are several striking differences between these trees, in particular concerning strain BtYu-RmYN02, which occupies different positions depending on the genomic region considered.

CoV genomes thus contain traces of recombinations, and the evolutionary history of coronaviruses should be interpreted in the light of this well-described mosaicism.

Of bats and men… plus some pangolins?

If we zoom in on the S gene (Fig. 2b), we can see a decrease in the PIP between the bat strain RaTG13 and SARS-CoV-2 in the RBD-coding genomic region. In particular, at positions 1200 to 1600 of this gene, the PIP falls to 70%, while it is higher than 96% over the rest of the genome. In the same region, the closest sequence to SARS-CoV-2 is that of a metagenome (MP789), obtained by assembling pangolin samples (Lam et al. 2020).

Upstream of the RBD, the PIP between SARS-CoV-2 and MP789 is fairly low (60%), whereas downstream it exceeds 90%. This led Xiao and co-workers (Xiao et al. 2020) to hypothesize that SARS-CoV-2 could result from recombinations between viruses infecting bats and pangolins, respectively (Fig. 1d). It should, however, be noted that, even in pangolins, there is currently no known non-human CoV whose PIP with SARS-CoV-2 exceeds 89% in the RBD region. This level of identity is much lower than the percentages observed between human viruses and strains of the last animal intermediates during previous zoonotic transmissions. For example, the identity rate between the human SARS-CoV genome and that of the closest civet strain is 99.52%.

The initial hypothesis was thus that SARS-CoV-2 results from multiple recombinations between different bat and pangolin CoVs, followed by an adaptation that would have increased its capacity for human-to-human transmission. Transmission to humans would come from contact with the intermediate host sold on the Wuhan market (Liu et al. 2020). However, this hypothesis raises many questions. On the one hand, the first identified patients did not attend the Wuhan market (Huang et al. 2020). On the other hand, despite the search for viruses in the animal species sold on this market (Zhang et al. 2020), to date no intermediate virus has been identified that may result from the supposed recombination between a bat virus and a pangolin virus. Moreover, the source of the pangolin samples is still unclear, and a warning has recently been posted on Nature website (on November 11, 2020) about Xiao's article (Xiao et al. 2020), indicating "additional actions will be taken once this matter is resolved."

Until the last hypothetical recombinant has been identified and its genome sequenced, it will not be known for certain in which species this recombination has taken place: a bat, a pangolin, another species? And above all, in which conditions? It is conceivable that the recombination took place in farm or laboratory animals rather than in wild pangolins or bats: In the former case, transmission to humans would be favored by closer and more frequent contact. Furthermore, the human ACE2 (angiotensin converting enzyme 2) protein, which is used by SARS-CoV-2 as a receptor for cell infection, is closer to the homologous protein of numerous farm animals than to the ACE2 proteins of pangolins and bats (Fig. 4). Another hypothesis is that the similarity between the RBD sequences of pangolin and SARS-CoV-2 results from a convergent evolution.

The hypothesis promoted by most specialists is that the virus has a zoonotic origin. This hypothesis relies on phylogenetic studies suggesting two main scenarios to explain the origin of SARS-CoV-2: (i) adaptation in an animal host before zoonotic transfer or (ii) adaptation in humans after zoonotic transfer (Latinne et al. 2020; Zhou et al. 2020a; Lam et al. 2020; Zhang et al. 2020; Xiao et al. 2020; Andersen et al. 2020). However, in the absence of evidence regarding the last animal intermediate before human contamination (the "proximal" origin of the virus), some authors suggested that SARS-CoV-2 may have been manufactured in a laboratory (synthetic origin) (Segreto and Deigin 2020; Relman 2020). Others suggested that SARS-CoV-2 may result from a chiropteran virus that became adapted to other species in laboratory animal models and then escaped from the laboratory (Sirotkin and Sirotkin 2020). It might also be envisaged that it comes from a viral strain cultured on human cells in a laboratory in order to study its infectious potential, and that has been progressively "humanized" (adapted to humans) by selection of the viruses having the highest ability to spread in these conditions.

Regardless of the mechanism of appearance of the virus, it is important to understand how it crossed the species barrier and became highly transmissible from human to human, in order to prevent new outbreaks (Cheng et al. 2007). In conclusion, there is little evidence that supports pangolins as the intermediate host in a zoonosis and, in the absence of such evidence, additional virus strains should be collected

from wild sites and from animal farms in order to elucidate the transmission path from bat to human.

Protein S is a major player in the evolution of CoVs and the crossing of the species barrier

The S gene codes for the spike protein, which is located in the viral envelope and forms characteristic crown-like structures on the viral surface, from which the name of coronavirus derives (Fig. 3a). The spike protein plays a decisive role in the initiation of the viral cycle because it participates in the recognition of the ACE2 receptors of the host cell, which then enables the delivery of the viral genome into the cells. This receptor, present in all mammal species, is located at the outer surface of different human cell types, including alveolar cells in the lung, enterocytes in the small intestine, arterial and venous endothelial cells and arterial smooth muscle cells in most organs. ACE2 messenger RNA is also detected in the cerebral cortex, striatum, hypothalamus and brainstem. Interferons, which are signaling proteins produced in response to viral infections, also increase ACE2 expression, thereby promoting the systemic spread of the virus (Ziegler et al. 2020).

The S protein is synthesized as an inactive precursor, which requires two successive proteolytic cleavages to ensure its biological function (Fig. 3b). The first cleavage, called "*priming*", generates the S1 and S2 subunits. The second cleavage occurs within S2 and releases the end of a *fusion peptide* located at the N-terminus of the S2' subunit. These two proteolytic cleavages are likely catalyzed, respectively, by furin and other proteases such as TMPRSS2 (transmembrane protease 2) (Hoffmann et al. 2020). The S protein cleavage is essential for the formation of infectious viral particles, as they favor ACE2 receptor recognition and allow fusion between the viral and cell membranes (Fig. 3a).

The S1 protein of SARS-CoV and SARS-CoV-2 contains the RBD domain (Figs. 3b, 4) which ensures the recognition of the ACE2 receptor by the virus (Wrapp et al. 2020; Wu et al. 2020; Lam et al. 2020). It also bears most of the exposed sites on the virus surface (Fig. 3c), including the

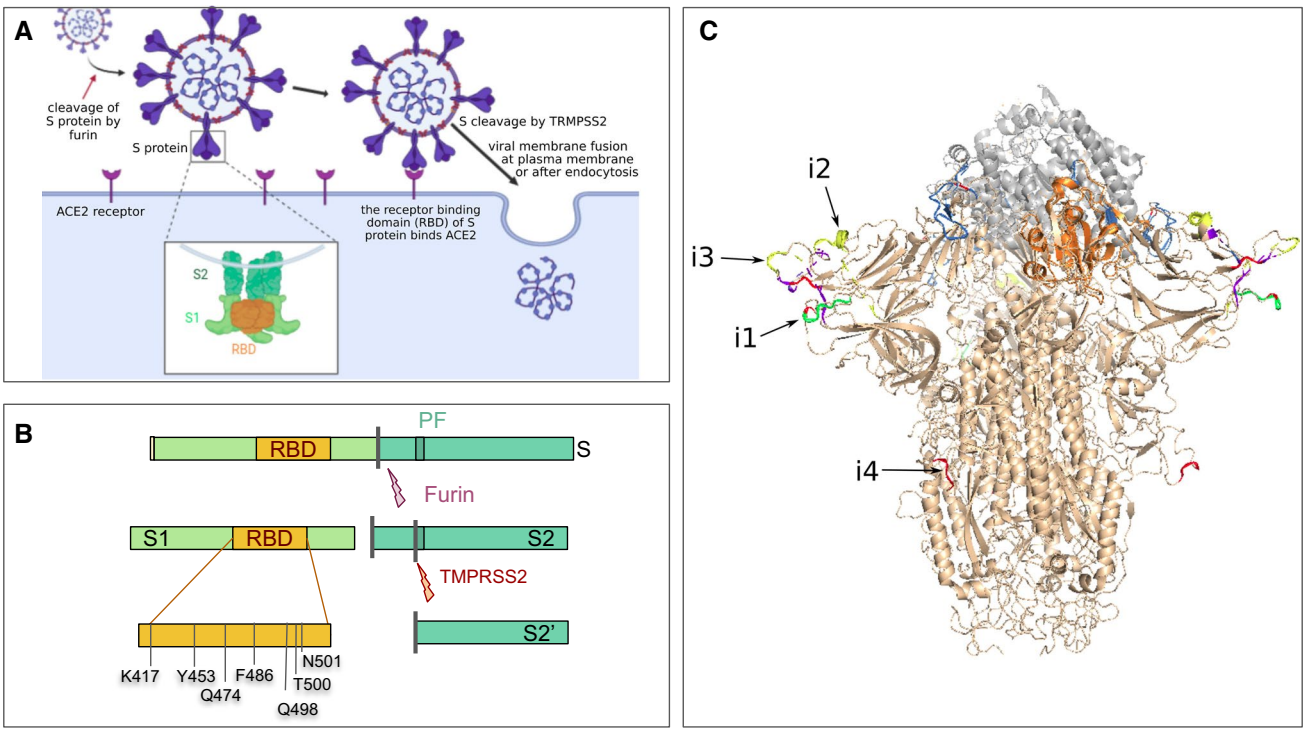


Fig. 3 Structure and function of the spike protein (S protein). **a** SARS-CoV-2 S protein specifically recognizes the ACE2 receptor of the host cells and thereby starts the infection cycle. **b** The S protein undergoes 2 maturation steps by proteolytic cleavage (respectively catalyzed by the furin and the TMPRSS2 proteins), which are required to activate the protein and to unlock the fusion peptide. **c** Structure of the viral S protein bound to the host ACE2 receptor. The SARS-CoV-2 S protein structure (beige) was produced by running SWISS-MODEL on the SARS-CoV homolog (Protein Data Bank entry 6acc) and aligned on the structure of an RBD domain (orange) interacting with ACE2 (gray) from the PDB model 6m0j. The SARS-CoV-2 insertions are highlighted in colors, with a coloring scale reflecting the taxonomic scope of the insertion: red (only found in human SARS-CoV-2, yellow, green, blue and purple (insertion found in most sarbecoviruses)

major antigens that can be recognized by antibodies produced by infected hosts (Ni et al. 2020). The sequence of these exposed sites shows a high variability between virus species, which results from the selection of mutations enabling viruses to escape the immune response.

The RBD residues directly involved in the recognition of ACE2 are also subject to strong evolutionary constraints (Fig. 4). Some key residues are required for efficient infection of chiropterans, the intermediate host or humans (Lu et al. 2015; Letko et al. 2020; Yan et al. 2020), and SARS-CoV-2 may have acquired its epidemic propensity through mutations in these key residues. Phylogenetic analysis of the S protein is therefore particularly informative to understand the evolution of CoVs and their ability to cross the species barrier.

In this context, the identification of new SARS-CoV-2-like coronavirus sequences isolated from Malayan pangolins was an important step forward. Indeed, although the whole genome PIP between these viruses and SARS-CoV-2 does not exceed 89% (Fig. 2a, strain MP789) vs. 96% for RaTG13, the amino acid identity is 98% in the RBD domain (Xiao et al. 2020). This difference of similarity between nucleic acids and proteins can be explained by the fact that almost all mutations in this region are synonymous, suggesting a strong selective pressure, presumably related to the key function of RBD in the infection. It therefore appears that some CoVs that infect pangolins possess an RBD domain very close to SARS-CoV-2, and may thus have a strong affinity for the human ACE2 receptor and thereby infect human cells more effectively than bat viruses (Fig. 4a).

With the currently available sequences, analyses based on phylogenies of complete virus genomes are not sufficient to draw firm conclusions on the evolutionary origin of SARS-CoV-2. This leads to various alternative hypotheses about a possible synthetic origin of this virus. For example, it was proposed that SARS-CoV-2 was reconstructed from metagenomic sequences obtained from bat fecal samples. Concerns have also been raised in relation to genetic manipulations of viruses led in order to understand the mechanisms driving the crossing of species barriers. Indeed, experiments of serial passage between model animals and/or cultured cells would lead to a fast evolution by selecting adaptive traits resulting from spontaneous mutations (Sirotkin and Sirotkin 2020).

Genetic manipulations of viruses and gain-of-function experiments

The issue of the natural or synthetic origin of SARS-CoV-2 deserves to be examined in more detail on the basis of available evidence. Hypotheses must be examined knowing which types of genetic manipulations are currently carried out in laboratories. Indeed, the manipulation of genomes of potentially pathogenic viruses is a common practice, which aims at understanding the mechanisms of replication and emergence of these viruses, and at developing new antiviral or vaccine strategies. Due to the risks of unexpected species cross-contamination of a new host (especially humans) and accidental dissemination of artificial recombinant viruses,

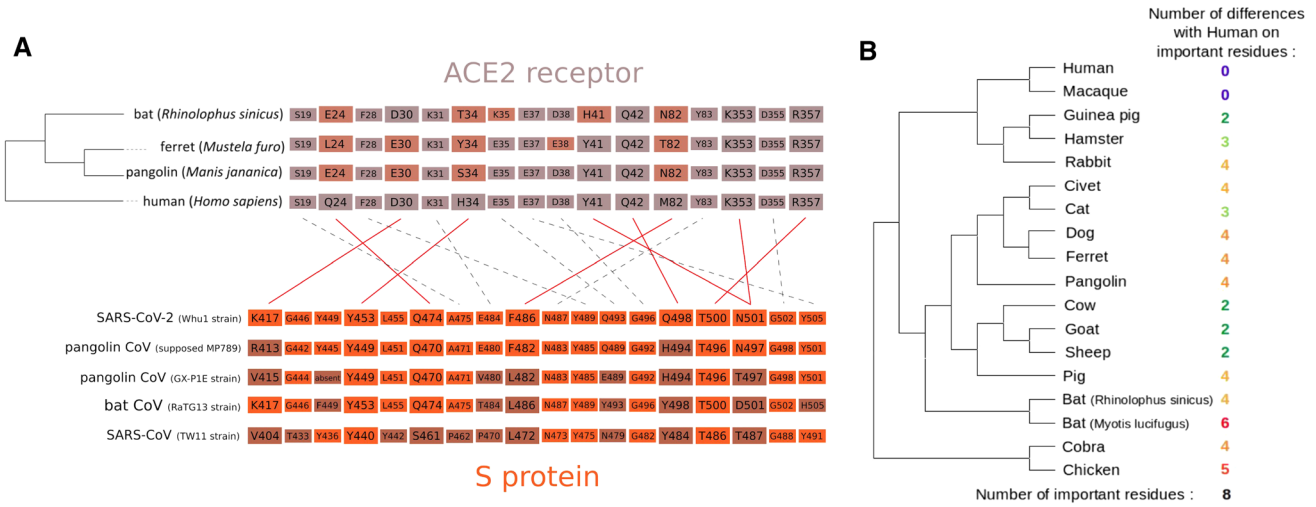
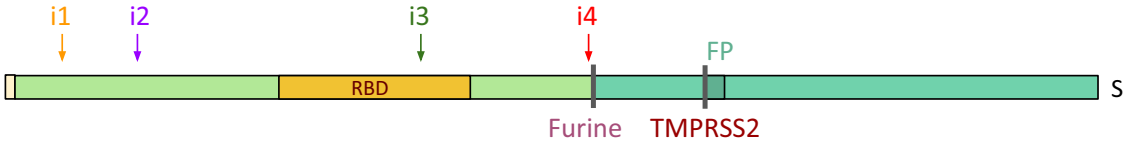
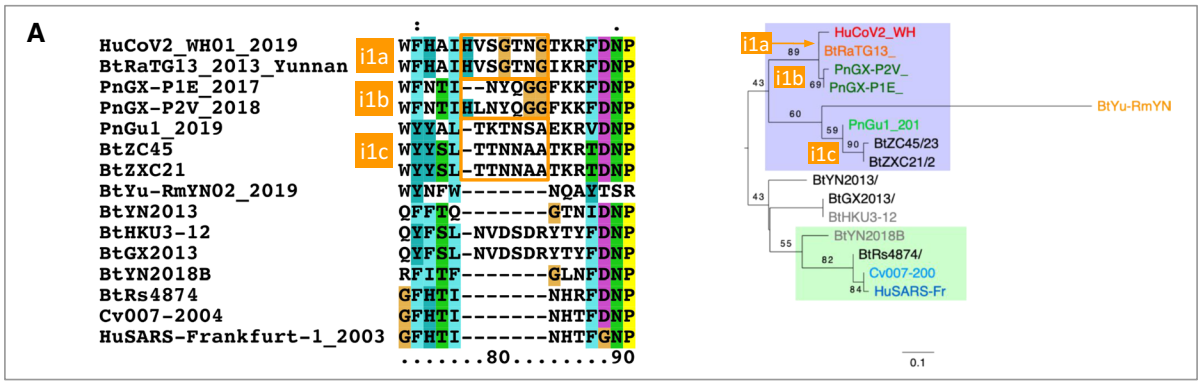
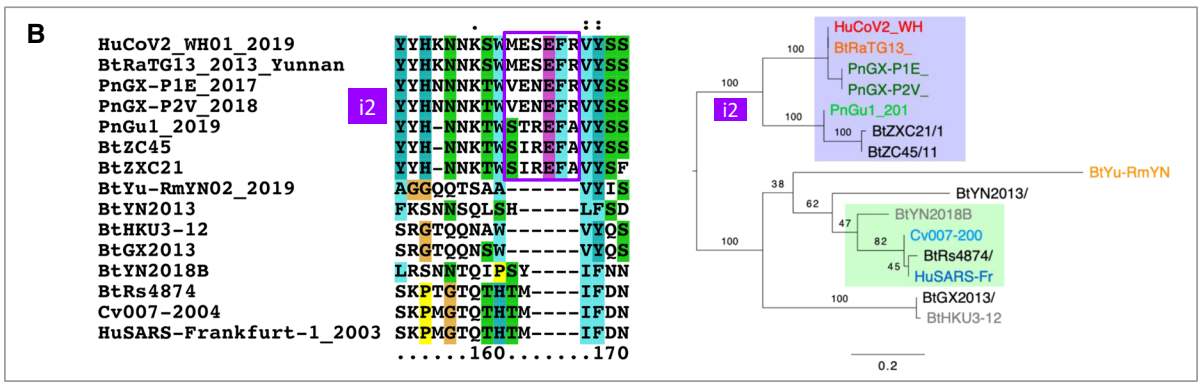
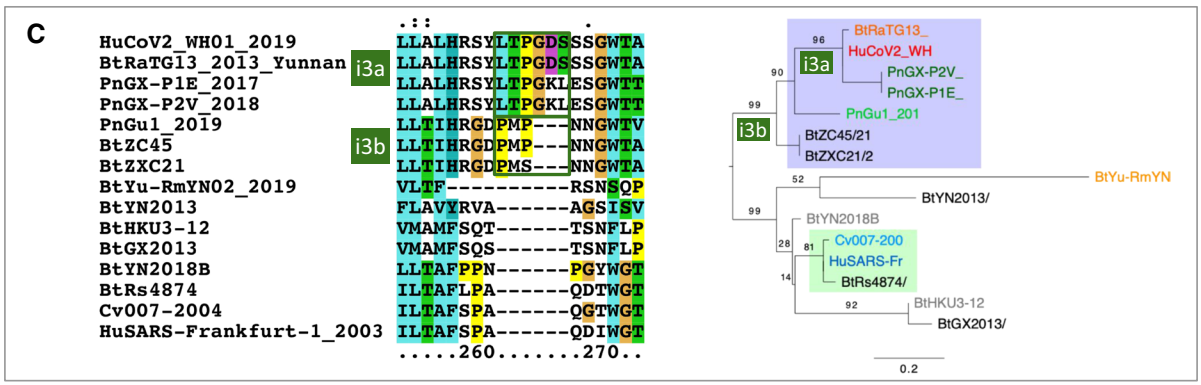


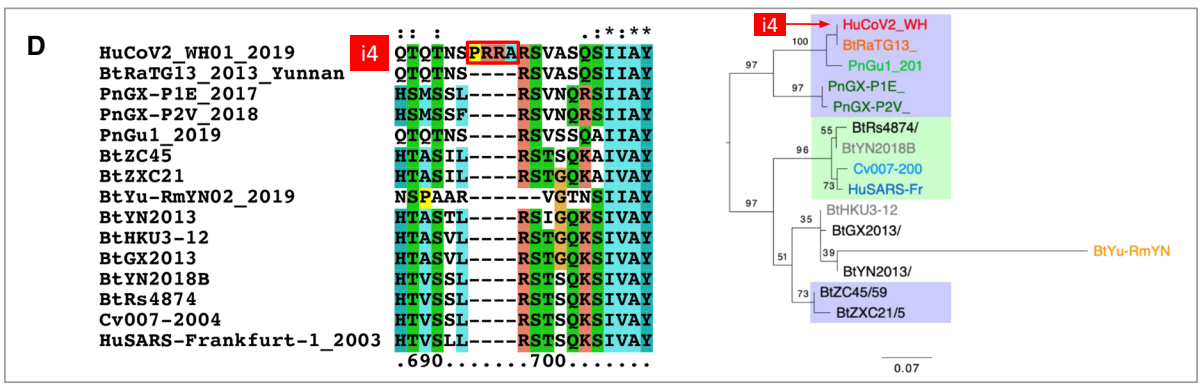
Fig. 4 Conservation of ACE2 proteins and interactions with the viral S protein. **a** Interactions between ACE2 and S and conservation of the key residues [adapted from Wang et al. (2020), Yan et al. (2020)] in different viral strains and animal species. The key interactions between S and ACE2 residues are denoted by solid lines, and weaker interactions by dotted lines. **b** Number of differences between human ACE2 and its ortholog in several animal species for the key residues involved in the interactions with the S protein. [adapted from Yan et al. (2020)]


i1
i2
i3
i4
FP
RBD
S
Furine
TMPRSS2


A
HuCoV2_WH01_2019
BtRaTG13_2013_Yunnan
PnGX-P1E_2017
PnGX-P2V_2018
PnGu1_2019
BtZC45
BtZXC21
BtYu-RmYN02_2019
BtYN2013
BtHKU3-12
BtGX2013
BtYN2018B
BtRs4874
Cv007-2004
HuSARS-Frankfurt-1_2003
i1a
i1b
i1c
.....80.....90
HuCoV2_WH
BtRaTG13_
PnGX-P2V_
PnGX-P1E_
PnGu1_201
BtZC45/23
BtZXC21/2
BtYN2013/
BtGX2013/
BtHKU3-12
BtYN2018B
BtRs4874/
Cv007-200
HuSARS-Fr
BtYu-RmYN
0.1


B
HuCoV2_WH01_2019
BtRaTG13_2013_Yunnan
PnGX-P1E_2017
PnGX-P2V_2018
PnGu1_2019
BtZC45
BtZXC21
BtYu-RmYN02_2019
BtYN2013
BtHKU3-12
BtGX2013
BtYN2018B
BtRs4874
Cv007-2004
HuSARS-Frankfurt-1_2003
i2
.....160.....170
HuCoV2_WH
BtRaTG13_
PnGX-P1E_
PnGX-P2V_
PnGu1_201
BtZXC21/1
BtZC45/11
BtYu-RmYN
BtYN2013/
BtYN2018B
Cv007-200
BtRs4874/
HuSARS-Fr
BtGX2013/
BtHKU3-12
0.2


C
HuCoV2_WH01_2019
BtRaTG13_2013_Yunnan
PnGX-P1E_2017
PnGX-P2V_2018
PnGu1_2019
BtZC45
BtZXC21
BtYu-RmYN02_2019
BtYN2013
BtHKU3-12
BtGX2013
BtYN2018B
BtRs4874
Cv007-2004
HuSARS-Frankfurt-1_2003
i3a
i3b
.....260.....270..
BtRaTG13_
HuCoV2_WH
PnGX-P2V_
PnGX-P1E_
PnGu1_201
BtZC45/21
BtZXC21/2
BtYu-RmYN
BtYN2013/
BtYN2018B
Cv007-200
HuSARS-Fr
BtRs4874/
BtHKU3-12
BtGX2013/
0.2


D
HuCoV2_WH01_2019
BtRaTG13_2013_Yunnan
PnGX-P1E_2017
PnGX-P2V_2018
PnGu1_2019
BtZC45
BtZXC21
BtYu-RmYN02_2019
BtYN2013
BtHKU3-12
BtGX2013
BtYN2018B
BtRs4874
Cv007-2004
HuSARS-Frankfurt-1_2003
i4
.690.....700.....
HuCoV2_WH
BtRaTG13_
PnGu1_201
PnGX-P1E_
PnGX-P2V_
BtRs4874/
BtYN2018B
Cv007-200
HuSARS-Fr
BtHKU3-12
BtGX2013/
BtYu-RmYN
BtYN2013/
BtZC45/59
BtZXC21/5
0.07

◂**Fig. 5** Taxonomic coverage of the insertions observed in SARS-CoV-2 S protein. Each panel shows multiple alignments of amino acid sequences around the insertion (left) and the likely occurrence of the evolutionary event on the phylogenetic trees inferred from the amino acid sequences surrounding the insertions (right). The insertions, respectively, cover the positions 153–158 (**a**), 245–251 (**b**), 445–449, (**c**) and 680–683 (**d**) of SARS-CoV-2 S protein. The schema on the top of the panels indicates the respective positions of the four insertions. Except for insertion i3b, the sequences sharing a same insertion appear grouped in the phylogenetic tree, suggesting a distinct origin for each insertion. The deep difference between tree topologies indicates that these regions of insertions result from different evolutionary stories. The values on the bifurcations denote the bootstrap score (on a scale from 0 to 100), which indicate the robustness of the corresponding branching. A weak bootstrap value (< 50) means that the corresponding branching has a weak reliability. Note that the weak values are often attached to BtYuRmYN02, which results from the metagenomic assembly of a large number of samples for various sources. Consistently, this metagenome is strongly inconsistent between the different aligned fragments, which questions its biological relevance

these investigations are conducted in high-security laboratories (BSL3 or BSL4) subject to strict control and transparency procedures.

The controversy on gain-of-function experiments (increase in virulence or infectivity of viruses by genetic manipulation) began in 2011, following the work of the teams of Ron Fouchier (Russell et al. 2012) and Yoshihiro Kawaoka (Imai et al. 2012) on the influenza virus. In order to understand the virulence factors of influenza, these researchers had tested the effect of mutations that could increase the transmissibility of the H5N1 virus in different animal models. The US Department of Health's National Science Advisory Board for Biosecurity (NSABB), alerted by these experiments in December 2011, asked the journals Nature and Science not to disclose the results of this work on behalf of the significant death toll expected in case of intentional (bioterrorism) or accidental release of these viruses from the laboratory. Because of the importance of the results for public health and the research communities, the NSABB ultimately recommended the general findings to be published, but recommended that the manuscripts should not include "methodological and other details that might allow replication of the experiments by those who would seek to do harm" (Institute of Medicine and National Research Council 2013).

The risk of accidental escape of new potentially pandemic pathogens is increased by the proliferation of high biosafety laboratories (BSL-3 and BSL-4) in densely populated areas (Van Boeckel et al. 2013). In addition, experiments on viruses such as avian influenza viruses or SARS from chiropterans, that are currently unable to infect humans, are allowed in BSL-3 laboratories: It increases the risk of accidents because selection or mutagenesis can confer an epidemic potential to these viruses (Enserink 2003; Normile 2004; Henkel et al. 2012).

Prior to 2002, although they caused major epidemics in livestock, CoVs were considered to be viruses of low public health significance, as they were mainly responsible for benign diseases such as seasonal colds. Since the emergence of SARS-CoV in 2002, studies conducted in the USA and in China have tested the possibility of zoonotic transfer of bat CoVs to humans and attempted to elucidate the processes leading to the emergence of new pathogens (Ren et al. 2008; Zeng et al. 2016; Menachery et al. 2015; Hu et al. 2017).

Recombinant viruses potentially adapted to humans have been constructed from bat CoVs, including through replacement of the bat RBD with the RBD of human SARS-CoV in US and Chinese laboratories (Zeng et al. 2016; Menachery et al. 2015; Hu et al. 2017). Among other discoveries, these experiments nevertheless revealed that infection of human cells is often limited because the activation of the S protein requires specific proteolysis, which is incompletely performed by human cells (Fig. 3b). This difficulty can be circumvented by treating viruses with trypsin (Menachery et al. 2020) or by adding a furin proteolysis site downstream of the RBD domain at the S1/S2 processing site, which can be cleaved by human cells (Follis et al. 2006; Belouzard et al. 2009). These investigations indicate as expected that it is possible to adapt bat viruses to infect human cells or various animal models, and that chiropteran CoVs have the potential for direct zoonotic transmission to humans, particularly if they acquire an adapted proteolysis site, which requires only a few mutations or the insertion of a short sequence rich in basic amino acids (Hu et al. 2017). This hypothesis has been put forward by Sirotkin and Sirotkin, who developed the hypothesis that the virus might have arisen from serial passages, and accidental escape from the laboratory (Sirotkin and Sirotkin 2020).

The spectacular progress in synthetic biology and reverse genetics methods over the last 20 years also increases the risks associated with gain-of-function experiments: it is now possible to assemble a viral genome in about ten days from different DNA fragments synthesized on the basis of sequences from one or more wild virus genomes and to obtain a "new" virus in less than a month (Zeng et al. 2016; Thao et al. 2020; Iseni and Tournier 2020).

HIV sequences and a furin cleavage site inserted into the SARS-CoV-2 S gene?

Doubts about the zoonotic origin of SARS-CoV-2 were raised following the observation of four insertions of short sequences (noted i1 to i4 in Figs. 2b, 3c and 5a–d) within the S1 protein. The fourth insertion (i4) is particularly noteworthy, because it is unique among all the coronaviruses of the SARS group, and because it confers a particular property to the protein (Coutard et al. 2020). This insertion adds

4 amino acids at the precise cleavage site between S1 and S2, immediately upstream of an arginine (Fig. 5d), which creates a sequence RRAR, corresponding to the consensus recognition motif of the furin protease. Similar changes in the cleavage site of viral envelope proteins are known to promote infectivity of different respiratory viruses (e.g., influenza or Sendaï), by facilitating their spread through the respiratory tract and systemic dissemination (Moulard and Decroly 2000; Sun et al. 2010).

The uniqueness of this furin cleavage site in the spike protein of SARS-CoV-2, as well as its conservation in all the isolates of SARS-CoV-2 circulating in human populations, suggests that it has favored, if not allowed, the crossing of the species barrier and/or the evolution of the ancestral virus into a human-to-human transmissible virus. The importance of this conservation for human-to-human transmission is supported by two further observations: First, this proteolysis site is unstable when the virus is grown on some cultured simian cells (VeroE6 strain), and second, experiments on hamsters show reduced symptom severity when the furin site is deleted (Lau et al. 2020). This suggests that a strong selection pressure is exerted on this furin site for the spread of SARS-CoV-2 in humans.

It should also be noted that the appearance of furin cleavage sites in human CoVs is not an exceptional event. Similar sites have been observed in other human CoVs outside the SARS-CoV group, such as MERS, HKU1 and OC43 (Matsuyama et al. 2018; Coutard et al. 2020). Such insertion could result from the presence of palindromic sequences found around the furin cleavage site, thereby providing a natural mechanism to explain the insertion of proteolytic cleavage site (Gallaher 2020).

Three other insertions were identified (Fig. 5a–c). These short sequences are present in SARS-CoV-2 but absent from some chiropteran isolates (i.e., CoVZC45 and CoVZXC21) and from SARS-CoV. The authors of a pre-publication (Pradhan et al. 2020) pointed out a fact they qualify uncanny: at these four insertions, the SARS-CoV-2 S protein shows similarities with fragments of the HIV-1 virus ENV and GAG proteins. However, following critical comments regarding methodological and interpretation weaknesses, the authors withdrew their manuscript from the bioRxiv site.

This "uncanny fact" should therefore have remained anecdotical. Nevertheless, in April 2020, Professor Luc Montagnier, recipient of the 2008 Medicine Nobel Prize for his contribution to the discovery of HIV, made the headlines by claiming on several media that these insertions could not result from natural recombination or accident, but of man-made genetic manipulations, carried out intentionally, presumably as part of a research aimed at developing HIV vaccines. These assertions were immediately challenged by numerous scientists, who argued that the similar sequences between HIV and SARS-CoV-2 are so short (about 30 nucleotides in a genome of 30,000) that their similarity is likely coincidental. The controversy further amplified, in a politically tense context where the President of the USA accused China of having let the virus escape from a BSL-4 laboratory in Wuhan.

Such a controversial climate does not favor a rational analysis of the facts, and paradoxically, no in-depth analysis has been published to date on the origin of these insertions. Yet, as we show below, bioinformatics and molecular phylogeny approaches can provide interesting new information.

Luc Montagnier's hypothesis is based on an analysis of sequence similarities between a fragment of the SARS-CoV-2 S gene and the HIV genome. The most significant alignment obtained by replicating this analysis is shown in Fig. 6a. The significance of the alignment is reflected by BLAST *expect* score, which estimates the statistical expectation, i.e., the number of matches of the same level of similarity that would be found if random sequences were used as queries. A similarity between two sequences is considered significant when the "*expect* score" is much lower than 1. For example, when comparing homologous gene sequences, scores in the order of 10^{-150} are frequently found. On the other hand, an *expect* score higher than 1—such as observed here—means that the similarity is insufficient to support a common ancestral origin of the sequences. Hence, with an *expect* score of 7.5, the alignment of HIV and SARS-CoV-2 sequences does not indicate any sign of homology. This can easily be tested by running the same query with a randomized sequence obtained by shuffling the residues of the S gene. Figure 6b shows the result of this test: The random sequence returns matches as significant as the actual coronavirus gene (Fig. 6a). This confirms that the similarities between coronavirus and HIV are not significant.

In addition, phylogenetic inferences carried out in the vicinity of the insertions (Fig. 5) show that the four insertions found in SARS-CoV-2 cover different sub-groups of coronavirus strains, suggesting that they occurred independently at different times of coronavirus diversification. In particular, the first three insertions are observed in virus sequences isolated not only from human and bats (RaTG13), but also from pangolins from China or Malaysia. The hypothesis that these insertions are the result of recent experimental manipulations would not explain the presence of these sequences in several virus isolates from different species, collected at different locations, especially since these insertions occurred at different times during the evolution of these virus strains.

In this context, how can we understand the appearance and function of these insertions? The analysis of S protein alignments shows that insertions occur very frequently in the coronavirus S gene. Moreover, its structure, resolved by electron cryo-microscopy (Wrapp et al. 2020), indicates that the four SARS-CoV-2 insertions are located on its surface (Fig. 3c), suggesting that they may participate in the escape of the virus to the infection control by antiviral immunity.

A **HIV-1 isolate 19828.PPH11 from Netherlands envelope glycoprotein (env) gene, partial cds**

Sequence ID: HQ644953.1		Length: 1143	Number of Matches: 1	Range 1: 967 to 994
Score	Expect	Identities	Gaps	Strand
38.3 bits(41)	7.5	25/28(89%)	0/28(0%)	Plus/Plus

```
Query  86    AATGGTACTAAGAGGTTTGATAACCCTG  113
            ||||||||||| ||||| |||||| |||
Sbjct  967   AATGGTACTAAAAGGTTAGATAACACTG  994
```

B **HIV-1 isolate patient B clone 16.3 from Netherlands envelope glycoprotein (env) gene, complete cds**

Sequence ID: HQ386166.1		Length: 2580	Number of Matches: 1	Range 1: 2493 to 2523
Score	Expect	Identities	Gaps	Strand
39.2 bits(42)	2.1	27/31(87%)	0/31(0%)	Plus/Minus

```
Query  351   CCTAAAAGTTCTTTGTAATAACTGTATTATT  381
             ||||||||||||||||||||  | ||| |||
Sbjct  2523  CCTAAAAGTTCTTTGTAATATTTCTATAATT  2493
```

Fig. 6 Matches between S gene and HIV genome. **a** Top-ranking alignment between the S gene and the HIV genome. **b** Top-ranking alignment between the randomized query sequence (shuffled nucleotides) and the HIV genome. Note the value of the expect score, which indicates the number of false positives expected by chance. The comparison shows that the alignment between the coding sequence of S protein and the HIV genome is not significant, since the expect score is higher than 1, and even higher for the actual gene than for a random sequence. The alignments were performed on NCBI BLAST server (https://blast.ncbi.nlm.nih.gov/Blast.cgi)

In conclusion, the hypothesis of HIV insertions inside the S gene can be ruled out. The phylogeny indicates that insertions/deletions occur frequently in CoV spike, and that the furin cleavage site, which is unique to SARS-CoV-2 among SARS-like coronaviruses, is the most recent insertion. This insertion plays a key role in SARS-CoV-2 spreading, but unfortunately, the data currently available do not enable us to conclude when, where and how this insertion appeared. Identifying the proximal animal host before the zoonosis might bring an answer to this question.

Current status: the jury is still out

A puzzling question is the origin of the specific features of RBD. It is clear that this RBD cannot come from the 2002 SARS-CoV virus, since SARS-CoV's RBD is genetically very distant from SARS-CoV-2, as shown by PIP profiles as well as the phylogenetic analysis of CoV RBD domains (Fig. 2b, e). In addition, the SARS-CoV-2 residues that play a key role in the recognition of the ACE2 receptor are not conserved in the SARS-CoV (Fig. 4). These sequence differences entail a 20-fold higher affinity for the receptor in SARS-CoV-2 than in SARS-CoV (Walls et al. 2020). However, SARS-CoV-2 binding to target cells is comparable to that of SARS-CoV, as the accessibility of the RBD is suboptimal in the protein at the surface of the virus (Shang et al. 2020a, b).

The article "The proximal origin of Sars-coV-2" (Andersen et al. 2020) is recurrently put forward as a proof for the natural origin of SARS-CoV-2. However, their reasoning does not rely on an actual positive proof of the zoonotic origin. Indeed, as discussed above, the divergence between RaTG13 and SARS-CoV-2 dates from several decades and we still do not know any suitable candidates for the animal hosting the proximal viral strains. Rather, the rationale of this article consists in opposing two mutually exclusive options, which are implicitly considered as exhaustive: either a "natural proximal origin" (i.e., a recent zoonosis), or a virus intentionally designed on the basis of prior knowledge, and constructed by reverse engineering (design hypothesis). They provide two arguments against the design hypothesis (the prior knowledge was insufficient to conceive the RBD, and there is no trace of reverse engineering in the genome) and thus conclude that the virus must be of natural origin. This reasoning is, however, flawed, because it restricts the choice to a dichotomy, whereas several other hypotheses are conceivable. In particular, the authors discard the possibility that the virus would result from laboratory selection through successive passages between animal species or cells, because they consider that the pangolin hypothesis is more parsimonious. However, these two hypotheses are so different that they cannot be evaluated in terms of maximum parsimony. They should rather be compared on the basis of their respective likelihood, but these would currently be very difficult to estimate, in the absence of key information,

in particular on the precise experiments performed in China on closely related viruses before the pandemic. Besides, the pangolin hypothesis has now been strongly questioned (Lee et al. 2020; Choo et al. 2020; Frutos et al. 2020). Regarding the hypothesis of reverse engineering, even though it is not obvious to identify any trace a posteriori there are currently several traceless options for genetic engineering. In conclusion, the arguments supporting the natural proximal origin are so far inconclusive and, albeit this hypothesis has been widely supported by the scientific community (Calisher et al. 2020), alternative hypotheses about a possible laboratory origin cannot be formally ruled out (Relman 2020). This question should thus be re-opened, and all the hypotheses should be evaluated and weighted according to the different elements of information at our disposal.

Discussion and perspectives

We have shown above that bioinformatics analysis can shed light on the possible origins of SARS-CoV-2, the virus responsible for the COVID-19 pandemic. This article reports only preliminary analyses, and further studies are currently being carried out in laboratories to dig into the available data and extract all relevant information. It is hoped that new data will soon be available that will resolve the remaining unanswered questions. The current understanding is therefore incomplete and provisional, but it is useful to ask which conclusions can already be drawn on the basis of available data, and what kind of new results or analyses would provide us with additional information, or even enable us to ascertain the origins of the virus.

The first question is that of the last animal host before man. Phylogenetic analyses indicate that CoVs from chiropterans frequently circulate between different bat species and are occasionally transmitted to other mammals. Virus co-evolution with their host and adaptation to new hosts involve point mutations but also recombinations, which are frequent in coronaviruses. These raise particular difficulties because whole genome-based phylogenetic inference is biased by the mosaicism, since the resulting tree would reflect a mixture of the distinct evolutionary trajectories followed by the different genomic fragments. It is therefore important to identify the recombinant fragments and to perform separate phylogenetic inferences for each one. Available data suggest that SARS-CoV-2 is derived from multiple recombination events between chiropteran CoVs that undoubtedly represent the primary reservoir of the virus. The effect of recombinations is particularly important for the adaptability of the S protein because of its key role in the interaction with the host ACE2 protein and virus entry.

The possible role of pangolin viruses in this process remains uncertain because, although the possible importance of the RBD identified in the CoV from pangolin is established, the region of strong similarity between pangolin virus and SARS-CoV-2 is short and the likelihood of pangolin-to-human transmission could be very low. Furthermore, even the pangolin viruses that are the closest to SARS-CoV-2 (such as MP789), as well as its bat-CoV relatives (notably RaTG13 and RmYN02) display a relatively low identity rate with SARS-CoV-2, suggesting that closer relatives and potentially more recent intermediate hosts remain to be discovered. The discovery of animal viruses sharing a very high similarity with SARS-CoV-2 would validate its natural origin. Consequently, the sequencing of new CoV genomes potentially involved in zoonosis (those circulating in chiropterans and in species in contact with human populations) is requested. It would be necessary to focus primarily on mammalian species whose ACE2 receptor better matches the key characteristics of the human receptor than chiroptera, such as pigs, goats, sheep, cows or cats (Fig. 4b). By bringing people into contact with wildlife in nature or in farms, wildlife trafficking and deforestation should also be questioned. In China, pangolin farms and intensive breeding of minks and raccoon dogs have been spreading, raising new health issues, beyond the questions about the feasibility of such domestication (Hua et al. 2015). In addition, these new exotic farms come alongside all intensive farming of domestic animals (poultry, pigs, etc.), which also creates reservoirs of viruses (influenza, etc.) in areas with high human density (Gibbs et al. 2009).

It should be kept in mind that the reliability of the results depends on the quality of sequencing, metagenomic reconstructions, public accessibility of the data and the accuracy of the annotations in sequence databases (Hassanin 2020).

The insertion between the S1 and S2 subunits of the S protein created a furin-sensitive proteolytic cleavage site which appears to contribute to its infectivity and/or epidemic propensity in humans. This insertion must be recent since it is absent from all the close relatives of SARS-CoV-2. This observation is crucial as this site probably played a key role in the species barrier crossing and/or in the efficiency of human-to-human transmission, which is a prerequisite for the emergence of epidemics.

Knowing that several laboratories are conducting and publishing gain-of-function experiments to characterize the interactions between coronavirus RBD and transmembrane receptors such as ACE2, it has been suggested that SARS-CoV-2 would result from experiments to “humanize” an animal virus of the RaTG13 type. To date, no convincing evidence has been reported from the initial studies carried out by the scientific community. However, bioinformatic analyses revealed biases in codon usages that might reflect some genetic manipulation (Gu et al. 2020). Segreto and Deigin develop the hypothesis of a genome modified by

molecular engineering (Segreto and Deigin 2020). More thorough analyses are warranted to clarify this issue.

Beyond the frame of existing national regulations (e.g., for France the Microorganisms and Toxins Regulations, MOT), at the global level, the identification, the isolation and the culture of these new respiratory viruses must be carried out under the safest possible experimental conditions, with unquestionable traceability, in order to prevent zoonotic transmission. Considering the impact of infectious risks, civil society and the scientific community will have to re-examine the practice of gain-of-function experiments and adaptation to humans in the laboratory, of viral strains cultured in intermediate animal hosts. In 2015, aware of this problem, the US federal agencies froze funding for any new study involving these experiments ("Statement on Funding Pause on Certain Types of Gain-of-Function Research" 2015). This moratorium ended in 2017 (Burki 2018). A new assessment of risks versus potential benefits of these practices should be done. Of course, it is desirable to avoid the pitfall of overly strict regulations that would impede the study of the molecular mechanisms involved in the spread of viruses and thereby prevent the development of antivirals and vaccines. Regardless of its origin, the study of the molecular mechanisms involved in the emergence of potentially pandemic viruses is and will remain essential to develop therapeutic and vaccine strategies.

To conclude, on the basis of currently available data it is not possible to determine whether the emergence of SARS-CoV-2 is the result of a zoonosis from a wild viral strain or an accidental escape of experimental strains. Answering this question is of crucial importance to establish future policies of prevention and biosafety. Indeed, a recent zoonosis would justify enforcing the sampling in natural ecosystems and/or farms and breeding facilities in order to prevent new spillover. Conversely, the perspective of a laboratory escape would call for an in-depth revision of the risk/benefit balance of some laboratory practices, as well as an enforcement of biosafety regulations. As the international team of 10 experts mandated by the WHO enters in China to investigate on SARS-CoV-2 origins (Mallapaty 2020), all the rational hypotheses should be envisaged in an open minded way.

Materials and methods

Reproducibility of the analyses

All the analyses to produce the results and figures of this article follow the FAIR principles (Findable, Accessible, Interoperable and Reusable). The software environment, sequence data, commented code and examples can be downloaded as a github repository (https://github.com/jvanheld/SARS-CoV-2_origins), and the main results can be browsed on the github web pages (https://jvanheld.github.io/SARS-CoV-2_origins/). The software environment is fully described in a yaml-formatted conda configuration file, enabling to re-run all the analyses on Linux or Mac OS X operating systems. The release of the code corresponding to this article is available on zenodo (https://zenodo.org/record/3931505).

A few sequences could, however, not be made available in the github repository because they were downloaded from the GISAID server, which does not allow to redistribute the data and metadata. Since these sequences were crucial to reproduce some key elements of the current debate about SARS-CoV-2 origins, we incorporated them in our analyses. These sequences can be found on the GISAID server (https://www.gisaid.org/) with the IDs EPI_ISL_412977 (Bat virus metagenome RmYN02), EPI_ISL_410544 (Pangolin virus Gu-P2S_2019) and EPI_ISL_410721 (Pangolin virus genome Gu1_2019), respectively.

Collections of viral strains

We defined two collections of viral strains enabling us to highlight different aspects of SARS-CoV-2 origins: (1) "around-CoV-2" regroups human SARS-CoV-2 with 18 other strains from Bat or Pangolins that are closer to SARS-CoV-2 than to any other coronavirus genome; (2) "selected" includes the latter collection plus 23 additional strains representative of other coronavirus groups, including SARS-CoV, MERS-CoV and a few more distant strains. The strains of the collection "around CoV-2" are described in Table 1.

Sequences

Viral sequence genomes were collected from the NCBI Nucleotide database (https://www.ncbi.nlm.nih.gov/nuccore/). A workbook with the identifiers and descriptions of the sequences is included in the github and zenodo releases. S gene sequences were extracted based on the annotation of their coordinates in the NCBI annotations. S protein sequences were obtained from the NCBI protein database (https://www.ncbi.nlm.nih.gov/protein/).

PIP profiles

Profiles of Percent Identical Proteins were computed with an original R script available on the github repository, which enables to draw PIP profiles for either nucleic or peptidic sequences.

For genomic PIP profiles, each viral sequence of interest ("around-cov-2" or "selected" collections) was aligned onto

Table 1 Sources and publication dates for the viral strains discussed in this article

Strain	Host	Isolate origin	Isolate date	Publication date	Precisions concerning the origin of the sample
BtBM48-31	Bat	Bulgaria	2008	October 1, 2010	
BtGX2013	Bat	China	2013	July 7, 2017	
BtHKU3-12	Bat	China (unspecified)	unspecified	April 5, 2010	China according to publication but origin not indicated in NCBI
BtRaTG13_2013_Yunnan	Bat	Yunnan	Jul 24, 2013	March 24, 2020	Sequence published in 2020, annotated as isolated in 2013. Partial genomic sequences (RoRp region) published by Shi group of 2016 i have 100% identity with RaTG13
BtRs4874	Bat	China	Jul 21, 2013	December 18, 2017	Shi's group in Wuhan (Hubei Province, China)
BtYN2013	Bat	China	2013	July 7, 2017	
BtYN2018B	Bat	China	Sep 1, 2016	June 30, 2019	
BtYu-RmYN02_2019	Bat	China, Yunnan—Xishuang-banna	Jun 25, 2019	February 3, 2020	Metagenome constructed by sequencing a mixture of 11 fecal samples from Rhinolophus malayanus bats
BtZC45	Bat	Zhoushan	2017		
BtZXC21	Bat	Zhoushan	2015	February 5, 2020	
Cv007-2004	Civet	China: Guangzhou in Guangdong Province	2019	December 1, 2005	Civet virus closest to the 2003 SARS-CoV. Quoted from the article: "These cases were not linked to any laboratory accident"
HuCoV2_WH01_2019	Human	China, Hubei, Wuhan	Dec 23, 2019	February 11, 2020	Pandemic reference genome for COVID-19
HuSARS-Frankfurt-1_2003	Human	Frankfurt	2003	March 16, 2004	Reference genome for the 2003 SARS epidemic
PnGu-P2S_2019	Pangolin	China, Guangdong	2019	February 17, 2020	Sequence available in GISAID, very close to MP789. Pre-publication version ?
PnMP789	Pangolin	China: smuggled Malayan pangolins, Guangdong customs	Mar 29, 2019	April 23, 2020	Metagenome assembled from samples of 3 pangolins collected in March and July 2019
PnGu1_2019	Pangolin	China, Guangdong	2019	February 18, 2020	
PnGX-P1E_2017	Pangolin	Chinese customs on a flight from Malaysia	2017	April 23, 2020	
PnGX-P2V_2018	Pangolin	Chinese customs on a flight from Malaysia	2018	April 23, 2020	Collected from pangolin, this strain has been cultured on human cells (and therefore presumably suitable for human infection)

the reference genome (SARS-CoV-2) using the Needleman-Wunsch global pairwise alignment algorithm. PIP profiles for coding sequences were based on translation-based multiple alignments of the nucleic sequences with the R function DECIPHER::AlignTranslation(). The PIP was measured on the resulting aligned nucleic sequences, as well as on the aligned protein sequences (not shown in this article).

Phylogenetic analysis

For nucleotide as well as amino acid sequences, we performed multiple alignments with clustalw v2.1 (Larkin et al. 2007) followed by maximum likelihood-based phylogenetic inferences with PhyML v3.3.20190909 (Guindon et al. 2010). We assumed a GTR substitution model for nucleotide sequences and an LG substitution model for amino acid

sequences, with gamma-distributed substitution rates. The other PhyML parameters were left to their default value.

Structural analyses of the spike protein

A model of the full SARS-CoV-2 spike protein was built by aligning the sequence of SARS-CoV-2 spike protein on the PDB 6acc.1.A model of the full SARS-CoV spike protein (the most complete model of a mature coronavirus spike trimmer available to date) using the SWISS-MODEL online tool.

Structural analyses were conducted using the pymol software and the scripts available on https://github.com/jvanheld/SARS-CoV-2_origins/tree/master/scripts/pymol. The 6m0j model and the model we built were aligned, and the insertions identified by running the script https://jvanheld.github.io/SARS-CoV-2_origins/scripts/python/detection_insertion.py on the alignment of all sarbecovirus spike sequences (https://jvanheld.github.io/SARS-CoV-2_origins/results/spike_protein/muscle_alignments/selected_coronavirus_spike_proteins_aligned_muscle.clw) with the SARS-CoV-2 spike as the reference sequence were colored depending on the number of coronaviruses in which this insertion is found.

Supplementary Information The online version of this article (https://doi.org/10.1007/s10311-020-01151-1) contains supplementary material, which is available to authorized users.

Acknowledgements We would like to thank the following colleagues for their careful revision of the early drafts of the French manuscript and their many suggestions for its improvement: Cathy Bellan, Mathias Bonal, Bruno Canard, Bruno Coutard, Hélène Chiapello, Denis Gerlier, Catherine Nguyen, Nadia Rabah, Annick Stevens, Denis Thieffry.

References

Andersen KG, Andrew Rambaut W, Lipkin I, Holmes EC, Garry RF (2020) The proximal origin of SARS-CoV-2. Nat Med 26(4):450–452. https://doi.org/10.1038/s41591-020-0820-9

Belouzard S, Chu VC, Whittaker GR (2009) Activation of the SARS coronavirus spike protein via sequential proteolytic cleavage at two distinct sites. Proc Natl Acad Sci USA 106(14):5871–5876. https://doi.org/10.1073/pnas.0809524106

Burki T (2018) Ban on gain-of-function studies ends. Lancet Infectious Diseases 18(2):148–149. https://doi.org/10.1016/S1473-3099(18)30006-9

Calisher C, Carroll D, Colwell R, Corley RB, Daszak P, Drosten C, Enjuanes L et al (2020) Statement in support of the scientists, public health professionals, and medical professionals of China combatting COVID-19. Lancet 395(10226):e42–e43. https://doi.org/10.1016/S0140-6736(20)30418-9

Casane D, Policarpo M, Laurenti P (2019) Pourquoi le taux de mutation n'est-il jamais égal à zéro ? Médecine/Sciences 35(3):245–251. https://doi.org/10.1051/medsci/2019030

Cheng Vincent C C, Lau Susanna K P, Woo PCY, Yuen KY (2007) Severe acute respiratory syndrome coronavirus as an agent of emerging and reemerging infection. Clin Microbiol Rev 20(4):660–694. https://doi.org/10.1128/CMR.00023-07

Choo SW, Zhou J, Tian X, Zhang S, Qiang S, O'Brien SJ, Tan KY, Platto S, Koepfli KP, Antunes A, Sitam FT (2020) Are pangolins scapegoats of the COVID-19 outbreak-CoV transmission and pathology evidence? Conserv Lett 13(6):e12754

Coutard B, Valle C, de Lamballerie X, Canard B, Seidah NG, Decroly E (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. https://doi.org/10.1016/j.antiviral.2020.104742

Cui J, Li F, Shi Z-L (2019) Origin and evolution of pathogenic coronaviruses. Nat Rev Microbiol 17(3):181–192. https://doi.org/10.1038/s41579-018-0118-9

Drosten C, Günther S, Preiser W, van der Werf S, Brodt H-R, Becker S, Rabenau H et al (2003) Identification of a novel coronavirus in patients with severe acute respiratory syndrome. N Engl J Med 348(20):1967–1976. https://doi.org/10.1056/NEJMoa030747

Eckerle LD, Becker MM, Halpin RA, Li K, Venter E, Xiaotao L, Scherbakova S et al (2010) Infidelity of SARS-CoV Nsp14-exonuclease mutant virus replication is revealed by complete genome sequencing. PLoS Pathog 6(5):e1000896. https://doi.org/10.1371/journal.ppat.1000896

Enserink M (2003) Singapore lab faulted in SARS case. Science 301(5641):1824. https://doi.org/10.1126/science.301.5641.1824b

Ferron F, Subissi L, Morais ATSD, Le NTT, Sevajol M, Gluais L, Decroly E et al (2018) Structural and molecular basis of mismatch correction and ribavirin excision from coronavirus RNA. Proc Natl Acad Sci USA 115(2):E162–E171. https://doi.org/10.1073/pnas.1718806115

Follis KE, York J, Nunberg JH (2006) Furin cleavage of the SARS coronavirus spike glycoprotein enhances cell-cell fusion but does not affect virion entry. Virology 350(2):358–369. https://doi.org/10.1016/j.virol.2006.02.003

Frutos R, Serra-Cobo J, Chen T, Devaux CA (2020) COVID-19: time to exonerate the pangolin from the transmission of SARS-CoV-2 to humans. Infect Genet Evol 84:104493

Gallaher WR (2020) A palindromic RNA sequence as a common breakpoint contributor to copy-choice recombination in SARS-COV-2. Adv Virol 165(10):2341–2348. https://doi.org/10.1007/s00705-020-04750-z

Ge X-Y, Wang N, Zhang W, Ben H, Li B, Zhang Y-Z, Zhou J-H et al (2016) Coexistence of multiple coronaviruses in several bat colonies in an abandoned mineshaft. Virologica Sinica 31(1):31–40. https://doi.org/10.1007/s12250-016-3713-9

Gibbs AJ, Armstrong JS, Downie JC (2009) From where did the 2009 "swine-origin" influenza A virus (H1N1) emerge? Virol J 6(1):207. https://doi.org/10.1186/1743-422X-6-207

Graham RL, Baric RS (2010) Recombination, reservoirs, and the modular spike: mechanisms of coronavirus cross-species transmission. J Virol 84(7):3134–3146. https://doi.org/10.1128/JVI.01394-09

Gu H, Chu DK, Peiris M, Poon LL (2020) Multivariate analyses of codon usage of SARS-CoV-2 and other betacoronaviruses. Virus Evol 6(1):veaa032. https://doi.org/10.1093/ve/veaa032

Guan Y, Zheng BJ, He YQ, Liu XL, Zhuang ZX, Cheung CL, Luo SW et al (2003) Isolation and characterization of viruses related to the SARS coronavirus from animals in Southern China. Science (New York, NY) 302(5643):276–278. https://doi.org/10.1126/science.1087139

Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol 59(3):307–321. https://doi.org/10.1093/sysbio/syq010

Hassanin A (2020) The SARS-CoV-2-like virus found in captive pangolins from Guangdong should be better sequenced. BioRxiv 2020.05.07.077016. Cold Spring Harbor Laboratory

Henkel RD, Miller T, Weyant RS (2012) Monitoring select agent theft, loss and release reports in the United States—2004–2010. Appl Biosaf. https://doi.org/10.1177/153567601201700402

Hoffmann M, Kleine-Weber H, Schroeder S, Krüger N, Herrler T, Erichsen S, Schiergens TS et al (2020) SARS-CoV-2 cell entry depends on ACE2 and TMPRSS2 and is blocked by a clinically proven protease inhibitor. Cell 181(2):271–280.e8. https://doi.org/10.1016/j.cell.2020.02.052

Hu B, Zeng L-P, Yang X-L, Ge X-Y, Zhang W, Li B, Xie J-Z et al (2017) Discovery of a rich gene pool of Bat SARS-related Coronaviruses provides new insights into the origin of SARS Coronavirus. PLoS Pathog 13(11):e1006698. https://doi.org/10.1371/journal.ppat.1006698

Hua L, Gong S, Wang F, Li W, Ge Y, Li X, Hou F (2015) Captive breeding of pangolins: current status, problems and future prospects. ZooKeys 507:99–114. https://doi.org/10.3897/zookeys.507.6970

Huang C, Wang Y, Li X, Ren L, Zhao J, Yi H, Zhang L et al (2020) Clinical features of patients infected with 2019 Novel Coronavirus in Wuhan, China. Lancet 395(10223):497–506. https://doi.org/10.1016/S0140-6736(20)30183-5

Imai M, Watanabe T, Hatta M, Das SC, Ozawa M, Shinya K, Zhong G et al (2012) Experimental adaptation of an influenza H5 HA confers respiratory droplet transmission to a reassortant H5 HA/H1N1 virus in ferrets. Nature 486(7403):420–428. https://doi.org/10.1038/nature10831

Institute of Medicine and National Research Council (2013) Perspectives on research with H5N1 Avian influenza: scientific inquiry, communication, controversy: summary of a workshop. The National Academies Press, Washington, DC. https://doi.org/10.17226/18255

Iseni F, Tournier J-N (2020) Une course contre la montre: Création du SARS-CoV-2 en laboratoire, un mois après son émergence ! Médecine/Sciences. https://doi.org/10.1051/medsci/2020124

Lam TT-Y, Shum MH-H, Zhu H-C, Tong Y-G, Ni X-B, Liao Y-S, Wei W et al (2020) Identifying SARS-CoV-2 related Coronaviruses in Malayan pangolins. Nature. https://doi.org/10.1038/s41586-020-2169-0

Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG (2007) Clustal W and Clustal X version 2.0. Bioinformatics 23(21):2947–2948

Latinne A, Ben H, Olival KJ, Zhu G, Zhang L, Li H, Chmura AA et al (2020) Origin and cross-species transmission of bat Coronaviruses in China. BioRxiv. https://doi.org/10.1101/2020.05.31.116061

Lau S-Y, Wang P, Mok BW-Y, Zhang AJ, Chu H, Lee AC-Y, Deng S et al (2020) Attenuated SARS-CoV-2 variants with deletions at the S1/S2 junction. Emerg Microbes Infect 9(1):837–842. https://doi.org/10.1080/22221751.2020.1756700

Lee J, Hughes T, Lee M-H, Field H, Rovie-Ryan JJ, Sitam FT, Sipangkui S, Nathan SK, Ramirez D, Kumar SV, Lasimbang H, Epstein JH, Daszak P (2020) No evidence of coronaviruses or other potentially zoonotic viruses in Sunda pangolins (Manis javanica) entering the wildlife trade via Malaysia. EcoHealth 17(3):406–418

Letko M, Marzi A, Munster V (2020) Functional assessment of cell entry and receptor usage for SARS-CoV-2 and other Lineage B Betacoronaviruses. Nat Microbiol 5(4):562–569. https://doi.org/10.1038/s41564-020-0688-y

Liu P, Jiang J-Z, Wan X-F, Hua Y, Li L, Zhou J, Wang X et al (2020) Are pangolins the intermediate host of the 2019 Novel Coronavirus (SARS-CoV-2)? PLoS Pathog 16(5):e1008421. https://doi.org/10.1371/journal.ppat.1008421

Lu G, Wang Q, Gao GF (2015) Bat-to-human: spike features determining 'Host Jump' of Coronaviruses SARS-CoV, MERS-CoV, and beyond. Trends Microbiol 23(8):468–478. https://doi.org/10.1016/j.tim.2015.06.003

Lu R, Zhao X, Li J, Niu P, Yang B, Honglong W, Wang W et al (2020) Genomic characterisation and epidemiology of 2019 Novel Coronavirus: implications for virus origins and receptor binding. Lancet (London, England) 395(10224):565–574. https://doi.org/10.1016/S0140-6736(20)30251-8

Luis AD, Hayman David T S, O'Shea TJ, Cryan PM, Gilbert AT, Pulliam Juliet R C, Mills JN et al (2013) A comparison of bats and rodents as reservoirs of zoonotic viruses: are bats special? Proc Biol Sci 280(1756):20122753. https://doi.org/10.1098/rspb.2012.2753

Luk Hayes K H, Li X, Fung J, Lau SKP, Woo PCY (2019) Molecular epidemiology, evolution and phylogeny of SARS Coronavirus. Infect Genet Evol 71:21–30. https://doi.org/10.1016/j.meegid.2019.03.001

Mallapaty S (2020) Meet the scientists investigating the origins of the COVID pandemic. Nature 588(7837):208–208

Matsuyama S, Shirato K, Kawase M, Terada Y, Kawachi K, Fukushi S, Kamitani W (2018) Middle East respiratory syndrome coronavirus spike protein is not activated directly by cellular furin during viral entry into target cells. J Virol. https://doi.org/10.1128/JVI.00683-18

Menachery VD, Yount BL, Debbink K, Agnihothram S, Gralinski LE, Plante JA, Graham RL et al (2015) A SARS-like cluster of circulating bat coronaviruses shows potential for human emergence. Nat Med 21(12):1508–1513. https://doi.org/10.1038/nm.3985

Menachery VD, Dinnon KH, Yount BL, McAnarney ET, Gralinski LE, Hale A, Graham RL et al (2020) Trypsin treatment unlocks barrier for zoonotic bat coronavirus infection. J Virol. https://doi.org/10.1128/JVI.01774-19

Moulard M, Decroly E (2000) Maturation of HIV envelope glycoprotein precursors by cellular endoproteases. Biochim Biophys Acta (BBA) Rev Biomembr 1469(3):121–132. https://doi.org/10.1016/S0304-4157(00)00014-9

Ni L, Ye F, Cheng M-L, Feng Y, Deng Y-Q, Zhao H, Wei P, Ge J, Gou M, Li X, Sun L, Cao T, Wang P, Zhou C, Zhang R, Liang P, Guo H, Wang X, Qin C-F, Chen F, Dong C (2020) Detection of SARS-CoV-2-specific humoral and cellular immunity in COVID-19 convalescent individuals. Immunity 52(6):971–977

Normile D (2004) Lab accidents prompt calls for new containment program. Science 304(5675):1223–1225. https://doi.org/10.1126/science.304.5675.1223a

Pradhan P, Pandey AK, Mishra A, Gupta P, Tripathi PK, Menon MB, Gomes J, Vivekanandan P, Kundu B (2020) Uncanny similarity of unique inserts in the 2019-NCoV spike protein to HIV-1 Gp120 and gag. BioRxiv. https://doi.org/10.1101/2020.01.30.927871

Relman DA (2020) Opinion: to stop the next pandemic, we need to unravel the origins of COVID-19. Proc Nat Acad Sci 117(47):29246–29248

Ren W, Xiuxia Q, Li W, Han Z, Meng Yu, Zhou P, Zhang S-Y, Wang L-F, Deng H, Shi Z (2008) Difference in receptor usage between severe acute respiratory syndrome (SARS) coronavirus and

SARS-like coronavirus of bat origin. J Virol 82(4):1899–1907. https://doi.org/10.1128/JVI.01085-07

Rissanen I, Ahmed AA, Azarm K, Beaty S, Hong P, Nambulli S, Duprex WP, Lee B, Bowden TA (2017) Idiosyncratic Mòjiāng virus attachment glycoprotein directs a host-cell entry pathway distinct from genetically related henipaviruses. Nat Commun 8(1):1–11

Russell CA, Fonville JM, Brown André E X, Burke DF, Smith DL, James SL, Herfst S et al (2012) The potential for respiratory droplet-transmissible A/H5N1 influenza virus to evolve in a mammalian host. Science (New York, N.Y.) 336(6088):1541–1547. https://doi.org/10.1126/science.1222526

Sabir Jamal S M, Lam Tommy T-Y, Ahmed Mohamed M M, Li L, Shen Y, Abo-Aba Salah E M, Qureshi MI et al (2016) Co-circulation of three camel coronavirus species and recombination of MERS-CoVs in Saudi Arabia. Science (New York, N.Y.) 351(6268):81–84. https://doi.org/10.1126/science.aac8608

Segreto R, Deigin Y (2020) The genetic structure of SARS-CoV-2 does not rule out a laboratory origin: SARS-COV-2 chimeric structure and furin cleavage site might be the result of genetic manipulation. BioEssays. https://doi.org/10.1002/bies.202000240

Shang J, Wan Y, Luo C, Ye G, Geng Q, Auerbach A, Li F (2020a) Cell entry mechanisms of SARS-CoV-2. Proc Nat Acad Sci 117(21):11727–11734

Shang J, Ye G, Shi K, Wan Y, Luo C, Aihara H, Geng Q, Auerbach A, Li F (2020b) Structural basis of receptor recognition by SARS-CoV-2. Nature 581(7807):221–224

Sirotkin K, Sirotkin D (2020) Might SARS-CoV-2 have arisen via serial passage through an animal host or cell culture? A potential explanation for much of the novel coronavirus' distinctive genome. BioEssays. https://doi.org/10.1002/bies.202000091

Song H-D, Chang-Chun T, Zhang G-W, Wang S-Y, Zheng K, Lei L-C, Chen Q-X et al (2005) Cross-host evolution of severe acute respiratory syndrome coronavirus in palm civet and human. Proc Natl Acad Sci USA 102(7):2430–2435. https://doi.org/10.1073/pnas.0409608102

Statement on Funding Pause on Certain Types of Gain-of-Function Research. 2015. National Institutes of Health (NIH). 20 janvier 2015. https://www.nih.gov/about-nih/who-we-are/nih-director/statements/statement-funding-pause-certain-types-gain-function-research

Sun X, Tse LV, Damon Ferguson A, Whittaker GR (2010) Modifications to the hemagglutinin cleavage site control the virulence of a neurotropic H1N1 influenza virus. J Virol 84(17):8683–8690. https://doi.org/10.1128/JVI.00797-10

Thao TT, Nhu FL, Ebert N, V'kovski P, Stalder H, Portmann J, Kelly J et al (2020) Rapid reconstruction of SARS-CoV-2 using a synthetic genomics platform. Nature. https://doi.org/10.1038/s41586-020-2294-9

Van Boeckel TP, Tildesley MJ, Linard C, Halloy J, Keeling MJ, Gilbert M (2013) The Nosoi commute: a spatial perspective on the rise of BSL-4 laboratories in cities. arXiv:1312.3283 [q-bio], décembre. http://arxiv.org/abs/1312.3283

Walls AC, Park Y-J, Tortorici MA, Wall A, McGuire AT, Veesler D (2020) Structure, function, and antigenicity of the SARS-CoV-2 spike glycoprotein. Cell 181(2):281–292

Wang Q, Zhang Y, Lili W, Niu S, Song C, Zhang Z, Guangwen L et al (2020) Structural and functional basis of SARS-CoV-2 entry by using human ACE2. Cell 181(4):894–904.e9. https://doi.org/10.1016/j.cell.2020.03.045

Wrapp D, Wang N, Corbett KS, Goldsmith JA, Hsieh C-L, Abiona O, Graham BS, McLellan JS (2020) Cryo-EM structure of the 2019-NCoV spike in the prefusion conformation. Science (New York, N.Y.) 367(6483):1260–1263. https://doi.org/10.1126/science.abb2507

Wu Z, Yang L, Yang F, Ren X, Jiang J, Dong J, Sun L, Zhu Y, Zhou H, Jin Q (2014) Novel henipa-like virus, Mojiang paramyxovirus, in rats, China, 2012. Emerg Infect Dis 20(6):1064

Wu F, Zhao S, Bin Yu, Chen Y-M, Wang W, Song Z-G, Yi H et al (2020) A new coronavirus associated with human respiratory disease in China. Nature 579(7798):265–269. https://doi.org/10.1038/s41586-020-2008-3

Xiao K, Zhai J, Feng Y, Niu Zhou X, Zhang J-JZ, Li N et al (2020) Isolation of SARS-CoV-2-related coronavirus from malayan pangolins. Nature 583(7815):286–289. https://doi.org/10.1038/s41586-020-2313-x

Yan R, Zhang Y, Yaning Li L, Xia YG, Zhou Q (2020) Structural basis for the recognition of SARS-CoV-2 by full-length human ACE2. Science 367(6485):1444–1448. https://doi.org/10.1126/science.abb2762

Zaki AM, van Boheemen S, Bestebroer TM, Osterhaus Albert DME, Fouchier Ron AM (2012) Isolation of a novel coronavirus from a man with pneumonia in Saudi Arabia. N Engl J Med 367(19):1814–1820. https://doi.org/10.1056/NEJMoa1211721

Zeng L-P, Gao Y-T, Ge X-Y, Zhang Q, Peng C, Yang X-L, Tan B et al (2016) Bat severe acute respiratory syndrome-like coronavirus WIV1 encodes an extra accessory protein, ORFX, involved in modulation of the host immune response. J Virol 90(14):6573–6582. https://doi.org/10.1128/JVI.03079-15

Zhang T, Qunfu W, Zhang Z (2020) Probable pangolin origin of SARS-CoV-2 associated with the COVID-19 outbreak. Curr Biol 30(7):1346–1351.e2. https://doi.org/10.1016/j.cub.2020.03.022

Zhou H, Chen X, Tao H, Li J, Song H, Liu Y, Wang P et al (2020a) 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. https://doi.org/10.1016/j.cub.2020.05.023

Zhou P, Yang X-L, Wang X-G, Ben H, Zhang L, Zhang W, Si H-R et al (2020b) A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature 579(7798):270–273. https://doi.org/10.1038/s41586-020-2012-7

Zhou P, Yang X-L, Wang X-G, Ben H, Zhang L, Zhang W, Si H-R et al (2020c) Addendum: a pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature. https://doi.org/10.1038/s41586-020-2951-z

Ziegler CGK, Allon SJ, Nyquist SK, Mbano IM, Miao VN, Tzouanas CN, Cao Y et al (2020) SARS-CoV-2 receptor ACE2 is an interferon-stimulated gene in human airway epithelial cells and is detected in specific cell subsets across tissues. Cell. https://doi.org/10.1016/j.cell.2020.04.035