



HAL
open science

Tracing the origins of SARS-CoV-2 in coronavirus phylogenies

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

► **To cite this version:**

Erwan Sallard, José Halloy, Didier Casane, Etienne Decroly, Jacques van Helden. Tracing the origins of SARS-CoV-2 in coronavirus phylogenies. In press. hal-02891455v7

HAL Id: hal-02891455

<https://hal.science/hal-02891455v7>

Submitted on 6 Aug 2020 (v7), last revised 24 Nov 2020 (v8)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution - NonCommercial 4.0 International License

Tracing the origins of SARS-COV-2 in coronavirus phylogenies

Translation of the French manuscript to appear in Médecine/Sciences, Aug/Sept 2020

Erwan Sallard¹, José Halloy², Didier Casane^{3,4}, Etienne Decroly^{5*} and Jacques van Helden^{6,7**}

- 1) École Normale Supérieure de Paris, 45 rue d'Ulm, 75005 Paris, France.
ORCID: [0000-0002-2324-3633](https://orcid.org/0000-0002-2324-3633).
- 2) Université de Paris, CNRS, LIED UMR 8236, 85 bd Saint-Germain, 75006 Paris, France.
ORCID: [0000-0003-1555-2484](https://orcid.org/0000-0003-1555-2484)
- 3) Université Paris-Saclay, CNRS, IRD, UMR Évolution, Génomes, Comportement et Écologie, 91198, Gif-sur-Yvette, France. ORCID: [0000-0001-5463-1092](https://orcid.org/0000-0001-5463-1092)
- 4) Université de Paris, UFR Sciences du Vivant, F-75013 Paris, France.
- 5) Aix-Marseille Univ, CNRS, UMR 7257, AFMB, Case 925, 163 Avenue de Luminy, 13288 Marseille Cedex 09, France. ORCID: [0000-0002-6046-024X](https://orcid.org/0000-0002-6046-024X)
- 6) CNRS, Institut Français de Bioinformatique, IFB-core, UMS 3601, Evry, France.
ORCID: [0000-0002-8799-8584](https://orcid.org/0000-0002-8799-8584)
- 7) Aix-Marseille Univ, INSERM, lab. Theory and Approaches of Genome Complexity (TAGC), Marseille, France.

* etienne.decroly@afmb.univ-mrs.fr, ** Jacques.van-Helden@univ-amu.fr

The last two authors have equally contributed

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 540,000 deaths in six months. Understanding the origin of this virus is an important issue and it is necessary to determine the mechanisms of its 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 evoked to account for the origin - natural or synthetic - of the virus. Based on available data, it is currently impossible to firmly assert whether SARS-CoV2 results from a natural zoonotic emergence or from an accidental escape of a laboratory strain. Regardless of its origin, studying the evolution of the molecular mechanisms involved in the emergence of this pandemic virus is essential to develop therapeutic and vaccine strategies.

The COVID-19 pandemic

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 into a pandemic in early 2020. By July 5, 2020, more than 11.2 million infections had been reported with at least 530,000 deaths. The etiological agent of COVID-19 was rapidly identified and by 26 January 2020, 10 viral genomes had been sequenced (R. 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 (R. 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 the bat (P. Zhou et al. 2020; R. Lu et al. 2020).

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 (Menachery et al. 2015; Hu et al. 2017; Luk et al. 2019). 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, Li, et Shi 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 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 strains isolated from the last intermediate hosts before transmission to humans: civets for SARS-CoV in 2003 (**Figure 1B**) (Guan et al. 2003; Song et al. 2005), and camels for MERS-CoV (**Figure 1C**) (Sabir et al. 2016). In the latter case, several zoonotic transmissions (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, which enables its entry into the host cell (Hu et al. 2017).

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 faeces 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 (P. Zhou et al. 2020). Unfortunately, the precise location of the sample collection is documented neither in the article nor in the sequence databases. However, 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).

More recently, a metagenome (RmYN02) was assembled from faeces 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 (**Figure 2**) (H. Zhou et al. 2020).

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 are about 10,000 and 19,000 nucleotides, respectively). CoVs are able to maintain such a long a genome thanks to a replication error correction system unique in the world of RNA viruses, depending on a viral exonuclease that ensured a proofreading mechanism limiting the mutation rate (Eckerle et al. 2010; Ferron et al. 2018; Casane, Policarpo, et Laurenti 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 (**Figure 2A**).

The CoV viral polymerase, in addition to its canonical RdRp activity, is able to jump between different RNA strands during replication, 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 et Baric 2010). It is believed that the SARS-CoV genome is a "mosaic" genome composed of pieces of at least two pre-existing CoVs.

A recombinant genome can be detected by comparing its Percent of Identical Positions (PIP) profiles 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 (**Figure 2 A-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 inference should thus be made for each recombinant region separately, as illustrated in **Figures 2C to 2E**, where we have inferred evolutionary trees respectively from the genomic sequences of the ORF1ab polyprotein (**Figure 2C**), the S1 subunit (**Figure 2D**), and the receptor binding domain (RBD) (**Figure 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.

Of bats and men... plus some pangolins?

If we zoom in on the S gene (**Figure 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 (K. Wu et al. 2012; 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. 2020a) to hypothesize that SARS-CoV-2 could result from recombinations between viruses infecting bats and pangolins respectively (**Figure 1D**). It should however be noted that, even in Pangolins, there is currently no known non-human virus 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 commonly accepted hypothesis is 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 infected patients did not attend the Wuhan market (Huang et al. 2020), and on the other hand, despite the search for viruses in the animal species sold on this market (Zhang, Wu, et Zhang 2020), to date no intermediate virus has been identified that may result from the supposed recombination between a bat virus and a pangolin virus .

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 pangolin or wild bats: in the former case, transmission to humans would be favoured 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 (**Figure 4**). Another hypothesis is that the similarity between the RBD sequences of pangolin and SARS-CoV-2 results from a convergent evolution.

In the absence of evidence regarding the last animal intermediate before human contamination, some authors suggested that SARS-CoV-2 may have been manufactured in a laboratory (synthetic origin). However, this hypothesis has been refuted by many specialists, notably on the basis of phylogenetic studies that suggest 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 (Andersen et al. 2020; Latinne et al. 2020; P. Zhou et al. 2020; Lam et al. 2020; Xiao, Zhai, Feng, Zhou, Zhang, Zou, Li, Guo, Li, Shen, Zhang, Shu, Huang, Li, Zhang, Chen, Wu, Peng, Huang, Xie, Cai, Hou, Chen, et al. 2020; Zhang, Wu, et Zhang 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. 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 Human) 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 protect against new outbreaks (Cheng et al. 2007).

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 protrusions on the viral surface, from which the name of coronavirus derives (**Figure 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 allows the viral genome penetrating 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 (**Figure 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 beginning of the S2' subunit. These two proteolytic cleavages are likely catalysed respectively by furin and TMPRSS2 (transmembrane protease 2) (Hoffmann et al. 2020). The S protein cleavage is essential for the formation of infectious viral particles, as they allow fusion between the viral and cell membranes (**Figure 3A**).

The S1 protein of SARS-CoV and SARS-CoV-2 contains the RBD domain (**Figure 3B** and **4**) which ensures the recognition of the ACE2 receptor by the virus (Wrapp et al. 2020; F. Wu et al. 2020; Lam et al. 2020). It also bears most of the exposed sites on the virus surface (**Figure 3C**), including the 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 (**Figure 4**). Some key residues are required for efficient infection of chiropterans, the intermediate host, or humans (G. Lu, Wang, et Gao 2015; Letko, Marzi, et Munster 2020; Wang 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% (**Figure 2A**, strain MP789) vs. 96% for RaTG13, the amino acid identity is 98% in the RBD domain (Xiao, Zhai, Feng, Zhou, Zhang, Zou, Li, Guo, Li, Shen, Zhang, Shu, Huang, Li, Zhang, Chen, Wu, Peng, Huang, Xie, Cai, Hou, Liu, et al. 2020). For the same domain, the nucleotide identity rate is only 89% between MP789 and SARS-CoV-2: this difference 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 (**Figure 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 speculations 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 faecal samples. Similar concerns have been raised in relation to genetic manipulations of viruses in order to understand the mechanisms driving the the species barrier crossing.

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 type 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 crossing, contamination of a new host (especially humans) and accidental dissemination of artificial recombinant viruses, 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 on the influenza virus (Imai et al. 2012). 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 U.S. 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" (Committee on Science et al. 2013).

The risk of accidental escape of new pathogens are 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, Miller, et Weyant 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 United States 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).

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 (Menachery et al. 2015; Zeng et al. 2016). 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 (**Figure 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, which can be cleaved by human cells (Follis, York, et Nunberg 2006; Belouzard, Chu, et Whittaker 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).

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 (Zeng et al. 2016; Thao et al. 2020) and to obtain a "new" virus in less than a month.

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

Doubts about the natural origin of SARS-CoV-2 were raised following the observation of 4 insertions of short sequences (noted i1 to i4 in **Figures 3C, 2B and 5A-D**) within the S1 protein. The fourth insertion (i4) is particularly noteworthy, because it is unique to CoV-SARS2, 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 (**Figure 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 Sendai), by facilitating their spread through the respiratory tract and systemic dissemination (Moulard et Decroly 2000; Sun et al. 2010).

The conservation of this specific site of furin cleavage in all the isolates of SARS-CoV-2 circulating in human populations suggests that it has favoured, if not allowed, the crossing of the species barrier and/or the evolution of the epidemic form of the 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 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 CoV-2-CoV-SARS 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 human CoVs outside the SARS-CoV group, such as MERS, HKU1, OC43 (Matsuyama et al. 2018; Coutard et al. 2020).

Three other insertions were identified (**Figure 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 prepublication (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 anecdotal. 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 United States accused China of having let the virus escape from a BSL-4 laboratory in Wuhan.

Such a controversial climate does not favour a rational analysis of the facts and, paradoxically, no in-depth analysis has been published to date on the origin of these insertions. Yes, 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 **Figure 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* 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 randomised sequence obtained by shuffling the residues of the spike protein. **Figure 6B** shows the result of this test: the random sequence returns matches as significant as the coronavirus gene (**Figure 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 (**Figure 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 protein. Moreover, its structure, recently resolved by electron cryo-microscopy (Wrapp et al. 2020), indicates that the four SARS-CoV-2 insertions are located on its surface (**Figure 3C**), suggesting that they participate in the escape of the virus to the infection control by antiviral immunity.

A second, frequently raised hypothesis is that this virus could result from a man-made recombination between a RaTG13-like bat virus and an RBD domain with high affinity for human ACE2, cloned from SARS-CoV in a laboratory. However, this hypothesis is also inconsistent with phylogenetic analyses of the RBD domains of CoV, because SARS-CoV's RBD is genetically very distant from SARS-CoV-2 (**Figure 2B**). 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 (**Figure 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 former (Shang et al. 2020).

Current status 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 coevolution 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. 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.

The possible role of pangolin viruses in this process remains uncertain because, although its functional importance 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 (**Figure 4B**). 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.

By bringing people into contact with wildlife in nature or in farms, wildlife trafficking and deforestation should also be questioned. In China, pangolin farms 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, Armstrong, et Downie 2009).

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, is 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). More thorough analyses might contribute 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 coronaviruses and thereby prevent the development of antivirals and vaccines. However, high-risk practices should be re-evaluated and regulated at the international level.

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 natural zoonosis or an accidental escape of experimental strains. 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.

Material 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 sequences. 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.

Sequences

Viral sequence genomes were collected from the NCBI Nucleotide database (<https://www.ncbi.nlm.nih.gov/nucleotide/>). 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 by translating the corresponding gene sequences.

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 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/elected_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 files

Table 1. Metadata about the viral genomes and S proteins used in this article.

File: selected_coronavirus_genomes-and-S-proteins_2020-07-06.xlsx

Acknowledgements

We would like to thank the following colleagues for their careful revisions of the early drafts of the manuscript, and their many suggestions for its improvement: Mathias Bonal, Bruno Canard, Bruno Coutard, H el ene Chiapello, Denis Gerlier, Catherine Nguyen, Nadia Rabah, Annick Stevens, Denis Thieffry.

References

- Andersen, Kristian G., Andrew Rambaut, W. Ian Lipkin, Edward C. Holmes, et Robert F. Garry. 2020. « The Proximal Origin of SARS-CoV-2 ». *Nature Medicine* 26 (4): 450-52. <https://doi.org/10.1038/s41591-020-0820-9>.
- Belouzard, Sandrine, Victor C. Chu, et Gary R. Whittaker. 2009. « Activation of the SARS Coronavirus Spike Protein via Sequential Proteolytic Cleavage at Two Distinct Sites ». *Proceedings of the National Academy of Sciences of the United States of America* 106 (14): 5871-76. <https://doi.org/10.1073/pnas.0809524106>.
- Burki, Talha. 2018. « Ban on Gain-of-Function Studies Ends ». *The Lancet Infectious Diseases* 18 (2): 148-49. [https://doi.org/10.1016/S1473-3099\(18\)30006-9](https://doi.org/10.1016/S1473-3099(18)30006-9).
- Casane, Didier, Maxime Policarpo, et Patrick Laurenti. 2019. « Pourquoi le taux de mutation n'est-il jamais  egal  a z ero ? ». *m edecine/sciences* 35 (3): 245-51. <https://doi.org/10.1051/medsci/2019030>.
- Cheng, Vincent C. C., Susanna K. P. Lau, Patrick C. Y. Woo, et Kwok Yung Yuen. 2007. « Severe Acute Respiratory Syndrome Coronavirus as an Agent of Emerging and Reemerging Infection ». *Clinical Microbiology Reviews* 20 (4): 660-94. <https://doi.org/10.1128/CMR.00023-07>.
- Coutard, B., C. Valle, X. de Lamballerie, B. Canard, N. G. Seidah, et E. Decroly. 2020. « The Spike Glycoprotein of the New Coronavirus 2019-NCoV Contains a Furin-like Cleavage Site Absent in CoV of the Same Clade ». *Antiviral Research* 176: 104742. <https://doi.org/10.1016/j.antiviral.2020.104742>.
- Cui, Jie, Fang Li, et Zheng-Li Shi. 2019. « Origin and evolution of pathogenic coronaviruses ». *Nature Reviews Microbiology* 17 (3): 181-92. <https://doi.org/10.1038/s41579-018-0118-9>.
- Drosten, Christian, Stephan G unther, Wolfgang Preiser, Sylvie van der Werf, Hans-Reinhard

- Brodthorn, Stephan Becker, Holger Rabenau, et al. 2003. « Identification of a Novel Coronavirus in Patients with Severe Acute Respiratory Syndrome ». *New England Journal of Medicine* 348 (20): 1967-76. <https://doi.org/10.1056/NEJMoa030747>.
- Eckerle, Lance D., Michelle M. Becker, Rebecca A. Halpin, Kelvin Li, Eli Venter, Xiaotao Lu, Sana Scherbakova, et al. 2010. « Infidelity of SARS-CoV Nsp14-Exonuclease Mutant Virus Replication Is Revealed by Complete Genome Sequencing ». *PLoS Pathogens* 6 (5): e1000896. <https://doi.org/10.1371/journal.ppat.1000896>.
- Enserink, Martin. 2003. « Singapore Lab Faulted in SARS Case ». *Science* 301 (5641): 1824-1824. <https://doi.org/10.1126/science.301.5641.1824b>.
- Ferron, François, Lorenzo Subissi, Ana Theresa Silveira De Moraes, Nhung Thi Tuyet Le, Marion Sevajol, Laure Gluais, Etienne Decroly, et al. 2018. « Structural and Molecular Basis of Mismatch Correction and Ribavirin Excision from Coronavirus RNA ». *Proceedings of the National Academy of Sciences of the United States of America* 115 (2): E162-71. <https://doi.org/10.1073/pnas.1718806115>.
- Follis, Kathryn E., Joanne York, et Jack H. Nunberg. 2006. « Furin Cleavage of the SARS Coronavirus Spike Glycoprotein Enhances Cell-Cell Fusion but Does Not Affect Virion Entry ». *Virology* 350 (2): 358-69. <https://doi.org/10.1016/j.virol.2006.02.003>.
- Ge, Xing-Yi, Ning Wang, Wei Zhang, Ben Hu, Bei Li, Yun-Zhi Zhang, Ji-Hua Zhou, 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, Adrian J., John S. Armstrong, et Jean C. Downie. 2009. « From where did the 2009 “swine-origin” influenza A virus (H1N1) emerge? ». *Virology Journal* 6 (1): 207. <https://doi.org/10.1186/1743-422X-6-207>.
- Graham, Rachel L., et Ralph S. Baric. 2010. « Recombination, Reservoirs, and the Modular Spike: Mechanisms of Coronavirus Cross-Species Transmission ». *Journal of Virology* 84 (7): 3134-46. <https://doi.org/10.1128/JVI.01394-09>.
- Guan, Y., B. J. Zheng, Y. Q. He, X. L. Liu, Z. X. Zhuang, C. L. Cheung, S. W. Luo, et al. 2003. « Isolation and Characterization of Viruses Related to the SARS Coronavirus from Animals in Southern China ». *Science (New York, N.Y.)* 302 (5643): 276-78. <https://doi.org/10.1126/science.1087139>.
- Guindon, Stéphane, Jean-François Dufayard, Vincent Lefort, Maria Anisimova, Wim Hordijk, et Olivier Gascuel. 2010. « New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0 ». *Systematic Biology* 59 (3): 307-21. <https://doi.org/10.1093/sysbio/syq010>.
- Henkel, Richard D., Thomas Miller, et Robbin S. Weyant. 2012. « Monitoring Select Agent Theft, Loss and Release Reports in the United States—2004–2010 ». *Applied Biosafety*, décembre. <https://doi.org/10.1177/153567601201700402>.
- Hoffmann, Markus, Hannah Kleine-Weber, Simon Schroeder, Nadine Krüger, Tanja Herrler, Sandra Erichsen, Tobias S. Schiergens, 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, Ben, Lei-Ping Zeng, Xing-Lou Yang, Xing-Yi Ge, Wei Zhang, Bei Li, Jia-Zheng Xie, et al. 2017. « Discovery of a Rich Gene Pool of Bat SARS-Related Coronaviruses Provides New Insights into the Origin of SARS Coronavirus ». *PLoS Pathogens* 13 (11): e1006698. <https://doi.org/10.1371/journal.ppat.1006698>.
- Hua, Liushuai, Shiping Gong, Fumin Wang, Weiye Li, Yan Ge, Xiaonan Li, et Fanghui Hou. 2015. « Captive Breeding of Pangolins: Current Status, Problems and Future Prospects ». *ZooKeys*, n° 507: 99-114. <https://doi.org/10.3897/zookeys.507.6970>.
- Huang, Chaolin, Yeming Wang, Xingwang Li, Lili Ren, Jianping Zhao, Yi Hu, Li Zhang, et al. 2020. « Clinical Features of Patients Infected with 2019 Novel Coronavirus in Wuhan, China ». *The Lancet* 395 (10223): 497-506.

- [https://doi.org/10.1016/S0140-6736\(20\)30183-5](https://doi.org/10.1016/S0140-6736(20)30183-5).
- Imai, Masaki, Tokiko Watanabe, Masato Hatta, Subash C. Das, Makoto Ozawa, Kyoko Shinya, Gongxun Zhong, 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): 420-28. <https://doi.org/10.1038/nature10831>.
- Lam, Tommy Tsan-Yuk, Marcus Ho-Hin Shum, Hua-Chen Zhu, Yi-Gang Tong, Xue-Bing Ni, Yun-Shi Liao, Wei Wei, et al. 2020. « Identifying SARS-CoV-2 Related Coronaviruses in Malayan Pangolins. » *Nature*, mars. <https://doi.org/10.1038/s41586-020-2169-0>.
- Latinne, Alice, Ben Hu, Kevin J. Olival, Guangjian Zhu, Libiao Zhang, Hongying Li, Aleksei A. Chmura, et al. 2020. « Origin and Cross-Species Transmission of Bat Coronaviruses in China ». *BioRxiv*, mai, 2020.05.31.116061. <https://doi.org/10.1101/2020.05.31.116061>.
- Lau, Siu-Ying, Pui Wang, Bobo Wing-Yee Mok, Anna Jinxia Zhang, Hin Chu, Andrew Chak-Yiu Lee, Shaofeng Deng, et al. 2020. « Attenuated SARS-CoV-2 variants with deletions at the S1/S2 junction ». *Emerging Microbes & Infections* 9 (1): 837-42. <https://doi.org/10.1080/22221751.2020.1756700>.
- Letko, Michael, Andrea Marzi, et Vincent Munster. 2020. « Functional Assessment of Cell Entry and Receptor Usage for SARS-CoV-2 and Other Lineage B Betacoronaviruses ». *Nature Microbiology* 5 (4): 562-69. <https://doi.org/10.1038/s41564-020-0688-y>.
- Liu, Ping, Jing-Zhe Jiang, Xiu-Feng Wan, Yan Hua, Linmiao Li, Jiabin Zhou, Xiaohu Wang, et al. 2020. « Are Pangolins the Intermediate Host of the 2019 Novel Coronavirus (SARS-CoV-2)? » *PLoS Pathogens* 16 (5): e1008421. <https://doi.org/10.1371/journal.ppat.1008421>.
- Lu, Guangwen, Qihui Wang, et George F. Gao. 2015. « Bat-to-Human: Spike Features Determining 'Host Jump' of Coronaviruses SARS-CoV, MERS-CoV, and Beyond ». *Trends in Microbiology* 23 (8): 468-78. <https://doi.org/10.1016/j.tim.2015.06.003>.
- Lu, Roujian, Xiang Zhao, Juan Li, Peihua Niu, Bo Yang, Honglong Wu, Wenling Wang, et al. 2020. « Genomic Characterisation and Epidemiology of 2019 Novel Coronavirus: Implications for Virus Origins and Receptor Binding ». *Lancet (London, England)* 395 (10224): 565-74. [https://doi.org/10.1016/S0140-6736\(20\)30251-8](https://doi.org/10.1016/S0140-6736(20)30251-8).
- Luis, Angela D., David T. S. Hayman, Thomas J. O'Shea, Paul M. Cryan, Amy T. Gilbert, Juliet R. C. Pulliam, James N. Mills, et al. 2013. « A Comparison of Bats and Rodents as Reservoirs of Zoonotic Viruses: Are Bats Special? » *Proceedings. Biological Sciences* 280 (1756): 20122753. <https://doi.org/10.1098/rspb.2012.2753>.
- Luk, Hayes K. H., Xin Li, Joshua Fung, Susanna K. P. Lau, et Patrick C. Y. Woo. 2019. « Molecular Epidemiology, Evolution and Phylogeny of SARS Coronavirus ». *Infection, Genetics and Evolution: Journal of Molecular Epidemiology and Evolutionary Genetics in Infectious Diseases* 71: 21-30. <https://doi.org/10.1016/j.meegid.2019.03.001>.
- Matsuyama, Shutoku, Kazuya Shirato, Miyuki Kawase, Yutaka Terada, Kengo Kawachi, Shuetsu Fukushi, et Wataru Kamitani. 2018. « Middle East Respiratory Syndrome Coronavirus Spike Protein Is Not Activated Directly by Cellular Furin during Viral Entry into Target Cells ». *Journal of Virology* 92 (19). <https://doi.org/10.1128/JVI.00683-18>.
- Menachery, Vineet D., Kenneth H. Dinnon, Boyd L. Yount, Eileen T. McAnarney, Lisa E. Gralinski, Andrew Hale, Rachel L. Graham, et al. 2020. « Trypsin Treatment Unlocks Barrier for Zoonotic Bat Coronavirus Infection ». *Journal of Virology* 94 (5). <https://doi.org/10.1128/JVI.01774-19>.
- Menachery, Vineet D., Boyd L. Yount, Kari Debbink, Sudhakar Agnihothram, Lisa E. Gralinski, Jessica A. Plante, Rachel L. Graham, et al. 2015. « A SARS-like Cluster of Circulating Bat Coronaviruses Shows Potential for Human Emergence ». *Nature Medicine* 21 (12): 1508-13. <https://doi.org/10.1038/nm.3985>.
- Moulard, Maxime, et Etienne Decroly. 2000. « Maturation of HIV Envelope Glycoprotein Precursors by Cellular Endoproteases ». *Biochimica et Biophysica Acta (BBA) - Reviews on*

- Biomembranes* 1469 (3): 121-32. [https://doi.org/10.1016/S0304-4157\(00\)00014-9](https://doi.org/10.1016/S0304-4157(00)00014-9).
- Normile, Dennis. 2004. « Lab Accidents Prompt Calls for New Containment Program ». *Science* 304 (5675): 1223-25. <https://doi.org/10.1126/science.304.5675.1223a>.
- Ren, Wuze, Xiuxia Qu, Wendong Li, Zhenggang Han, Meng Yu, Peng Zhou, Shu-Yi Zhang, Lin-Fa Wang, Hongkui Deng, et Zhengli Shi. 2008. « Difference in Receptor Usage between Severe Acute Respiratory Syndrome (SARS) Coronavirus and SARS-Like Coronavirus of Bat Origin ». *Journal of Virology* 82 (4): 1899-1907. <https://doi.org/10.1128/JVI.01085-07>.
- Russell, Colin A., Judith M. Fonville, André E. X. Brown, David F. Burke, David L. Smith, Sarah L. James, Sander Herfst, 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): 1541-47. <https://doi.org/10.1126/science.1222526>.
- Sabir, Jamal S. M., Tommy T.-Y. Lam, Mohamed M. M. Ahmed, Lifeng Li, Yongyi Shen, Salah E. M. Abo-Aba, Muhammd I. Qureshi, 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>.
- Shang, Jian, Yushun Wan, Chuming Luo, Gang Ye, Qibin Geng, Ashley Auerbach, et Fang Li. 2020. « Cell Entry Mechanisms of SARS-CoV-2 ». *Proceedings of the National Academy of Sciences*, mai. <https://doi.org/10.1073/pnas.2003138117>.
- Song, Huai-Dong, Chang-Chun Tu, Guo-Wei Zhang, Sheng-Yue Wang, Kui Zheng, Lian-Cheng Lei, Qiu-Xia Chen, et al. 2005. « Cross-Host Evolution of Severe Acute Respiratory Syndrome Coronavirus in Palm Civet and Human ». *Proceedings of the National Academy of Sciences of the United States of America* 102 (7): 2430-35. <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, Xiangjie, Longping V. Tse, A. Damon Ferguson, et Gary R. Whittaker. 2010. « Modifications to the Hemagglutinin Cleavage Site Control the Virulence of a Neurotropic H1N1 Influenza Virus ». *Journal of Virology* 84 (17): 8683-90. <https://doi.org/10.1128/JVI.00797-10>.
- Thao, Tran Thi Nhu, Fabien Labrousseau, Nadine Ebert, Philip V'kovski, Hanspeter Stalder, Jasmine Portmann, Jenna Kelly, et al. 2020. « Rapid Reconstruction of SARS-CoV-2 Using a Synthetic Genomics Platform ». *Nature*, mai. <https://doi.org/10.1038/s41586-020-2294-9>.
- Van Boeckel, Thomas P., Michael J. Tildesley, Catherine Linard, José Halloy, Matt J. Keeling, et Marius Gilbert. 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, Alexandra C., Young-Jun Park, M. Alejandra Tortorici, Abigail Wall, Andrew T. McGuire, et David Veasler. 2020. « Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein ». *Cell* 181 (2): 281-292.e6. <https://doi.org/10.1016/j.cell.2020.02.058>.
- Wang, Qihui, Yanfang Zhang, Lili Wu, Sheng Niu, Chunli Song, Zengyuan Zhang, Guangwen Lu, et al. 2020. « Structural and Functional Basis of SARS-CoV-2 Entry by Using Human ACE2 ». *Cell*, avril. <https://doi.org/10.1016/j.cell.2020.03.045>.
- Wrapp, Daniel, Nianshuang Wang, Kizzmekia S. Corbett, Jory A. Goldsmith, Ching-Lin Hsieh, Olubukola Abiona, Barney S. Graham, et Jason S. McLellan. 2020. « Cryo-EM Structure of the 2019-NCoV Spike in the Prefusion Conformation ». *Science (New York, N.Y.)* 367 (6483): 1260-63. <https://doi.org/10.1126/science.abb2507>.

- Wu, Fan, Su Zhao, Bin Yu, Yan-Mei Chen, Wen Wang, Zhi-Gang Song, Yi Hu, et al. 2020. « A New Coronavirus Associated with Human Respiratory Disease in China ». *Nature* 579 (7798): 265-69. <https://doi.org/10.1038/s41586-020-2008-3>.
- Wu, Kailang, Guiqing Peng, Matthew Wilken, Robert J. Geraghty, et Fang Li. 2012. « Mechanisms of Host Receptor Adaptation by Severe Acute Respiratory Syndrome Coronavirus ». *Journal of Biological Chemistry* 287 (12): 8904-11. <https://doi.org/10.1074/jbc.m111.325803>.
- Xiao, Kangpeng, Junqiong Zhai, Yaoyu Feng, Niu Zhou, Xu Zhang, Jie-Jian Zou, Na Li, Yaqiong Guo, Xiaobing Li, Xuejuan Shen, Zhipeng Zhang, Fanfan Shu, Wanyi Huang, Yu Li, Ziding Zhang, Rui-Ai Chen, Ya-Jiang Wu, Shi-Ming Peng, Mian Huang, Wei-Jun Xie, Qin-Hui Cai, Fang-Hui Hou, Yahong Liu, et al. 2020. « Isolation and Characterization of 2019-NCoV-like Coronavirus from Malayan Pangolins ». Preprint. Microbiology. <https://doi.org/10.1101/2020.02.17.951335>.
- Xiao, Kangpeng, Junqiong Zhai, Yaoyu Feng, Niu Zhou, Xu Zhang, Jie-Jian Zou, Na Li, Yaqiong Guo, Xiaobing Li, Xuejuan Shen, Zhipeng Zhang, Fanfan Shu, Wanyi Huang, Yu Li, Ziding Zhang, Rui-Ai Chen, Ya-Jiang Wu, Shi-Ming Peng, Mian Huang, Wei-Jun Xie, Qin-Hui Cai, Fang-Hui Hou, Wu Chen, et al. 2020. « Isolation of SARS-CoV-2-Related Coronavirus from Malayan Pangolins ». *Nature*, mai, 1-7. <https://doi.org/10.1038/s41586-020-2313-x>.
- Yan, Renhong, Yuanyuan Zhang, Yaning Li, Lu Xia, Yingying Guo, et Qiang Zhou. 2020. « Structural Basis for the Recognition of SARS-CoV-2 by Full-Length Human ACE2 ». *Science* 367 (6485): 1444-48. <https://doi.org/10.1126/science.abb2762>.
- Zeng, Lei-Ping, Yu-Tao Gao, Xing-Yi Ge, Qian Zhang, Cheng Peng, Xing-Lou Yang, Bing Tan, 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 ». *Journal of Virology* 90 (14): 6573-82. <https://doi.org/10.1128/JVI.03079-15>.
- Zhang, Tao, Qunfu Wu, et Zhigang Zhang. 2020. « Probable Pangolin Origin of SARS-CoV-2 Associated with the COVID-19 Outbreak ». *Current Biology* 30 (7): 1346-1351.e2. <https://doi.org/10.1016/j.cub.2020.03.022>.
- Zhou, Hong, Xing Chen, Tao Hu, Juan Li, Hao Song, Yanran Liu, Peihan Wang, et al. 2020. « A Novel Bat Coronavirus Closely Related to SARS-CoV-2 Contains Natural Insertions at the S1/S2 Cleavage Site of the Spike Protein ». *Current Biology* 0 (0). <https://doi.org/10.1016/j.cub.2020.05.023>.
- Zhou, Peng, Xing-Lou Yang, Xian-Guang Wang, Ben Hu, Lei Zhang, Wei Zhang, Hao-Rui Si, et al. 2020. « A Pneumonia Outbreak Associated with a New Coronavirus of Probable Bat Origin ». *Nature* 579 (7798): 270-73. <https://doi.org/10.1038/s41586-020-2012-7>.
- Ziegler, Carly G. K., Samuel J. Allon, Sarah K. Nyquist, Ian M. Mbanjo, Vincent N. Miao, Constantine N. Tzouanas, Yuming Cao, 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*, avril. <https://doi.org/10.1016/j.cell.2020.04.035>.

Tables

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	Oct 1, 2010	
BtGX2013	Bat	China	2013	Jul 7, 2017	
BtHKU3-12	Bat	China (unspecified)	unspecified	Apr 5, 2010	China according to publication but origin not indicated in NCBI
BtRaTG13_2013_Yunnan	Bat	Yunnan	Jul 24, 2013	Mar 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	Dec 18, 2017	Shi's group in Wuhan (Hubei Province, China)
BtYN2013	Bat	China	2013	Jul 7, 2017	
BtYN2018B	Bat	China	Sep 1, 2016	Jun 30, 2019	
BtYu-RmYN02_2019	Bat	China, Yunnan - Xishuangbanna	Jun 25, 2019	Feb 3, 2020	Metagenome constructed by sequencing a mixture of 11 faecal samples from Rhinolophus malayanus bats
BtZC45	Bat	Zhoushan	2017		
BtZXC21	Bat	Zhoushan	2015	Feb 5, 2020	
Cv007-2004	Civet	China : Guangzhou in Guangdong Province	2019	Dec 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	Feb 11, 2020	Pandemic reference genome for COVID-19
HuSARS-Frankfurt-1_2003	Human	Frankfurt	2003	Mar 16, 2004	Reference genome for the 2003 SARS epidemic
PnGu-P2S_2019	Pangolin	China, Guangdong	2019	Feb 17, 2020	Sequence available in GISAID, very close to MP789. Pre-publication version ?
PnMP789	Pangolin	China : smuggled Malayan pangolins, Guangdong customs	Mar 29, 2019	Apr 23, 2020	Metagenome assembled from samples of 3 pangolins collected in March and July 2019.
PnGu1_2019	Pangolin	China, Guangdong	2019	Feb 18, 2020	
PnGX-P1E_2017	Pangolin	Chinese customs on a flight from Malaysia	2017	Apr 23, 2020	
PnGX-P2V_2018	Pangolin	Chinese customs on a flight from Malaysia	2018	Apr 23, 2020	Collected from pangolin, this strain has been cultured on human cells (and therefore presumably suitable for human infection).

Figure legends

Figure 1. Phylogeny and emergence of coronaviruses. (A) Tree inferred from complete coronavirus genomes, based on progressive 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 (BrYuRmYN02, 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 intermediate host is the civet. 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)** Pandemic COVID-19. 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).

Figure 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 (200bp 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)**.

Figure 3. Structure and function of the spike protein (S protein).**(A)** SARS-CoV-2 S protein specifically recognises 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 catalysed 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 SWISSMODEL on the SARS-CoV homolog (Protein Data Bank entry 6acc), and aligned on the structure of an RBD domain (orange) interacting with ACE2 (grey) 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).

Figure 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. Conservation de la protéine ACE2 et interactions avec la protéine S. (adapted from (Yan et al. 2020)).

Figure 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 4 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 indicate 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.

Figure 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 randomised 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. The alignments were performed on NCBI BLAST server (<https://blast.ncbi.nlm.nih.gov/Blast.cgi>).

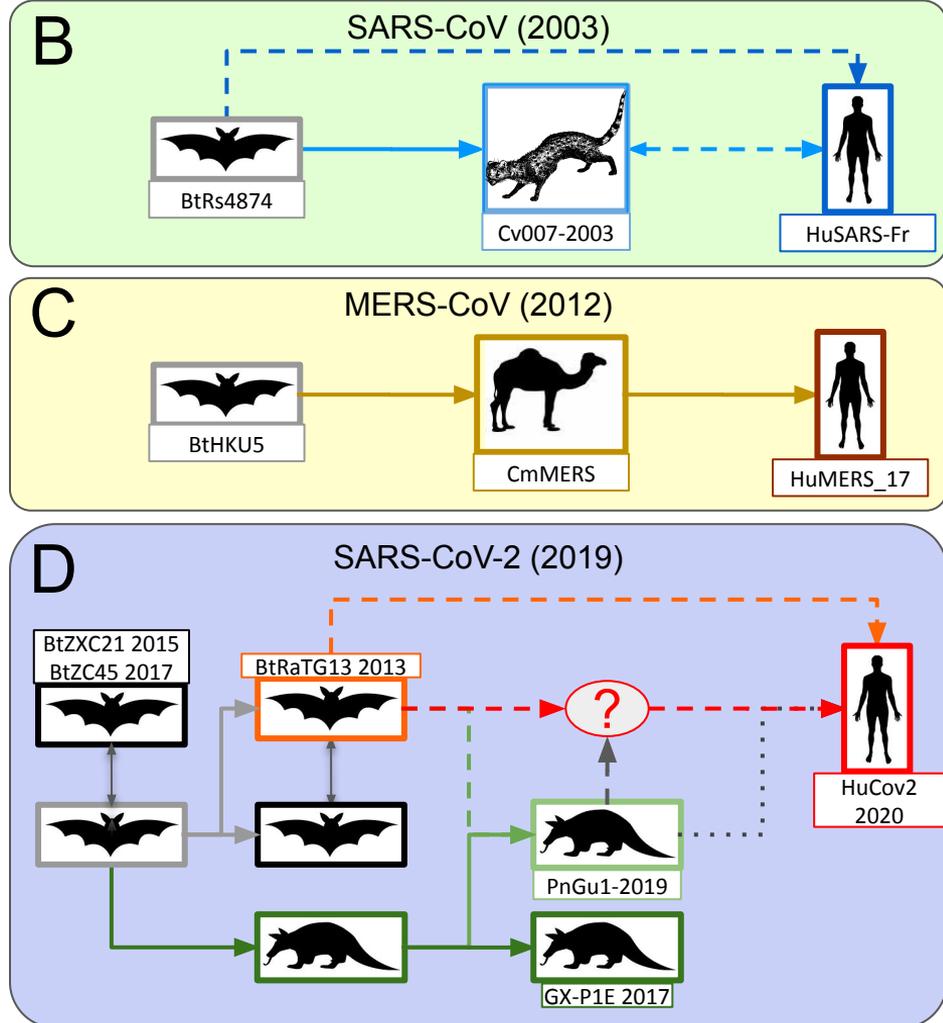
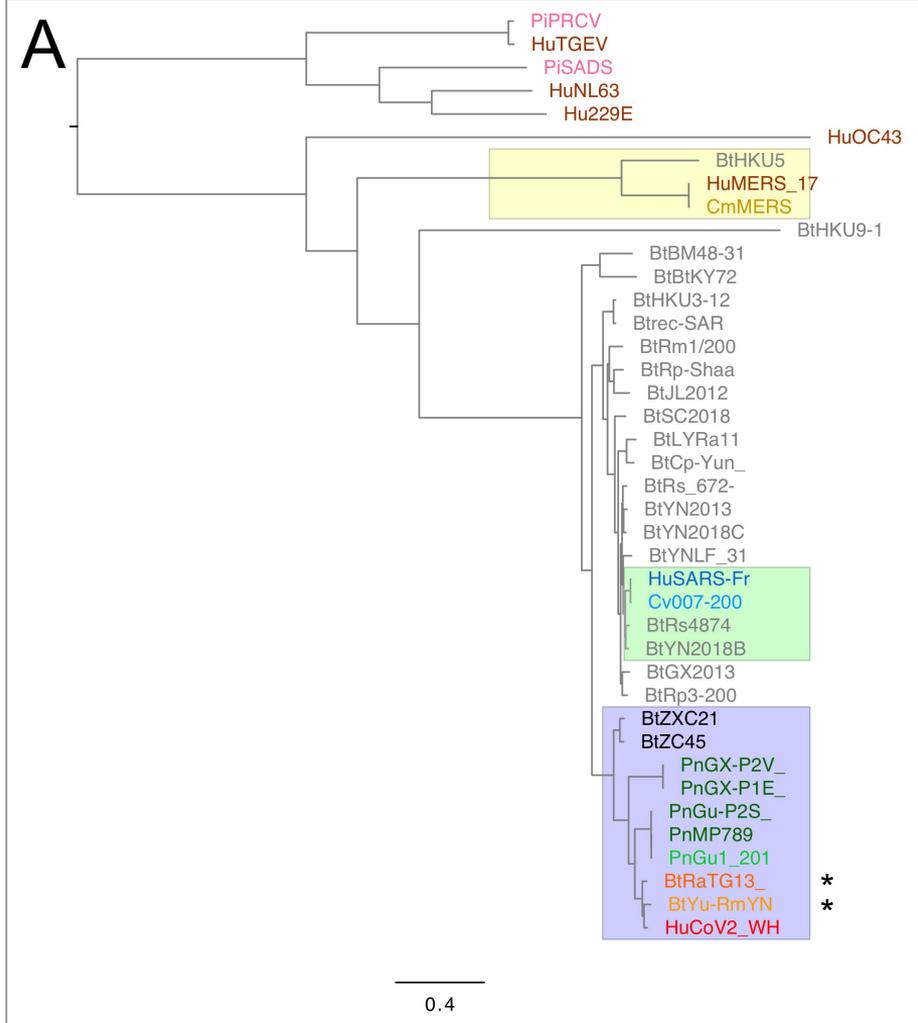


Figure 1

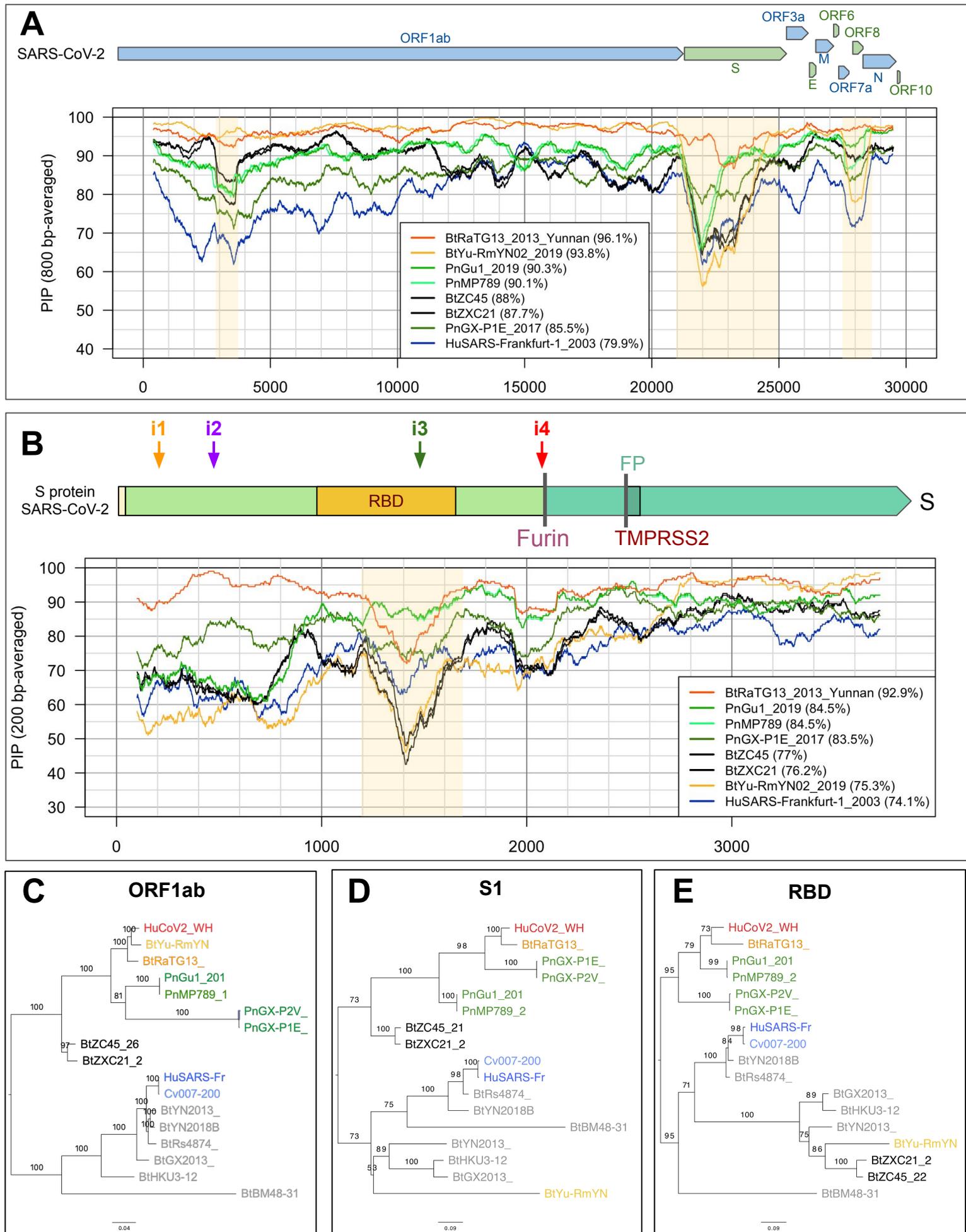


Figure 2

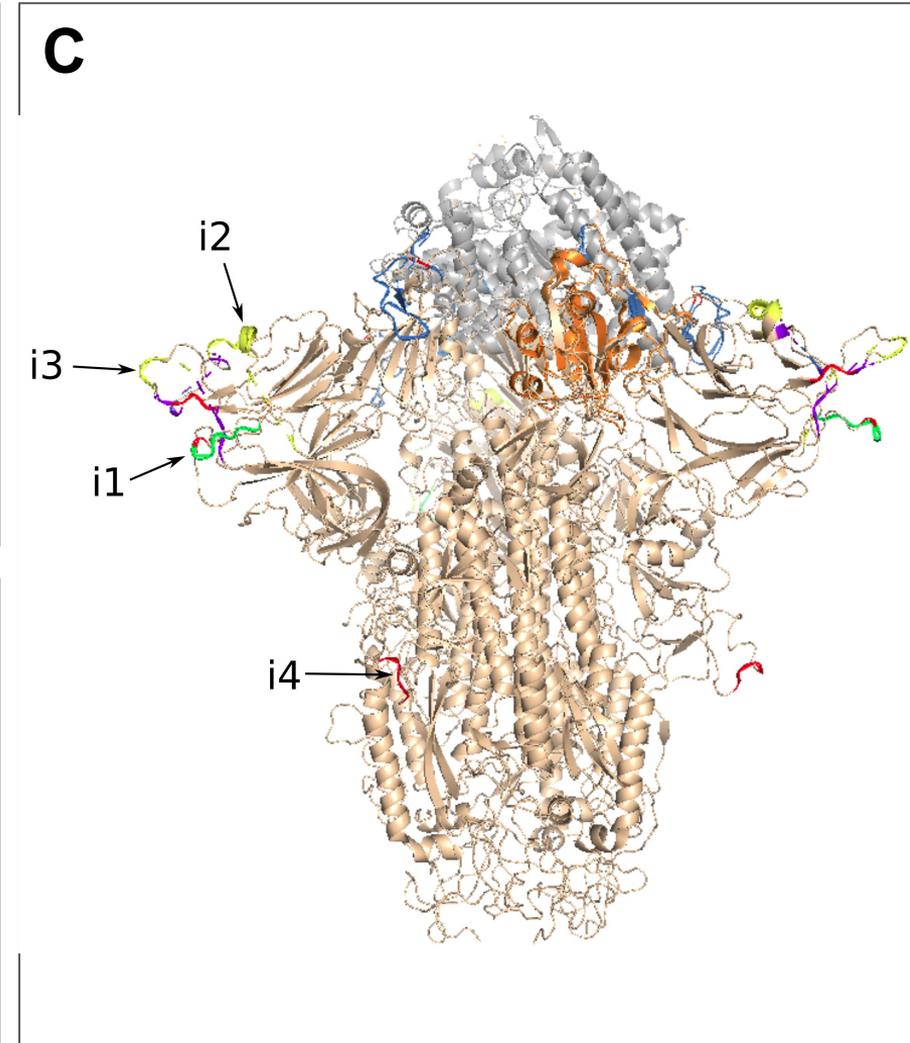
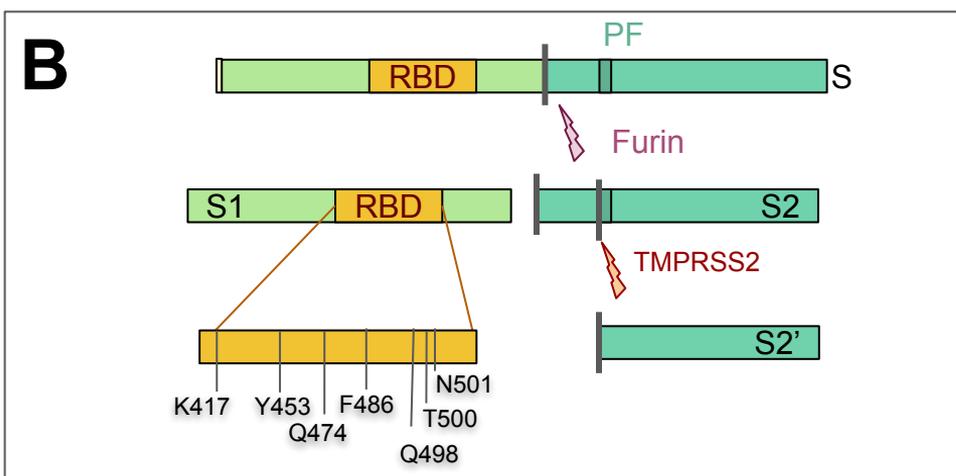
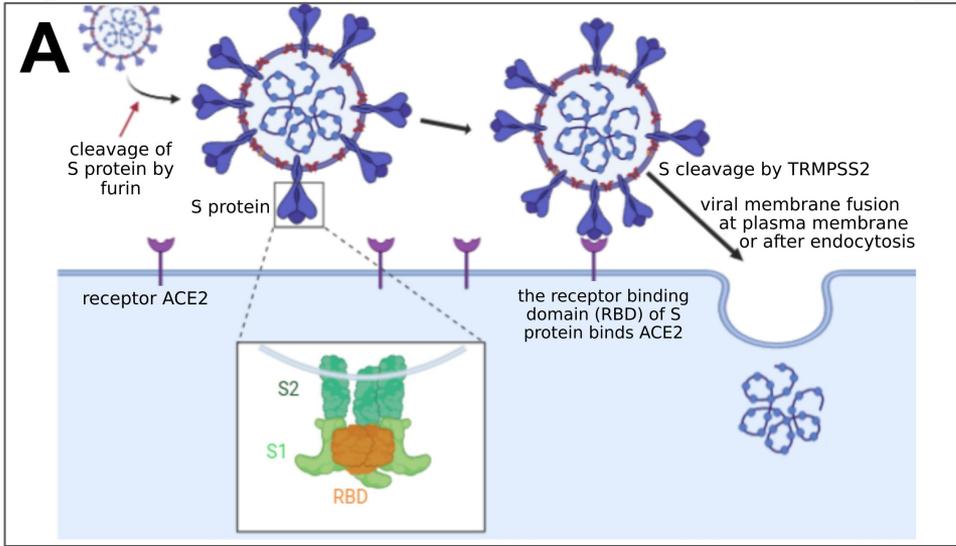
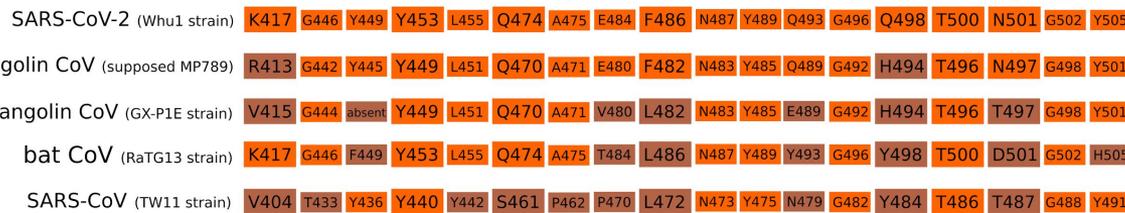


Figure 3

A

ACE2 receptor



S protein

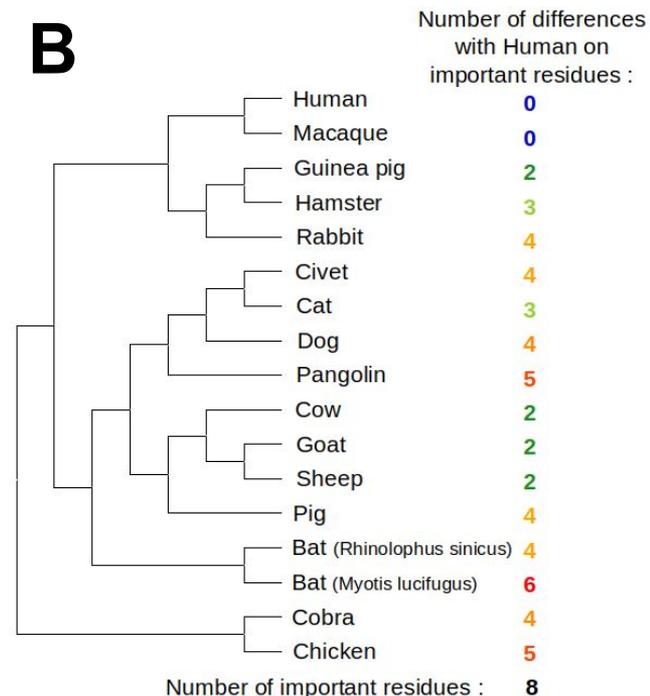
B

Figure 4

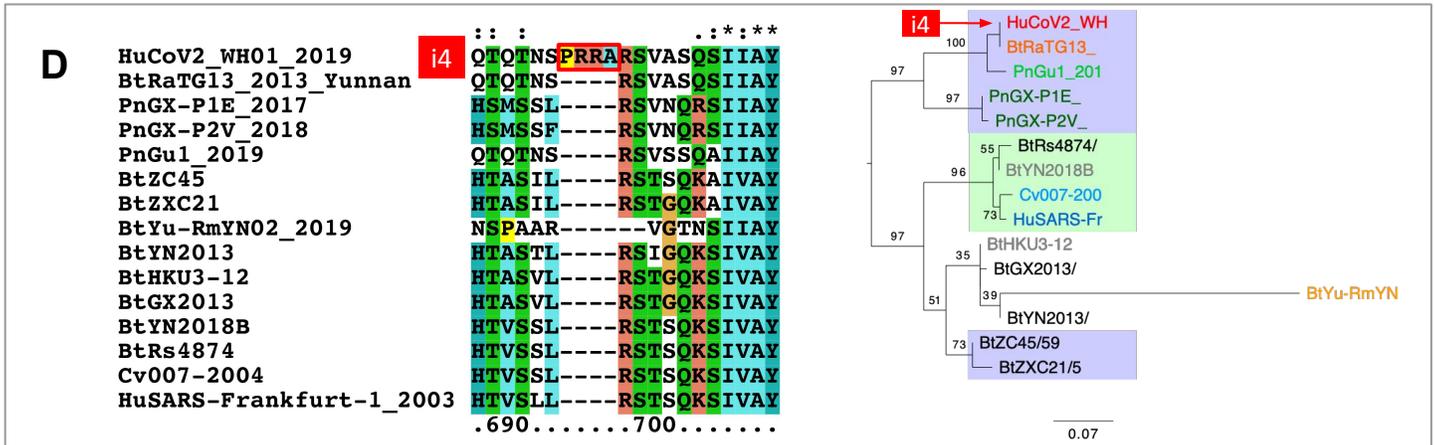
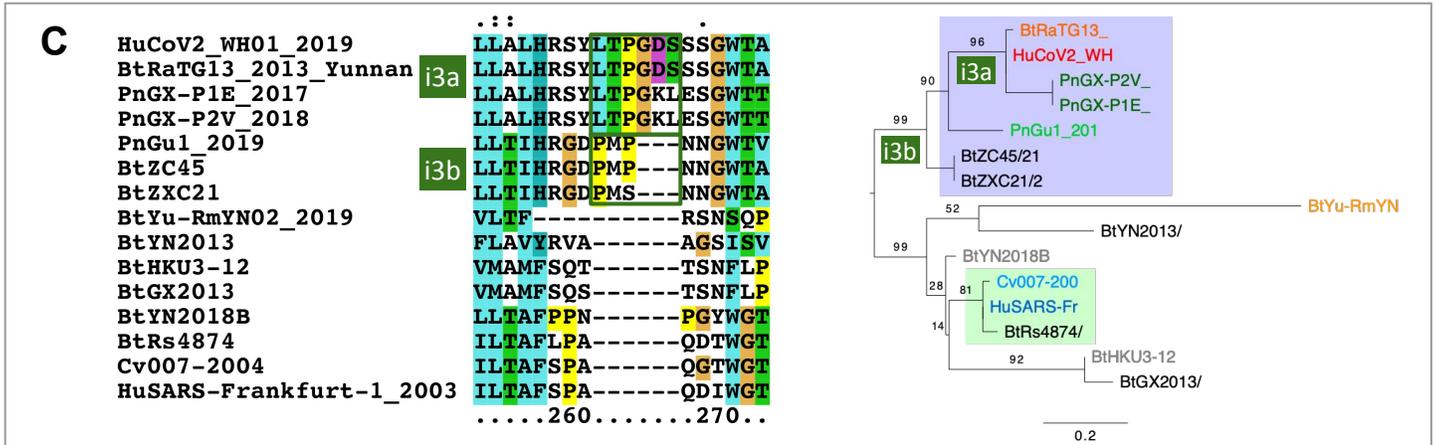
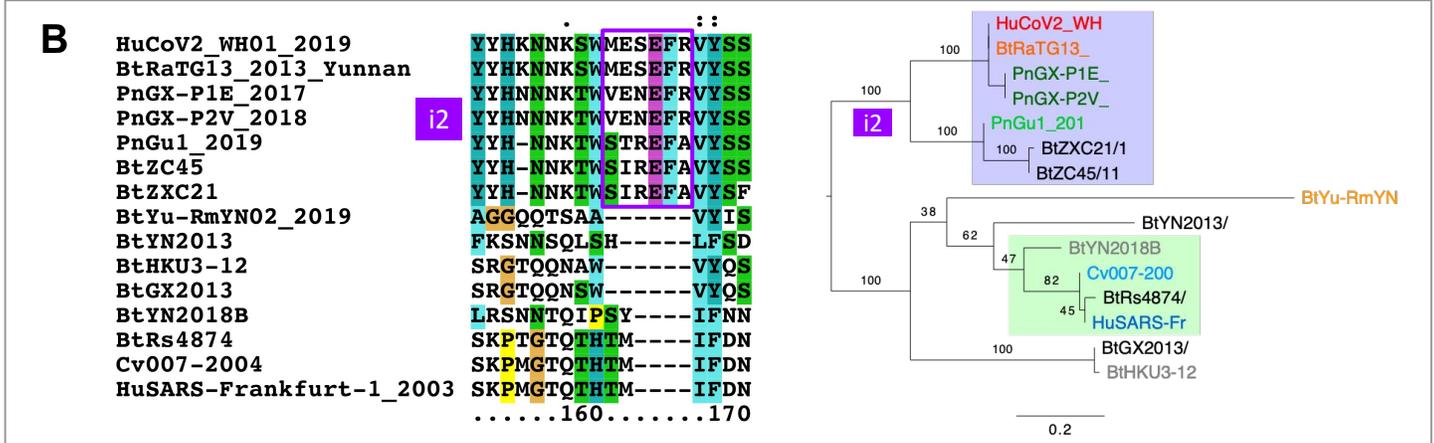
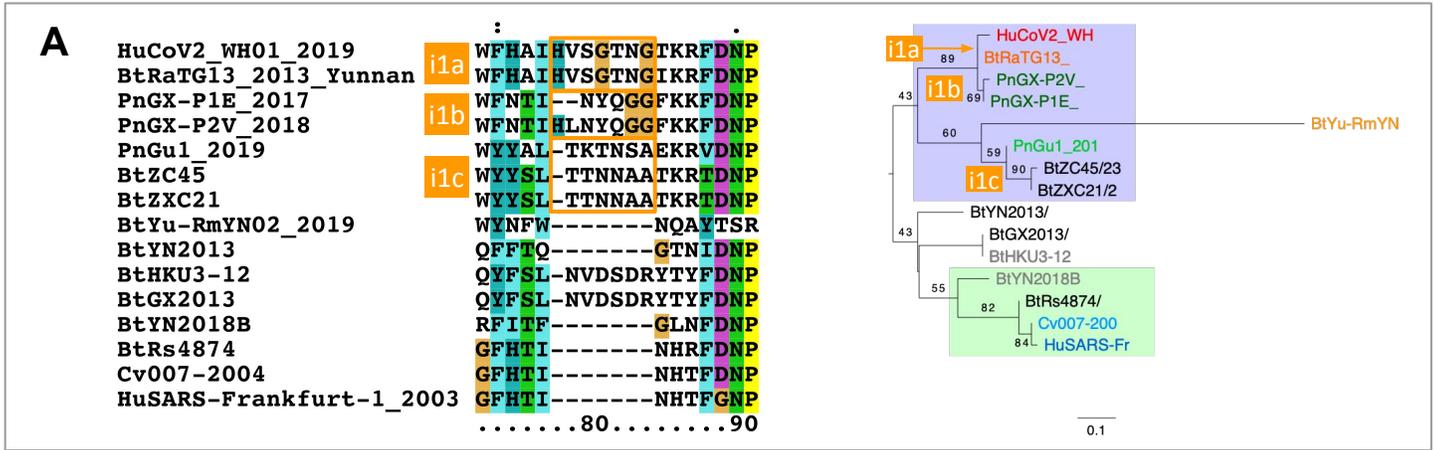
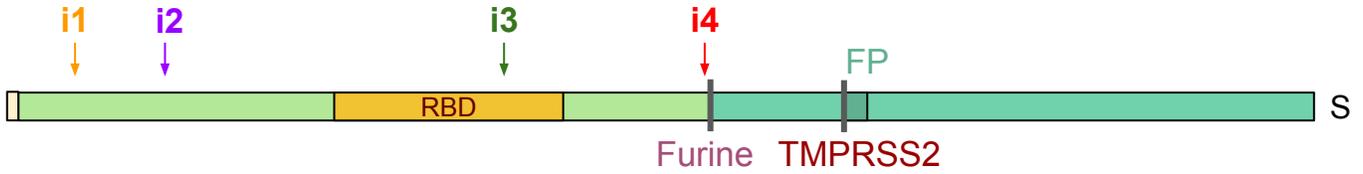


Figure 5

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	CCTAAAAGTTCCTTTGTAATAACTGTATTATT	381	
Sbjct	2523	CCTAAAAGTTCCTTTGTAATATTTCTATAATT	2493	

Figure 6