# Retrouver les origines du SARS-CoV-2 dans les phylogénies de coronavirus

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

# Retrouver les origines du SARS-COV-2 dans les phylogénies de coronavirus

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

1) École Normale Supérieure de Paris, 45 rue d'Ulm, 75005 Paris, France. ORCID: 0000-0002-2324-3633
2) Université de Paris, CNRS, LIED UMR 8236, 85 bd Saint-Germain, 75006 Paris, France.ORCID: 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
4) Université de Paris, UFR Sciences du Vivant, F-75013 Paris, France.
5) CNRS, Institut Français de Bioinformatique, IFB-core, UMS 3601, Evry, France.
6) Aix-Marseille Univ, Inserm, laboratoire *Theory and approaches of genome complexity* (TAGC), Marseille, France. ORCID: 0000-0002-8799-8584
7) AFMB, CNRS, Aix-Marseille Univ, UMR 7257, Case 925, 163 Avenue de Luminy, 13288 Marseille Cedex 09, France. ORCID: 0000-0002-6046-024X

* Les deux derniers auteurs ont contribué de façon égale à cet article

Jacques.van-Helden@univ-amu.fr

etienne.decroly@afmb.univ-mrs.fr

## Résumé

Le SARS-CoV-2 est un nouveau coronavirus (CoV) humain. Il a émergé en Chine fin 2019 et est responsable de la pandémie mondiale de Covid-19 qui a causé plus de 540 000 décès en six mois. La compréhension de l'origine de ce virus est une question importante et il est nécessaire de déterminer les mécanismes de sa dissémination afin de pouvoir se prémunir de nouvelles épidémies. En nous fondant sur des inférences phylogénétiques, l'analyse des séquences et les relations structure-fonction des protéines de coronavirus, éclairées par les connaissances actuellement disponibles, nous discutons les différents scénarios évoqués pour rendre compte de l'origine – naturelle ou synthétique – du virus.

## La pandémie de Covid-19

Le SARS-CoV-2 est le troisième coronavirus humain (CoV) responsable d'un syndrome respiratoire sévère à avoir émergé au cours des 20 dernières années, les deux précédents étant le SARS-CoV en 2002 [1] et le MERS-CoV en 2012 [2]. Le SARS-CoV-2, qui provoque chez l'homme la maladie Covid-19, s'est propagé en pandémie début 2020. Fin juin , plus de 10 millions d'infections étaient recensées avec au moins 500 000 morts. L'agent étiologique

de la Covid-19 a rapidement été identifié et dès le 26 janvier 2020, 10 génomes viraux ont été séquencés [3]. La comparaison de leur séquences donne un taux d'identité de 99,98 % entre paires de séquences génomiques, ce qui est caractéristique d'une émergence récente.

Au moment du séquençage des premiers isolats de SARS-CoV-2, les coronavirus les plus proches disponibles dans les bases de données étaient les souches bat-SL-CoVZXC21 et bat-SL-CoVZC45, isolées en 2015 et 2017 à partir de chauves-souris de la région de Zhoushan, en Chine de l'Est, et dont les génomes présentent 88 % d'identité avec le SARS-CoV-2 [3]. La séquence du génome du SARS-CoV-2 est plus distante de celles du SARS-CoV (79 % d'identité) et du MERS-CoV (50 % d'identité), responsables des épidémies humaines précédentes. Il fut alors conclu que le SARS-CoV-2 était un nouvel agent infectieux à transmission interhumaine appartenant à la famille des SARS-CoV, dont le réservoir animal était la chauve-souris.

## Origine évolutive du nouveau virus

L'origine zoonotique (issu d'un hôte animal avec transmission à l'homme) des CoV est largement documentée. Cette famille de virus infecte plus de 500 espèces de chiroptères (ordre de mammifères constitué de plus de 1 200 espèces de chauves-souris) qui représentent un réservoir important pour son évolution en permettant, entre autres, la recombinaison des génomes, chez des animaux infectés simultanément par différentes souches virales [4–6]. Il est admis que la transmission zoonotique des CoV à l'homme passe par une espèce hôte intermédiaire, dans laquelle des virus mieux adaptés aux récepteurs humains peuvent être sélectionnés, favorisant ainsi le franchissement de la barrière d'espèce [7]. Les vecteurs de la transmission zoonotique peuvent être identifiés en examinant les relations phylogénétiques entre les nouveaux virus et ceux isolés à partir de virus d'espèces animales vivant dans les régions d'émergence.

La ***figure 1A***, qui présente l'arbre phylogénétique produit à partir des alignements des génomes complets de différents CoV, montre la grande proximité (99 % d'identité des génomes) entre les coronavirus responsables des deux épidémies précédentes (SARS-CoV et MERS-CoV) et les souches isolées à partir des derniers hôtes intermédiaires avant l'homme : la civette pour le SARS-CoV de 2003 **(*Figure 1B*)**[8], et le dromadaire pour le MERS-CoV **(*Figure 1C*)** [9], pour lequel plusieurs transmissions zoonotiques ont été démontrées.

Bien qu'aucune épidémie liée à la transmission directe de la chauve-souris à l'homme n'ait été mise en évidence à ce jour, des études expérimentales ont démontré que plus de 60 CoV de chiroptères sont capables d'infecter les cellules humaines en culture *in vitro* [4,10]. L'identification, en 2017, d'isolats viraux très similaires au SARS-CoV chez les chauves-souris pose la question de la possibilité d'une transmission directe des chiroptères à l'homme, qui

pourrait résulter d'une évolution du domaine de liaison du virus au récepteur permettant son entrée dans la cellule [5].

## Le SARS-CoV-2 : du Yunnan à Wuhan ?

L'origine du SARS-CoV-2 fait l'objet de controverses. Les études bioinformatiques ont révélé qu'il possède une identité de 96,2 % avec un génome de CoV (souche RaTG13) qui a été reconstruit à partir d'échantillons de fèces et de prélèvements anaux d'une chauves-souris (*Rhinolophus affinis)* effectués en 2013, mais dont la séquence a été publiée début février 2020 [11]. Malheureusement, le lieu précis de récolte de ces échantillons n'est documenté ni dans l'article, ni dans les bases de données de séquences. Nous avons constaté que cette séquence correspond exactement à un fragment de 370 nucléotides (KP876546, seule partie de ce génome publiée en 2016), qui code pour un domaine de la polymérase BtCoV/4991, séquencé à partir d'isolats collectés dans un puits de mine de la province du Yunnan suite au décès de 3 mineurs d'une pneumonie atypique [12].

Plus récemment, un métagénome (RmYN02) a été assemblé à partir d'échantillons issus des fèces de 11 chauves-souris de l'espèce *Rhinolophus malayanus*, collectés en 2019 dans la province de Yunnan. La séquence de RmYN02 présente 97,2 % d'identité avec les deux premiers tiers du génome de SARS-CoV-2 (correspondant à l'ORF 1ab). Il diverge néanmoins assez fortement sur le tiers restant, en particulier au niveau de la séquence codant pour la protéine S1 et de l'ORF 8 **(*Figures 1A, 2A, 2B*)** [13].

## Une histoire évolutive par fragments

Les CoV ont un génome d'environ 30 000 nucléotides, ce qui est exceptionnellement long pour un virus à ARN (le virus du sida comporte 10 000 nucléotides, et celui d'Ebola 19 000). Un génome d'une telle taille est possible car les CoV disposent d'un système de correction d'erreurs de réplication unique dans le monde des virus à ARN, assuré par une exonucléase virale qui limite le taux de mutations [14,15]. Les deux premiers tiers du génome correspondent à un gène unique, ORF1ab, qui code pour un précurseur polyprotéique ensuite clivé en 16 protéines formant le complexe de réplication/transcription. Le dernier tiers contient 9 gènes codant pour des protéines produites à partir d'ARN subgénomiques synthétisés par la polymérase virale **(*Figure 2A*)**.

Cette polymérase est capable de réaliser des sauts de brins lors de la synthèse des ARN, une propriété qui joue probablement un rôle clé dans la capacité de recombinaison des CoV et favorise leur évolution et le changement d'hôte. Les recombinaisons génomiques sont fréquentes chez les CoV de chiroptères [5]. On suppose qu'elles ont joué une rôle dans l'origine du SARS-CoV, responsable de l'épidémie de SRAS en 2002 [16], dont le génome est

une mosaïque constituée de morceaux provenant d'au moins deux CoV différents de chauve-souris.

Les régions ayant fait l'objet de recombinaisons peuvent être détectées en comparant les profils de pourcentage de positions identiques (PPI) dans les séquences de différents génomes **(*Figure 2 A-B*)**. Elles se manifestent par le croisement des profils de différentes souches. En comparant le SARS-CoV-2 et d'autres virus génétiquement proches, on observe ainsi des recombinaisons dans les régions 2 900 à 3 800, 21 000 à 24 000, 27 500 à 28 500 (marquées par un fond jaune sur la ***Figure 2A***). Ce mosaïcisme dû aux recombinaisons biaise la recherche de l'origine d'un virus par une analyse phylogénétique reposant sur des génomes complets. L'arbre phylogénétique obtenu dans ce cas est en effet une combinaison d'histoires évolutives différentes suivies par les fragments recombinés. On est donc amené à réaliser des inférences phylogénétiques différentes pour chacune des régions recombinées. Cette approche est illustrée par les ***Figures 2C à 2E***, dans lesquelles sont montrés les arbres évolutifs réalisés à partir de l'ORF1ab **(*Figure 2C*)**, de la région codant pour la sous-unité S1 de la protéine S ***(Figure 2D)***, et du domaine de liaison de cette sous-unité au récepteur de la cellule de l'hôte (RBD, pour "Receptor Binding Domain") **(*Figure 2E*)**. Il existe plusieurs différences frappantes entre ces arbres, en particulier pour le métagénome BtYu-RmYN02, qui occupe des positions différentes selon la région génomique considérée.

## Des chauves-souris et des hommes... et quelques pangolins ?

Si l'on réalise un agrandissement centré sur le gène S **(*Figure 2B*)**, on constate une diminution du pourcentage d'identité entre la séquence nucléotidique de la souche de chauve-souris RaTG13 et celle du SARS-Cov-2, plus spécifiquement dans la région codant pour le domaine de liaison au récepteur cellulaire (RBD). En particulier, aux positions 1200 à 1 600 de ce gène, le pourcentage de positions identiques entre la souche RaTG13 et le SARS-Cov-2 tombe à 70 %, alors qu'il est supérieur à 96 % pour le reste du génome. Dans la même région, la séquence la plus proche de celle du SARS-CoV-2 est celle d'un autre métagénome (MP789), obtenu par assemblage d'échantillons de pangolins [17,18]. Le taux d'identité de MP789 est assez faible en amont du RBD (60 %), mais il dépasse 90 % en aval. Cette observation a conduit à l'hypothèse, émise par certains, selon laquelle le SARS-CoV-2 pourrait résulter de recombinaisons entre des souches de virus infectant respectivement les chauve-souris et les pangolins **(*Figure 1D*)**. Notons cependant que même pour le pangolin, le taux d'identité des nucléotides dans la région RBD atteint à peine 89 %, ce qui est bien inférieur aux taux observés entre les souches de virus isolées de l'homme et celles infectant les derniers intermédiaires animaux lors des précédentes transmissions zoonotiques : le

taux d'identité entre le génome du SARS-CoV humain et celui de la souche de civette la plus proche était en effet de 99,52 %.

L'hypothèse couramment admise est que le virus SARS-CoV-2 résulterait de recombinaisons multiples entre différents CoV circulant dans la faune sauvage, ce qui conduirait à une adaptation ayant augmenté la capacité de transmission inter-humaine du virus. La recombinaison se serait produite entre un virus de pangolin et un virus des chauves-souris. La transmission à l'homme proviendrait secondairement du contact avec l'hôte intermédiaire, éventuellement vendu sur le marché de Wuhan [19,20]. Cette hypothèse soulève cependant de nombreuses questions. En effet, les premiers patients infectés ne fréquentaient pas tous le marché de Wuhan [21]. De plus, en dépit des recherches de virus dans les espèces animales vendues sur ce marché, aucun virus intermédiaire, qui résulterait de la recombinaison supposée entre un virus de pangolin et un virus de chauve-souris, n'a pu être identifié à ce jour.

Tant que ce dernier recombinant hypothétique n'aura pas été identifié et son génome séquencé, des questions resteront en suspens : chez quelle espèce cette recombinaison a-t-elle eu lieu ? Une chauve-souris, un pangolin, une autre espèce ? Et surtout, dans quelles conditions ? Il est aussi envisageable que la recombinaison ait eu lieu chez des animaux d'élevage, ou de laboratoire, plutôt que chez le pangolin ou la chauve-souris sauvages. Dans le premier cas, la transmission à l'homme serait favorisée par des contacts plus étroits et fréquents, et par une plus grande similarité du récepteur ACE2 humain que celui du pangolin au niveau des résidus importants pour la fixation du SARS-CoV-2 **(*Figure 4A*)**. Une autre hypothèse est que la ressemblance entre les séquences de RBD du virus de pangolin et celle du SARS-CoV-2 résulte d'une évolution convergente.

En l'absence d'éléments probants concernant le dernier intermédiaire animal avant la contamination humaine, certains auteurs suggèrent que ce virus pourrait avoir été fabriqué dans un laboratoire (origine synthétique). Mais ces assertions ont été réfutées par de nombreux spécialistes, notamment sur la base d'études phylogénétiques qui suggèrent deux scénarios prépondérants pour expliquer l'origine du SARS-CoV-2 : (1) l'adaptation chez un animal hôte avant le transfert zoonotique, ou (2) l'adaptation chez l'homme après le transfert zoonotique [11, 17 18 20, 22]. D'autres pensent qu'il pourrait s'agir d'un virus de chiroptère qui se serait adapté à d'autres espèces dans des modèles animaux élevés en laboratoire, dont il se serait ensuite échappé. Il est également envisageable que ce virus provienne d'une souche virale cultivée sur des cellules au laboratoire afin d'étudier son potentiel infectieux. Ce virus cultivé se serait progressivement "humanisé" (adapté à l'humain) par sélection des virus les plus aptes à se propager dans ces conditions.

Quel que soit le mécanisme présidant à son apparition, il est important de comprendre comment ce virus a passé la barrière d'espèce et est devenu hautement transmissible d'homme à homme, cela afin de se prémunir de nouvelles émergences [23].

## La protéine S, une actrice majeure de l'évolution des CoV et du franchissement de la barrière d'espèce

Le gène S code pour la protéine Spike (ou spicule en français), qui est localisée au niveau de l'enveloppe virale et forme à la surface du virus des protubérances caractéristiques évoquant une couronne, d'où le nom de coronavirus **(*Figure 3*)**. La protéine Spike joue un rôle déterminant dans l'initiation du cycle viral. Elle participe à la reconnaissance par le virus des récepteurs exprimés par les cellules de l'hôte, ACE2 (enzymes de conversion de l'angiotensine 2), ce qui permet, ensuite, son entrée **(*Figure 3A*)**. Ce récepteur, présent chez les différentes espèces infectées, est localisé à la membrane plasmique de différents types de cellules, notamment les cellules alvéolaires du poumon, les entérocytes de l'intestin grêle, les cellules endothéliales artérielles et veineuses, et les cellules des muscles lisses artériels de la plupart des organes. L'ARN messager d'ACE2 est également détecté dans le cortex cérébral, le striatum, l'hypothalamus et le tronc cérébral. L'expression d'ACE2 est par ailleurs augmentée en réponse aux interférons, des cytokines produites lors des infections virales, ce qui favorise la propagation systémique du virus [24].

La protéine S est synthétisée sous la forme d'un précurseur inactif. Deux clivages protéolytiques successifs sont nécessaires pour assurer sa fonction biologique **(*Figure 3B*)**. Le premier clivage appelé *priming* génère deux sous-unités, S1 et S2. Le second libère l'extrémité d'un peptide, dit de fusion, localisé au début de la sous-unité S2. Ces clivages protéolytiques, catalysés respectivement par la furine et la protéine TMPRSS2 (*transmembrane serine protease 2*) [25], permettent la fusion entre les membranes virale et cellulaire. Ils sont donc indispensables à l'entrée et à la réplication virale à l'origine de la formation des nouveaux virions.

La protéine S1 de SARS-CoV et SARS-CoV-2, produit du clivage de S, contient le domaine RBD **(*Figure 2B, 3B et 5*)** qui assure la reconnaissance du récepteur cellulaire ACE2 par le virus, [17,26–28]. Elle contient également des sites antigéniques qui sont exposés à la surface du virus et accessibles au système immunitaire, constituant des antigènes pouvant être reconnus par les anticorps produits par les hôtes infectés [29]. Cependant, la séquence génomique codant pour ces sites potentiellement antigéniques présente une grande variabilité entre les espèces virales **(*Figure 3C*)** qui résulte de la sélection de mutations génétiques permettant aux virus d'échapper à la réponse immunitaire de l'hôte.

Les résidus du RBD impliqués directement dans la reconnaissance d'ACE2 subissent, eux aussi, des contraintes évolutives fortes **(*Figure 4*)**. Certains de ces résidus sont requis pour une infection efficace chez les chiroptères, ou chez l'hôte intermédiaire, ou encore chez

l'homme [28,30,31] et c'est grâce à des mutations de ces domaines RBD que le virus acquiert sa propension épidémique.

L'analyse phylogénétique de la protéine S des CoV est donc particulièrement instructive pour comprendre l'évolution des CoV et leur capacité à franchir la barrière d'espèce.

Dans ce contexte, l'identification de séquences de nouveaux coronavirus proches du SARS-CoV-2, isolés à partir de pangolins de Malaisie, a constitué une avancée importante. En effet, bien que, globalement, les séquences des ARN de ce virus n'aient qu'un taux d'identité modéré avec le génome complet du SARS-CoV-2 (89 % pour la souche MP789, à comparer aux 96 % pour RaTG13) **(*Figure 2A*)**, le taux d'identité des acides aminés constituant le domaine RBD des deux virus s'élève à 98 % [18]. Or, le taux d'identité nucléotidique entre les séquences codant pour le RBD du MP789 et du SARS-CoV-2 n'est que de 89 %. Cette différence d'identité entre génome et protéine s'explique par le fait que presque toutes les mutations touchant cette région sont synonymes (mutation d'un codon en un autre codant pour le même acide aminé). Ceci suggère une forte pression sélective, vraisemblablement liée à la fonction importante du RBD lors de l'infection. Certains CoV, qui infectent les pangolins, possèdent donc un domaine RBD de séquence très proche de celle de SARS-CoV-2. Leur protéine S pourrait donc avoir une forte affinité pour le récepteur ACE2 humain et ainsi permettre au virus de pangolin d'infecter plus efficacement les cellules humaines que le virus de chauve-souris **(*Figure 4B*)**.

En fonction des séquences actuellement disponibles, les analyses fondées sur la phylogénie des génomes complets de virus ne permettent pas de conclure définitivement quant à l'origine évolutive du SARS-CoV-2. Cette difficulté entraîne certaines spéculations se référant à une possible origine synthétique du virus. Une hypothèse est que le SARS-CoV-2 pourrait être une reconstruction à partir des séquences métagénomiques obtenues à partir d'échantillons fécaux de chauves-souris. Des inquiétudes ont été formulées, considérant des travaux portant sur des manipulations génétiques de virus qui avaient pour but de comprendre les mécanismes viraux leur permettant le franchissement de la barrière d'espèce.

## Manipulations génétiques des virus et gain de fonction

La question de l'origine naturelle ou synthétique du SARS-CoV-2 mérite d'être examinée en nous fondant sur les éléments tangibles qui sont à notre disposition. Les hypothèses proposées nécessitent de prendre en compte le contexte des manipulations génétiques qu'il est actuellement possibles de réaliser dans les laboratoires de virologie. La manipulation du génome de virus potentiellement pathogènes est une pratique courante. Elle vise, entre autres, à comprendre les mécanismes de réplication et d'émergence de ces virus et à développer de nouvelles stratégies antivirales ou vaccinales. Les risques de franchissement

inopiné de la barrière d'espèce, de contamination d'un nouvel hôte (en particulier l'homme) et de dissémination accidentelle de virus recombinants artificiels sont une problématique prise en compte lors de ces manipulations. Ceci implique de mener ces approches dans des laboratoires de haute sécurité (BSL3 ou BSL4[1]) soumis à de strictes procédures de contrôle.

La polémique sur les expériences de gain de fonction (augmentation de la virulence ou de l'infectiosité du virus par manipulation génétique) a été initiée en 2011 par les travaux des équipes de Ron Fouchier [32] et de Yoshihiro Kawaoka [33] sur le virus de la grippe. Afin de comprendre les facteurs de virulence du virus, ces chercheurs avaient testé l'effet de mutations susceptibles d'accroître la transmissibilité du virus H5N1 dans différents modèles animaux. Le *National Science Advisory Board for Biosecurity* (NSABB) du ministère de la Santé des États-Unis, alerté de ces expériences en décembre 2011, demanda aux revues *Nature* et *Science* de ne pas divulguer les résultats de ces travaux, au nom des risques qu'ils feraient courir aux populations de façon intentionnelle (bioterrorisme) ou non (sortie accidentelle du laboratoire). En raison de l'importance des résultats pour la santé publique et les communautés de recherche, le NSABB a finalement recommandé que les conclusions générales résultant de ces expériences soient publiées, mais que les manuscrits n'incluent pas « les détails méthodologiques et autres qui pourraient permettre la reproduction des expériences par ceux qui chercheraient à faire du mal » [34].

Ces risques d'échappement accidentel proviennent, Notamment, de l'accroissement du nombre de laboratoires de haute sécurité biologique (BSL3 et BSL4) qui sont principalement implantés dans des zones densément peuplées [35]. Les expériences utilisant des agents pathogènes qui n'infectent pas initialement l'homme, tels que les virus de grippes aviaires ou les SRAS de chiroptères, sont par ailleurs autorisées dans les laboratoires de type BSL3 (un niveau de sécurité biologique moindre que les laboratoires BSL4). Cette possibilité peut donc accroître les risques d'accidents [36–38], l'introduction de mutations ou leur sélection pouvant conférer à ces virus un potentiel épidémique.

Avant 2002, et bien qu'à l'origine d'épidémies importantes chez les animaux de rente, les CoV étaient considérés comme des virus de faible intérêt en santé publique : ils n'étaient principalement responsables que de pathologies bénignes, comme les rhumes saisonniers. Depuis l'émergence du SARS-CoV, en 2002-2003, des études ont testé la possibilité de transfert zoonotique des virus de chauves-souris (Bat-SCoV) chez l'homme et ont tenté d'élucider les processus conduisant à l'émergence de nouveaux pathogènes [39]. À partir du génome de ces virus Bat-SCoV, des virus recombinants potentiellement adaptés à l'espèce humaine ont été construits dans des laboratoires américains et chinois, notamment en remplaçant le RBD de chauve-souris par celui du SARS-CoV humain [40,41]. Ces expériences ont révélé que l'infection des cellules humaines reste cependant souvent limitée car

[1] Niveaux de sécurité des laboratoires, le niveau 4 étant le plus extrême.

l'activation de la protéine S nécessite une protéolyse par des protéases exprimées par les cellules de l'hôte, qui est incomplètement réalisée par les cellules humaines pour des virus animaux **(*Figure 3A*)**. Cette difficulté peut néanmoins être contournée en laboratoire en traitant les virus par la trypsine (une protéase) [40], ou en ajoutant, en aval du domaine RBD du génome viral, un site de protéolyse par la furine [42,43]. Ces manipulations ont été réalisées et on en retiendra que, d'une part, il est possible d'adapter les virus de chauves-souris pour infecter des cellules humaines ou différents modèles animaux, et que, d'autre part, les CoV de chiroptères ont un potentiel de transmission zoonotique directe vers l'homme, notamment s'ils acquièrent un site de protéolyse adapté, ce qui ne nécessite que quelques mutations ou l'insertion d'une courte séquence d' acides aminés basiques.

Un facteur aggravant le risque lié aux manipulations génétiques qui produisent des gains de fonction doit aujourd'hui être pris en considération : les progrès spectaculaires des méthodes de biologie synthétique et de génétique inverse réalisés ces 20 dernières années permettent en effet d'assembler, en une dizaine de jours, un génome viral à partir de différents fragments d'ADN synthétisés à partir de séquences d'un ou plusieurs génomes de virus sauvages [41,44]. On obtient ainsi un « nouveau » virus en moins d'un mois [58] (M/S voir 58).

## Des séquences de VIH et un site de clivage par la furine insérées dans le gène S de SARS-CoV-2 ?

Un doute sur l'origine du SARS-CoV-2 a été soulevé par l'observation de 4 insertions de courtes séquences (***Figures 3C, 2B et 5A-D***, i1 à i4) au sein de la séquence codant pour la protéine S . La quatrième insertion (i4) est particulièrement remarquable : elle est présente uniquement chez SARS-CoV-2, et elle confère une propriété particulière à la protéine [45]. Il s'agit de l'addition, au niveau du site de clivage entre S1 et S2, juste en amont d'une arginine, de quatre acides aminés **(*Figure 5D*)**, qui crée une séquence RRAR, qui correspond au site spécifique de protéolyse par la furine. Des modifications similaires, touchant le site de clivage de protéines d'enveloppe virale, favorisent l'infectiosité de différents virus respiratoires, comme l'influenza ou le virus Sendaï, en facilitant leur propagation à travers le tractus respiratoire et leur dissémination systémique [46,47].

La conservation de ce site de clivage, spécifique de la furine exprimée par l'hôte, dans tous les isolats de SARS-CoV-2 circulant dans les populations humaines suggère qu'elle a favorisé, sinon permis, le passage de la barrière d'espèce et/ou l'évolution de la forme épidémique du virus. L'importance de cette conservation pour la transmission entre humains est étayée par deux autres observations : ce site de protéolyse est instable quand on cultive le virus sur cellules VeroE6 (des cellules de singe), et des expériences sur les hamsters ont montré que la gravité des symptômes était atténuée lorsque le site furine était supprimé [48]. Une forte

pression de sélection s'exerce donc sur ce site ciblé par la furine afin de favoriser la propagation du SARS-CoV-2 chez l'homme.

Il faut également noter que l'apparition de sites de clivage par la furine chez les CoV humains n'est pas un événement exceptionnel. Des sites similaires sont présents chez d'autres CoV humains, différents de ceux du groupe des SARS-CoV, comme le MERS, HKU1, OC43 [45,49].

Trois autres insertions ont également été détectées **(*Figure 5A-C*)**. Elles se manifestent sous la forme de courtes séquences qui sont présentes dans le génome de SARS-CoV-2 mais absentes des génomes d'isolats de chiroptères (comme CoVZC45 et CoVZXC21) et du SARS-CoV de 2002. Dans une prépublication [50], les auteurs soulignaient un fait qu'ils qualifiaient de troublant : au niveau de ces insertions, la protéine S de SARS-CoV-2 présente des similarités avec des séquences de fragments des protéines ENV et GAG du virus VIH-1 (virus de l'immunodéficience humaine). Suite à des commentaires critiques concernant des faiblesses méthodologiques et d'interprétation, le manuscrit a été rétracté du site bioRxiv.

Ce "fait troublant," aurait donc dû rester anecdotique. Néanmoins, en avril 2020, le Professeur Luc Montagnier, Prix Nobel de médecine pour sa contribution à la découverte du VIH, défraie la chronique en proclamant que ces insertions ne résulteraient pas d'une recombinaison naturelle ou d'un accident, mais d'un vrai travail de génétique, effectué intentionnellement, vraisemblablement dans le cadre de recherches visant à développer des vaccins contre le VIH. Ces affirmations ont été immédiatement contestées par un grand nombre de scientifiques, qui ont rétorqué que les séquences similaires entre VIH et SARS-CoV-2 étaient tellement courtes (une trentaine de nucléotides sur un génome qui en compte 30 000) que leur ressemblance était vraisemblablement fortuite. La controverse s'est amplifiée, dans un contexte politique tendu où le président des États-Unis accusait la Chine d'avoir laissé échapper le virus manipulé d'un laboratoire P4 à Wuhan.

Ce type de polémique ne favorise pas une analyse sereine des faits. De façon paradoxale, à ce jour, aucune analyse approfondie n'a été publiée concernant l'origine de ces insertions. Des approches de bioinformatique et de phylogénie moléculaire sont pourtant susceptibles de nous apporter un éclairage intéressant, comme nous le montrons ci-dessous.

L'hypothèse de Luc Montagnier repose sur une analyse des similarités de séquences entre un fragment du gène S du SARS-CoV-2 et le génome du VIH. Nous reproduisons le résultat de cette analyse en ***Figure 6A***. L'alignement est caractérisé par un score qui estime l'espérance statistique (*expect*), autrement dit le nombre de similarités du même ordre qu'on s'attendrait à trouver si l'on avait aligné des séquences aléatoires : une ressemblance entre deux séquences est significative quand ce score est nettement inférieur à 1. Généralement, quand on compare des séquences de gènes homologues (résultant d'un gène ancestral commun), des scores aussi bas que $10^{-150}$ sont souvent atteints. Dans le cas qui nous intéresse, le score supérieur à 1 indique que l'alignement obtenu est fortuit, et ne

peut pas être considéré comme un indice d'homologie entre les séquences de VIH et de CoV. On peut aisément vérifier cela en soumettant à la même analyse une séquence dont les lettres ont été mélangées aléatoirement. La ***Figure 6B*** montre le résultat de ce test: on obtient des alignements aussi bons avec une séquence aléatoire **(*Figure 6A*)** qu'avec le gène de coronavirus **(*Figure 6B*)**. Les similarités entre coronavirus et VIH ne sont donc pas significatives.

D'après notre analyse phylogénétique, les quatre insertions que l'on observe chez le SARS-CoV-2 se trouvent chacune dans un sous-groupe différent de souches de coronavirus. Elles seraient donc apparues indépendamment, à différents moments de la diversification des virus **(*Figure 5)***. En particulier, les trois premières insertions sont observées dans les séquences de virus isolées à partir non seulement de chauve-souris (RaTG13), mais également de pangolins, provenant de Chine ou de Malaisie. L'hypothèse selon laquelle ces insertions résulteraient de manipulations expérimentales récentes ne permet donc pas d'expliquer leur présence dans plusieurs isolats viraux provenant de différentes espèces, prélevés à des endroits divers, et ce d'autant moins qu'elles se sont produites à différents moments au cours de l'évolution de ces souches virales.

Comment peut-on, dans ce contexte, comprendre l'apparition et le rôle de ces insertions ? L'analyse des alignements des protéines S montre que des insertions sont très fréquentes au sein de cette protéine chez les coronavirus. La structure de la protéine S, récemment résolue par cryo-microscopie électronique [26], nous indique que ces quatre insertions sont localisées à sa surface **(*Figure 3C*)**, ce qui suggère qu'elles participent à l'échappement du virus au contrôle de l'infection par l'immunité antivirale.

Une seconde hypothèse, régulièrement formulée, est que ce virus pourrait résulter d'une recombinaison produite en laboratoire entre un virus de chauves-souris du type RaTG13 et un domaine RBD de haute affinité pour l'homme, cloné à partir du SARS-CoV. Cette hypothèse s'avère également incohérente avec les analyses phylogénétiques des domaines RBD de CoV, le domaine RBD du SARS-CoV étant génétiquement très éloigné du SARS-CoV-2 **(*Figure 2B*)**. Les résidus jouant un rôle déterminant dans la reconnaissance du récepteur ACE2 de SARS-CoV ne sont, par ailleurs, pas conservés chez le SARS CoV-2 **(*Figure 4B*)**. Ces différences de séquences conduisent à une affinité du RBD pour le récepteur ACE2 20 fois plus élevée chez SARS-CoV-2 que chez SARS-CoV [27]. Toutefois, l'affinité du virus pour les cellules de l'hôte reste comparable à celle du SARS-CoV car l'accessibilité du RDB de la protéine S SARS-CoV-2 n'est pas optimale [51].

## État des lieux et perspectives

Nous avons montré que des analyses phylogénétiques pouvaient apporter un éclairage concernant les origines possibles du SARS-CoV-2, le virus responsable de la pandémie

Covid-19. Il ne s'agit que d'une première analyse et des études plus approfondies sont actuellement menées dans des laboratoires pour examiner les données disponibles et en extraire toutes les informations utiles. On peut espérer disposer prochainement de nouvelles données qui résoudront les questions restées sans réponse. Le bilan qui peut être fait à ce stade reste donc incomplet et provisoire, mais il est utile de se demander ce que l'on peut déjà conclure, sur la base des données qui sont en notre possession, et quels types de nouveaux résultats ou d'analyses nous apporteraient des informations complémentaires, voire nous permettraient de statuer définitivement quant aux origines du SARS-CoV-2.

Une première question est celle du dernier hôte animal ayant été infecté avant que le virus ne soit transmis à l'homme. Les analyses phylogénétiques indiquent que les CoV circulent fréquemment entre différentes espèces de chiroptères et passent occasionnellement chez d'autres mammifères. La coévolution des virus de ces animaux avec les hôtes potentiels et leur adaptation à ces nouveaux hôtes repose sur des mutations ponctuelles ainsi qu e sur des recombinaisons, fréquentes chez les coronavirus. Ces dernières posent des difficultés dans la mesure où une inférence phylogénétique fondée sur le génome complet s'avère biaisée par le mélange de fragments génomiques résultant de trajectoires évolutives différentes. Il est donc important d'identifier les régions recombinées, et de mener une étude phylogénétique séparément sur chacune d'entre elles. Les données disponibles suggèrent que le SARS-CoV-2 est issu de recombinaisons multiples de CoV de chiroptères. L'effet des recombinaisons est particulièrement important pour l'adaptabilité de la protéine S à son hôte, en raison de son rôle-clé dans l'interaction avec le récepteur ACE2 de l'hôte.

Le rôle éventuel des virus de pangolins dans ce processus de recombinaisons multiples reste incertain : bien que son importance fonctionnelle soit établie, la région de forte similarité est de taille réduite et la possibilité de transmission entre espèces n'est pas évidente à établir. La fiabilité des résultats dépend de la qualité du séquençage, des reconstructions métagénomiques, de l'accessibilité publique des données et de la précision de leur documentation dans des bases de données [52]. La clarification des hypothèses concernant l'émergence de SARS-CoV-2 nécessitera probablement le séquençage de nouveaux génomes de CoV potentiellement impliqués dans cette zoonose. Ceci nécessitera de chercher les virus qui circulent chez les chiroptères et dans des espèces en contact avec les populations humaines. Il serait souhaitable de se focaliser en priorité sur des espèces de mammifères dont le récepteur ACE2 présente des caractéristiques plus proches de l'ACE2 humain que celui des chiroptères, comme le porc, la chèvre, le mouton, la vache ou le chat **(*Figure 4A*)**.

Une insertion entre les fragments S1 et S2 de la protéine S a créé un site de clivage protéolytique sensible à la furine. Cette insertion est récente, puisque aucun virus connu proche de SARS-CoV-2 ne la contient. Le nouveau site furine qui en résulte semble contribuer à l'infectiosité du virus et/ou à sa propension épidémique chez l'homme. Cette

observation est cruciale car ce site a probablement joué un rôle déterminant dans le franchissement de la barrière d'espèce et/ou dans la transmission interhumaine du virus, une condition *sine qua non* pour l'émergence des épidémies.

L'origine animale du virus reste problématique. Une piste serait d'intensifier la collecte d'échantillons chez des espèces sauvages ou domestiques. L'étrange puzzle recombinatoire qu'est le génome de SARS-CoV-2 reste une énigme... Il nous faudra sans doute le résoudre pour comprendre ses origines. Même les souches les plus proches de SARS-CoV-2 (RATG13, RmYN02 pour la chauve-souris et MP789 pour le pangolin) présentent un taux de différences entre génomes par rapport à celui du de SARS-CoV-2 beaucoup plus élevé que ce qu'on attendrait chez un virus qui aurait été à l'origine de la dissémination humaine. La découverte de virus animaux présentant une très forte similarité avec SARS-CoV-2 fournirait un élément décisif pour valider son origine naturelle.

Le trafic des animaux sauvages et la déforestation, qui mettent les populations en contact avec la faune sauvage, ainsi que la création de nouveaux élevages à partir d'animaux sauvages reste également une question importante. En Chine, des élevages de pangolins ont été mis en place, qui posent non seulement des questions de faisabilité [54], mais aussi de nouvelles questions sanitaires. Ces nouveaux élevages exotiques s'ajoutent à tous les élevages intensifs d'animaux domestiques (volailles, porcs, etc.), qui constituent également des réservoirs de virus (pour la grippe, etc.), souvent localisés dans des zones à forte densité humaine [55].

Plusieurs laboratoires réalisent des expériences ciblées sur des gains de fonction afin de comprendre les interactions entre le domaine RBD des coronavirus et les récepteurs transmembranaires (comme ACE2) qui constituent leur point d'entrée dans les cellules de l'hôte. Dans ce contexte, pourrait-on envisager que le SARS-CoV-2 résulterait d'expériences visant à "humaniser" (adapter à l'humain) un virus animal, comme le RaTG13 ? À ce jour, les premières recherches effectuées par la communauté scientifique n'ont apporté aucun élément déterminant qui conforterait cette hypothèse. Toutefois des analyses bioinformatiques ont révélé des biais d'usage de codons alternatifs suggérant une possible manipulation génétique [53]. Des analyses plus approfondies seraient susceptibles de clarifier cette question.

Bien que des réglementations nationales existent (en France, réglementation Microorganismes et toxines, MOT), sur le plan mondial, la recherche, l'isolement et la culture de ces nouveaux virus respiratoires nécessitent d'être réalisés dans des conditions expérimentales les plus sécurisées possibles, avec une traçabilité irréprochable pour éviter toute transmission zoonotique. Au vu des risques infectieux, la société civile et la communauté scientifique devront réinterroger la pratique d'expériences de gain de fonction et l'adaptation, en laboratoire, des souches virales dans des hôtes animaux intermédiaires. En 2015, conscientes de ce problème, les agences fédérales américaines avaient gelé le

financement de toute nouvelle étude impliquant ces expériences [56]. Ce moratoire, relativement respecté, du financement de ce type de recherche, a pris fin en 2017 [57]. Une nouvelle évaluation des risques au regard des bénéfices potentiels de ces pratiques s'impose. Certes, il est souhaitable d'éviter l'écueil de réglementations trop strictes qui pourraient devenir stérilisantes pour l'étude des mécanismes moléculaires impliqués dans la propagation de ces virus et pour le développement d'antiviraux et de vaccins. Toutefois, les pratiques à haut risque devraient être repensées et encadrées au niveau international.

Sur la base des données actuelles ***(voir Tableau I)***, il est actuellement difficile de statuer à propos de l'émergence du SARS-CoV-2 et de déterminer s'il est le fruit d'une transmission zoonotique naturelle ou d'une fuite accidentelle à partir de souches expérimentales. Quelle que soit son origine, l'étude des mécanismes d'évolution et des processus moléculaires impliqués dans l'émergence de ce virus pandémique est et restera essentielle afin d'élaborer des stratégies thérapeutiques et vaccinales.

## Reproductibilité des analyses

Les analyses réalisées pour produire les résultats et les figures de cette article respectent les principes du code FAIR (facile à trouver, accessible, interopérable et réutilisable). L'environnement logiciel, les données de séquence, le code commenté et les exemples d'utilisation sont accessibles en libre accès (https://github.com/jvanheld/SARS-CoV-2_origins).

## Remerciements

Nous remercions les collègues suivants pour leur relecture attentive des premières versions du manuscrit, et leurs nombreuses suggestions qui ont permis de l'améliorer : Cathy Bellan, Mathias Bonal, Bruno Canard, Bruno Coutard, Hélène Chiapello, Denis Gerlier, Catherine Nguyen, Nadia Rabah, Annick Stevens, Denis Thieffry.

## Summary.

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

SARS-CoV-2 is a new human coronavirus (CoV), which emerged in China at the end of 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 be able to contain new epidemics. Based on phylogenetic inferences, sequence analysis and structure-function relationships of coronavirus proteins, informed by the knowledge currently available, we discuss the different scenarios evoked to account for the origin - natural or synthetic - of the

virus. On the basis of currently available data, it is impossible to determine whether SARS-CoV2 is the result of a natural zoonotic emergence or an accidental escape from experimental strains. Regardless of its origin, the study of the evolution of the molecular mechanisms involved in the emergence of this pandemic virus is essential to develop therapeutic and vaccine strategies.

**Full English translation available at https://hal.archives-ouvertes.fr/hal-02891455.**

## Références


1. Drosten C, Günther S, Preiser W, *et al.* Identification of a Novel Coronavirus in Patients with Severe Acute Respiratory Syndrome. *New England Journal of Medicine* 2003 ; 348 : 1967–1976.
2. Zaki AM, Boheemen S van, Bestebroer TM, *et al.* Isolation of a Novel Coronavirus from a Man with Pneumonia in Saudi Arabia. *New England Journal of Medicine* 2012 ; 367 : 1814–1820.
3. Lu R, Zhao X, Li J, *et al.* Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *Lancet* 2020 ; 395 : 565–574.
4. Menachery VD, Yount BL, Debbink K, *et al.* A SARS-like cluster of circulating bat coronaviruses shows potential for human emergence. *Nat. Med.* 2015 ; 21 : 1508–1513.
5. Hu B, Zeng L-P, Yang X-L, *et al.* Discovery of a rich gene pool of bat SARS-related coronaviruses provides new insights into the origin of SARS coronavirus. *PLoS Pathog.* 2017 ; 13 : e1006698.
6. Luk HKH, Li X, Fung J, *et al.* Molecular epidemiology, evolution and phylogeny of SARS coronavirus. *Infect. Genet. Evol.* 2019 ; 71 : 21–30.
7. Cui J, Li F, Shi Z-L. Origin and evolution of pathogenic coronaviruses. *Nature Reviews Microbiology* 2019 ; 17 : 181–192.
8. Song H-D, Tu C-C, Zhang G-W, *et al.* Cross-host evolution of severe acute respiratory syndrome coronavirus in palm civet and human. *Proc. Natl. Acad. Sci. U.S.A.* 2005 ; 102 : 2430–2435.
9. Sabir JSM, Lam TT-Y, Ahmed MMM, *et al.* Co-circulation of three camel coronavirus species and recombination of MERS-CoVs in Saudi Arabia. *Science* 2016 ; 351 : 81–84.
10. Luis AD, Hayman DTS, O'Shea TJ, *et al.* A comparison of bats and rodents as reservoirs of zoonotic viruses: are bats special? *Proc. Biol. Sci.* 2013 ; 280 : 20122753.
11. Zhou P, Yang X-L, Wang X-G, *et al.* A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* 2020 ; 579 : 270–273.
12. Ge X-Y, Wang N, Zhang W, *et al.* Coexistence of multiple coronaviruses in several bat colonies in an abandoned mineshaft. *Virol. Sin.* 2016 ; 31 : 31–40.
13. Zhou H, Chen X, Hu T, *et al.* A Novel Bat Coronavirus Closely Related to SARS-CoV-2 Contains Natural Insertions at the S1/S2 Cleavage Site of the Spike Protein. *Current Biology* 2020 ; 30 : 2196-2203.e3.
14. Ferron F, Subissi L, Silveira De Morais AT, *et al.* Structural and molecular basis of mismatch correction and ribavirin excision from coronavirus RNA. *Proc. Natl. Acad. Sci. U.S.A.* 2018 ; 115 : E162–E171.
15. Casane D, Policarpo M, Laurenti P. Pourquoi le taux de mutation n'est-il jamais égal à zéro ? *Med Sci (Paris)* 2019 ; 35 : 245–251.
16. Graham RL, Baric RS. Recombination, reservoirs, and the modular spike: mechanisms

of coronavirus cross-species transmission. *J. Virol.* 2010 ; 84 : 3134–3146.
17. Lam TT-Y, Shum MH-H, Zhu H-C, *et al.* Identifying SARS-CoV-2 related coronaviruses in Malayan pangolins. *Nature* 2020.
18. Xiao K, Zhai J, Feng Y, *et al.* Isolation of SARS-CoV-2-related coronavirus from Malayan pangolins. *Nature* 2020 ; 1–7.
19. Liu P, Jiang J-Z, Wan X-F, *et al.* Are pangolins the intermediate host of the 2019 novel coronavirus (SARS-CoV-2)? *PLoS Pathog.* 2020 ; 16 : e1008421.
20. Zhang T, Wu Q, Zhang Z. Probable Pangolin Origin of SARS-CoV-2 Associated with the Covid-19 Outbreak. *Current Biology* 2020 ; 30 : 1346-1351.e2.
21. Huang C, Wang Y, Li X, *et al.* Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *The Lancet* 2020 ; 395 : 497–506.
22. Andersen KG, Rambaut A, Lipkin WI, *et al.* The proximal origin of SARS-CoV-2. *Nature Medicine* 2020 ; 26 : 450–452.
23. Cheng VCC, Lau SKP, Woo PCY, *et al.* Severe Acute Respiratory Syndrome Coronavirus as an Agent of Emerging and Reemerging Infection. *Clinical Microbiology Reviews* 2007 ; 20 : 660–694.
24. Ziegler CGK, Allon SJ, Nyquist SK, *et al.* 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* 2020 ; 181 : 1016-1035.e19.
25. Hoffmann M, Kleine-Weber H, Schroeder S, *et al.* SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor. *Cell* 2020 ; 181 : 271-280.e8.
26. Wrapp D, Wang N, Corbett KS, *et al.* Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. *Science* 2020 ; 367 : 1260–1263.
27. Walls AC, Park Y-J, Tortorici MA, *et al.* Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein. *Cell* 2020 ; 181 : 281-292.e6.
28. Wang Q, Zhang Y, Wu L, *et al.* Structural and Functional Basis of SARS-CoV-2 Entry by Using Human ACE2. *Cell* 2020 ; 181 : 894-904.e9.
29. Ni L, Ye F, Cheng M-L, *et al.* Detection of SARS-CoV-2-specific humoral and cellular immunity in Covid-19 convalescent individuals. *Immunity* 2020.
30. Letko M, Marzi A, Munster V. Functional assessment of cell entry and receptor usage for SARS-CoV-2 and other lineage B betacoronaviruses. *Nature Microbiology* 2020 ; 5 : 562–569.
31. Yan R, Zhang Y, Li Y, *et al.* Structural basis for the recognition of SARS-CoV-2 by full-length human ACE2. *Science* 2020 ; 367 : 1444–1448.
32. Russell CA, Fonville JM, Brown AEX, *et al.* The potential for respiratory droplet-transmissible A/H5N1 influenza virus to evolve in a mammalian host. *Science* 2012 ; 336 : 1541–1547.
33. Imai M, Watanabe T, Hatta M, *et al.* Experimental adaptation of an influenza H5 HA confers respiratory droplet transmission to a reassortant H5 HA/H1N1 virus in ferrets. *Nature* 2012 ; 486 : 420–428.
34. Committee on Science T, Affairs P and G, Sciences B on L, *et al. Official Statements*. National Academies Press (US), 2013.
35. Van Boeckel TP, Tildesley MJ, Linard C, *et al.* The Nosoi commute: a spatial perspective on the rise of BSL-4 laboratories in cities. *arXiv:1312.3283 [q-bio]* 2013.
36. Enserink M. Singapore Lab Faulted in SARS Case. *Science* 2003 ; 301 : 1824–1824.
37. Normile D. Lab Accidents Prompt Calls for New Containment Program. *Science* 2004 ; 304 : 1223–1225.
38. Henkel RD, Miller T, Weyant RS. Monitoring Select Agent Theft, Loss and Release Reports in the United States—2004–2010: *Applied Biosafety* 2012.
39. Ren W, Qu X, Li W, *et al.* Difference in Receptor Usage between Severe Acute

Respiratory Syndrome (SARS) Coronavirus and SARS-Like Coronavirus of Bat Origin. *JVI* 2008 ; 82 : 1899–1907.
40. Menachery VD, Dinnon KH, Yount BL, *et al.* Trypsin Treatment Unlocks Barrier for Zoonotic Bat Coronavirus Infection. *J. Virol.* 2020 ; 94.
41. Zeng L-P, Gao Y-T, Ge X-Y, *et al.* 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* 2016 ; 90 : 6573–6582.
42. Follis KE, York J, Nunberg JH. Furin cleavage of the SARS coronavirus spike glycoprotein enhances cell-cell fusion but does not affect virion entry. *Virology* 2006 ; 350 : 358–369.
43. Belouzard S, Chu VC, Whittaker GR. Activation of the SARS coronavirus spike protein via sequential proteolytic cleavage at two distinct sites. *Proc. Natl. Acad. Sci. U.S.A.* 2009 ; 106 : 5871–5876.
44. Thao TTN, Labroussaa F, Ebert N, *et al.* Rapid reconstruction of SARS-CoV-2 using a synthetic genomics platform. *Nature* 2020 ; 1–8.
45. Coutard B, Valle C, Lamballerie X de, *et al.* The spike glycoprotein of the new coronavirus 2019-nCoV contains a furin-like cleavage site absent in CoV of the same clade. *Antiviral Res.* 2020 ; 176 : 104742.
46. Moulard M, Decroly E. Maturation of HIV envelope glycoprotein precursors by cellular endoproteases. *Biochimica et Biophysica Acta (BBA) - Reviews on Biomembranes* 2000 ; 1469 : 121–132.
47. Sun X, Tse LV, Ferguson AD, *et al.* Modifications to the Hemagglutinin Cleavage Site Control the Virulence of a Neurotropic H1N1 Influenza Virus. *Journal of Virology* 2010 ; 84 : 8683–8690.
48. Lau S-Y, Wang P, Mok BW-Y, *et al.* Attenuated SARS-CoV-2 variants with deletions at the S1/S2 junction. *Emerging Microbes & Infections* 2020 ; 9 : 837–842.
49. Matsuyama S, Shirato K, Kawase M, *et al.* Middle East Respiratory Syndrome Coronavirus Spike Protein Is Not Activated Directly by Cellular Furin during Viral Entry into Target Cells. *J. Virol.* 2018 ; 92.
50. Pradhan P, Pandey AK, Mishra A, *et al.* Uncanny similarity of unique inserts in the 2019-nCoV spike protein to HIV-1 gp120 and Gag. *bioRxiv* 2020 ; 2020.01.30.927871.
51. Shang J, Wan Y, Luo C, *et al.* Cell entry mechanisms of SARS-CoV-2. *PNAS* 2020 ; 117 : 11727–11734.
52. Hassanin A. The SARS-CoV-2-like virus found in captive pangolins from Guangdong should be better sequenced. *bioRxiv* 2020 ; 2020.05.07.077016.
53. Gu H, Chu D, Peiris M, *et al.* Multivariate Analyses of Codon Usage of SARS-CoV-2 and other betacoronaviruses. *bioRxiv* 2020 ; 2020.02.15.950568.
54. Hua L, Gong S, Wang F, *et al.* Captive breeding of pangolins: current status, problems and future prospects. *Zookeys* 2015 ; 99–114.
55. Gibbs AJ, Armstrong JS, Downie JC. From where did the 2009 “swine-origin” influenza A virus (H1N1) emerge? *Virology Journal* 2009 ; 6 : 207.
56. Statement on Funding Pause on Certain Types of Gain-of-Function Research *National Institutes of Health (NIH)* 2015.
57. Burki T. Ban on gain-of-function studies ends. *The Lancet Infectious Diseases* 2018 ; 18 : 148–149.
58. Iseni F, Tournier JN. Une course contre la montre : Création du SARS-CoV-2 en laboratoire, un mois après son émergence ! *Med/Sci (Paris)* 2020 ; 36 : xxx-xxx.

# Tableaux

## *Tableau I. Origines et dates de publication des souches de virus utilisées dans cet article.*

| Souche | Hôte | Origine de l'isolat | Date de l'isolat | Publication séquence | Précisions concernant l'origine de l'échantillon |
|---|---|---|---|---|---|
| BtBM48-31 | Chauve-souris | Bulgarie | 2008 | 1er octobre 2010 | |
| BtGX2013 | Chauve-souris | Chine | 2013 | 7 juillet 2017 | |
| BtHKU3-12 | Chauve-souris | Chine (non précisé) | non précisé | 5 avril 2010 | Chine d'après la publication mais origine non indiquée dans NCBI |
| BtRaTG13_2013_Yunnan | Chauve-souris | Yunnan, Chine | 24 juillet 2013 | 24 mars 2020 | Séquence publiée en 2020, annotée comme isolée en 2013.<br>Les séquences génomiques partielles (région RdRp) publiées par groupe de Shi de 2016 i ont 100% d'identité avec RaTG13. Génome reconstruit à partir d'échantillons de 6 espèces de chauve-souris |
| BtRs4874 | Chauve-souris | Chine | 21 juillet 2013 | 18 décembre 2017 | Groupe de Shi à Wuhan (province de Hubei, Chine) |
| BtYN2013 | Chauve-souris | Chine | 2013 | 7 Juillet 2017 | |
| BtYN2018B | Chauve-souris | Chine | 1er septembre, 2016 | Juin 30, 2019 | |
| BtYu-RmYN02_2019 | Chauve-souris | Chine Yunnan - Xishuangbanna | 25 juin 2019 | Février 3, 2020 | Métagénome construit par séquençage d'un mélange de 11 échantillons provenant de fèces de chauves-souris de l'espèce *Rhinolophus malayanus* |
| BtZC45 | Chauve-souris | Zhoushan | 2017 | | |
| BtZXC21 | Chauve-souris | Zhoushan | 2015 | 5 février 2020 | |
| Cv007-2004 | Civette | Chine : Guangzhou dans la province de Guangdong | 2019 | 1er décembre 2005 | Virus de civette le plus proche de celui du SARS-CoV de 2003.<br>Mentionné dans l'article : "These cases were not linked to any laboratory accident." |
| HuCoV2_WH01_2019 | Humain | Chine, Hubei, Wuhan | 23 décembre 2019 | 11 février 2020 | Génome de référence pour la pandémie Covid-19 |
| HuSARS-Frankfurt-1_2003 | Humain | Francfort | 2003 | 16 Mars 2004 | Génome de référence pour l'épidémie SRAS de 2003 |
| PnGu-P2S_2019 | Pangolin | Chine, Guangdong | 2019 | 17 février 2020 | Séquence disponible dans GISAID, très proche de MP789. Version pré-publication ? |
| PnMP789 | Pangolin | Chine: pangolins malaisiens de contrebande, douanes du Guangdong | 29 mars 2019 | 23 avril 2020 | Métagénome assemblé à partir d'échantillons de 3 pangolins récoltés en mars et en juillet 2019. |
| PnGu1_2019 | Pangolin | Chine, Guangdong | 2019 | 18 février 2020 | |
| PnGX-P1E_2017 | Pangolin | Douanes chinoises sur un vol en provenance de malaisie | 2017 | 23 avril 2020 | |
| PnGX-P2V_2018 | Pangolin | Douanes chinoises sur un vol en provenance de malaisie | 2018 | 23 avril 2020 | Prélevée chez le pangolin, cette souche a été cultivée sur cellules humaines (et donc vraisemblablement adaptée à l'infection d'humains) |

## Légende des figures

**Figure 1. Phylogénie et émergence des coronavirus. (A)** Arbre de génomes complets de coronavirus, en se basant sur un alignement multiple (clustalw) suivi d'une inférence en maximum de vraisemblance (PhyML). Les génomes assemblés à partir de données métagénomiques sont marqués d'une étoile. Le préfixe des virus correspond aux espèces: Bt (chauve-souris), Hu (humain), Pn (pangolin), Cv (civette), Cm (dromadaire), Pi (porc). On constate que les distances entre HuCoV2 et les souches virales les plus proches (BtYuRmYN02, BtRaTG13) sont plus élevées que pour SARS-CoV (humain - civette) ou MERS-CoV (humain-dromadaire) **(B-D)** Hypothèses de transmission du réservoir animal (chauve-souris) jusqu'à l'homme, basées sur la phylogénie moléculaire des génomes viraux. **(B)** Pour la pandémie SARS-Cov de 2003, l'hôte intermédiaire est la civette. Une transmission directe de la chauve-souris à l'homme est également envisagée. **(C)** Pandémie MERS-CoV de 2012, avec le dromadaire comme hôte intermédiaire. Plusieurs événements de transmission directe ont été documentés. **(D)** Pandémie Covid-19. Plusieurs scénarios sont proposés concernant le dernier hôte avant la transmission à l'humain.

**Figure 2. Le taux d'identité entre SARS-CoV-2 et les autres coronavirus varie selon la position dans le génome.** Le niveau 100% correspond au génome de référence du SARS-CoV-2. **(A)** Pourcentage d'identité entre SARS-CoV-2 et d'autres coronavirus le long du génome entier (fenêtres glissantes de 800 nucléotides). **(B)** Pourcentage d'identité le long du gène S (fenêtres glissantes de 200 nucléotides). **(C-E)** Impact des recombinaisons sur la topologie des arbres phylogénétiques inférés à partir de différentes régions génomiques: gène ORF1ab **(C)**, région codante de la partie S1 de la protéine S **(D)**, et RBD **(E)**.

**Figure 3. Structure et fonctions de la protéine S** (spicule, spike en anglais). **(A)** Représentation schématique de l'infection des cellules par le SARS-CoV-2 après fixation de la protéine S au récepteur ACE2. **(B)** La protéine S subit deux étapes de maturation par clivage protéolytique (par les protéases furine puis TMPRSS2) nécessaires à son activation et à la libération du peptide de fusion. **(C)** Structure de la protéine S fixée au récepteur ACE2. La structure de la protéine S de SARS-CoV-2 (en beige) est obtenue grâce au logiciel SWISSMODEL sur la base de la structure 6acc de SARS-CoV (disponible dans protein data bank), et alignée sur la structure d'un domaine RBD (en orange) interagissant avec ACE2 (en gris) issue du modèle 6m0j (disponible dans protein data bank). Les sites d'insertion sont indiqués en couleurs. Les résidus sont colorés en fonction de l'ordre de conservation des insertions, en passant du rouge (insertion présente uniquement chez SARS-CoV-2), au jaune, vert, bleu clair puis indigo (insertion présente chez la majorité des sarbécovirus).

**Figure 4. Conservation de la protéine ACE2 et interactions avec la protéine S. (A)** Divergences entre la protéine ACE2 humaine et celle de plusieurs espèces animales au niveau des résidus impliqués dans la fixation de SARS-CoV-2 (d'après [31]). **(B)** Interactions entre ACE2 et S, et conservation des résidus importants (d'après [31]) chez différentes souches virales et espèces animales. Les principales interactions entre résidus de S et d'ACE2 sont indiquées par des traits pleins, certaines interactions plus faibles sont indiquées en pointillés.

**Figure 5. Étude des insertions observées dans la protéine S de SARS-CoV-2.** (A-D) Alignements des séquences protéiques de plusieurs coronavirus au niveau d'insertions dans la protéine S de SARS-CoV-2. **(A)** Insertion correspondant aux résidus 153-158 de la protéine S de SARS-CoV-2. **(B)** Insertion correspondant aux résidus 245-251. **(C)** Insertion correspondant aux résidus 445-449. **(D)** Insertion correspondant aux résidus 680-683. Le schéma du haut indique la position des 4 insertions dans le gène S. Chaque panneau **(A-D)** montre les alignements de séquences d'acides aminés aux alentours de l'insertion (gauche), et l'origine probable de l'insertion dans la phylogénie inférée sur la base des séquences encadrant l'insertion (droite). A part pour l'insertion i3b, les phylogénies regroupent ensemble les séquences qui partagent une même insertion, ce qui indique une origine unique de chaque insertion. Les profondes différences entre les phylogénies indiquent que les régions d'insertions ont des histoires évolutives très différentes. Les valeurs associées aux bifurcations indiquent la robustesse des branchements sur une échelle de 1 à 100. Les valeurs faibles (<50) indiquent qu'un branchement particulier a une faible fiabilité. On note que les valeurs faibles sont souvent associées à BtYuRmYN02, qui résulte d'un assemblage métagénomique composé d'un grand nombre d'échantillons de sources diverses. Ce métagénome est également celui qui montre le plus d'incohérence entre les différents fragments alignés, ce qui met en cause sa pertinence biologique.

**Figure 6. Recherche de similarités entre les séquences codant pour la protéine spike de CoV2 et le génome de HIV. (A)** Alignement le plus significatif entre la séquence codant pour la protéine S de SARS-CoV-2 (query) et le génome du HIV (subject). **(B)** Contrôle négatif : alignement le plus significatif entre une séquence aléatoire, obtenue en mélangeant les nucléotides de la séquence précédente, et le génome du HIV. Noter la valeur du score "expect", qui indique le nombre de faux-positifs attendus au hasard. Ce score présente pour les deux alignements des valeurs supérieures à 1, et est même plus élevé pour l'alignement de la séquence de CoV que pour la séquence aléatoire. On peut en conclure que la similarités entre la séquence codante de la protéine S et le génome du HIV n'est pas significative. Les alignements ont été réalisés sur le site BLAST du NCBI (https://blast.ncbi.nlm.nih.gov/Blast.cgi).

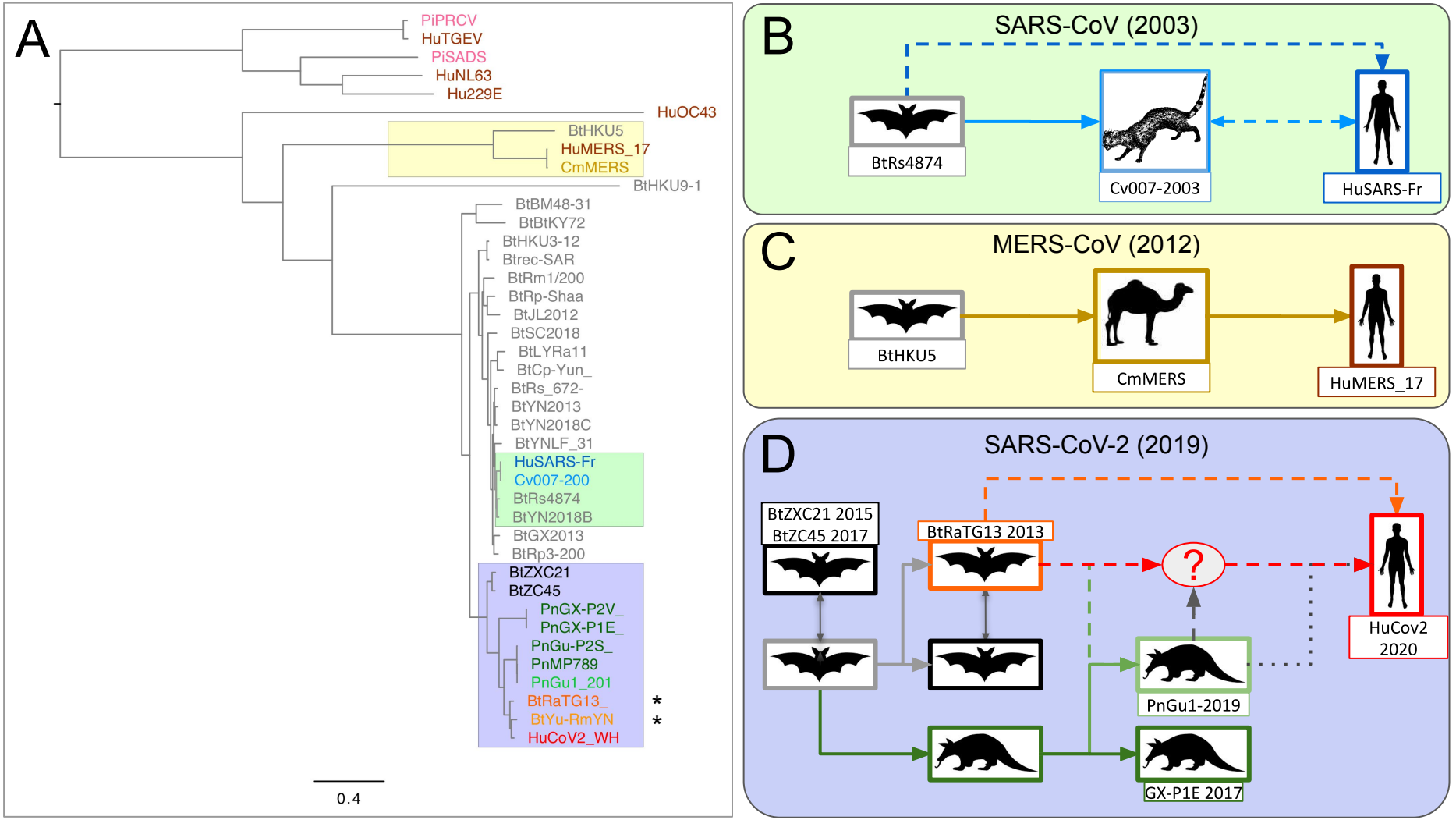
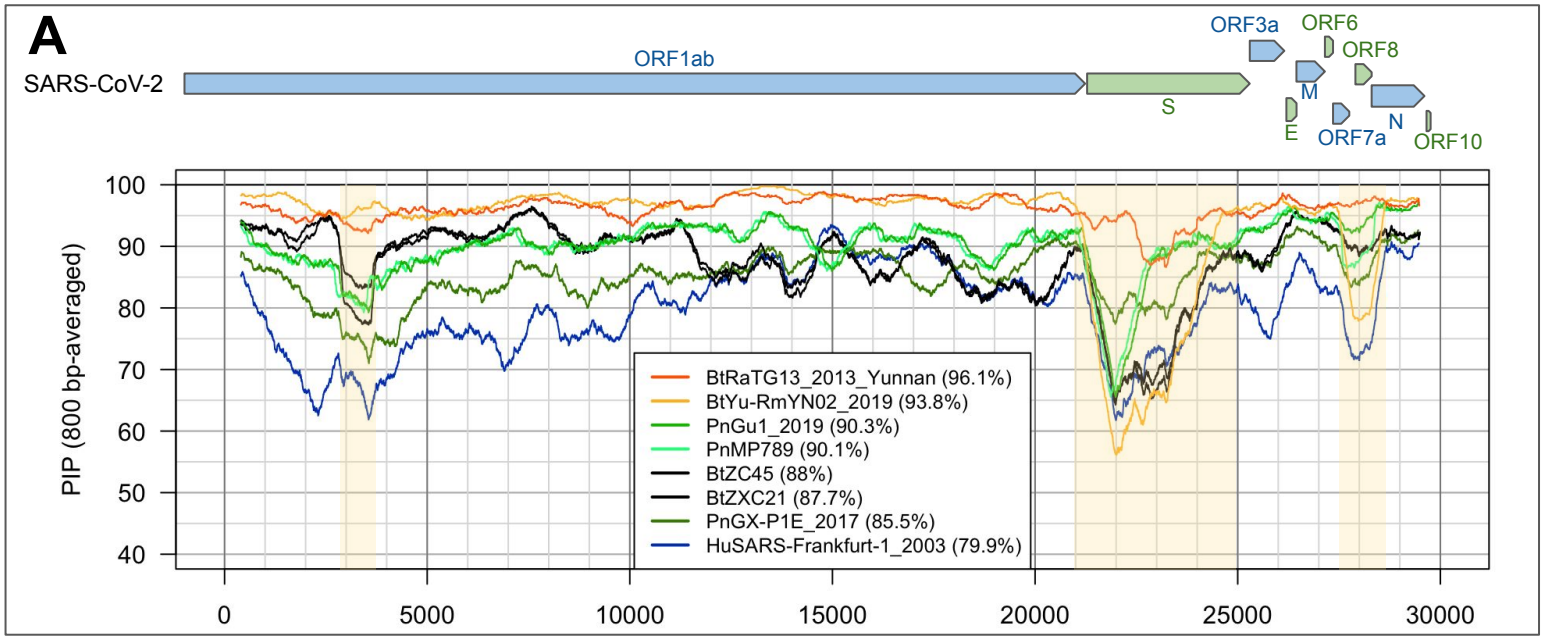
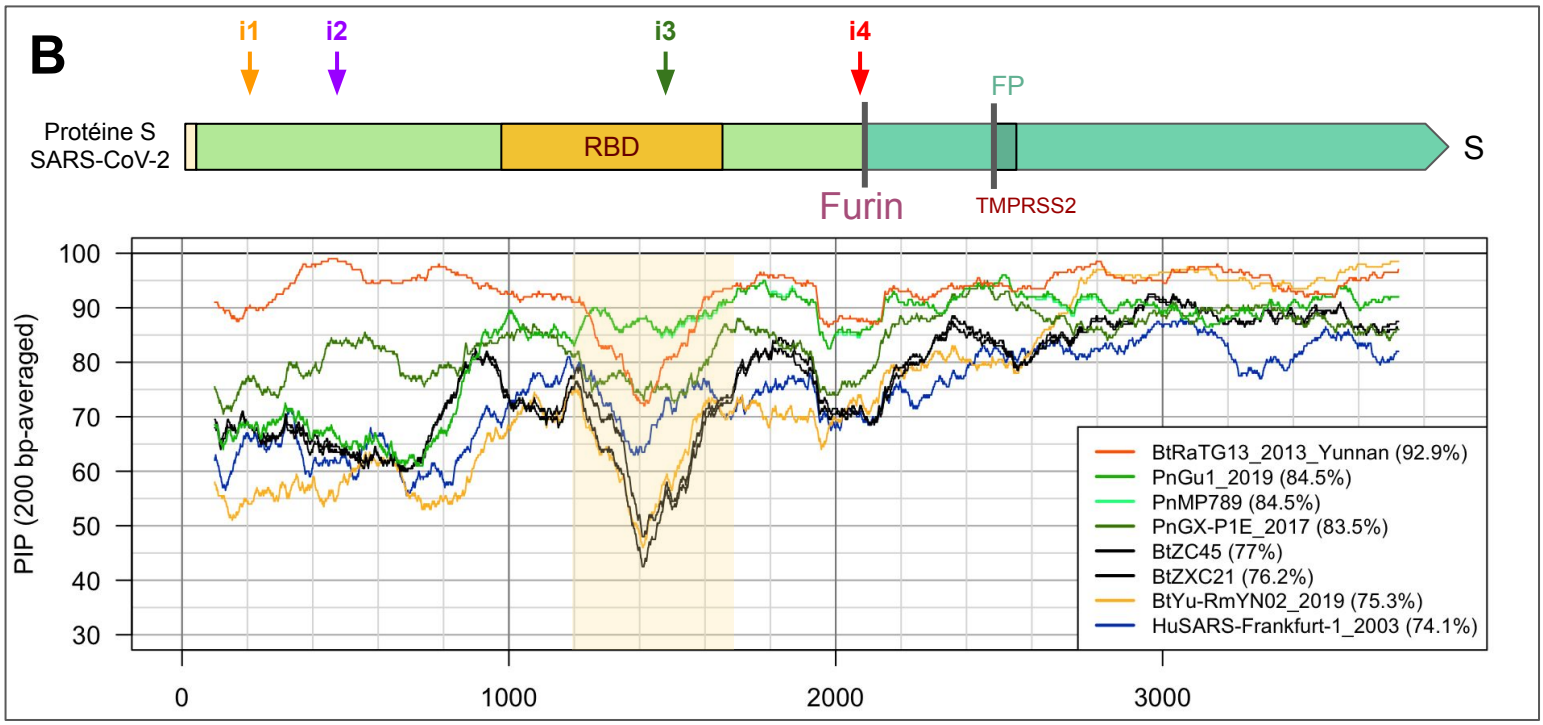
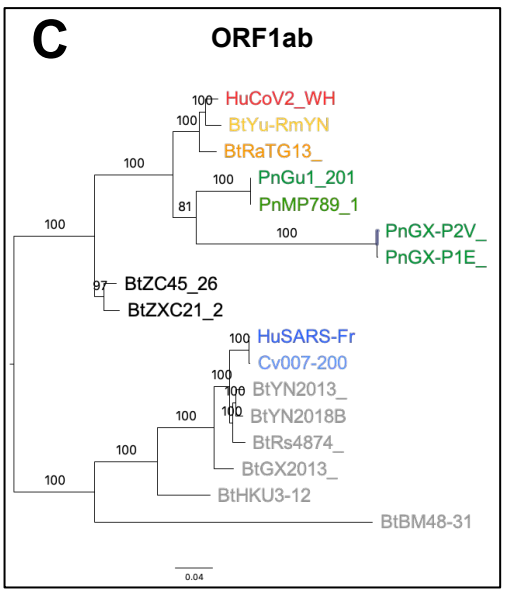
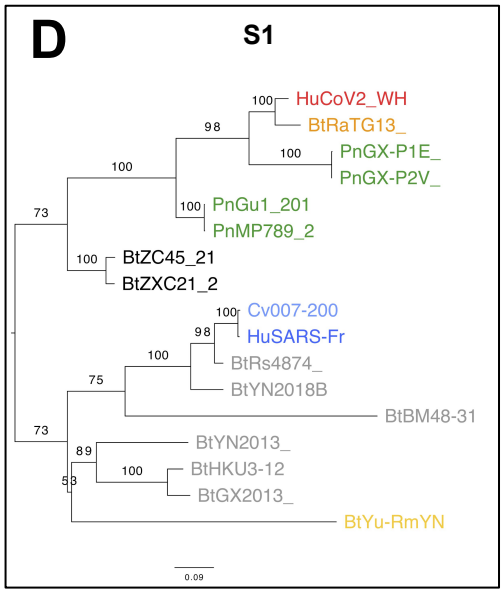

A
PiPRCV
HuTGEV
PiSADS
HuNL63
Hu229E
HuOC43
BtHKU5
HuMERS_17
CmMERS
BtHKU9-1
BtBM48-31
BtBtKY72
BtHKU3-12
Btrec-SAR
BtRm1/200
BtRp-Shaa
BtJL2012
BtSC2018
BtLYRa11
BtCp-Yun_
BtRs_672-
BtYN2013
BtYN2018C
BtYNLF_31
HuSARS-Fr
Cv007-200
BtRs4874
BtYN2018B
BtGX2013
BtRp3-200
BtZXC21
BtZC45
PnGX-P2V_
PnGX-P1E_
PnGu-P2S_
PnMP789
PnGu1_201
BtRaTG13_
BtYu-RmYN
HuCoV2_WH
*
*
0.4
B
SARS-CoV (2003)
BtRs4874
Cv007-2003
HuSARS-Fr
C
MERS-CoV (2012)
BtHKU5
CmMERS
HuMERS_17
D
SARS-CoV-2 (2019)
BtZXC21 2015
BtZC45 2017
BtRaTG13 2013
?
HuCov2
2020
PnGu1-2019
GX-P1E 2017

Figure 1


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
Protéine S
SARS-CoV-2
RBD
FP
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

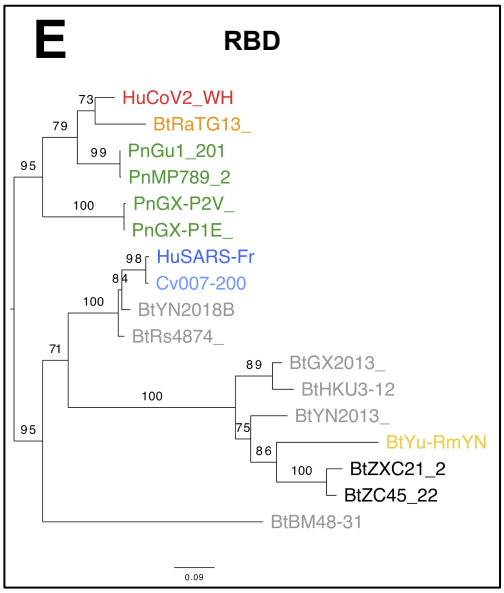
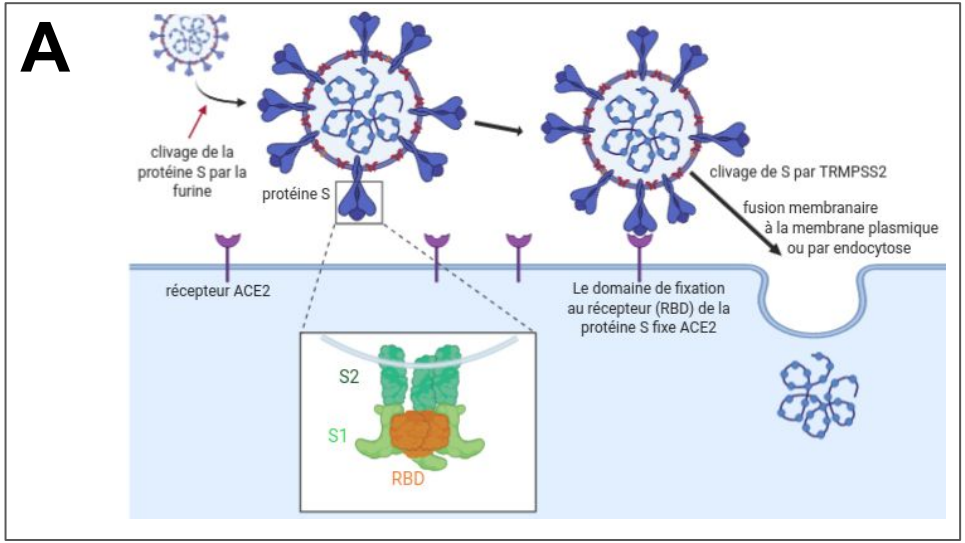
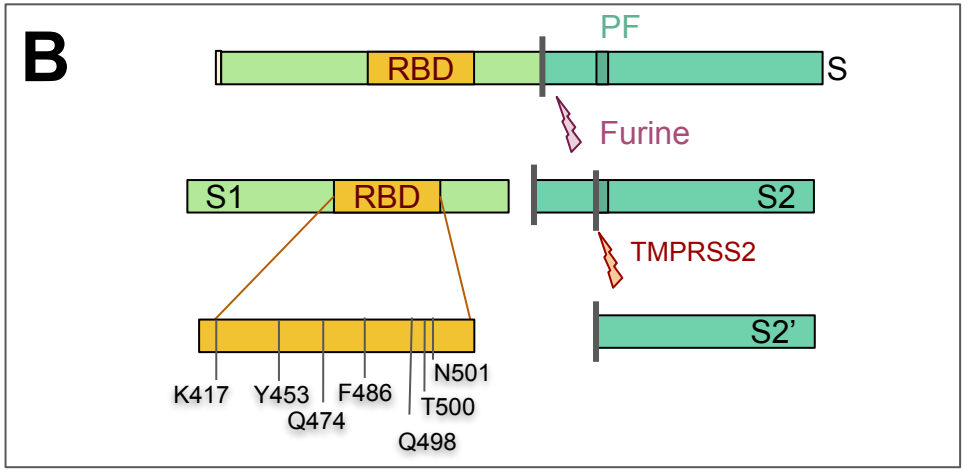

E
RBD
HuCoV2_WH
BtRaTG13_
PnGu1_201
PnMP789_2
PnGX-P2V_
PnGX-P1E_
HuSARS-Fr
Cv007-200
BtYN2018B
BtRs4874_
BtGX2013_
BtHKU3-12
BtYN2013_
BtYu-RmYN
BtZXC21_2
BtZC45_22
BtBM48-31

Figure 2


A
clivage de la protéine S par la furine
protéine S
clivage de S par TRMPSS2
fusion membranaire à la membrane plasmique ou par endocytose
récepteur ACE2
Le domaine de fixation au récepteur (RBD) de la protéine S fixe ACE2
S2
S1
RBD


B
PF
RBD
S
Furine
S1
RBD
S2
TMPRSS2
S2'
K417
Y453
F486
N501
T500
Q474
Q498

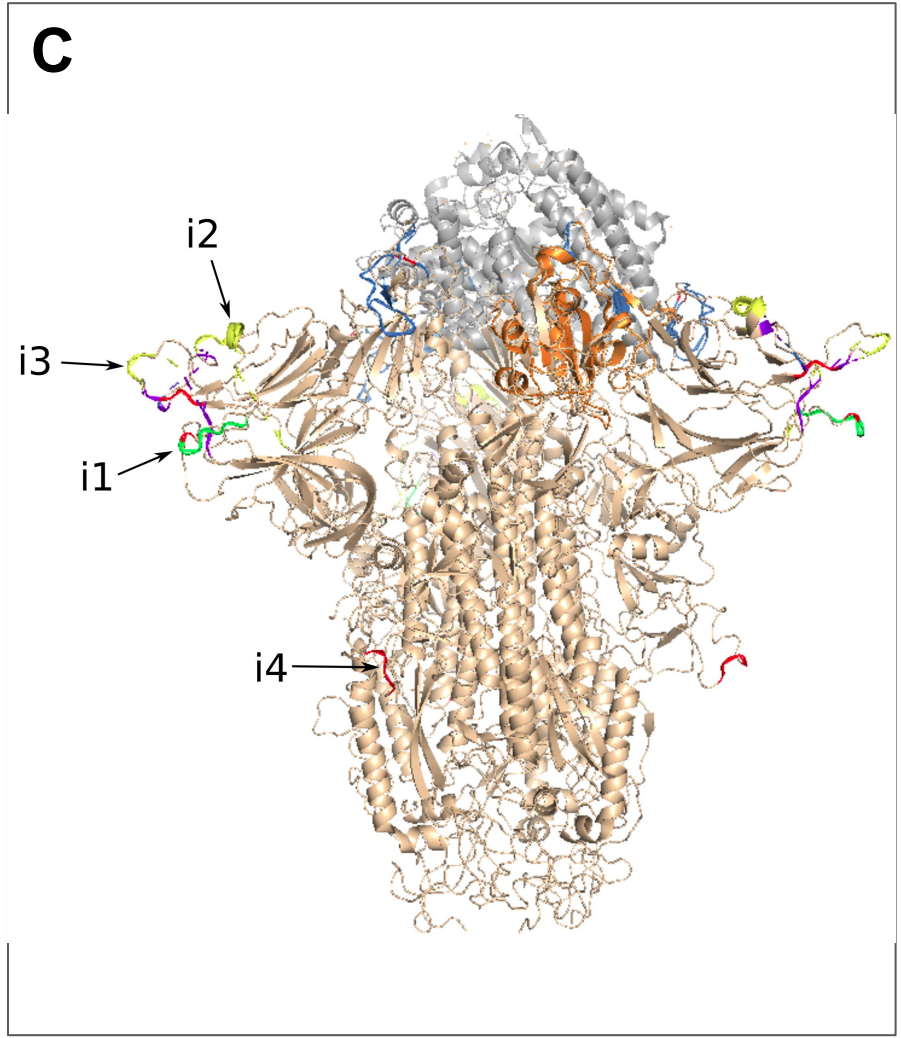
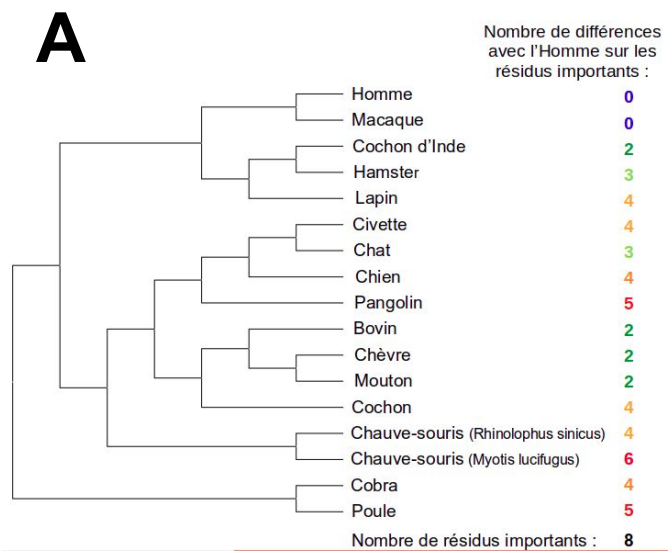

C
i2
i3
i1
i4

Figure 3


A
Nombre de différences avec l'Homme sur les résidus importants :
Homme 0
Macaque 0
Cochon d'Inde 2
Hamster 3
Lapin 4
Civette 4
Chat 3
Chien 4
Pangolin 5
Bovin 2
Chèvre 2
Mouton 2
Cochon 4
Chauve-souris (Rhinolophus sinicus) 4
Chauve-souris (Myotis lucifugus) 6
Cobra 4
Poule 5
Nombre de résidus importants : 8

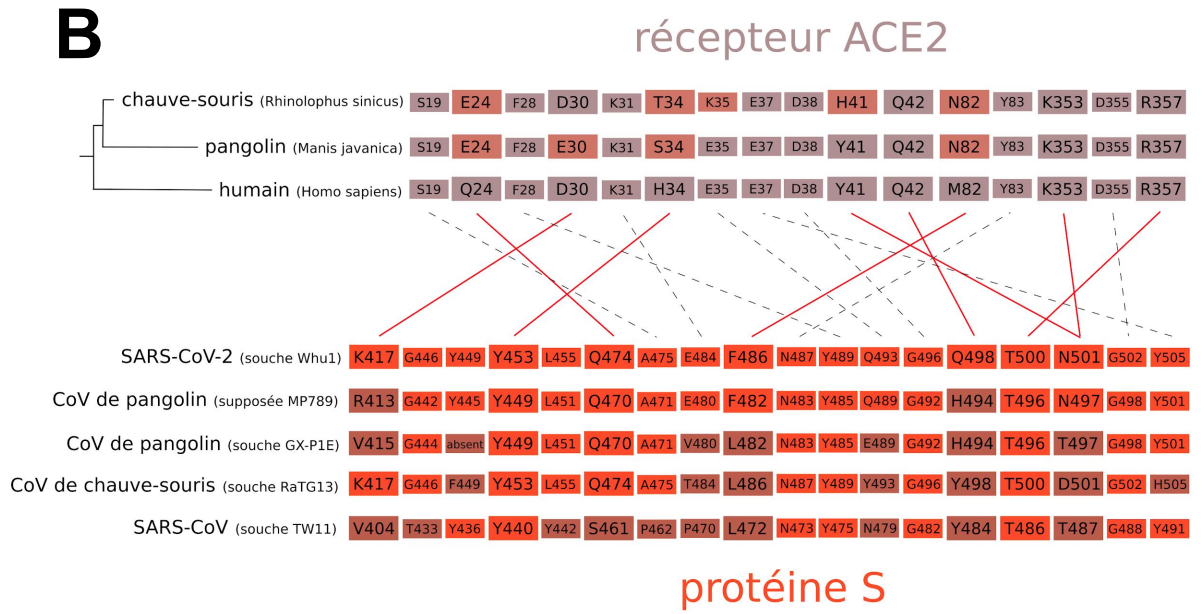
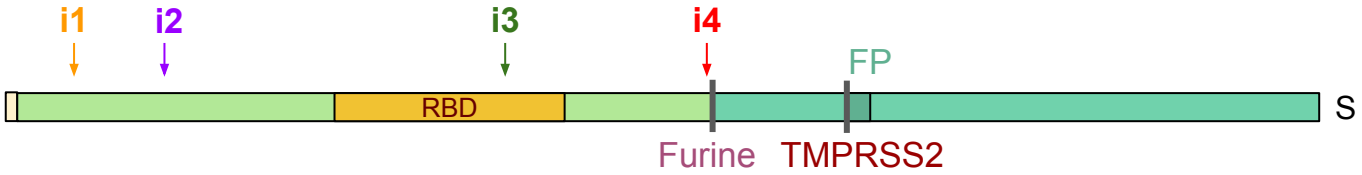
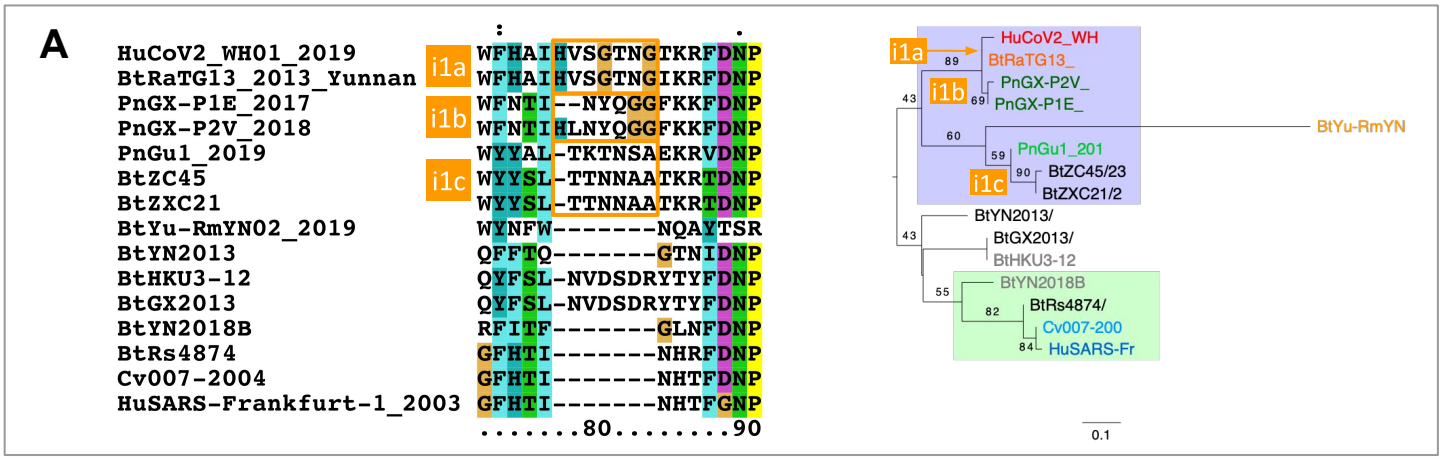
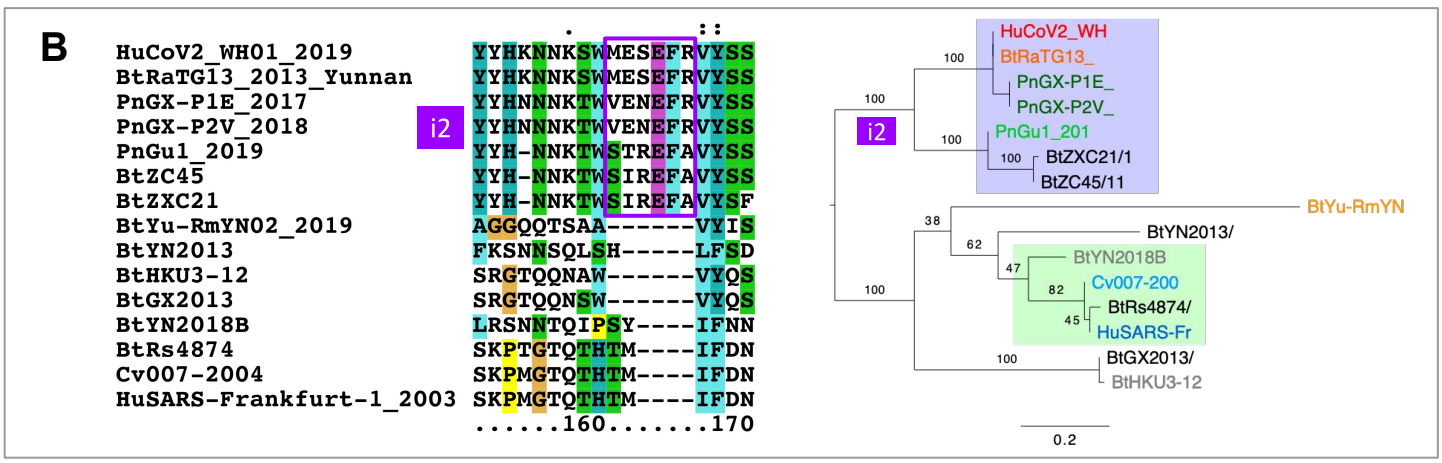
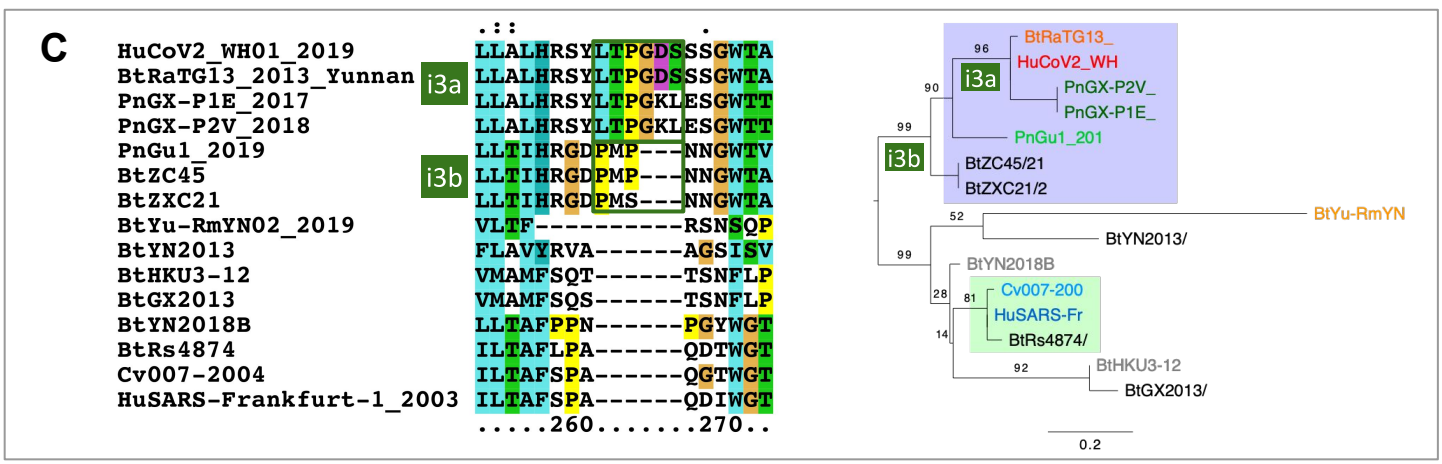

B
récepteur ACE2
chauve-souris (Rhinolophus sinicus) S19 E24 F28 D30 K31 T34 K35 E37 D38 H41 Q42 N82 Y83 K353 D355 R357
pangolin (Manis javanica) S19 E24 F28 E30 K31 S34 E35 E37 D38 Y41 Q42 N82 Y83 K353 D355 R357
humain (Homo sapiens) S19 Q24 F28 D30 K31 H34 E35 E37 D38 Y41 Q42 M82 Y83 K353 D355 R357
SARS-CoV-2 (souche Whu1) K417 G446 Y449 Y453 L455 Q474 A475 E484 F486 N487 Y489 Q493 G496 Q498 T500 N501 G502 Y505
CoV de pangolin (supposée MP789) R413 G442 Y445 Y449 L451 Q470 A471 E480 F482 N483 Y485 Q489 G492 H494 T496 N497 G498 Y501
CoV de pangolin (souche GX-P1E) V415 G444 absent Y449 L451 Q470 A471 V480 L482 N483 Y485 E489 G492 H494 T496 T497 G498 Y501
CoV de chauve-souris (souche RaTG13) K417 G446 F449 Y453 L455 Q474 A475 T484 L486 N487 Y489 Y493 G496 Y498 T500 D501 G502 H505
SARS-CoV (souche TW11) V404 T433 Y436 Y440 Y442 S461 P462 P470 L472 N473 Y475 N479 G482 Y484 T486 T487 G488 Y491
protéine S

Figure 4


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
WFHAIHVSGTNGIKRFDNP
WFNTI--NYQGGFKKFDNP
WFNTIHLNYQGGFKKFDNP
WYYAL-TKTNSAEKRVDNP
WYYSL-TTNNAATKRTDNP
WYYSL-TTNNAATKRTDNP
WYNFW-------NQAYTSR
QFFTQ-------GTNIDNP
QYFSL-NVDSDRYTYFDNP
QYFSL-NVDSDRYTYFDNP
RFITF-------GLNFDNP
GFHTI-------NHRFDNP
GFHTI-------NHTFDNP
GFHTI-------NHTFGNP
.....80.....90
i1a
i1b
i1c
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
YYHKNNKSWMESEFRVYSS
YYHKNNKSWMESEFRVYSS
YYHNNNKTWVENEFRVYSS
YYH-NNKTWSTREFAVYSS
YYH-NNKTWSIREFAVYSS
YYH-NNKTWSIREFAVYSF
AGGQQTSAA------VYIS
FKSNNSQLSH-----LFSD
SRGTQQNAW------VYQS
SRGTQQNSW------VYQS
LRSNNTQIPSY----IFNN
SKPTGTQTHTM----IFDN
SKPMGTQTHTM----IFDN
SKPMGTQTHTM----IFDN
......160.....170
i2
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
LLALHRSYLTPGDSSSGWTA
LLALHRSYLTPGDSSSGWTA
LLALHRSYLTPGKLESGWTT
LLALHRSYLTPGKLESGWTT
LLTIHRGDPMP---NNGWTV
LLTIHRGDPMP---NNGWTA
LLTIHRGDPMS---NNGWTA
VLTF----------RSNSQP
FLAVYRVA------AGSISV
VMAMFSQT------TSNFLP
VMAMFSQS------TSNFLP
LLTAFPPN------PGYWGT
ILTAFLPA------QDTWGT
ILTAFSPA------QGTWGT
ILTAFSPA------QDIWGT
.....260.....270..
i3a
i3b
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

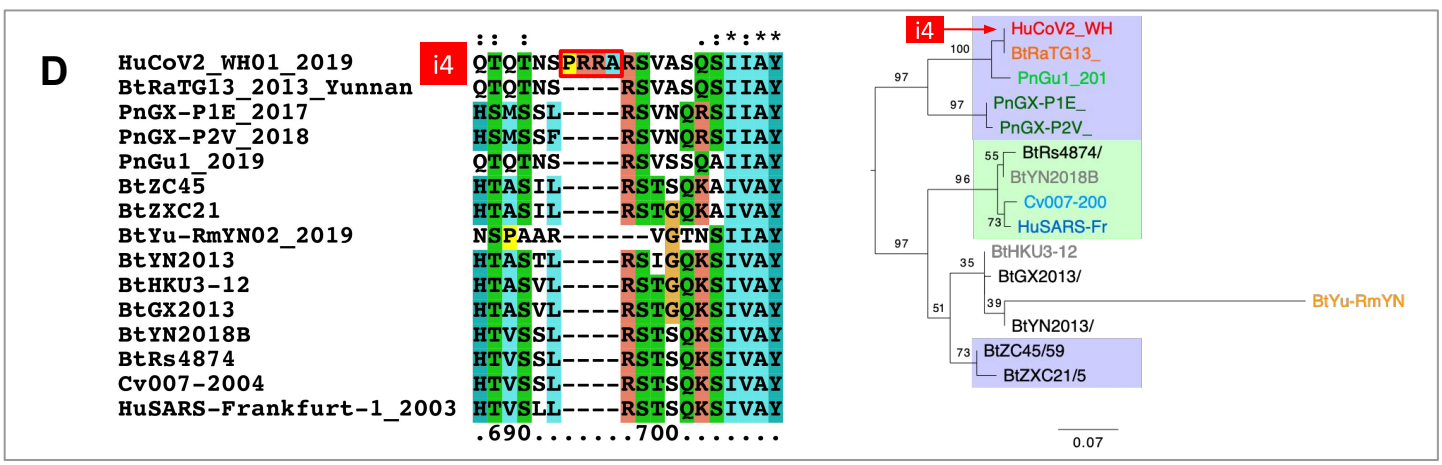

D
HuCoV2_WH01_2019
BtRaTG13_2013_Yunnan
PnGX-P1E_2017
PnGX-P2V_2018
PnGu1_2019
BtZC45
BtZXC21
BtYu-RmYN02_2019
BtYN2013
BtHKU3-12
BtGX2013
BtYN2018B
BtRs4874
Cv007-2004
HuSARS-Frankfurt-1_2003
i4
QTQTNSPRRARSVASQSIIAY
QTQTNS----RSVASQSIIAY
HSMSSL----RSVNQRSIIAY
HSMSSF----RSVNQRSIIAY
QTQTNS----RSVSSQAIIAY
HTASIL----RSTSQKAIVAY
HTASIL----RSTGQKAIVAY
NSPAAR------VGTNSIIAY
HTASTL----RSIGQKSIVAY
HTASVL----RSTGQKSIVAY
HTASVL----RSTGQKSIVAY
HTVSSL----RSTSQKSIVAY
HTVSSL----RSTSQKSIVAY
HTVSSL----RSTSQKSIVAY
.690.....700.....
i4
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

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

Figure 6