

Les problèmes de classification de SARS-CoV-2

Daniel Parrochia

▶ **To cite this version:**

Daniel Parrochia. Les problèmes de classification de SARS-CoV-2. 2021. hal-03238987

HAL Id: hal-03238987
https://hal.science/hal-03238987

Preprint submitted on 27 May 2021

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

Les problèmes de classification de SARS-CoV-2

Daniel Parrochia

Université de Lyon (France)

Résumé. Deux types de méthodes permettent actuellement d'analyser la proximité des vivants ou des virus : la classification automatique et les arbres phylogénétiques. La classification automatique permet d'évaluer la distances entre certaines séquences de génomes apparentés afin d'établir la proximité relative des éléments qui les portent. Les arbres phylogénétiques sont des représentations schématiques de l'évolution entre différentes espèces apparentées en fonction de leurs similitudes génétiques et physiques. Les virus sont des espèces biologiques qui comprennent des séquences d'acides nucléiques. Ils sont sujets à un large éventail de changements évolutifs, y compris des mutations de leurs séquences. La capacité de celles-ci à accumuler des changements par mutations ou recombinaison avec d'autres espèces donne naissance à de nouvelles lignées virales. Les données dont on dispose aujourd'hui rendent complexes une attribution définitive de l'espèce SARS-Cov-2, à l'origine de la pandémie CoVid-19, et permettent difficilement de répondre à la question de son origine. Certaines particularités du virus laissent notamment planer un doute sur son origine naturelle.

Mots clés : Covid-19, Coronavirus, SARS-Cov-2, RaTG-13, classification, arbre phylogénétique, chauves-souris, Chine.

1 Introduction

Lors d'une épidémie virale, il est important d'établir rapidement si l'épidémie est causée par un nouveau virus ou un virus déjà connu , ce qui permet ensuite de décider des approches et des actions les plus appropriées pour le détecter, contrôler sa transmission et, d'une façon générale, le combattre. Il faut donc pouvoir situer ce virus par rap-

port à d'autres, évaluer sa proximité ou sa distance vis à vis de ceux-là, en un mot : le classer. Une telle évaluation n'a pas simplement des implications théoriques comme la dénomination des virus. Pratiquement, même si c'est à une autre échelle, elle pourra aider à définir les priorités de recherche en virologie et en santé publique. Ce genre d'incidences pragmatiques fait, d'une façon générale, l'intérêt des classifications. Au-delà de la recherche d'un ordre économique qui résume la connaissance, c'est là l'une de leurs principales motivations (voir [Dagognet 70] ;[Parrochia 13]).

S'agissant d'un nouveau virus à situer, on va en général tenter d'évaluer son degré de parenté avec des virus précédemment identifiés infectant le même hôte ou avec des groupes monophylétiques de virus bien établis, souvent connus sous le nom de génotypes ou clades [1], qui peuvent inclure ou non des virus d'hôtes différents.

Cette question est formellement traitée dans un cadre officiel. La taxonomie virale, supervisée et coordonnée par l'ICTV (l'*International Committee on Taxonomy of Viruses*) regroupe les virus en taxons dans un schéma hiérarchique de rangs dans lequel l'espèce représente le rang le plus bas et le plus peuplé, contenant les groupes de virus (taxons) les moins divergents. Un suivi est organisé, impliquant des groupes d'étude pour chaque famille de virus. Bien qu'il existe de plus en plus de virus qui ne sont liés à aucune maladie connue chez leurs hôtes respectifs, on notera tout de même que, pour l'homme, les virus étant souvent des agents responsables de maladies, les noms des virus sont liés aux maladies qu'ils provoquent et l'OMS est chargée de nommer les maladies causées par les nouveaux virus humains émergents.

Dans l'immensité des virus, les Coronavirus (virus à ARN, à brin positif, enveloppés) forment une famille de 39 espèces réparties en 27 sous-genres, 5 genres et 2 sous-familles relevant de la famille des Coronaviridae, sous-ordre des Cornidovirineae, ordre des Nidovirales et règne des Riboviria (voir [Siddell 19] ; [Ziebuhr 19]).

Le SARS-Cov-2 a été classé dans le sous-genre des Sarbecoronavirus, lui-même contenu dans le genre des BétaCoronavirus, situé dans la sous-famille des Orthocoronavirinae, elle-même comprise dans la famille des Coronaviridae (voir Fig. 1).

Cette classification renvoie à des critères particuliers, d'ailleurs évolutifs dans le temps, et qui n'ont pas, pour l'instant, de motivation mathématiquement indiscutable.

Comme le rapporte le *Coronaviridae Study Group of the International Committee on Taxonomy of Viruses* (CSJCTV) (voir [CSJCTV 20]), la classification des Corona-

1. Rappelons qu'un *clade* est un terme désignant un groupe d'organismes qui proviennent tous d'un ancêtre commun.

virus au départ, était largement fondée sur les réactivités sérologiques (croisées) à la protéine de pointe (protéine "spike") virale. Elle se fonde plutôt maintenant sur des analyses de séquences comparatives des protéines réplicatives. Le choix des protéines et les méthodes utilisées pour les analyser ont, là aussi, progressivement évolué depuis le début du 21^e siècle. Cinq domaines essentiels étant les seuls conservés dans tous les virus de l'ordre Nidovirales (3CLpro, NiRAN, RdRp, ZBD et HEL1), ils ont été retenus pour servir à sa constitution.

Depuis 2011, un logiciel (DEMARC pour DivErsity pArtitioning by hieRarchical Clustering) permet de définir les taxons et les rangs. Il s'agit de fixer des critères de démarcation, quelle que soit la taille de l'échantillonnage, autrement dit qu'il s'agisse d'un seul virus ou de centaines d'entre eux. L'algorithme est celui de l'agrégation par saut minimum [2] et la persistance des seuils face à l'augmentation de l'échantillonnage viral est interprétée comme reflétant les forces biologiques et les facteurs environnementaux.

Les virologues estiment que la recombinaison homologue, courante chez les Coronavirus, est limitée dans les régions du génome codant pour les protéines les plus essentielles, telles que celles utilisées pour la classification, ainsi qu'aux membres de la même espèce virale. Cette restriction favoriserait la diversité intra-espèce et contribuerait à la séparation inter-espèces.

Pour faciliter l'utilisation de seuils de rang, ceux-ci sont convertis en distances par paires (pair-wise distances) et exprimés en pourcentage, que les chercheurs utilisent couramment pour arriver à une attribution provisoire d'un virus donné dans une classe de la taxonomie des Coronavirus, après une analyse phylogénétique conventionnelle des virus sélectionnés.

C'est ce protocole qu'on retrouve dans toutes les classifications actuellement proposées.

2 La situation globale de SARS-CoV-2

La génomique comparative précédemment décrite, qui permet de quantifier et de partitionner la variation des protéines réplicatives les plus conservées aboutit à différentes arborescences. Les estimations de distance entre le SARS-CoV-2 et les Co-

2. On sait que son défaut est d'écraser un peu l'ultramétrique, et donc la hauteur de la classification.

ronavirus les plus étroitement apparentés peuvent évidemment varier d'une étude à l'autre en fonction du choix de ce qu'on mesure (les distances entre nucléotides ou acide aminés) et de la région du génome. En conséquence, il n'y a pas encore d'accord sur la position taxonomique exacte du SARS-CoV-2 dans le sous-genre Sarbecoronavirus.

Cependant, la variation génomique des virus connus de l'espèce Coronavirus liée au syndrome respiratoire aigu sévère (SARS) est plus faible que celle d'autres espèces comparativement bien échantillonnées – par exemple, celles prototypées par le MERS-CoV, le Coronavirus humain OC43 (HCoV-OC43) et le virus de la bronchite infectieuse (IBV) – et l'espèce SARS est bien séparée des autres espèces connues de Coronavirus dans l'espace des séquences. Ces deux caractéristiques facilitent l'attribution sans ambiguïté du SARS-CoV-2 à l'espèce SARS et aucune perturbation dans la classification des 2500 espèces de Coronavirus n'est intervenue lorsqu'on l'y a incluse.

Côté intra-spécifique, les distances du SARS-CoV-2 aux autres virus font partie des 25% de distances les plus grandes, la plus importante étant celle entre le SARS-CoV-2 et un isolat du virus d'une chauve-souris africaine (SARSr-CoV BtKY72), ces opposés représentant 2 lignées de base au sein de l'espèce Coronavirus liée au syndrome respiratoire aigu sévère (qui comprend peu de virus connus). De telles relations contrastent avec la ramification peu profonde de la lignée la plus peuplée de l'espèce, qui comprend tous les isolats humains de SARS-CoV collectés lors de l'épidémie de 2002-2003 ainsi que les virus de chauve-souris étroitement apparentés d'origine asiatique identifiés dans la recherche de sources zoonotiques potentielles de cette épidémie.

Evidemment, le petit nombre de Coronavirus liés au SARS et actuellement connus fait que la connaissance de la diversité naturelle de l'espèce peut être biaisée. La considération d'un plus grand nombre de virus avec des analyses génétiques approfondies donnerait sans doute une meilleure vision de l'espèce. On peut malgré tout déjà apporter un certain nombre d'informations sur la situation réciproque de ces virus.

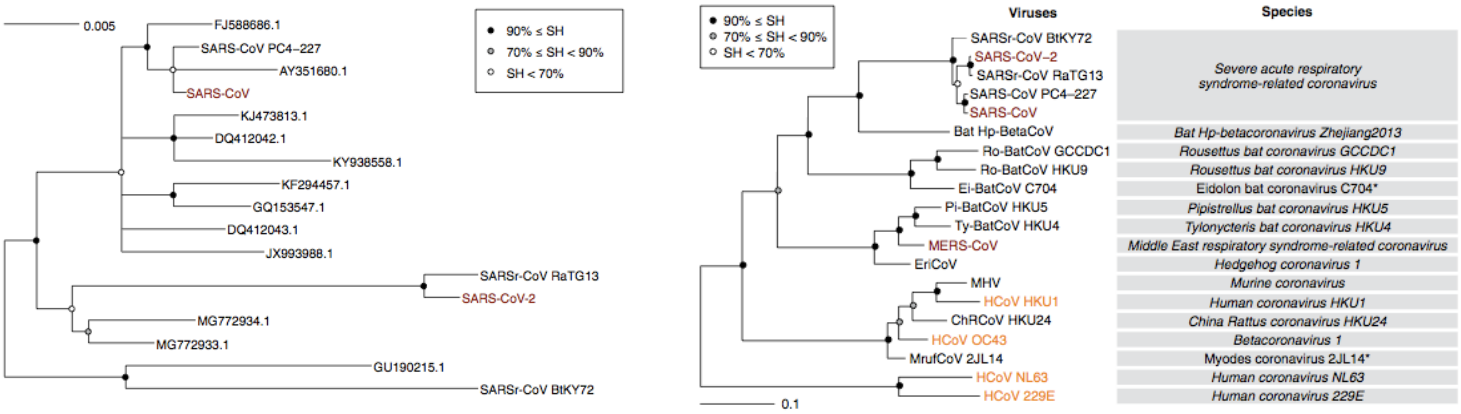


FIGURE 1 – Taxinomie des SARS et des virus apparentés (d'après le CSGICTV)

3 SARS-Cov-2, RaTG-13 et les autres

Contrairement à d'autres virus qui ont émergé au cours des deux dernières décennies, les Coronavirus sont hautement recombinants. Analysant l'histoire évolutive du SARS-CoV-2 à l'aide des données génomiques disponibles sur les Sarbecoronavirus, des chercheurs du Centre d'études de la dynamique des maladies infectieuses de Pennsylvanie (voir [Boni 20]) ont démontré que les Sarbecoronavirus qui circulent chez les chauves-souris "fer à cheval" avaient des antécédents de recombinaison complexes, ce qui n'empêche pas, cependant, d'évaluer leur proximité ou leur éloignement.

Globalement, le SARS-CoV-2 est à 96,2% proche du RaTG-13, ce qui fait de ce dernier le plus proche parent du virus responsable de la pandémie. Ce fait se traduit parfaitement dans les arbres suivants, prenant en compte le domaine N-terminal, le domaine S1 sans la variable loop, et le domaine S2 de ces virus (voir Fig. 2).

Toutefois, malgré l'acquisition par la lignée SARS-CoV-2 de certains éléments, dans son domaine de liaison au récepteur (RBD [3]) de la protéine Spike (S), qui permettent une transmission à l'homme, le récepteur ACE2 et son RBD sont génétiquement plus proches d'un virus de pangolin que de RaTG-13. C'est ce que montre en tout cas l'étude de la sous-région à boucle variable de la protéine S.

3. Receptor Binding Domain.

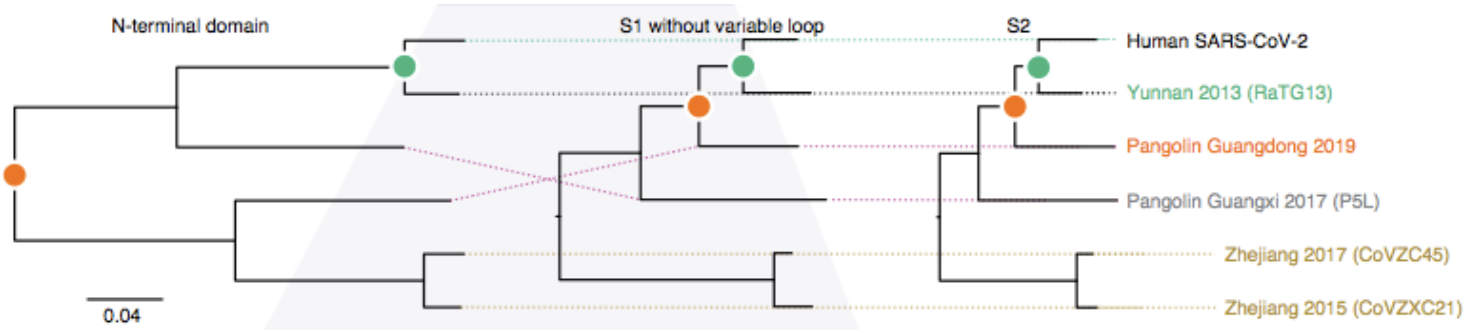


FIGURE 2 – La parenté entre SARS-CoV-2 et RaTG-13 en dehors de la variable loop de S1

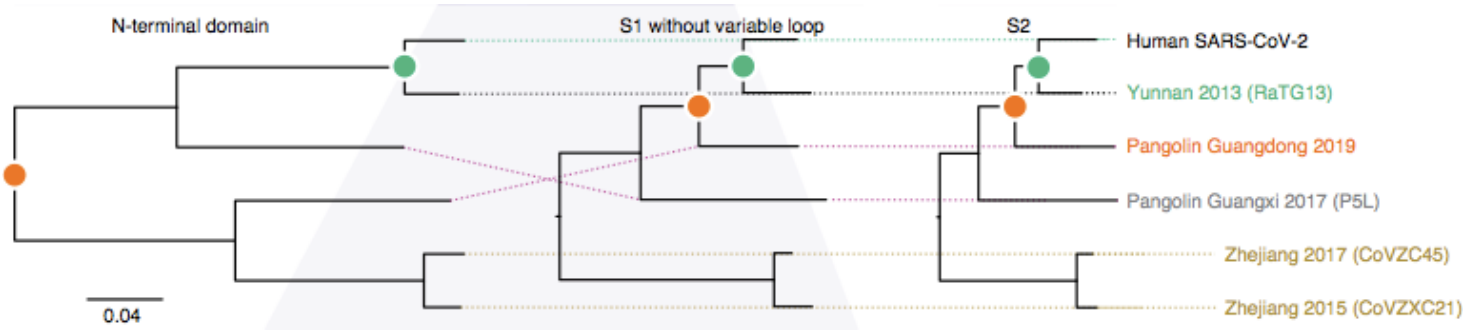


FIGURE 3 – La parenté entre SARS-CoV-2 et RaTG-13 dans la variable loop de S1

Au premier examen, cela suggérerait que le SARS-CoV-2 est un recombinant d'un ancêtre du Pangolin-2019 et de RaTG-13, ce qui a d'ailleurs été soutenu (voir [Xiao 20] ; [Li 20]).

Cependant, à y regarder de plus près, les divergences relatives dans l'arbre phylogénétique montrent que SARS-CoV-2 est peu susceptible d'avoir acquis la boucle variable d'un ancêtre du Pangolin-2019 car ces deux séquences ont approximativement 10 à 15% de divergence dans toute la protéine S (à l'exclusion du domaine N-terminal). C'est RaTG-13 qui est le plus divergent dans la région à boucle variable. Du point de vue génétique, il serait donc susceptible d'être le produit d'une recombinaison lui ayant fait acquérir une boucle variable divergente à partir d'un Sarbecoronavirus de

chauve-souris non échantillonné jusqu'à présent.

Mais voilà le plus remarquable : la région à boucle variable contient les six résidus de contact-clés dans le RBD qui confèrent au SARS-CoV-2 sa spécificité de liaison ACE2. Or ces résidus sont également présents dans la séquence génétique correspondante du Pangolin Guangdong 2019. L'explication la plus économique de la présence partagée de ces résidus consiste donc à dire qu'ils étaient présents chez les ancêtres communs de SARS-CoV-2, de RaTG-13 et du Pangolin Guangdong 2019, et ont été par la suite perdus par recombinaison dans la lignée menant à RaTG-13. Ceci fournit un argument convaincant pour que la lignée SARS-CoV-2 soit la conséquence d'un saut zoonotique direct ou presque direct des chauves-souris à l'homme, car les principaux résidus de liaison ACE2 étaient présents dans les virus circulant chez les chauves-souris.

En utilisant l'approche la plus conservatrice pour l'identification d'une région génomique non recombinante (NRR1), on peut penser que le SARS-CoV-2 forme une lignée sœur avec RaTG-13, et se trouve lié à des lignées cousines de Coronavirus génétiquement apparentées, échantillonnées chez des pangolins dans les provinces du Guangdong et du Guangxi (Fig. 3). Étant donné que ces virus de pangolin sont des prédécesseurs de l'ancêtre de la lignée RaTG-13/SARS-CoV-2, il est donc probable qu'ils se soient combinés avec des virus de chauves-souris. Bien que les pangolins puissent agir en tant qu'hôtes intermédiaires pour que les virus de chauve-souris pénètrent chez l'homme, il n'y a cependant aucune preuve que l'infection par le pangolin soit une condition nécessaire pour que les virus de chauve-souris se transmettent aux humains.

Les auteurs de l'étude que nous commentons (voir [Boni 20] et al.) sont également parvenus à relier certaines sous-régions génétiques des Sarbecoronavirus à des zones géographiques précises de la Chine, aboutissant ainsi à deux clades phylogénétiques. Un clade géographique comprend des virus provenant de provinces du sud (Guangxi, Yunnan, Guizhou et Guangdong), tandis que son principal clade frère est constitué de virus provenant de provinces du nord (Shanxi, Henan, Hebei et Jilin), ainsi que de la province du Hubei dans le centre de la Chine et de la province du Shaanxi dans le nord-ouest. Plusieurs séquences recombinantes dans ces arbres montrent néanmoins que des événements de recombinaison se produisent dans l'ensemble de la géographie. Le virus du Sichuan (SC2018) semble être ainsi un recombinant des virus du nord, du centre et du sud, tandis que les deux virus du Zhejiang (CoVZXC21 et CoVZC45) semblent porter une région recombinante provenant du sud ou du centre de la Chine.

Des estimations de temps de divergence ont aussi été menées par les auteurs. En utilisant l'approche la plus prudente, l'estimation de la date de divergence pour le SARS-CoV-2 et le RaTG-13 donnerait l'année 1969, tandis que celle entre le SARS-CoV et sa séquence de chauves-souris la plus étroitement liée remonterait à 1962. On notera que ces dates, sont plus anciennes que d'autres estimations obtenues à l'aide d'une collection de génomes du SARS-CoV d'hôte humain et de civette (ainsi que de quelques génomes de chauve-souris étroitement liés), ce qui implique que les taux d'évolution étaient principalement informés par l'échelle d'épidémie de SARS à court terme et probablement biaisés à la hausse.

L'idée générale est donc que les Sarbecoronavirus circulent dans une variété d'espèces de chauves-souris "fer à cheval" dont les aires de répartition se chevauchent largement et que des virus étroitement liés au SARS-CoV-2 sont portés par ces chauves-souris depuis de nombreuses décennies. La diversité non échantillonnée issue de l'ancêtre commun du SARS-CoV-2/RaTG-13 forme un clade de Sarbecoronavirus de chauve-souris aux propriétés généralistes – en ce qui concerne en tout cas leur capacité à infecter une gamme de cellules de mammifères – qui a pu faciliter son saut vers l'homme et pourrait le faire à nouveau.

Il reste que le RBD humain compatible ACE2 – que les auteurs supposent présent dans toute une lignée de Sarbecoronavirus de chauve-souris et qui aurait finalement conduit au SARS-CoV-2 – n'a jusqu'à présent été trouvée que dans quelques virus pangolins. En outre, l'autre caractéristique-clé considérée comme déterminante dans la capacité du SARS-CoV-2 à infecter les humains – une insertion de site de clivage polybasique dans le protéine S – n'a pas encore été observée chez un autre proche parent du virus SARS-CoV- 2 de la chauve-souris. Ces deux faits ne laissent pas d'être assez mystérieux et conduisent à admettre la thèse d'une évolution naturelle d'un virus de chauve-souris vers SARS-CoV-2 sur des décennies avec une certaine réserve.

4 La question de la recombinaison

Un coup d'œil sur la répartition des différents virus de chauve-souris, de pangolin en Chine, incluant le virus SARS-CoV-2 humain, montre la disposition suivante (voir Fig. 4), que nous empruntons à une étude menée par des chercheurs de Hong-Kong (voir [Lau 20]).

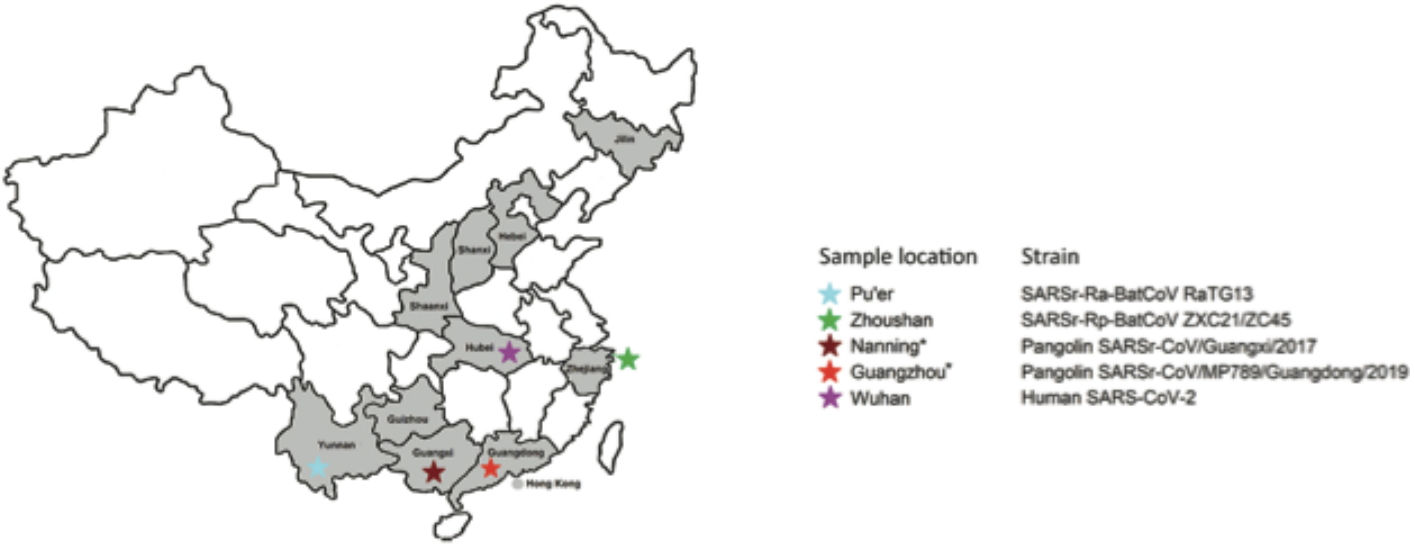


FIGURE 4 – Répartition des virus de chauve-souris et de pangolin en Chine

A priori, le virus RaTG-13 de la mine de Mojiang (Yunnan) n'est pas tout proche (au plan géographique) de celui du pangolin du Guandong, la distance Mojiang-Guangzhou avoisinant les 1500 km à vol d'oiseau.

Il est vrai – l'étude précédente (voir [Boni 20]) l'a montré – que les virus sont fortement recombinants, avec des clades de recombinaisons qu'on peut mettre en correspondance avec la géographie de la Chine. Cela donne la carte suivante (voir Fig. 5), à poser en regard de la précédente :

Ainsi qu'on l'a vu précédemment, à quelques exceptions près, les recombinaisons se font plutôt, comme on pouvait évidemment s'y attendre, de proche en proche. Or si le Yunnan et le Guandong sont séparés par le Guangxi, ce sont tout de même des régions du sud de la Chine. Et même si, à la dimension de ce pays, les distances sont importantes. il est tout de même très surprenant que l'on ne trouve précisément aucun virus de type SARS-CoV-2 dans ces régions, et que le seul qu'on trouve se situe complètement ailleurs, à Wuhan, dans la Chine du centre. Les villes de Mojiang, Guangzhou et Wuhan forment les sommets d'un triangle rectangle dont l'hypothènuse est la distance Mojiang-Wuhan (1800 km), en l'occurrence le plus court chemin de transport du virus – et l'on sait précisément que des échantillons ont précisément

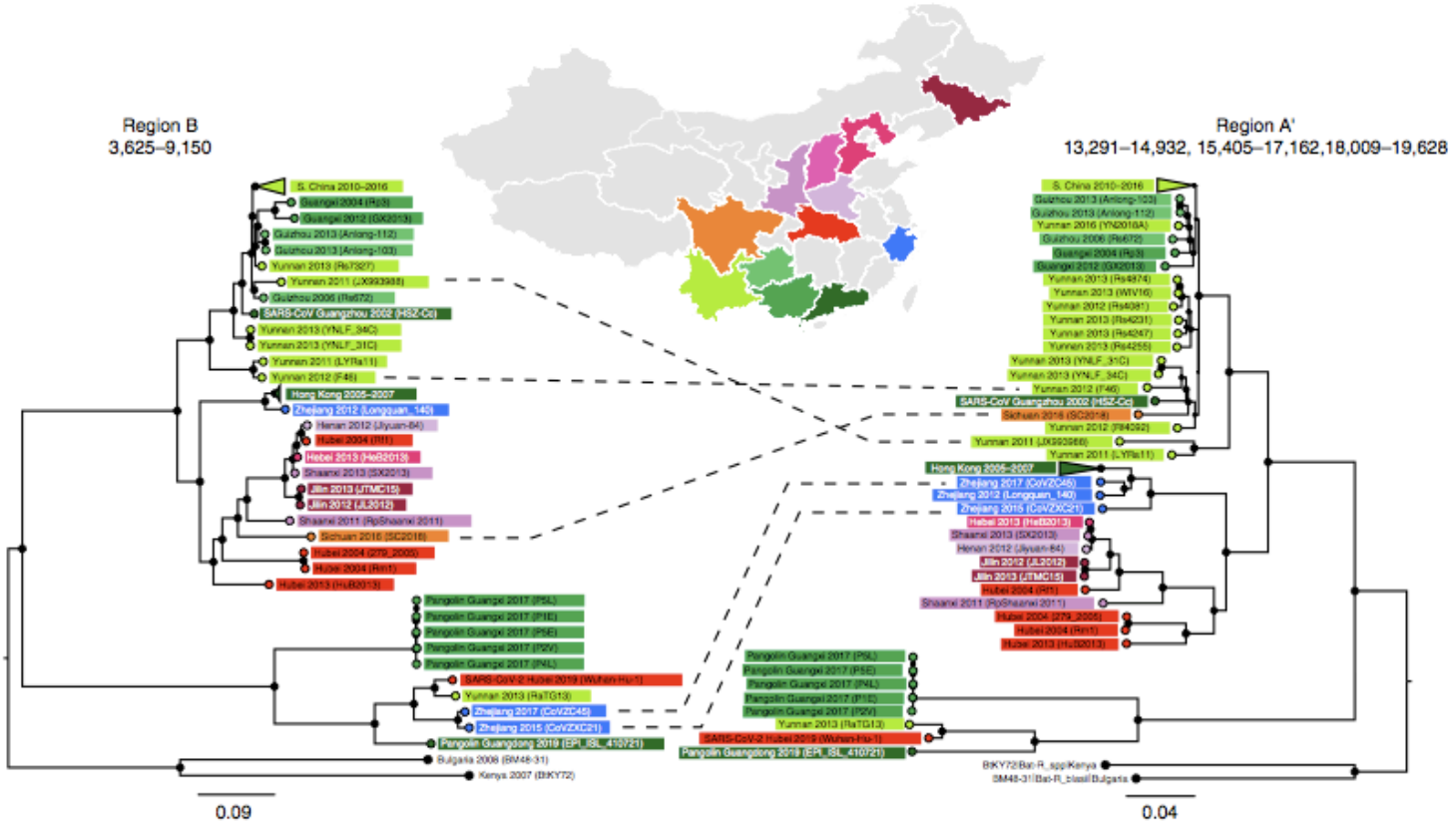


FIGURE 5 – Répartition des virus de chauve-souris et de pangolin en Chine

emprunté ce chemin pour aller au laboratoire de virologie de Shi Zhengli à Wuhan (le Wuhan Institute of Virology ou WIV), où ils ont été étudiés.

Si recombinaison il y a eu, il faut donc tenir compte de ces faits pour, éventuellement, pouvoir l'expliquer. Mais, pour l'instant, malgré plus d'un an d'enquête, aucun virus recombinant à la fois des caractéristiques de RaTG-13 (du Yunnan) et du virus du pangolin (du Guandong) n'a été trouvé dans la nature au sein des états du sud ou des régions limitrophes, de sorte que le virus de Wuhan semble être, pour ainsi dire, sorti d'un chapeau.

Aussi bien, jusqu'à présent, l'espoir porté par Andersen (et al.) (voir [Andersen 20]) de trouver des séquences virales apparentées à SARS-CoV-2 à partir de sources animales n'a pas été satisfait. "Par exemple, écrivaient-ils, une observation future d'un site de clivage polybasique intermédiaire ou entièrement formé dans un virus de type SRAS-CoV-2 provenant d'animaux apporterait un soutien encore plus grand aux

hypothèses de sélection naturelle. Il serait également utile d'obtenir davantage de données génétiques et fonctionnelles sur le SARS-CoV-2, y compris à partir d'études sur les animaux. L'identification d'un hôte intermédiaire potentiel du SARS-CoV-2, ainsi que le séquençage du virus à partir de cas très précoces, seraient également très instructifs". Aucun de ces souhaits n'a pu jusqu'ici être satisfait, et on peut douter qu'ils le soient un jour.

Une autre étude, qui confirme la parenté de SARS-CoV-2 et de RaTG-13, fait observer que, si SARS-CoV-2 provient de RaTG-13, l'augmentation de la substitution synonymique entre la souche SARS-CoV-2 et RaTG13 suggère que le génome de SARS-CoV-2 semble soumis à une sorte de sélection purifiante de type négatif. En revanche, certains sites dans la protéine de pointe connaissent une sélection positive. Ces observations peuvent évidemment suggérer que l'évolution adaptative pourrait avoir contribué à ces changements. Cependant, remarquent les auteurs, "comme l'ARN mutagène (par exemple 5-FU), facile à trouver en kit, pourrait induire le même modèle de mutation, le mécanisme du modèle de mutation observé entre SARS-CoV-2 et RaTG-13 devrait être étudié plus avant dans les travaux futurs" (voir [Longxian 20]).

Rien n'est donc dit, finalement, concernant l'origine naturelle ou (partiellement) artificielle de SARS-CoV-2. Mais si le Yunnan (potentiellement ou non accompagné du Guandgong) est la lointaine origine du virus de Wuhan, alors il faut expliquer pourquoi l'ensemble des mutations (naturelles comme artificielles s'il en a existé) a conduit à l'apparition du virus précisément à Wuhan et pas ailleurs. Dans tous les cas, l'explication la plus simple est encore qu'il ait été transporté du Sud de la Chine en cet endroit par un moyen artificiel quelconque, car ce ne peut être ni le vol d'une chauve souris, ni la course d'un pangolin, et encore moins le vent qui peuvent expliquer ce voyage. Il faudra en outre comprendre comment il a pu se répandre dans l'environnement local (aucun des animaux du fameux marché de Wuhan n'était infecté du SARS-Cov-2 et certaines personnes infectées n'avaient jamais mis les pieds dans ce marché). L'analyse en réseau peut aider à matérialiser cette possibilité.

5 Analyse en réseau, diffusion de l'épidémie et rétrodiction

La méthode de construction d'arbres phylogénétique ne facilitant pas toujours une interprétation sans ambiguïté des données, on a eu recours, dès les années 1990,

à des méthodes complémentaires. La construction de réseaux phylogénétiques – et non plus de simples arbres – a l'avantage de permettre la visualisation simultanée d'une multitude d'arbres optimaux. Parmi ses succès, on peut noter la reconstitution des mouvements de population préhistoriques ayant colonisé la planète (4, 5) ainsi que la reconstruction de la préhistoire du langage (6). L'application de cette approche aux données virologiques dans le cas du Coronavirus a donné lieu en mars 2020, a une publication (voir [Forster 20]) fondée sur la base de données GISAID, contenantune compilation de 253 génomes complets et partiels du coronavirus SARS-CoV-2 recueillis depuis décembre 2019. Pour comprendre l'évolution de ce virus chez l'homme, et pour aider à tracer les voies d'infection et à concevoir des stratégies préventives, les auteurs ont présenté un réseau phylogénétique de 160 génomes du SARS-Cov-2 en grande partie complets. Zhou et coll. (7) ayant récemment signalé un coronavirus de chauve-souris étroitement apparenté, avec une similitude de séquence de 96,2% avec le virus humain, ce virus de la chauve-souris a été utilisé comme un groupe externe, la racine du réseau étant placée dans un groupe de lignées appelé «A». Dans l'ensemble, le réseau, comme prévu dans une épidémie en cours, montre des génomes viraux ancestraux existant à côté de leurs génomes filles nouvellement mutés.

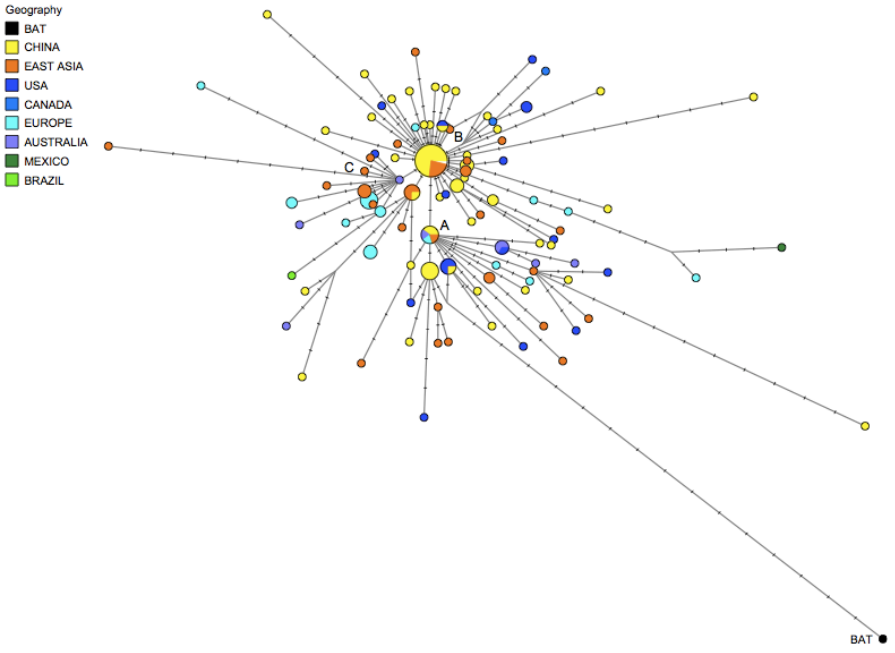


FIGURE 6 – Réseau phylogénétique de 160 génomes de SARS-CoV-2

Selon le commentaire des auteurs, "il existe deux sous-groupes de A qui se distinguent par la mutation synonyme T29095C. Dans le sous-groupe des allèles T, quatre individus chinois (de la province côtière du sud de la Chine du Guangdong) portent le génome ancestral, tandis que trois patients japonais et deux patients américains en diffèrent par un certain nombre de mutations. Ces patients américains auraient eu des antécédents de résidence dans la source présumée de l'épidémie à Wuhan. Le sous-cluster de l'allèle C possède des branches mutationnelles relativement longues et comprend cinq individus de Wuhan, dont deux sont représentés dans le nœud ancestral, et huit autres Asiatiques de l'Est de Chine et des pays adjacents. Il est à noter que près de la moitié (15/33) des types de ce sous-cluster, cependant, se trouvent en dehors de l'Asie de l'Est, principalement aux États-Unis et en Australie. Deux nœuds de réseau dérivés sont frappants en termes de nombre d'individus inclus dans le type nodal et dans les branches mutationnelles rayonnant à partir de ces nœuds. Nous avons étiqueté ces clusters phylogénétiques B et C."

Pour le type B, tous sauf 19 des 93 génomes ont été échantillonnés à Wuhan, dans d'autres parties de l'est de la Chine et, sporadiquement, dans les pays asiatiques adjacents. Une dizaine seulement ont été trouvés dans des génomes viraux de pays non-asiatiques (États-Unis, Canada, Mexique, France, Allemagne, Italie ou Australie). Le nœud B est dérivé de A par deux mutations : la mutation synonyme T8782C et la mutation non synonyme C28144T changeant une leucine en sérine. Le groupe B est frappant en ce qui concerne la longueur des branches mutationnelles (laquelle ne semble pas être due à un décalage temporel). Un scénario fondateur complexe est une possibilité, mais une autre explication à considérer serait que le virus ancestral de type B de Wuhan est immunologiquement ou écologiquement adapté à une grande partie de la population d'Asie de l'Est, et peut avoir besoin de muter pour vaincre la résistance. en dehors de l'Asie de l'Est.

Selon les auteurs également, "le type C diffère de son type parent B par la mutation non synonyme G26144T qui transforme une glycine en valine. Dans l'ensemble de données, il s'agit du principal type européen (n = 11), avec des représentants en France, en Italie, en Suède et en Angleterre, ainsi qu'en Californie et au Brésil. Il est absent de l'échantillon de la Chine continentale, mais évident à Singapour (n = 5) et également trouvé à Hong Kong, à Taiwan et en Corée du Sud."

Une application pratique du réseau phylogénétique est de reconstruire les voies d'infection là où elles sont inconnues et présentent un risque pour la santé publique. Ainsi, le 25 février 2020, le premier Brésilien aurait été infecté à la suite d'une visite en Italie, ce que l'algorithme du réseau reflète bien avec un lien mutationnel entre un italien et son génome viral brésilien dans le cluster C. Dans un autre cas, un homme

de l'Ontario avait voyagé de Wuhan dans le centre de la Chine à Guangdong dans le sud, puis est rentré au Canada, où il est tombé malade. Son génome viral se ramifie à partir d'un nœud ancestral reconstruit, avec des variantes virales dérivées à Foshan et Shenzhen (villes toutes deux dans le Guangdong, en accord avec ses antécédents de voyage. D'autres Américains du nord seront infectés à partir de là. Le cas du génome viral mexicain unique dans le réseau est une infection documentée, diagnostiquée le 28 février 2020 chez un voyageur mexicain en Italie. Le réseau confirme l'origine italienne du virus mexicain mais montre aussi que ce virus italien dérive de la première infection allemande documentée le 27 janvier 2020 chez un employé travaillant à Munich. Celui-ci, à son tour, avait contracté l'infection d'un collègue chinois de Shanghai qui avait reçu lui-même du fait de la visite de ses parents venus de Wuhan. Ce voyage viral de Wuhan au Mexique, d'une durée d'un mois, est documenté par 10 mutations dans le réseau phylogénétique.

Ce réseau viral est donc un instantané des premiers stades de la pandémie. La question se pose évidemment de savoir si, reparcourant le réseau à l'envers (rétrodiction) ; l'enracinement de l'évolution virale peut être obtenu à ce stade précoce en utilisant comme racine le génome échantillonné le plus ancien disponible. Comme le font remarquer les auteurs, le premier génome du virus qui a été échantillonné le 24 décembre 2019 est déjà éloigné de la racine associée à l'exogroupe A lié au virus de chauve-souris de Mojiang (RaTG-13).

Le travail de Forster et al. a été fortement critiqué, au motif que le réseau en question n'est ni phylogénétique, ni évolutif : non seulement un tel réseau ne reflète pas les caractéristiques biologiques importantes censées sous-tendre l'évolution virale, telles que la recombinaison et le transfert horizontal de gènes, mais, n'étant pas orienté, il ne peut pas désigner une racine ni polariser la transformations des caractères. Par exemple, le groupe externe ne peut pas s'enraciner en A, car A lui-même peut être dérivé de l'un des deux virus ancestraux possibles. Quant à l'option de l'algorithme utilisée pour l'enracinement, elle relie simplement la séquence «outgroup» (c'est-à-dire, en l'occurrence, le coronavirus de chauve-souris RaTG-13 de Mojiang) à la séquence la plus similaire du réseau «endogroup» déjà produite. On ne peut donc pas en tirer comme conséquence que A est l'origine de la pandémie (voir [Sánchez 20]). Outre cet enracinement potentiellement incorrect, on a pu également reprocher aussi à l'article un biais d'échantillonnage (voir [Mavian 20]). Les auteurs contestent que RaTG-13 puisse être un ancêtre de SARS-CoV-2 au motif que son identité de séquence avec ce dernier n'est que de 96,2%, "ce qui implique que ces génomes viraux (qui ont près de 30 000 nucléotides de long) diffèrent de plus de 1 000 mutations. Il est peu probable, affirment-ils, qu'un groupe externe aussi éloigné fournisse une

racine fiable pour le réseau. Le faux enracinement du réseau viendrait du fait que la branche du virus de la chauve-souris n'a que 16 ou 17 mutations de longueuret qu'un virus de Wuhan de la semaine 0 (24 décembre 2019) est décrit comme un descendant d'un clade de virus collectés dans les semaines 1 à 9 (vraisemblablement de beaucoup d'endroits en dehors de la Chine), ce qui n'a aucun sens évolutif ou épidémiologique. La distinction des trois types de virus est également contestée car les mutations synonymes ou non qui permettent de séparer A et C ou B et C et ne concernent que quelques nucléotides ou protéines, ce qui peut correspondre à des mutations aléatoires pouvant être propagées sans qu'elles soient sélectionnées ou avantageuses, les virus ayant coutume de muter sans arrêt. Par ailleurs, le fait que les séquences du SRAS-CoV-2 présentent un certain regroupement géographique n'est pas nouveau, mais ne peut être utilisé comme une preuve de différences biologiques à moins d'être étayé par des données expérimentales solides. Or les résultats de Forster et al. sont basées sur un ensemble de données non représentatives de 160 génomes, sans corrélation significative entre la prévalence des cas confirmés et le nombre de souches séquencées par pays. Leur «réseau est supposer tracer fidèlement les voies d'infection pour les cas documentés de COVID-19, mais, en réalité, il ne prend pas en compte la diversité virale manquante, n'évalue pass plusieurs hypothèses de transmission qui seraient cohérentes avec les données de séquence, et ne fournit aucun argument sur la robustesse du schéma de branchement dans leur réseau. En fin de compte, aucune conclusion ferme ne peut donc être tirée sans évaluer la probabilité de voies de diffusion alternatives.

6 Conclusion

Il est prouvé que la séquence du RBD de SARS-CoV-2 n'est pas le produit d'une sélection positive, mais plutôt d'une recombinaison (voir [Cagliani 20]). Reste à savoir, cependant, s'il s'agit d'une recombinaison naturelle, utilisant un virus de pangolin malais, comme cela a été suggéré ([Liu 19] ; [Wong 20]), ou s'il s'agit d'autre chose [4].

Vu les incertitudes liées aux différentes méthodes destinées à préciser la nature et la situation du SARS-CoV-2 (qu'il s'agisse de classification automatique, d'arbres phylogénétiques ou d'étude en réseaux), la présente situation met en évidence de façon cruciale l'absence d'une théorie générale des classifications (voir [**?**]), qui permettrait

4. Nous avons exprimé nos doutes et nos hypothèses sur les origines possibles de SARS-CoV-2 dans un autre article (voir [**?**]).

de placer le SARS-CoV-2 dans un contexte parfaitement objectif, selon des critères biogénétiques univoques et indiscutables. Au lieu de ça, nous sommes livrés à des hypothèses invérifiables ou des interprétations discutables, le seul fait incontestable étant l'étrange adaptation à l'homme du virus.

De ce point de vue, la conclusion du rapport de l'OMS (World Health Organization ou WHO en anglais) selon laquelle une origine du virus par fuite d'un laboratoire est "hautement improbable" (voir [WHO 21[1]] ; 118) est, pour le moins, assez prématurée. Elle a d'ailleurs été relativisée par la déclaration du directeur de cette organisation, qui a réclamé des études plus approfondies (voir [WHO 21[2]]). En l'absence de toute possibilité de remonter de façon fiable à l'origine de la pandémie, même par des méthodes modernes d'analyses, le mystère demeure sur la provenance du SARS-CoV-2, même si la Chine semble un foyer évident.

Références

[Andersen 20] Rambaut , A., Lipkin, W. I., Holmes , E. C., et Garry, R. F., "The proximal origin of SARS-CoV-2", *Nature Medicine*, Vol. 26, 450-455, Apri 2020.

[Boni 20] Boni, M. F., Lemey, P., Jiang, X., Lam, T. T.-Y., Perry, B., Castoe, T., Rambaut, A. and Robertson D. L., "Evolutionary origins of the SARS-CoV-2 sarbecovirus lineage responsible for the COVID-19 pandemic", *Nature Microbiology*, 5 :1408-1417, 2020.

[Cagliani 20] Cagliani, R., Forni, D., Clerici, M. Sironia, M., "Computational Inference of Selection Underlying the Evolution of the Novel Coronavirus, Severe Acute Respiratory Syndrome Coronavirus 2", *Journal of Virology*, Volume 94, Issue 12, e00411-20, Juin 2020.

[CSJCTV 20] Coronaviridae Study Group of the International Committee on Taxonomy of Viruses (Gorbalyena, A. E. et al.), "The species Severe acute respiratory syndrome-related Coronavirus : classifying 2019-nCoV and naming it SARS-CoV-2", *Nature Microbiology*, Vol. 5, Mars 2020, 536-544 | Consensus statement : https ://doi.org/10.1038/s41564-020-0695-z.

[Dagognet 70] Dagognet, F., *Le catalogue de la Vie*, P. U. F., 1970.

[Forster 20] Forster, P., Forster, L., Renfrew, C., Forster, M., "Phylogenetic network analysis of SARS-CoV-2 genomes", *Proc. Natl. Acad. Sci. U S A*, 28 avril 2020, 9241-9243, doi : 10.1073/pnas.2004999117. Epub 2020 Apr 8.

[Lau 20] Susanna K.P. Lau, S. K. P., Luk, H. K. H., Wong, A. C. P., Li, K. S. M., Zhu, He, L. Z., Fung, J., Chan, T. T. Y., Fung, K. S. C., Woo, P. C . Y., "Possible Bat Origin of Severe Acute Respiratory Syndrome Coronavirus 2", *Emerging Infectious Diseases*, www.cdc.gov/eid, Vol. 26, No. 7, July 2020.

[Longxian 20] Longxian, L.V., Gaolei Li, Jinhui Chen, Xinle Liang, Yudong Li, "Comparative genomic analysis revealed specific mutation pattern between human coronavirus SARS-CoV-2 and Bat-SARSr-CoV RaTG13", *Biorxiv*, Mars, 2, 2020 : https ://doi.org/10.1101/2020.02.27.969006 ;

[Li 20] Li, X. et al., "Emergence of SARS-CoV-2 through recombination and strong purifying selection", *Sci. Adv.* 6, eabb9153, 2020.

[Liu 19] Liu P., Chen W, Chen J.P.., "Viral Metagenomics Revealed Sendai Virus and Coronavirus Infection of Malayan Pangolins (Manis javanica)", *Viruses*,11(11) : 979, 2019.

[Mavian 20] Mavian C., Pond S. K., Marini S. et al., "Sampling bias and incorrect rooting make phylogenetic network tracing of SARS-COV-2 infections unreliable", *Proc Natl Acad Sci U S A*, 117(23), 12522-12523, 9 juin 2020 ; doi : 10.1073/pnas.2007295117. Epub 2020 May 7.

[Parrochia 13] Parrochia, D., Neuville, P., *Towards a general theory of classifications*, Bâle, Birkhaüser, 2013.

[Parrochia 21] Parrochia, D., "Sur les origines possibles de l'épidémie CoVid-19", https ://hal.archives-ouvertes.fr/hal-03189187, 10 avril 2021.

[Sánchez 20] Sánchez-Pacheco, S. , Kong S., Pulido-Santacruz, P., Murphy, R. W. et Kubatko, L., "Median-joining network analysis of SARS-CoV-2 genomes is neither phylogenetic nor evolutionary", *PNAS* 117 (23), 12518-12519, juin 2020 ; publié pour la première fois le 7 mai 2020 ; https ://doi.org/10.1073/pnas.2007062117.

[Siddell 19] Siddell, S. G. et al., "Additional changes to taxonomy ratified in a special vote by the International Committee on Taxonomy of Viruses (October 2018)", *Arch. Virol.*, 164, 943-946, 2019.

[WHO 21[1]] *WHO-convened Global Study of Origins of SARS-CoV-2 : China Part*, Joint WHO-China Study, 14 January-10 February 2021, Joint Report, https ://www.who.int/health-topics/coronavirus/origins-of-the-virus

[WHO 21[2]] Ghebreyesus T. A., *WHO Director-General's remarks at the Member State Briefing on the report of the international team studying the origins of SARS-CoV-2*, 30 March 2021 : https ://www.who.int/director-general/speeches/detail/who-director-general-s-remarks-at-the-member-state-briefing-on-the-report-of-the-international-team-studying-the-origins-of-sars-cov-2.

[Wong 20] Wong, M. C., Javornik Cregeen, S. J., Ajami, N. J., Petrosino, J. F., "Evidence of recombination in coronaviruses implicating pangolin origins of nCoV-2019", doi : https ://doi.org/10.1101/2020.02.07.93920, February 13, 2020.

[Xiao 20] Xiao, K. et al., "Isolation of SARS-CoV-2-related Coronavirus from Malayan pangolins", *Nature* 583, 286-289, 2020.

[Ziebuhr 19] Ziebuhr, J. et al., "Proposal 2019.021S.Ac.v1 : Create ten new species and a new genus in the subfamily Orthocoronavirinae of the family Coronaviridae and five new species and a new genus in the subfamily Serpentovirinae of the family Tobaniviridae", *ICTV*, 2019 : https ://ictv.global/proposal/2019.Nidovirales/.