INSTITUT DE FRANCE
Académie des sciences

Origine et évolution du SARS-Cov-2

11 juin 2020

Séquençage intensif, modélisation et analyses bioinformatiques

Le SARS-CoV-2 est un virus à ARN. Son génome comporte environ 30 000 bases, ce qui en fait le plus long des génomes de virus à ARN connus. Par comparaison, le génome de la grippe a une longueur de 10 à 15000 bases, et le VIH (un rétrovirus) une longueur d'environ 10 000 bases. Les premières séquences du génome du SARS-CoV-2 ont été disponibles fin décembre, toutes issues de Wuhan (Chine), la première le 23/12, la seconde le 26/12, puis 16 autres le 30/12. A partir de mi-janvier a commencé le séquençage hors Chine, et fin janvier on disposait d'environ 200 séquences issues de nombreux pays (Thaïlande, Népal, Japon, Canada, Etats-Unis, France, Allemagne, Italie, Australie…). Le séquençage a d'abord été assez lent, puisque fin Mars on ne disposait toujours que de quelques centaines de séquences. Il s'est considérablement accéléré en Avril avec le confinement massif à la surface du globe et les dizaines puis centaines de milliers de cas déclarés en Europe et aux Etats-Unis. Certains jours d'Avril-Mai plusieurs milliers de nouvelles séquences ont été déposées sur le site du GISAID (*Global Initiative on Sharing Avian Influenza Data*, www.gisaid.org/), qui recueille et rend publique les séquences du monde entier. Aujourd'hui (25 Mai) il y a plus de 30 000 séquences sur ce site, et des sites nationaux se mettent en place (Royaume Uni, Canada), si bien qu'il est difficile de faire un décompte exact. Cependant certains pays semblent avoir peu séquencé (ou gardé leurs séquences sans les rendre public), l'Italie par exemple avec moins de 100 génomes disponibles sur le GISAID malgré l'impact de la pandémie. La France n'a pas non plus fourni beaucoup de séquences, notamment en comparaison du Royaume Uni (environ 400 versus 15000 sur le GISAID aujourd'hui). En particulier, on ne dispose que de très peu de séquences du Grand Est, alors que c'est un foyer majeur. Il est essentiel de corriger cette faiblesse du système français en matière de données génomiques d'intérêt médical, qu'on retrouve sur d'autres virus (VIH par exemple) et en général.

L'ensemble des résultats sur l'origine et l'évolution du SARS-CoV-2 vient de l'analyse informatique de ces séquences, couplée aux métadonnées associées comme la date et le lieu de séquençage, la technique de séquençage, etc. Dans certains cas on a pu faire du suivi de contacts et déterminer quelle était l'origine précise d'un virus trouvé à un endroit donné. C'est par exemple le cas pour les deux premiers génomes Thaïlandais, dont on a montré en suivant les itinéraires des patients qu'ils venaient de Chine. On a aussi pu retracer l'origine des premiers cas en Italie comme venant de Munich et avant de Chine, à l'occasion d'un congrès scientifique. Mais ces informations sont très partielles. Pour l'essentiel nos connaissances viennent de l'analyse des séquences, qui se base sur des modèles (par exemple pour décrire les mutations, leur rythme et leur régularité au cours du temps), ainsi que sur des algorithmes, aujourd'hui

confrontés à des masses et flux de données hors normes. On doit garder en tête que les conclusions que l'on tire de ces analyses dépendent des modèles, des approximations et de ces algorithmes généralement heuristiques, du fait de la masse de données et la complexité des problèmes. L'évolution du virus ayant commencé sous nos yeux il y a environ 6 mois, les souches montrent encore peu de différences ce qui limite certaines analyses, par exemple la recherche de recombinants ou de traces d'adaptation. Finalement, certaines analyses sont compliquées par le faible séquençage de certaines régions du monde, comme l'Italie voire la France (cf. ci-dessus) et la qualité de certaines séquences. Malgré ces limites, on a aujourd'hui des réponses très claires sur un certain nombre de questions, par exemple sur l'origine naturelle du virus et le fait qu'il n'est pas issu d'un laboratoire.

Origine et phylogénie

Dès l'obtention des premières séquences on a construit des phylogénies pour retrouver l'origine du SARS-CoV-2 (qui ne s'appelait pas encore ainsi et est nommé hCoV-19 par certains). C'est un Betacoronavirus, membre des Sarbecovirus, sous-genre viral incluant le virus responsable de l'épidémie du SRAS en 2003, et nommé SARS-Cov-1 (pour *severe acute respiratory syndrome-related coronavirus*). Cette famille infecte non seulement les humains, mais de nombreux mammifères, notamment les chauves-souris et les pangolins. La phylogénie de ce virus et ses variants (Fig. 1) montre que :

- Les virus connus les plus proches du SARS-CoV-2 viennent de deux chauves-souris Rhinolophe ou « fer à cheval », trouvées dans le Yunnan en 2013 et 2019. L'identité entre les génomes est d'environ 96% pour l'une (RaTG13) et 93% pour l'autre (RmYN02), mais ce taux d'identité varie le long du génome. En particulier, il est assez faible (60%-70%) dans le région RBD (Region Binding Domain, ~60 acides aminés) de liaison à la protéine humaine ACE2, qui permet l'entrée dans la cellule hôte.
- Plus éloigné globalement (90% d'identité), se trouve un virus de pangolin, dont la région RBD est à l'inverse très proche du SARS-CoV-2, avec une seule mutation en acide aminé, contre une douzaine pour la chauve-souris.
- Toutes les autres souches apparentées au SARS-CoV-2 sont beaucoup plus éloignées, notamment le SARS-CoV-1 (80% d'identité).

Les analyses montrent qu'au sein des Sarbecovirus, les recombinaisons sont nombreuses. Celles-ci sont très possiblement à l'origine du SARS-CoV-2, mais à ce jour on ne peut l'affirmer car dans les réservoirs naturels on n'a pas trouvé de génomes ou portions significatives de génomes qui soient très proches de la forme humaine. Dans un premier temps, on a suggéré qu'en raison des similarités observées (cf. ci-dessus), le SARS-CoV-2 serait un recombinant dans la région RBD de virus de chauve-souris et de pangolin. Mais depuis, l'observation de l'évolution de cette région dans les souches humaines a montré que celle-ci mute rapidement, avec une vingtaine de mutations en acide aminé présentes parmi les souches

humaines disponibles aujourd'hui (25 Mai 2020). L'hypothèse alternative d'une adaptation du virus de chauve-souris dans cette région, plutôt qu'une recombinaison avec le virus de pangolin, est donc tout à fait crédible, et la question n'est pas tranchée.

On a beaucoup lu que le SARS-CoV-2 est issu de la chauve-souris et on tend à penser que le passage à l'humain est récent. En réalité il y a 4% de différences entre les deux génomes, soit environ 1200 mutations. Depuis Décembre 2019 on voit évoluer le virus chez l'humain. En se basant sur le nombre de mutations observées aujourd'hui par rapport aux toutes premières séquences, et en rapportant ce nombre au temps écoulé, on trouve que les génomes évoluent au rythme d'une ou deux mutations par mois, ce qui est lent si on compare à la grippe ou au VIH. Les 1200 mutations observées correspondent donc à 50 à 100 ans d'évolution et une date comprise entre 1970 et 1995 pour l'ancêtre commun de SARS-CoV-2 et RaTG13. Des analyses Bayésiennes plus poussées indiquent des dates analogues voire plus anciennes, avec une part d'incertitude recouvrant tout le 20ème siècle. Quoiqu'il en soit ce résultat indique et confirme qu'entre cet ancêtre commun et le SARS-CoV-2 il y a eu de nombreux intermédiaires, chez la chauve-souris ou d'autres mammifères comme le pangolin, et que ces intermédiaires restent à découvrir. Les coronavirus ayant franchi trois fois de manière marquante la barrière d'espèces au cours des 20 dernières années, il est probable qu'ils le fassent à nouveau, d'où l'importance de rechercher ces réservoirs animaux.

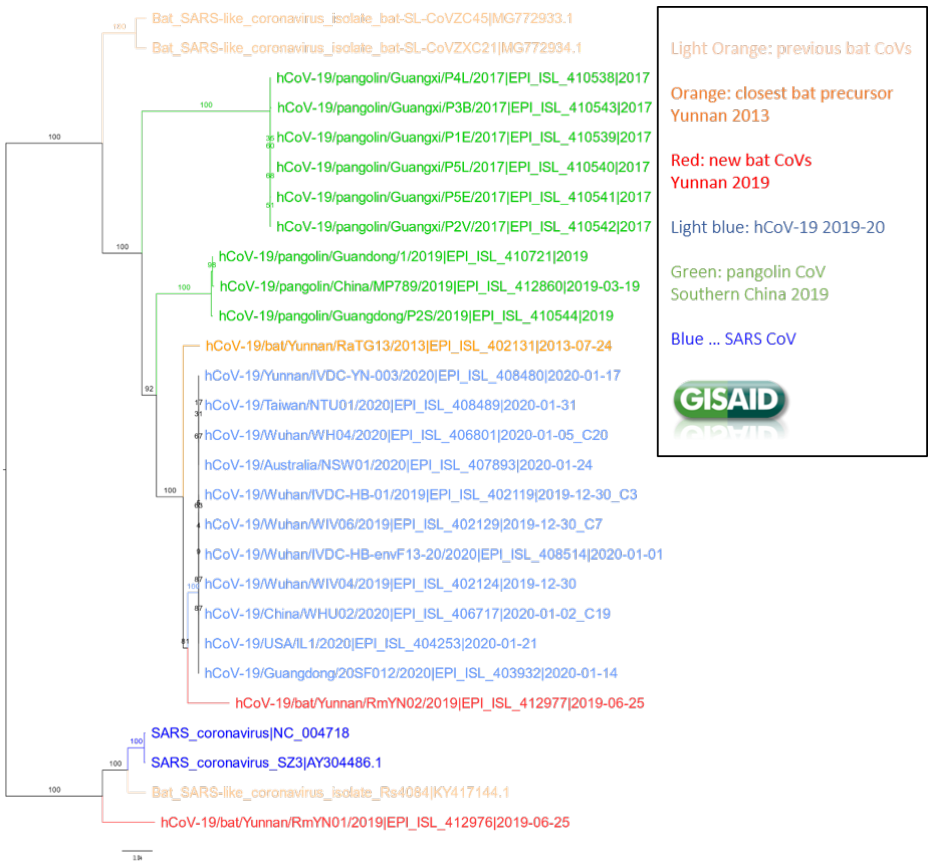

Figure 1 – Phylogénie des souches apparentées au SARS-CoV-2 (GISAID, 10/5/2020)

Une origine naturelle

On a très tôt trouvé dans la presse grand-public des allégations comme quoi le SARS-CoV-2 serait un produit de laboratoire. Tous les résultats et chiffres donnés ci-dessus démontrent le contraire. Les 1200 mutations qui séparent le SARS-CoV-2 de la souche la plus proche chez la chauve-souris (RaTG13) sont réparties aléatoirement le long du génome, alors que le génie biologique aurait produit un assemblage de fragments connus (sans mutations par rapport à certaines souches connues), avec des mutations ponctuelles dans des régions stratégiques, par exemple RBD. En se basant sur des similarités locales entre le génome du SARS-CoV-2 et celui du VIH, on a aussi prétendu qu'il s'agissait d'une tentative de vaccin contre le VIH. Mais ces similarités locales n'expliquent pas l'origine du reste du génome, et elles ne sont pas significatives. Elles portent sur un court segment de 38 nucléotides où les deux virus ont 87% d'identité. Mais en comparant le génome du SARS-CoV-2 à d'autres génomes on trouve de nombreuses similarités de cet ordre, par exemple 89% sur un segment de longueur 44 avec un génome de plante. On voit simplement ici l'effet du bricolage de l'évolution, qui utilise et réutilise les mêmes solutions pour bâtir le vivant. En aucun cas la marque d'une significativité statistique, bien difficile à estimer.

Evolution chez l'hôte humain

On s'est très tôt posé la question de l'origine, de la date et du « patient zéro » de la pandémie humaine. Les seules séquences ne permettent pas de répondre complétement à cette question. En phylogénie on utilise la méthode du groupe externe pour enraciner le groupe d'intérêt ; par exemple, pour enraciner l'arbre des mammifères on utilise un génome de reptile ou d'oiseau qui en sont les espèces les plus proches. Mais cette méthode ne fonctionne pas ici, le nombre de mutations observées entre le virus humain et celui de la chauve-souris (1200, cf. ci-dessus) étant sans commune mesure avec le nombre de mutations observées parmi les séquences humaines (quelques dizaines). Autrement dit toutes les séquences humaines sont peu ou prou à la même distance de celle de la chauve-souris et on ne peut pas avec cette méthode désigner avec certitude la racine de la pandémie.

On a de meilleures garanties en s'appuyant non seulement sur les séquences mais aussi sur les dates et l'histoire. En effet, le 30 décembre 2019 on a trouvé chez plusieurs patients exactement la même séquence, et celle-ci a rapidement été trouvée en Thaïlande, au Japon et aux USA, et elle était toujours présente fin Mars au Royaume Uni. C'est donc une bonne candidate pour être la « séquence zéro ». Elle est utilisée comme telle par de nombreuses équipes et de nombreux logiciels ou sites web, notamment GISAID, NEXSTRAIN (https://nextstrain.org/), etc. Comme cette séquence (WIV04/2019) a été trouvée en différents points du globe, il est impossible de dire quelle est son origine géographique, même si l'histoire semble indiquer la Chine. Diverses méthodes de datation indiquent un début de la diffusion de l'épidémie entre fin Octobre et début Décembre 2019. Il est difficile à l'heure actuelle d'avoir plus de certitudes. Comme avec

d'autres virus, VIH notamment, l'acquisition de nouvelles séquences, possiblement prélevées dans des échantillons anciens, devrait permettre d'affiner ces estimations et vraisemblablement trouver une origine plus précoce de la pandémie humaine.

Les premières séquences ne montraient que très peu voire aucune différence. Aujourd'hui, après environ 6 mois d'évolution, les séquences les plus récentes sont séparées de la « séquence zéro » par au plus 25 mutations en nucléotides et 12 en acides aminés. Ces chiffres sont pour une part incertains, car il est parfois difficile de distinguer mutation et erreur de séquençage. On observe quelques délétions, parfois longues et trouvées chez plusieurs patients, comme une délétion de 382b dans l'ORF8 et sa région régulatrice, échantillonnée une vingtaine de fois entre Singapour et Taiwan. Cette délétion a été observée sous une forme proche dans le SARS-CoV-1 et pourrait être une marque d'adaptation. A l'inverse on n'a trouvé aucune insertion marquante et partagée entre les virus de patients différents.

On a très rapidement cherché des mutations qui pourraient résulter d'une adaptation à l'hôte humain, ou être attachées à une plus grande virulence ou sévérité. La rareté des mutations rend ces analyses difficiles. De manière générale, on considère que les mutations observées aujourd'hui ont un impact neutre sur le phénotype, et sont issues d'un processus aléatoire lié aux erreurs de réplication. Suivant cette hypothèse communément admise, il n'existerait donc pas de souches plus virulentes que les autres. Cependant, la mutation D614G (transformation d'un résidu D en G à la position 614) de la protéine Spike semble correspondre à une transmissibilité accrue, si l'on se base sur la fréquence croissante de cette mutation dans les données mondiales. Alors que la pandémie a commencé avec la version D614 et s'est transmise dans cette version à de nombreux pays, ceux-ci sont aujourd'hui presque tous majoritairement affectés par la version G614 (par exemple en France), à l'exception notable de la Chine (très peu de variants G) et de l'Islande (retour du D après une phase exclusivement G). Les tenants de cette hypothèse avancent des différences génétiques chez l'hôte humain pour expliquer ces résultats, qui semblent différencier les populations humaines. Cette hypothèse doit cependant être regardée avec précaution, car les variants D et G pourraient correspondre à des vagues successives dans milieux socio-culturels différents. Par ailleurs, on ne voit aucune différence de sévérité entre les deux variants.

Alors que les coronavirus recombinent abondamment, on n'a pas trouvé pour l'instant de marqueurs fiables de recombinaison parmi les souches humaines. Celle-ci pourrait se produire en cas de co-infection par des souches significativement différentes, mais la probabilité de co-infection est très faible, et les souches qui circulent actuellement sont trop proches pour que ce phénomène puisse être détecté si tant est qu'il se produise. De ce point de vue, le SARS-CoV-2 semble se distinguer nettement de la grippe, qui évolue par réassortiment de sous-types différents, pouvant aboutir à des changements radicaux présentant des risques pandémiques majeurs.

Inférences phylogéographiques, clades et sous-types

De même qu'on fait des phylogénies inter-espèces des séquences (cf. Fig. 1), on infère en routine des phylogénies des séquences humaines, qui donnent une représentation des chaînes de transmission, depuis la « séquence zéro » à la racine, jusqu'aux dernières séquences observées aujourd'hui aux feuilles de l'arbre. Les caractéristiques de l'évolution des séquences (peu de mutations, absence de recombinaison) font que les arbres ainsi obtenus sont relativement robustes. Il en va de même pour les inférences phylogéographiques. En se basant sur des modèles de migration, on peut généralement attribuer avec une confiance relativement élevée une origine géographique aux nœuds ancestraux d'une phylogénie. En regroupant en clusters (ou foyers) de transmission les feuilles et nœuds d'un même sous-arbre partageant la même origine géographique, on en déduit des scénarios comme celui représenté dans la Figure 2. On peut aussi projeter ces inférences sur une carte du monde, comme dans la Figure 3. Globalement on observe une origine chinoise de la pandémie. Il est clair dans ces reconstructions (et l'histoire des faits) que la pandémie a démarré en Chine, à Wuhan, et qu'elle a rapidement circulé dans le monde entier, avec des introductions multiples, par exemple au Royaume Uni. On observe plusieurs effets fondateurs, comme en France, avec un cluster majoritaire et de nombreuses introductions indépendantes. On observe aussi des retours vers la Chine, ou des circulations bidirectionnelles entre la Belgique et les Pays Bas, par exemple.

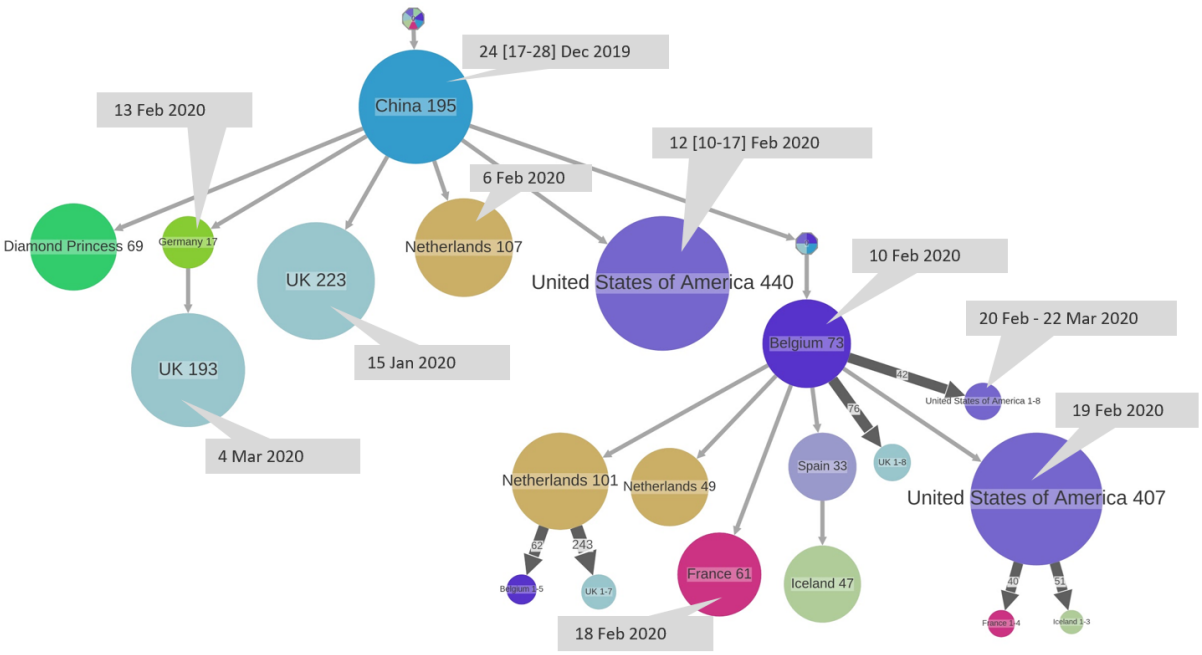

Figure 2 - Scénario phylogéographique montrant les principaux flux de transmission. Les nœuds correspondent à des clusters (ou foyers) de transmission partageant une même origine géographique. Les chiffres donnent le nombre de virus séquencés dans ces clusters (par exemple 61 dans le principal cluster français). Les dates sont celles de l'origine de la transmission au

sein du cluster (par exemple entre le 17 et le 28 décembre pour le cluster chinois initial). Les flèches minces montrent la transmission par un seul patient d'un pays à un autre (par exemple une origine belge pour le principal cluster français). Les flèches épaisses indiquent des transmissions multiples indépendantes et leur nombre (par exemple 243 transmissions depuis la Hollande vers le Royaume Uni). Données : GISAID, 24/04/2020 ; 3 500 séquences ; les plus petits clusters ne sont pas représentés. Logiciels : RaxML-NG, LSD2 et PastML. Auteurs : Anna Zhukova et Olivier Gascuel, Institut Pasteur et CNRS.

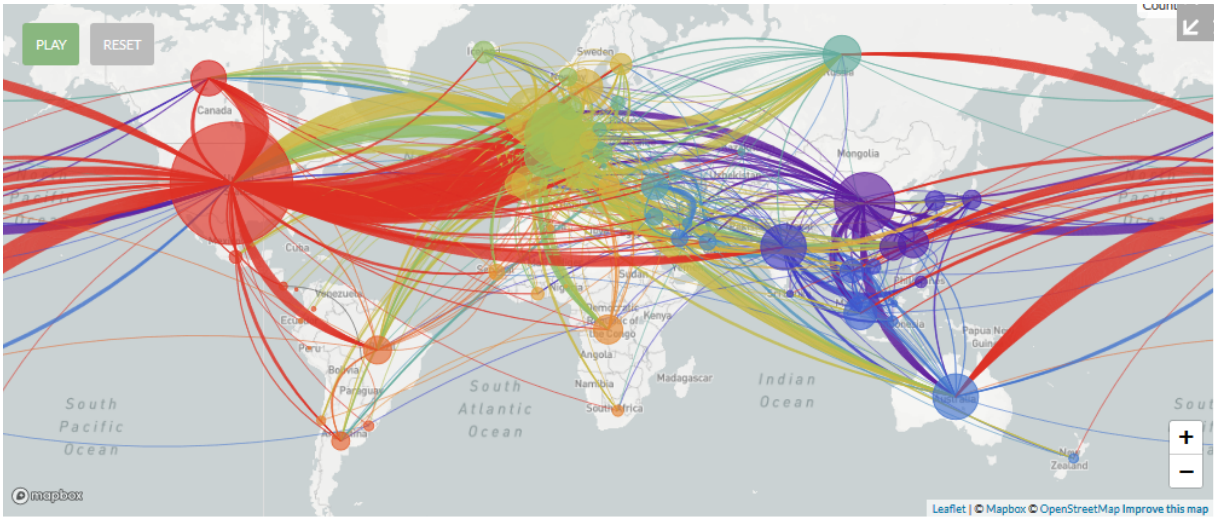

Figure 3 – Projection sur une carte d'un scénario phylogéographique (NEXTSTRAIN).

Finalement, se pose la question des clades ou sous-types, correspondant à des séquences distinctes et possédant des caractères épidémiologiques d'intérêt. Plusieurs groupes ont proposé des classifications et nomenclatures différentes. La distinction la plus convaincante est associée à la mutation D614G discutée plus haut. Les séquences comportant la version G614, accompagnée de deux mutations au niveau de l'ARN, constituent le clade G du GISAID, nommé A2 par NEXSTRAIN. Ce clade a une différence phylogénétique claire avec les autres séquences, et il présente un grand intérêt épidémiologique potentiel (transmissibilité accrue), même si les preuves formelles manquent encore. Au sein de ce clade on distingue des sous-clades, mais sans intérêt épidémiologique évident. Les séquences situées en dehors du clade G constituent pour l'essentiel le clade S (GISAID) ou B (NEXTSTRAIN), qui contient la « séquence zéro » et les premières séquences observées en décembre/janvier. Environ 15% des séquences sont non classées (du moins dans le système du GISAID). Après 6 mois d'évolution on est donc très loin de la séparation en groupes et sous-types du VIH-1, séparation qui représente plus d'un siècle d'évolution, avec des différences marquées et stables et des caractères épidémiologiques affirmés (répartition géographique par exemple).

En résumé

- L'analyse des séquences indique très clairement une origine naturelle du SARS-CoV-2 et aucune ressemblance significative avec le VIH, comme cela a pu être suggéré.
- De nouvelles données, issues de réservoirs encore inexplorés ou d'échantillons anciens, devraient permettre de faire progresser nos connaissances sur l'origine du virus, y compris en termes de date d'apparition et de circulation dans la population humaine.
- Les séquences du SARS-CoV-2 mutent et présentent de nombreux variants, en nucléotides ainsi qu'en acides aminés. Rien ne prouve à ce jour que ces mutations aient un impact sur la virulence ou la sévérité, mais elles induisent vraisemblablement des variations dans les réponses immunitaires, dont l'impact sur les vaccins potentiels et les tests devra être évalué.
- Plus de séquences, portant sur des périodes plus longues d'évolution et avec une représentation plus exhaustive des différentes populations humaines, permettront d'asseoir l'étude de ces mutations (ponctuelles, délétions, insertions, recombinaisons…) en termes de pression de sélection, convergence, adaptation à l'hôte humain, virulence, sévérité et risque pandémique.
- Sous la pression de cette pandémie et de ses données massives, les méthodes, les algorithmes et les modèles progressent rapidement et devraient asseoir l'épidémiologie moléculaire comme un domaine clé dans l'étude et la lutte contre les pandémies virales à venir.

Cette fiche a été conçue et rédigée par la cellule de crise Coronavirus de l'Académie des sciences. Créée à l'initiative de Pascale Cossart, Secrétaire perpétuel de l'Académie, celle-ci réunit des académiciens experts du domaine : Jean-François Bach, Pierre Corvol, Dominique Costagliola, Pascale Cossart (coordinatrice), Patrick Couvreur, Olivier Faugeras, Olivier Gascuel, Daniel Louvard, Félix Rey, Philippe Sansonetti, Alain-Jacques Valleron.

Les informations qui figurent sur cette fiche ont été produites collectivement et sont susceptibles d'évoluer. Elles seront éventuellement réactualisées en fonction des avancées des connaissances scientifiques.