*Module d'enseignement "Nature et culture 3*
*Licence Sciences et Humanités*
*Aix-Marseille Université*
*Novembre 2020*

# Enquête bioinformatique sur les origines de SARS-CoV-2

# CM1 – Analyse des génomes de coronavirus

*Cours donné par **Jacques van Helden**, **Emese Meglécz** et **Gabriel Neve***

*Sur base d'une enquête menée par Erwan Salard, José Haloy, Didier Casane, Etienne Decroly et Jacques van Helden*

| | | | | |
|---|---|---|---|---|
| 03/11 | 10:00 | 12:00 | **CM1** | (1) La question des origines de SARS-CoV-2.<br>(2) Biologie de SARS-CoV-2.<br>(3) Événements évolutifs.<br>(4) Bases de données biologiques.<br>(5) Alignement de séquences par paires.<br>(6) Recherche de séquences par similarité. |
| 10/11 | 8:30 | 11:30 | **TP1 groupe 1** | (1) Bases de données de séquences biologiques (Uniprot, NCBI).<br>(2) Alignement par paires (needle).<br>(3) Recherche de similarités (BLAST). |
| | 12:30 | 15:30 | **TP1 groupe 2** | cf groupe 1 |
| | 15:45 | 18:00 | **CM2** | (1) Interprétation des résultats du TP 1.<br>(2) Profils de pourcents de positions identiques.<br>(3) Alignements multiples.<br>(4) Inférence phylogénétique |
| 17/11 | 8:30 | 11:30 | **TP2 groupe 1** | (1) Alignements 1 à N et profils de PPI (PIPprofileR)<br>(2) Alignements multiples (clustal).<br>(3) Inférence phylogénétique (phylogeny.fr) |
| | 12:30 | 15:30 | **TP2 groupe 2** | cf groupe 1 |
| | 16:00 | 17:00 | **CM3** | Interprétation, résumé et conclusion du cours |

*Pourquoi l'origine de SARS-CoV-2 suscite-t-elle des débat ?*

# *Pourquoi la question des origines se pose-t-elle ?*

- Dès le début de la pandémie, la question des origines du virus a suscité de fortes controverses : origine naturelle (transfert d'animal à humain) ou artificielle (produit dans un laboratoire) ?
- Les éléments de la controverse
  - Le virus a émergé dans la ville de Wuhan, où est situé le laboratoire de référence pour les recherches sur les coronavirus
  - Ce laboratoire réalise des expériences sur les coronavirus, notamment de gain de fonction
  - L'hypothèse d'une contamination dans le marché de Wuhan est fortement remise en cause
  - Certains chercheurs affirment que le génome de SARS-CoV-2 résulte de manipulations génétiques
  - Certains chercheurs pensent que le virus pourrait s'être échappé accidentellement d'un laboratoire. D'autres affirment carrément qu'il a été conçu pour servir d'arme biologique
  - Ces affirmations sont fortement médiatisées par les réseaux sociaux, mais ne sont pas publiées dans les revues scientifiques avec comité de lecture
  - Le débat scientifique est fortement perturbé par le contexte politique et géopolitique (relations Chine - USA, rôle de l'OMS, …)
- Comment distinguer le vrai du faux ?
  - **Approche: ignorer délibérément les débats médiatiques et analyser les séquences des virus**

# *A la recherche de l'hôte manquant*

- Les coronavirus responsables des dernières émergences épidémiques chez l'humain trouvaient leur origine dans un "réservoir naturel" constitué par les chauves-souris.
- La transmission à l'homme passe généralement par un hôte animal intermédiaire: la civette pour le SRAS (2002) et le dromadaire pour le MERS (2012).
- Pour la pandémie COVID-19 on a invoqué le pangolin comme hôte intermédiaire, en suggérant que le passage à l'homme provenait d'animaux vendus sur le marché de Wuhan.
- Cette hypothèse est cependant remise en cause pour différentes raisons.
  - Les premiers patients ne fréquentaient pas le marché
  - Les génomes des coronavirus de pangolin dont on dispose sont beaucoup plus éloignés de SARS-CoV-2 que ceux des chauves-souris

# *Expériences de gain de fonction*

- Les coronavirus responsables des dernières émergences épidémiques chez l'humain trouvaient leur origine dans un "réservoir naturel" constitué par les chauves-souris.
- La transmission à l'homme passe généralement par un hôte animal intermédiaire: la civette pour le SRAS (2002) et le dromadaire pour le MERS (2012).
- Pour la pandémie COVID-19 on a invoqué le pangolin comme hôte intermédiaire, en suggérant que le passage à l'homme provenait d'animaux vendus sur le marché de Wuhan.

- Certains laboratoires de virologie pratiquent des expériences dites "de gain de fonction" qui consistent à modifier un virus pour le rendre plus virulent ou plus contagieux, afin d'étudier les mécanismes moléculaires de l'infection virale.
- Ceci suscite de fortes réticences au sein même de leur communauté.
- 2011: Ron Fouchier annonce qu'il a produit un virus H5N1 modifié (9 mutations ponctuelles) pour augmenter sa contagiosité chez le furet.
- Moratoire:
  - En 2014 le NIH annonce qu'il ne financera plus les expériences de gain de fonction
  - En 2017 il lève cette mesure, en invoquant l'intérêt de ces expériences pour comprendre les mécanismes de l'infection virale.
- Plusieurs incidents ont déjà été rapportés concernant des fuites accidentelles de virus de laboratoires.

# *Arguments en faveur d'une zoonose naturelle*

- Andersen et collègues publient dès janvier 2020 un article affirmant que SARS-CoV-2 est sans aucun doute d'origine naturelle.
- Argument principal : la séquence du domaine de liaison au récepteur (RBD) est optimale pour se lier au récepteur ACE2, mais d'une façon différente de celles qu'on connaissait jusqu'alors. D'après les auteurs, si on avait conçu un virus dans le but de le rendre infectieux pour l'homme, on n'aurait pas pu concevoir cette séquence.
- Autre argument: on ne trouve pas dans ce génome de traces d'ingénierie moléculaire (par exemple des sites de restriction)


*While the analyses above suggest that SARS-CoV-2 may bind human ACE2 with high affinity, computational analyses predict that the interaction is not ideal and that the RBD sequence is different from those shown in SARS-CoV to be optimal for receptor binding. Thus, the high-affinity binding of the SARS-CoV-2 spike protein to human ACE2 is most likely the result of natural selection on a human or human-like ACE2 that permits another optimal binding solution to arise. This is strong evidence that SARS-CoV-2 is not the product of purposeful manipulation.*

Andersen, K.G., Rambaut, A., Lipkin, W.I., Holmes, E.C., and Garry, R.F. (2020). The proximal origin of SARS-CoV-2. Nature Medicine 26, 450–452.

Andersen, K.G., Rambaut, A., Lipkin, W.I., Holmes, E.C. & Garry, R.F. 2020. The proximal origin of SARS-CoV-2. Nature Medicine 26: 450–452. Nature Publishing Group.

# *Arguments en faveur d'une zoonose naturelle*

- Andersen et collègues publient dès janvier 2020 un article affirmant que SARS-CoV-2 est sans aucun doute d'origine naturelle.
- Argument principal : la séquence du domaine de liaison au récepteur (RBD) est optimale pour se lier au récepteur ACE2, mais d'une façon différente de celles qu'on connaissait jusqu'alors. D'après les auteurs, si on avait conçu un virus dans le but de le rendre infectieux pour l'homme, on n'aurait pas pu concevoir cette séquence.
  - ***Oui mais** les auteurs n'envisagent pas une autre voie possible : en cultivant des virus sur des cellules (en labo) on peut réaliser une sélection artificielle, qui peut déboucher sur des souches adaptées sans nécessiter de connaissance a priori des séquences.*
- Autre argument: on ne trouve pas dans ce génome de traces d'ingénierie moléculaire (par exemple des sites de restriction)
  - ***Oui mais** les techniques de biologie synthétique permettent depuis 15 ans de générer une molécule d'ADN de novo, sans recourir à des enzymes de restriction (et donc sans trace).*


correspondence

Check for updates

The proximal origin of SARS-CoV-2

NATURE MEDICINE | VOL 26 | APRIL 2020 | 450–455 | www.nature.com/naturemedicine

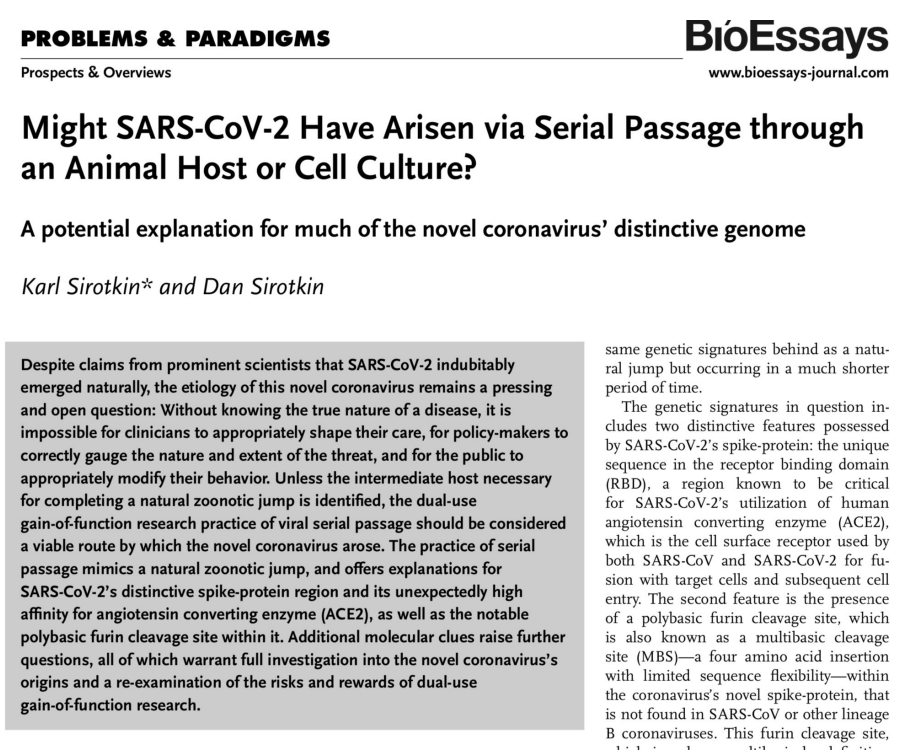
*While the analyses above suggest that SARS-CoV-2 may bind human ACE2 with high affinity, computational analyses predict that the interaction is not ideal and that the RBD sequence is different from those shown in SARS-CoV to be optimal for receptor binding. Thus, the high-affinity binding of the SARS-CoV-2 spike protein to human ACE2 is most likely the result of natural selection on a human or human-like ACE2 that permits another optimal binding solution to arise. This is strong evidence that SARS-CoV-2 is not the product of purposeful manipulation.*

Andersen, K.G., Rambaut, A., Lipkin, W.I., Holmes, E.C., and Garry, R.F. (2020). The proximal origin of SARS-CoV-2. Nature Medicine 26, 450–452.

Andersen, K.G., Rambaut, A., Lipkin, W.I., Holmes, E.C. & Garry, R.F. 2020. The proximal origin of SARS-CoV-2. Nature Medicine 26: 450–452. Nature Publishing Group.

# *Oui mais ...*

- Sirotkin & Sirotkin discutent des faiblesses des arguments d'Andersen
- Ils soulignent également l'absence générale d'évaluation sérieuse des hypothèses alternatives concernant les possibilité d' échappement de laboratoire.
- Ils développent l'historique des accidents de laboratoire, et évaluent les scénarios qui permettraient également de comprendre la nature des données en notre possession.
- Des passages successifs d'une souche virale d'une espèce à l'autre (en culture cellulaire ou sur animaux) donneraient le même effet : une divergence accélérée par rapport aux taux de mutations en milieu naturel.



**PROBLEMS & PARADIGMS**
Prospects & Overviews
BioEssays
www.bioessays-journal.com

## Might SARS-CoV-2 Have Arisen via Serial Passage through an Animal Host or Cell Culture?

### A potential explanation for much of the novel coronavirus' distinctive genome

*Karl Sirotkin* and Dan Sirotkin*

**Despite claims from prominent scientists that SARS-CoV-2 indubitably emerged naturally, the etiology of this novel coronavirus remains a pressing and open question: Without knowing the true nature of a disease, it is impossible for clinicians to appropriately shape their care, for policy-makers to correctly gauge the nature and extent of the threat, and for the public to appropriately modify their behavior. Unless the intermediate host necessary for completing a natural zoonotic jump is identified, the dual-use gain-of-function research practice of viral serial passage should be considered a viable route by which the novel coronavirus arose. The practice of serial passage mimics a natural zoonotic jump, and offers explanations for SARS-CoV-2's distinctive spike-protein region and its unexpectedly high affinity for angiotensin converting enzyme (ACE2), as well as the notable polybasic furin cleavage site within it. Additional molecular clues raise further questions, all of which warrant full investigation into the novel coronavirus's origins and a re-examination of the risks and rewards of dual-use gain-of-function research.**

same genetic signatures behind as a natural jump but occurring in a much shorter period of time.

The genetic signatures in question includes two distinctive features possessed by SARS-CoV-2's spike-protein: the unique sequence in the receptor binding domain (RBD), a region known to be critical for SARS-CoV-2's utilization of human angiotensin converting enzyme (ACE2), which is the cell surface receptor used by both SARS-CoV and SARS-CoV-2 for fusion with target cells and subsequent cell entry. The second feature is the presence of a polybasic furin cleavage site, which is also known as a multibasic cleavage site (MBS)—a four amino acid insertion with limited sequence flexibility—within the coronavirus's novel spike-protein, that is not found in SARS-CoV or other lineage B coronaviruses. This furin cleavage site,

Sirotkin, K. & Sirotkin, D. 2020. Might SARS-CoV-2 Have Arisen via Serial Passage through an Animal Host or Cell Culture?: A potential explanation for much of the novel coronavirus' distinctive genome. BioEssays 2000091.

# *Un virus synthétique avec des bouts de HIV ?*

Le 17 avril 2020, le Professeur Luc Montagnier, Prix Nobel de médecine pour sa contribution à la découverte du HIV (le virus responsable du SIDA), défraie la chronique en annonçant sur plusieurs médias (Pourquoi Docteur, CNEWS) que le génome du coronavirus SARS-CoV-2, agent de la pandémie COVID-19, comporte quatre fragments de séquences provenant du HIV. De plus, il affirme que la présence de ces séquences ne résulte pas d'une recombinaison naturelle (fréquente chez les virus) ou d'un accident, mais d'un vrai travail d'ingénieur, effectué intentionnellement, vraisemblablement dans le cadre de recherches visant à développer des vaccins contre le HIV.

Pour appuyer sa théorie, Luc Montagnier cite deux études :

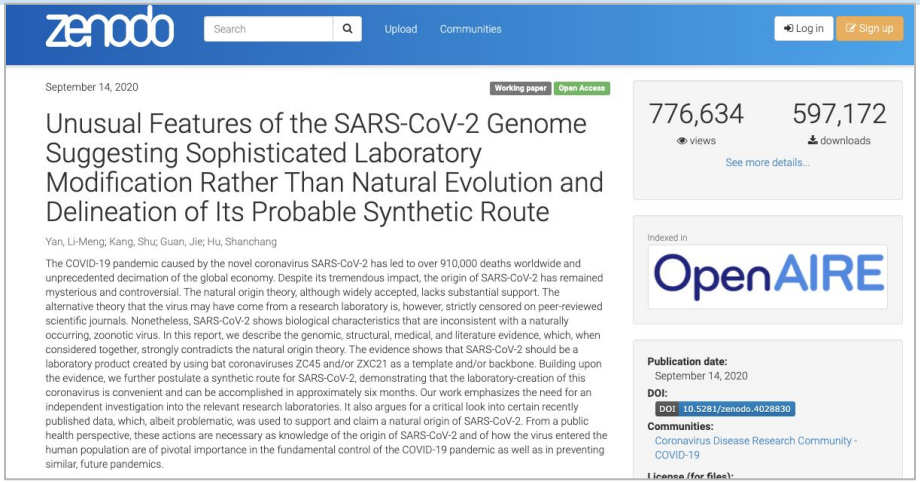
- le travail d'un collègue mathématicien, Jean-Claude Perez, qui "a fouillé les moindres détails de la séquence",
- une analyse des séquences génomiques et protéiques des coronavirus préalablement publiée par une équipe indienne, qui a, selon lui, "été forcée de rétracter" sa publication.

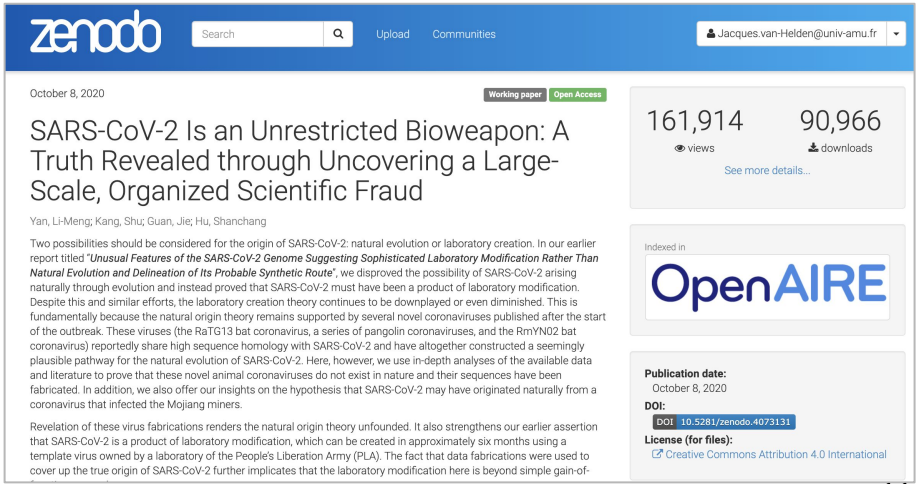
Professeur Luc Montagnier : Le virus covid19 est une manipulation humaine (https://www.youtube.com/watch?v=qSWCLHIOiMo).

*"Je suis arrivé à la conclusion qu'il y avait eu une manipulation de ce virus. [...] Il y a un modèle qui est évidemment le virus classique, et là c'était un modèle venant de la chauve-souris, et là, à ce modèle on a par-dessus ajouté les séquences du VIH, du SIDA. … Non, ce n'est pas naturel, c'était un travail de professionnel, de biologiste moléculaire, très minutieux, on peut dire d'horloger, au niveau des séquences. Dans quel but ce n'est pas clair. Mon travail c'est d'exposer les faits, c'est tout. Je n'accuse personne, je ne sais pas qui a fait ça et pourquoi. La possibilité c'est qu'on a voulu faire un vaccin contre le SIDA. Donc on a pris des petites séquences du virus [HIV] et on les a installées dans la séquence plus grande du coronavirus. [...] Il y a quand même une volonté d'étouffement, nous ne sommes pas les premiers. Un groupe de chercheurs indiens très renommés avaient publié la même chose, on les a forcés à rétracter. Si vous regardez leur publication vous voyez une grande bande "annulé". "*

# *Une arme biologique ?*

- Li-Meng Yan
  - chercheuse chinoise réfugiée aux Etats-Unis
  - travaillait dans le laboratoire de référence de l'OMS pour la Chine
- En septembre et octobre 2020, elle publie sur zenodo (dépôt d'oeuvres électroniques, sans révision par des experts) deux articles, où elle affirme que
  - le virus résulte de manipulations génétiques reposant sur les méthodes classiques de biologie moléculaire (recombinaison d'ADN au moyen d'enzymes de restriction) ;
  - ce virus est une arme biologique.
- Ses articles n'ont pas encore été publiés, mais ils sont téléchargés et fortement médiatisés.
- Les arguments sous-jacents font cependant l'objet de critiques par les spécialistes.



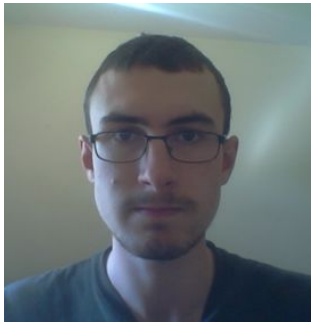




Yan, Li-Meng, Kang, Shu, Guan, Jie & Hu, Shanchang. 2020a. SARS-CoV-2 Is an Unrestricted Bioweapon: A Truth Revealed through Uncovering a Large-Scale, Organized Scientific Fraud. , doi: 10.5281/ZENODO.4073131. Zenodo.

Yan, Li-Meng, Kang, Shu, Guan, Jie & Hu, Shanchang. 2020b. Unusual Features of the SARS-CoV-2 Genome Suggesting Sophisticated Laboratory Modification Rather Than Natural Evolution and Delineation of Its Probable Synthetic Route. , doi: 10.5281/ZENODO.4028830. Zenodo.

# A l'origine de ce cours

- Ce cours repose en grande partie sur un travail d'équipe mené durant le confinement d'avril-mai 2020.
- Nous remercions les nombreux collègues qui nous ont suggéré des améliorations sur les premières versions du manuscrit.
- Afin d'assurer la traçabilité et la reproductibilité de nos analyses, les données utilisées et les logiciels développés sont en libre accès
  - https://jvanheld.github.io/SARS-CoV-2_origins/

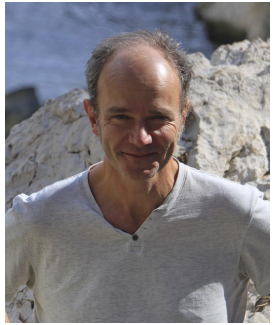


**Erwan Sallard**
Étudiant à l'Institut de Biologie
Ecole Normale Supérieure, Paris

**Didier Casane**
Biologie évolutive
Professeur, Université de Paris

**José Halloy**
Modélisation des systèmes biologiques
Professeur, Université de Paris

**Etienne Decroly**
Virologie
Dir. de Recherches, CNRS

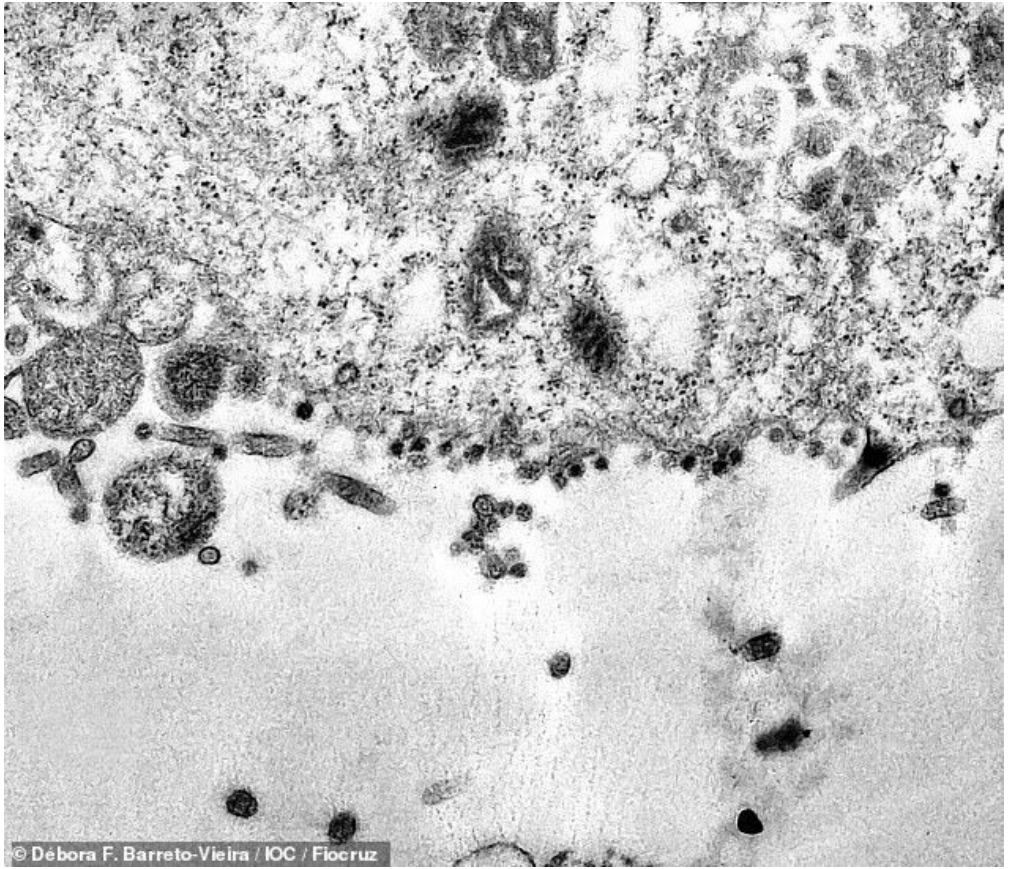


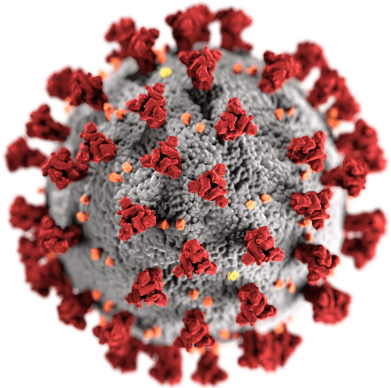
**Jacques van Helden**
Bioinformatique
Professeur, Aix-Marseille Université

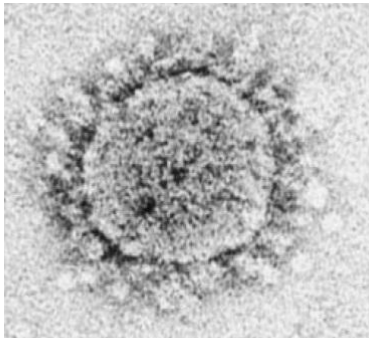

- Sallard, E., Halloy, J., Casane, D., van Helden, J. & Decroly, É. 2020. **Retrouver les origines du SARS-CoV-2 dans les phylogénies de coronavirus**. Med Sci (Paris) 36: 783–796.
- English version : Erwan Sallard, José Halloy, Didier Casane, Etienne Decroly, Jacques van Helden. **Tracing the origins of SARS-CoV-2 in coronavirus phylogenies**. hal-02891455

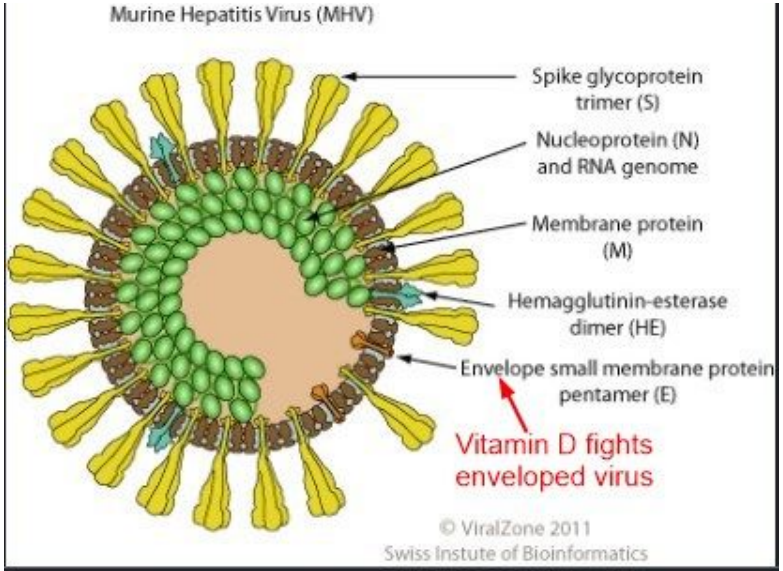
# *Biologie de SARS-CoV-2*

# *Coronavirus*





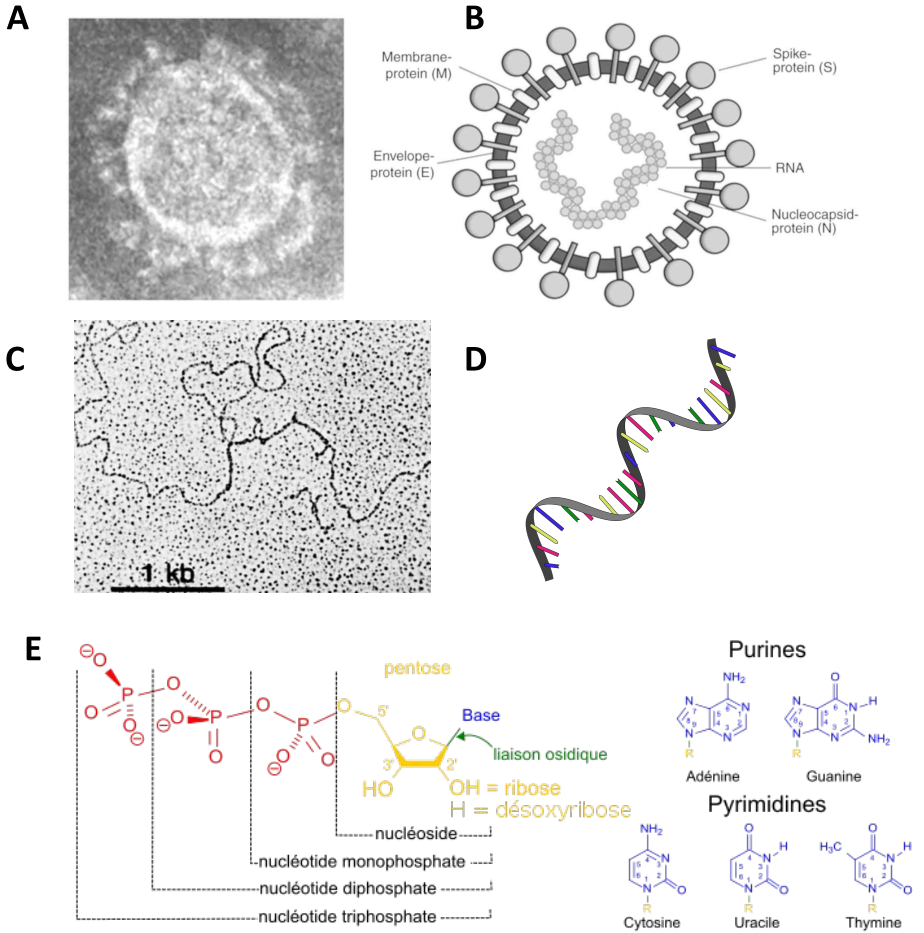


# *Le génome des coronavirus est constitué d'ARN*

- Le génome est l'ensemble du matériel génétique d'un organisme
- Le génome des coronavirus est constitué d'acide ribonucléique (ARN).

**Légende de la figure**

A. Micrographie électronique d'un virion de coronavirus.
B. Schéma de la structure du virion. Le chapelet à l'intérieur symbolise l'ARN.
C. Micrographie électronique de l'ARN viral extrait de son enveloppe
D. Schéma de la structure de l'ARN. Les bâtonnets de couleur représentent les nucléotides
    Adénine
    Cytosine
    Guanine
    Uracile
E. Structure chimique des nucléotides




Parvez, M., Jagirdar, R., Purty, R., Venkata, S., Agrawal, V., Kumar, J., et al. 2020. COVID-19 pandemic: Understanding the emergence, pathogenesis and containment (Review). World Acad Sci J, doi: 10.3892/wasj.2020.59.

# *Génome de SARS-CoV-2 de référence*

- Début de la séquence génomique de SARS-CoV-2
- La taille totale fait 29 899 nucléotides
- Un des plus grands génomes parmi les virus à ARN
  - A Adénine
  - C Cytosine
  - G Guanine
  - T Uracile (on remplace le U par un T)

>MT019529.1 Severe acute respiratory syndrome coronavirus 2 isolate BetaCoV/Wuhan/IPBCAMS-WH-01/2019, complete genome

ATTAAAGGTTTATACCTTCCCAGGTAACAAACCAACCAACTTTCGATCTCTTGTAGATCTGTTCTCTAAA
CGAACTTTAAAATCTGTGTGGCTGTCACTCGGCTGCATGCTTAGTGCACTCACGCAGTATAATTAATAAC
TAATTACTGTCGTTGACAGGACACGAGTAACTCGTCTATCTTCTGCAGGCTGCTTACGGTTTCGTCCGTG
TTGCAGCCGATCATCAGCACATCTAGGTTTCGTCCGGGTGTGACCGAAAGGTAAGATGGAGAGCCTTGTC
CCTGGTTTCAACGAGAAAACACACGTCCAACTCAGTTTGCCTGTTTTACAGGTTCGCGACGTGCTCGTAC
GTGGCTTTGGAGACTCCGTGGAGGAGGTCTTATCAGAGGCACGTCAACATCTTAAAGATGGCACTTGTGG
CTTAGTAGAAGTTGAAAAAGGCGTTTTGCCTCAACTTGAACAGCCCTATGTGTTCATCAAACGTTCGGAT
GCTCGAACTGCACCTCATGGTCATGTTATGGTTGAGCTGGTAGCAGAACTCGAAGGCATTCAGTACGGTC
GTAGTGGTGAGACACTTGGTGTCCTTGTCCCTCATGTGGGCGAAATACCAGTGGCTTACCGCAAGGTTCT
TCTTCGTAAGAACGGTAATAAAGGAGCTGGTGGCCATAGTTACGGCGCCGATCTAAAGTCATTTGACTTA
GGCGACGAGCTTGGCACTGATCCTTATGAAGATTTTCAAGAAAACTGGAACACTAAACATAGCAGTGGTG
TTACCCGTGAACTCATGCGTGAGCTTAACGGAGGGGCATACACTCGCTATGTCGATAACAACTTCTGTGG
CCCTGATGGCTACCCTCTTGAGTGCATTAAAGACCTTCTAGCACGTGCTGGTAAAGCTTCATGCACTTTG
TCCGAACAACTGGACTTTATTGACACTAAGAGGGGTGTATACTGCTGCCGTGAACATGAGCATGAAATTG
CTTGGTACACGGAACGTTCTGAAAAGAGCTATGAATTGCAGACACCTTTTGAAATTAAATTGGCAAAGAA
ATTTGACACCTTCAATGGGGAATGTCCAAATTTTGTATTTCCCTTAAATTCCATAATCAAGACTATTCAA
CCAAGGGTTGAAAAGAAAAAGCTTGATGGCTTTATGGGTAGAATTCGATCTGTCTATCCAGTTGCGTCAC
CAAATGAATGCAACCAAATGTGCCTTTCAACTCTCATGAAGTGTGATCATTGTGGTGAAACTTCATGGCA
GACGGGCGATTTTGTTAAAGCCACTTGCGAATTTTGTGGCACTGAGAATTTGACTAAAGAAGGTGCCACT
ACTTGTGGTTACTTACCCCAAAATGCTGTTGTTAAAATTTATTGTCCAGCATGTCACAATTCAGAAGTAG
GACCTGAGCATAGTCTTGCCGAATACCATAATGAATCTGGCTTGAAAACCATTCTTCGTAAGGGTGGTCG
CACTATTGCCTTTGGAGGCTGTGTGTTCTCTTATGTTGGTTGCCATAACAAGTGTGCCTATTGGGTTCCA
CGTGCTAGCGCTAACATAGGTTGTAACCATACAGGTGTTGTTGGAGAAGGTTCCGAAGGTCTTAATGACA
ACCTTCTTGAAATACTCCAAAAAGAGAAAGTCAACATCAATATTGTTGGTGACTTTAAACTTAATGAAGA
GATCGCCATTATTTTGGCATCTTTTTCTGCTTCCACAAGTGCTTTTGTGGAAACTGTGAAAGGTTTGGAT
TATAAAGCATTCAAACAAATTGTTGAATCCTGTGGTAATTTTAAAGTTACAAAAGGAAAAGCTAAAAAAG
GTGCCTGGAATATTGGTGAACAGAAATCAATACTGAGTCCTCTTTATGCATTTGCATCAGAGGCTGCTCG

https://www.ncbi.nlm.nih.gov/nuccore/MT019529.1

# *Apports de la génomique*

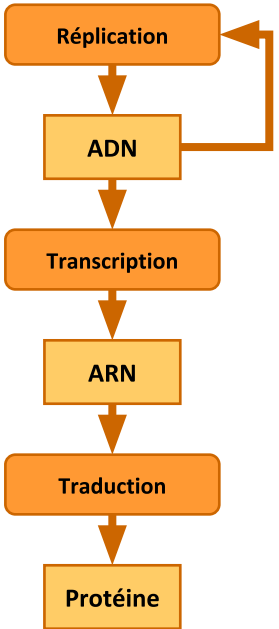
- Le séquençage d'un génome viral représente un coût modique et se fait très rapidement.
- On dispose de
  - plusieurs dizaines de génomes de  coronavirus non-humains
  - plusieurs dizaines de milliers de génomes de SARS-CoV-2 (humain)
- Ces données permettent de poser des questions concernant
  - l'origine animale de SARS-CoV-2
  - sa propagation dans la population humaine

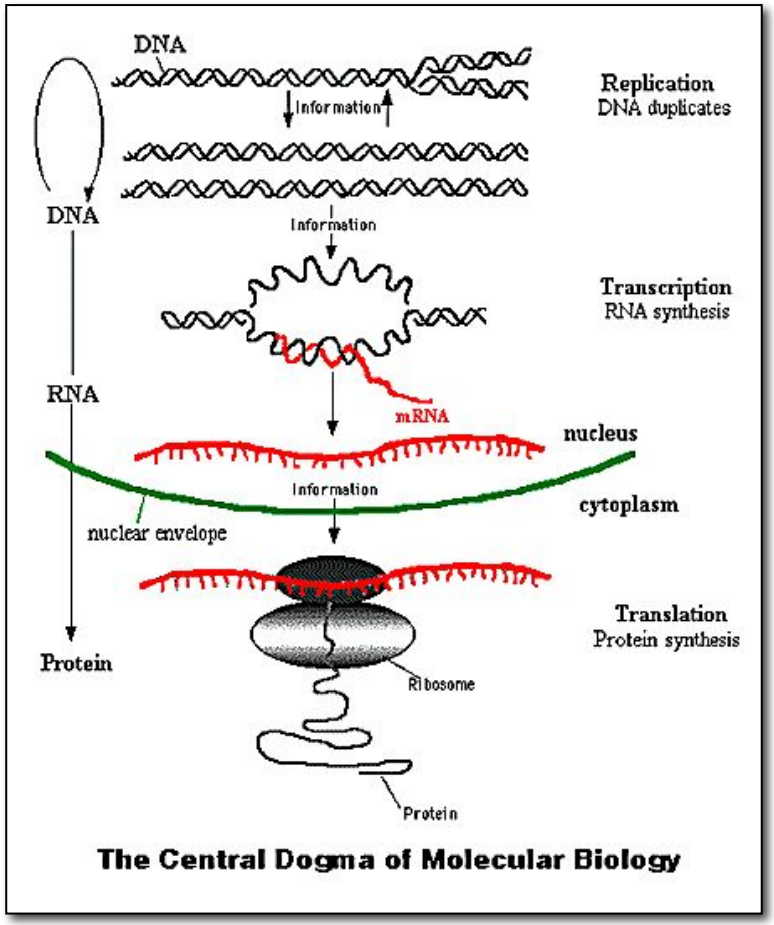
# Le “dogme central” de la biologie

- Formulé en 1958 par Francis Crick
  - Crick, F. H. (1958). On protein synthesis. Symp Soc Exp Biol 12, 138-63.
  - Je recommande également de lire cette discussion ultérieure : Crick, F. (1970). Central dogma of molecular biology. Nature 227, 561-3.
- On le résume souvent de la façon suivante
  - ***DNA makes RNA makes protein***
    *(l’ADN fait l’ARN qui fait les protéines)*
- Cependant le dogme ne dit pas exactement cela. Il énonce les transferts d’information qui sont possibles ou impossibles entre les séquences d’acides nucléiques et celles des protéines.

***Le dogme central*** *stipule que, une fois que l’ « information » est passée dans la protéine elle ne peut pas en ressortir. Plus précisément, le transfert d’information serait possible d’acide nucléique à acide nucléique, ou d’acide nucléique à protéine, mais le transfert de protéine à protéine, ou de protéine à acide nucléique est impossible. Information signifie ici la détermination précise de la séquence, soit des bases dans l’acide nucléique, soit des résidus aminoacides dans la protéine.*

Crick, F. H. (1958). On protein synthesis. Symp Soc Exp Biol 12, 138-63.



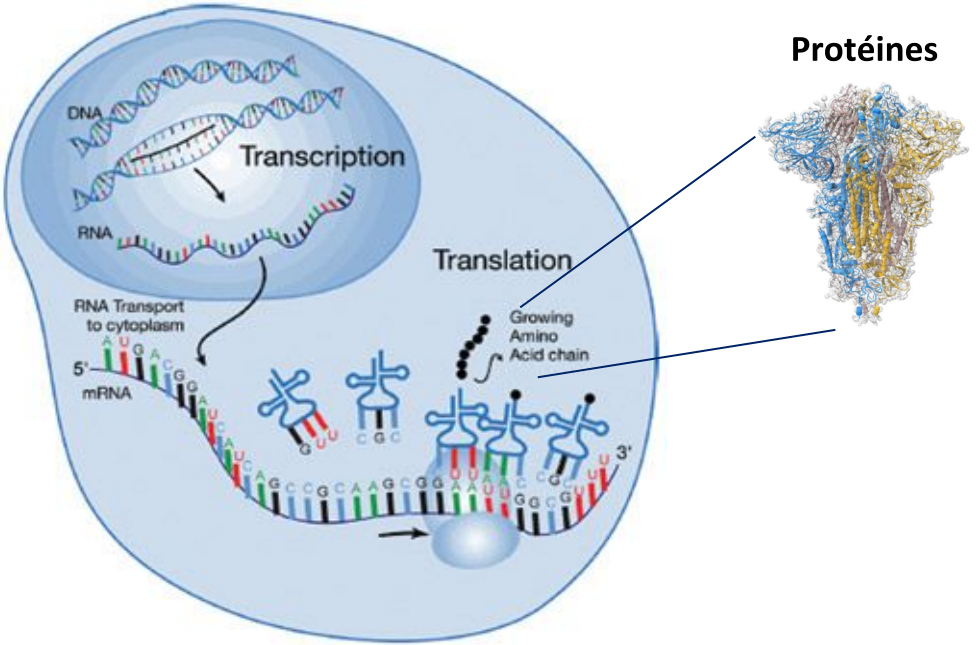




http://www.accessexcellence.org/AB/GG/central.html

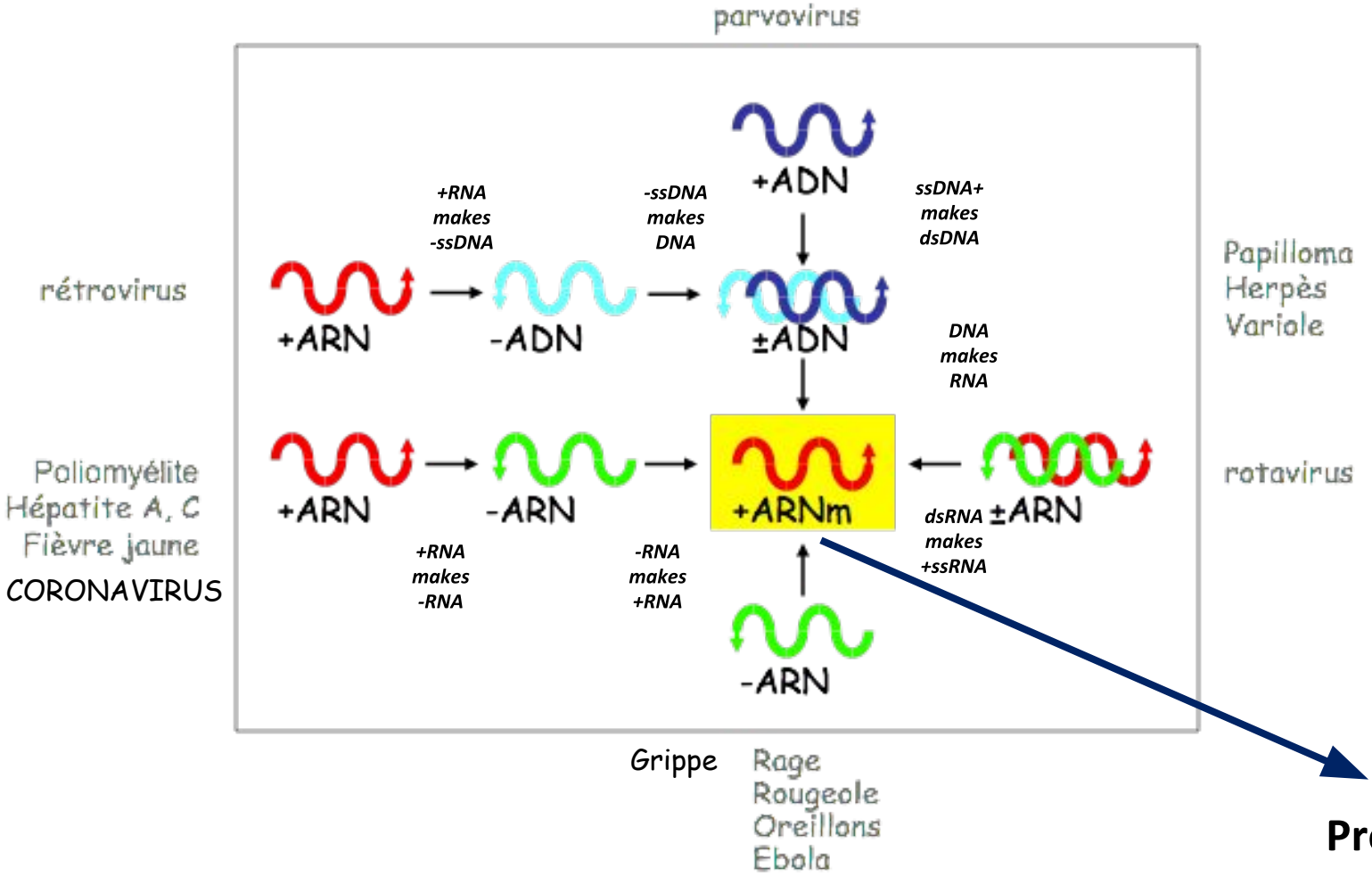
# *DNA makes RNA makes protein**

** Formulation compacte des flux de l'information moléculaire, par Francis Crick*

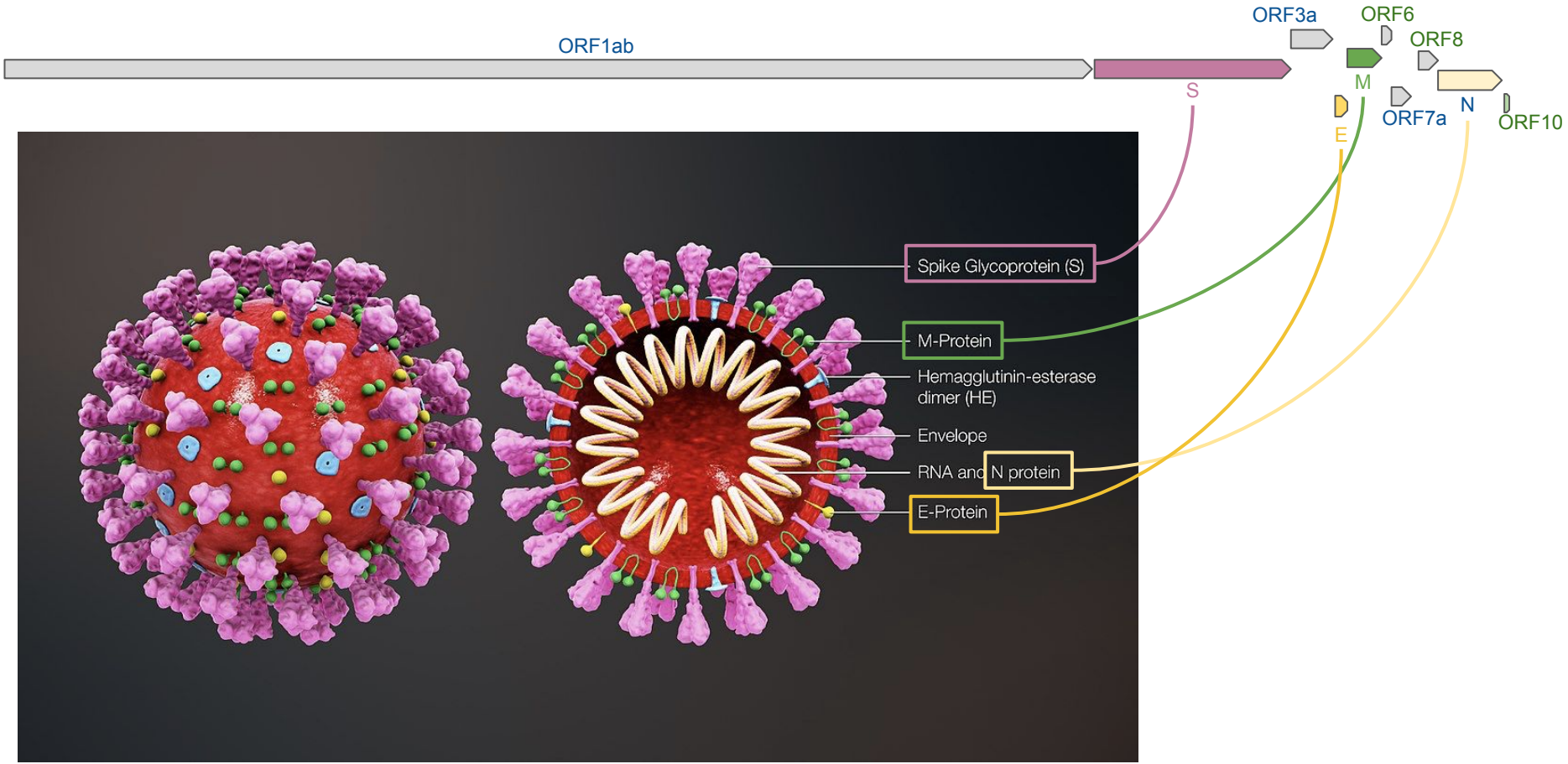
- Dans tous les organismes cellulaires
  - L'ADN sert de modèle à la synthèse d'ARN (***transcription***)
  - L'ARN sert de modèle à la synthèse des protéines (*traduction*)
- Les protéines sont les principaux acteurs moléculaires des organismes vivants
  - Enzymes
  - Transporteurs
  - Régulateurs
  - Cycle cellulaire
  - Différenciation cellulaire
  - … un tas d'autres fonctions



# *Le matériel génétique des virus*



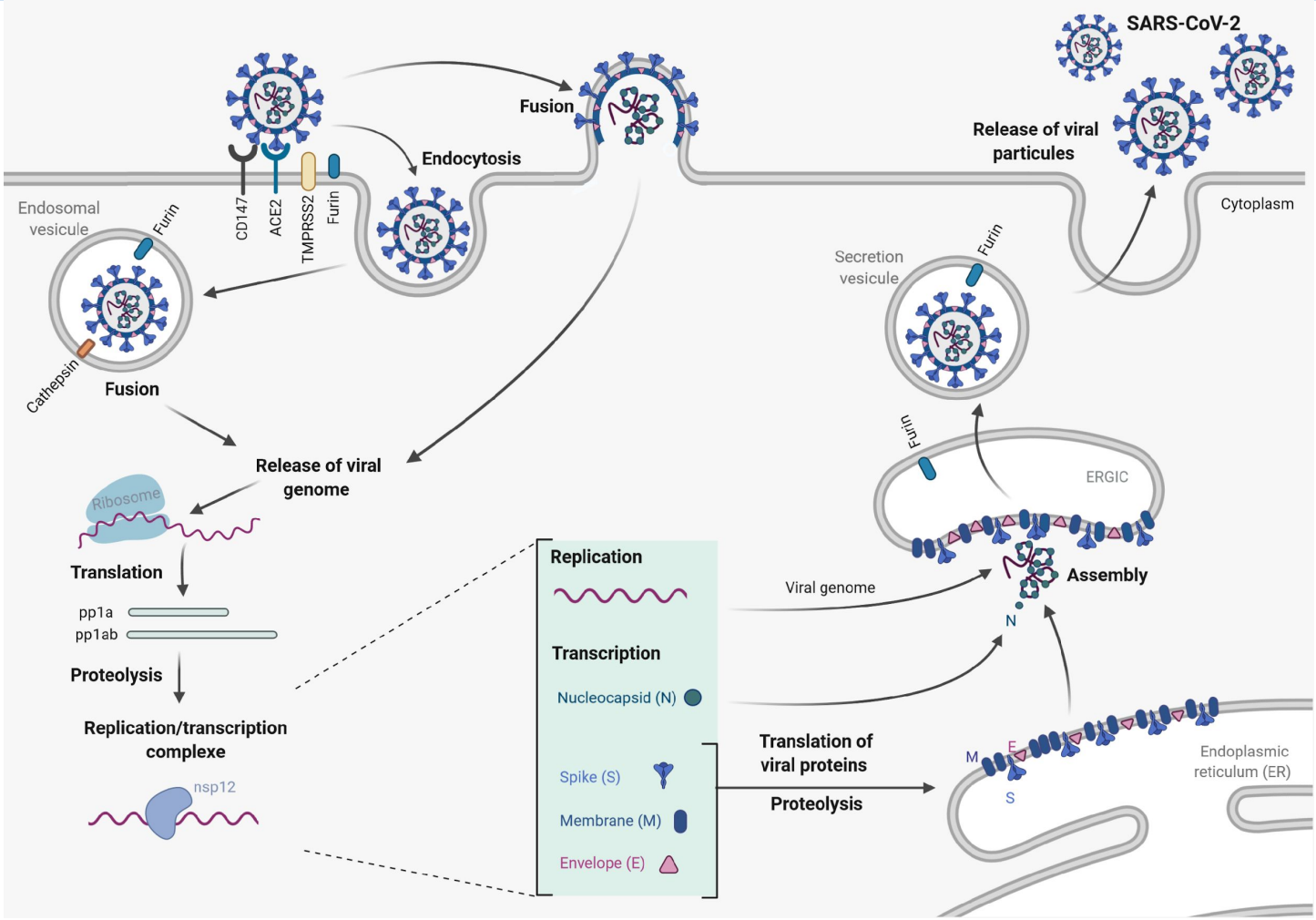
# *Fonction des gènes de SARS-CoV-2 – Gènes structuraux*



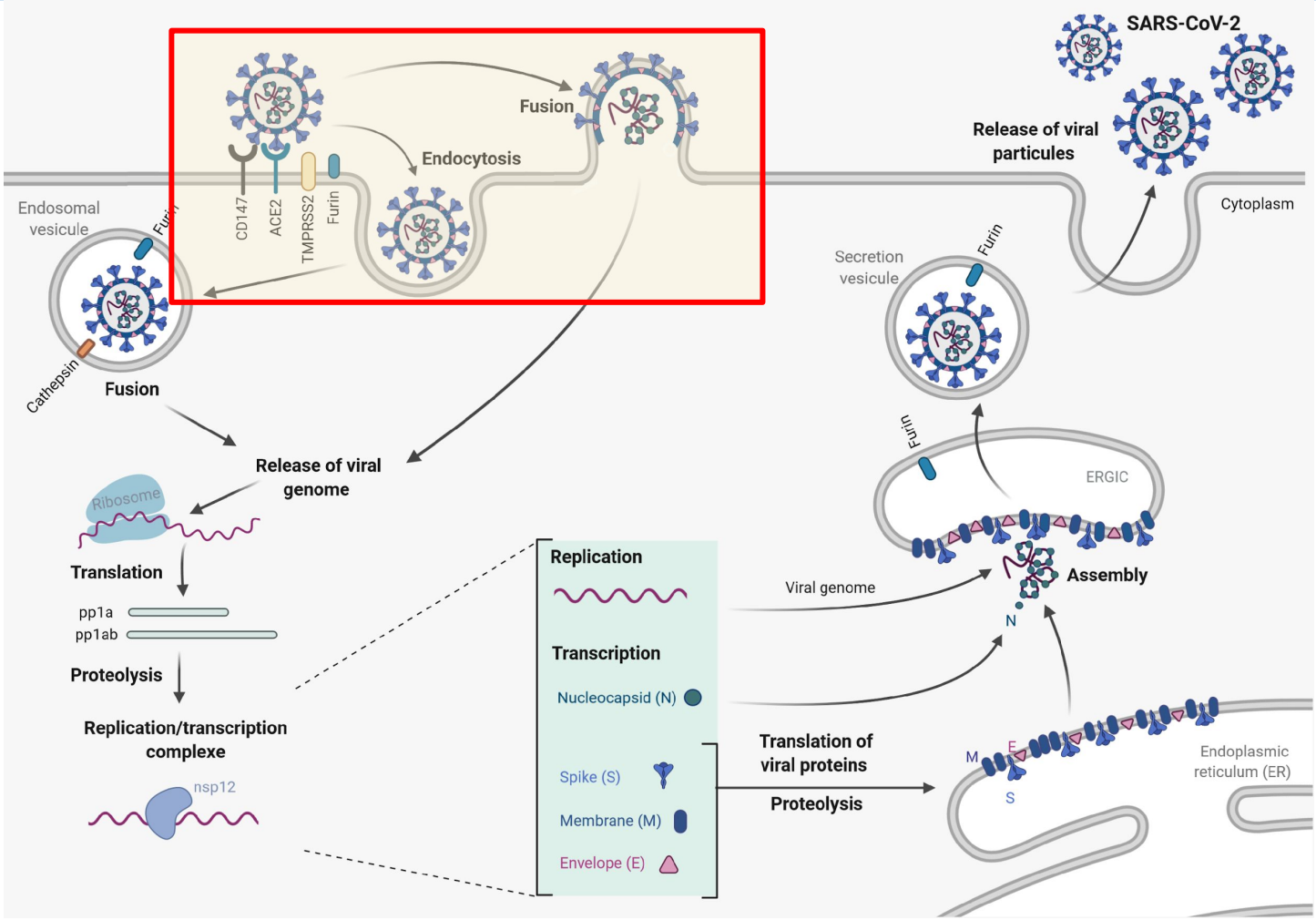

Source de la figure: https://commons.wikimedia.org/wiki/File:3D_medical_animation_coronavirus_structure.jpg

# *Mécanismes d'infection*

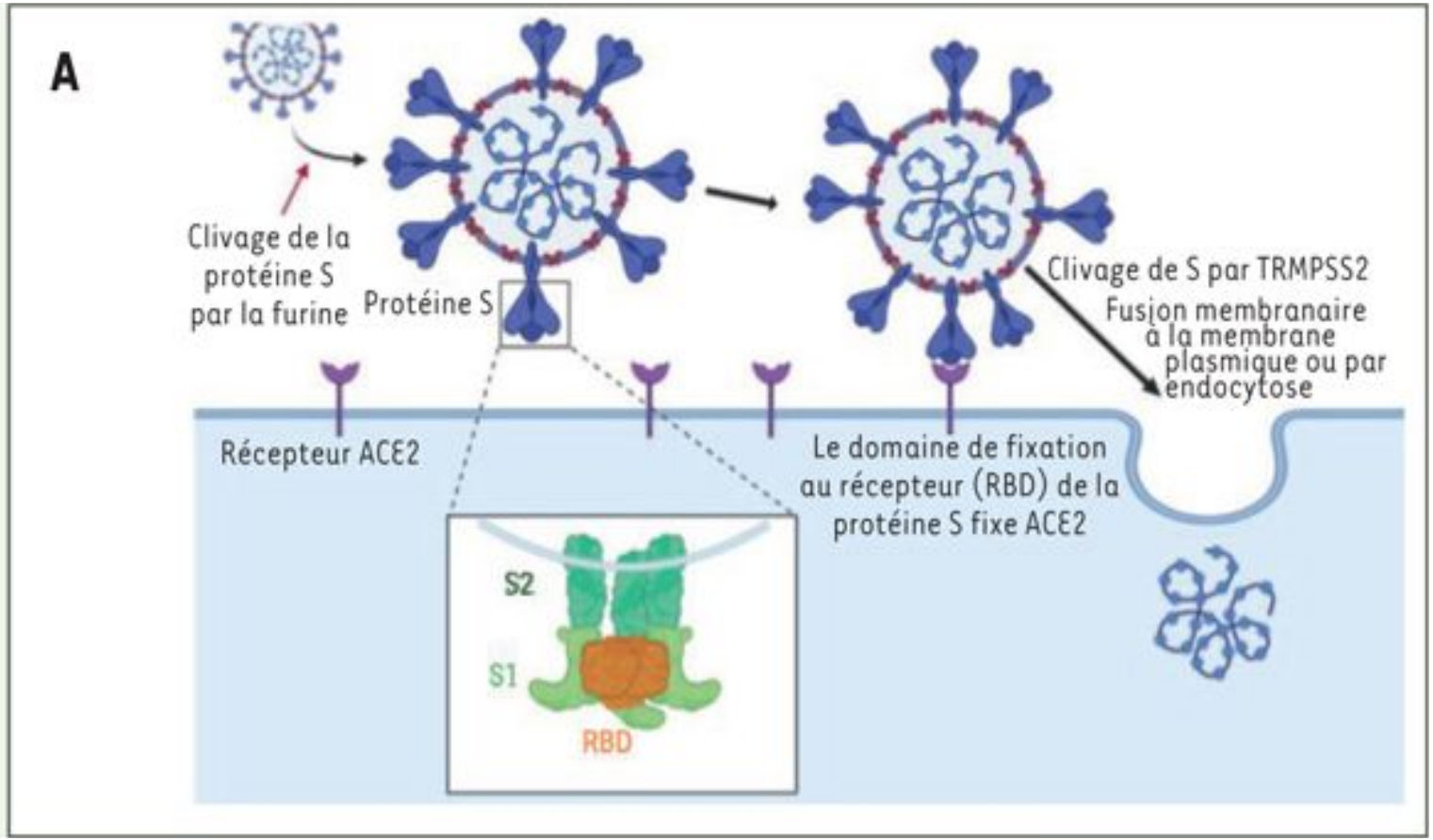
# Cycle de réplication des coronavirus



# Cycle de réplication des coronavirus - Entrée dans la cellule hôte

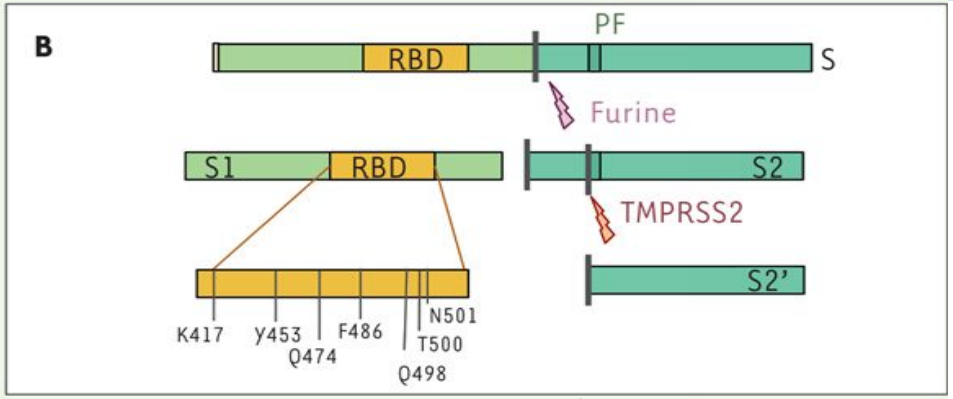


# *La protéine S reconnaît un récepteur cellulaire (SARS likes ACE2)*

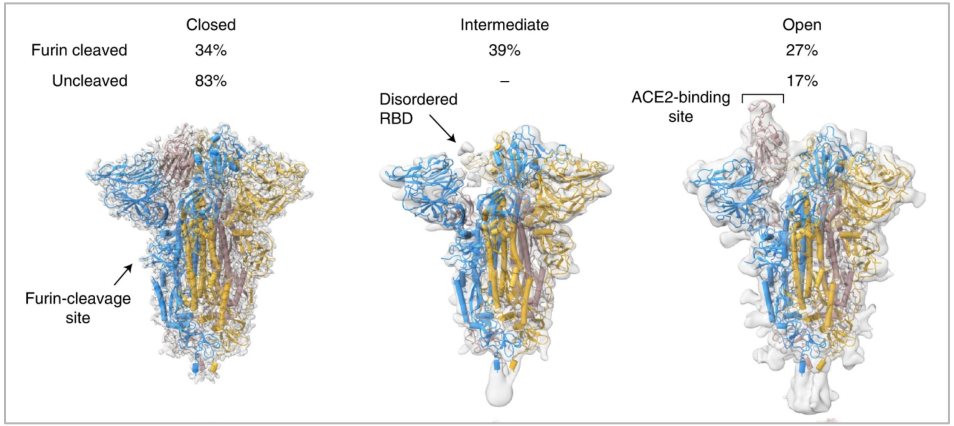


# *La protéine S doit être activée par un clivage protéolytique (priming)*

- Juste après la traduction, le produit du gène S est une protéine inactive.
- L'activation requiert le clivage (coupure) de la protéine en deux parties (S1, S2).
- Chez SARS-CoV-2, ce clivage est réalisé par une enzyme appelée furine.
- Ceci a une grave conséquence, car cette enzyme est ubiquitaire dans les cellules humaines, ce qui explique en partie que les symptômes de la Covid-19 ne se limitent pas aux voies respiratoires.



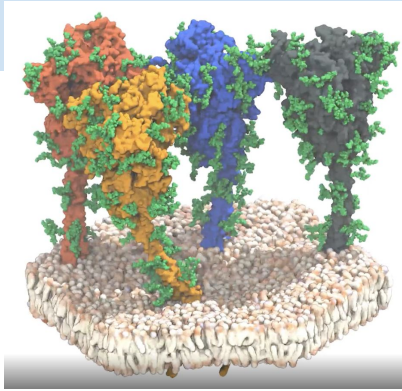

Sallard, E., Halloy, J., Casane, D., van Helden, J. & Decroly, É. 2020. **Retrouver les origines du SARS-CoV-2 dans les phylogénies de coronavirus**. Med Sci (Paris) 36: 783–796




Wrobel, A.G., Benton, D.J., Xu, P., Roustan, C., Martin, S.R., Rosenthal, P.B., et al. 2020. SARS-CoV-2 and bat RaTG13 spike glycoprotein structures inform on virus evolution and furin-cleavage effects. Nat Struct Mol Biol 27: 763–767.

Coutard, B., Valle, C., de Lamballerie, X., Canard, B., Seidah, N.G. & Decroly, E. 2020. The spike glycoprotein of the new coronavirus 2019-nCoV contains a furin-like cleavage site absent in CoV of the same clade. Antiviral Res. 176: 104742.

# *Structure de la protéine S  & dynamique moléculaire*

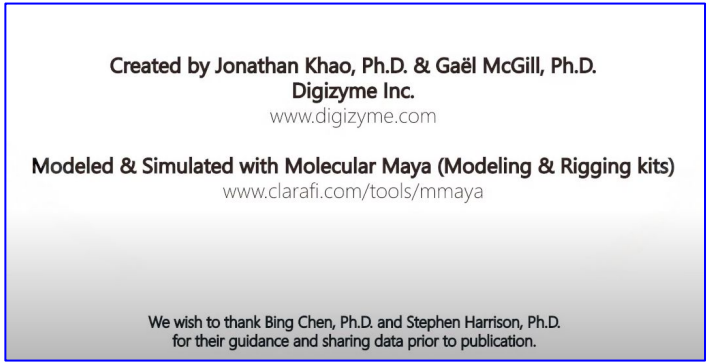


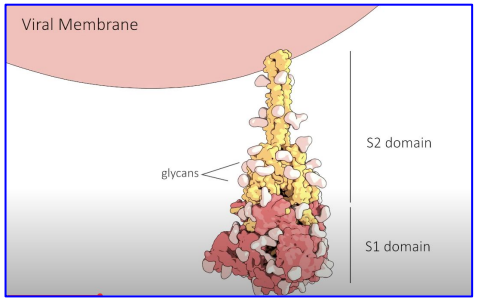
https://www.nature.com/immersive/d41586-020-02493-0/assets/bCb4ohBwz0/pia23468_1.mp4

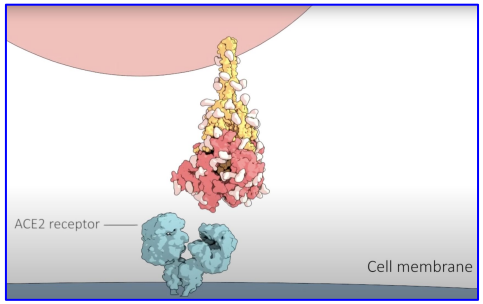
# *Simulation moléculaire de l'interaction spicule - récepteur*

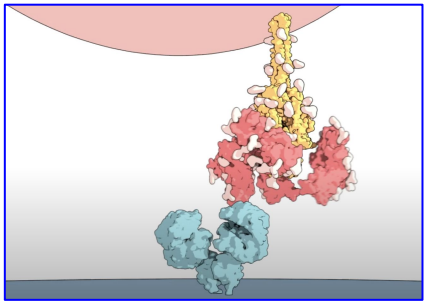
https://youtu.be/e2Qi-hAXdJo?t=12

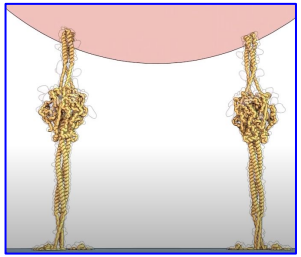
Un modèle visuel de simulation dynamique illustrant la façon dont la protéine de spicule (spike protein) assure la fusion des membranes virale et cellulaire.

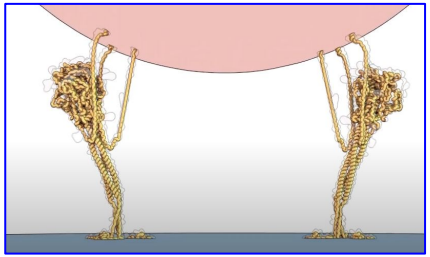


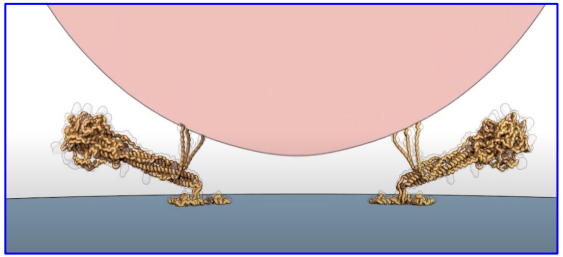



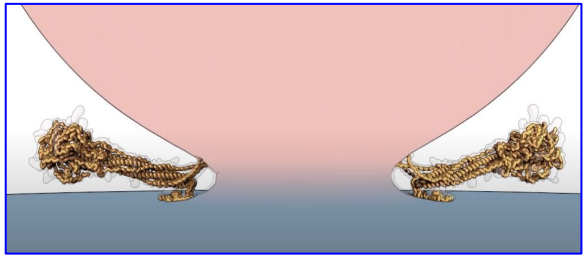



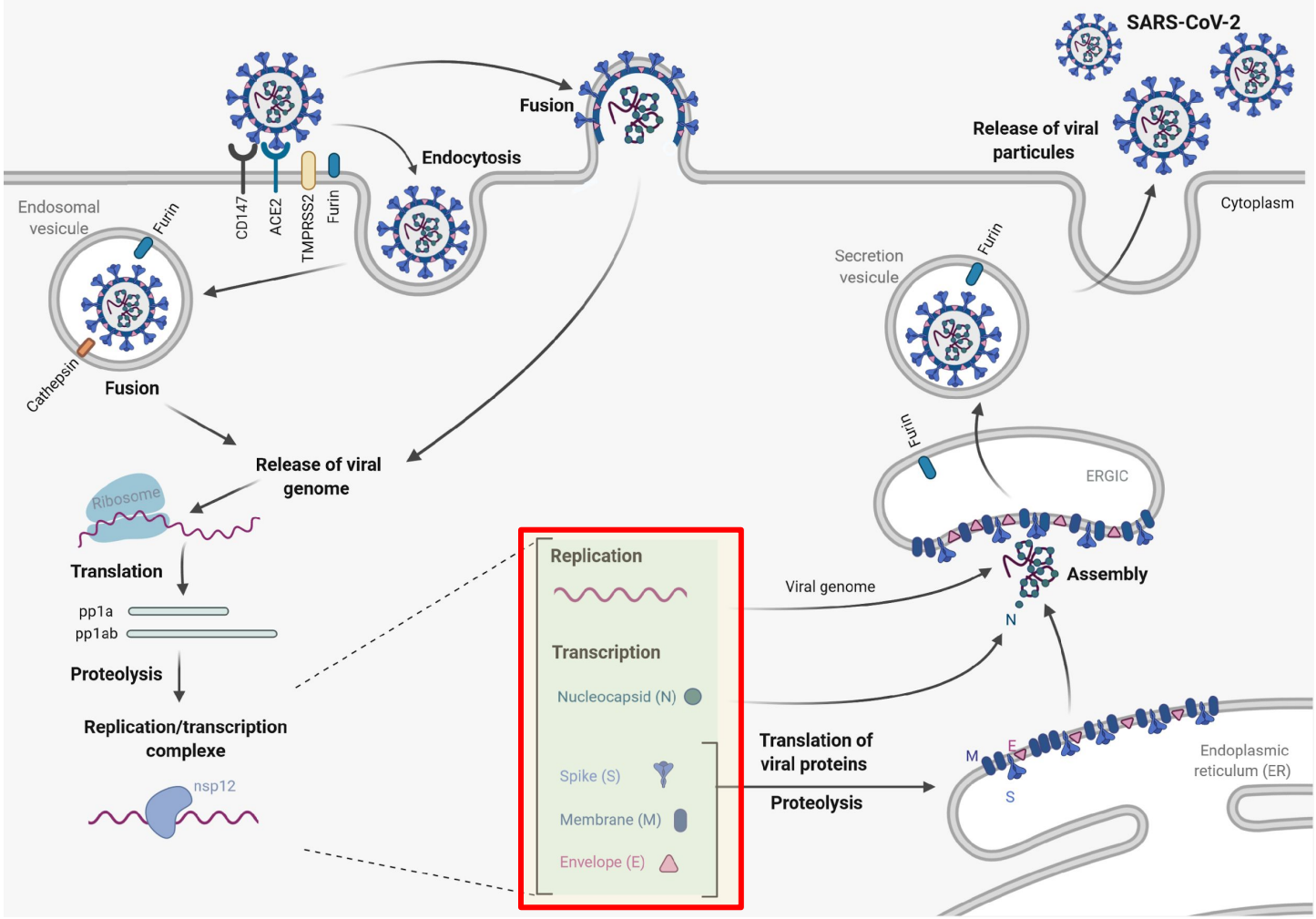








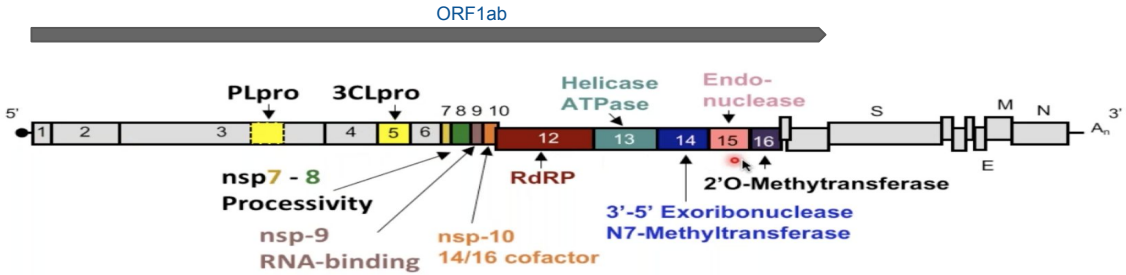
# Cycle de réplication des coronavirus - Réplication et transcription

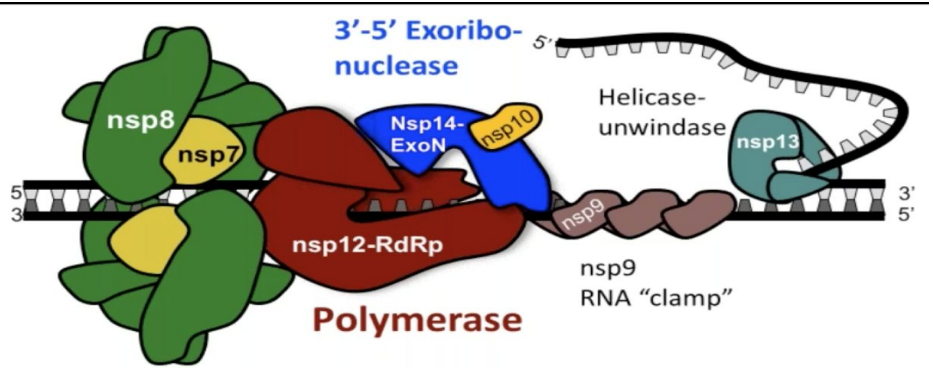


# *Le complexe réplication/transcription*

**SARS-CoV-2**


**SARS-CoV**



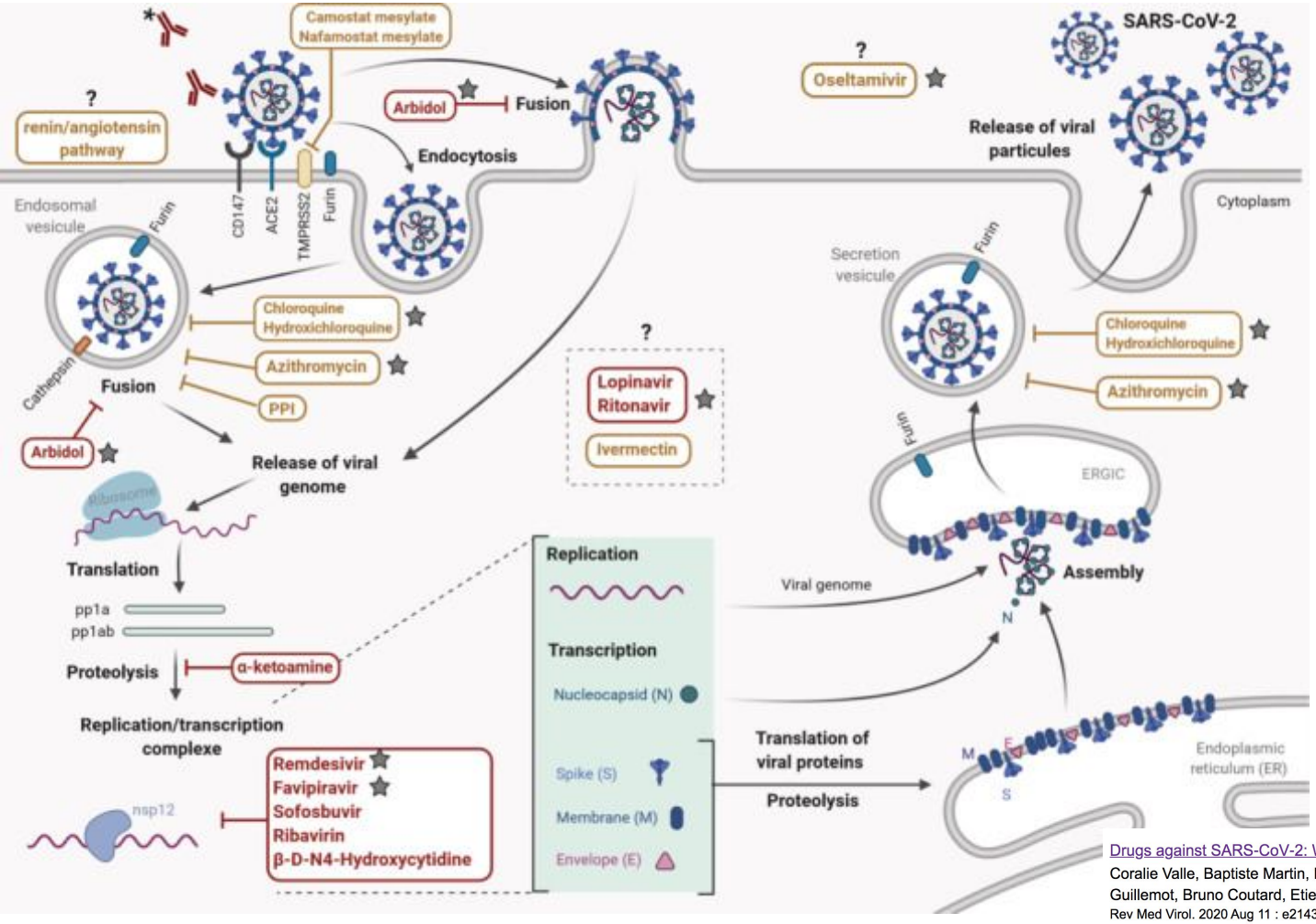




- Le gène ORF1ab code pour une "polyprotéine", qui contient 16 protéines distinctes
- Une dizaine de ces protéines forment un complexe (figure du bas) qui assure la réplication et la transcription de l'ARN.

# *SARS-CoV : le complexe réplication/transcription*

**Structure** of **replicating SARS-CoV-2 polymerase**. Hillen HS, Kokic G, Farnung L, Dienemann C, Tegunov D, Cramer P. Nature. 2020 Aug;584(7819):154-156.

# *Mécanismes d'action des médicaments contre le SARS-CoV-2*



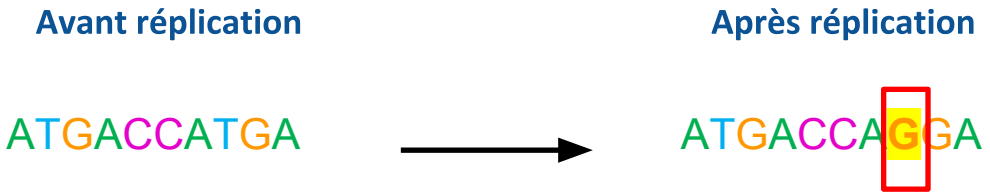

- Des études sont menées dans un grand nombre de laboratoires pour identifier des molécules qui pourraient bloquer l'infection par SARS-CoV-2.
- Ces études partent de médicaments antiviraux connus, qui interagissent à différents niveau du cycle infectieux des virus.

Drugs against SARS-CoV-2: What do we know about their mode of action?
Coralie Valle, Baptiste Martin, Franck Touret, Ashleigh Shannon, Bruno Canard, Jean-Claude Guillemot, Bruno Coutard, Etienne Decroly
Rev Med Virol. 2020 Aug 11 : e2143. doi: 10.1002/rmv.2143 [Epub ahead of print]

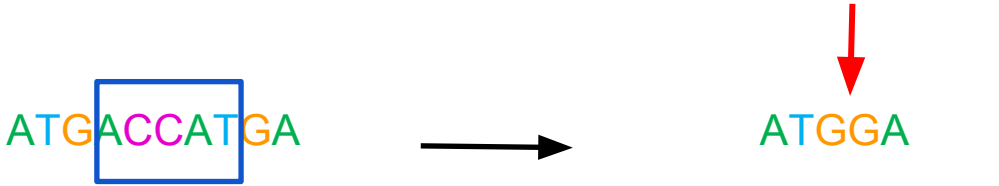
# *Événements évolutifs*

# *Typologie des mutations*

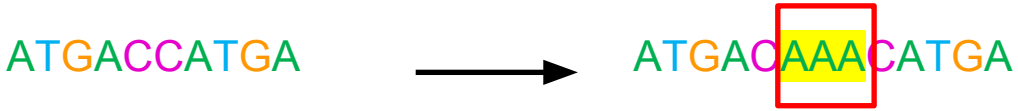
- Substitution
  - Remplacement d'un résidu (une lettre) par un autre




- Délétion
  - Perte d'un fragment de la molécule

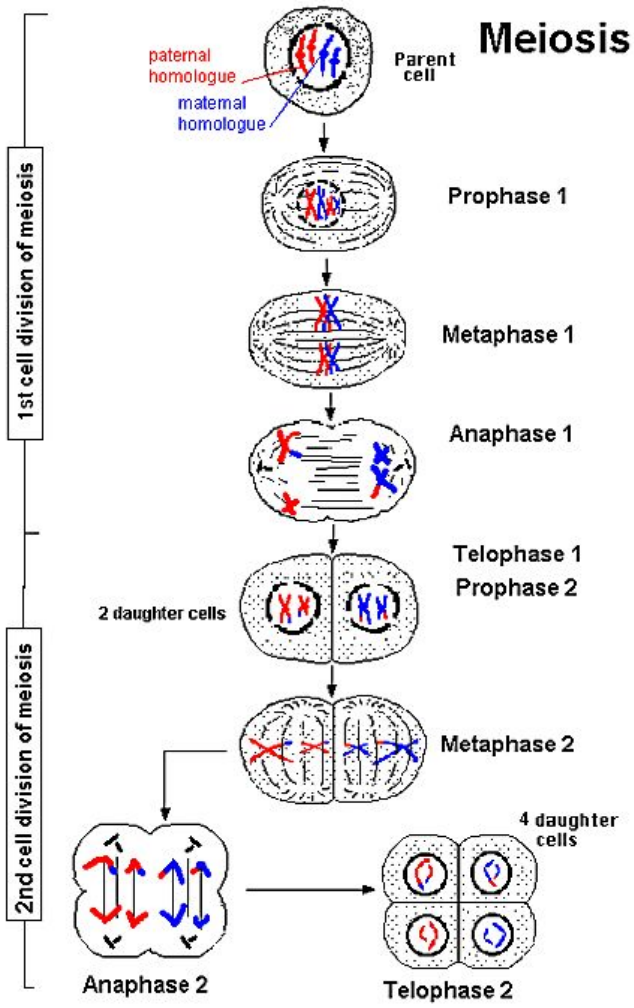



- Insertion
  - Ajout d'un fragment de molécule



# Recombinaison

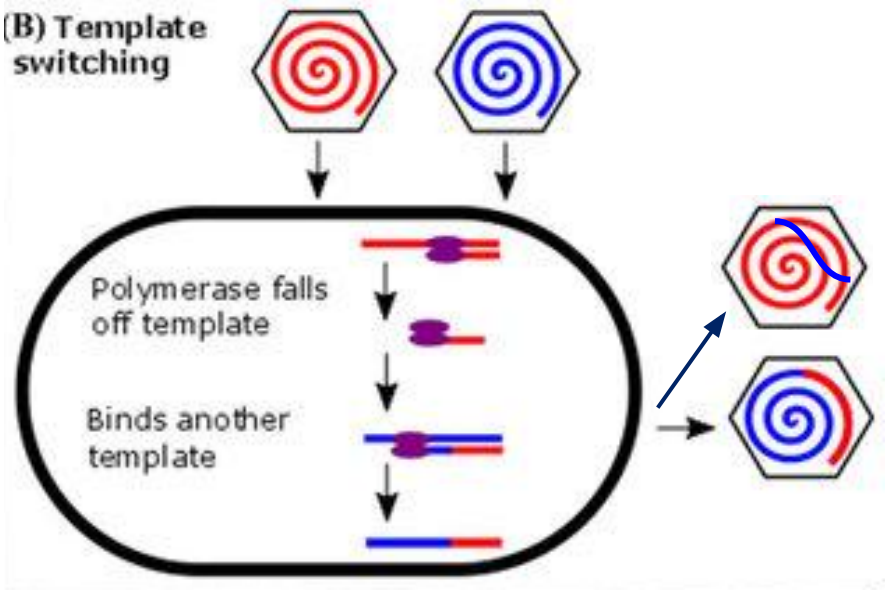
- Chez les organismes cellulaires, lors de la méiose, une cellule mère diploïde forme 4 cellules-filles haploïdes.
- Les chromosomes parentaux sont distribués aléatoirement et de façon indépendante entre les 4 cellules-filles.
- Des événements de “crossing-over” (croisements) provoquent une recombinaison de fragments de chromosomes.
- La liaison génétique entre les gènes d'un même chromosome n'est pas complète.




http://www.accessexcellence.org/AB/GG/

# *Recombinaisons chez les coronavirus*

- Une chauve-souris peut se retrouver infectée par plusieurs coronavirus en même temps.
- Pendant la réplication des coronavirus, il arrive que la polymérase de l'ARN "saute" d'un virus à l'autre.
- Ceci donne naissance à un virus "chimérique", dont le génome est composé de fragments d'origines différentes.
- Ceci complique l'analyse de la phylogénie des virus, car différents fragments génomiques résultent d'histoires évolutives différentes.

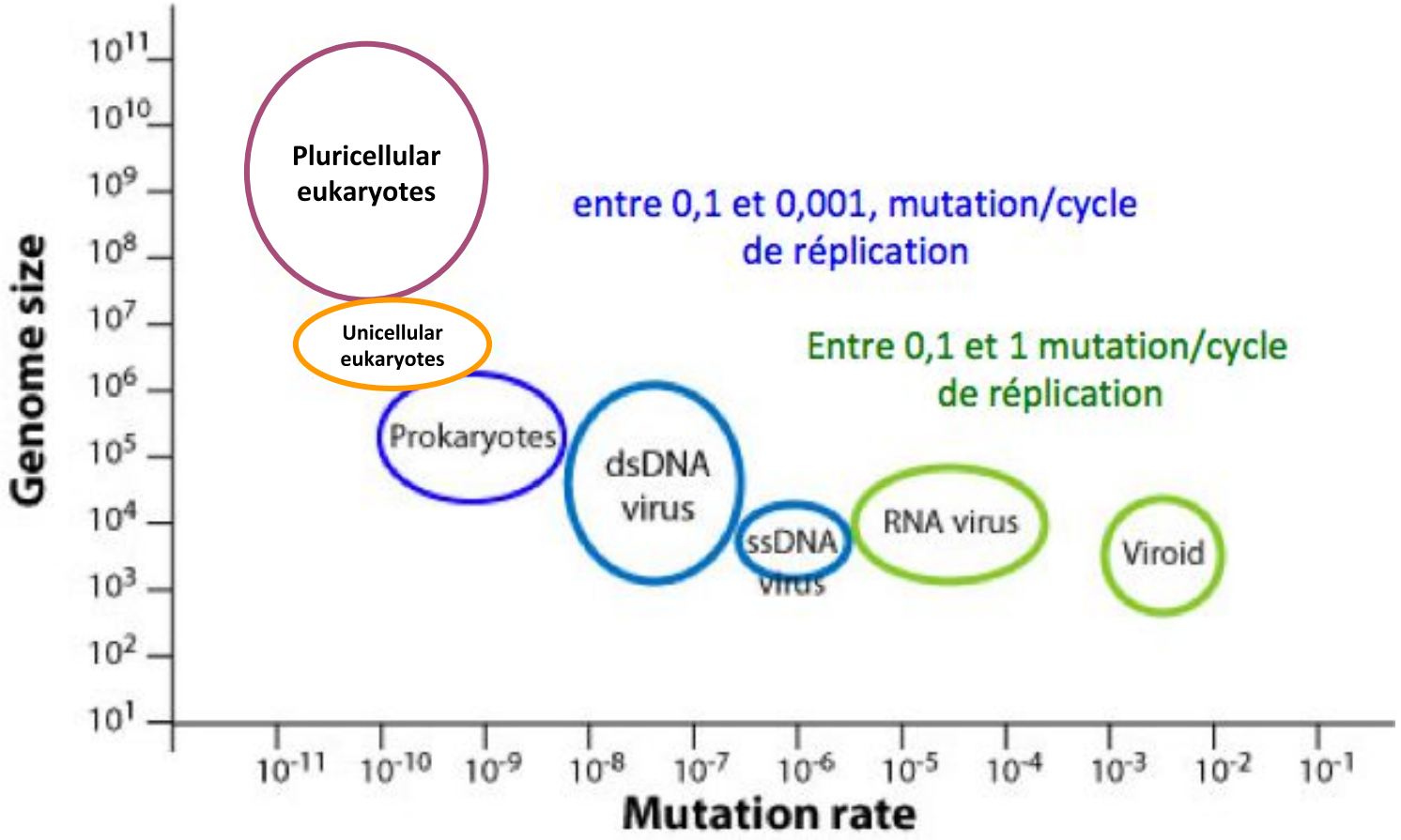



Discovery of a rich gene pool of bat SARS-related coronaviruses provides new insights into the origin of SARS coronavirus

Ben Hu ,et al Plos path : November 30, 2017

J. Dennehy, Evolutionary ecology of virus emergence: Virus emergence, 2016, Annals of the New York Academy of Science

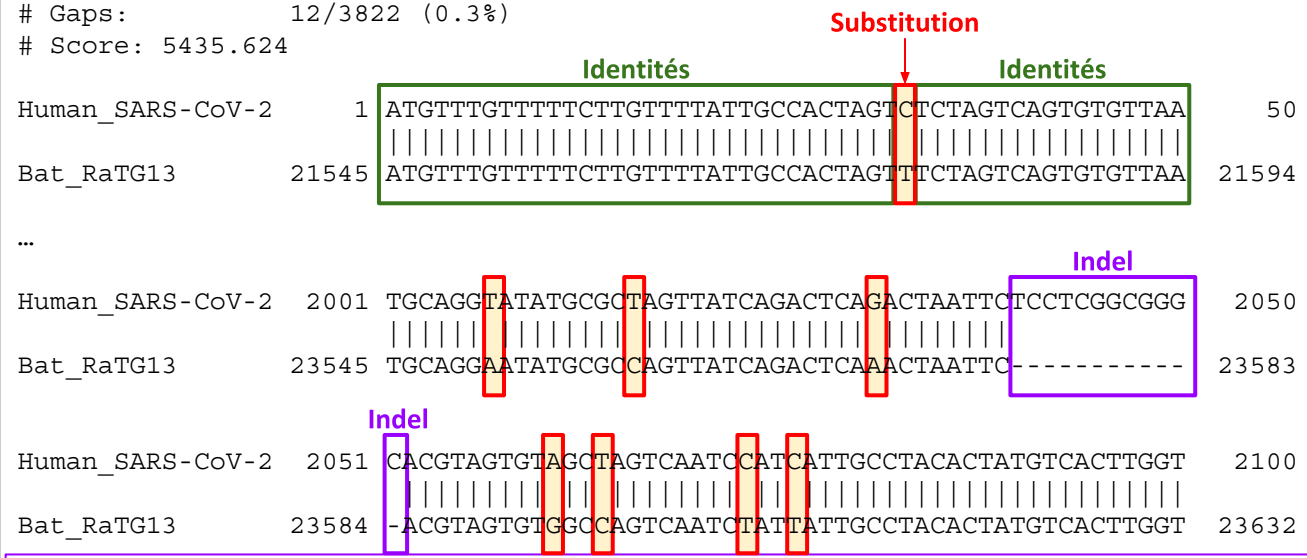
# *Taux de substitutions représentatifs chez différents groupes taxonomiques*




Adapté d'après une diapo d'Etienne Decroly

# *Alignement de séquences – Gènes S de SARS-CoV-2 et RaTG13*

```
# Aligned_sequences: 2
# 1: Human_SARS-CoV-2_BetaCoV/Wuhan/IPBCAMS-WH-01/2019
# 2: Bat_RaTG13
#
# Length: 3822
# Identity:    3549/3822 (92.9%)
# Similarity:    NA/3822 (NA%)
# Gaps:           12/3822 (0.3%)
# Score: 5435.624
```

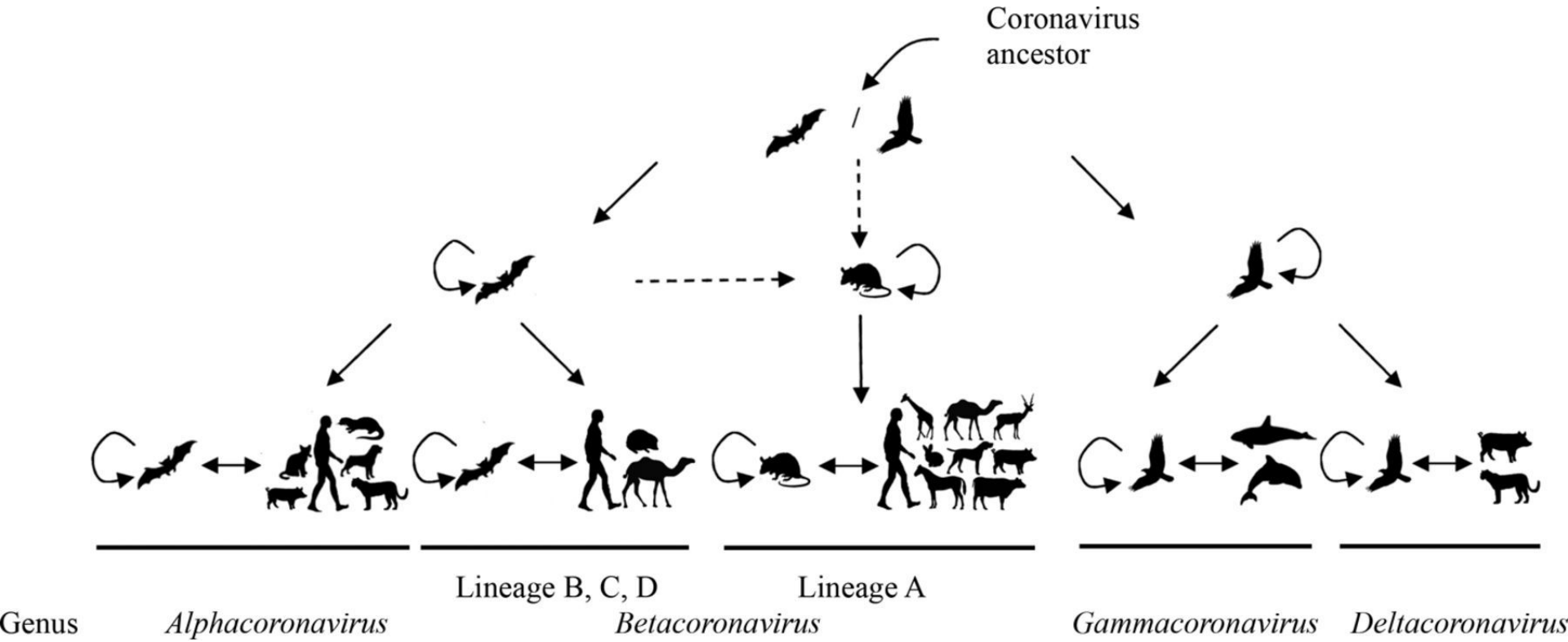



**Note**

- "**Indel**" signifie "Insertion ou délétion"
- Sur base de ce résultat, la différence observée peut provenir soit d'une insertion chez un ancêtre de SARS-CoV-2, soit d'une délétion chez un ancêtre de RaTG13

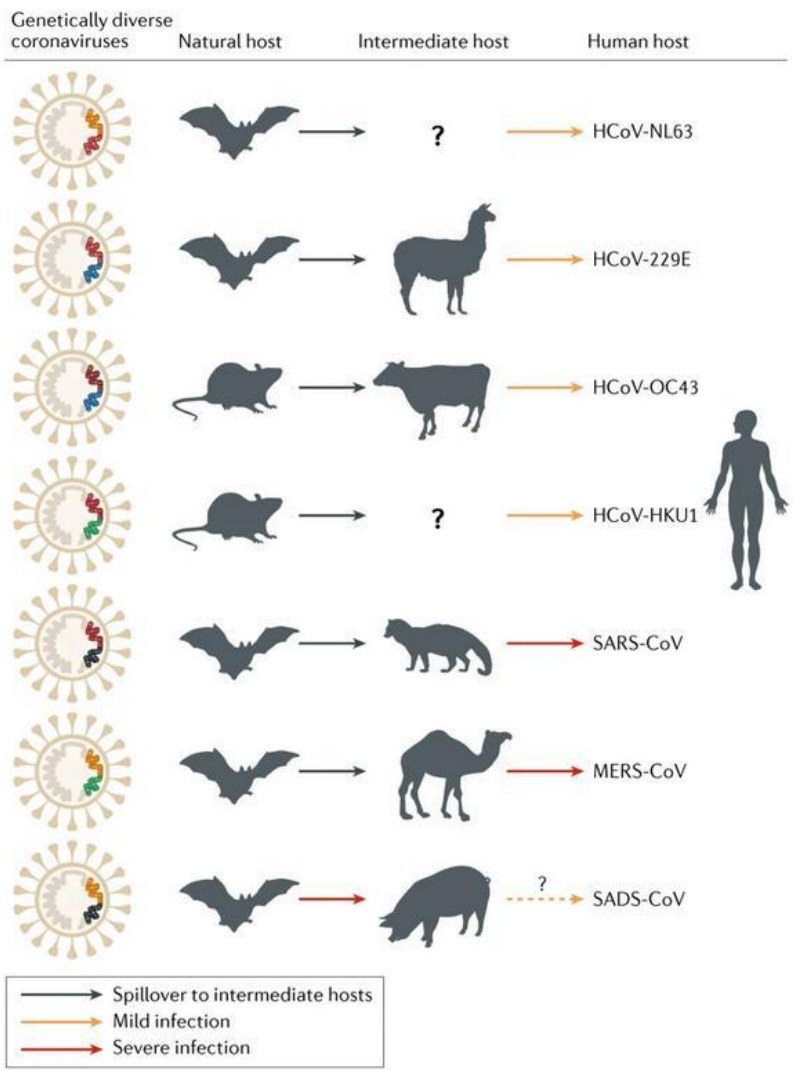
*Des chauves-souris et des hommes*

# *Cycle zoonotique des coronavirus*



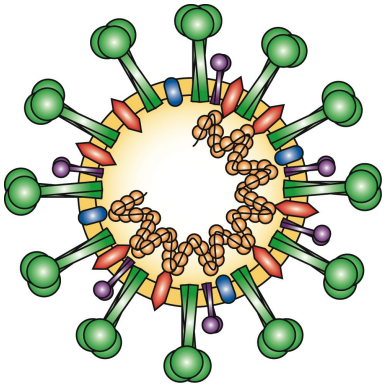

- Modèle du “débordement”
- Modèle de la circulation de « quasi espèces”

# *Mécanismes d'émergence des CoV humains*




Origin and evolution of pathogenic coronaviruses. Jie Cui, Fang Li & Zheng-Li Shi. Nature Reviews Microbiology volume 17, pages 181–192(2019)

# *Emergences de coronavirus humains*



Common cold
OC43 can infect lower respiratory track

HKU1; Pneumonia
NL63; Bronchiolitis

**SARS-CoV**

**MERS-CoV**

**2019-nCoV**

229E
OC43
HKU1
NL63

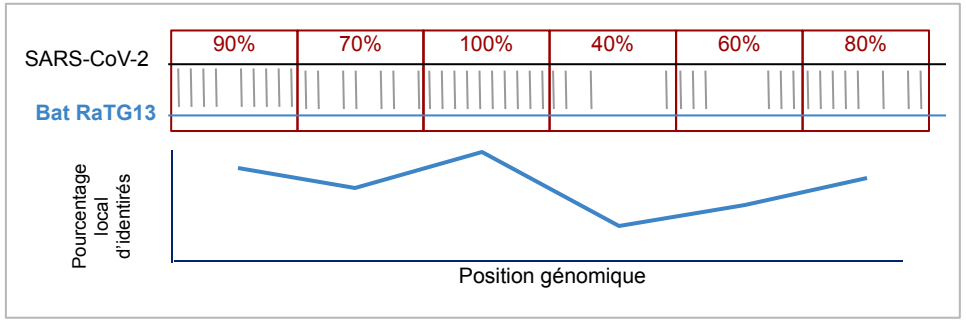
1967/1970 — 1980 — 2002 — 2004/2005 — 2012 — 2019

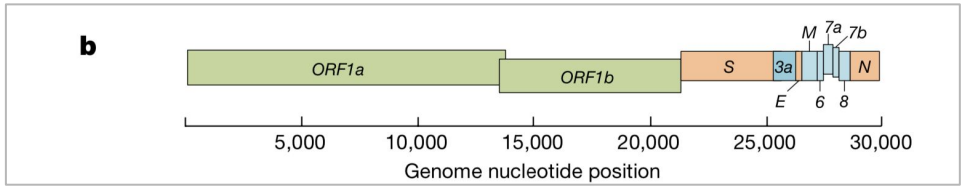
OC43 genome similar to Bovine coronavirus

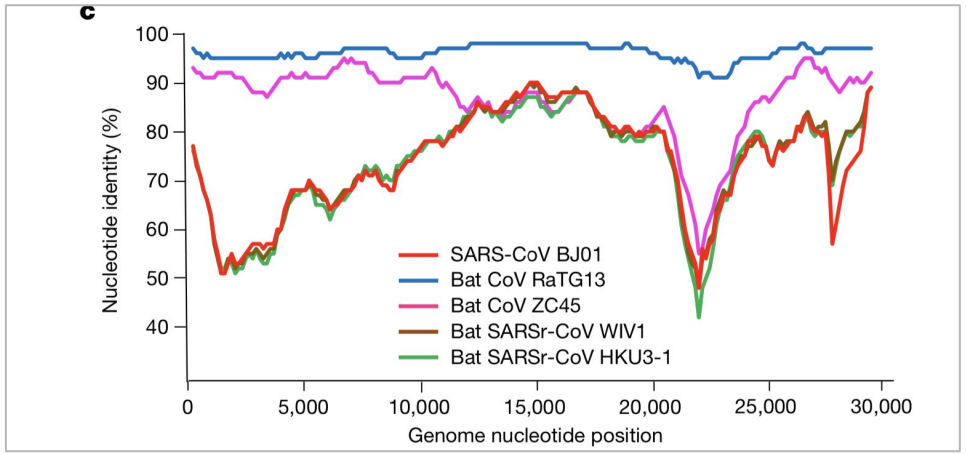
**SARS-CoV. MERS-CoV, 2019-nCoV**
**Severe respiratory disease**

# Profils de positions identiques (PPI)

- Haut: principe de calcul du PPI
  - alignement d'une paire de séquences
  - découpage de la séquence en "fenêtres"
  - calcul du pourcentage local d'identité de chaque fenêtre
  - dessin du profil de positions identiques (PPI)
- Milieu : positions des gènes de SARS-CoV-2 sur le génome
- Bas: PPI de quelques génomes de coronavirus sur celui de SARS-CoV-2
- Commentaires dans la diapo suivante



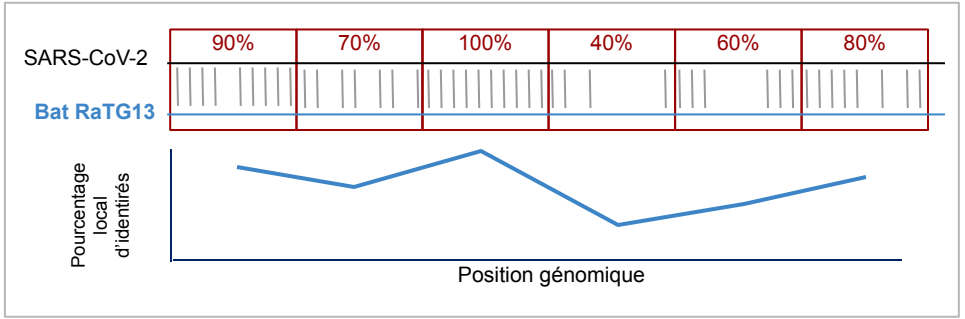



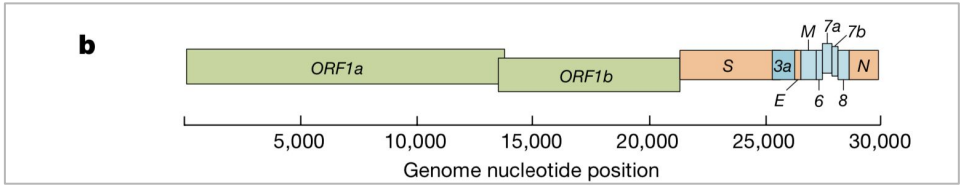



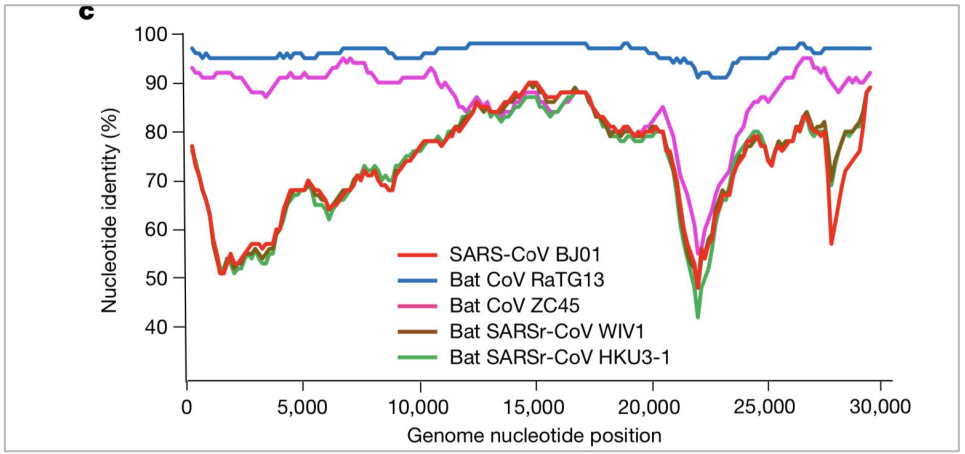

Zhou, P., Yang, X.-L., Wang, X.-G., Hu, B., Zhang, L., Zhang, W., et al. 2020. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature 579: 270–273.

# Profils de positions identiques (PPI) de génomes de coronavirus

- RaTG13 (virus de chauve-souris) est le génome le plus proche de SARS-CoV-2
- On a identifié d'autres virus de chauve-souris relativement proches de SARS-CoV-2 (Cov ZC45)
- Les virus SARS-CoV humains (pandémie 2002-2003) sont moins proches
- Pour chaque espèce, on observe des fluctuations le long du profil de PPI
- Entre 22.000 et 25.000 : chute brutale des PPI




Zhou, P., Yang, X.-L., Wang, X.-G., Hu, B., Zhang, L., Zhang, W., et al. 2020. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature 579: 270–273.

# *Une origine probable: la chauve-souris*

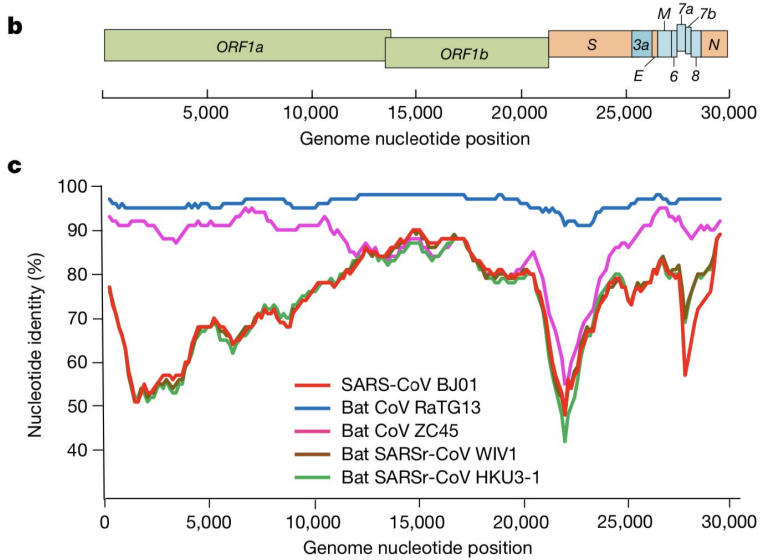
- 3 février 2020: publication du génome complet de SARS-CoV-2
- Recherche de virus similaires dans les bases de données de séquence
  - Les virus les plus proches sont des virus de chauves-souris (Bat CoV ZC45)
- Dans le même article, les auteurs décrivent un nouveau génome de chauve-souris: **RaTG13**
  - A ce jour la souche virale la plus proche de SARS-CoV-2 connue
- Figures du bas: profil de positions identiques (PPI), expliqué ci-après.

**Article**

## A pneumonia outbreak associated with a new coronavirus of probable bat origin

https://doi.org/10.1038/s41586-020-2012-7
Received: 20 January 2020
Accepted: 29 January 2020
Published online: 3 February 2020
Open access

Peng Zhou[1,5], Xing-Lou Yang[1,5], Xian-Guang Wang[2,5], Ben Hu[1], Lei Zhang[1], Wei Zhang[1], Hao-Rui Si[1,3], Yan Zhu[1], Bei Li[1], Chao-Lin Huang[2], Hui-Dong Chen[2], Jing Chen[1,3], Yun Luo[1,3], Hua Guo[1,3], Ren-Di Jiang[1,3], Mei-Qin Liu[1,3], Ying Chen[1,3], Xu-Rui Shen[1,3], Xi Wang[1,3], Xiao-Shuang Zheng[1,3], Kai Zhao[1,3], Quan-Jiao Chen[1], Fei Deng[1], Lin-Lin Liu[4], Bing Yan[1], Fa-Xian Zhan[4], Yan-Yi Wang[1], Geng-Fu Xiao[1] & Zheng-Li Shi[1]✉




Zhou, P., Yang, X.-L., Wang, X.-G., Hu, B., Zhang, L., Zhang, W., et al. 2020. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature 579: 270–273.

# *De Yunnan à Wuhan*

- 2013
  - Pneumonie atypique chez 6 mineurs dans la province de Yunnan, 3 décès
  - Plusieurs pistes sont évoquées (levures, virus) dont un coronavirus
  - Collecte d'échantillons de chauves-souris dans la mine
- 2016 : publication d'un fragment de séquence (360 nucléotides, 1% du génome) de virus de *Rhinolophus affinis*
- 2018 : dépôt des fragments de séquençage (reads) dans une base de données, pour ~90% du génome
- 2020 :
  - publication de la séquence complète du génome viral, sous l'identifiant "**BatCoV RaTG13**".
  - Ce génome est le plus proche connu de celui de SARS-CoV-2 (96,2%  nucléotides identiques).
- Note: l'implication de coronavirus dans le décès des mineurs fait encore l'objet de débats



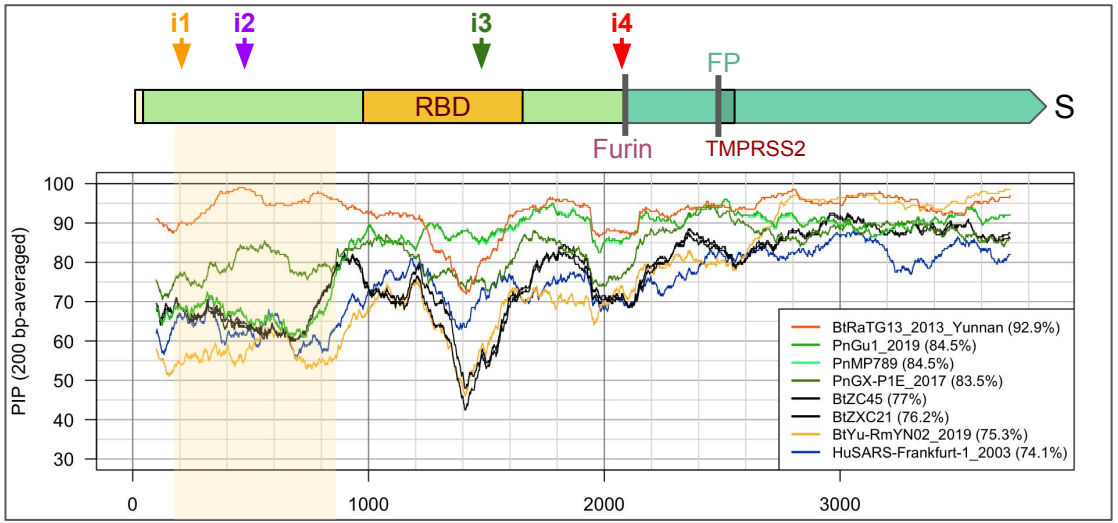


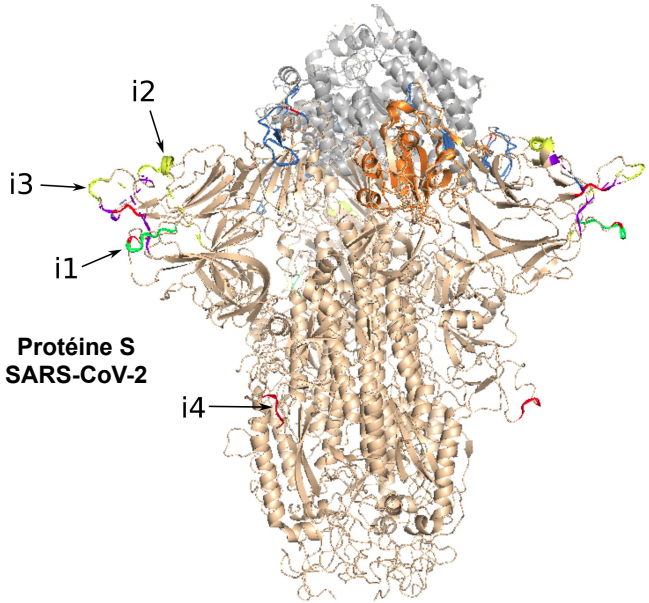

Zhou, P., Yang, X.-L., Wang, X.-G., Hu, B., Zhang, L., Zhang, W., et al. 2020. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature 579: 270–273.

*Insertions dans les séquences du gène S*

# *Quatre insertions dans le gène S de SARS-CoV-2*

- Les flèches indiquent la position des 4 insertions sur le gène S (gauche) et sur la protéine spicule (droite).
- Les 3 premières sont situées à l'extérieur de la protéine, dans des régions "exposées".



# *Le site de clivage par la furine*

- Le site de clivage par la furine qu'on observe dans la protéine spicule de SARS-CoV-2 ne se trouve dans aucun autre coronavirus.
- Il résulte de l'insertion de 12 nucléotide à un endroit particulier du le gène S.

EL PAÍS

CORONAVIRUS

ccu cgg cgg gca

The 12 letters that changed the world

The genome of the new coronavirus harbors a short sequence suspected of being the main culprit of its uniquely infectious and aggressive nature

MANUEL ANSEDE | ARTUR GALOCHA | MARIANO ZAFRA

19 MAY 2020 - 18:25 CEST

Coutard, B., Valle, C., de Lamballerie, X., Canard, B., Seidah, N.G. & Decroly, E. 2020. The spike glycoprotein of the new coronavirus 2019-nCoV contains a furin-like cleavage site absent in CoV of the same clade. Antiviral Res. 176: 104742.

# *Des insertions bizarres?*

- Figure from Pradhan et al (2020), initially published on bioRxiv and retracted.
- The “multiple alignment” is actually a pairwise alignment + a consensus.
- The gaps obtained from a multiple alignment overlap with these ones, but they start and end at different positions.
- It is precisely because they did not do a multiple alignment that they did not realize that 3 of these insertions were not unique to SARS-CoV-2.

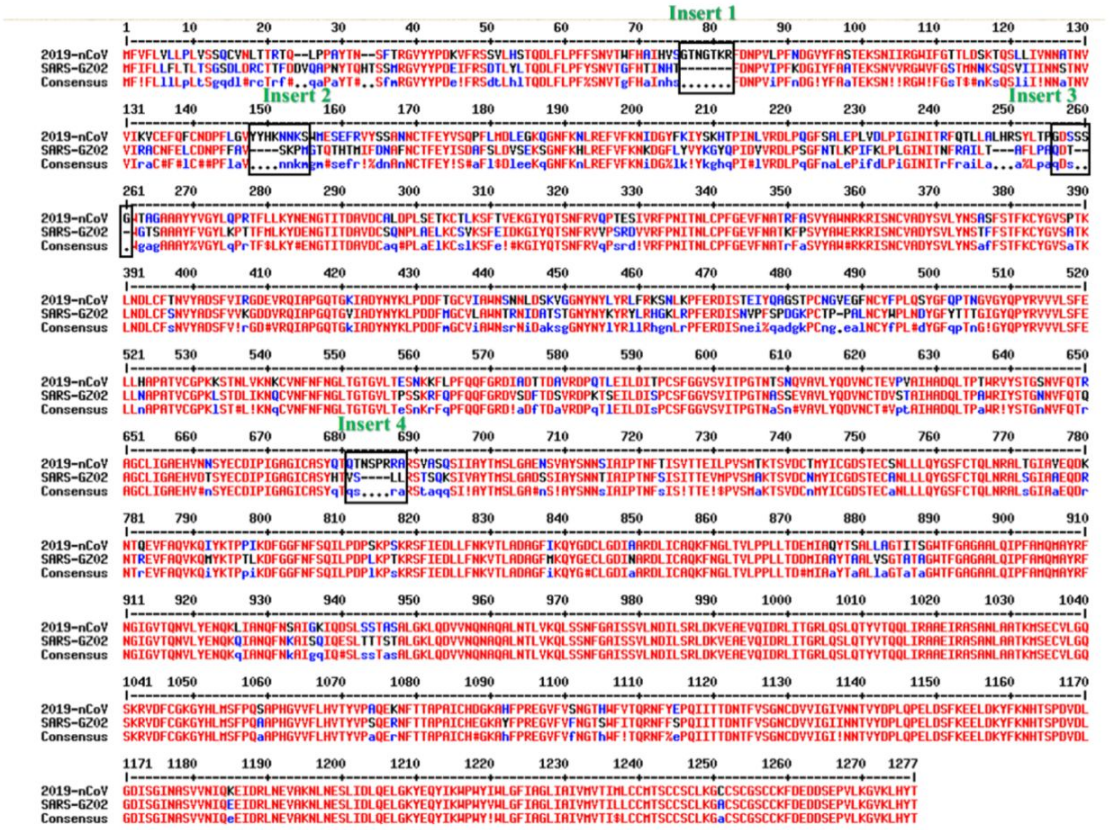
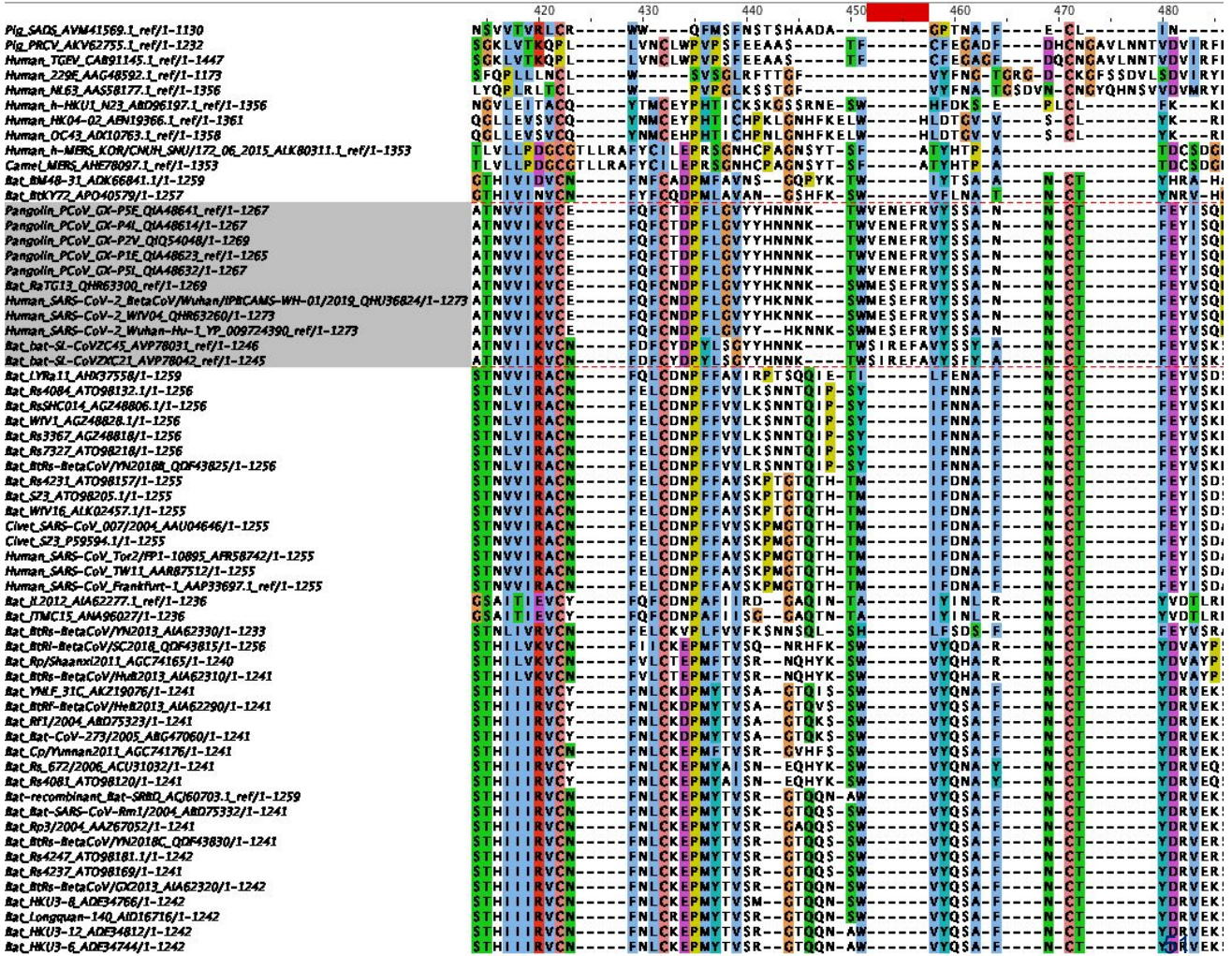



**Figure 2: Multiple sequence alignment between spike proteins of 2019-nCoV and SARS.** The sequences of spike proteins of 2019-nCoV (Wuhan-HU-1, Accession NC_045512) and of SARS CoV (GZ02, Accession AY390556) were aligned using MultiAlin software. The sites of difference are highlighted in boxes.

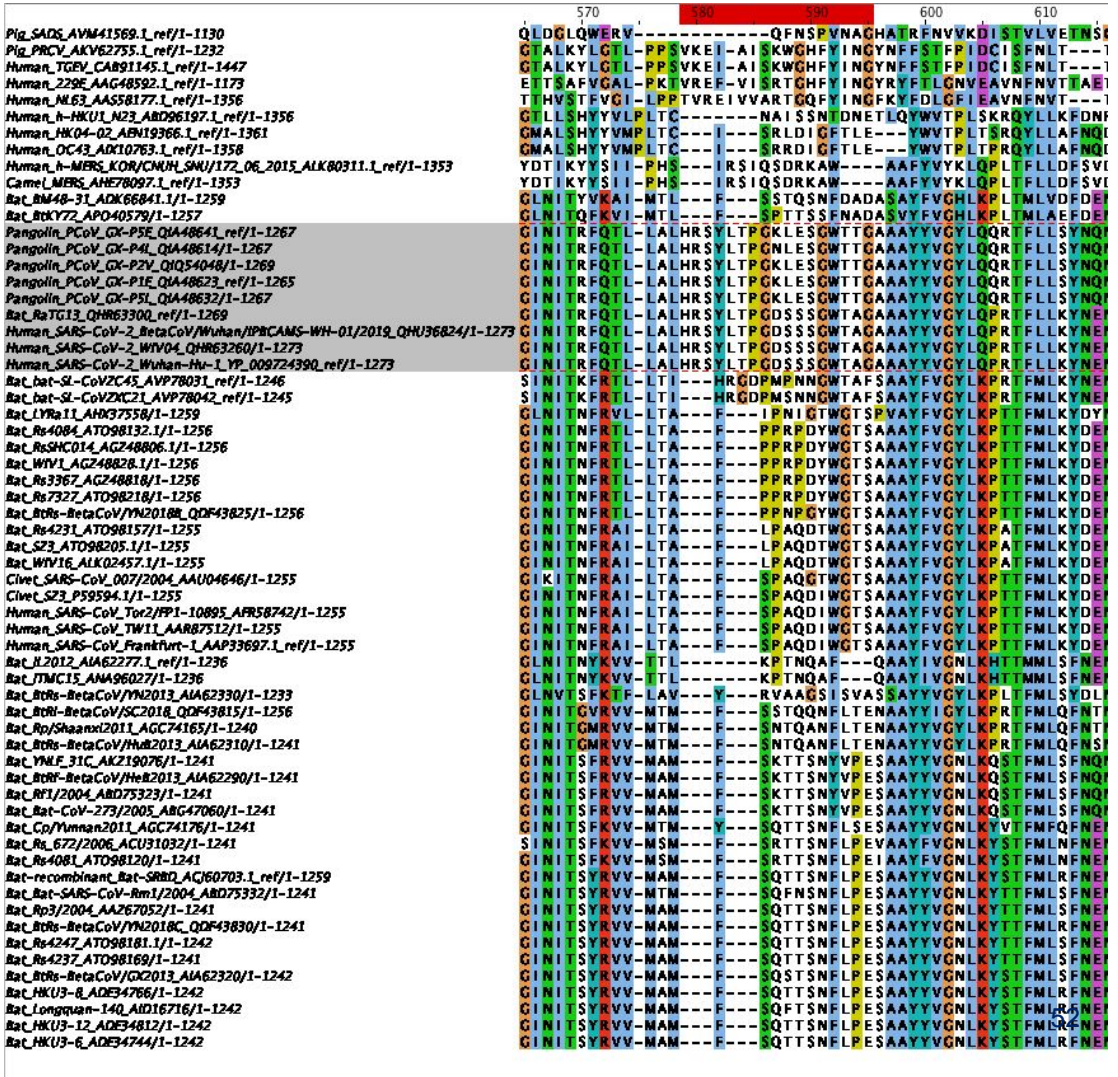
# *Insertion partagée entre tous les virus du groupe CoV-2*

- Position: 153-158 de SARS-CoV-2
- Cette insertion se trouve chez les virus de pangolin + plusieurs chauve-souris
- Les résidus sont identiques entre SARS-CoV-2 et la souche RaTG13 de chauve-souris (la plus proche de SARS-CoV-2)
- Par contre elle présente 3 substitutions entre les souches de pangolin et SARS-CoV-2.



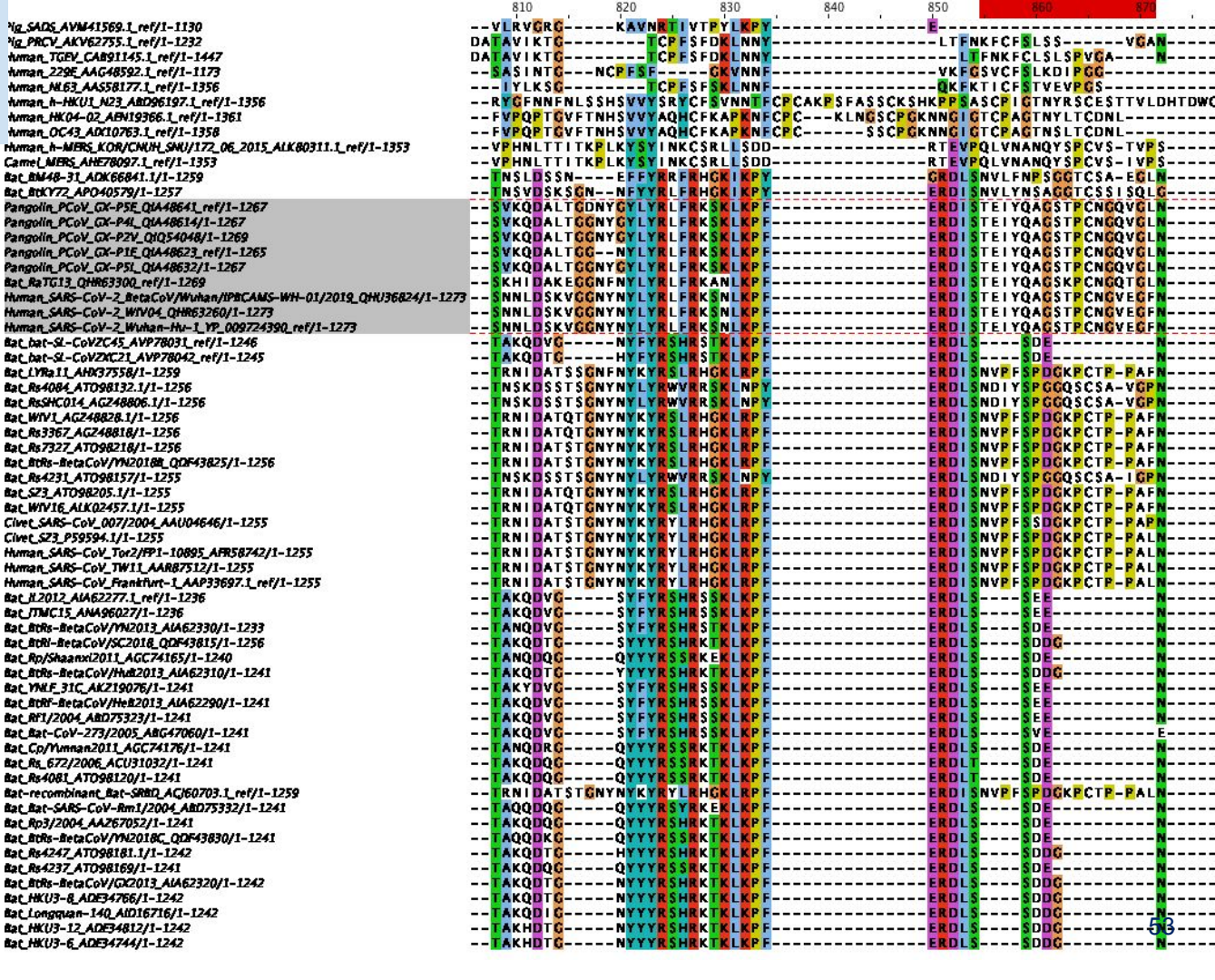
# *Insertion partagée par la majorité des virus du groupe CoV-2*

- Position: 245-251 de SARS-CoV-2
- Cette insertion se trouve chez les virus de pangolin + la souche RaTG13 de chauve-souris
- Elle est cependant absente de 2 souches de chauves-souris appartenant au groupe CoV-2 : CoVZC45 et CoVZXC21
- Au-delà de l'insertion on trouve un bloc conservé (jusqu'à la position 595 de l'alignement).
- Au sein de ce bloc, une paire de résidus distingue les pangolins du groupe SARS2 + Bat RaTG13 .



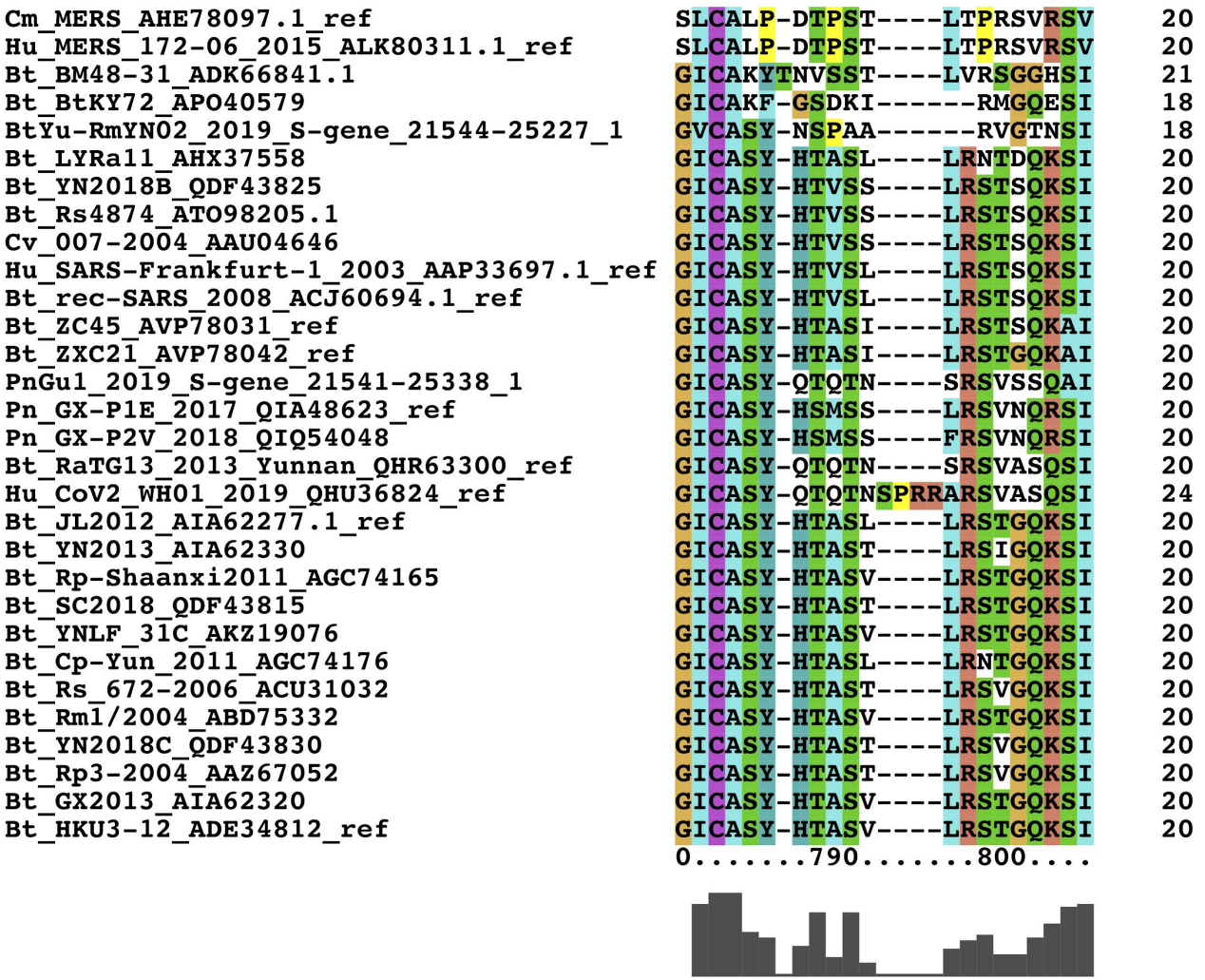
# *Insertion i3*

- Position
  - 470-486 de SARS-CoV-2
  - 855-872 sur l'alignement
- Commune au groupe pangolin + Bat_RaTG13 + SARS-CoV-2
- 2 substitutions uniques à Bat_RaTG13



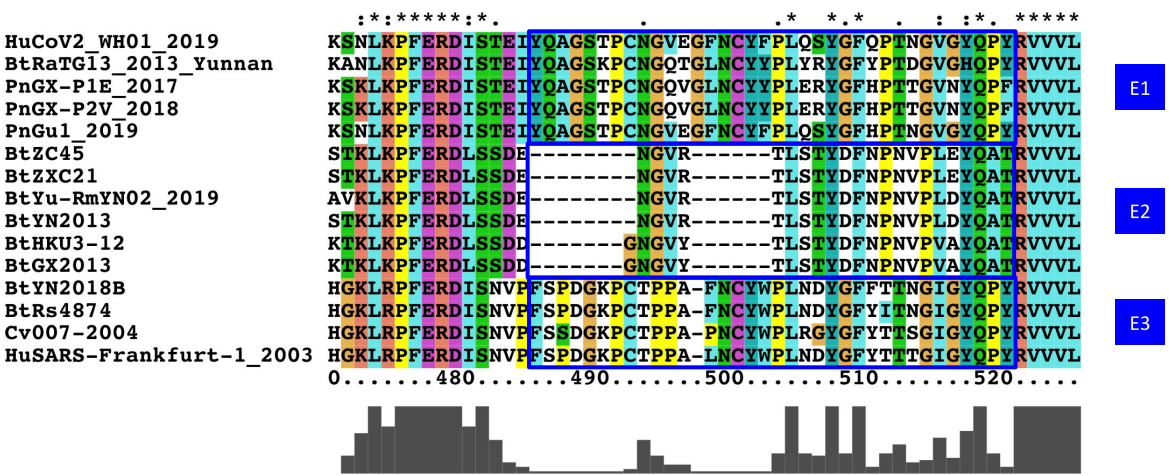
# *Insertion d'un site Furine (i4)*

- Positions : 1181-1184 de l'alignement
- On trouve chez SARS-CoV-2 un site unique SPRRAR, qui résulte d'une insertion SPRR et d'une substitution L -> A
- La séquence PRRA correspond au motif reconnu par la furine (protéase).
- Cette insertion est à l'origine du site de clivage responsable du caractère particulièrement virulent de SARS-CoV-2

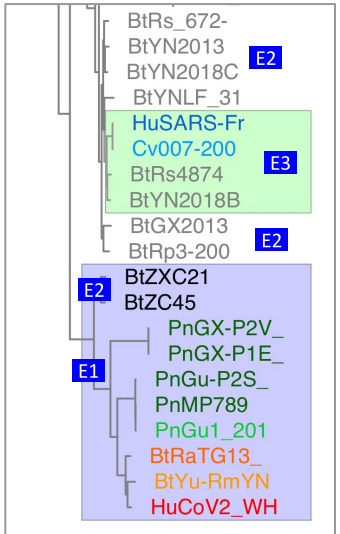


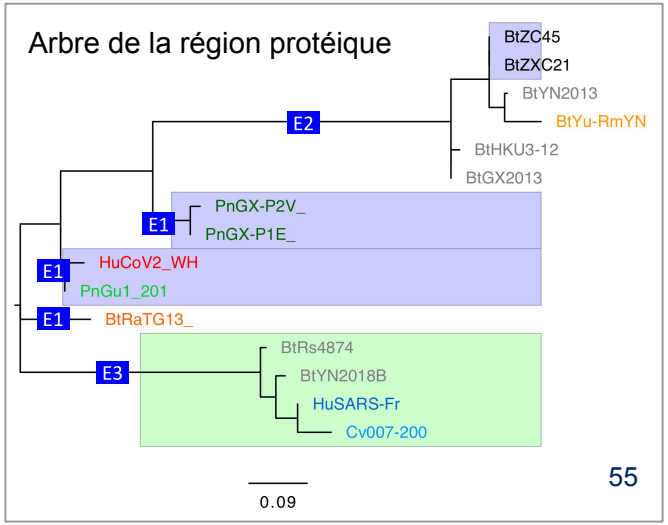
# *Un site recombinant?*

- L'interprétation de ce site est plus complexe.
- Il existe clairement trois groupes de séquences.
- Ceux-ci s'étendent au-delà des deux indels.
- L'arbre construit à partir de cette région est peu robuste, et incohérent avec celui des génomes.
- La répartition des sous-blocs de séquences est plus cohérente avec l'arbre des espèces.
- Cette région a échappé à Pradhan et al. parce qu'ils ont réalisé un alignement par paire SARS-CoV-2 vs SARS plutôt qu'un alignement multiple.




Arbre des génomes



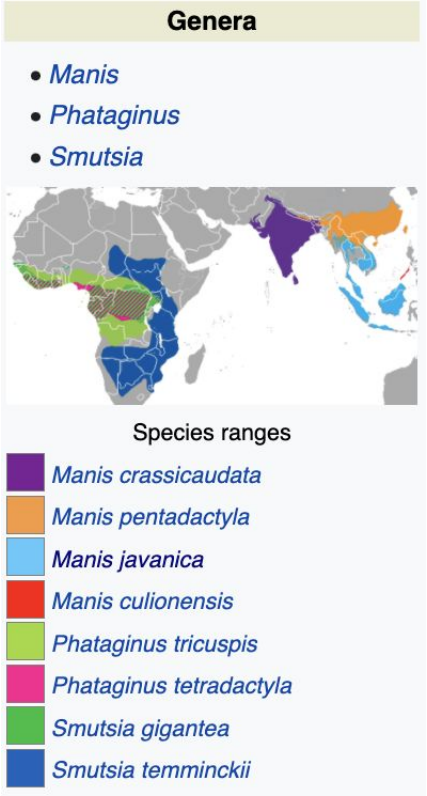



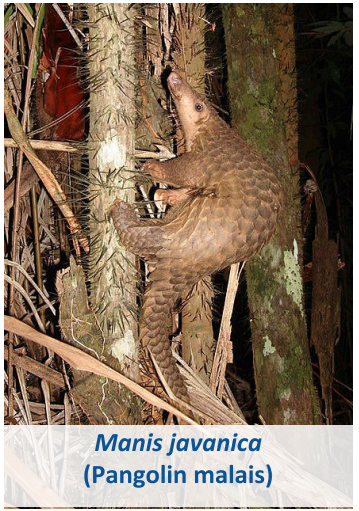
*Des chauves-souris et des hommes … et des pangolins ?*

# *Pangolins*

- Mode de vie
- Aire géographique
- Migration
- Contacts avec les chauves-souris
- Contacts avec l'humain

**Répartition géographique de différentes espèces de pangolin**



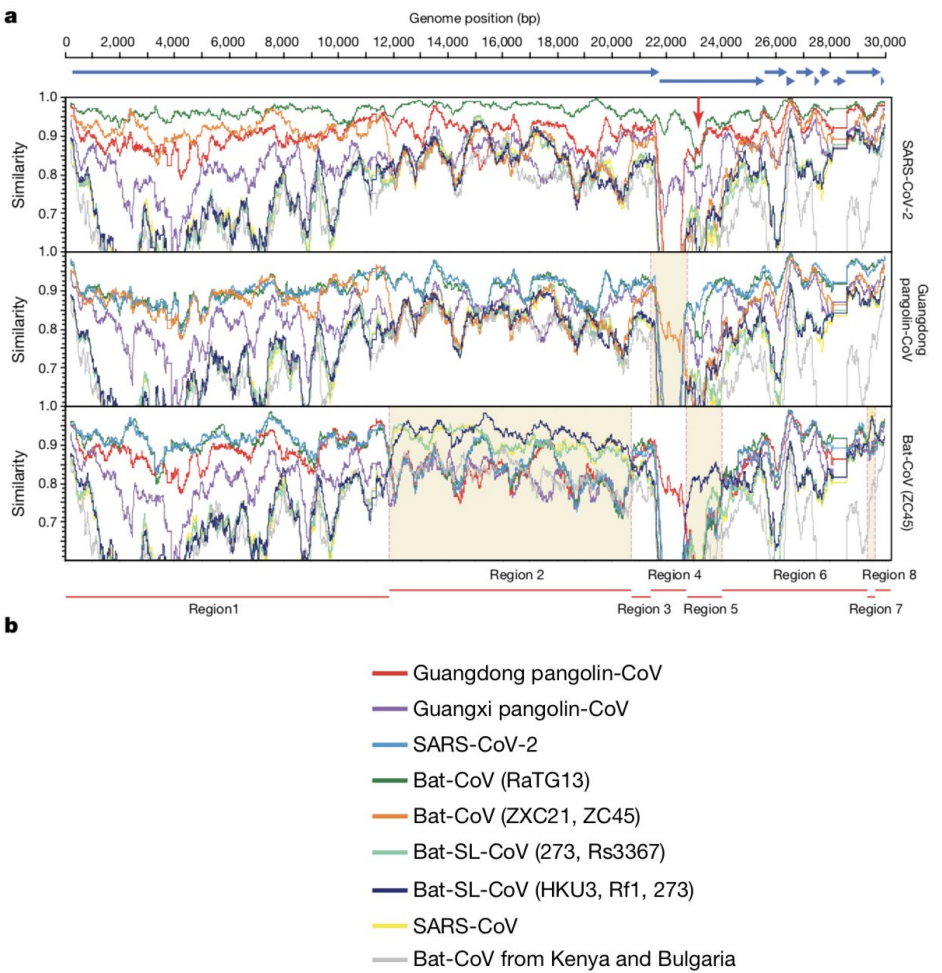



***Manis javanica***
**(Pangolin malais)**

**Aire de répartition de *Manis javanica***

# … et des pangolins ?

- Lam et collègues comparent le génome de SARS-CoV-2 à des génomes de virus isolés à partir de pangolins.
- Ces virus sont globalement plus éloignés de SARS-CoV-2 que ceux de chauve-souris.
- Cependant, on observe une identité plus élevée dans la région particulière où les PPI des autres coronavirus s'affaissent.
- Ceci suggère la possibilité d'une recombinaison entre des virus de chauve-souris et de pangolin.




Lam, T.T.-Y., Shum, M.H.-H., Zhu, H.-C., Tong, Y.-G., Ni, X.-B., Liao, Y.-S., et al. 2020. Identifying SARS-CoV-2 related coronaviruses in Malayan pangolins. Nature, doi: 10.1038/s41586-020-2169-0.

# *Et si les pangolins n'y étaient pour rien ?*

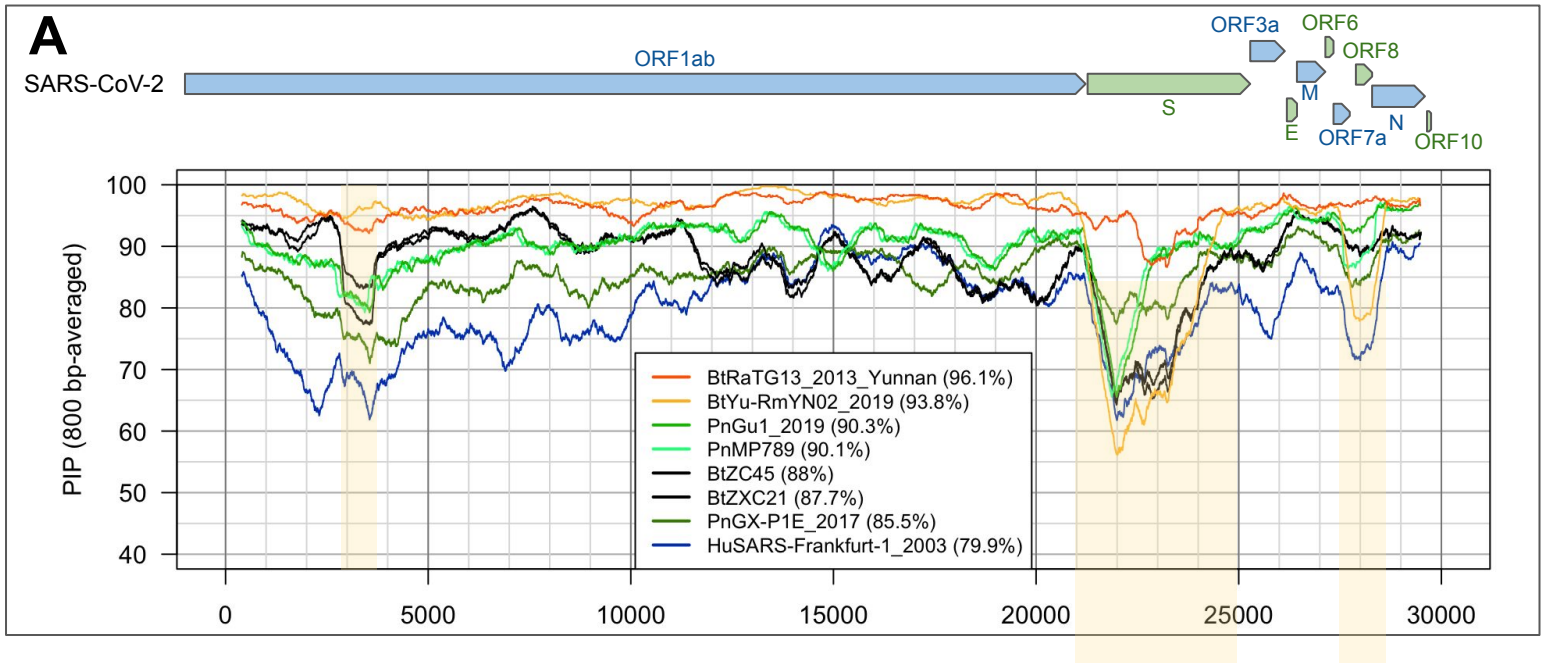
L'hypothèse du pangolin est fortement remise en cause pour plusieurs raisons.

- Les génomes des virus de pangolin les plus proches dont on dispose sont plus éloignés de SARS-CoV-2 que ceux des génomes de chauves-souris.
- La région où la similarité est la plus forte correspondent au gène S, qui code pour la protéine spicule. Or la protéine spicule de ces virus de pangolin n'est pas capable d'adhérer au récepteur de cellules humaines.

*Détecter les recombinaisons par comparaison de génomes*

# *Détection des recombinaisons génomiques*

Les chutes brutales d'identité sur les profils PPI (fond jaune) dénotent des régions résultant vraisemblablement de recombinaisons.

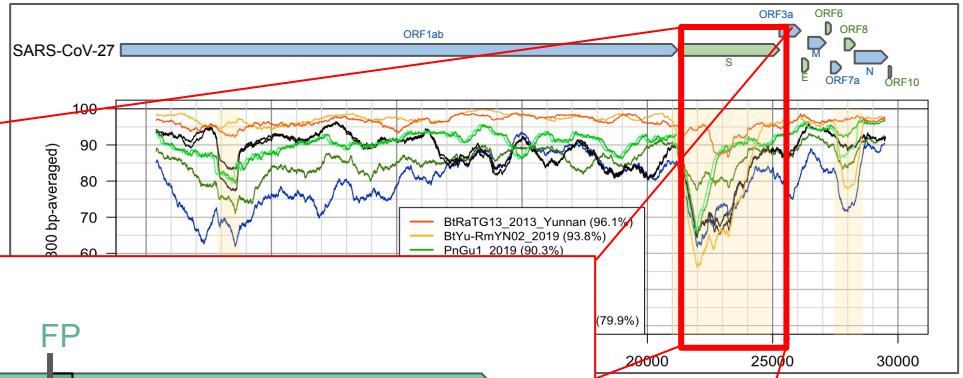


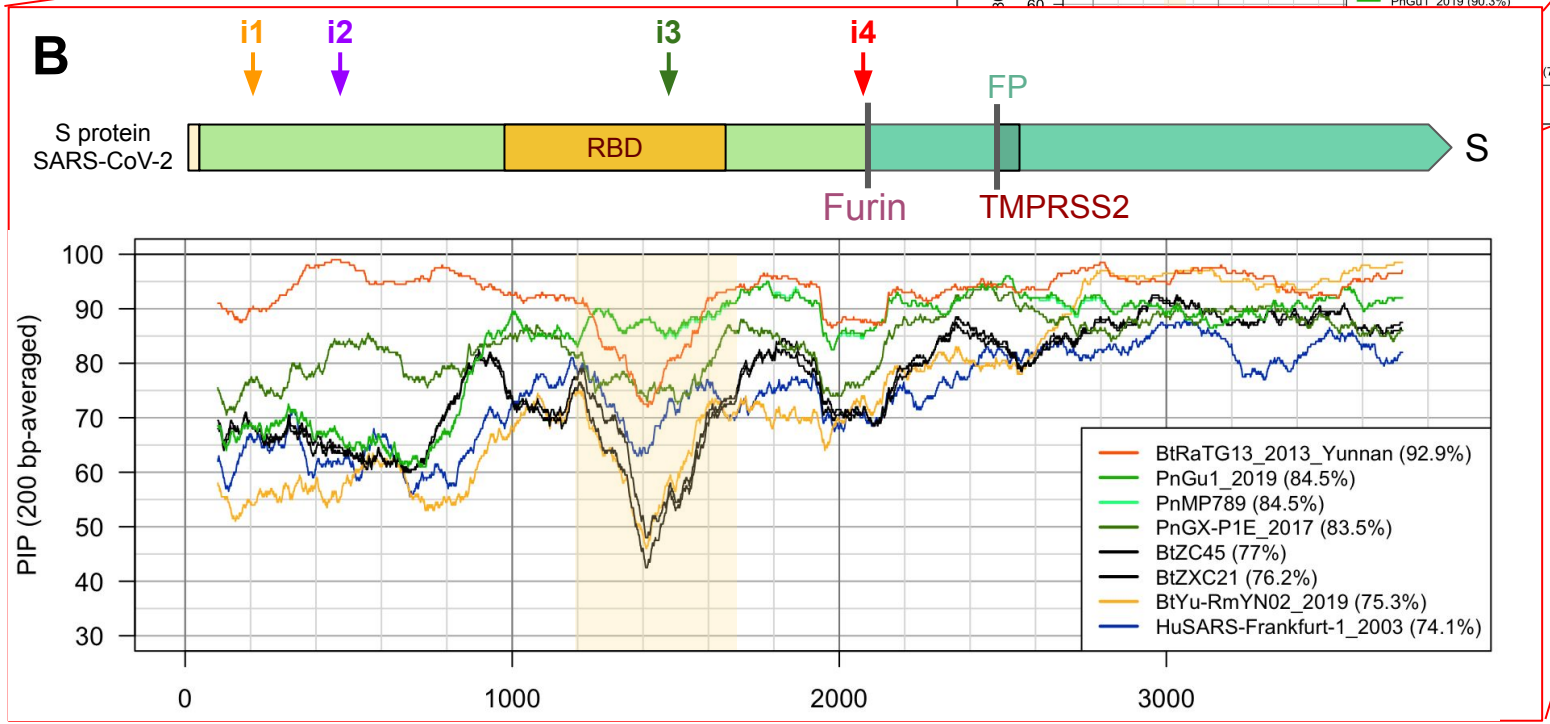


- Sallard, E., Halloy, J., Casane, D., van Helden, J. & Decroly, É. 2020. **Retrouver les origines du SARS-CoV-2 dans les phylogénies de coronavirus**. Med Sci (Paris) 36: 783–796.
- English version : Erwan Sallard, José Halloy, Didier Casane, Etienne Decroly, Jacques van Helden. **Tracing the origins of SARS-CoV-2 in coronavirus phylogenies**. hal-02891455

# Comparaison entre coronavirus - gène S

Profils de pourcentages de positions identiques (PPI) entre régions génomiques du gène S de différents coronavirus et SARS-CoV-2 (la référence à 100%).




Sallard, E., Halloy, J., Casane, D., van Helden, J., Decroly, E. Retrouver les origines du SARS-COV-2 dans les phylogénies de coronavirus. Médecine/Sciences 36, Sept-Août 2020.

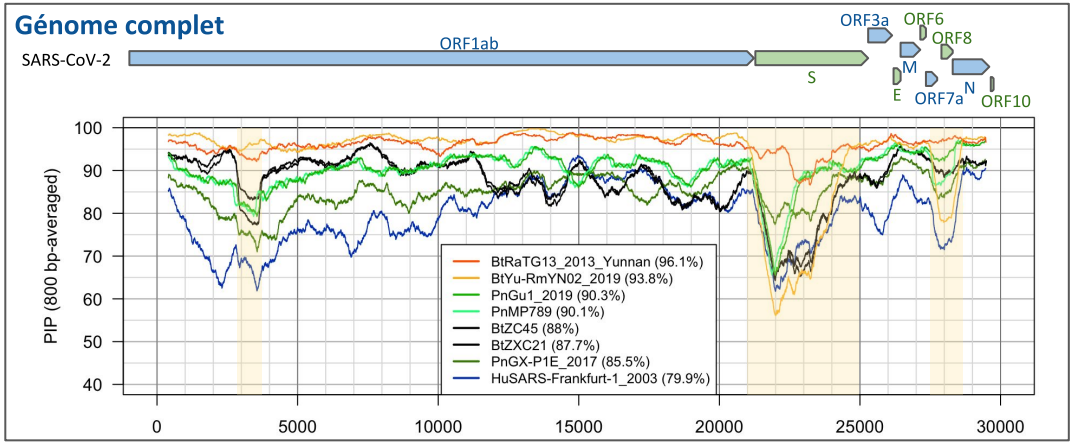
# *Recombinaisons génomiques dans les génomes de coronavirus*

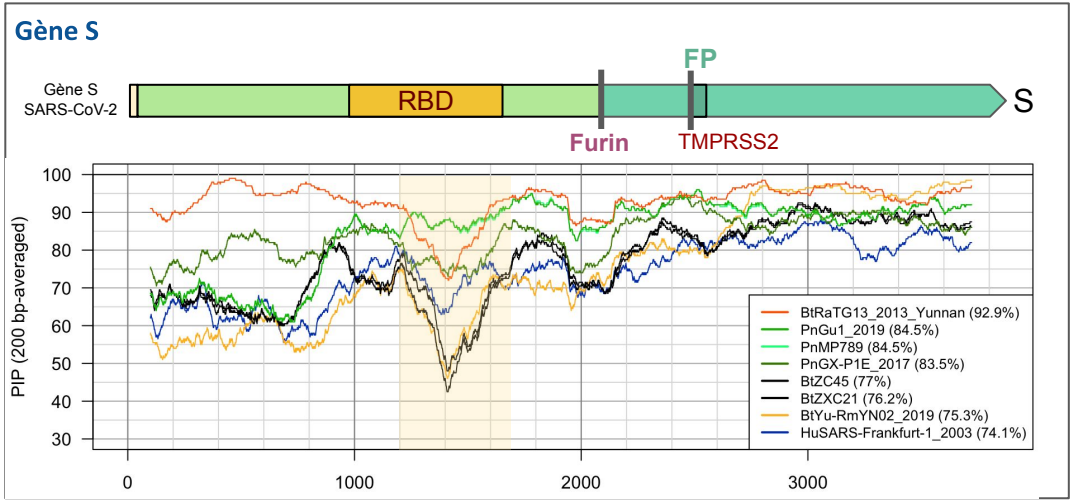
**Le profil génomique**

- régions ayant vraisemblablement fait l'objet de recombinaisons (fond jaune).

**Profil PPI du gène spicule (S)**

- **S**: spicule
- **RBD**: receptor binding domain
- Dans la région du RBD forte baisse des PPI
- Le RBD est en évolution rapide, pourquoi ?
  - Immunogène → forte pression sélective en faveur de variations qui permettent d' échapper à l'immunité
  - Spécificité d'espèce → modifications permettent de changer d'hôte





- Sallard, E., Halloy, J., Casane, D., van Helden, J. & Decroly, É. 2020. **Retrouver les origines du SARS-CoV-2 dans les phylogénies de coronavirus**. Med Sci (Paris) 36: 783–796.

# *Un métagénome de virus de chauve-souris (RmYN02)*

- Hong Zhou et coll. (2019)
- Séquence métagénomique d'un coronavirus de chauve-souris assemblée à partir de 11 échantillons de chauves-souris de Yunnan
- Génome très proche de SARS-CoV-2 ***sauf*** dans la région du gène spicule (S).
  - ce virus serait donc un recombinant de RaTG13
  - ***oui mais*** ce métagénome est reconstruit à partir de 11 échantillons différents. La recombinaison pourrait donc être un artéfact
- Les auteurs soulignent aussi la présence d'une insertion similaire au site furine de SARS-CoV-2.
  - ***Oui mais*** ce scénario est peu vraisemblable, car il aurait nécessité quatre événements évolutifs (une insertion et trois délétions).

Report

Current Biology

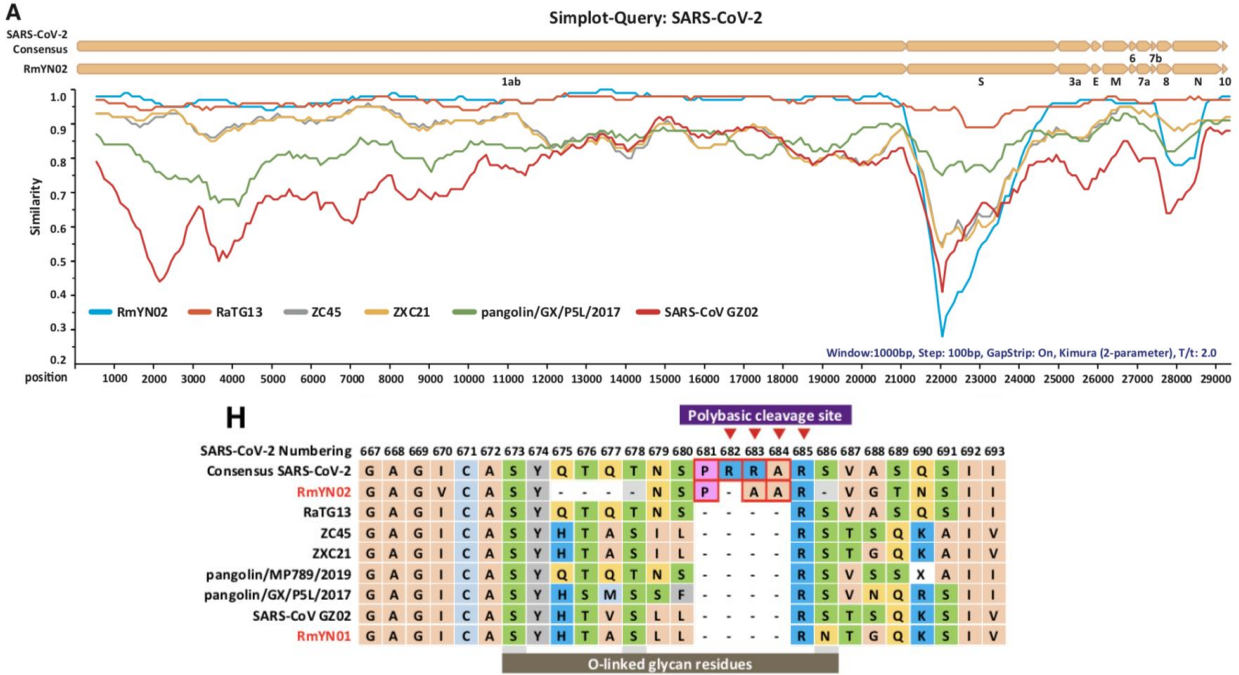
**A Novel Bat Coronavirus Closely Related to SARS-CoV-2 Contains Natural Insertions at the S1/S2 Cleavage Site of the Spike Protein**

Highlights

- Metagenomic analysis identified a novel coronavirus, RmYN02, from *R. malayanus*

Authors

Hong Zhou, Xing Chen, Tao Hu, ...,
Alice C. Hughes, Yuhai Bi, Weifeng Shi




Zhou, H., Chen, X., Hu, T., Li, J., Song, H., Liu, Y., et al. 2020. A Novel Bat Coronavirus Closely Related to SARS-CoV-2 Contains Natural Insertions at the S1/S2 Cleavage Site of the Spike Protein. Current Biology 30: 2196-2203.e3. Elsevier.

# *Bases de données de séquences biologiques*

# *Uniprot / Swiss-prot KB*

- KB: knowledge base
- Swiss-prot contient des séquences protéiques “annotées” par des biologistes. L’annotation consiste à associer à une séquence les connaissances résultant d’expérimentation.
- Deux limitations
  - Le nombre de publications augmente tellement qu’il n’est pas possible à l’équipe de Swiss-prot de tout annoter
  - Le nombre de séquences augmente de façon tellement rapide qu’il est impossible de toutes les caractériser expérimentalement
- TREMBL: annotation automatique des séquences traduites de la base de données EMBL.
- Uniprot = Swiss-prot + TREMBL

# *NCBI Entrez*

Le NCBI (USA) propose une série de bases de données couvrant les différents domaines de la biologie : taxonomie, séquences protéiques, séquences nucléiques, publications biologiques, …

Nous utiliserons le site Web “Entrez” pour consulter les séquences génomiques de coronavirus et leurs annotations.

# *Alignement d'une paire de séquences*

# Exercice

- On dispose des deux séquences suivantes
  - `Seq1 TTTGCGTTAAATCGTGTAGCAATTTAA`
  - `Seq2 AAGAATGGCGTTTTTAATAGCAATAT`
- Questions
  1. En décalant progressivement les séquences, identifiez le(s) décalage(s) qui révèlent des régions de similarité.
  2. A chaque position de décalage, identifiez les segments parfaitement conservés (successions ininterrompue de résidus identiques).
  3. Au vu du résultat, pensez-vous que l'insertion d'un gap permettrait d'augmenter le score d'alignement?

# Solution de l'exercice

- On dispose des deux séquences suivantes
  - `Seq1  TTTGCGTTAAATCGTGTAGCAATTTAA`
  - `Seq2  AAGAATGGCGTTTTTAATAGCAATAT`
- Questions
  1. En décalant progressivement les séquences, identifiez le(s) décalage(s) qui révèlent des régions de similarité.
  2. A chaque position de décalage, identifiez les segments parfaitement conservés (successions ininterrompue de résidus identiques).
  3. Au vu du résultat, pensez-vous que l'insertion d'un gap permettrait d'augmenter le score d'alignement?
- Décalage -4
  - `Position -4   123456789`
  - `Seq1    1234TTTGCGTTAAATCGTGTAGCAATTTAA`
  - `Seq2    AAGAATGGCGTTTTTAATAGCAATAT`
- Décalage -1
  - `Seq1    TTTGCGTTAAATCGTGTAGCAATTTAA`
  - `Seq2  AAGAATGGCGTTTTTAATAGCAATAT`

# Alignement avec « gaps » (brèches)

- Les alignements sans gaps sont rarement pertinents, car les divergences entre séquences incluent souvent des insertions et délétions.

- Les gaps permettent de mettre en évidence les régions de similarités multiples.

```
----TTTGCGTT--AAATCGTGTAGCAATTTAA     s=substitution; |=identité
1111s|s|||||11s||22222|||||||s|22     1=gap dans la 1ère séquence
AAGAATGGCGTTTTTAA-----TAGCAATAT--     2=gap dans la 2de séquence
```

- Gaps, insertions et délétions
  - Les "***gaps***" (***brèches***) reflètent soit une insertion dans l'une des séquences, soit une délétion dans l'autre.
  - Sur seule base de l'alignement d'une paire de séquences, on ne peut pas déterminer si un gap correspond à une délétion ou une insertion.
  - On utilise le terme ***indel*** pour désigner cet événement évolutif de nature indéterminée (insertion ou délétion) qui a donné lieu à un gap observé dans un alignement.

# Alignements globaux (Needleman-Wunsch) versus locaux (Smith-Waterman)

- Alignement global
  - Approprié, par exemple, pour les protéines homologues qui sont conservées sur toute leur longueur.
  - L'alignement final inclut obligatoirement les deux séquences complètes.

```
LQGPSKTGKGS-SRSWDN
|----|-|||---|--|-
LN-ITKAGKGAIMRLGDA
```

  - Algorithme: ***Needleman-Wunsch*** (1970).
  - Outil web EMBOSS : ***needle*** (nucleic acids (nucleic acids or proteins).
- Alignement local
  - Approprié, par exemple, pour les protéines qui partagent un domaine commun, restreint à un segment de chaque séquence.

```
LQGPSSKTGKGS-SSRIWDN
      |-|||
LN-ITKKAGKGAIMRLGDA
```

  - L'alignement final est restreint aux segments conservés.

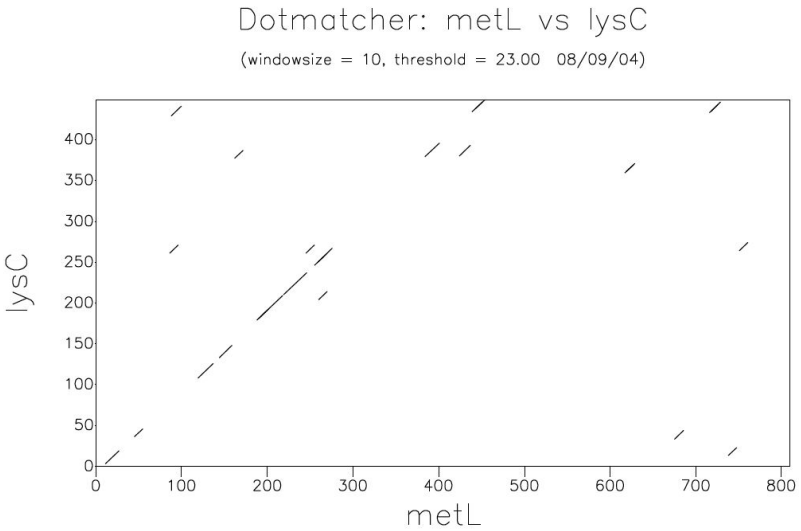
```
KTGKG
|-|||
KAGKG
```

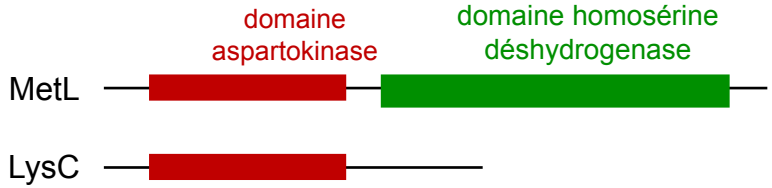
  - Algorithme: ***Smith-Waterman*** (1981).
  - Outil Web EMBOSS : ***water*** (nucleic acids(nucleic acids or proteins)


- Needleman, S. B. & Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 48, 443-53.
- Smith, T. F. & Waterman, M. S. (1981). Identification of common molecular subsequences. J Mol Biol 147, 195-7.

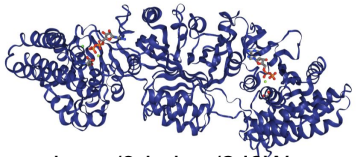
# Aspartokinases: dot plot avec matrice de substitution (BLOSUM62)

- Avec le logiciel *dotmatcher*, une matrice de substitution est utilisée pour assigner un score à chaque paire de résidus. Les segments de lignes indiquent des régions de correspondance entre séquences.
- Ceci révèle la similarité entre les domaines aspartokinase de LysC (l'ensemble de la séquence) et de MetL (positions 1 à ~450).
- La région de similarité ne recouvre que la partie N-terminale (gauche) de MetL, car il s'agit d'une enzyme bi-fonctionnelle. La région C-terminale de MedL contient un domaine homosérine déshydrogenase qui est absent de la protéine LysC.
- Sur base de ce dessin, on comprend qu'un alignement local sera plus pertinent qu'un alignement global car il révèlera le domaine que ces protéines ont en commun.




Structure 3D du domaine aspartokinase



http://www.rcsb.org/3d-view/2J0W

# Alignement global : exemple (Needleman-Wunsch)

```
# Matrix: EBLOSUM62
# Gap_penalty: 10.0
# Extend_penalty: 0.5
# Length: 867
# Identity:     254/867 (29.3%)
# Similarity:   423/867 (48.8%)
# Gaps:         104/867 (12.0%)
# Score: 929.0

metL              1 MSVIAQAGAKGRQLHKFGGSSLADVKCYLRVAGIMAEYSQPDDM-MVVSA     49
                        .::.||||:|:|:.:.:||||.|:...:::..: .|:||
thrA              1           MRVLKFGGTSVANAERFLRVADILESNARQGQVATVLSA     39

metL             50 AGSTTNQLINWLK----------LSQTDRLSAHQVQQTLRRYQCDLISG     88
                    ....||.|:.::            :|..:|:.|            :|::|
thrA             40 PAKITNHLVAMIEKTISGQDALPNISDAERIFA----------ELLTG     77

metL             89 LLPAEEADSL--ISAFV-SDLERLAALLDSGIN------DAVYAEVVGHG    129
                    |..|:....|  :..|| .:.::.:.:| .||:      |::.|.::..|
thrA             78 LAAAQPGFPLAQLKTFVDQEFAQIKHVL-HGISLLGQCPDSINAALICRG    126

metL            130 EVWSARLMSAVLNQQGLPAAWLD-AREFLRAERAAQPQVD--EGLSYPLL    176
                    |..|..:|:.||..:|.....:| ..:.|.....:..||  |.......
thrA            127 EKMSIAIMAGVLEARGHNVTVIDPVEKLLAVGHYLESTVDIAESTRRIAA    176

metL            177 QQLLVQHPGKRLVVTGFISRNNAGETVLLGRNGSDYSATQIGALAGVSRV    226
                    .:....|   .::::.||.:.|..||.|:|||||||||..:.|.......
thrA            177 SRIPADH---MVLMAGFTAGNEKGELVVLGRNGSDYSAAVLAACLRADCC    223

metL            227 TIWSDVAGVYSADPRKVKDACLLPLLRLDEASELARLAAPVLHARTLQPV    276
                    .||:||.|||:.|||:|.||.||..:...||.||:..|.|||.||:.|:
```

- Alignement des protéines metL et thrA d'E.coli avec l'algorithme de Needleman-Wunsch.
- Barres verticales « | »
  - **Identité**: les deux résidus alignés sont identiques.
- Doubles points « : »
  - **Substitution « conservative »**
  - Les deux résidus alignés sont différents mais ***similaires*** (la paire de résidus a un score positif dans la matrice de substitution utilisée (ici, BLOSUM62). Voir plus loin pour comprendre ces matrices.
- Points « . »
  - **Substitution non-conservative**
  - Cette paire de résidus (distincts) a un score négatif dans la matrice de substitution.
- Espace: « »
  - **Gap**: les résidus d'une des deux séquences ne correspondent à aucun résidu sur l'autre.
  - Le gap peut provenir soit d'une délétion, soit d'une insertion, on parle donc d'***indel,*** pour désigner l' événement évolutif d'où provient ce gap.

# Needleman-Wunsch with partial similarities

```
# Matrix: EBLOSUM62
# Gap_penalty: 10.0
# Extend_penalty: 0.5
# Length: 854
# Identity:     136/854 (15.9%)
# Similarity:   209/854 (24.5%)
# Gaps:         449/854 (52.6%)
# Score: 351.0
metL                1 MSVIAQAGAKGRQLHKFGGSSLADVKCYLRVAGIMAEYSQPDDMMVVSAA     50
                      ||.|.        :.|||||:|:||.....|.|.|:...:.. .::|:||:
lysC                1 MSEIV--------VSKFGGTSVADFDAMNRSADIVLSDANV-RLVVLSAS     41

metL               51 GSTTNQLINWLK-LSQTDRLSAHQVQQTLRRYQCDLISGL----LPAEEA     95
                      ...||.|:...: |...:|.   :....:|..|..::..|    :..||.
lysC               42 AGITNLLVALAEGLEPGERF---EKLDAIRNIQFAILERLRYPNVIREEI     88

metL               96 DSLISAFVSDLERLAALLDSGINDAVYAEVVGHGEVWSARLMSAVLNQQG    145
                      :.|:.. ::.|...|||..|   .|:..|:|.|||:.|..|...:|.::.
lysC               89 ERLLEN-ITVLAEAAALATS---PALTDELVSHGELMSTLLFVEILRERD    134

metL              146 LPAAWLDAREFLRA-ERAAQPQVDEGLSYPLLQQLLVQHPGKRLVVT-GF    193
                      :.|.|.|.|:.:|. :|..:.:.|......|....|:....:.||:| ||
lysC              135 VQAQWFDVRKVMRTNDRFGRAEPDIAALAELAALQLLPRLNEGLVITQGF    184

metL              194 ISRNNAGETVLLGRNGSDYSATQIGALAGVSRVTIWSDVAGVYSADPRKV    243
                      |...|.|.|..|||.||||:|..:.....|||.||:||.|:|:|.|||.|
lysC              185 IGSENKGRTTTLGRGGSDYTAALLAEALHASRVDIWTDVPGIYTTDPRVV    234

metL              244 KDACLLPLLRLDEASELARLAAPVLHARTLQPVSGSEIDLQLRCSYTPDQ    293
```

- Alignment of *E.coli* lysC and metL proteins with Needleman-Wunsch algorithm.
- metL contains two domains: aspartokinase and homoserine dehydrogenase.
- LysC only contains the aspartokinase domains.
- With Smith-Waterman, the %similarity is calculated over the whole length of the alignment (854aa), which gives 24.5%.
- Actually, most of the alignment length is in the terminal gap (the homoserine dehydrogenase domain of metL).
- This percentage is lower than the usual threshold for considering two proteins as homolog.

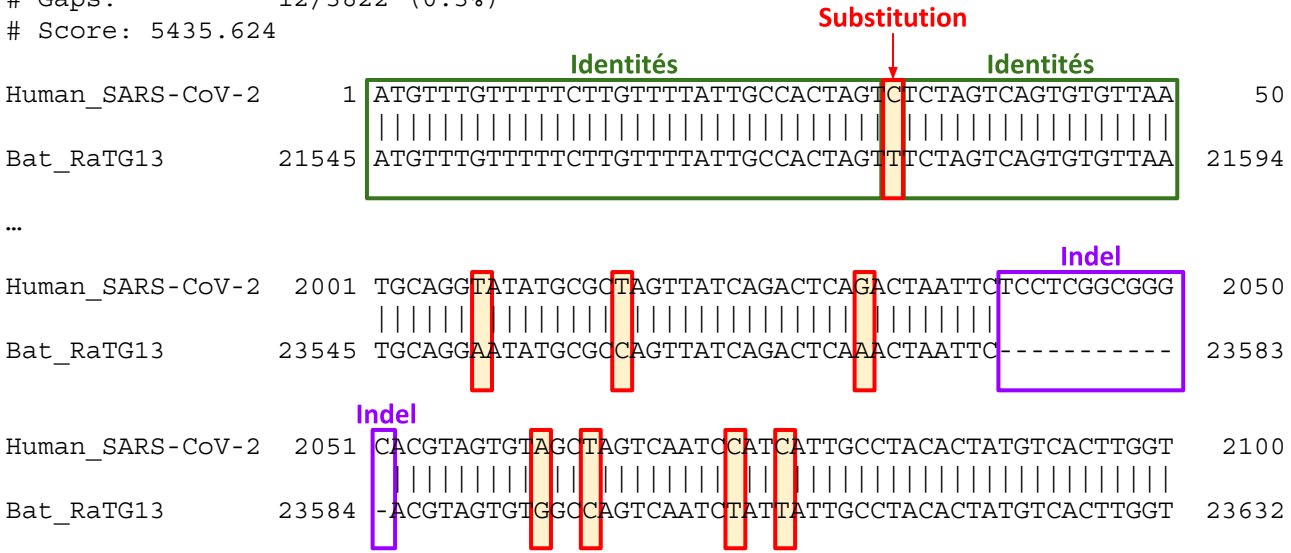
# Smith-Waterman with partial similarities

```
# Matrix: EBLOSUM62
# Gap_penalty: 10.0
# Extend_penalty: 0.5
# Length: 482
# Identity:     133/482 (27.6%)
# Similarity:   205/482 (42.5%)
# Gaps:          85/482 (17.6%)
# Score: 353.5
metL               16 KFGGSSLADVKCYLRVAGIMAEYSQPDDMMVVSAAGSTTNQLINWLK-LS     64
                      ||||:|:||.....|.|.|:...:. .::|:||:...||.|:...: |.
lysC                8 KFGGTSVADFDAMNRSADIVLSDANV-RLVVLSASAGITNLLVALAEGLE     56

metL               65 QTDRLSAHQVQQTLRRYQCDLISGL----LPAEEADSLISAFVSDLERLA    110
                      ..:|.   :....:|..|..:.:..|    :..||.:.|:.. ::.|...|
lysC               57 PGERF---EKLDAIRNIQFAILERLRYPNVIREEIERLLEN-ITVLAEAA    102

metL              111 ALLDSGINDAVYAEVVGHGEVWSARLMSAVLNQQGLPAAWLDAREFLRA-    159
                      ||..|   .|:..|:|.|||:.|..|...:|.:::.:.|.|.|.|:.:|.
lysC              103 ALATS---PALTDELVSHGELMSTLLFVEILRERDVQAQWFDVRKVMRTN    149

metL              160 ERAAQPQVDEGLSYPLLQQLLVQHPGKRLVVT-GFISRNNAGETVLLGRN    208
                      :|..:.:.|......|....|:....:.||:| |||...|.|.|..|||.
lysC              150 DRFGRAEPDIAALAELAALQLLPRLNEGLVITQGFIGSENKGRTTTLGRG    199

metL              209 GSDYSATQIGALAGVSRVTIWSDVAGVYSADPRKVKDACLLPLLRLDEAS    258
                      ||||:|..:.....|||.||:||.|:|:.|||.|..|..:.:...||:
lysC              200 GSDYTAALLAEALHASRVDIWTDVPGIYTTDPRVVSAAKRIDEIAFAEAA    249

metL              259 ELARLAAPVLHARTLQPVSGSEIDLQLRCSYTPDQGSTRI---------E    299
```

- Alignment of *E.coli* lysC and metL proteins with Smith-Waterman algorithm.
- The alignment is almost identical to the one reported by Needleman-Wunsch, but the score is now considered on the aligned segments only (482 aa).
- On this region, there is 42.5% of similarity.

# Alignement de séquences – Gènes S de SARS-CoV-2 et RaTG13

```
# Aligned_sequences: 2
# 1: Human_SARS-CoV-2_BetaCoV/Wuhan/IPBCAMS-WH-01/2019
# 2: Bat_RaTG13
#
# Length: 3822
# Identity:    3549/3822 (92.9%)
# Similarity:    NA/3822 (NA%)
# Gaps:           12/3822 (0.3%)
# Score: 5435.624
```




**Note**

- "**Indel**" signifie "Insertion ou délétion"
- Sur base de ce résultat, la différence observée peut provenir soit d'une insertion chez un ancêtre de SARS-CoV-2, soit d'une délétion chez un ancêtre de RaTG13

# *Recherche de séquences par similarité*

- Les diapos précédentes illustraient des alignements entre une paire de séquences choisies a priori.
- Dans certains cas on dispose de la séquence d'un gène ou d'une protéine d'intérêt, et on désire l'aligner avec ***toutes*** les séquences connues.
- Ceci permet par exemple de détecter des ressemblances entre une séquence inconnue (par exemple une séquence codante qu'on vient d'identifier dans un génome nouvellement séquencé) et des séquences déjà présentes dans une base de données, et dont la fonction est connue.
- La recherche par similarité est aujourd'hui l'approche principale pour assigner des fonctions aux gènes (alignement d'ADN) et aux protéines (alignement de séquences peptidiques).

# Homologie et analogie

- La similarité entre deux traits (organes, séquences) peut s'interpréter par deux hypothèses alternatives: homologie et analogie.
- **Homologie**
  - La similarité s'explique par le fait que les deux séquences divergent d'un ancêtre commun.
  - Les différences entre les deux caractères homologues résultent de l'accumulation de mutations à partir de l'ancêtre commun. Il s'agit donc d'une évolution par ***divergence évolutive***.
- **Analogie**
  - Ressemblance entre deux traits (organes, séquence) qui ne résulte pas d'une origine ancestrale commune (par opposition à l'homologie).
  - Les traits similaires sont apparus de façon ***indépendante***. Leur ressemblance peut éventuellement manifester l'effet d'une pression évolutive qui a sélectionné les mêmes propriétés.
  - Dans ce cas, on parle de ***convergence évolutive***.

# Matrices de substitutions

- Une ***matrice de substitution*** associe un score à chaque paire de résidus qu'on peut trouver dans un alignement.
  - Chaque ligne et chaque colonne représente l'un des résidus (4 nucléotides, 20 acide aminés).
  - La diagonale correspond aux identités.
  - Le triangle inférieur correspond à des substitutions.
  - Le triangle supérieur est symétrique au triangle inférieur, il n'est pas nécessaire d'indiquer les nombres.
- Les ***scores négatifs*** sont considérés comme des pénalités associées à certaines substitutions qu'on n'observe que rarement dans les alignements. Les algorithmes d'alignements tenteront donc d'éviter ces substitutions.
- Les ***scores positifs*** correspondent à des substitutions qu'on observe plus souvent que prévu, dans les alignements d'un grand nombre de séquences. Ceci suggère que ces substitutions particulières sont moins dommageable que d'autres, et on les qualifie donc de « ***substitutions conservatives*** » ou encore de « ***mutations ponctuelles acceptées*** » (***PAM***).
- Au sein d'un alignement, le terme ***similarité*** désigne les positions où se superposent des résidus ayant un score positif dans la matrice de substitution (identité ou substitution conservative).

## Matrice de substitutions entre nucléotides

| | A | C | G | T |
|---|---|---|---|---|
| A | 2 | | | |
| C | -2 | 2 | | |
| G | -2 | -2 | 2 | |
| T | -1 | -2 | -2 | 2 |

## Matrice de substitutions entre acides aminés

| | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Ala | A | 4 | | | | | | | | | | | | | | | | | | |
| Arg | R | -1 | 5 | | | | | | | | | | | | | | | | | |
| Asn | N | -2 | 0 | 6 | | | | | | | | | | | | | | | | |
| Asp | D | -2 | -2 | 1 | 6 | | | | | | | | | | | | | | | |
| Cys | C | 0 | -3 | -3 | -3 | 9 | | | | | | | | | | | | | | |
| Gln | Q | -1 | 1 | 0 | 0 | -3 | 5 | | | | | | | | | | | | | |
| Glu | E | -1 | 0 | 0 | 2 | -4 | 2 | 5 | | | | | | | | | | | | |
| Gly | G | 0 | -2 | 0 | -1 | -3 | -2 | -2 | 6 | | | | | | | | | | | |
| His | H | -2 | 0 | 1 | -1 | -3 | 0 | 0 | -2 | 8 | | | | | | | | | | |
| Ile | I | -1 | -3 | -3 | -3 | -1 | -3 | -3 | -4 | -3 | 4 | | | | | | | | | |
| Leu | L | -1 | -2 | -3 | -4 | -1 | -2 | -3 | -4 | -3 | 2 | 4 | | | | | | | | |
| Lys | K | -1 | 2 | 0 | -1 | -3 | 1 | 1 | -2 | -1 | -3 | -2 | 5 | | | | | | | |
| Met | M | -1 | -1 | -2 | -3 | -1 | 0 | -2 | -3 | -2 | 1 | 2 | -1 | 5 | | | | | | |
| Phe | F | -2 | -3 | -3 | -3 | -2 | -3 | -3 | -3 | -1 | 0 | 0 | -3 | 0 | 6 | | | | | |
| Pro | P | -1 | -2 | -2 | -1 | -3 | -1 | -1 | -2 | -2 | -3 | -3 | -1 | -2 | -4 | 7 | | | | |
| Ser | S | 1 | -1 | 1 | 0 | -1 | 0 | 0 | 0 | -1 | -2 | -2 | 0 | -1 | -2 | -1 | 4 | | | |
| Thr | T | 0 | -1 | 0 | -1 | -1 | -1 | -1 | -2 | -2 | -1 | -1 | -1 | -1 | -2 | -1 | 1 | 5 | | |
| Trp | W | -3 | -3 | -4 | -4 | -2 | -2 | -3 | -2 | -2 | -3 | -2 | -3 | -1 | 1 | -4 | -3 | -2 | 11 | |
| Tyr | Y | -2 | -2 | -2 | -3 | -2 | -1 | -2 | -3 | 2 | -1 | -1 | -2 | -1 | 3 | -3 | -2 | -2 | 2 | 7 |
| Val | V | 0 | -3 | -3 | -3 | -1 | -2 | -2 | -3 | -3 | 3 | 1 | -2 | 1 | -1 | -2 | -2 | 0 | -3 | -1 | 4 |
| | | Ala | Arg | Asn | Asp | Cys | Gln | Glu | Gly | His | Ile | Leu | Lys | Met | Phe | Pro | Ser | Thr | Trp | Tyr | Val |
| | | A | R | N | D | C | Q | E | G | H | I | L | K | M | F | P | S | T | W | Y | V |

# Utilisation d'une matrice de substitution pour calculer le score d'un alignement

$$S = \sum_{i=1}^{L} s_{r_{1,i} r_{2,i}}$$

| | | Ala | Arg | Asn | Asp | Cys | Gln | Glu | Gly | His | Ile | Leu | Lys | Met | Phe | Pro | Ser | Thr | Trp | Tyr | Val |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Ala | A | 4 | | | | | | | | | | | | | | | | | | | |
| Arg | R | -1 | 5 | | | | | | | | | | | | | | | | | | |
| Asn | N | -2 | 0 | 6 | | | | | | | | | | | | | | | | | |
| Asp | D | -2 | -2 | 1 | 6 | | | | | | | | | | | | | | | | |
| Cys | C | 0 | -3 | -3 | -3 | 9 | | | | | | | | | | | | | | | |
| Gln | Q | -1 | 1 | 0 | 0 | -3 | 5 | | | | | | | | | | | | | | |
| Glu | E | -1 | 0 | 0 | 2 | -4 | 2 | 5 | | | | | | | | | | | | | |
| Gly | G | 0 | -2 | 0 | -1 | -3 | -2 | -2 | 6 | | | | | | | | | | | | |
| His | H | -2 | 0 | 1 | -1 | -3 | 0 | 0 | -2 | 8 | | | | | | | | | | | |
| Ile | I | -1 | -3 | -3 | -3 | -1 | -3 | -3 | -4 | -3 | 4 | | | | | | | | | | |
| Leu | L | -1 | -2 | -3 | -4 | -1 | -2 | -3 | -4 | -3 | 2 | 4 | | | | | | | | | |
| Lys | K | -1 | 2 | 0 | -1 | -3 | 1 | 1 | -2 | -1 | -3 | -2 | 5 | | | | | | | | |
| Met | M | -1 | -1 | -2 | -3 | -1 | 0 | -2 | -3 | -2 | 1 | 2 | -1 | 5 | | | | | | | |
| Phe | F | -2 | -3 | -3 | -3 | -2 | -3 | -3 | -3 | -1 | 0 | 0 | -3 | 0 | 6 | | | | | | |
| Pro | P | -1 | -2 | -2 | -1 | -3 | -1 | -1 | -2 | -2 | -3 | -3 | -1 | -2 | -4 | 7 | | | | | |
| Ser | S | 1 | -1 | 1 | 0 | -1 | 0 | 0 | 0 | -1 | -2 | -2 | 0 | -1 | -2 | -1 | 4 | | | | |
| Thr | T | 0 | -1 | 0 | -1 | -1 | -1 | -1 | -2 | -2 | -1 | -1 | -1 | -1 | -2 | -1 | 1 | 5 | | | |
| Trp | W | -3 | -3 | -4 | -4 | -2 | -2 | -3 | -2 | -2 | -3 | -2 | -3 | -1 | 1 | -4 | -3 | -2 | 11 | | |
| Tyr | Y | -2 | -2 | -2 | -3 | -2 | -1 | -2 | -3 | 2 | -1 | -1 | -2 | -1 | 3 | -3 | -2 | -2 | 2 | 7 | |
| Val | V | 0 | -3 | -3 | -3 | -1 | -2 | -2 | -3 | -3 | 3 | 1 | -2 | 1 | -1 | -2 | -2 | 0 | -3 | -1 | 4 |
| | | Ala | Arg | Asn | Asp | Cys | Gln | Glu | Gly | His | Ile | Leu | Lys | Met | Phe | Pro | Ser | Thr | Trp | Tyr | Val |
| | | A | R | N | D | C | Q | E | G | H | I | L | K | M | F | P | S | T | W | Y | V |

- Les matrices de substitution sont utilisées pour calculer le score d'un alignement.
- Ce score est la somme, pour toutes les positions de l'alignement ($i$ *de 1 à* $L$) , des scores des paires de résidus ($r_{1,I}$ et $r_{2,I}$).
- Les "gaps" sont traités par une règle spécifique reposant sur deux paramètres de pénalité:
  - Pénalité d'ouverture de gap (***go***)
    - Valeurs typiques: entre -10 et -15
  - Pénalité d'extension de gap (***ge***)
    - Valeurs typiques: entre -0.5 et -2

| *i* | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | R | L | A | S | V | E | T | D | M | P | - | - | - | - | - | L | T | L | R | Q | H |
| | T | L | T | S | L | Q | T | T | L | K | N | L | K | E | M | A | H | L | G | T | H |

S

# Utilisation d'une matrice de substitution pour calculer le score d'un alignement

$$S = \sum_{i=1}^{L} s_{r_{1,i} r_{2,i}}$$

| | | Ala | Arg | Asn | Asp | Cys | Gln | Glu | Gly | His | Ile | Leu | Lys | Met | Phe | Pro | Ser | Thr | Trp | Tyr | Val |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | A | R | N | D | C | Q | E | G | H | I | L | K | M | F | P | S | T | W | Y | V |
| Ala | A | 4 | | | | | | | | | | | | | | | | | | | |
| Arg | R | -1 | 5 | | | | | | | | | | | | | | | | | | |
| Asn | N | -2 | 0 | 6 | | | | | | | | | | | | | | | | | |
| Asp | D | -2 | -2 | 1 | 6 | | | | | | | | | | | | | | | | |
| Cys | C | 0 | -3 | -3 | -3 | 9 | | | | | | | | | | | | | | | |
| Gln | Q | -1 | 1 | 0 | 0 | -3 | 5 | | | | | | | | | | | | | | |
| Glu | E | -1 | 0 | 0 | 2 | -4 | 2 | 5 | | | | | | | | | | | | | |
| Gly | G | 0 | -2 | 0 | -1 | -3 | -2 | -2 | 6 | | | | | | | | | | | | |
| His | H | -2 | 0 | 1 | -1 | -3 | 0 | 0 | -2 | 8 | | | | | | | | | | | |
| Ile | I | -1 | -3 | -3 | -3 | -1 | -3 | -3 | -4 | -3 | 4 | | | | | | | | | | |
| Leu | L | -1 | -2 | -3 | -4 | -1 | -2 | -3 | -4 | -3 | 2 | 4 | | | | | | | | | |
| Lys | K | -1 | 2 | 0 | -1 | -3 | 1 | 1 | -2 | -1 | -3 | -2 | 5 | | | | | | | | |
| Met | M | -1 | -1 | -2 | -3 | -1 | 0 | -2 | -3 | -2 | 1 | 2 | -1 | 5 | | | | | | | |
| Phe | F | -2 | -3 | -3 | -3 | -2 | -3 | -3 | -3 | -1 | 0 | 0 | -3 | 0 | 6 | | | | | | |
| Pro | P | -1 | -2 | -2 | -1 | -3 | -1 | -1 | -2 | -2 | -3 | -3 | -1 | -2 | -4 | 7 | | | | | |
| Ser | S | 1 | -1 | 1 | 0 | -1 | 0 | 0 | 0 | -1 | -2 | -2 | 0 | -1 | -2 | -1 | 4 | | | | |
| Thr | T | 0 | -1 | 0 | -1 | -1 | -1 | -1 | -2 | -2 | -1 | -1 | -1 | -1 | -2 | -1 | 1 | 5 | | | |
| Trp | W | -3 | -3 | -4 | -4 | -2 | -2 | -3 | -2 | -2 | -3 | -2 | -3 | -1 | 1 | -4 | -3 | -2 | 11 | | |
| Tyr | Y | -2 | -2 | -2 | -3 | -2 | -1 | -2 | -3 | 2 | -1 | -1 | -2 | -1 | 3 | -3 | -2 | -2 | 2 | 7 | |
| Val | V | 0 | -3 | -3 | -3 | -1 | -2 | -2 | -3 | -3 | 3 | 1 | -2 | 1 | -1 | -2 | -2 | 0 | -3 | -1 | 4 |

- Les matrices de substitution sont utilisées pour calculer le score d'un alignement.
- Ce score est la somme, pour toutes les positions de l'alignement ($i$ *de 1 à* $L$) , des scores des paires de résidus ($r_{1,I}$ et $r_{2,I}$).
- Les "gaps" sont traités par une règle spécifique reposant sur deux paramètres de pénalité:
  - Pénalité d'ouverture de gap (***go***)
    - Valeurs typiques: entre -10 et -15
  - Pénalité d'extension de gap (***ge***)
    - Valeurs typiques: entre -0.5 et -2

| *i* | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | R | L | A | S | V | E | T | D | M | P | - | - | - | - | - | L | T | L | R | Q | H | |
| | . | \| | . | \| | : | : | \| | . | : | . | go | ge | ge | ge | ge | . | . | \| | . | . | \| | |
| | T | L | T | S | L | Q | T | T | L | K | N | L | K | E | M | A | H | L | G | T | H | |
| S | -1 | +4 | +0 | +4 | +1 | +2 | +5 | -1 | +2 | -1 | -10 | -1 | -1 | -1 | -1 | -1 | -2 | +4 | -2 | -1 | +8 | = 7 |

# Résultat de BLAST – Requête peptidique vs DB de peptides

```
>gi|16127996|ref|NP_414543.1| bifunctional: aspartokinase I
          (N-terminal); homoserine dehydrogenase I (C-terminal)
          [Escherichia coli K12]
         Length = 820

 Score =  344 bits (882), Expect = 2e-95
Identities = 247/821 (30%)  Positives = 410/821 (49%)  Gaps = 44/821 (5%)

Query: 16   KFGGSSLADVKCYLRVAGIMAEYSQPDDMM-VVSAAGSTTNQLINWLKLSQTDRLSAHQV 74
            KFGG+S+A+ + +LRVA I+   ++    +  V+SA    TN L+  ++ + + + +   +
Sbjct: 5    KFGGTSVANAERFLRVADILESNARQGQVATVLSAPAKITNHLVAMIEKTISGQDALPNI 64

Query: 75   QQTLRRYQCDLISGLLPAEEADSL--ISAFVSDLERLAALLDSGIN------DAVYAEVV 126
                R +  +L++GL   A+     L  +  FV          +  GI+      D++ A ++
Sbjct: 65   SDAERIF-AELLTGLAAAQPGFPLAQLKTFVDQEFAQIKHVLHGISLLGQCPDSINAALI 123

Query: 127  GHGEVWSARLMSAVLNQQGLPAAWLDAREFLRAER---AAQPQVDEGLSYPLLQQLLVQH 183
              GE   S  +M+ VL  +G     +D  E L A       +   + E         ++   H
Sbjct: 124  CRGEKMSIAIMAGVLEARGHNVTVIDPVEKLLAVGHYLESTVDIAESTRRIAASRIPADH 183

Query: 184  PGKRLVVTGFISRNNAGETVLLGRNGSDYSATQIGALAGVSRVTIWSDVAGVYSADPRKV 243
               +++ GF + N  GE V+LGRNGSDYSA  + A         IW+DV GVY+ DPR+V
Sbjct: 184  ---MVLMAGFTAGNEKGELVVLGRNGSDYSAAVLAACLRADCCEIWTDVDGVYTCDPRQV 240

Query: 244  KDACLLPLLRLDEASELARLAAPVLHARTLQPVSGSEIDLQLRCSYTPDQ-----GSTRI 298
             DA LL  +   EA EL+   A VLH RT+ P++  +I   ++ +  P       G++R
Sbjct: 241  PDARLLKSMSYQEAMELSYFGAKVLHPRTITPIAQFQIPCLIKNTGNPQAPGTLIGASRD 300

Query: 299  ERVLASGTGARIVTSHDDVCLIEFQVPASQDFKLAHKEIDQILKRAQVRPLAVGVHNDRQ 358
            E  L      + +++ +++ +     P  +         +  + RA++  + +   +
Sbjct: 301  EDELP----VKGISNLNNMAMFSVSGPGMKGMVGMAARVFAAMSRARISVVLITQSSSEY 356
```

- La ligne entre les séquences “Query” et “Sbjct” indique les correspondances entre acides aminés.
- Identités
- Substitutions “conservatives”: paires de résidus distincts mais dont la substitution est *généralement* moins délétère que pour d’autres paires de résidus.
- Substitutions non conservatives
- Positives: identités + substitutions conservatives.
- Gaps: lacunes insérées dans une séquence afin d’optimiser l’alignement des fragments avoisinants.

# Résultat de BLAST – Requête peptidique vs DB de peptides

```
>gi|16127996|ref|NP_414543.1| bifunctional: aspartokinase I
          (N-terminal); homoserine dehydrogenase I (C-terminal)
          [Escherichia coli K12]
         Length = 820

 Score =  344 bits (882), Expect = 2e-95
 Identities = 247/821 (30%), Positives = 410/821 (49%), Gaps = 44/821 (5%)

Query: 16  KFGGSSLADVKCYLRVAGIMAEYSQPDDMM-VVSAAGSTTNQLINWLKLSQTDRLSAHQV 74
           KFGG+S+A+ + +LRVA I+    ++    +  V+SA    TN L+  ++ + + + +   +
Sbjct: 5   KFGGTSVANAERFLRVADILESNARQGQVATVLSAPAKITNHLVAMIEKTISGQDALPNI 64

Query: 75  QQTLRRYQCDLISGLLPAEEADSL--ISAFVSDLERLAALLDSGIN------DAVYAEVV 126
               R +  +L++GL  A+    L  +  FV          +  GI+      D++ A ++
Sbjct: 65  SDAERIF-AELLTGLAAAQPGFPLAQLKTFVDQEFAQIKHVLHGISLLGQCPDSINAALI 123

Query: 127 GHGEVWSARLMSAVLNQQGLPAAWLDAREFLRAER---AAQPQVDEGLSYPLLQQLLVQH 183
             GE  S  +M+ VL  +G     +D  E L A      +   + E         ++   H
Sbjct: 124 CRGEKMSIAIMAGVLEARGHNVTVIDPVEKLLAVGHYLESTVDIAESTRRIAASRIPADH 183

Query: 184 PGKRLVVTGFISRNNAGETVLLGRNGSDYSATQIGALAGVSRVTIWSDVAGVYSADPRKV 243
               +++ GF + N  GE V+LGRNGSDYSA  + A         IW+DV GVY+ DPR+V
Sbjct: 184 ---MVLMAGFTAGNEKGELVVLGRNGSDYSAAVLAACLRADCCEIWTDVDGVYTCDPRQV 240

Query: 244 KDACLLPLLRLDEASELARLAAPVLHARTLQPVSGSEIDLQLRCSYTPDQ-----GSTRI 298
            DA LL  +   EA EL+   A VLH RT+ P++  +I   ++ +  P       G++R
Sbjct: 241 PDARLLKSMSYQEAMELSYFGAKVLHPRTITPIAQFQIPCLIKNTGNPQAPGTLIGASRD 300

Query: 299 ERVLASGTGARIVTSHDDVCLIEFQVPASQDFKLAHKEIDQILKRAQVRPLAVGVHNDRQ 358
           E  L      + +++ +++ +     P  +         +   + RA++  + +  +
Sbjct: 301 EDELP----VKGISNLNNMAMFSVSGPGMKGMVGMAARVFAAMSRARISVVLITQSSSEY 356
```

- Exemple de résultat de recherche par similarité de séquences.
  - Requête (query): metA
  - Protéine identifiée dans la base de données: (subject): thrA.
- Le premier critère d'évaluation d'un résultat de BLAST:
  - La e-valeur (**expect**) indique le nombre de faux-positifs attendus au hasard, si l'on plaçait le seuil au niveau du score observé (344 bits dans ce cas-ci).
  - Plus la e-valeur est faible, plus le résultat est fiable.
  - Si la e-valeur est >= 1, le résultat est peu fiable (on aurait facilement obtenu un « aussi bon » alignement avec des séquences aléatoires.

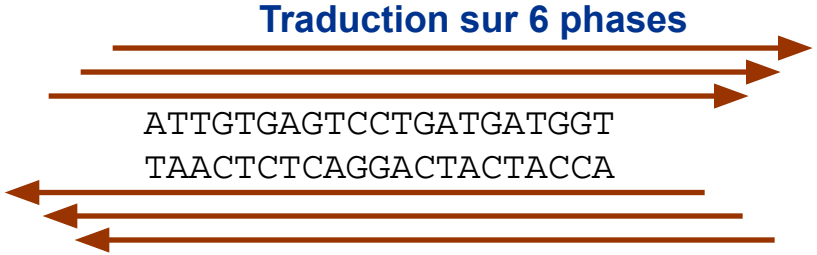
# *Modalités de BLAST*

# DNA versus protein searches

- When the query is a coding DNA sequence, it is recommended to apply searches with the translated rather than raw DNA sequences
  - This allows to introduce a substitution matrix (PAM, BLOSUM, ...), which better reflects the evolutionary changes.
  - It has been shown that some distant relationships can be detected with translated searches, but escape detection with the DNA search.
  - It is easier to filter out low complexity regions from proteins than from DNA sequences.

# Traduction d'une séquence nucléique dans les 6 phases

- Si l'on dispose d'une séquence nucléique, on peut facilement déduire la séquence de la protéine qui pourrait être produite par sa traduction, sur chacun des 6 brins.
- Si cette séquence n'est pas codante, on s'attend à trouver des codons stop assez fréquemment (3 codons sur 64).
- Cependant, rien n'empêche d'aligner les 6 séquences ainsi produites avec d'autres séquences peptidiques.



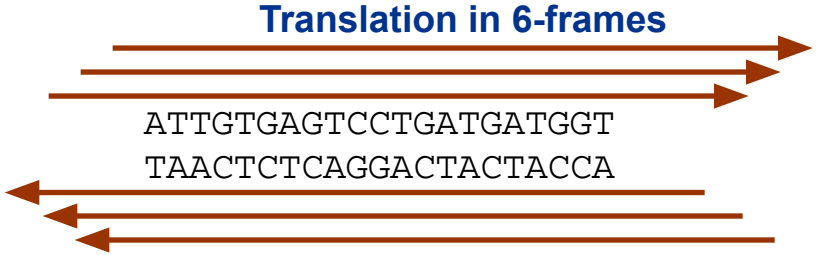

**Résultat**

```
F1       I   V   S   P   D   D   G
F2        L   *   V   L   M   M   V
F3         C   E   S   *   *   W   X
      1 ATTGTGAGTCCTGATGATGGT 21
        ----:----|----:----|-
      1 TAACACTCAGGACTACTACCA 21
F6        X   T   L   G   S   S   P
F5       X   Q   S   D   Q   H   H
F4         N   H   T   R   I   I   T
```

- Outil: http://www.ebi.ac.uk/Tools/st/emboss_sixpack/

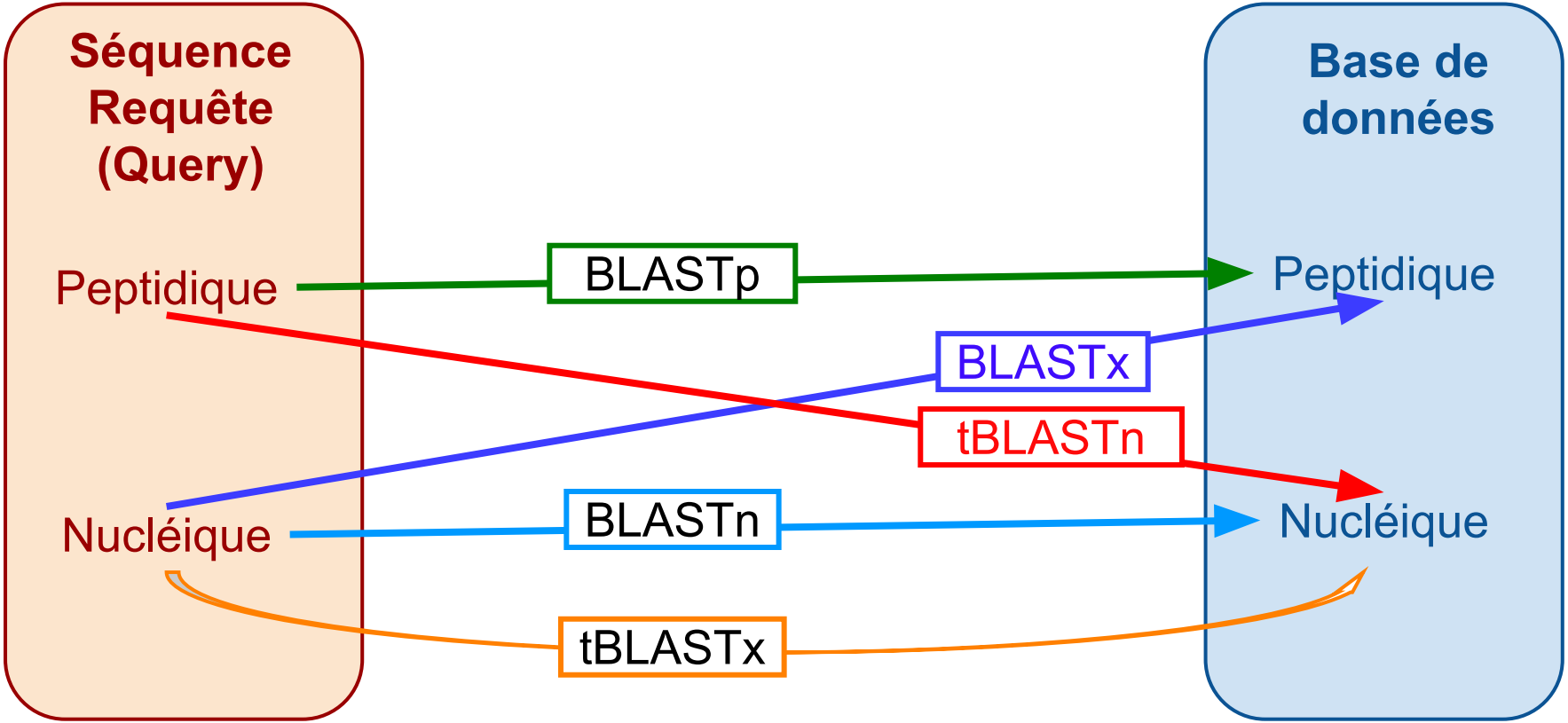
# BLAST - a family of purpose-specific programs

- Different program names exist, depending on the type (protein or nucleic acid) of query and database sequences.
- For comparison between nucleic acids and proteins, the nucleic acid is translated in the 6 frames (3 frames per strand)




| Query | Database | Program | Application examples | Study cases |
|---|---|---|---|---|
| protein | protein | blastp | Starting from a protein of known function detect putative homologs in the whole Uniprot database. | Collect sequences similar to the blue-sensitive opsin in all human proteins. |
| nucleic acid | nucleic acid | blastn | Match RNAi against a genome.<br>Match mRNA (or EST) against a genome. | |
| nucleic acid (translated) | protein | blastx | After having sequenced a piece of DNA, search potentially coding fragments + their putative homologs without any prior knowledge of gene positions in the query sequence. | |
| protein | nucleic acid (translated) | tblastn | - Identify a genomic region likely to code for an homolog of a protein of interest.<br>- Identify pseudo-genes (defective genes, with many stop codons) for a protein of interest in a genome. | Do cats see colors ? Get Human blue-sensitive opsin protein, connect UCSC genome browser, use BLAT to find similarities in Cat genome |
| nucleic acid (translated) | nucleic acid (translated) | tblastx | | |

# Les modalités de BLAST




Source: Emese Meglecz

*Comment interpréter la similarité entre séquences de HIV et de SARS-CoV-2 ?*

# *Alignement de séquences de SARS-CoV-2 sur le génome du HIV*

Haut: fragment le plus significatif de l'alignement de la séquence du gène S sur le génome du VIH. Noter le score Expect = 7.5. Ce score n'est significatif que s'il est nettement inférieur à 1.

Bas: fragment le plus significatif de l'alignement d'une séquence aléatoire sur le génome du VIH. Noter le score Expect = 2.1, supérieur à 1 et donc non-significatif (comme on s'y attend, puisque la séquence est aléatoire).

Conclusion: l'alignement sur lequel s'appuient Perez et Luc Montagnier correspond à ce qu'on s'attend à trouver par hasard en alignant des séquences de cette taille.

**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
```

**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
```

Sallard, E., Halloy, J., Casane, D., van Helden, J., Decroly, E. Retrouver les origines du SARS-COV-2 dans les phylogénies de coronavirus. Médecine/Sciences 36, Sept-Août 2020.

*Travaux pratiques*

# *Travaux pratiques*

Nous mobiliserons une série d'outils bioinformatiques accessibles en ligne pour analyser les séquences de coronavirus et pour tenter de trouver des éléments informatifs concernant l'origine de SARS-CoV-2.

| | | |
|---|---|---|
| Supports de ce cours | Diapos, tuto, données | https://jvanheld.github.io/shnc-origines-sars-cov-2/ |
| Uniprot | Base de donnée de séquences protéiques | https://www.uniprot.org/ |
| NCBI Entrez | Bases de données biologiques | https://www.ncbi.nlm.nih.gov/ |
| EMBOSS needle | Alignement de paires de séquences | https://www.ebi.ac.uk/Tools/psa/emboss_needle/ |
| NCBI BLAST | Recherche de séquences par similarité | https://blast.ncbi.nlm.nih.gov/Blast.cgi |
| PIPprofileR | Profils de pourcentages de positions identiques | https://pipprofiler.france-bioinformatique.fr/ |
| Clustal | Alignement de séquences multiples | https://www.ebi.ac.uk/Tools/msa/clustalo/ |
| phylogeny.fr | Phylogénie moléculaire | https://www.phylogeny.fr/ |
| AMU | page AMETICE de N&C3 | https://ametice.univ-amu.fr/course/view.php?id=62928 |