Module d'enseignement "Nature et culture 3
Licence Sciences et Humanités
Aix-Marseille Université
Novembre 2021

Alignements de séquences multiples

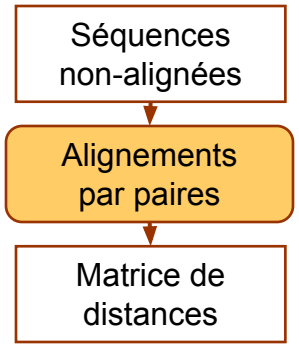
Jacques van Helden

Alignement progressif

- L'approche la plus courante pour aligner des séquences multiples est de réaliser un ***alignement progressif***.
- L'algorithme procède en plusieurs étapes (détaillées dans les diapos suivantes):
  - Calculer une ***matrice de distances***, qui indique la distance entre chaque paire de séquences.
  - Construire un ***arbre guide*** qui regroupe en premier lieu les séquences les plus proches, et remonte en regroupant progressivement les séquences les plus éloignées.
  - Utiliser ce arbre pour aligner progressivement les séquences.
- Il s'agit d'une approche ***heuristique***
  - Cette approche est praticable pour un grand nombre de séquences, mais ne peut pas garantir de retourner l'alignement optimal.

Alignement progressif – 1ère étape: construction de la matrice de distance

- On effectue un alignement par paires entre chaque paire de protéines
  - Alignement par programmation dynamique ou par BLAST.
  - Nombre d'alignements = $n * (n - 1) / 2$
- A partir de chaque alignement par paire, calculer la distance entre les deux séquences.
  - $d_{i,j}=s_{i,j}/L_{j,j}$
    - $d_{j,j}$ distance entre les séquences *i* and *j*
    - $L_{j,j}$ longueur de l'alignement
    - $s_{j,j}$ nombre de substitutions
- Remarques
  - Les gaps ne sont pas pris en compte dans la métrique de distance
  - La matrice est symétrique: $d_{i,j} = d_{j,i}$
  - Les éléments diagonaux sont nuls: $d_{i,i} = 0$



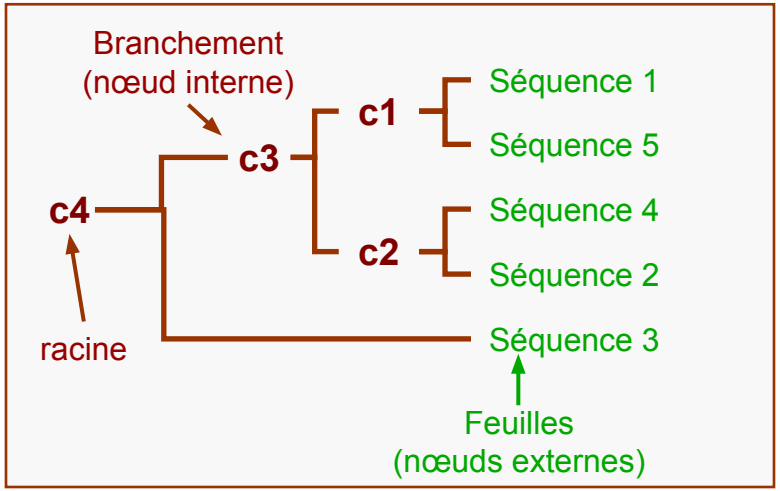
	seq 1	seq 2	...	seq n
seq 1	d1,1	d1,2	...	d1,n
seq 2	d2,1	d2,2	...	d2,n
...	...	...	...	...
seq n	dn,1	dn,2	...	dn,n

Principe de la construction de l'arbre-guide – Méthode UPGMA

Matrice de distance

	séquence 1	séquence 2	séquence 3	séquence 4	séquence 5
séquence 1	0.00	4.00	6.00	3.50	1.00
séquence 2	4.00	0.00	6.00	2.00	4.50
séquence 3	6.00	6.00	0.00	5.50	6.50
séquence 4	3.50	2.00	5.50	0.00	4.00
séquence 5	1.00	4.50	6.50	4.00	0.00

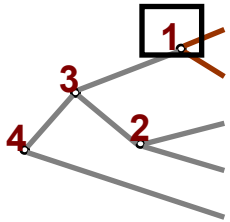
Arbre



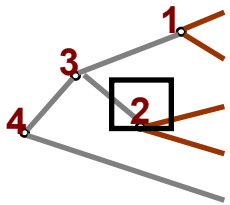
- Le clustering hiérarchique est une méthode de clustering agrégative.
  - Prend une matrice de distance en entrée
  - Regroupe progressivement les objets en allant des plus proches aux plus distants.
- Il existe plusieurs possibilités pour établir une règle d'agglomération, qui définit la distance entre deux groupes.
  - Liaison simple (***single linkage***): distance entre groupes A et B est la distance entre les plus proches de leurs éléments respectifs.
  - Liaison moyenne (***average linkage***): distance moyenne entre tous les objets des deux groupes (=**UPGMA**).
  - Liaison complète (***complete linkage***): distance entre les éléments les plus éloignés des groupes A et B.
- Algorithme
  - 1. Assigner chaque objet à un cluster séparé.
  - 2. Identifier la paire de clusters les plus proches, et les regrouper en un seul.
  - 3. Répéter la seconde étape jusqu'à ce qu'il ne reste qu'un seul cluster.
- Le résultat est un arbre, dont les nœuds intermédiaires correspondent aux clusters.
  - N objets → N-1 nœuds intermédiaires
- Les longueurs des branches représentent les distances entre clusters.

Alignement progressif – 3ème étape: alignement multiple

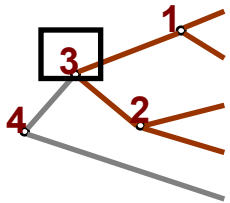
- On construit un alignement multiple en incorporant progressivement les séquences selon leur ordre de branchement dans l'arbre guide, en remontant des plus proches aux plus éloignées.



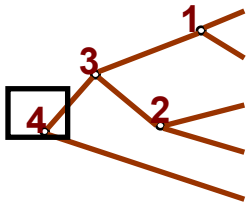
Seq5 **GATTGTAGTA**
Seq1 **GATGGTAGTA**
Seq4
Seq2
Seq3



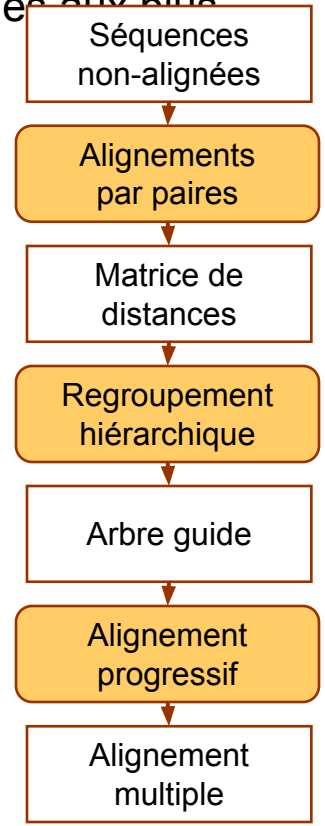
Seq5 GATTGTAGTA
Seq1 GATGGTAGTA
Seq4 **GATTGTTC--GTA**
Seq2 **GATTGTTCGGGTA**
Seq3



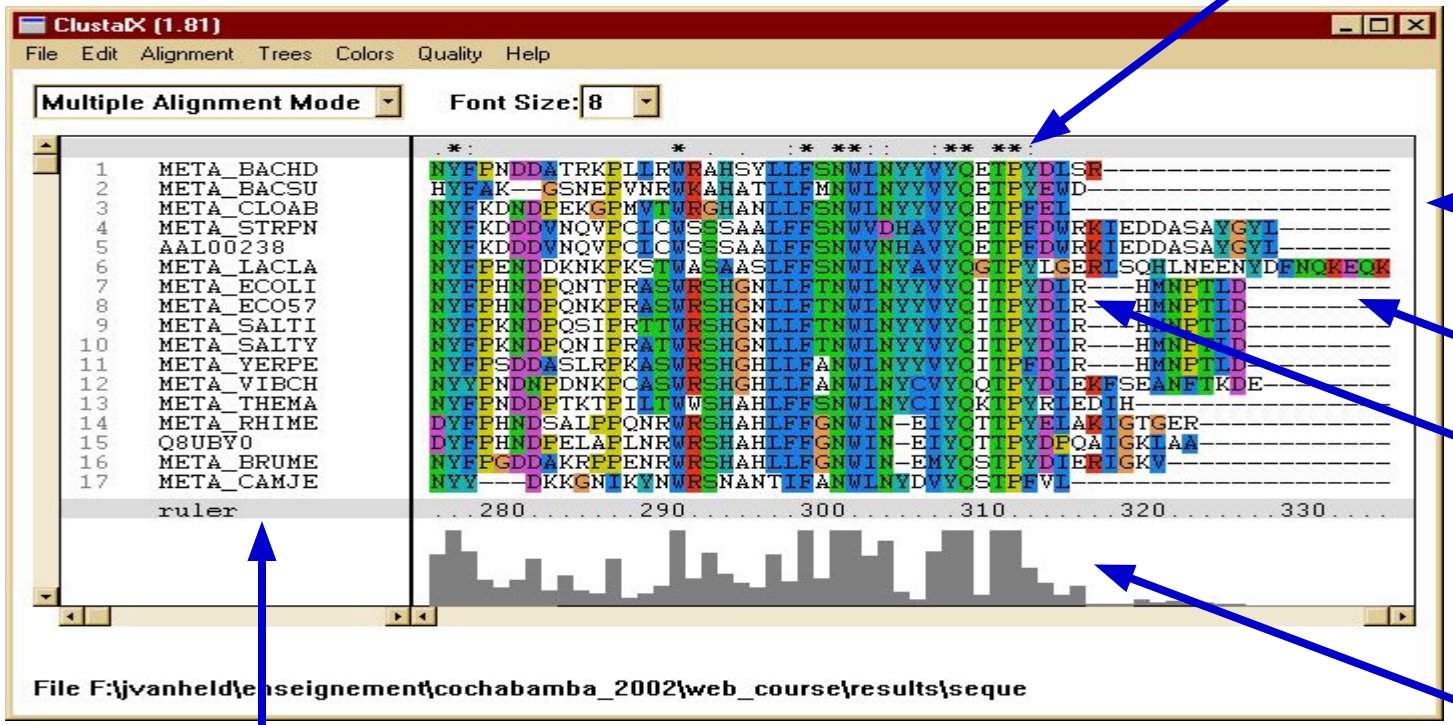
Seq5 **GATTGTA---GTA**
Seq1 **GATGGTA---GTA**
Seq4 **GATTGTTC--GTA**
Seq2 **GATTGTTCGGGTA**
Seq3



Seq5 GATTGTA-----GTA
Seq1 GATGGTA-----GTA
Seq4 GATTGTTC----GTA
Seq2 GATTGTTCGG--GTA
Seq3 **GATGGTAGGCGTGTA**



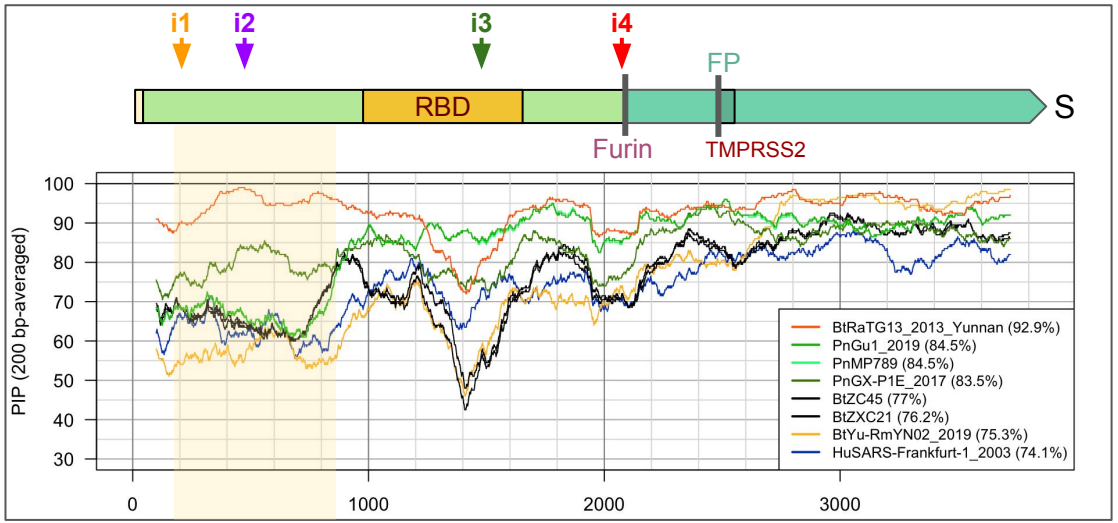
Alignement multiple global : Homoserine-O-dehydrogenase

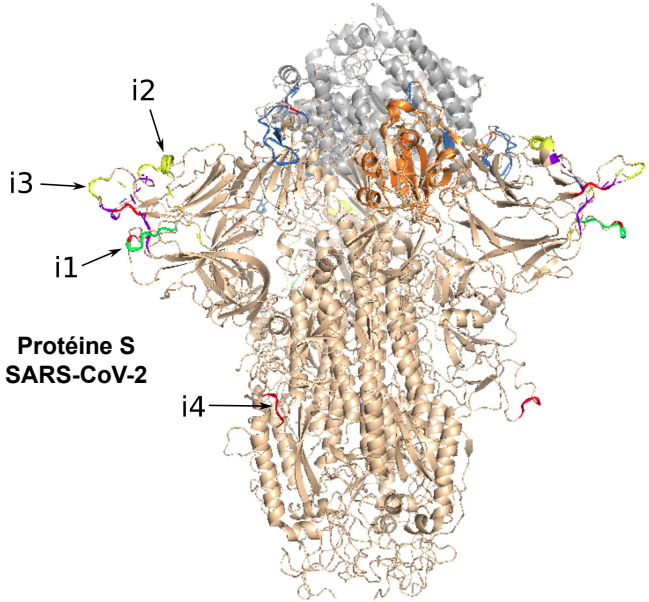


Insertions dans les séquences du gène S

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.

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". "

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.

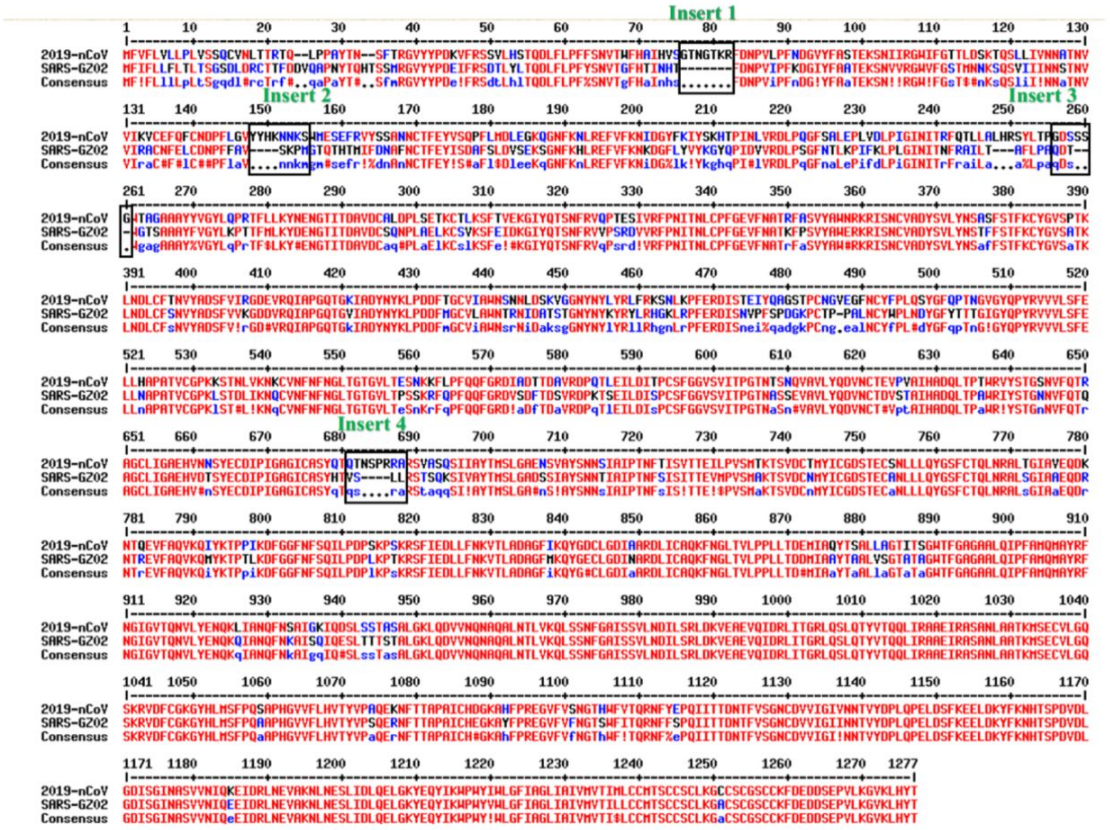
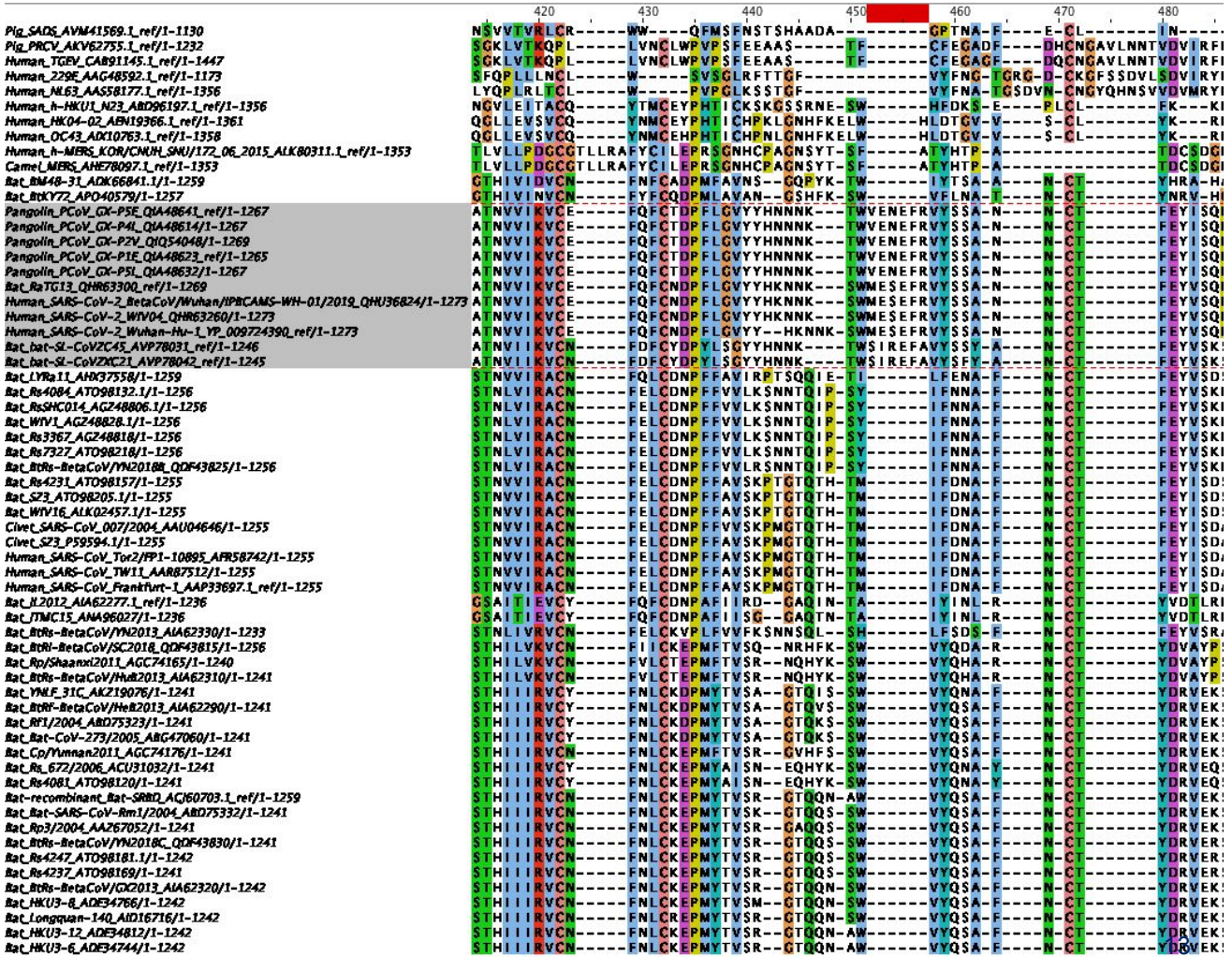


Figure 2: Multiple sequence alignment between spike proteins of 2019-nCoV and SARS. The sequences of spike proteins of 2019-nCoV (Wuhan-HU-1, Accession NC_045512) and of SARS CoV (GZ02, Accession AY390556) were aligned using MultiAlin software. The sites of difference are highlighted in boxes.

Etude des insertions / délétions dans les alignements multiples de la protéine spicule (spike)

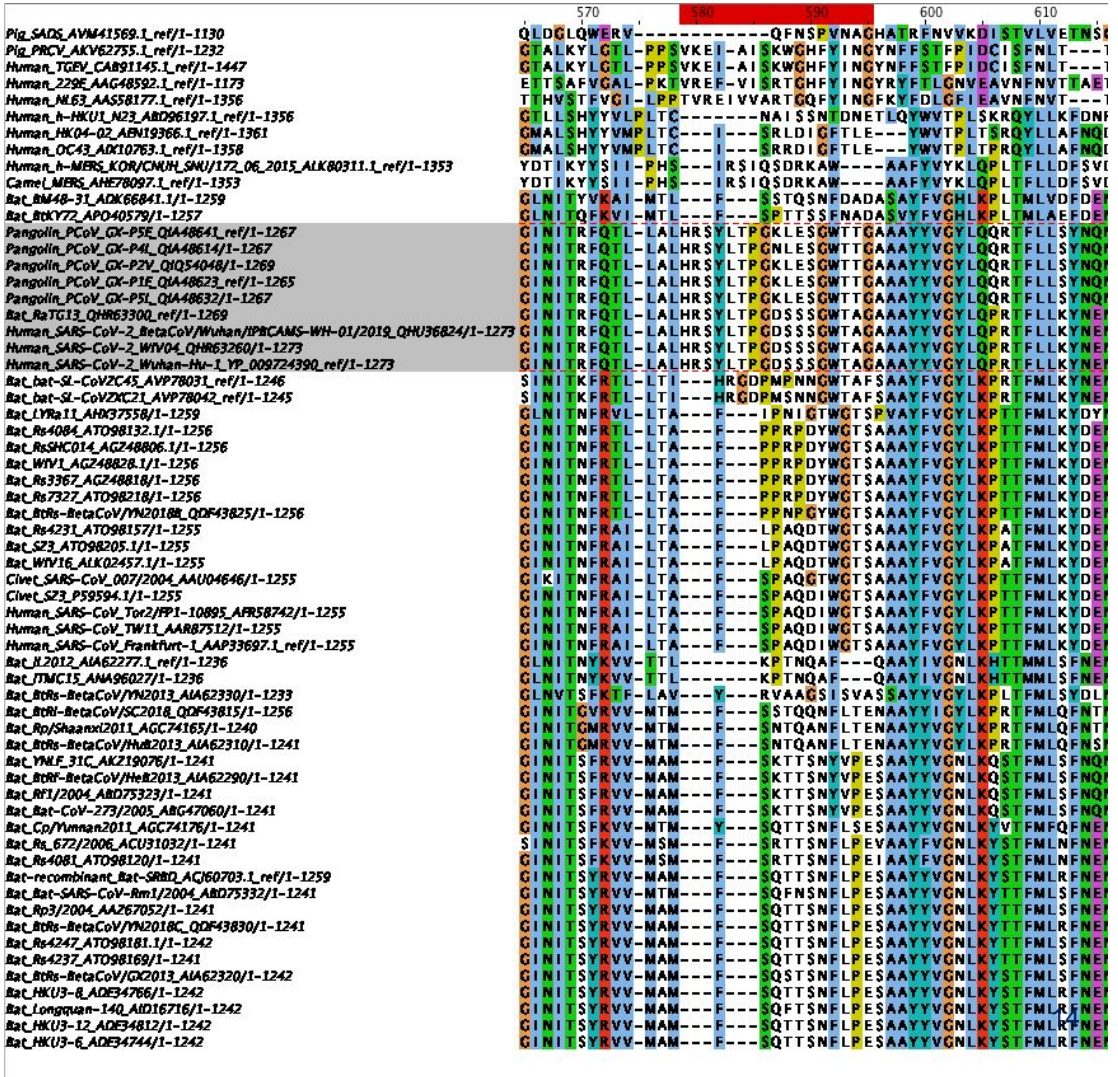
Insertion partagée entre tous les virus du groupe CoV-2

- 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.



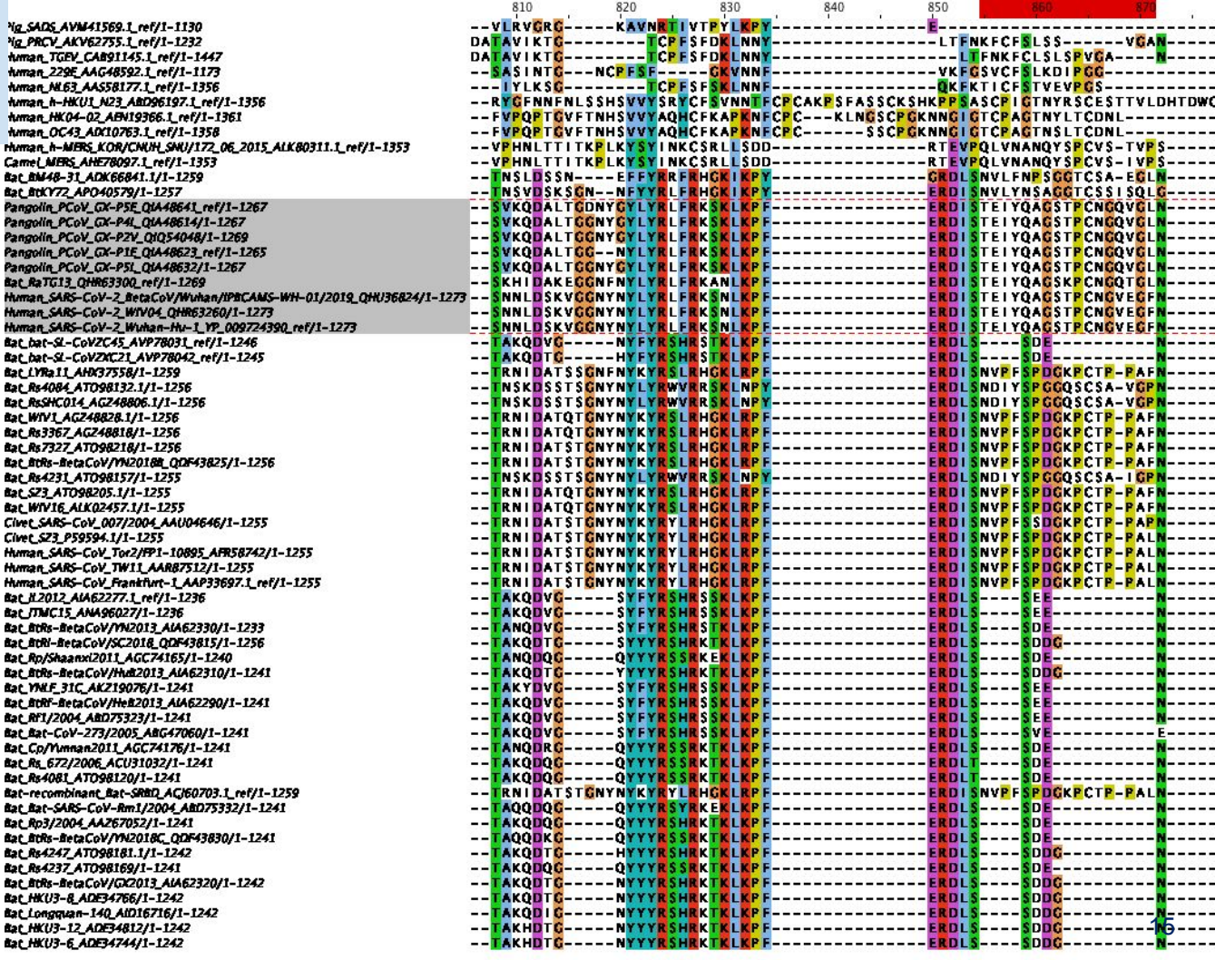
Insertion partagée par la majorité des virus du groupe CoV-2

- 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 .



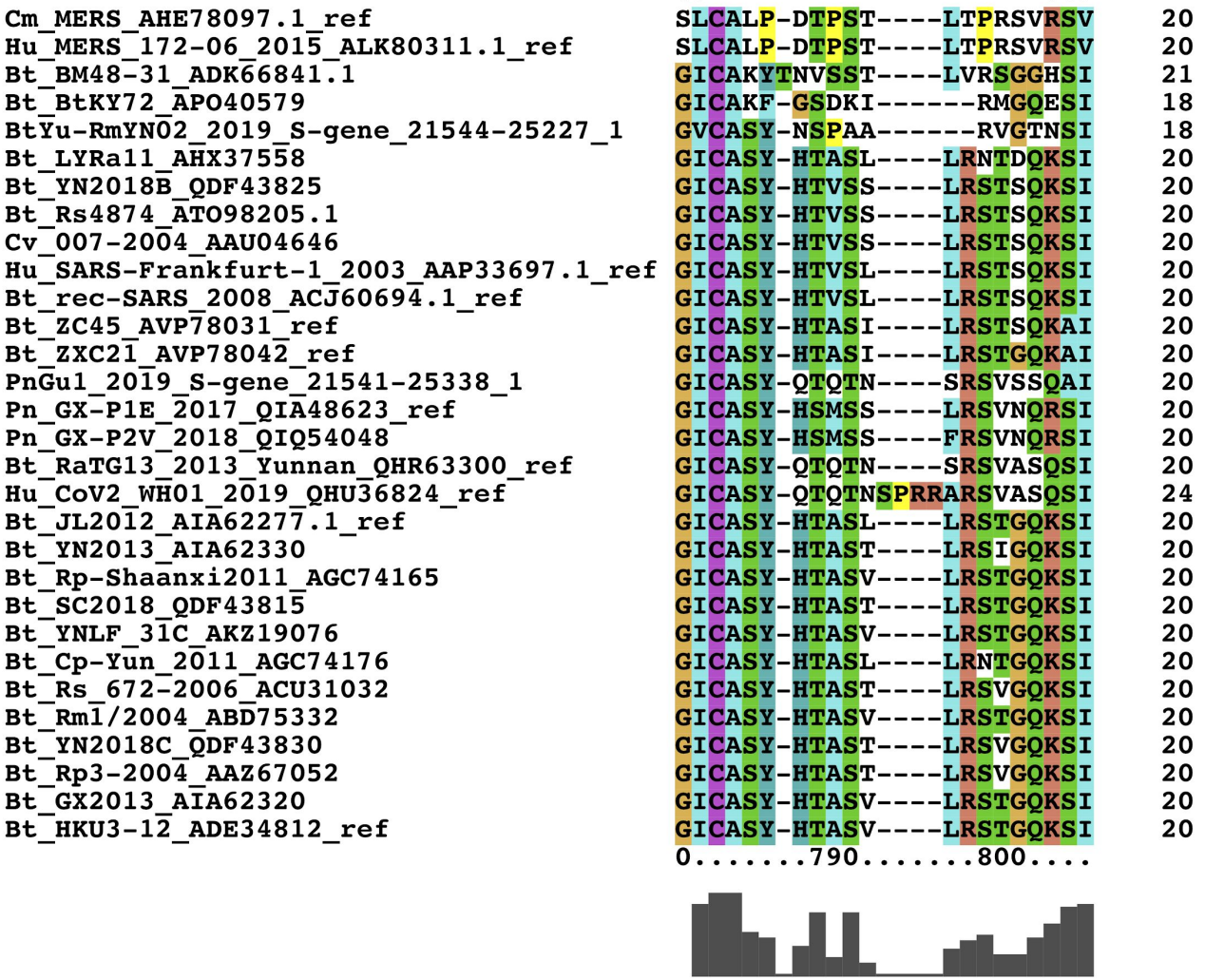
Insertion i3

- 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



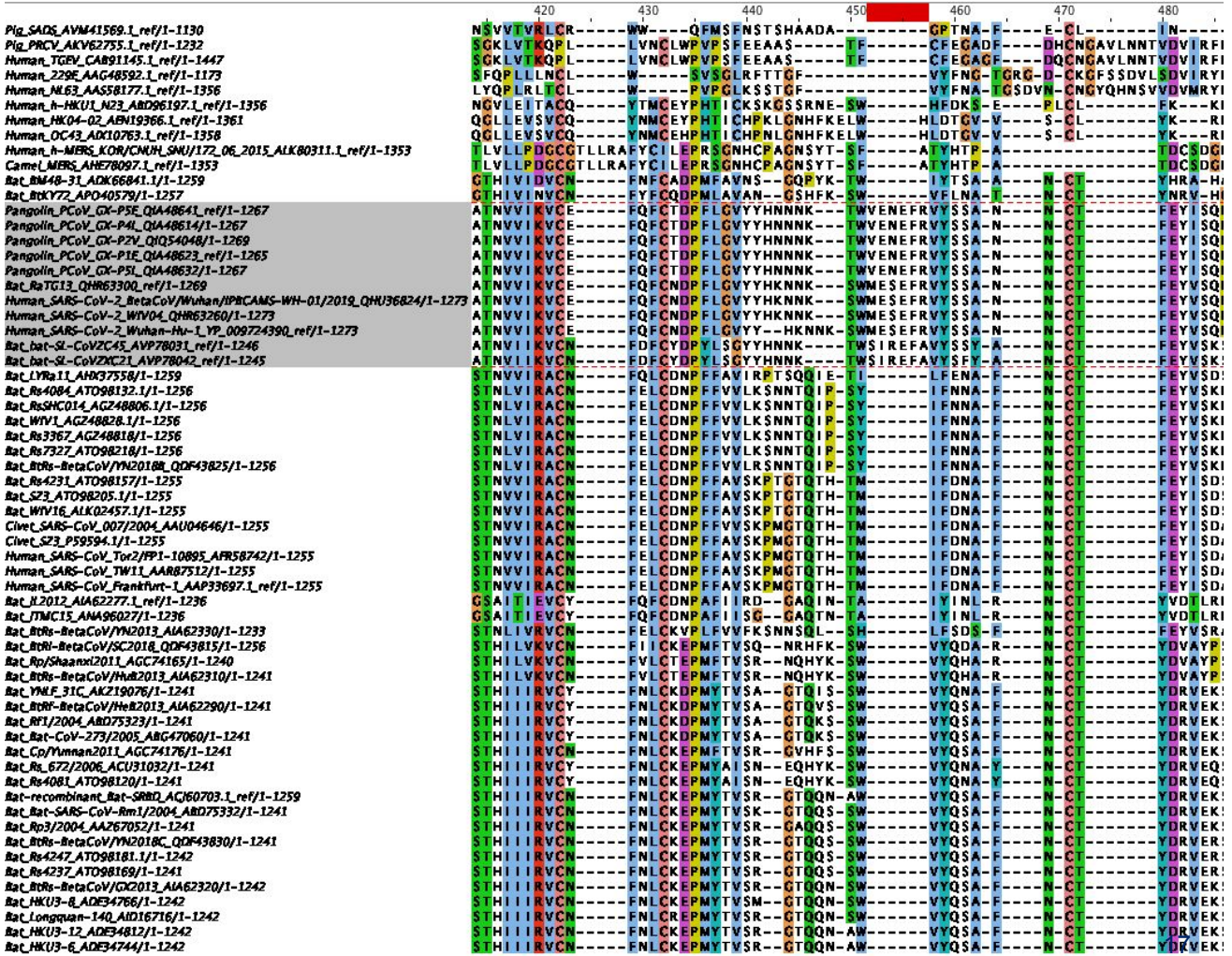
Insertion d'un site Furine (i4)

- 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



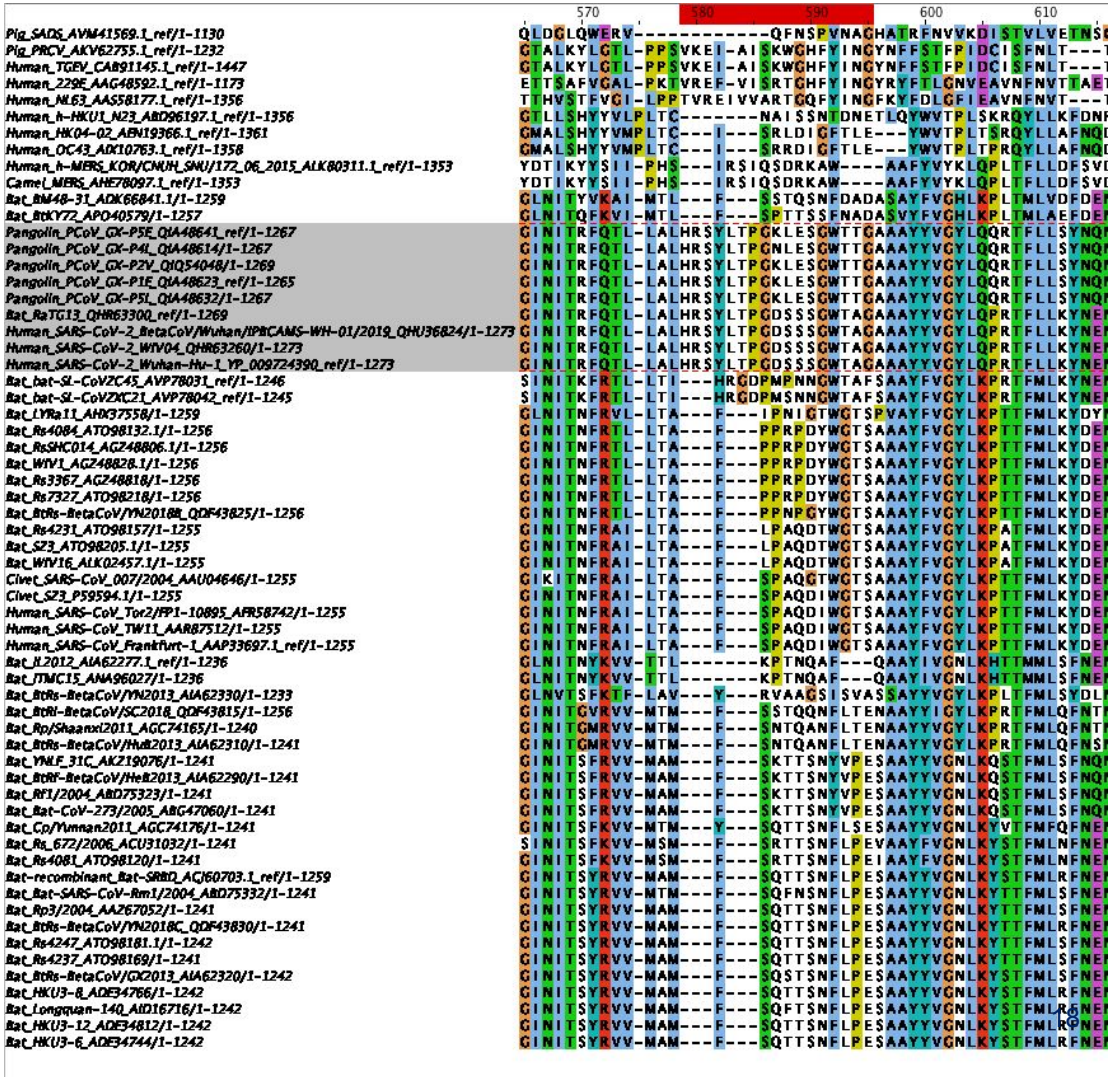
Insertion partagée entre tous les virus du groupe CoV-2

- 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.



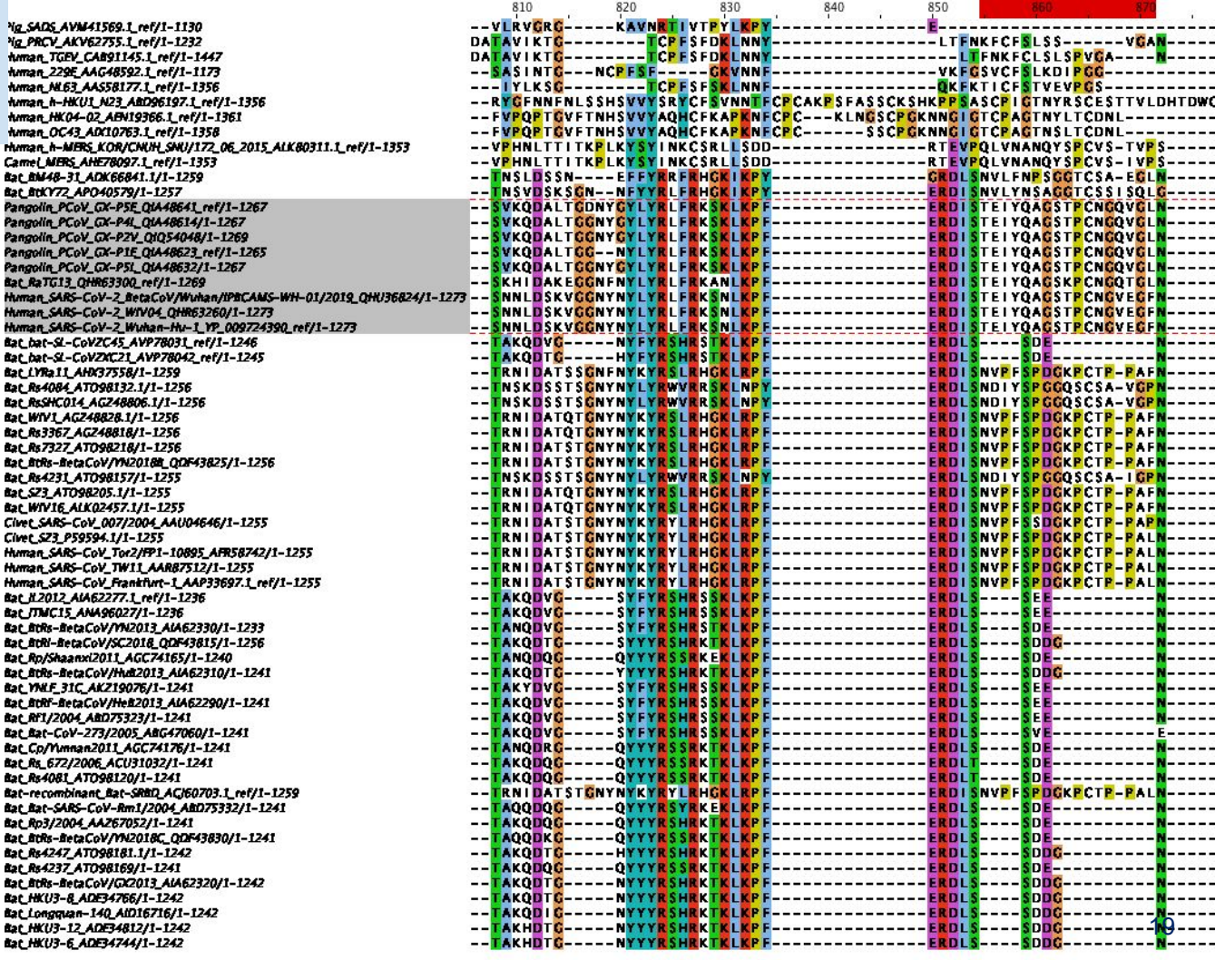
Insertion partagée par la majorité des virus du groupe CoV-2

- 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 .



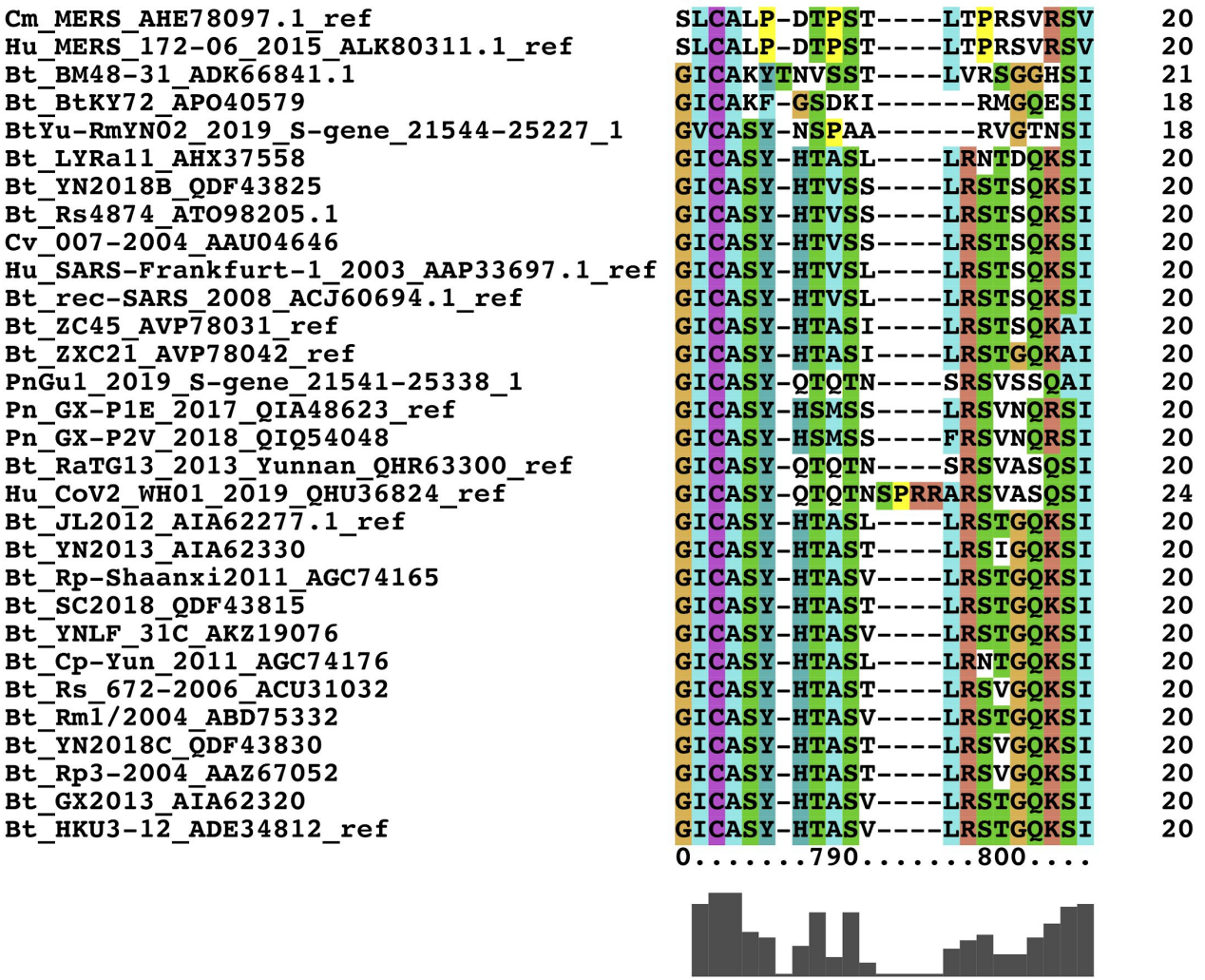
Insertion i3

- 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



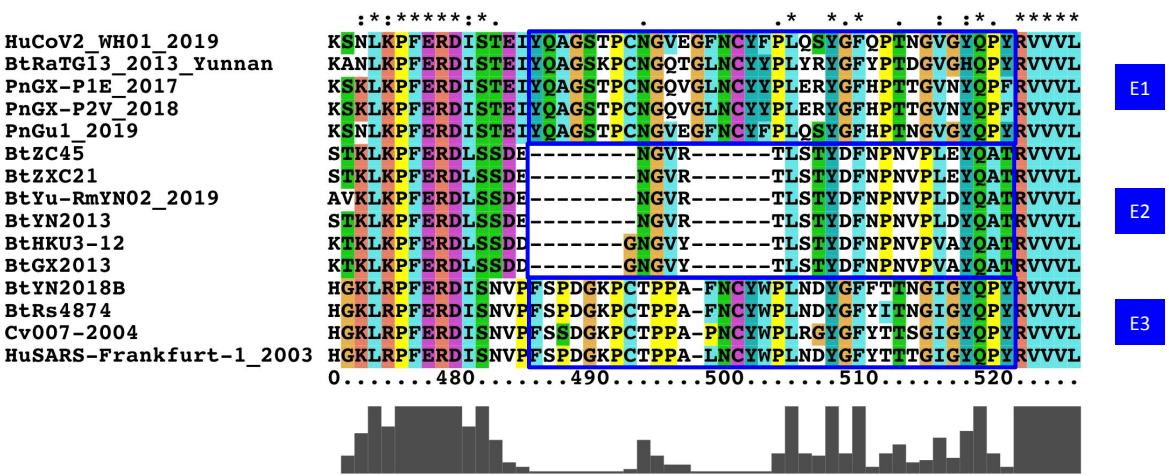
Insertion d'un site Furine (i4)

- 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

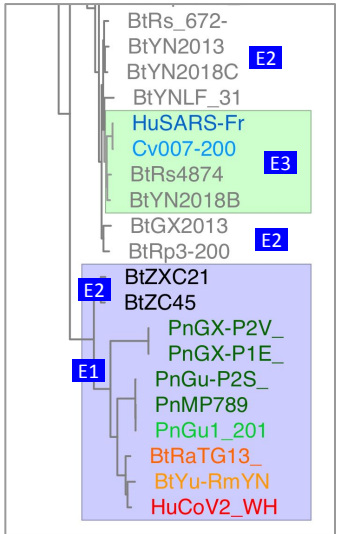


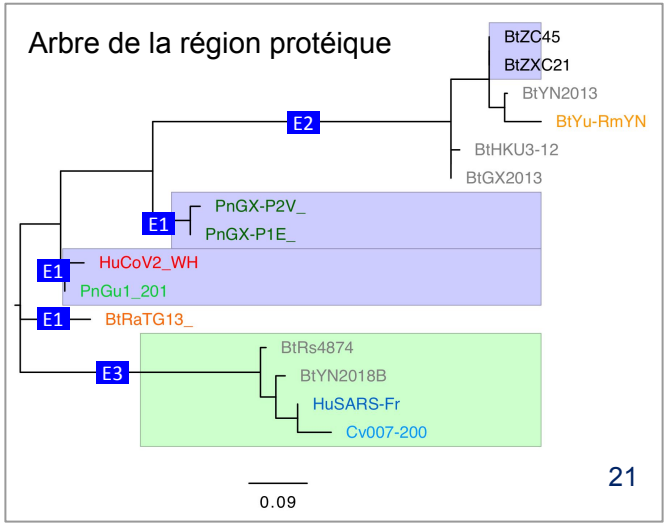
Un site recombinant?

- 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.



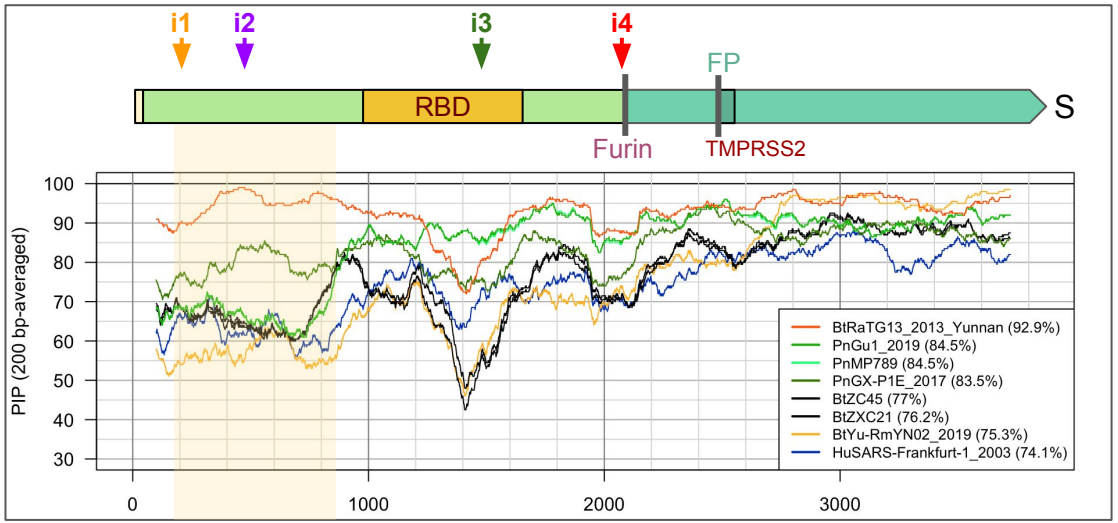
Arbre des génomes

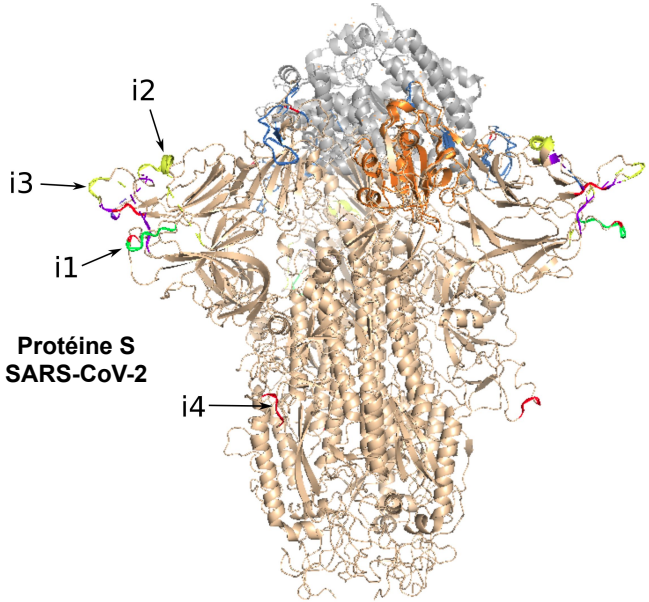




Four insertions in the S gene

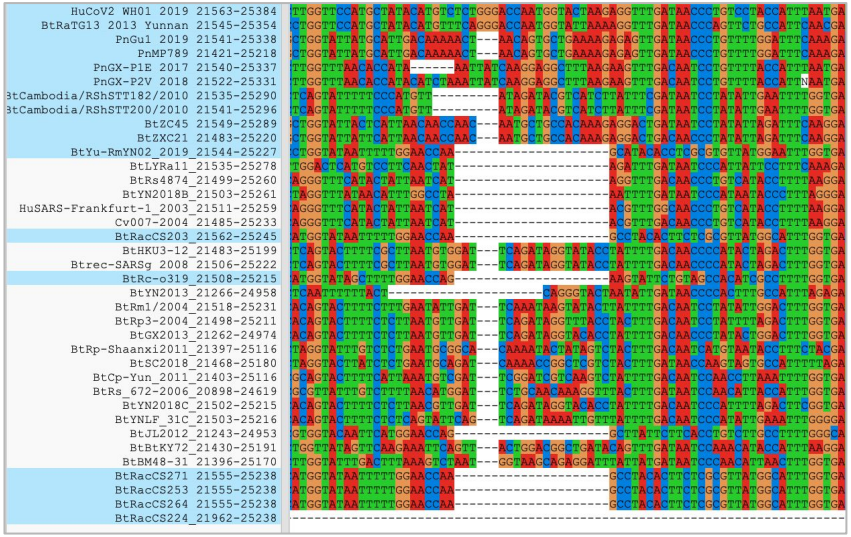
- Arrows indicate the 4 insertions in the S gene (left) and on the spike protein (left)
- Note:
  - the 3 first insertions are located in loops, at the exposed surface of the protein
  - These regions are known to be immunogenic, and highly variable (some mutations can trigger immune escape)

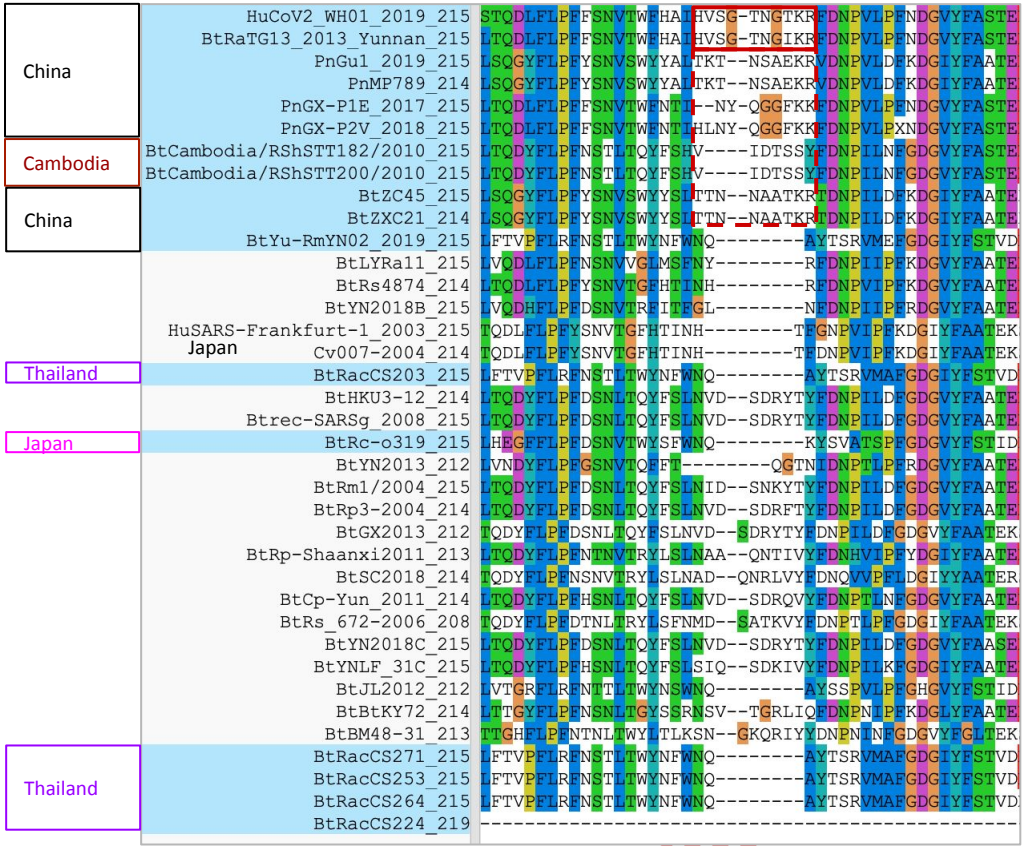




Insertion 1

- Found in all the SARS-CoV-2.
- Other insertions in the same region, in CoV-2 and CoV groups (dashed rectangles)
- No insertion in new Bat viruses from Japan or Thailand



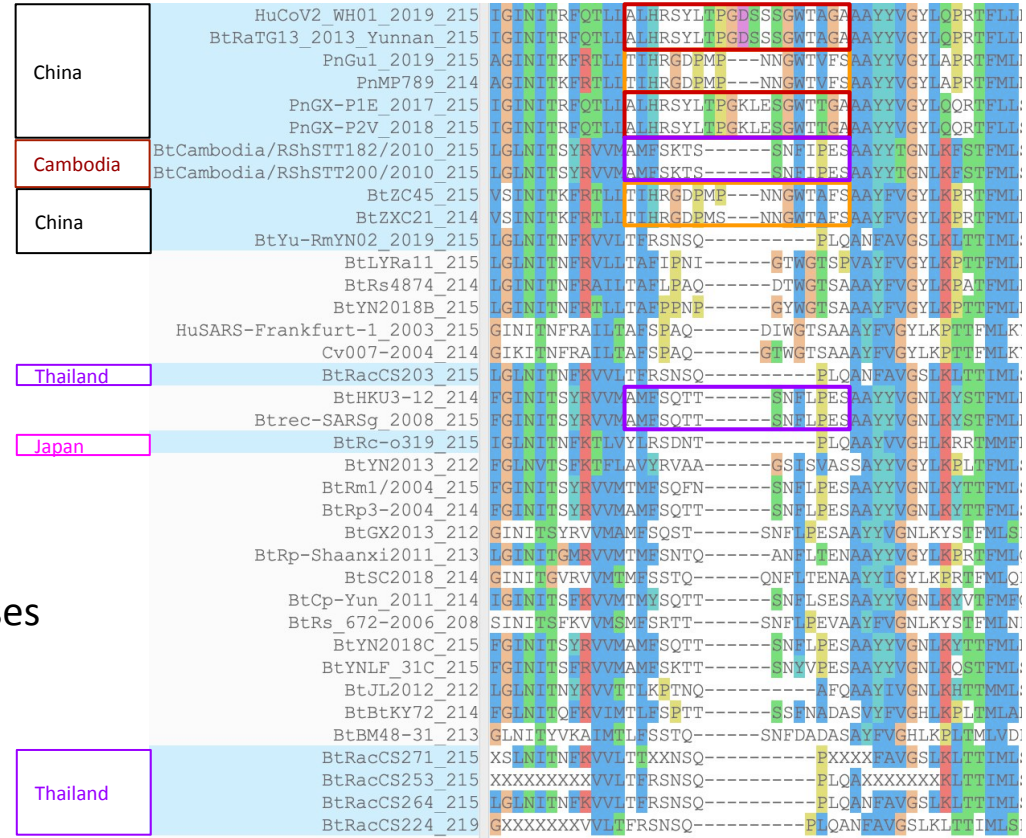


Insertion 2

- Region 144-157 of SARS-CoV-2
- Highly variable region between coronaviruses
- Shared between
  - SARS-CoV-2
  - RaTG13
  - Pangolin viruses
  - BtZXC21 et BtZC45
- Different insertions in the other genomes of Cov2 group (highlighted in blue)
  - RmYN02 (metagenome from 11 Bat fecal samples from Yunnan, China).
  - Bat virus from Japan
  - Bat viruses from Thailand

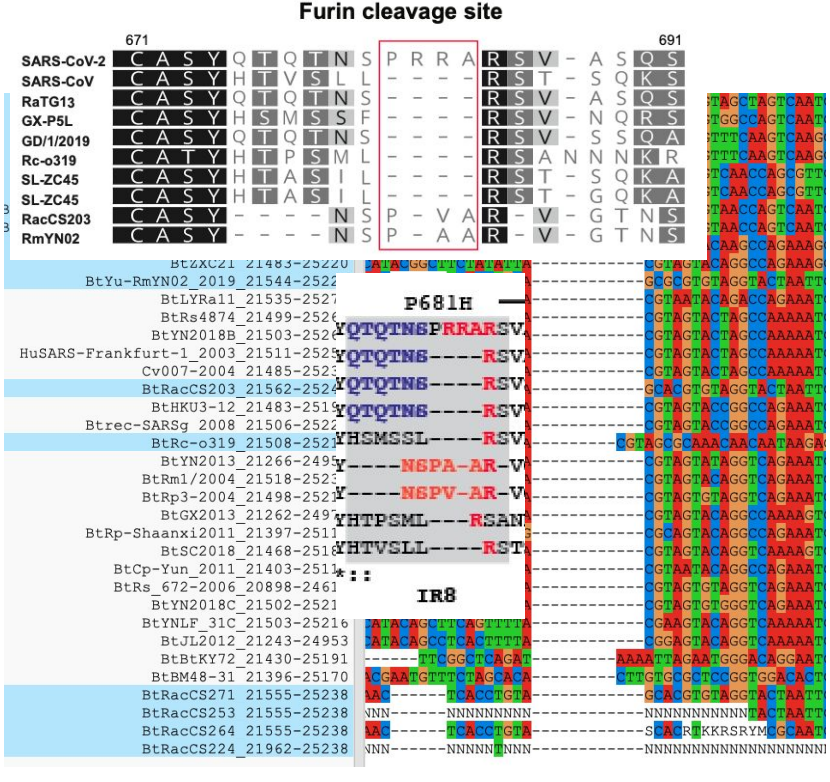
Insertion 3

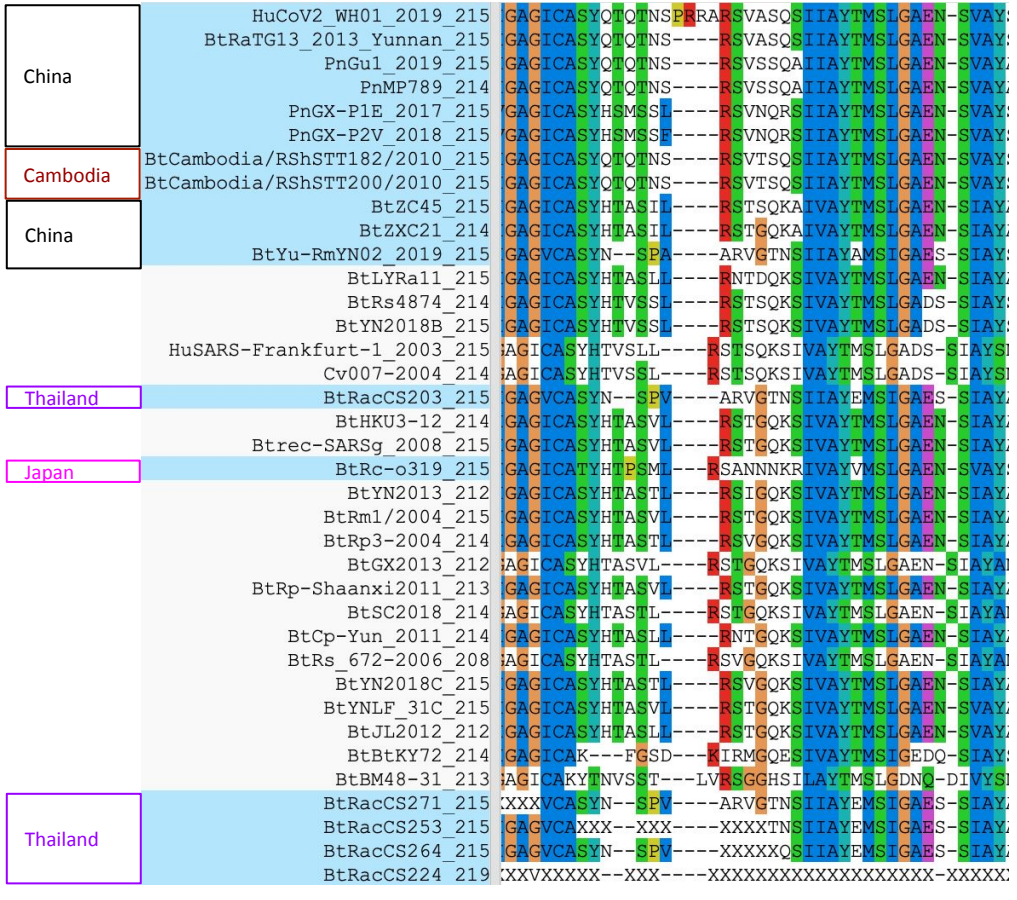
- Highly variable region
- Insertion shared between
  - SARS-CoV-2
  - RaTG13
  - GX Pangolins
- Another insertion shared between
  - GU Pangolins
  - BtZXC21 et BtZC45
- Note: insertion shared between
  - Cambodia (CoV2 group)
  - 2 bats from CoV1 group
- Yet different sequences in other Bat viruses of CoV2 group
  - Japanese (Rc-o319)
  - Thai viruses (RacCS*)
-

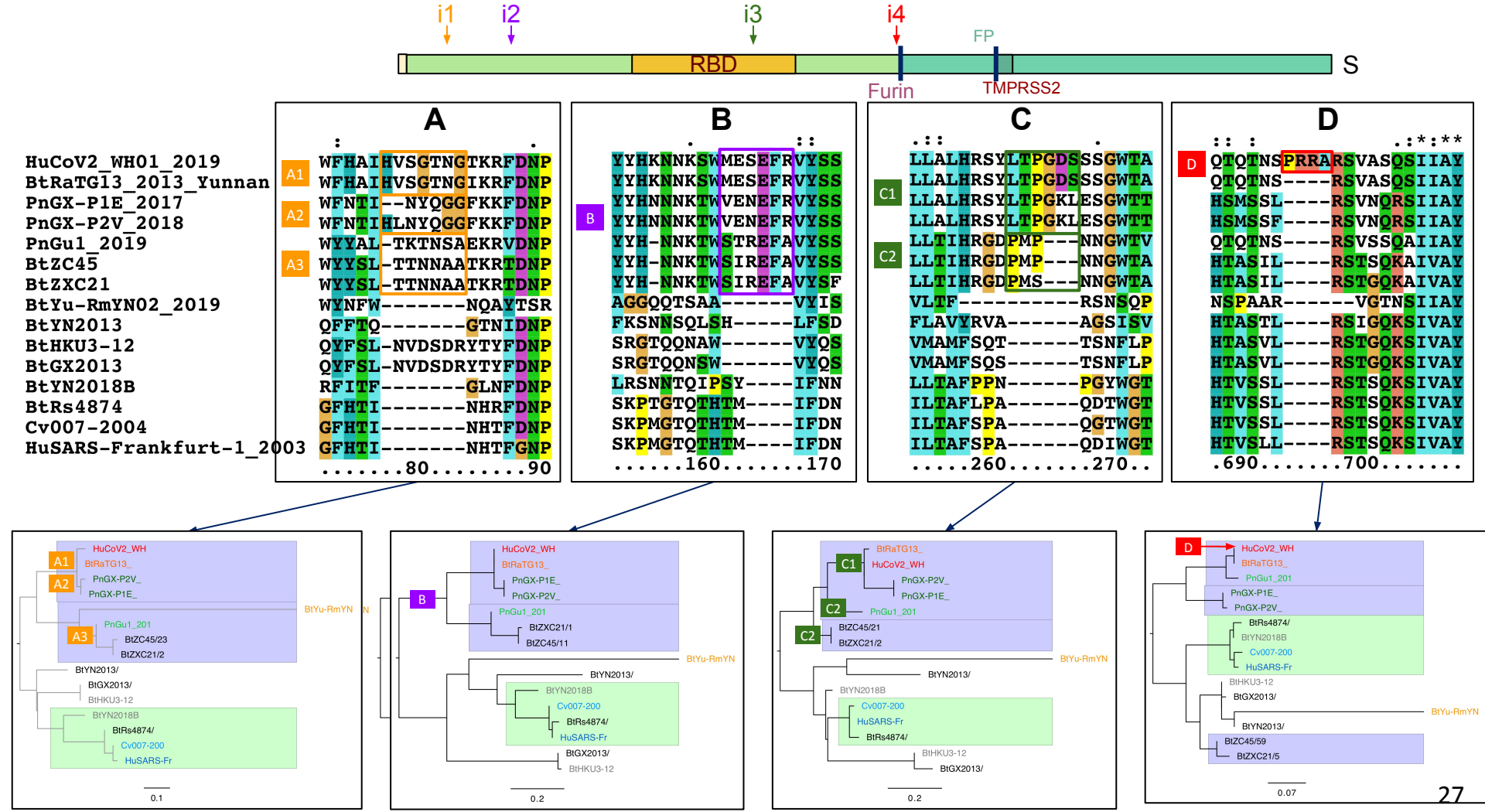


Insertion 4

Unique to SARS-CoV-2 : a furin cleavage site





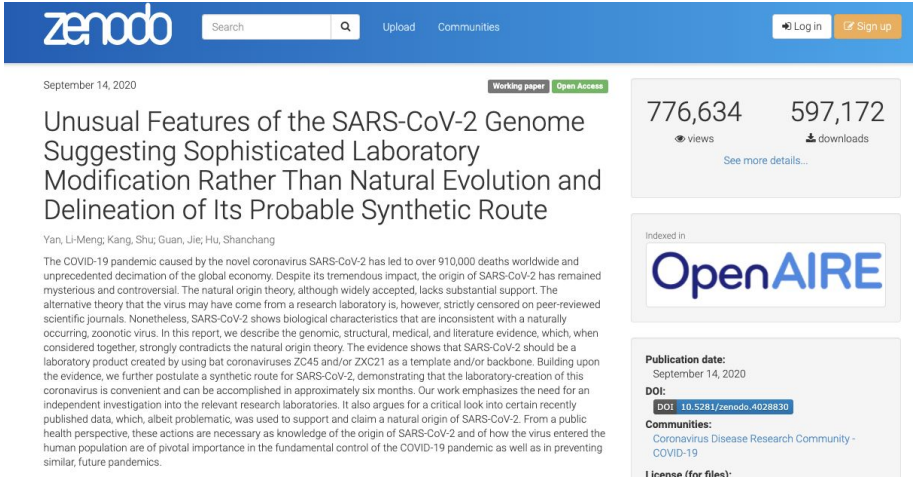


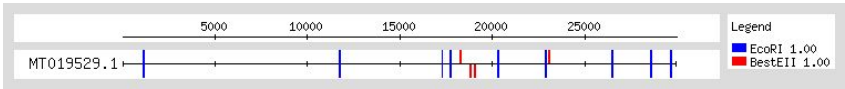
Sallard, E., Halloy, J., Casane, D., van Helden, J., Decroly, E. Retrouver les origines du SARS-COV-2 dans les phylogénies de coronavirus. Médecine/Sciences 36, Sept-Août 2020.

Un virus construit par ingénierie moléculaire ?

Un virus construit par ingénierie moléculaire ?

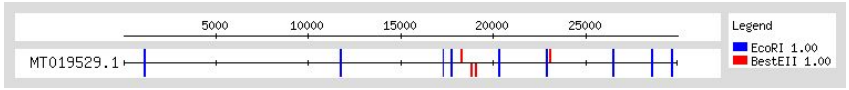
- Sept 2020: un preprint fait du bruit
- Li-Meng Yan
  - chercheuse chinoise
  - travaillait dans le laboratoire de référence de l'OMS pour la Chine
  - réfugiée aux Etats-Unis
- 600.000 téléchargements en 10 jours
- Arguments
  - Présence de sites de restriction dans le génome de SARS-Cov-2

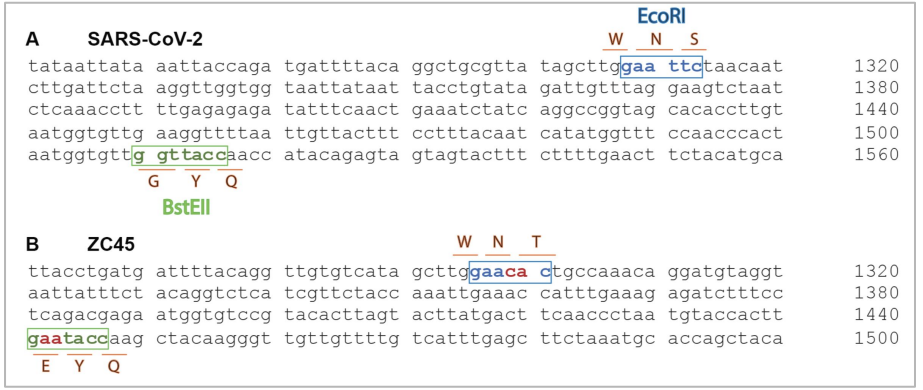




Des sites de restriction créés dans le génome de SARS-CoV-2 ?

- Li-Meng Yan détecte dans la séquence de SARS-CoV-2 des séquences correspondant aux sites de restriction de EcoRI et BstEII, souvent utilisés en biologie moléculaire pour créer de l'ADN recombinant au moyen d'enzymes de restriction ("ciseaux" moléculaires).
- Elle souligne que ces sites auraient pu être facilement créés à partir de sites très similaires présents dans le virus de chauve-souris Bat ZC45.
- Des questions se posent sur la signification de ce résultat
  - Probabilité de trouver ces sites aléatoirement ?
  - Les sites sont-ils uniques à SARS-CoV-2 dans le lignage des SARS-CoV ?
- Une première observation:
  - Le génome de SARS-CoV-2 contient 9 sites EcoRI et 4 sites BstEII
  - étant donné la courte taille des sites de restriction, et la taille du génome, on peut s'attendre à trouver de tels sites dans n'importe quelle séquence de 30 kilobases



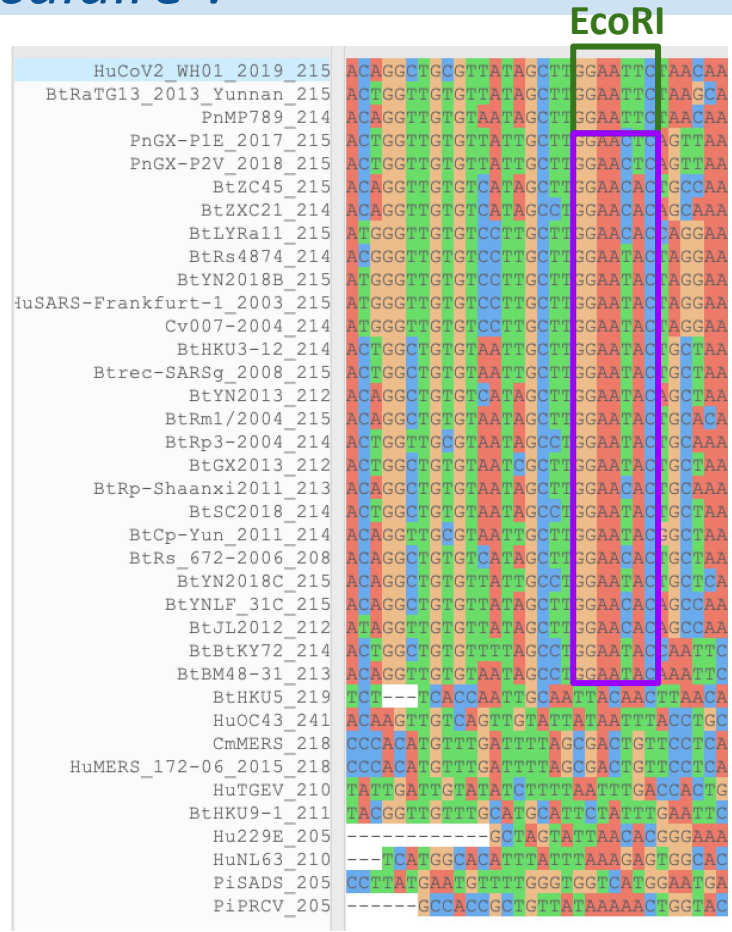


A. An EcoRI site is found at the 5'-end of the RBM and a BstEII site at the 3'-end.

B. Although these two restriction sites do not exist in the original spike gene of ZC45, they can be conveniently introduced given that the sequence discrepancy is small (2 nucleotides) in either case.

Un virus construit par ingénierie moléculaire ?

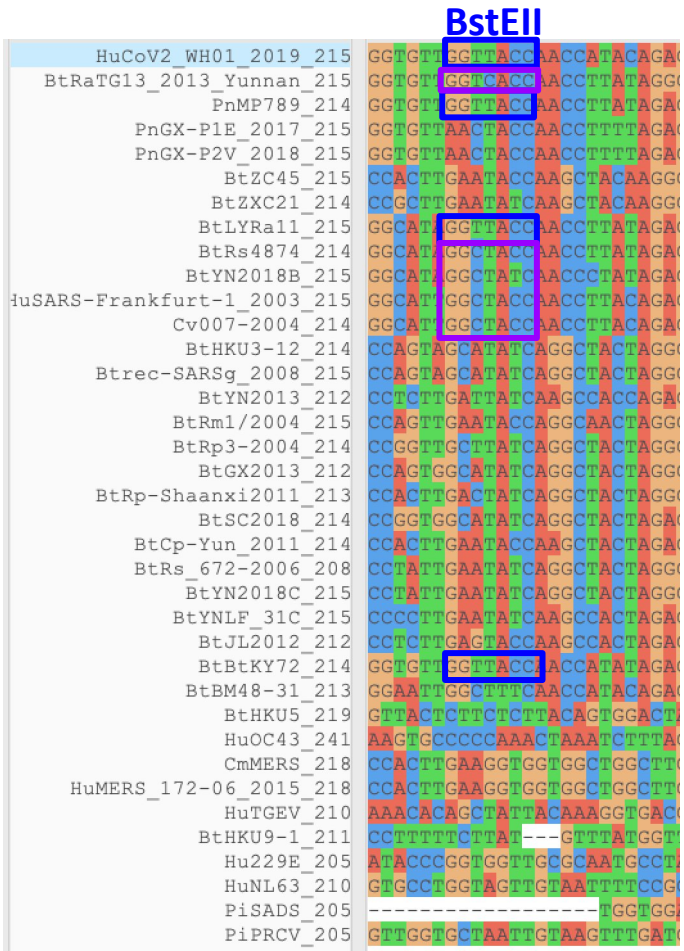
- Le site de restriction **EcoRI** (**séquence GAATTC**) se trouve dans le virus humain, mais également dans les virus les plus proches de chauve-souris (RatG13) et de pangolin (MP789).
- La majorité de ce site, ainsi que les séquences avoisinantes (**TTGGAAT*CT**) est conservée dans l'ensemble de la lignée SARS.
- Les séquences des autres SARS sont encore plus proche du site SARS-CoV-2 que celui de BtZ45 mis en avant par Li-Meng Yan.
- Il est donc plus vraisemblable que le site de SARS-CoV-2 soit apparu par une substitution d'un seul nucléotide à partir de n'importe quel SARS que par modification de 2 nucléotides du virus de chauve-souris BtZC45.



Quasi-identique à EcoRI

Un virus construit par ingénierie moléculaire ?

- Le site de restriction BstEII (**séquence GGTTACC**) se retrouve dans le virus humain, mais également dans les virus les plus proches de chauve-souris (RatG13) et de pangolin (MP789).
- Ce site se trouve également dans d'autres virus de la lignée, notamment ceux du SRAS humain (HuSARS-Frankfurt-1_2003), et de la civette (Cv007_2004).
- L'hypothèse de Li-Meng Yan (création à partir du génome de chauve-souris BtZC45) n'est donc pas convaincante.



Co-occurrence des sites de restriction qui encadrent le RBD