

Alignements de séquences multiples

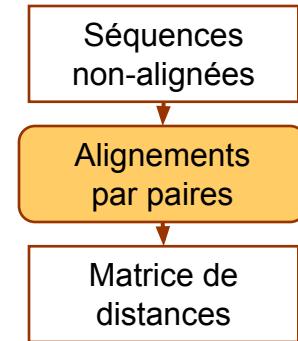
Jacques van Helden

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 – 1^{ère} 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.
 - $d_{i,j} = s_{i,j} / L_{j,j}$
 - $d_{i,j}$ distance entre les séquences i and j
 - $L_{j,j}$ longueur de l'alignement
 - $s_{i,j}$ 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: $d_{i,i} = 0$

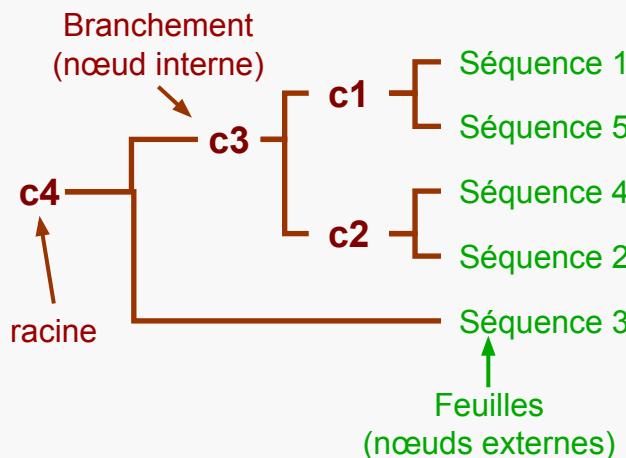
	seq 1	seq 2	...	seq n
seq 1	d1,1	d1,2	...	d1,n
seq 2	d2,1	d2,2	...	d2,n
...
seq n	dn,1	dn,2	...	dn,n

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

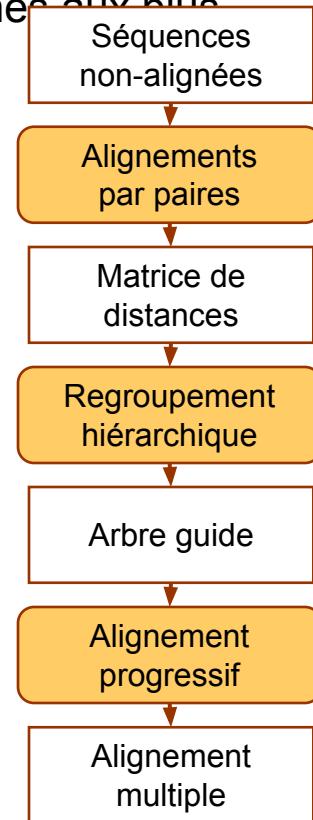
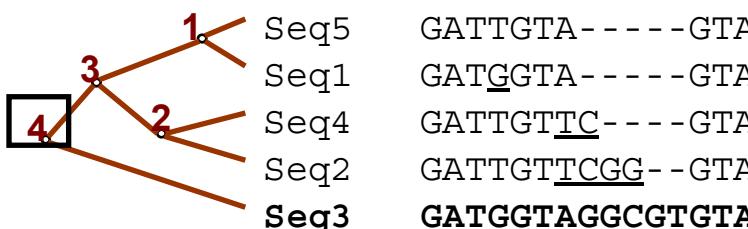
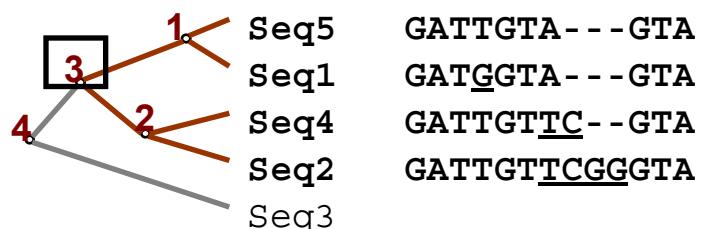
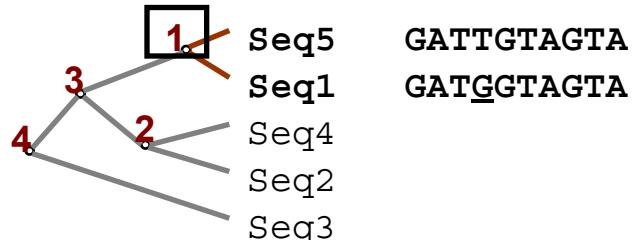
Arbre



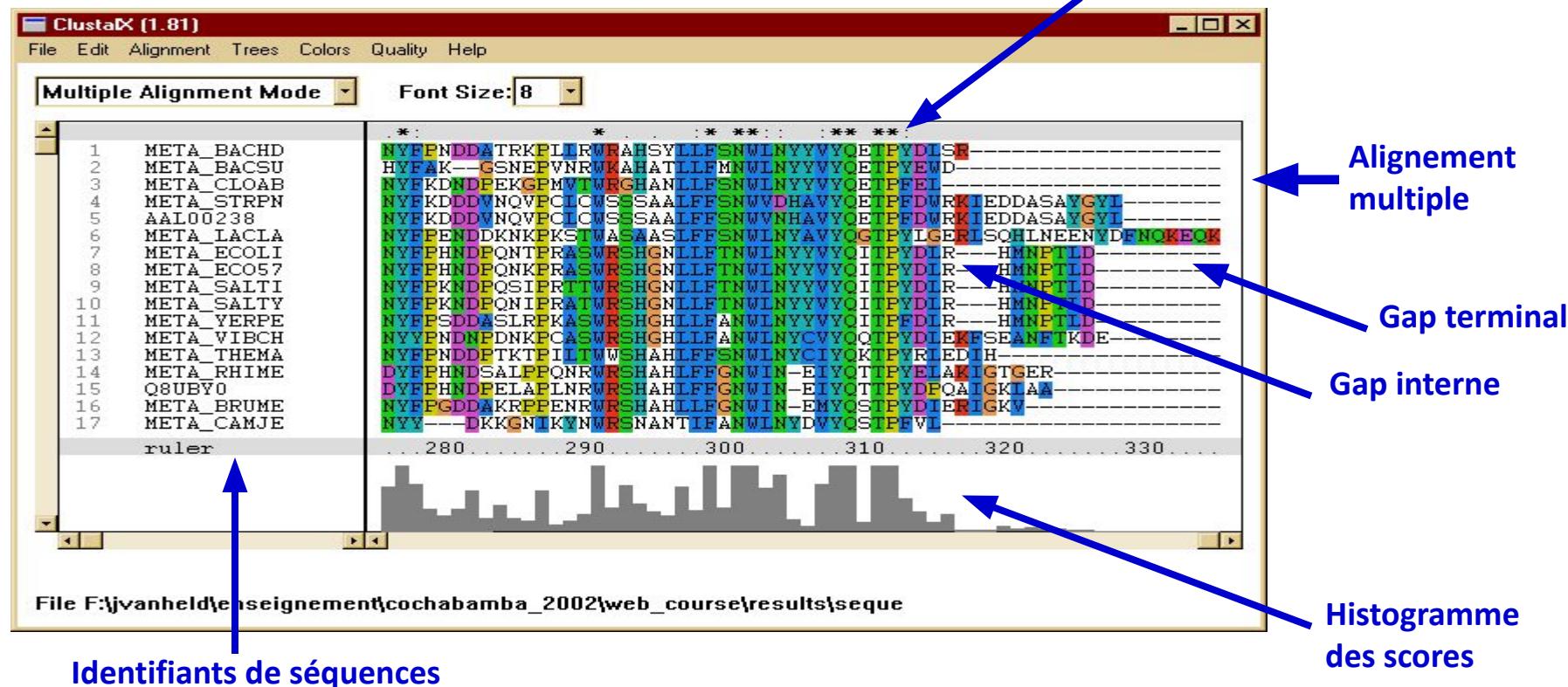
- Le clustering hiérarchique est une méthode de clustering agrégative.
 - Prend une matrice de distance en entrée
 - Regroupe progressivement les objets en allant des plus proches aux plus distants.
- Il existe plusieurs possibilités pour établir une règle d'agglomération, qui définit la distance entre 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 – 3^{ème} 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.



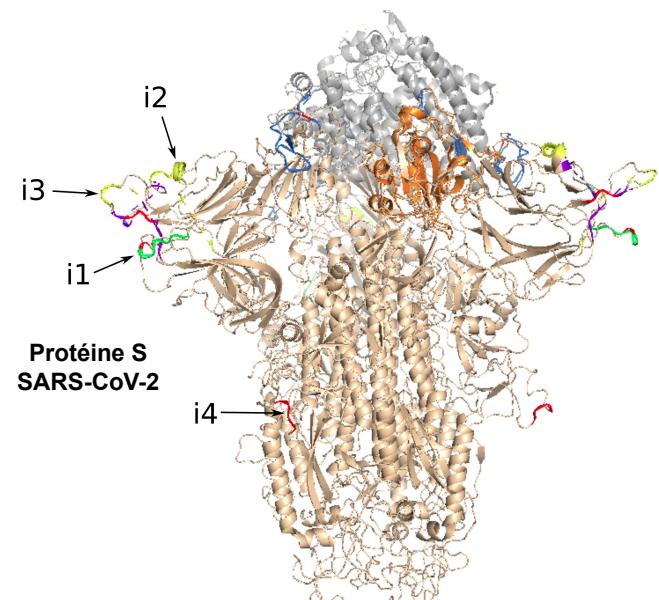
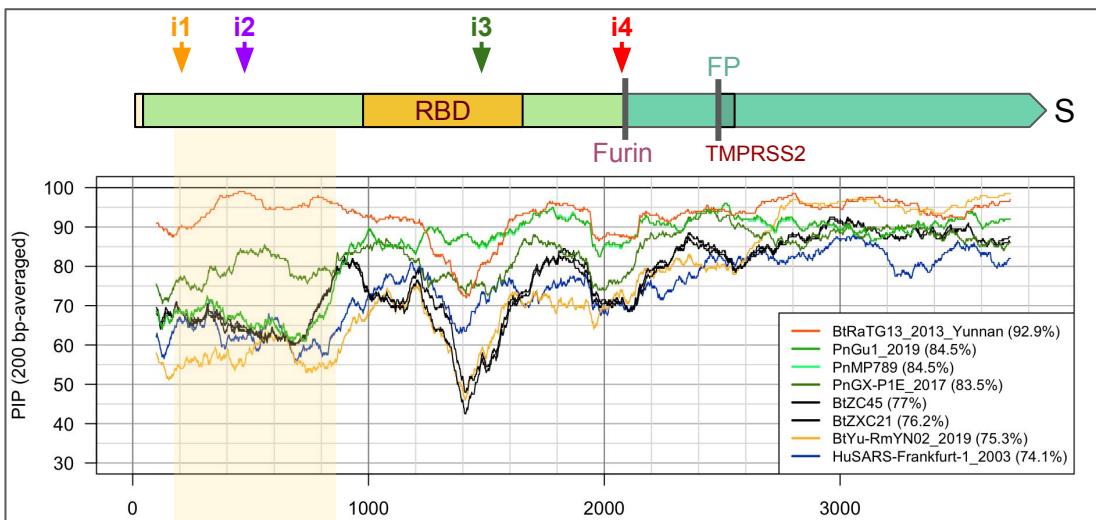
Alignement multiple global : Homoserine-O-dehydrogenase



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 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

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.

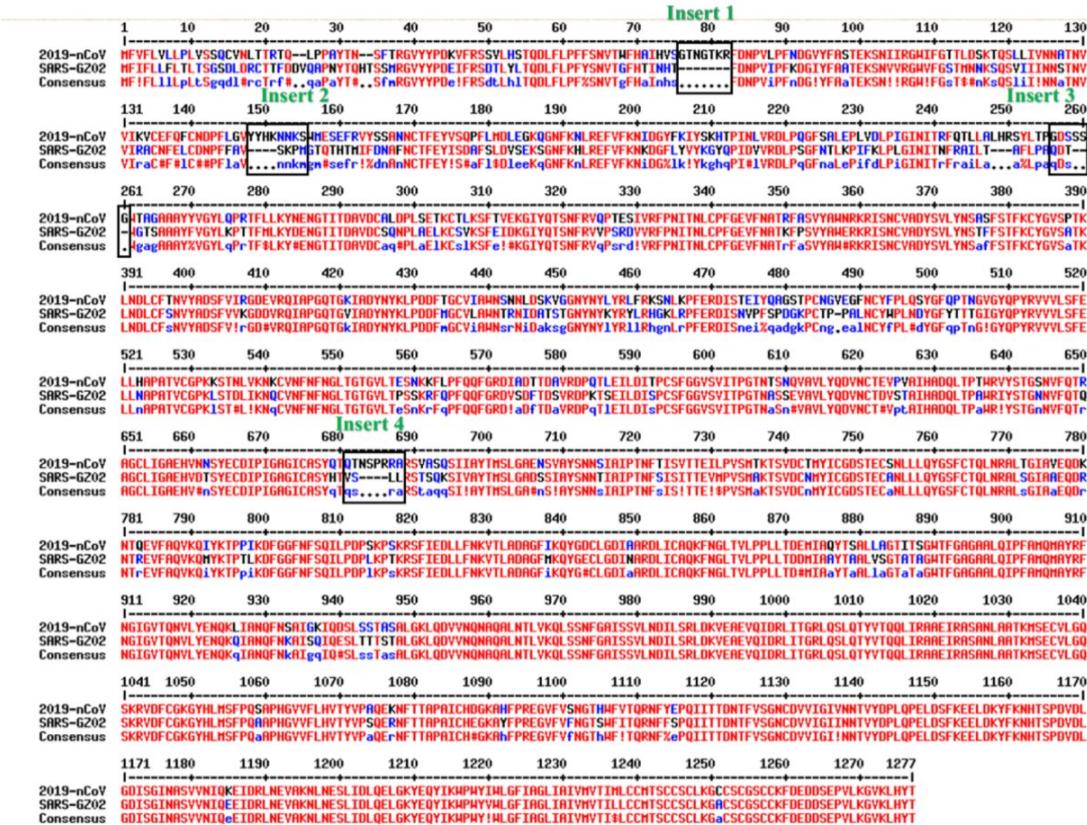
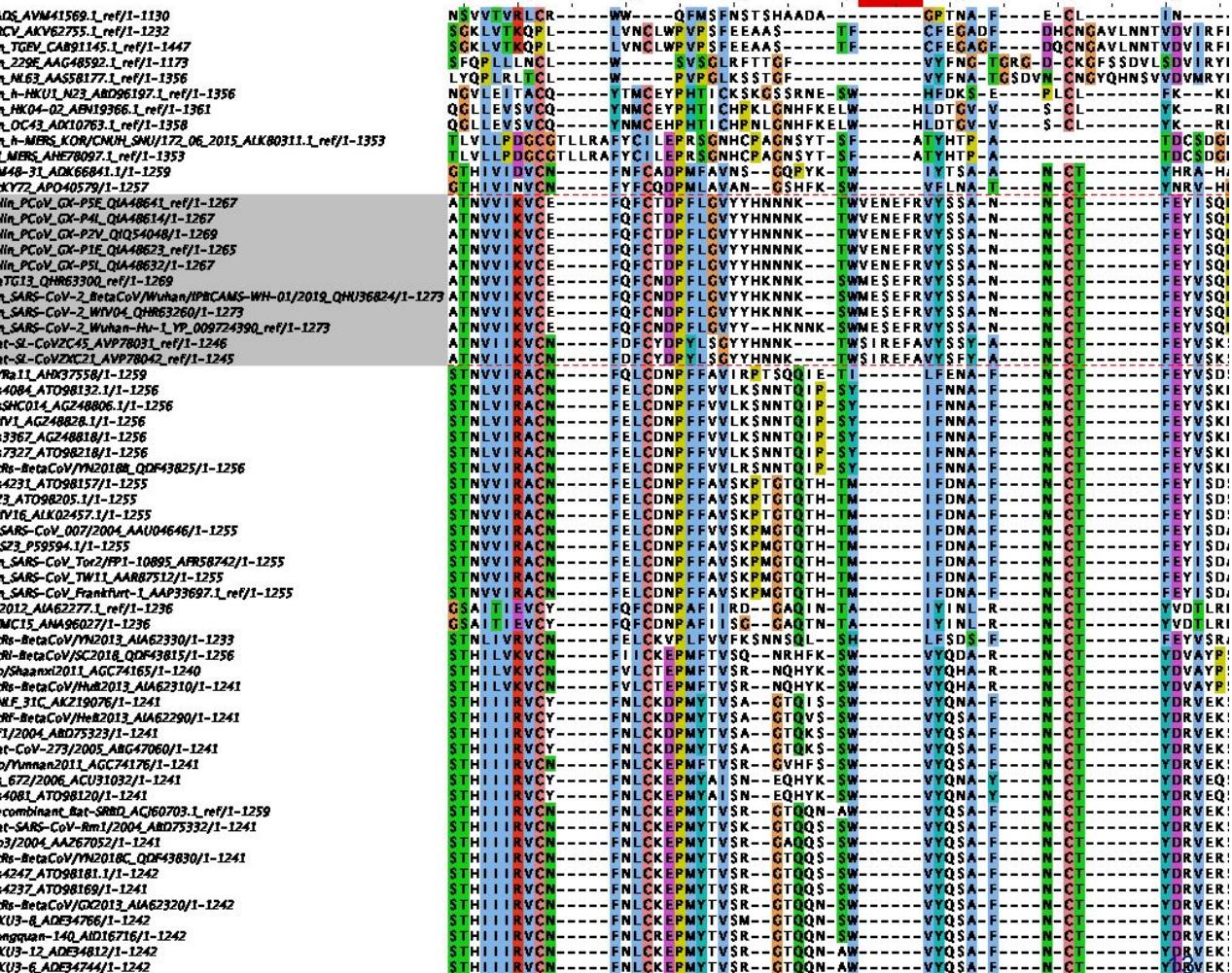


Figure 2: Multiple sequence alignment between spike proteins of 2019-nCoV and SARS. The sequences of spike proteins of 2019-nCoV (Wuhan-HU-1, Accession NC_045512) and of SARS CoV (GZ02, Accession AY390556) were aligned using MultiAlin software. The sites of difference are highlighted in boxes.

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

_SADS_AVAM41111_1_ref/1-1130
_PRCV_AKV627551_ref/1-1232
_man_TGEV_CAB911451_1_ref/1-1447
_man_229E_AA465921_1_ref/1-1173
_man_ML63_AA5581771_1_ref/1-1356
_man_h-HKU1_N23_AB0961971_1_ref/1-1356
_man_HK04_02_ABH193661_1_ref/1-1361
_man_OC43_ADX102631_1_ref/1-1358
_man_h-MERS_COV/NCU/SHN/172_06_2015_ALK80311.1_ref/1-1353
_man_MERS_AHE780971_1_ref/1-1353
_BM4-31_ABG66841/1-1259
_RKY72_AP0405791_1-1257
_gogin_PCoV_GX-P5E_QIA48641_ref/1-1267
_gogin_PCoV_GX-P4L_QIA48614/1-1267
_gogin_PCoV_GX-P2Y_QIO54048/1-1269
_gogin_PCoV_GX-P1E_QIA48623_ref/1-1265
_gogin_PCoV_GX-P5L_QIA48632/1-1267
_RaTG13_QHR6330Q_ref/1-1269
_man_SARS-CoV-2_Wuhan/WH-01/2019_QHU36824/1-1273
_man_SARS-CoV-2_WIV04_QHR63260/1-1273
_man_SARS-CoV-2_Wuhan-Hu-1_WP_009724390_ref/1-1273
_bat-SL_CoV2Z45_AVP78031_ref/1-1246
_bat-SL_CoV2ZC21_AVP78042_ref/1-1245
_LYm11_ANH375581/1-1259
_Re4084_AT098132/1/1-1256
_ReSHC014_AG248806/1/1-1256
_WIV1_LAG248828/1/1-1256
_Re3367_AG248818/1/1-1256
_Re7327_AT098218/1/1-1256
_BtRS-BetaCoV/WN2018N_QDF43825/1-1256
_Re4231_AT098157/1-1255
_SP3_AT098205/1/1-1255
_WV16_ALK02457.1/1-1255
_WV16_ALK02457.1/1-1255
_man_SARS-CoV_007/2004_AAU04646/1-1255
_SP2_P59594/1/1-1255
_man_SARS-CoV_Tor/FP1-10895_APR58742/1-1255
_man_SARS-CoV_TW11_AAR87512/1-1255
_man_SARS-CoV_Frankfurt-1_AAP39697.1_ref/1-1255
_H2012_ALA62277.1_ref/1-1236
_TMIC15_ANA96027/1-1236
_BtRS-BetaCoV/WN2013_AIA62330/1-1233
_BtRS-BetaCoV/SC2018_QDF43815/1-1236
_Rp/Shaanxi2011_AGC74165/1-1240
_BtRS-BetaCoV/Hu/d2013_AIA62310/1-1241
_YMLF_31C_AK219076/1-1241
_BtRS-BetaCoV/Hu/d2013_AIA62290/1-1241
_R1/2004_ABD75323/1-1241
_Bat-CoV-273/2005_AGC47060/1-1241
_Co/Yunnan2011_AGC74176/1-1241
_Re_672/2006_ACU31032/1-1241
_Re_4081_AT098120/1-1241
_recombinant_Bat-SRBC_AJC60703_1_ref/1-1259
_Bat-SARS-CoV-Rml/2004_ABD75332/1-1241
_Rp/2/2004_AA267052/1-1241
_BtRS-BetaCoV/WN2018C_QDF43830/1-1241
_Re4247_AT098181/1/1-1242
_Re4237_AT098169/1/1-1241
_BtRS-BetaCoV/GDZ013_AIA62320/1-1242
_HKU3-8_ADE34766/1-1242
_Longnan-140_ADI61716/1-1242
_HKU3-12_ADE54812/1-1242
_HKU3-6_ADE34744/1-1242

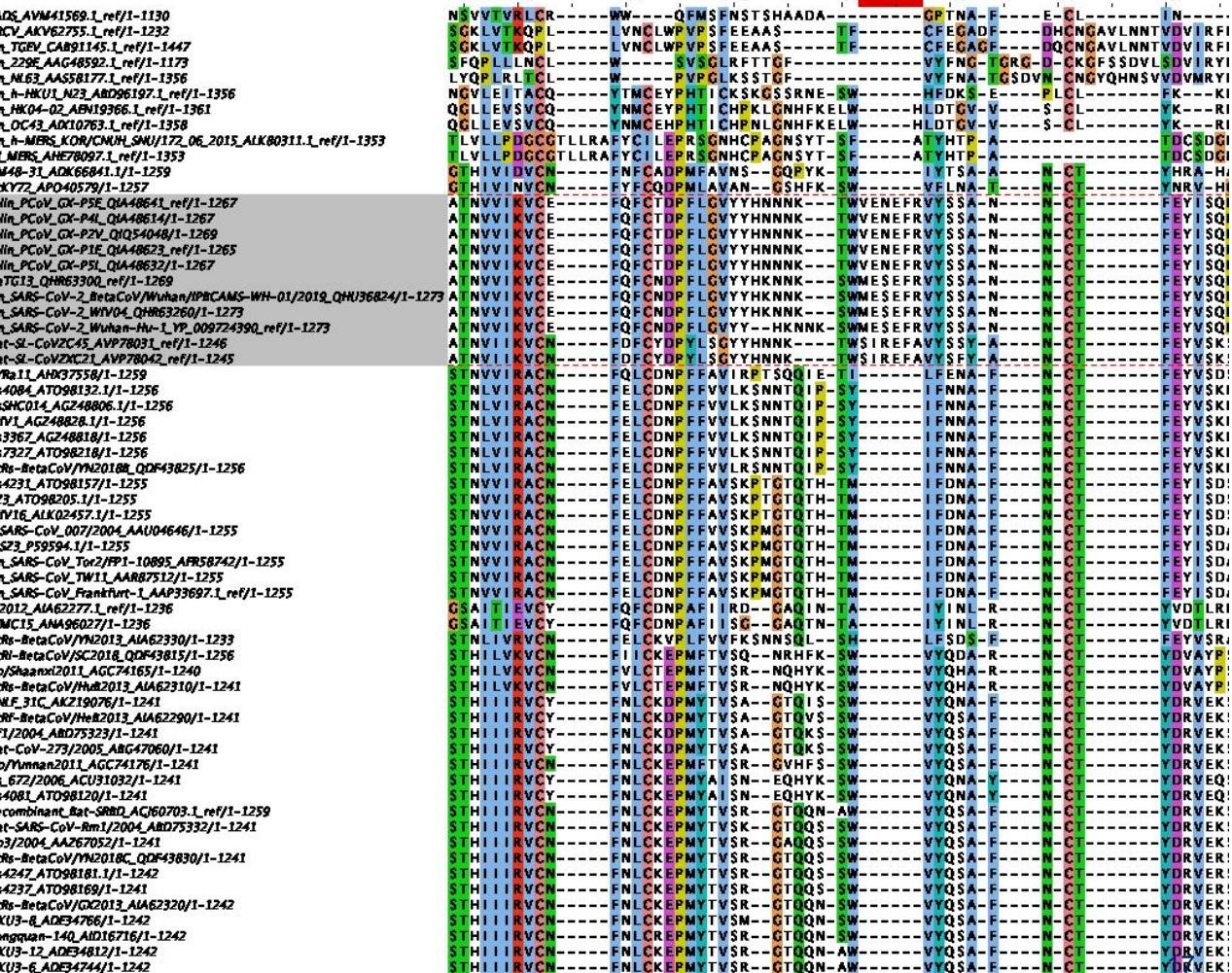
Insertion d'un site Furine (i4)

- 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 .

M41569_1.ref/1-1130	QDLDGLQWERV	-QFNNSPVNAGHAIRFNVVVKDITSTLVETNSI
KV62755_1.ref/1-1232	TALKYLGTL	-PPSVKEIAIKWCHFYINCYNNFSFPRTDCISFNLT--
_CA891145_1.ref/1-1447	TALKYLGTL	-PPSVKEIAIKWCHFYINCYNNFSFPRTDCISFNLT--
_AAG48592_1.ref/1-1173	ETTSASFVGAL	-PKTVREVRVIEIVRTGHFYINQYNTLGNVEAVNFNVT
_AA558177_1.ref/1-1356	TTTHSFTFGV	-LPPTRVREIVARTGFIINGFKYDFLGEAVNFNVT
_KU1_N23_ALB96197_1.ref/1-1356	GMALSHYYVPL	-NAISSNEDNETQYWPVTPISKRQYLLKQFNF
4-02_ABH19366_1.ref/1-1361	GMALSHYYVPL	-I--RLDIFTLE---YWVTPITPSRQYLLAFNQ
3_ADK10763_1.ref/1-1358	GMALSHYYVPL	-I--RRDIFTLE---YWVTPITPSRQYLLAFNQ
BS_CoV/CMHU/SW/172_06_2015_ALK80311_1.ref/1-1359	YDTIKYYSII	-PHS--IRS1QSDRKAM---AAFYVYKLQPLTFLLDFSVI
_AHE78097_1.ref/1-1353	YDTIKYYSII	-PHS--IRS1QSDRKAM---AAFYVYKLQPLTFLLDFSVI
_ADK66841_1-1259	GLNI	-TYVKAIA-MTL--F--STSQNSFDADASYYAVFGVQPLTMVLDF
AP04579_1-1257	GLNI	-TQFKVI-MTL--F--SPTTSSFNADADASYYVFGVQPLTMVLDF
_ov_GK-P56_QIA48641.ref/1-1267	GINI	-TRFQTL-LALHRSYLTPGKLESQWTGAAAYYVGYLQPRTFLLSYNQ
_ov_GK-P41_QIA48614/1-1267	GINI	-TRFQTL-LALHRSYLTPGKLESQWTGAAAYYVGYLQPRTFLLSYNQ
_ov_GK-P27_QD54048/1-1269	GINI	-TRFQTL-LALHRSYLTPGKLESQWTGAAAYYVGYLQPRTFLLSYNQ
_ov_GK-P16_QIA48623.ref/1-1265	GINI	-TRFQTL-LALHRSYLTPGKLESQWTGAAAYYVGYLQPRTFLLSYNQ
_ov_GK-P51_QIA48632/1-1267	GINI	-TRFQTL-LALHRSYLTPGKLESQWTGAAAYYVGYLQPRTFLLSYNQ
QH86330Q.ref/1-1269	GINI	-TRFQTL-LALHRSYLTPGDSSSGWTAGAAAAYVGYLQPRTFLLKYNE
_CoV-2_BetaCoV/Wuhan/IPCAMS-WH-01/2019_QH86284/1-1273	GINI	-TRFQTL-LALHRSYLTPGDSSSGWTAGAAAAYVGYLQPRTFLLKYNE
_CoV-2_WiV04_QH863260/1-1273	SINI	-TRFQTL-LALHRSYLTPGDSSSGWTAGAAAAYVGYLQPRTFLLKYNE
CoVZ4C5_AVP78031.ref/1-1246	SINI	-TRFQTL-LTR-HRGDPMNNNGWTASAAFYVFGYKPRTFMLKYNE
CoVNC21_AVP78042.ref/1-1245	SINI	-TRFQTL-LTR-HRGDPMNNNGWTASAAFYVFGYKPRTFMLKYNE
ADM37558_1-1259	GLNI	-TRFRVLT-LTA-F--IPNIGWTGTSVAVYFGYKPRTTFMLKYD
AT098132_1-1-1256	GLNI	-TRFRVLT-LTA-F--PPRPDYNGTSAAAYVFGYKPRTTFMLKYD
_14_AG248806_1/1-1256	GINI	-TRFRVLT-LTA-F--PPRPDYNGTSAAAYVFGYKPRTTFMLKYD
Z48828_1/1-1256	GINI	-TRFRVLT-LTA-F--PPRPDYNGTSAAAYVFGYKPRTTFMLKYD
AG248818/1-1256	GINI	-TRFRVLT-LTA-F--PPRPDYNGTSAAAYVFGYKPRTTFMLKYD
AT098218/1-1256	GINI	-TRFRVLT-LTA-F--PPRPDYNGTSAAAYVFGYKPRTTFMLKYD
taCoV/YN2018_QDF43825/1-1256	GINI	-TRFRVLT-LTA-F--PPRPDYNGTSAAAYVFGYKPRTTFMLKYD
AT098157/1-1255	GINI	-TRFRVLT-LTA-F--LPAQDTWTGTSAAAYVFGYKPRATTFMLKYD
098205_1/1-1255	GINI	-TRFRVLT-LTA-F--LPAQDTWTGTSAAAYVFGYKPRATTFMLKYD
AL024571_1/1-1255	GINI	-TRFRVLT-LTA-F--LPAQDTWTGTSAAAYVFGYKPRATTFMLKYD
Cov_007/2004_AAU04646/1-1255	GK	-TRFRVLT-LTA-F--SPAQDGWTGTSAAAYVFGYKPRATTFMLKYD
59594/1-1-1255	GK	-TRFRVLT-LTA-F--SPAQDGWTGTSAAAYVFGYKPRATTFMLKYD
_CoV_Tor2/FP1-10895_AFR58742/1-1255	GINI	-TRFRVLT-LTA-F--SPAQDINGTSAAAYVFGYKPRTTFMLKYD
_CoV_TW/11_AAR87512/1-1255	GINI	-TRFRVLT-LTA-F--SPAQDINGTSAAAYVFGYKPRTTFMLKYD
_CoV_Frankfurt_1_AAP33697_1.ref/1-1255	GINI	-TRFRVLT-LTA-F--SPAQDINGTSAAAYVFGYKPRTTFMLKYD
AI62277_1.ref/1-1256	GINI	-TRFRVLT-LTA-F--SPAQDINGTSAAAYVFGYKPRTTFMLKYD
AN496027/1-1256	GLNI	-TRFRVLT-LTR-HKPTNQAF--QAAIYVGNLKHTTMMLSFNE
taCoV/YN2013_ALA62230/1-1233	GLNI	-TRFRVLT-LTR-HKPTNQAF--QAAIYVGNLKHTTMMLSFNE
taCoV/SC2018_QDF43815/1-1256	GLNV	-SPAKDINGTSAAAYVFGYKPRTTFMLKYD
nx2011_ALGC74165/1-1240	GLNV	-SPAKDINGTSAAAYVFGYKPRTTFMLKYD
taCoV/HuB2013_ALA62230/1-1241	GLNV	-SPAKDINGTSAAAYVFGYKPRTTFMLKYD
C_AK219076/1-1241	GLNV	-SPAKDINGTSAAAYVFGYKPRTTFMLKYD
taCoV/HuB2013_ALA62290/1-1241	GLNV	-SPAKDINGTSAAAYVFGYKPRTTFMLKYD
_ABD75323/1-1241	GLNV	-SPAKDINGTSAAAYVFGYKPRTTFMLKYD
_273/2005_ABC47060/1-1241	GLNV	-SPAKDINGTSAAAYVFGYKPRTTFMLKYD
an2011_ALGC74176/1-1241	GLNV	-SPAKDINGTSAAAYVFGYKPRTTFMLKYD
2006_ACU31032/1-1241	GLNV	-SPAKDINGTSAAAYVFGYKPRTTFMLKYD
AT098120/1-1241	GLNV	-SPAKDINGTSAAAYVFGYKPRTTFMLKYD
nant_Bat-SRMQ_AJC60703_1.ref/1-1259	SINI	-SFKVV-MTM-Y--SQTTSNLFPESSAAAYVGNLKVTTFMQFNE
_CoV_Rm1/2004_ABD75332/1-1241	SINI	-SFKVV-MTM-Y--SQTTSNLFPESSAAAYVGNLKVTTFMQFNE
4_AA267052/1-1241	SINI	-SFKVV-MTM-Y--SQTTSNLFPESSAAAYVGNLKVTTFMQFNE
taCoV/YN2018C_QDF43830/1-1241	SINI	-SFKVV-MTM-Y--SQTTSNLFPESSAAAYVGNLKVTTFMQFNE
AT098181_1/1-1242	SINI	-SFKVV-MTM-Y--SQTTSNLFPESSAAAYVGNLKVTTFMQFNE
AT098169/1-1241	SINI	-SFKVV-MTM-Y--SQTTSNLFPESSAAAYVGNLKVTTFMQFNE
taCoV/GX2013_ALA62320/1-1242	SINI	-SFKVV-MTM-Y--SQTTSNLFPESSAAAYVGNLKVTTFMQFNE
_ADE34766_1-1242	SINI	-SFKVV-MTM-Y--SQTTSNLFPESSAAAYVGNLKVTTFMQFNE
_n-14_AID16716/1-1242	SINI	-SFKVV-MTM-Y--SQTTSNLFPESSAAAYVGNLKVTTFMQFNE
2_ADE4812/1-1242	SINI	-SFKVV-MTM-Y--SQTTSNLFPESSAAAYVGNLKVTTFMQFNE
4_ADE2744/1-1242	SINI	-SFKVV-MTM-Y--SQTTSNLFPESSAAAYVGNLKVTTFMQFNE

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

Cm_MERS_AHE78097.1_ref
Hu_MERS_172-06_2015_ALK80311.1_ref
Bt_BM48-31_ADK66841.1
Bt_BtKY72_AP040579
BtYu-RmYN02_2019_S-gene_21544-25227_1
Bt_LYRa11_AHX37558
Bt_YN2018B_QDF43825
Bt_Rs4874_AT098205.1
Cv_007-2004_AAU04646
Hu_SARS-Frankfurt-1_2003_AAP33697.1_ref
Bt_rec-SARS_2008_ACJ60694.1_ref
Bt_ZC45_AVP78031_ref
Bt_ZXC21_AVP78042_ref
PnGu1_2019_S-gene_21541-25338_1
Pn_GX-P1E_2017_QIA48623_ref
Pn_GX-P2V_2018_QIQ54048
Bt_RaTG13_2013_Yunnan_QHR63300_ref
Hu_Cov2_WH01_2019_QHU36824_ref
Bt_JL2012_AIA62277.1_ref
Bt_YN2013_AIA62330
Bt_Rp-Shaanxi2011_AGC74165
Bt_SC2018_QDF43815
Bt_YNLF_31C_AKZ19076
Bt_Cp-Yun_2011_AGC74176
Bt_Rs_672-2006_ACU31032
Bt_Rml/2004_ABD75332
Bt_YN2018C_QDF43830
Bt_Rp3-2004_AAZ67052
Bt_GX2013_AIA62320
Bt_HKU3-12_ADE34812_ref

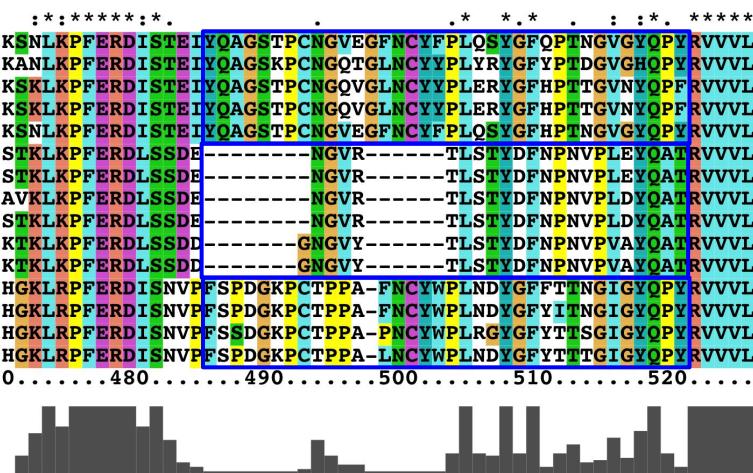
SLCALP	-DTPST-----	LTPRSVRSV	20
SLCALP	-DTPST-----	LTPRSVRSV	20
GICAKY	TNVSST-----	LVRSGGHSI	21
GICAKF	-GSDKI-----	RMQESI	18
GVCASY	-NSPAA-----	RVGTNSI	18
GICASY	-HTASL-----	LRNTDQKSI	20
GICASY	-HTVSS-----	LRSTSQKSI	20
GICASY	-HTVSS-----	LRSTSQKSI	20
GICASY	-HTVSS-----	LRSTSQKSI	20
GICASY	-HTVSL-----	LRSTSQKSI	20
GICASY	-HTVSL-----	LRSTSQKSI	20
GICASY	-HTASI-----	LRSTSQKAI	20
GICASY	-HTASI-----	LRSTGOKAI	20
GICASY	-QTQTN-----	SRSVSSQAI	20
GICASY	-HSMSS-----	LRSVNQRSI	20
GICASY	-HSMSS-----	FRSVNQRSI	20
GICASY	-QTQTN-----	SRSVASQSI	20
GICASY	-QTQTNSPRRA-----	SRSVASQSI	24
GICASY	-HTASL-----	LRSTGQKSI	20
GICASY	-HTAST-----	LRSIGQKSI	20
GICASY	-HTASV-----	LRSTGQKSI	20
GICASY	-HTAST-----	LRSTGQKSI	20
GICASY	-HTASV-----	LRSTGQKSI	20
GICASY	-HTASL-----	LRNTDQKSI	20
GICASY	-HTAST-----	LRSVGQKSI	20
GICASY	-HTASV-----	LRSTGQKSI	20
GICASY	-HTAST-----	LRSVGOKSI	20
GICASY	-HTAST-----	LRSVGOKSI	20
GICASY	-HTASV-----	LRSTGQKSI	20
GICASY	-HTASV-----	LRSTGQKSI	20
D790.....	800....	20



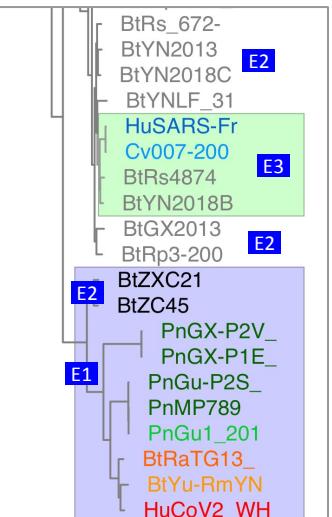
Un site recombinant?

- 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.

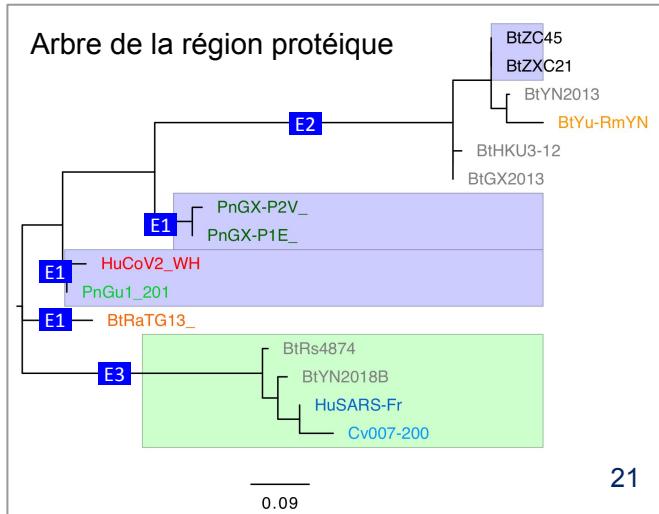
HuCoV2_WH01_2019
 BtRaTG13_2013_Yunnan
 PnGX_P1E_2017
 PnGX_P2V_2018
 PnGu1_2019
 BtZC45
 BtZXC21
 BtYu_RmYN02_2019
 BtYN2013
 BtHKU3_12
 BtGX2013
 BtYN2018B
 BtRs4874
 Cv007_2004
 HuSARS_Frankfurt_1_2003



Arbre des génomes

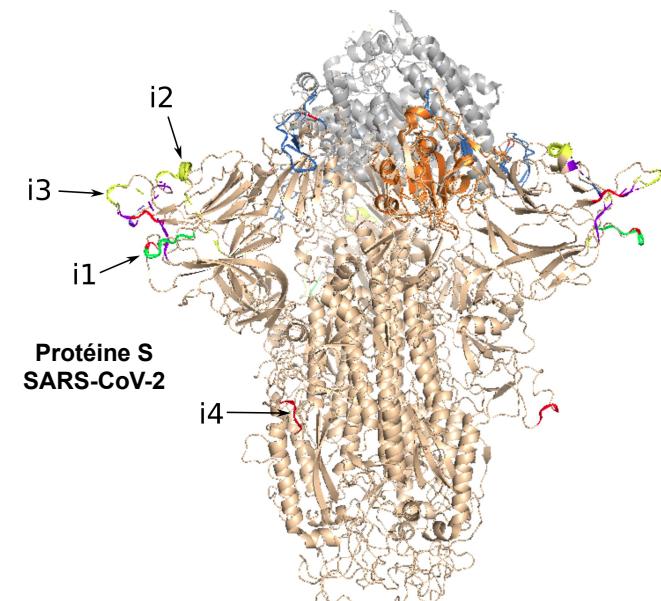
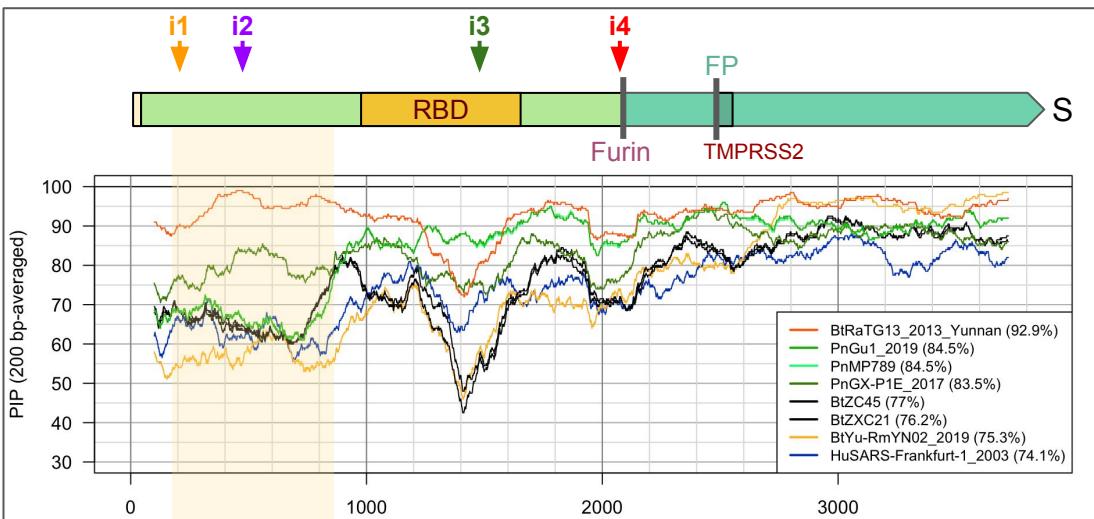


Arbre de la région protéique



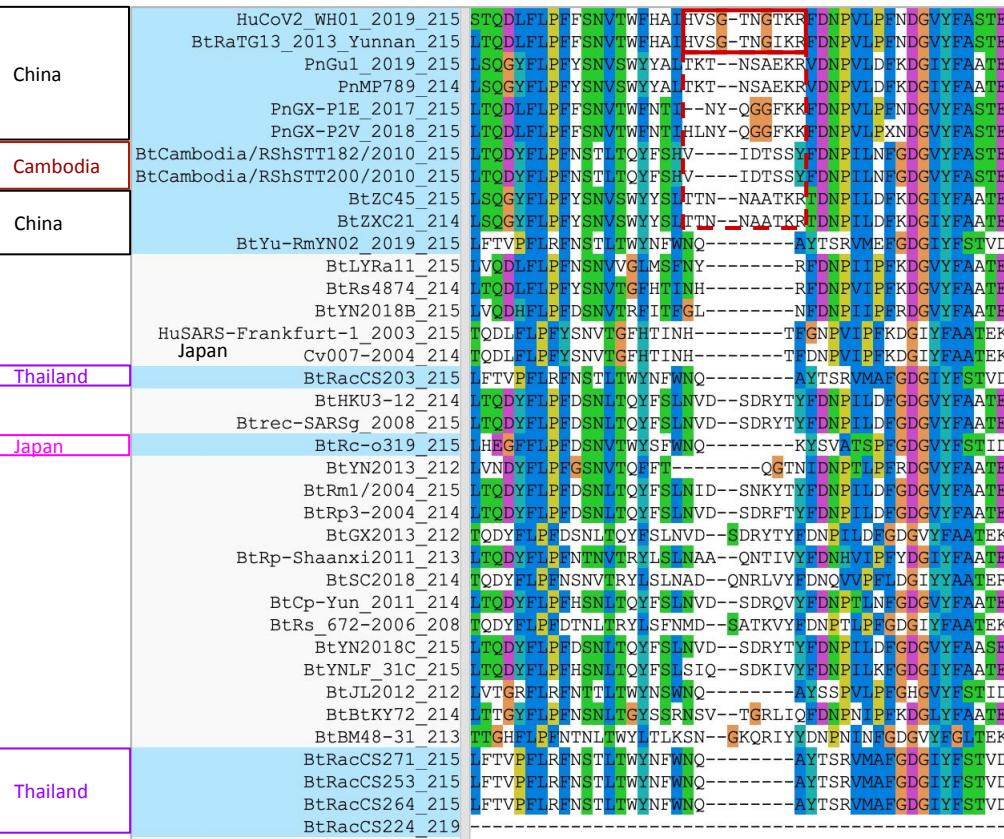
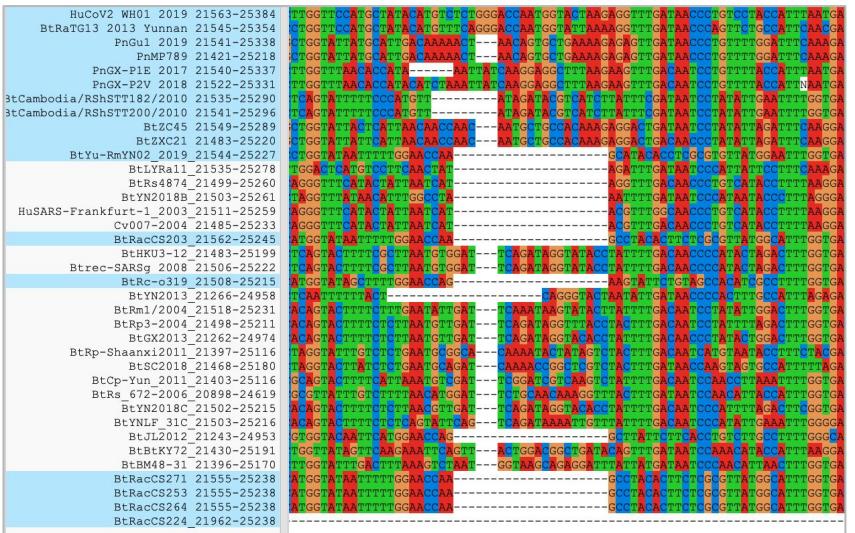
Four insertions in the S gene

- Arrows indicate the 4 insertions in the S gene (left) and on the spike protein (right)
- 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)



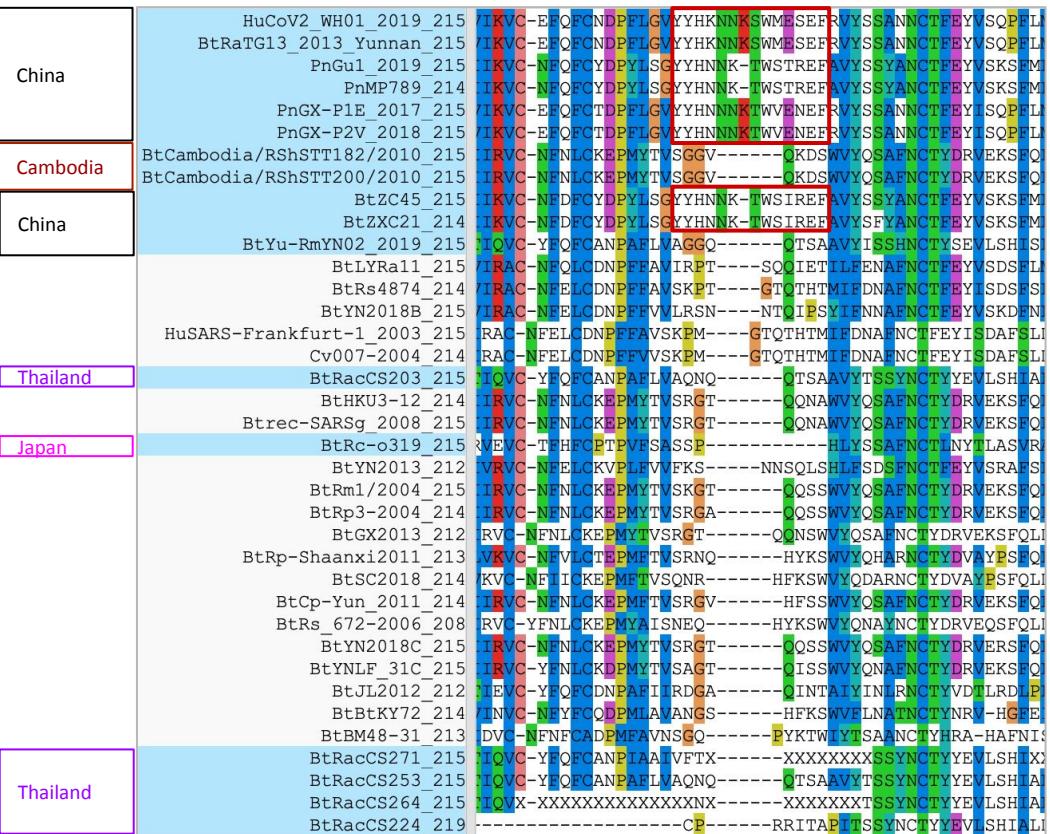
Insertion 1

- 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



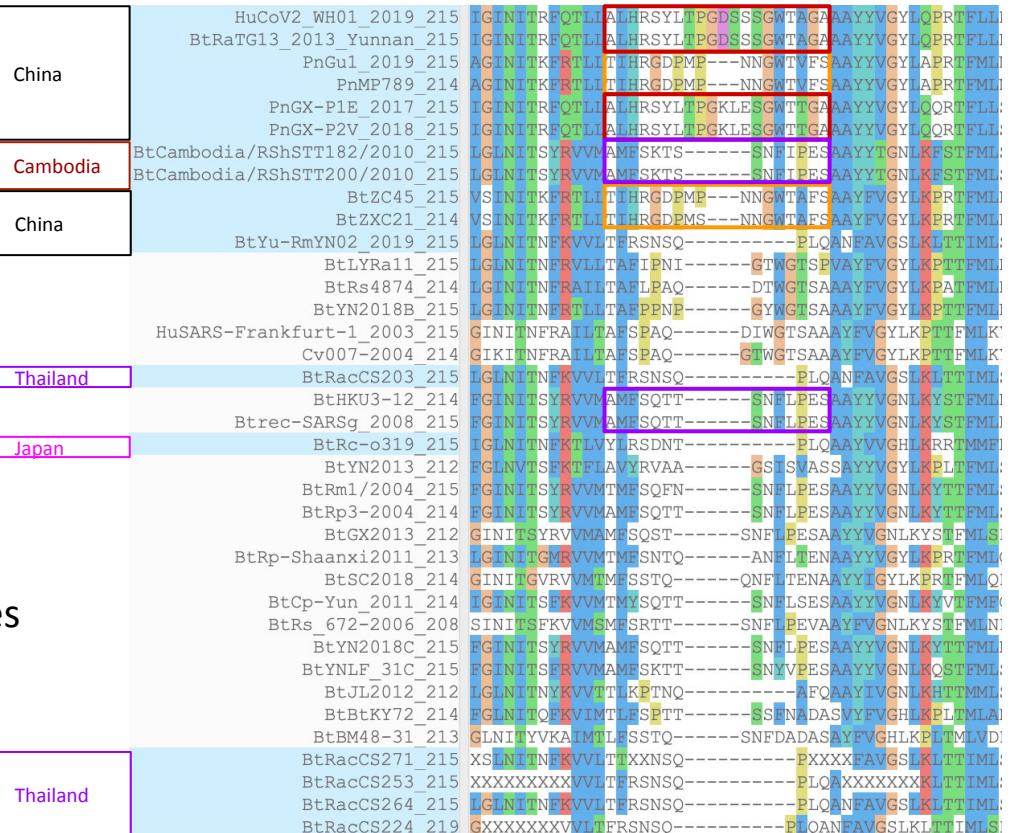
Insertion 2

- 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

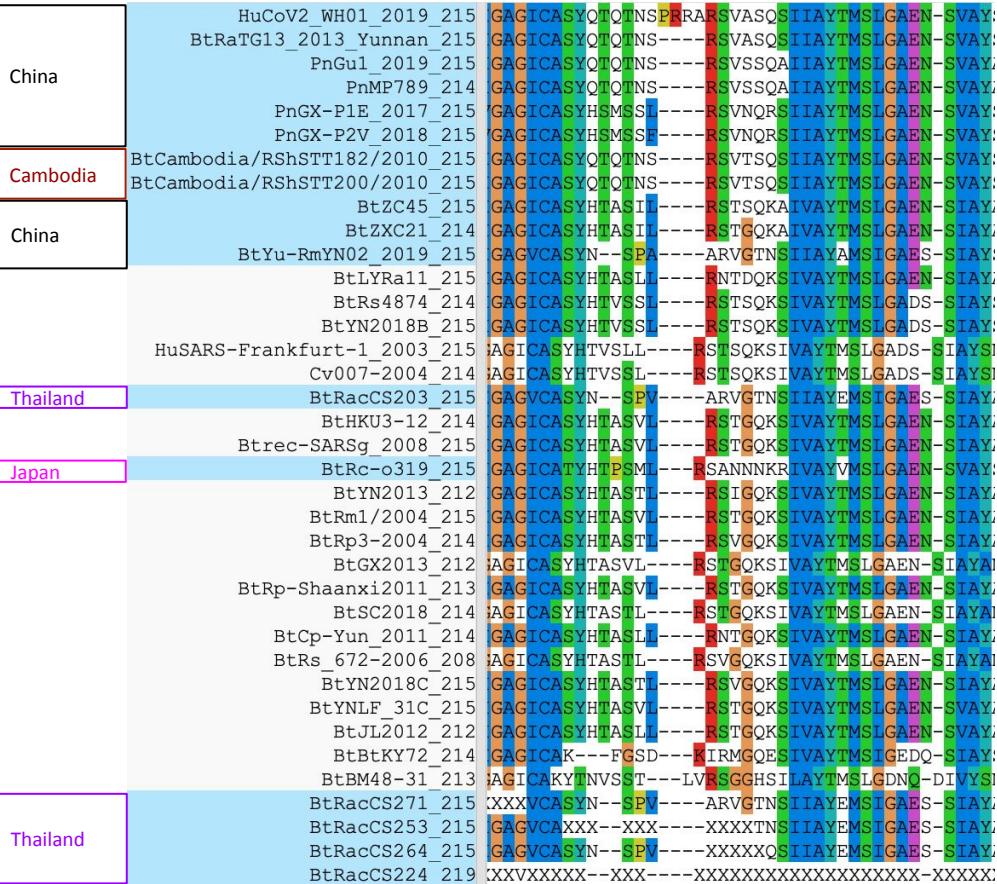
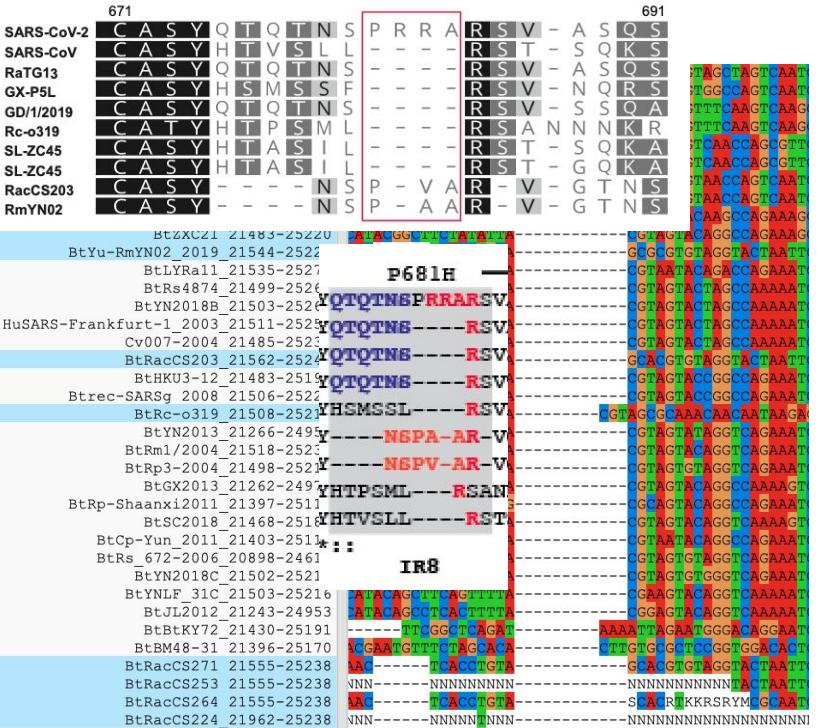
- Highly variable region
- Insertion shared between
 - SARS-CoV-2
 - RaTG13
 - GX Pangolins
- Another insertion shared between
 - GU Pangolins
 - BtZXC21 et BtZC45
- Note: insertion shared between
 - Cambodia (CoV2 group)
 - 2 bats from CoV1 group
- Yet different sequences in other Bat viruses of CoV2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

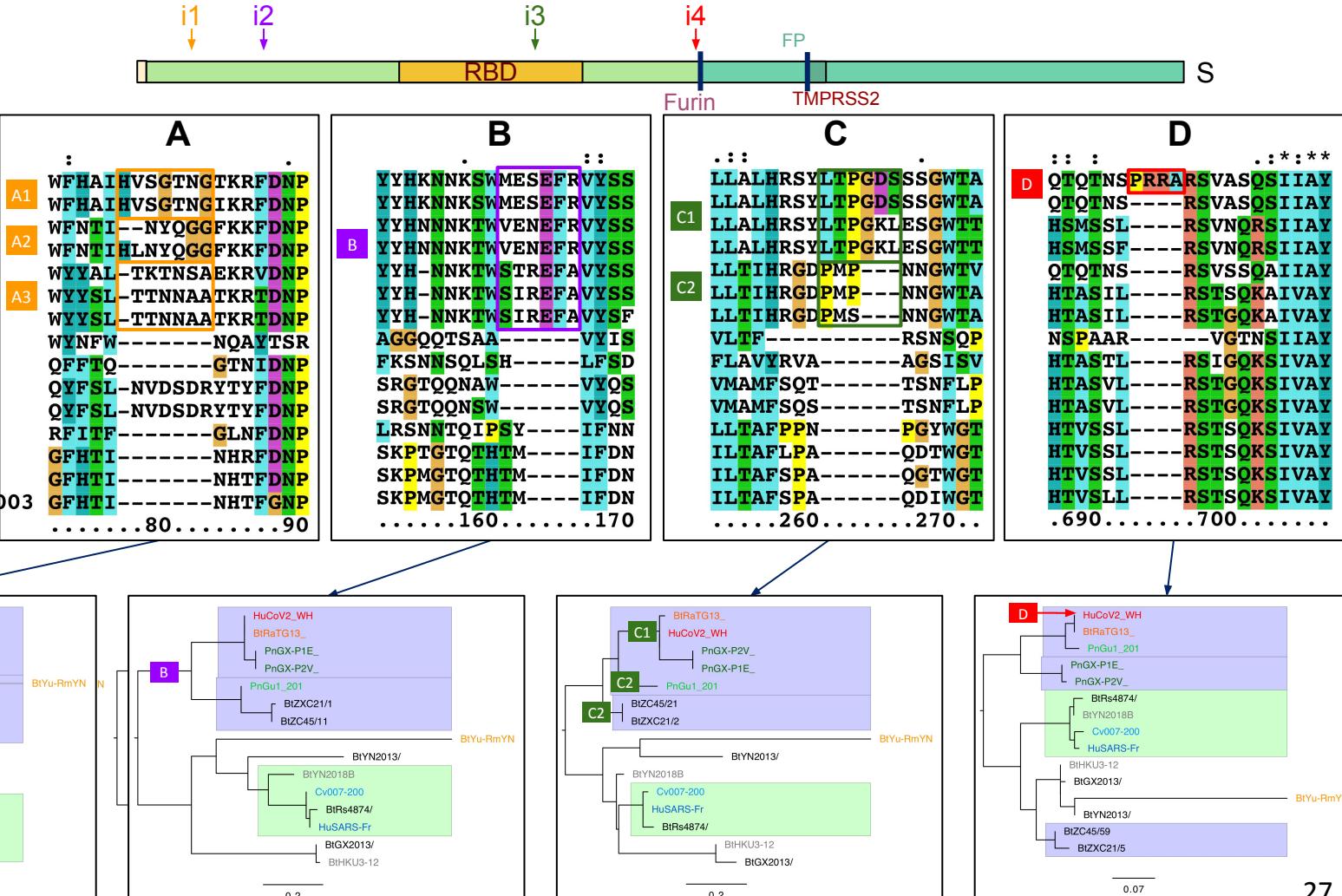


Insertion 4

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

Furin cleavage site





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

zenodo

Search Upload Communities

Log in Sign up

September 14, 2020

Working paper Open Access

776,634 views 597,172 downloads See more details...

Indexed in OpenAIRE

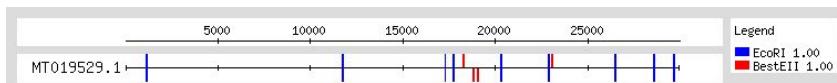
Publication date: September 14, 2020
DOI: DOI: 10.5281/zenodo.4028830
Communities: Coronavirus Disease Research Community - COVID-19
I license (for files):

Legend: EcoRI 1.00, BstEII 1.00

The figure shows a horizontal line representing the genome sequence MT019529.1. Above the line, a scale marks positions at 5000, 10000, 15000, 20000, and 25000. Below the line, vertical tick marks indicate cleavage sites. Blue tick marks represent EcoRI 1.00 sites, and red tick marks represent BstEII 1.00 sites. The pattern of cleavage sites is highly regular and symmetric, suggesting a synthetic origin.

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

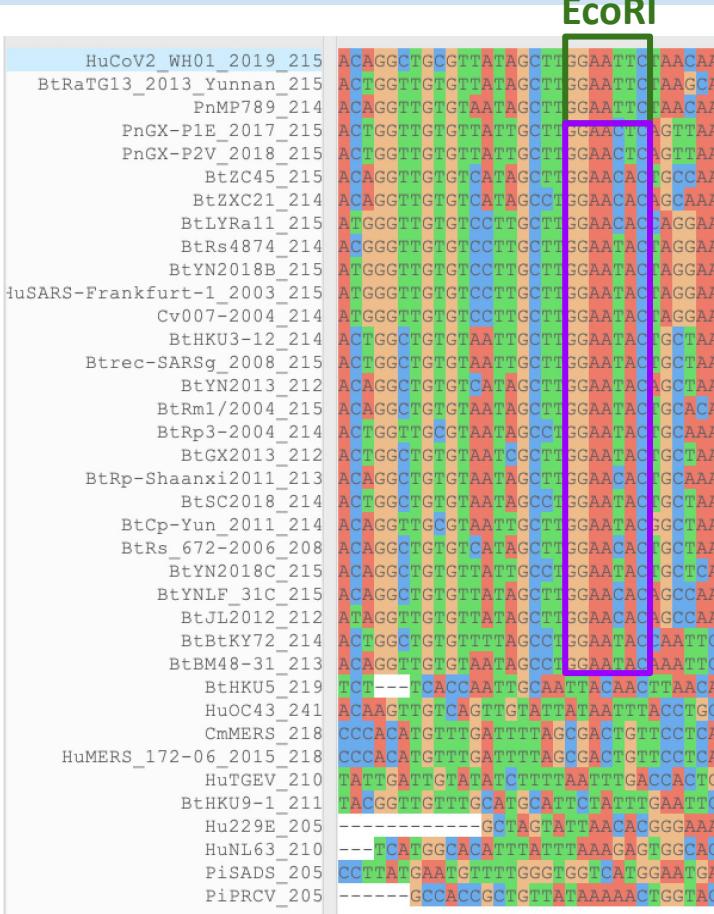


A SARS-CoV-2		EcoRI					
		W N S					
tataattata	aattaccaga	tgattttaca	ggctgcgtta	tagctt	gaa tt	taacaat	1320
cttgattcta	aggtttgtt	taattataat	tacctgtata	gatttttag	gaagtctaat		1380
ctcaaacctt	ttgagagaga	tatttcaact	gaaatctatc	aggccgttag	cacaccttgt		1440
aatggtgtt	aaggtttaa	ttgttacttt	ccttacaaat	catatggtt	ccaaccact		1500
aatggtgtt	g gt	tacccacc	atacagagta	gtagtactt	ctttgaact	tctacatgca	1560
		G Y Q					
B ZC45		BstEII					
		W N T					
ttacctgtat	atttacagg	ttgtgtata	gcttg	gaaca	ctgccaaaca	ggatgttaggt	1320
aattattttc	acaggctca	tcgttctacc	aaatttgaac	catttggaaag	actatctttcc		1380
tcagacgaga	atgggtccg	tacacttgc	actttagact	tcaaccctaa	tgttaccactt		1440
[gaatacc]	aaag	ctacaagggt	ttgttggtt	tcatttgagc	tcttaaatgc	accagctaca	1500
		E Y Q					

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

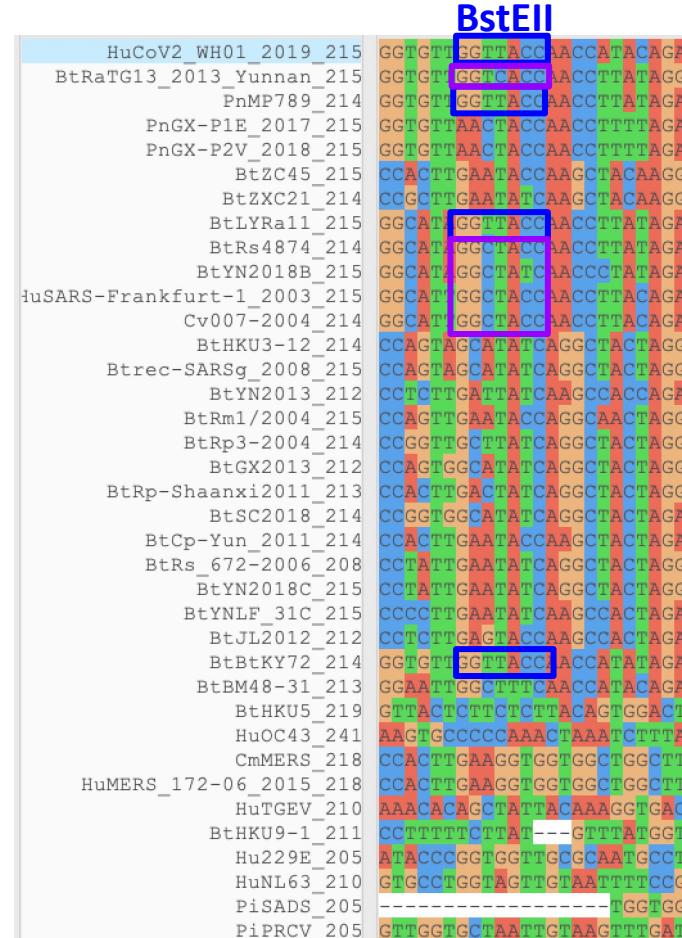
- Le site de restriction **EcoRI** (**séquence GAATTC**) se trouve dans le virus humain, mais également dans les virus les plus proches de chauve-souris ([RatG13](#)) et de pangolin ([MP789](#)).
- La majorité de ce site, ainsi que les séquences avoisinantes (**TTGGAAT*CT**) est conservée dans l'ensemble 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 GGTTTAC**) 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

