

# Alignements de séquences multiples

*Jacques van Helden*

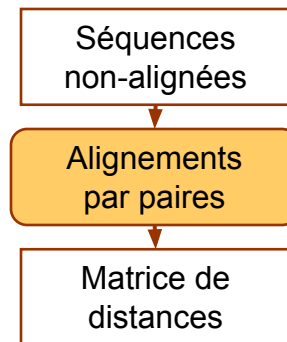
- L'approche la plus courante pour aligner des séquences multiples est de réaliser un **alignement progressif**.
- L'algorithme procède en plusieurs étapes (détaillées dans les diapos suivantes):
  - Calculer une **matrice de distances**, qui indique la distance entre chaque paire de séquences.
  - Construire 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.

- 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:  $d_{i,i} = 0$



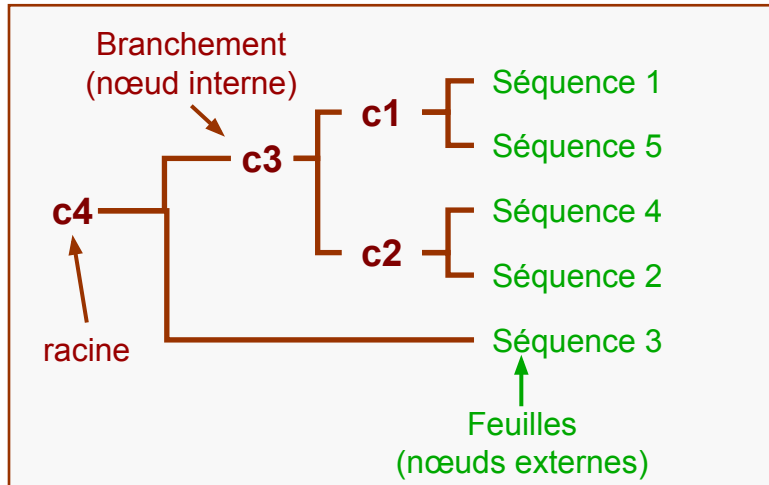
	seq 1	seq 2	...	seq n
seq 1	d1,1	d1,2	...	d1,n
seq 2	d2,1	d2,2	...	d2,n
...	...	...	...	...
seq n	dn,1	dn,2	...	dn,n

# 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

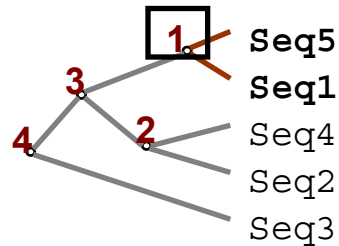


- Le clustering hiérarchique est une méthode de clustering agrégative.
  - Prend une matrice de distance en entrée
  - Regroupe progressivement les objets en allant des plus proches aux plus distants.
- Il existe plusieurs possibilités pour établir une règle d'agglomération, qui définit la distance entre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

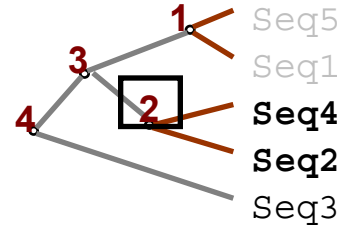


# Alignement progressif – 3<sup>ème</sup> étape: alignement multiple

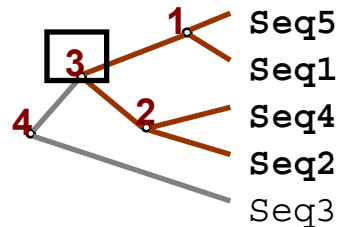
- On construit un alignement multiple en incorporant progressivement les séquences selon leur ordre de branchement dans l'arbre guide, en remontant des plus proches aux plus éloignées.



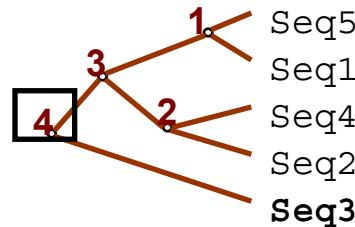
GATTGTAGTA  
GATGTAGTA



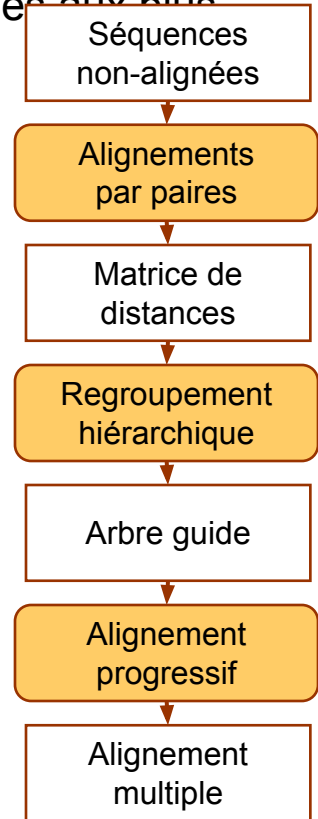
GATTGTAGTA  
GATGTAGTA  
GATTGTTC - - GTA  
GATTGTTCGGGTA



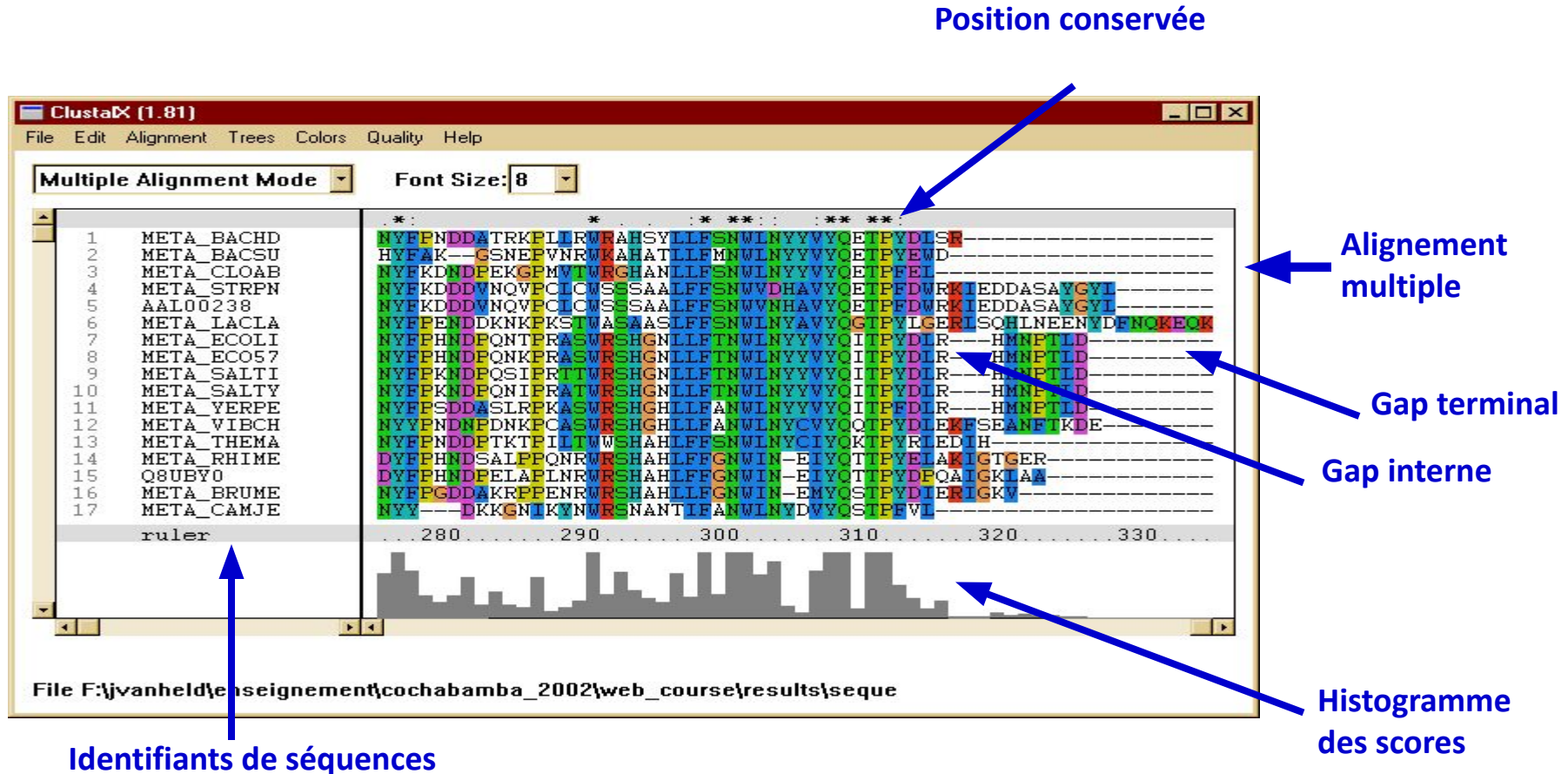
GATTGTA - - - GTA  
GATGGTA - - - GTA  
GATTGTTC - - GTA  
GATTGTTCGGGTA



GATTGTA - - - - GTA  
GATGGTA - - - - GTA  
GATTGTTC - - - - GTA  
GATTGTTCGG - - - GTA  
GATGGTAGGCGTGTA



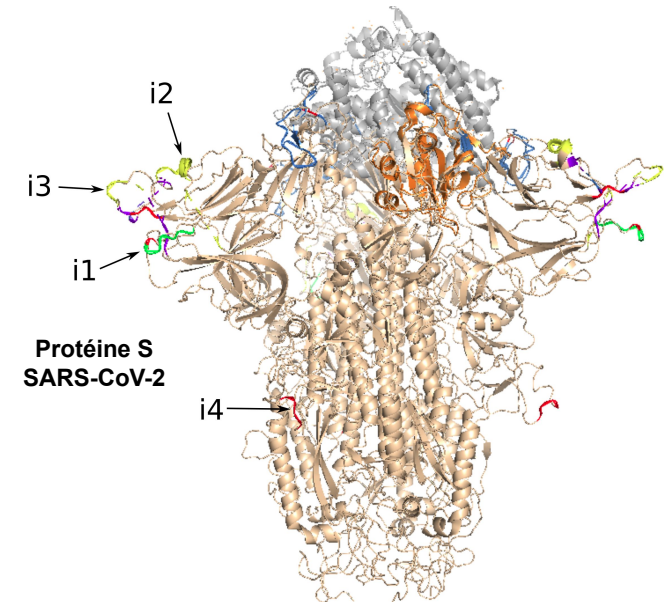
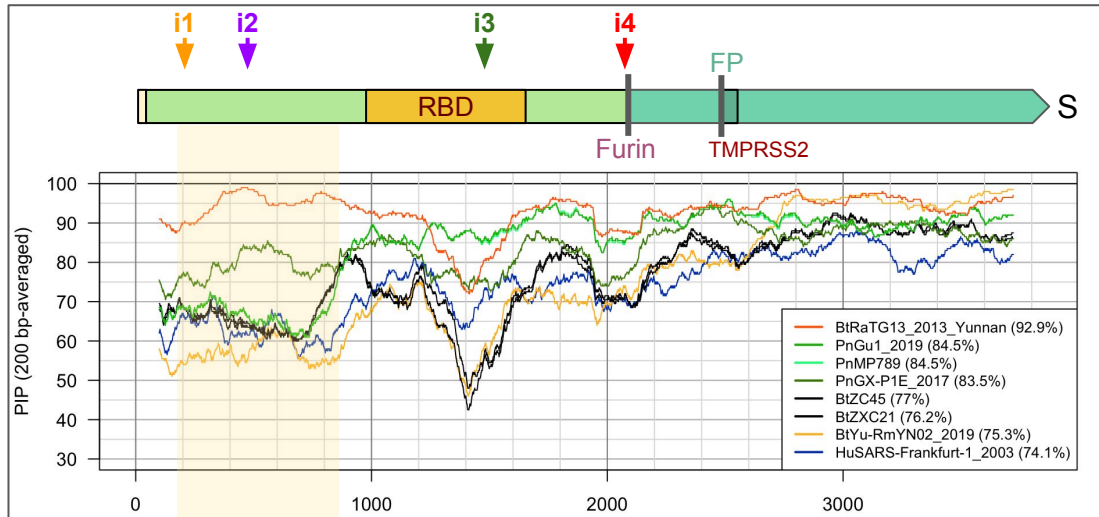
# Alignement multiple global : Homoserine-O-dehydrogenase



## *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 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

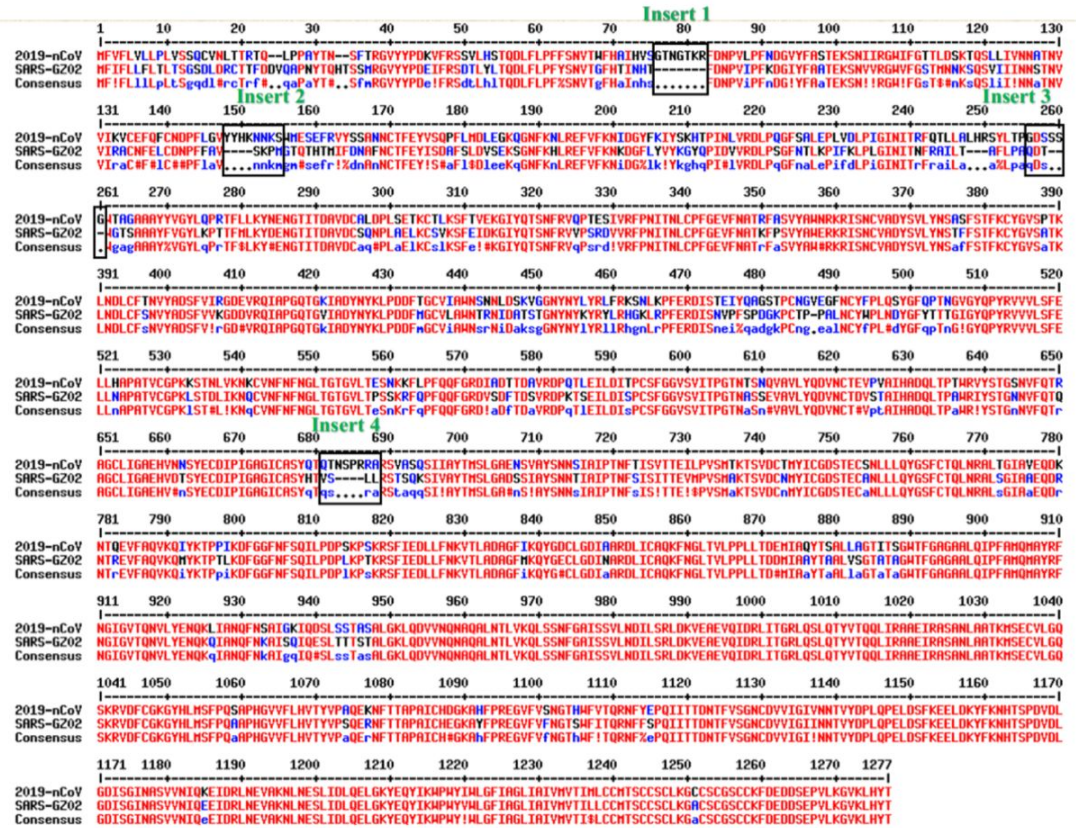
(<https://www.youtube.com/watch?v=qSWCLHIOiMo>).

*"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 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_APO40579
BtYu-RmYN02_2019_S-gene_21544-25227_1
Bt_LYRa11_AHX37558
Bt_YN2018B_QDF43825
Bt_Rs4874_ATO98205.1
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_Rm1/2004_ABD75332
Bt_YN2018C_QDF43830
Bt_Rp3-2004_AAZ67052
Bt_GX2013_AIA62320
Bt_HKU3-12_ADE34812_ref
    
```

```

SLCALP-DTPST----LTPRSVRSV 20
SLCALP-DTPST----LTPRSVRSV 20
GICAKYTNVSSST----LVRSGGHSI 21
GICAKF-GSDKI-----RMGOESI 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----LRSTGQKAI 20
GICASY-QTQTN----SRSVSSQAI 20
GICASY-HSMSS----LRSVNORSI 20
GICASY-HSMSS----FRSVNORSI 20
GICASY-QTQTN----SRSVASQSI 20
GICASY-QTQTNSPRRARSVASQSI 24
GICASY-HTASL----LRSTGQKSI 20
GICASY-HTAST----LRSIGQKSI 20
GICASY-HTASV----LRSTGQKSI 20
GICASY-HTAST----LRSTGQKSI 20
GICASY-HTASV----LRSTGQKSI 20
GICASY-HTASL----LRNTGQKSI 20
GICASY-HTAST----LRSVGQKSI 20
GICASY-HTASV----LRSTGQKSI 20
GICASY-HTAST----LRSVGQKSI 20
GICASY-HTAST----LRSVGQKSI 20
GICASY-HTASV----LRSTGQKSI 20
GICASY-HTASV----LRSTGQKSI 20
0.....790.....800....
    
```













## 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_APO40579
BtYu-RmYN02_2019_S-gene_21544-25227_1
Bt_LYRa11_AHX37558
Bt_YN2018B_QDF43825
Bt_Rs4874_ATO98205.1
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_Rm1/2004_ABD75332
Bt_YN2018C_QDF43830
Bt_Rp3-2004_AAZ67052
Bt_GX2013_AIA62320
Bt_HKU3-12_ADE34812_ref
    
```

```

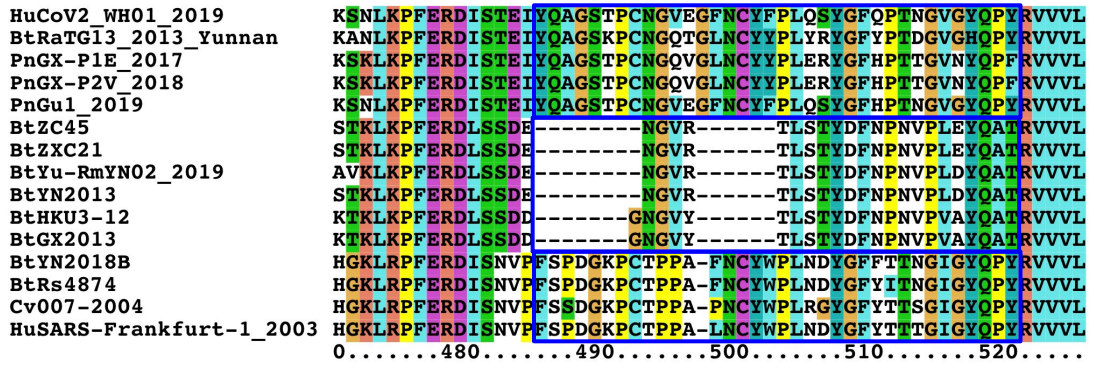
SLCALP-DTPST----LTPRSVRSV 20
SLCALP-DTPST----LTPRSVRSV 20
GICAKYTNVSSST----LVRSGGHSI 21
GICAKF-GSDKI-----RMGOESI 18
GVCASY-NSPAA-----RVGTNSI 18
GICASY-HTASL----LRNTDQKSI 20
GICASY-HTVSS----LRSTSQKSI 20
GICASY-HTVSS----LRSTSQKSI 20
GICASY-HTVSS----LRSTSQKSI 20
GICASY-HTVSL----LRSTSQKSI 20
GICASY-HTVSL----LRSTSQKSI 20
GICASY-HTASI----LRSTSQKAI 20
GICASY-HTASI----LRSTGQKAI 20
GICASY-QTQTN----SRSVSSQAI 20
GICASY-HSMSS----LRSVNORSI 20
GICASY-HSMSS----FRSVNORSI 20
GICASY-QTQTN----SRSVASQSI 20
GICASY-QTQTNSPRRARSVASQSI 24
GICASY-HTASL----LRSTGQKSI 20
GICASY-HTAST----LRSIGQKSI 20
GICASY-HTASV----LRSTGQKSI 20
GICASY-HTAST----LRSTGQKSI 20
GICASY-HTASV----LRSTGQKSI 20
GICASY-HTASL----LRNTGQKSI 20
GICASY-HTAST----LRSVGQKSI 20
GICASY-HTASV----LRSTGQKSI 20
GICASY-HTAST----LRSVGQKSI 20
GICASY-HTAST----LRSVGQKSI 20
GICASY-HTASV----LRSTGQKSI 20
GICASY-HTASV----LRSTGQKSI 20
0.....790.....800....
    
```





# Un site recombinant?

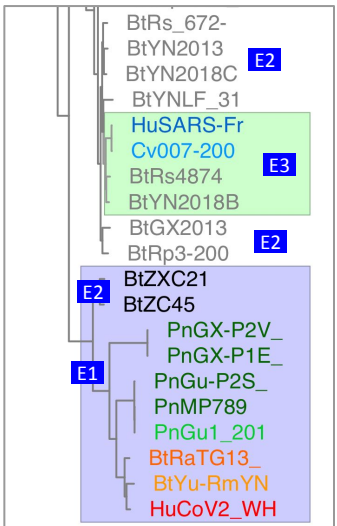
- L'interprétation de ce site est plus complexe.
- Il existe clairement trois groupes de séquences.
- Ceux-ci s'étendent au-delà des deux indels.
- L'arbre construit à partir de cette région est peu robuste, et incohérent avec celui des génomes.
- La répartition des sous-blocs de séquences est plus cohérente 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.



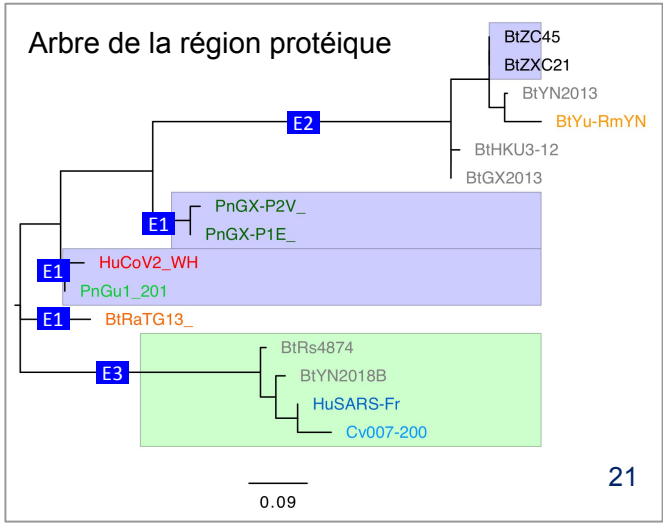
E1  
E2  
E3



Arbre des génomes

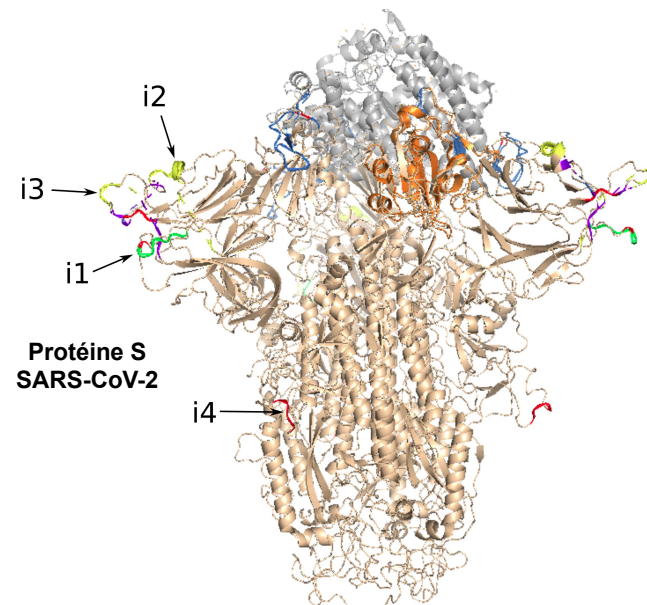
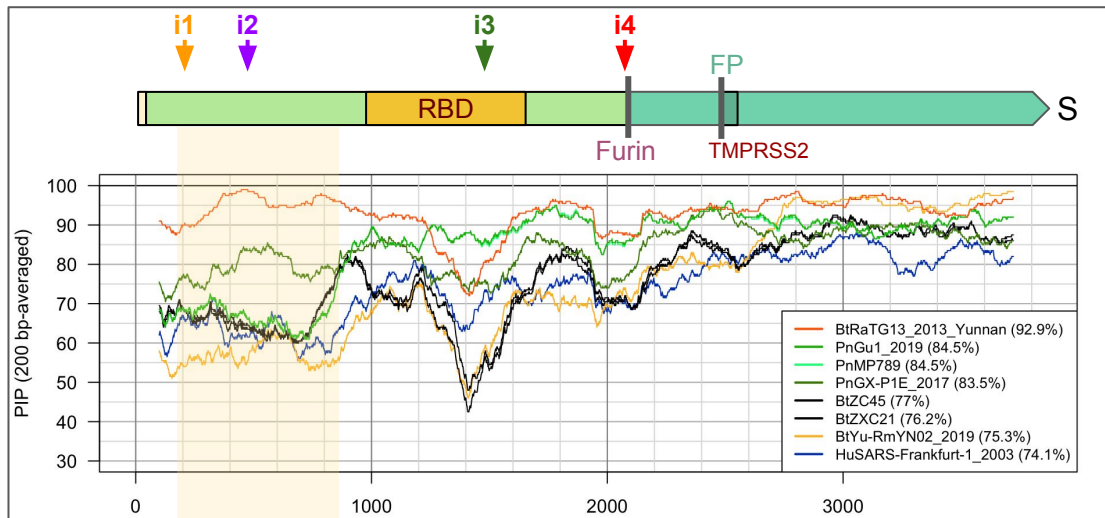


Arbre de la région protéique



# Four insertions in the S gene

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

HuCoV2 WH01_2019	21563-25384	ATGGGTCACAGGTAACATGCTCTGGGACCAATGGTACTAAGAGGTTTGATAACCGTGCTCCATTAAGT
BtRaTG13_2013_Yunnan	21545-25354	CTGGTTCATGCTATAAATGTTTCAGGGCCCAATGGTATTAAAGGTTTGATAACCGGTTCCCATTAAGT
PnGu1_2019	21541-25338	CTGGTATTATCCATCCCAAAAACCT---AAAGGGCTGAAAAGAGCGTGTGATAACCGGTTTGGATTCAAGA
PnMP789_21421-25218		CTGGTATTATCCATCCCAAAAACCT---AAAGGGCTGAAAAGAGCGTGTGATAACCGGTTTGGATTCAAGA
PnGX-P1E_2017	21540-25337	CTGGTATTATCCATCCCAAAAACCT---AAAGGGCTGAAAAGAGCGTGTGATAACCGGTTTGGATTCAAGA
PnGX-P2V_2018	21522-25331	CTGGTATTATCCATCCCAAAAACCT---AAAGGGCTGAAAAGAGCGTGTGATAACCGGTTTGGATTCAAGA
BtCambodia/RShST182/2010	21535-25290	TCAGTATTTTCCCATCTG---ATAGAGGTCATCTATTTGATAATCCCAATCCCAATTTGGGA
BtCambodia/RShST200/2010	21541-25289	TCAGTATTTTCCCATCTG---ATAGAGGTCATCTATTTGATAATCCCAATCCCAATTTGGGA
BtZC45_21549-25289		TCAGTATTTTCCCATCTG---ATAGAGGTCATCTATTTGATAATCCCAATCCCAATTTGGGA
BtZXC21_21483-25220		TCAGTATTTTCCCATCTG---ATAGAGGTCATCTATTTGATAATCCCAATCCCAATTTGGGA
BtYu-RmYN02_2019	21544-25227	CTGGTATAATTTTGGAAACCA---GCTATAGCGTCTGCGGTTTAGGATTGGGA
BtLYRa11_21535-25278		CTGGTATAATTTTGGAAACCA---GCTATAGCGTCTGCGGTTTAGGATTGGGA
BtRs4874_21499-25260		CTGGTATAATTTTGGAAACCA---GCTATAGCGTCTGCGGTTTAGGATTGGGA
BtYN2018B_21503-25261		CTGGTATAATTTTGGAAACCA---GCTATAGCGTCTGCGGTTTAGGATTGGGA
HuSARS-Frankfurt-1_2003	21511-25259	AGGGTTTACAGCTATAAATCA---AGGTTTACAGCTATAAATCA---AGGTTTACAGCTATAAATCA
Cv007-2004	21485-25233	AGGGTTTACAGCTATAAATCA---AGGTTTACAGCTATAAATCA---AGGTTTACAGCTATAAATCA
BtRacCS203_21562-25245		AGGGTTTACAGCTATAAATCA---AGGTTTACAGCTATAAATCA---AGGTTTACAGCTATAAATCA
BtHKU3-12_21483-25199		AGGGTTTACAGCTATAAATCA---AGGTTTACAGCTATAAATCA---AGGTTTACAGCTATAAATCA
Btrec-SARSg_2008	21506-25222	TCAGTATTTTCCCATCTG---ATAGAGGTCATCTATTTGATAATCCCAATCCCAATTTGGGA
BtRC-0319_21508-25215		TCAGTATTTTCCCATCTG---ATAGAGGTCATCTATTTGATAATCCCAATCCCAATTTGGGA
BtYN2013_21266-24958		TCAGTATTTTCCCATCTG---ATAGAGGTCATCTATTTGATAATCCCAATCCCAATTTGGGA
BtRm1/2004_21518-25231		TCAGTATTTTCCCATCTG---ATAGAGGTCATCTATTTGATAATCCCAATCCCAATTTGGGA
BtRp3-2004_21498-25211		TCAGTATTTTCCCATCTG---ATAGAGGTCATCTATTTGATAATCCCAATCCCAATTTGGGA
BtGX2013_21262-24974		TCAGTATTTTCCCATCTG---ATAGAGGTCATCTATTTGATAATCCCAATCCCAATTTGGGA
BtRp-Shaanxi2011_21397-25116		TAGGACTTTTGCCTGAAGGCGCA---AAAATAGTATAGTCTATTGGCAATCCCAATCCCAATTTGGGA
BtSC2018_21468-25180		TAGGACTTTTGCCTGAAGGCGCA---AAAATAGTATAGTCTATTGGCAATCCCAATCCCAATTTGGGA
BtCp-Yun_2011_21403-25116		TAGGACTTTTGCCTGAAGGCGCA---AAAATAGTATAGTCTATTGGCAATCCCAATCCCAATTTGGGA
BtRs_672-2006_20898-24619		TAGGACTTTTGCCTGAAGGCGCA---AAAATAGTATAGTCTATTGGCAATCCCAATCCCAATTTGGGA
BtYN2018C_21502-25215		TAGGACTTTTGCCTGAAGGCGCA---AAAATAGTATAGTCTATTGGCAATCCCAATCCCAATTTGGGA
BtYNLF_31C_21503-25216		TAGGACTTTTGCCTGAAGGCGCA---AAAATAGTATAGTCTATTGGCAATCCCAATCCCAATTTGGGA
BtJL2012_21243-24953		CTGGTATAATTTTGGAAACCA---GCTATAGCGTCTGCGGTTTAGGATTGGGA
BtBLK72_21430-25191		CTGGTATAATTTTGGAAACCA---GCTATAGCGTCTGCGGTTTAGGATTGGGA
BtBM48-31_21396-25170		CTGGTATAATTTTGGAAACCA---GCTATAGCGTCTGCGGTTTAGGATTGGGA
BtRacCS271_21555-25238		CTGGTATAATTTTGGAAACCA---GCTATAGCGTCTGCGGTTTAGGATTGGGA
BtRacCS253_21555-25238		CTGGTATAATTTTGGAAACCA---GCTATAGCGTCTGCGGTTTAGGATTGGGA
BtRacCS264_21555-25238		CTGGTATAATTTTGGAAACCA---GCTATAGCGTCTGCGGTTTAGGATTGGGA
BtRacCS224_21962-25238		CTGGTATAATTTTGGAAACCA---GCTATAGCGTCTGCGGTTTAGGATTGGGA

China	HuCoV2_WH01_2019_215	STQDLFLPFFFSNVITWHAHIVHSG-TNGTKRFDNPVLPFDGVIYFASTE
China	BtRaTG13_2013_Yunnan_215	LTQDLFLPFFFSNVITWHAHIVHSG-TNGIKRFDNPVLPFDGVIYFASTE
China	PnGu1_2019_215	LSQGYFLPFFYSNVSWYYALTKT--NSAEKRVNDNPVLPFDKGGIYFAATE
China	PnMP789_214	LSQGYFLPFFYSNVSWYYALTKT--NSAEKRVNDNPVLPFDKGGIYFAATE
China	PnGX-P1E_2017_215	LTQDLFLPFFFSNVITWNTLHLY--NY-QGGFKKFDNPVLPFDGVIYFASTE
China	PnGX-P2V_2018_215	LTQDLFLPFFFSNVITWNTLHLY--NY-QGGFKKFDNPVLPFDGVIYFASTE
Cambodia	BtCambodia/RShST182/2010_215	LTQDYFLPFDNSLTQYFSLNV---IDTSSYFDNPILDFGDIYFAATE
Cambodia	BtCambodia/RShST200/2010_215	LTQDYFLPFDNSLTQYFSLNV---IDTSSYFDNPILDFGDIYFAATE
China	BtZC45_215	LSQGYFLPFFYSNVSWYYSLTTN--NAATKRFDNPILDFKGGIYFAATE
China	BtZXC21_214	LSQGYFLPFFYSNVSWYYSLTTN--NAATKRFDNPILDFKGGIYFAATE
China	BtYu-RmYN02_2019_215	LFTVFFLRFNSTLTWYFNWQ-----AYTSRVMEFGDGIYFSTVD
China	BtLYRa11_215	LVDLFLPFDNSLVGLMSPNY-----RFDNPIPFKGGVYFAATE
China	BtRs4874_214	LTQDLFLPFFYSNVITGFEHTINH-----RFDNPVLPFDKGGVYFAATE
China	BtYN2018B_215	LVDHFLPFDNSVTRFTTFL-----NFDNPIPFKGGVYFAATE
China	HuSARS-Frankfurt-1_2003_215	TQDLFLPFFYSNVITGFHTINH-----TGNVPVLPFDKGGIYFAATEK
China	Japan Cv007-2004_214	TQDLFLPFFYSNVITGFHTINH-----TGNVPVLPFDKGGIYFAATEK
Thailand	BtRacCS203_215	LFTVFFLRFNSTLTWYFNWQ-----AYTSRVMEFGDGIYFSTVD
Thailand	BtHKU3-12_214	LTQDYFLPFDNSLTQYFSLNV---SDRYTYFDNPILDFGDIYFAATE
Thailand	Btrec-SARSg_2008_215	LTQDYFLPFDNSLTQYFSLNV---SDRYTYFDNPILDFGDIYFAATE
Japan	BtRc-0319_215	LHEGFFLPFDNSVITWYFVWQ-----KYSVATSPFGDGIYFSTID
Japan	BtYN2013_212	LVNDYFLPFDNSVITQFFI-----CGTNTFDNPILPFKGGVYFAATE
Japan	BtRm1/2004_215	LTQDYFLPFDNSLTQYFSLNID--SNKYTYFDNPILDFGDIYFAATE
Japan	BtRp3-2004_214	LTQDYFLPFDNSLTQYFSLNV---SDRFTYFDNPILDFGDIYFAATE
Japan	BtGX2013_212	TQDYFLPFDNSLTQYFSLNV---SDRYTYFDNPILDFGDIYFAATEK
Japan	BtRp-Shaanxi2011_213	LTQDYFLPFDNSVITWYFSLNLA---QNTIVYFDNHVLPFDGIYFAATE
Japan	BtSC2018_214	TQDYFLPFDNSVITWYFSLNAD--QNRVLYFDNPVLPFDGIYFAATE
Japan	BtCp-Yun_2011_214	LTQDYFLPFDNSLTQYFSLNV---SDRQVYFDNPILDFGDIYFAATE
Japan	BtRs_672-2006_208	TQDYFLPFDNSLTQYFSLNMD--SATKVYFDNPVLPFDGIYFAATEK
Japan	BtYN2018C_215	LTQDYFLPFDNSLTQYFSLNV---SDRYTYFDNPILDFGDIYFAATE
Japan	BtYNLF_31C_215	LTQDYFLPFDNSLTQYFSLISIQ--SDKIVYFDNPILDFGDIYFAATE
Japan	BtJL2012_212	LVTGRFLRFNSTLTWYFNWQ-----AYSSVLPFDGDIYFSTID
Japan	BtBK72_214	LTTGYFLPFDNSVITWYFVWQ---TGRLIHDNPIPFKGGVYFAATE
Japan	BtBM48-31_213	TTGHFLPFDNSLTWYTLKSN--GKQRIYFDNPVLPFDGVIYFLTEK
Thailand	BtRacCS271_215	LFTVFFLRFNSTLTWYFNWQ-----AYTSRVMEFGDGIYFSTVD
Thailand	BtRacCS253_215	LFTVFFLRFNSTLTWYFNWQ-----AYTSRVMEFGDGIYFSTVD
Thailand	BtRacCS264_215	LFTVFFLRFNSTLTWYFNWQ-----AYTSRVMEFGDGIYFSTVD
Thailand	BtRacCS224_219	-----



# 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

China	HuCoV2_WH01_2019_215	YIKVC-EFQFCNDPFLGVYYHKNKSWMESEFRVYSSANNCTFEYISQPFLL
	BtRaTG13_2013_Yunnan_215	YIKVC-EFQFCNDPFLGVYYHKNKSWMESEFRVYSSANNCTFEYISQPFLL
	PnGu1_2019_215	YIKVC-NFQFCYDPYISGYHNNK-TWSTREFAVYSSYANCTFEYISKSFEM
	PnMP789_214	YIKVC-NFQFCYDPYISGYHNNK-TWSTREFAVYSSYANCTFEYISKSFEM
	PnGX-P1E_2017_215	YIKVC-EFQFCTDPFLGVYYHNNK-TWVNEFRVYSSANNCTFEYISQPFLL
Cambodia	PnGX-P2V_2018_215	YIKVC-EFQFCTDPFLGVYYHNNK-TWVNEFRVYSSANNCTFEYISQPFLL
	BtCambodia/RShST182/2010_215	YIKVC-NFNLCKEPMYTVSGGV-----QKDSWVYQSAFNCTYDRVEKSFQ
China	BtCambodia/RShSTT200/2010_215	YIKVC-NFNLCKEPMYTVSGGV-----QKDSWVYQSAFNCTYDRVEKSFQ
	BtZC45_215	YIKVC-NFDFCYDPYISGYHNNK-TWSIREFVYSSYANCTFEYISKSFEM
	BtZXC21_214	YIKVC-NFDFCYDPYISGYHNNK-TWSIREFVYSSYANCTFEYISKSFEM
	BtYu-RmYN02_2019_215	YIQVC-YFQFCANPAFLVAGGQ-----QTSAAVYTSSSHNCTYSEVLSHSIS
	BtLYRa11_215	YIRAC-NFQLCDNPFPAVIRPT----SQIETILLFENAFNCTFEYISDSFSI
Thailand	BtRs4874_214	YIRAC-NFELCDNPFPAVSKPT----GTQHTMIFDNAFNCTFEYISDSFSI
	BtYN2018B_215	YIRAC-NFELCDNPFVVLRSN----NTQIISYIFNNAFNCTFEYISKDFNI
	HuSARS-Frankfurt-1_2003_215	YIRAC-NFELCDNPFPAVSKPM----GTQHTMIFDNAFNCTFEYISDAFSLI
	Cv007-2004_214	YIRAC-NFELCDNPFVVSXPM----GTQHTMIFDNAFNCTFEYISDAFSLI
	BtRacCS203_215	YIQVC-YFQFCANPAFLVAQNQ-----QTSAAVYTSSSHNCTYSEVLSHSIA
	BtHKU3-12_214	YIRVC-NFNLCKEPMYTVSRGT-----QONAWVYQSAFNCTYDRVEKSFQ
	Btrec-SARSg_2008_215	YIRVC-NFNLCKEPMYTVSRGT-----QONAWVYQSAFNCTYDRVEKSFQ
	BtRc-o319_215	YIEVC-TFHFCETPVVSASSP-----HLYSSAFNCTLYNLLASVR
	BtYN2013_212	YIRVC-NFELCKVPLFVVFKS-----NNSQLSHLFSDSFNCTYSEVLSHSIA
	BtRm1/2004_215	YIRVC-NFNLCKEPMYTVSKGT-----QOSSWVYQSAFNCTYDRVEKSFQ
Japan	BtRp3-2004_214	YIRVC-NFNLCKEPMYTVSRGA-----QOSSWVYQSAFNCTYDRVEKSFQ
	BtGX2013_212	YIRVC-NFNLCKEPMYTVSRGT-----QONSWYQSAFNCTYDRVEKSFQLI
	BtRp-Shaanxi2011_213	YIKVC-NFVLGTEPMFTVSRNQ-----HYKSWVYQHARNCTYDVVYPSFQ
	BtSC2018_214	YIKVC-NFTICKEPMFTVSONR-----HFKSWYQDARNCTYDVVAVPSFQLI
	BtCp-Yun_2011_214	YIRVC-NFNLCKEPMFTVSRGV-----HFSSWVYQSAFNCTYDRVEKSFQ
	BtRs_672-2006_208	YIRVC-YFNLCKEPMYVVISNEQ-----HYKSWVYQONAYNCTYDRVEQSFQLI
	BtYN2018C_215	YIRVC-NFNLCKEPMYTVSRGT-----QOSSWVYQSAFNCTYDRVEKSFQ
	BtYNLF_31C_215	YIRVC-YFNLCKEPMYTVSAGT-----QISSWVYQONAYNCTYDRVEKSFQ
	BtJL2012_212	YIEVC-YFQFCNDPAFIRLDGA-----QINTAIYINLRNCTYVDLRLDLE
	BtBtKY72_214	YINVC-NFYFCQDPMIAVANGS-----HFKSWVFLNANFNCTYVNRV-HGEIS
	BtBM48-31_213	YDVC-NFNFCADPMFAVNSGQ-----PYKTWLYYSAANCTYVHRA-HAFNIS
	BtRacCS271_215	YIQVC-YFQFCANPIAATVFTX-----XXXXXXXXXXSYNCTYSEVLSHSIXX
	BtRacCS253_215	YIQVC-YFQFCANPAFLVAQNQ-----QTSAAVYTSSSHNCTYSEVLSHSIA
	BtRacCS264_215	YIQVX-XXXXXXXXXXXXXXXXNX-----XXXXXXXXXXSYNCTYSEVLSHSIA
	BtRacCS224_219	-----CP-----RRITAPLTSYNCTYSEVLSHSIA

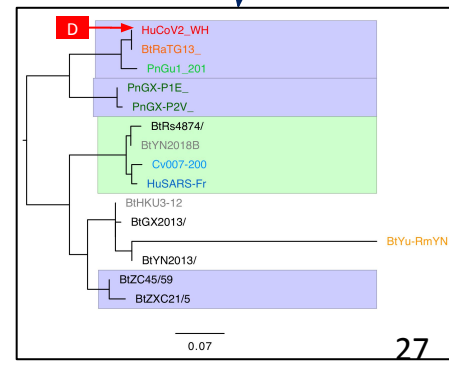
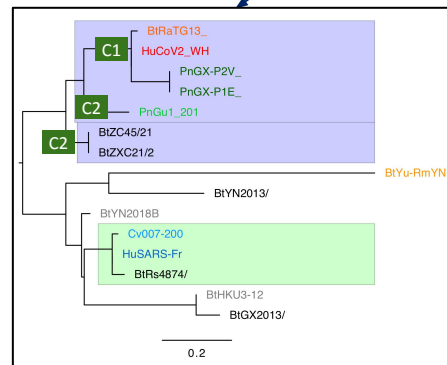
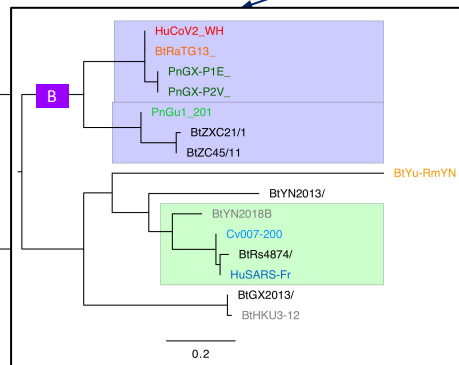
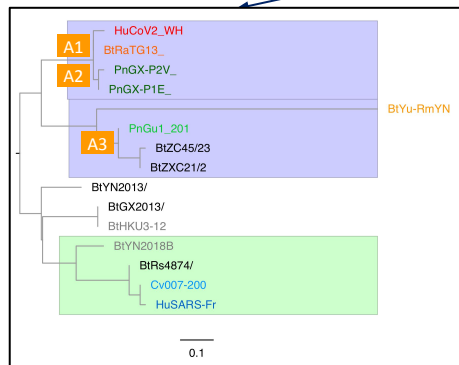
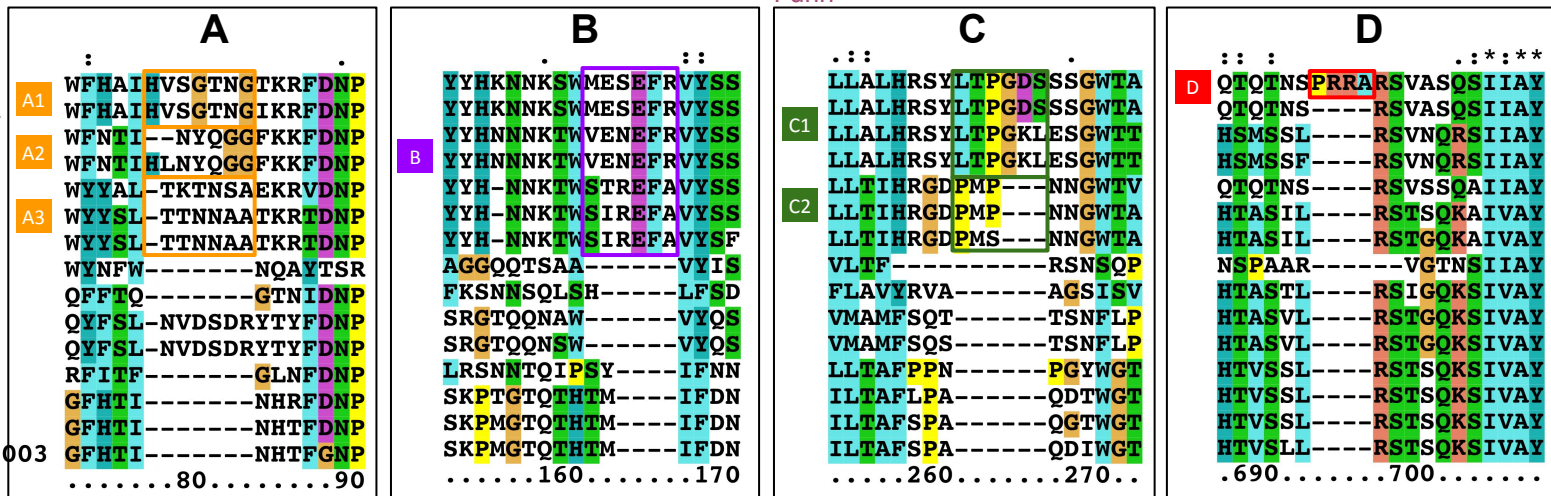
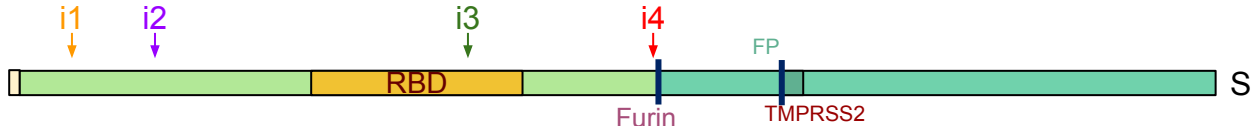
# Insertion 3

- 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\*)
- 

China	HuCoV2_WH01_2019_215	IGINITRFOTLIAALHRSYLTIPGDSSSGWTAGAAAYYVGYLQPRTEFLLI
	BtRaTG13_2013_Yunnan_215	IGINITRFOTLIAALHRSYLTIPGDSSSGWTAGAAAYYVGYLQPRTEFLLI
	PnGu1_2019_215	AGINITKFRLLITIHRCDEMP---NNGWTAFSAAYYVGYLAPRTFMLJ
	PnMP789_214	AGINITKFRLLITIHRCDEMP---NNGWTAFSAAYYVGYLAPRTFMLJ
	PnGX-PIE_2017_215	IGINITRFOTLIAALHRSYLTIPGKLESGWTTGAAAYYVGYLQQRTEFLLI
Cambodia	PnGX-P2V_2018_215	IGINITRFOTLIAALHRSYLTIPGKLESGWTTGAAAYYVGYLQQRTEFLLI
	BtCambodia/RShSTT182/2010_215	LGINITSYRVVMAMFSKTS-----SNFLPESAAAYYVGNLKYSTFMLJ
China	BtCambodia/RShSTT200/2010_215	LGINITSYRVVMAMFSKTS-----SNFLPESAAAYYVGNLKYSTFMLJ
	BtZC45_215	VSINITKFRLLITIHRCDEMP---NNGWTAFSAAYYVGYLKPRTFMLJ
	BtZXC21_214	VSINITKFRLLITIHRCDEMP---NNGWTAFSAAYYVGYLKPRTFMLJ
	BtYu-RmYN02_2019_215	LGINITNFKVVIITFRSNSQ-----PLOANFAVGSCLKLTTIMLJ
	BtLYRa11_215	LGINITNFRVLLTAFIPNI-----GTWGTSPVAYYVGYLKPRTFMLJ
	BtRs4874_214	LGINITNFRVLLTAFIPNI-----GTWGTSPVAYYVGYLKPRTFMLJ
	BtYn2018B_215	LGINITNFRVLLTAFIPNI-----GTWGTSPVAYYVGYLKPRTFMLJ
	BtYn2018B_215	LGINITNFRVLLTAFIPNI-----GTWGTSPVAYYVGYLKPRTFMLJ
	HuSARS-Frankfurt-1_2003_215	GININFRALLTAFSPAQ-----DIWGTSAAYYVGYLKPRTFMIK'
	Cv007-2004_214	GIKINFRALLTAFSPAQ-----DIWGTSAAYYVGYLKPRTFMIK'
Thailand	BtRacCS203_215	LGINITNFKVVIITFRSNSQ-----PLOANFAVGSCLKLTTIMLJ
Japan	BtHKU3-12_214	FGINITSYRVVMAMFSQTT-----SNFLPESAAAYYVGNLKYSTFMLJ
	Btrec-SARsg_2008_215	FGINITSYRVVMAMFSQTT-----SNFLPESAAAYYVGNLKYSTFMLJ
	BtRc-o319_215	IGLINITNFKTLVYLRSDNT-----PLOAAYVGHKRRITMFMJ
	BtYN2013_212	GLNIVTSFKTFLAVYRVAA-----GSTVASSAYYVGYLKPRTFMLJ
	BtRm1/2004_215	FGINITSYRVVMAMFSQFN-----SNFLPESAAAYYVGNLKYSTFMLJ
	BtRp3-2004_214	FGINITSYRVVMAMFSQTT-----SNFLPESAAAYYVGNLKYSTFMLJ
	BtGX2013_212	GINITSYRVVMAMFSQST-----SNFLPESAAAYYVGNLKYSTFMLJ
	BtRp-Shaanxi2011_213	LGINITGMRVVMAMFSQNTQ-----ANFLTENAAAYYVGYLKPRTFMLJ
	BtSC2018_214	GINITGVRVVMAMFSSTQ-----QNFLTENAAYYVGYLKPRTFMLJ
	BtCp-Yun_2011_214	IGINITSEKVVMTMYSQTT-----SNFLSESAAAYYVGNLKYSTFMLJ
Thailand	BtRs_672-2006_208	SINITSEKVVMTMYSQTT-----SNFLPEVAAYYVGNLKYSTFMLJ
	BtYN2018C_215	FGINITSYRVVMAMFSQTT-----SNFLPESAAAYYVGNLKYSTFMLJ
	BtYNLF_31C_215	FGINITSEKVVMTMYSQTT-----SNFLPESAAAYYVGNLKYSTFMLJ
	BtJL2012_212	LGINITNYKVVITLKEPTNQ-----AFQAAAYVGNLKYSTFMLJ
	BtBtKY72_214	GLNITIQEKVIMTLESETT-----SENADASVYVGHKLPRTFMLJ
	BtBM48-31_213	GLNITVYKAIMTLESTQ-----SNFDADASVYVGHKLPRTFMLJ
	BtRacCS271_215	XSLNITNFKVVIITFRSNSQ-----PXXXXFAVGSCLKLTTIMLJ
	BtRacCS253_215	XXXXXXXXXVVIITFRSNSQ-----PLOAXXXXXXKLTITIMLJ
	BtRacCS264_215	LGINITNFKVVIITFRSNSQ-----PLOANFAVGSCLKLTTIMLJ
	BtRacCS224_219	GXNXXXXXXXXVVIITFRSNSQ-----PLOANFAVGSCLKLTTIMLJ







*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

The screenshot shows the Zenodo preprint page for the paper. The header includes the Zenodo logo, search bar, and navigation links. The main content area displays the title, authors (Yan, Li-Meng; Kang, Shu; Guan, Jie; Hu, Shanchang), and a detailed abstract. The abstract discusses the COVID-19 pandemic and the origin of SARS-CoV-2, arguing for a laboratory origin based on genomic, structural, and medical evidence. On the right side, there are statistics for views (776,634) and downloads (597,172), along with an OpenAIRE logo and publication details including the date (September 14, 2020), DOI (10.5281/zenodo.4028830), and communities (Coronavirus Disease Research Community - COVID-19). At the bottom, there is a genomic map showing restriction sites for EcoRI and BestEII on the MT019529.1 sequence.

zenodo Search Upload Communities Log in Sign up

September 14, 2020 Working paper Open Access

## Unusual Features of the SARS-CoV-2 Genome Suggesting Sophisticated Laboratory Modification Rather Than Natural Evolution and Delineation of Its Probable Synthetic Route

Yan, Li-Meng; Kang, Shu; Guan, Jie; Hu, Shanchang

The COVID-19 pandemic caused by the novel coronavirus SARS-CoV-2 has led to over 910,000 deaths worldwide and unprecedented decimation of the global economy. Despite its tremendous impact, the origin of SARS-CoV-2 has remained mysterious and controversial. The natural origin theory, although widely accepted, lacks substantial support. The alternative theory that the virus may have come from a research laboratory is, however, strictly censored on peer-reviewed scientific journals. Nonetheless, SARS-CoV-2 shows biological characteristics that are inconsistent with a naturally occurring, zoonotic virus. In this report, we describe the genomic, structural, medical, and literature evidence, which, when considered together, strongly contradicts the natural origin theory. The evidence shows that SARS-CoV-2 should be a laboratory product created by using bat coronaviruses ZC45 and/or ZXC21 as a template and/or backbone. Building upon the evidence, we further postulate a synthetic route for SARS-CoV-2, demonstrating that the laboratory-creation of this coronavirus is convenient and can be accomplished in approximately six months. Our work emphasizes the need for an independent investigation into the relevant research laboratories. It also argues for a critical look into certain recently published data, which, albeit problematic, was used to support and claim a natural origin of SARS-CoV-2. From a public health perspective, these actions are necessary as knowledge of the origin of SARS-CoV-2 and of how the virus entered the human population are of pivotal importance in the fundamental control of the COVID-19 pandemic as well as in preventing similar, future pandemics.

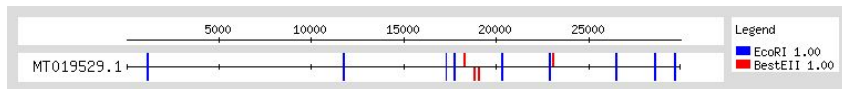
Indexed in OpenAIRE

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

Legend  
■ EcoRI 1.00  
■ BestEII 1.00

# 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	ggctgcgta	tagcttg	gaa	ttg	taacaat	1320	
cttgattcta	aggttggtgg	taattataat	tacctgtata	gattgtttag	gaagtcta	at		1380	
ctcaaacctt	ttagagagaga	tatttcaact	gaaatctatc	aggccggtag	cacaccttgt			1440	
aatggtgttg	aaggttttaa	ttgttacttt	cctttacaat	catatggttt	ccaaccact			1500	
aatggtgttg	gttacc	aacc	atacagagta	gtagtacttt	cttttgaact	tctacatgca		1560	
		G	Y	Q					
		BstEII							
B ZC45		EcoRI							
		W	N	T					
ttacctgatg	attttacagg	ttgtgtcata	gcttg	gaaca	ct	tgccaaaca	ggatgtaggt	1320	
aattatttct	acaggtctca	tcgttctacc	aaattgaaac	catttgaaag	agatccttcc			1380	
tcagacgaga	atggtgtccg	tacacttagt	acttatgact	tcaaccctaa	tgtaccactt			1440	
gaa	tacc	aaag	ctacaagggt	tgttgttttg	tcatttgagc	ttcta	aatgca	accagctaca	1500
		E	Y	Q					

- 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.

	EcoRI
HuCoV2_WH01_2019_215	ACAGGCTGCGTTATAGCTTGAATTCGAACAA
BtRaTG13_2013_Yunnan_215	ACTGGTTGTGTTATAGCTTGAATTCGAACAA
PnMP789_214	ACAGGTTGTGTAATAGCTTGAATTCGAACAA
PnGX-P1E_2017_215	ACTGGTTGTGTTATTGCTTGAATTCAGTTAA
PnGX-P2V_2018_215	ACTGGTTGTGTTATTGCTTGAATTCAGTTAA
BtZC45_215	ACAGGTTGTGCATAGCTTGAACACAGCCAA
BtZXC21_214	ACAGGTTGTGCATAGCTTGAACACAGCCAA
BtLYRa11_215	ATGGGTTGTGCTTGGCTTGAACACAGGAA
BtRs4874_214	ACGGGTTGTGCTTGGCTTGAATACAGGAA
BtYN2018B_215	ATGGGTTGTGCTTGGCTTGAATACAGGAA
HuSARS-Frankfurt-1_2003_215	ATGGGTTGTGCTTGGCTTGAATACAGGAA
Cv007-2004_214	ATGGGTTGTGCTTGGCTTGAATACAGGAA
BtHKU3-12_214	ACTGGCTGTGTAATTGCTTGAATACAGCTAA
Btrec-SARsg_2008_215	ACTGGCTGTGTAATTGCTTGAATACAGCTAA
BtYN2013_212	ACAGGCTGTGCATAGCTTGAATACAGCTAA
BtRm1/2004_215	ACAGGCTGTGTAATAGCTTGAATACAGCTAA
BtRp3-2004_214	ACTGGTTGGGTAATAGCTTGAATACAGCCAA
BtGX2013_212	ACTGGCTGTGTAATCGCTTGAATACAGCTAA
BtRp-Shaanxi2011_213	ACAGGCTGTGTAATAGCTTGAACACAGCCAA
BtSC2018_214	ACTGGCTGTGTAATAGCTTGAATACAGCTAA
BtCp-Yun_2011_214	ACAGGTTGGGTAATTGCTTGAATACAGCTAA
BtRs_672-2006_208	ACAGGCTGTGCATAGCTTGAACACAGCTAA
BtYN2018C_215	ACAGGCTGTGTTATTGCTTGAATACAGCTAA
BtYNLF_31C_215	ACAGGCTGTGTTATAGCTTGAACACAGCCAA
BtJL2012_212	ATAGGTTGTGTTATAGCTTGAACACAGCCAA
BtBtKY72_214	ACTGGCTGTGTTTATAGCTTGAATACAGCTAA
BtBM48-31_213	ACAGGTTGTGTAATAGCTTGAATACAGCTAA
BtHKU5_219	TCT---TCACCAATTGCAATTACAACCTAACAA
HuOC43_241	ACAAGTTGTCAGTTGTAATTAATTTACCTGTC
CmMERS_218	CCCACATGTTTATTTTATGCGACTGTTTCCCTCA
HuMERS_172-06_2015_218	CCCACATGTTTATTTTATGCGACTGTTTCCCTCA
HuTGEV_210	TATTGATTGATATATCTTTTAAATTTGACCAGT
BtHKU9-1_211	TACGGTTGTTTGCATGCAATCTTATTTGAATTC
Hu229E_205	-----GCTAGTATTAAACACGGGAAA
HuNL63_210	---TCATGGCACATTTATTTAAAGAGTGGCAC
PiSADS_205	CCTTATGAATGTTTGGGTTGTCATGGAATGA
PiPRCV_205	-----GCCACCGCTGTTATAAAAACCTGGTAC

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.

**BstEII**

HuCoV2_WH01_2019_215	GGTGTGGTTACCAACCATACAGAA
BtRaTG13_2013_Yunnan_215	GGTGTGGTACCAACCTTATAGGA
PnMP789_214	GGTGTGGTTACCAACCTTATAGAA
PnGX-P1E_2017_215	GGTGTAACTACCAACCTTTTAGAA
PnGX-P2V_2018_215	GGTGTAACTACCAACCTTTTAGAA
BtZC45_215	CCACTTGAATACCAAGCTACAAGG
BtZXC21_214	CCGCTTGAATATCAAGCTACAAGG
BtLYRa11_215	GGCATGGTTACCAACCTTATAGAA
BtRs4874_214	GGCATGGCTACCAACCTTATAGAA
BtYN2018B_215	GGCATGGCTATCAACCTTATAGAA
HuSARS-Frankfurt-1_2003_215	GGCATGGCTACCAACCTTACAGAA
Cv007-2004_214	GGCATGGCTACCAACCTTACAGAA
BtHKU3-12_214	CCAGTAGCATATCAGGCTACTAGGA
Btrec-SARSg_2008_215	CCAGTAGCATATCAGGCTACTAGGA
BtYN2013_212	CCCTCTGATTATCAAGCCACCAGAA
BtRm1/2004_215	CCAGTGAATACCAAGGCACTAGGA
BtRp3-2004_214	CCGGTTGCTTATCAGGCTACTAGGA
BtGX2013_212	CCAGTGGCATATCAGGCTACTAGGA
BtRp-Shaanxi2011_213	CCACTTGAATATCAGGCTACTAGGA
BtSC2018_214	CCGGTGGCATATCAGGCTACTAGAA
BtCp-Yun_2011_214	CCACTTGAATACCAAGCTACTAGAA
BtRs_672-2006_208	CCATATTGAATATCAGGCTACTAGGA
BtYN2018C_215	CCATATTGAATATCAGGCTACTAGGA
BtYNLF_31C_215	CCCCTTGAATATCAAGCCACTAGAA
BtJL2012_212	CCCTCTGAGTACCAAGCCACTAGAA
BtBtKY72_214	GGTGTGGTTACCAACCATATAGAA
BtBM48-31_213	GGAAATGGCTTTCAACCATACAGAA
BtHKU5_219	GTTACTCTTCTCTTACAGTGGACT
HuOC43_241	AAGTGCCCCCAAATAAATCTTTAA
CmMERS_218	CCACTTGAAGGTGGTGGCTGGCTT
HuMERS_172-06_2015_218	CCACTTGAAGGTGGTGGCTGGCTT
HuTGEV_210	AAACACAGCTATTACAAAGGTGAC
BtHKU9-1_211	CCTTTTCTTAT---GTTTATGGT
Hu229E_205	ATACCCGGTGGTTGCGCAATGCC
HuNL63_210	GTGCTTGGTAGTTGTAATTTCCG
PiSADS_205	-----GGTGG
PiPRCV_205	GTTGGTGTAAATGTAAGTTTGA



# Co-occurrence des sites de restriction qui encadrent le RBD

	EcoRI	BstEII
HuCoV2_WH01_2019_215	ACAGGCTGCGTTATAGCTTGGAAATCTAACAA	GGTGTGGTTACCACCATACAGAA
BtRaTG13_2013_Yunnan_215	ACTGGTTGTGTTATAGCTTGGAAATCTAAGCA	GGTGTGGGCACCAACCTTATAGGA
PnMP789_214	ACAGGTTGTGTAATAGCTTGGAAATCTAACAA	GGTGTGGTTACCACCTTATAGAA
PnGX-P1E_2017_215	ACTGGTTGTGTTATTGCTTGGAAACACAGTTAA	PnGX-P1E_2017_215 GGTGTAACTACCAACCTTTTAGAA
PnGX-P2V_2018_215	ACTGGTTGTGTTATTGCTTGGAACTCAGTTAA	PnGX-P2V_2018_215 GGTGTAACTACCAACCTTTTAGAA
BtZC45_215	ACAGGTTGTGTCATAGCTTGGAAACACAGCCAA	BtZC45_215 CCACTTGAATACCAAGCTACAAGG
BtZXC21_214	ACAGGTTGTGTCATAGCTTGGAAACACAGCAA	BtZXC21_214 CCGCTTGAATATCAAGCTACAAGG
BtLYRa11_215	ATGGGTTGTGTCCTTGCTTGGAAACACAGGAA	BtLYRa11_215 GGCATGGTTACCACCTTATAGAA
BtRs4874_214	ACGGGTTGTGTCCTTGCTTGGAAACTAGGAA	BtRs4874_214 GGCATGGCTACCACCTTATAGAA
BtYN2018B_215	ATGGGTTGTGTCCTTGCTTGGAAACTAGGAA	BtYN2018B_215 GGCATGGCTATCAACCTTATAGAA
HuSARS-Frankfurt-1_2003_215	ATGGGTTGTGTCCTTGCTTGGAAACTAGGAA	HuSARS-Frankfurt-1_2003_215 GGCATGGCTACCAACCTTACAGAA
Cv007-2004_214	ATGGGTTGTGTCCTTGCTTGGAAACTAGGAA	Cv007-2004_214 GGCATGGCTACCAACCTTACAGAA
BtHKU3-12_214	ACTGGCTGTGTAATTGCTTGGAAACTAGCTAA	BtHKU3-12_214 CCAGTAGCATATCAGGCTACTAGGA
Btrec-SARSg_2008_215	ACTGGCTGTGTAATTGCTTGGAAACTAGCTAA	Btrec-SARSg_2008_215 CCAGTAGCATATCAGGCTACTAGGA
BtYN2013_212	ACAGGCTGTGTCATAGCTTGGAAACTAGCTAA	BtYN2013_212 CCCTTGATATCAAGCCACCAGAA
BtRm1/2004_215	ACAGGCTGTGTAATAGCTTGGAAACTAGCTAA	BtRm1/2004_215 CCAGTTGAATACCAAGGCACTAGGA
BtRp3-2004_214	ACTGGTTGCGTAATAGCTTGGAAACTAGCTAA	BtRp3-2004_214 CCGGTTGCTTATCAGGCTACTAGGA
BtGX2013_212	ACTGGCTGTGTAATAGCTTGGAAACTAGCTAA	BtGX2013_212 CCAGTGGCATATCAGGCTACTAGGA
BtRp-Shaanxi2011_213	ACAGGCTGTGTAATAGCTTGGAAACTAGCTAA	BtRp-Shaanxi2011_213 CCACTTGCATATCAGGCTACTAGGA
BtSC2018_214	ACTGGCTGTGTAATAGCTTGGAAACTAGCTAA	BtSC2018_214 CCGGTGGCATATCAGGCTACTAGAA
BtCp-Yun_2011_214	ACAGGTTGCGTAATAGCTTGGAAACTAGCTAA	BtCp-Yun_2011_214 CCACTTGAATACCAAGCTACTAGAA
BtRs_672-2006_208	ACAGGCTGTGTCATAGCTTGGAAACTAGCTAA	BtRs_672-2006_208 CCATTTGAATATCAGGCTACTAGGA
BtYN2018C_215	ACAGGCTGTGTTATTGCTTGGAAACTAGCTAA	BtYN2018C_215 CCATTTGAATATCAGGCTACTAGGA
BtYNLF_31C_215	ACAGGCTGTGTTATTGCTTGGAAACTAGCTAA	BtYNLF_31C_215 CCCGTTGAATATCAAGCCACTAGAA
BtJL2012_212	ATAGGTTGTGTTATAGCTTGGAAACTAGCTAA	BtJL2012_212 CCCTTGAGTACCAAGCCACTAGAA
BtBtKY72_214	ACTGGCTGTGTTTATAGCTTGGAAACTAGCTAA	BtBtKY72_214 GGTGTGGTTACCACCTTATAGAA
BtBM48-31_213	ACAGGTTGTGTAATAGCTTGGAAACTAGCTAA	BtBM48-31_213 GGAAATGGCTTTCAACCATACAGAA
BtHKU5_219	TCT---TCACCAATTGCAATTACAACCTAACAA	BtHKU5_219 GTTACTCTTCTTACAGTGGACT
HuOC43_241	ACAAGTTGTGTCAGTTGATTTAATTTACCTGC	HuOC43_241 AAGTGCCCCCAAACCTAAATCTTTA
CmMERS_218	CCCACATGTTTGATTTTAGCGACTGTTCCCTCA	CmMERS_218 CCACTTGAAGGTTGGTGGCTGGCTT
HuMERS_172-06_2015_218	CCCACATGTTTGATTTTAGCGACTGTTCCCTCA	HuMERS_172-06_2015_218 CCACTTGAAGGTTGGTGGCTGGCTT
HuTGEV_210	TATTGATTGTATATCTTTTAAATTTGACCCTG	HuTGEV_210 AAACACAGCTATTACAAAGGTGAC
BtHKU9-1_211	TACGGTTGTTTGCATGCATTCTATTTGAATTC	BtHKU9-1_211 CCCTTTCTTAT---GTTTATGGTT
Hu229E_205	-----GCTAGTATTAACACGGGAAA	Hu229E_205 ATACCCGGTGGTTGCGCAATGCC
HuNL63_210	---TCATGGCACATTTATTTAAAGAGTGGCAC	HuNL63_210 GTGCGCTGGTAGTTGTAATTTCCG
PiSADS_205	CCTTATGAATGTTTTGGTGGTTCATGGAATGA	PiSADS_205 -----TGGTGG
PiPRCV_205	-----GCCACCGCTGTATAAAAACCTGGTAC	PiPRCV_205 TTTGGTGTAAATGTAAGTTTGA