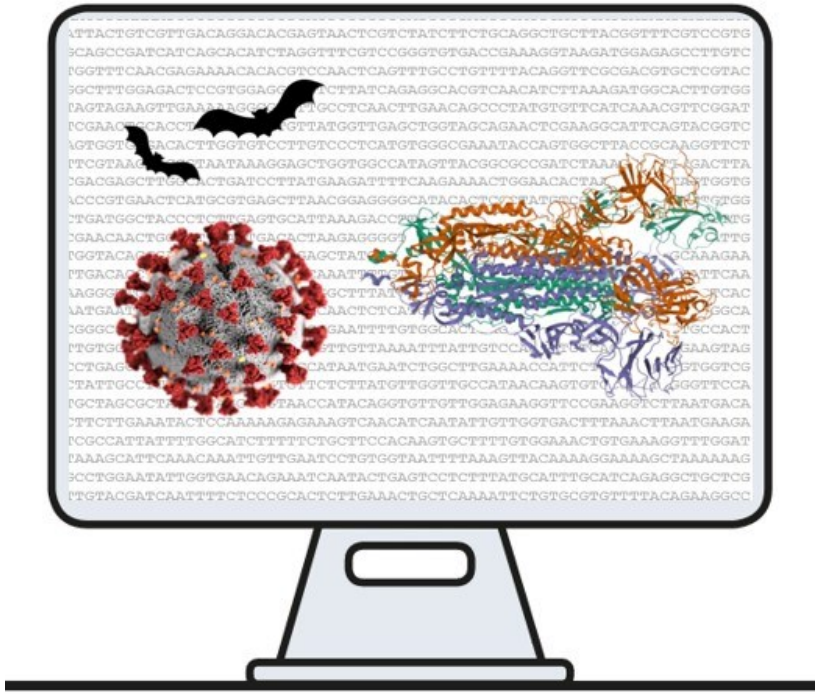
Swiss Institute of Bioinformatics

UNIL | Université de Lausanne
Service Culture et Médiation scientifique

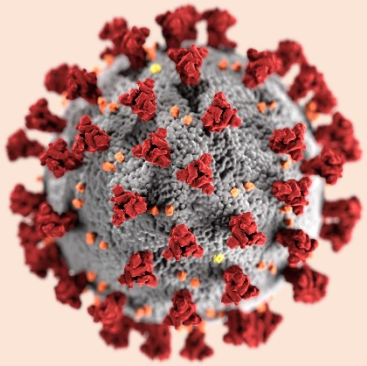
Hunting SARS-CoV-2, its variants and its origin

with the help of bioinformatics…

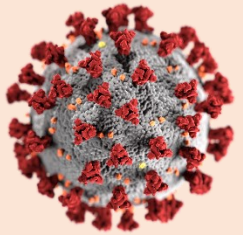
This workshop is an opportunity to discover SARS-CoV-2 and bioinformatics databases & tools used by researchers from all over the world!



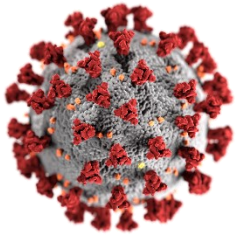
*…and to understand the importance of having an **open access to all this data** (open data)*



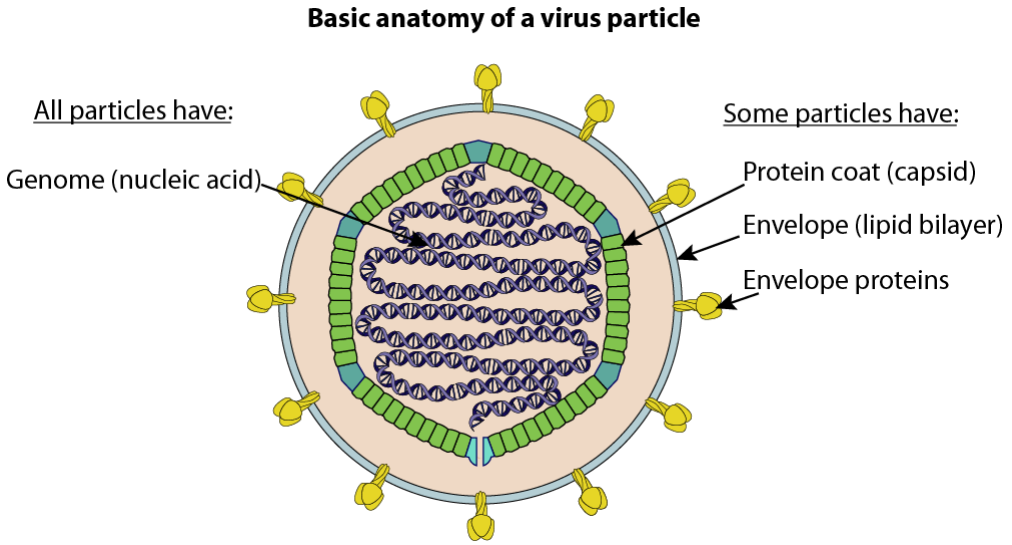
1 - Introduction



What is a virus?



A virus is a parasitic agent transmitted via microscopic virus particle that contains genetic material (DNA or RNA).



A virus can only replicate itself by entering a cell and using the cellular machinery. Virus particles are formally referred as virion.

Some viruses infect animals, other viruses infect plants. Viruses can also infect bacteria: these are called bacteriophages.

If a virus causes a disease, it is considered as pathogenic.

Viruses & some numbers …

8% of the human genome is of viral origin. A protein of viral origin, syncytin, played a key role in the evolution of mammals, allowing the placenta to appear.

Other human proteins of viral origin (UniProtKB)

The living world could not exist without viruses!

An article published in 2011 in Nature Microbiology estimated that there are more than one quintillion - one followed by 30 zeros - of viruses on Earth.

"We swallow over a billion viruses every time we go swimming."

"They are the most diverse organisms on our planet (...) and we still know nothing about them".

Viruses play a key role in the regulation of carbon in the ocean.

In a 2018 study, Suttle found that more than 800 million viruses were deposited on every square meter of the Earth every day.

A virus can produce more than 1000 copies of itself per day.

Viruses that infect bacteria (bacteriophages) play an important role in regulating populations of bacteria that are harmful to humans. They could be an alternative to antibiotics in some cases (phagotherapy).

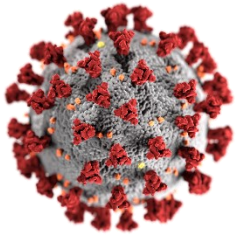
We touch hundreds of millions of viruses every day.

Source (FR)

Viruses that infect human…

“Biologists estimate that 380 trillion viruses are living on and inside human body right now—10 times the number of bacteria.”

The human virome (Scientific American)



There are about 200 types of viruses known to infect humans.
There are different host, modes of transmission, as well as different associated pathologies (Diseases).

https://viralzone.expasy.org/678

ViralZone search

Human viruses and associated pathologies

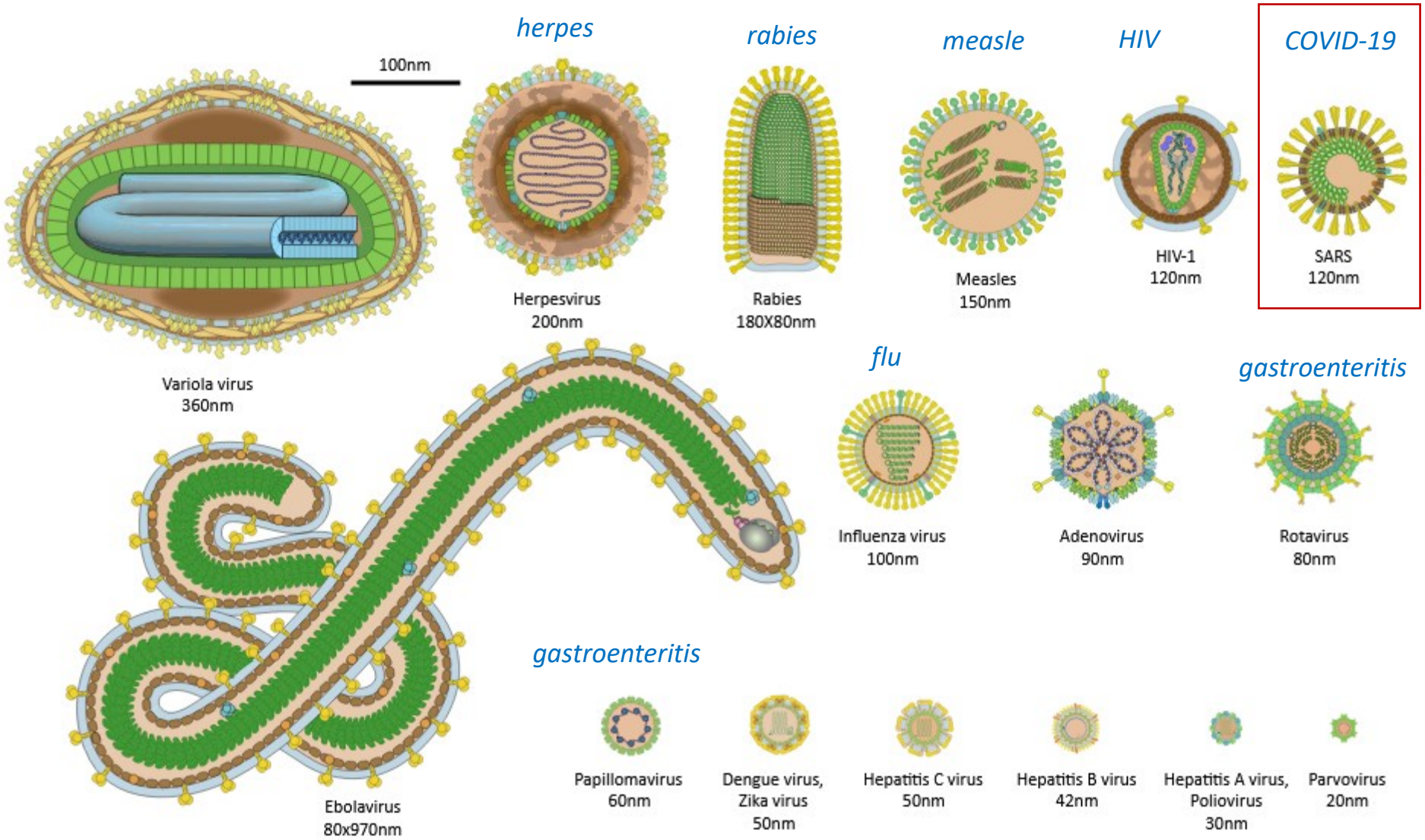
The table below displays the list of human viral pathogens, with transmission and general facts about associated pathologies.

(See human viruses by Baltimore classification) View in TEXT format

Virus	Genus, Family	Host	Transmission	Disease	genome	Proteome
Adeno-associated virus	Dependovirus, Parvoviridae	Human, vertebrates	Respiratory	None	Genome	Proteome
Aichi virus	Kobuvirus, Picornaviridae	Human	Fecal-oral	Gastroenteritis	Genome	Proteome
Australian bat lyssavirus	Lyssavirus, Rhabdoviridae	Human, bats	Zoonosis, animal bite	Fatal encephalitis	Genome	Proteome
BK polyomavirus	Polyomavirus, Polyomaviridae	Human	Respiratory fluids or urine	None	Genome	Proteome
Banna virus	Seadornavirus, Reoviridae	Human, cattle, pig, mosquitoes	Zoonosis, arthropod bite	Encephalitis	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11	Proteome
Barmah forest virus	Alphavirus, Togaviridae	Human, marsupials, mosquitoes	Zoonosis, arthropod bite	Fever, joint pain	Genome	Proteome
Bunyamwera virus	Orthobunyavirus, Bunyaviridae	Human, mosquitoes	Zoonosis, arthropod bite	Encephalitis	1, 2, 3	Proteome
Bunyavirus La Crosse	Orthobunyavirus, Bunyaviridae	Human, deer, mosquitoes, tamias	Zoonosis, arthropod bite	Encephalitis	1, 2, 3	Proteome
Bunyavirus snowshoe hare	Orthobunyavirus, Bunyaviridae	Human, rodents, mosquitoes	Zoonosis, arthropod bite	Encephalitis	S M L	Proteome
Cercopithecine herpesvirus	Lymphocryptovirus, Herpesviridae	Human, monkeys	Zoonosis, animal bite	Encephalitis	Genome	Proteome
Chandipura virus	Vesiculovirus, Rhabdoviridae	Human, sandflies	Zoonosis, athropod bite	Encephalitis	Not available	Proteome
Chikungunya virus	Alphavirus, Togaviridae	Human, monkeys, mosquitoes	Zoonosis, arthropod bite	Fever, joint pain	Genome	Proteome
Cosavirus A	Cosavirus, Picornaviridae	Human	Fecal-oral (probable)	-	Genome	Proteome
Cowpox virus	Orthopoxvirus, Poxviridae	Human, mammals	Zoonosis, contact	None	Genome	Proteome
Coxsackievirus	Enterovirus, Picornaviridae	Human	Fecal-oral	Meningitis, myocarditis, paralysis	Genome	Proteome

ViralZone

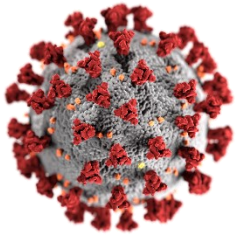
Examples of types of viruses that infect humans and their respective sizes



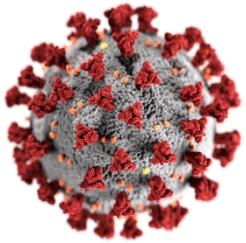
ViralZone

https://viralzone.expasy.org/5216

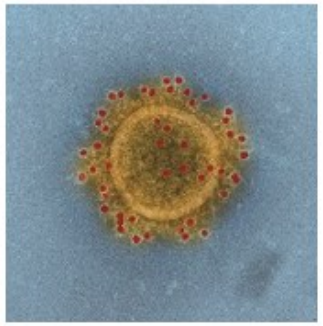
Coronaviruses and SARS-CoV-2



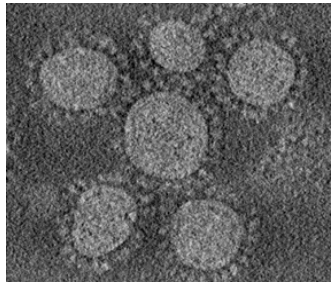
Coronaviruses belong to a large family of viruses (that includes more than 40 species) and most of them are harmless to humans. Four types of coronaviruses (called OC43, 229E, NL63 and HKU1) are endemic and are known to cause colds. Three types of coronaviruses can cause serious lung infections:



- Severe Acute Respiratory Syndrome-related Coronavirus 2
- **SARS-CoV-2**
- responsible for the Coronavirus disease-19 (COVID-19) since 2020
- reservoir: bat(?)

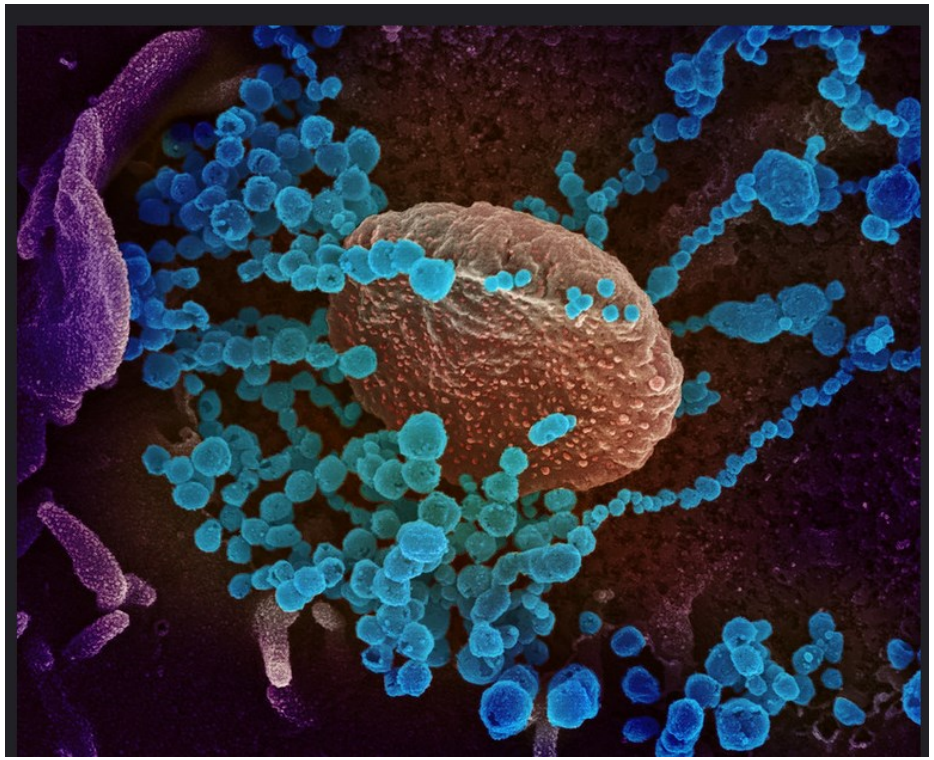


- Middle-East Respiratory Syndrome Coronavirus
- **MERS-CoV**
- responsible recurrent epidemic since 2012
- reservoir: camel, bat (?)

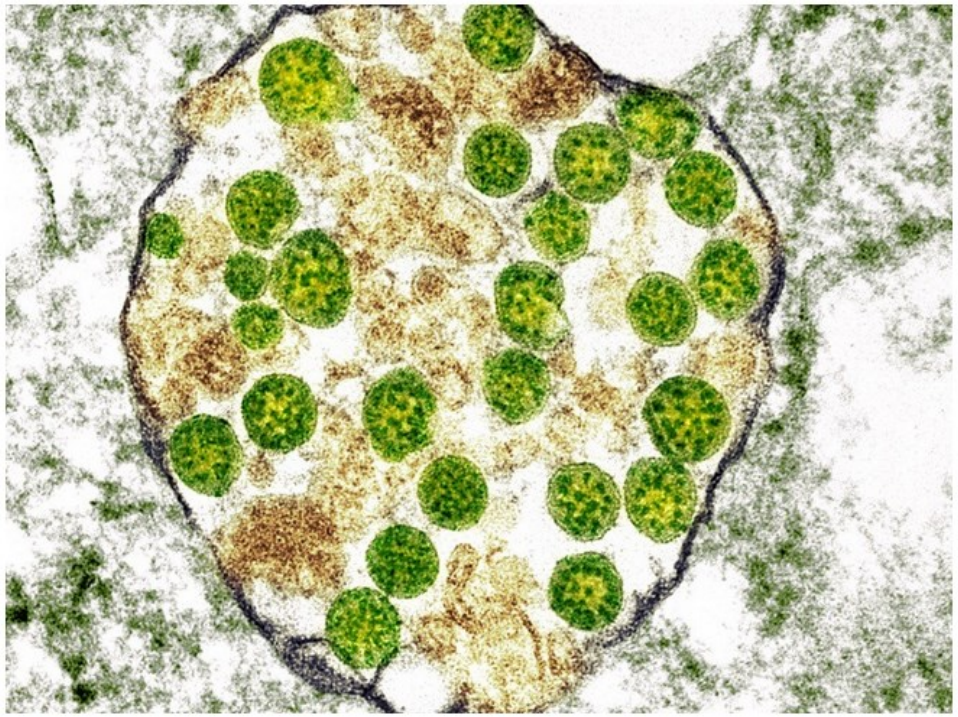


- Severe Acute Respiratory Syndrome-related Coronavirus
- **SARS-CoV**
- responsible for an epidemic in 2003, that affected more than 30 countries.
- reservoir: bat

SARS-CoV-2 images outside and inside a cell ...

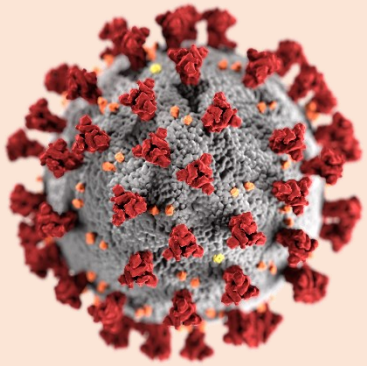


https://www.flickr.com/photos/niaid/49557550751/

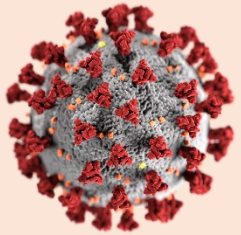


A cell infected with particles of SARS-CoV-2. Credit: Cynthia S. Goldsmith and Azaibi Tamin/CDC/SPL

https://www.nature.com/articles/d41586-020-00502-w



2 – The first sequence of SARS-CoV-2 genome

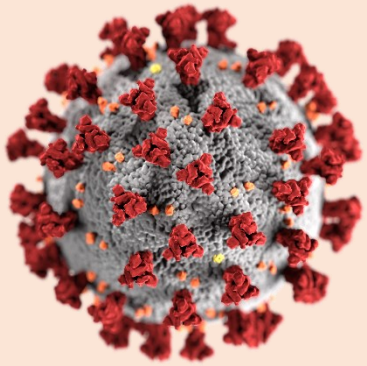


On 10 January 2020, a Chinese team submitted a first sequence SARS-CoV-2's genome to the GenBank database.

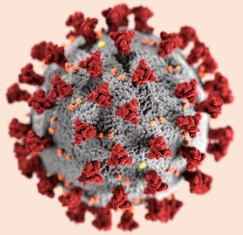
SARS-CoV-2's genome sequence contains 29,903 nucleotides*.

This sequence is considered as the **'reference sequence'**, meaning that the SARS-CoV-2 sequences collected world-wide are compared against it.

* *SARS-CoV-2's genome consists of single-stranded RNA. In databases such as GenBank, this type of genome is represented as a DNA sequence (T instead of U).*



3 – Setting up the RT-PCR test



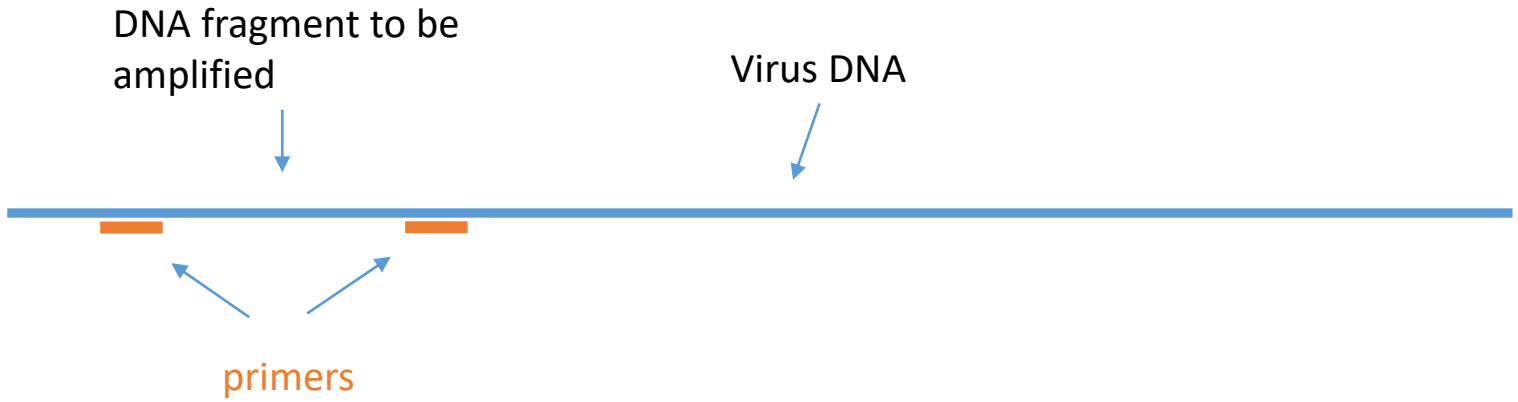
Sequencing the SARS-CoV-2 genome has allowed the rapid implementation of a PCR test to detect the presence of the virus in nasopharyngeal (nose) or oropharyngeal (throat) smears.

The test can be 'positive' for as many as 100 viruses present in the smear!

Photo d'illustration. • © JOE RAEDLE / GETTY IMAGES NORTH AMERICA / Getty Images via AFP

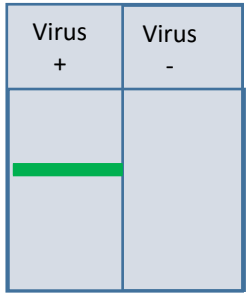
RT-PCR (Reverse Transcription-Polymerase Chain Reaction) is a lab technique that makes it possible to convert RNA into DNA, to selectively amplify (or 'photocopy') a DNA fragment, and then to make millions to billions of fragment copies.

Two small DNA sequences of about 20 nucleotides, called primers, are carefully selected.
They are complementary (100 % identical) to the DNA strand and 'frame' the fragment to be amplified.



Amplification of the fragment through several PCR cycles

If the virus is present in the tested sample, the amplified DNA fragments will be visible on an agarose gel, for example.
Note that quantitative PCR (qPCR, or real-time PCR) is much used in diagnostics. It consists of collecting data during PCR with fluorescently labeled primers.



In order to validate the test, several fragments (different regions in the genome) are amplified.

Test the specificity of the primers (1)

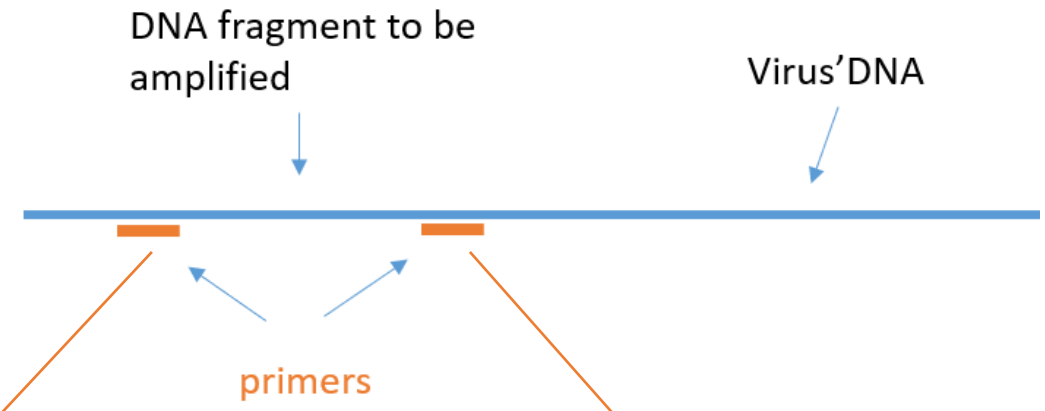
Examples of primers used for PCR testing:

- Primer no 1 CTCGAACTGCACCTCATGG
- Primer no 2 GGCATACACTCGCTATGTC

1. How often do you find the sequence of a primer in the virus's genome (manually)?

- Look for the primer sequence in SARS-CoV-2's genome (Use Ctrl F or command F for mac)

Test the specificity of the primers (1)



Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1, complete genome

NCBI Reference Sequence: NC_045512.2

GenBank Graphics

>NC_045512.2 Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1, complete genome

ATTAAAGGTTTATACCTTCCCAGGTAACAAACCAACCAACTTTCGATCTCTTGTAGATCTGTTCTCTAAA
CGAACTTTAAAATCTGTGTGGCTGTCACTCGGCTGCATGCTTAGTGCACTCACGCAGTATAATTAATAAC
TAATTACTGTCGTTGACAGGACACGAGTAACTCGTCTATCTTCTGCAGGCTGCTTACGGTTTCGTCCGTG
TTGCAGCCGATCATCAGCACATCTAGGTTTCGTCCGGGTGTGACCGAAAGGTAAGATGGAGAGCCTTGTC
CCTGGTTTCAACGAGAAAACACACGTCCAACTCAGTTTGCCTGTTTTACAGGTTCGCGACGTGCTCGTAC
GTGGCTTTGGAGACTCCGTGGAGGAGGTCTTATCAGAGGCACGTCAACATCTTAAAGATGGCACTTGTGG
CTTAGTAGAAGTTGAAAAAGGCGTTTTGCCTCAACTTGAACAGCCCTATGTGTTCATCAAACGTTCGGAT
GCTCGAACTGCACCTCATGGTCATGTTATGGTTGAGCTGGTAGCAGAACTCGAAGGCATTCAGTACGGTC
GTAGTGGTGAGACACTTGGTGTCCTTGTCCCTCATGTGGGCGAAATACCAGTGGCTTACCGCAAGGTTCT
TCTTCGTAAGAACGGTAATAAAGGAGCTGGTGGCCATAGTTACGGCGCCGATCTAAAGTCATTTGACTTA
GGCGACGAGCTTGGCACTGATCCTTATGAAGATTTTCAAGAAAACTGGAACACTAAACATAGCAGTGGTG
TTACCCGTGAACTCATGCGTGAGCTTAACGGAGGGGCATACACTCGCTATGTCGATAACAACTTCTGTGG
CCCTGATGGCTACCCTCTTGAGTGCATTAAAGACCTTCTAGCACGTGCTGGTAAAGCTTCATGCACTTTG
TCCGAACAACTGGACTTTATTGACACTAAGAGGGGTGTATACTGCTGCCGTGAACATGAGCATGAAATTG
CTTGGTACACGGAACGTTCTGAAAAGAGCTATGAATTGCAGACACCTTTTGAAATTAAATTGGCAAAGAA
ATTTGACACCTTCAATGGGGAATGTCCAAATTTTGTATTTCCCTTAAATTCCATAATCAAGACTATTCAA
CCAAGGGTTGAAAAGAAAAAGCTTGATGGCTTTATGGGTAGAATTCGATCTGTCTATCCAGTTGCGTCAC
CAAATGAATGCAACCAAATGTGCCTTTCAACTCTCATGAAGTGTGATCATTGTGGTGAAACTTCATGGCA

➢ Primer no 1 CTCGAACTGCACCTCATGG

Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1, complete genome

NCBI Reference Sequence: NC_045512.2

GenBank Graphics

>NC_045512.2 Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1, complete genome

ATTAAAGGTTTATACCTTCCCAGGTAACAAACCAACCAACTTTCGATCTCTTGTAGATCTGTTCTCTAAA
CGAACTTTAAAATCTGTGTGGCTGTCACTCGGCTGCATGCTTAGTGCACTCACGCAGTATAATTAATAAC
TAATTACTGTCGTTGACAGGACACGAGTAACTCGTCTATCTTCTGCAGGCTGCTTACGGTTTCGTCCGTG
TTGCAGCCGATCATCAGCACATCTAGGTTTCGTCCGGGTGTGACCGAAAGGTAAGATGGAGAGCCTTGTC
CCTGGTTTCAACGAGAAAACACACGTCCAACTCAGTTTGCCTGTTTTACAGGTTCGCGACGTGCTCGTAC
GTGGCTTTGGAGACTCCGTGGAGGAGGTCTTATCAGAGGCACGTCAACATCTTAAAGATGGCACTTGTGG
CTTAGTAGAAGTTGAAAAAGGCGTTTTGCCTCAACTTGAACAGCCCTATGTGTTCATCAAACGTTCGGAT
GCTCGAACTGCACCTCATGGTCATGTTATGGTTGAGCTGGTAGCAGAACTCGAAGGCATTCAGTACGGTC
GTAGTGGTGAGACACTTGGTGTCCTTGTCCCTCATGTGGGCGAAATACCAGTGGCTTACCGCAAGGTTCT
TCTTCGTAAGAACGGTAATAAAGGAGCTGGTGGCCATAGTTACGGCGCCGATCTAAAGTCATTTGACTTA
GGCGACGAGCTTGGCACTGATCCTTATGAAGATTTTCAAGAAAACTGGAACACTAAACATAGCAGTGGTG
TTACCCGTGAACTCATGCGTGAGCTTAACGGAGGGGCATACACTCGCTATGTCGATAACAACTTCTGTGG
CCCTGATGGCTACCCTCTTGAGTGCATTAAAGACCTTCTAGCACGTGCTGGTAAAGCTTCATGCACTTTG
TCCGAACAACTGGACTTTATTGACACTAAGAGGGGTGTATACTGCTGCCGTGAACATGAGCATGAAATTG
CTTGGTACACGGAACGTTCTGAAAAGAGCTATGAATTGCAGACACCTTTTGAAATTAAATTGGCAAAGAA
ATTTGACACCTTCAATGGGGAATGTCCAAATTTTGTATTTCCCTTAAATTCCATAATCAAGACTATTCAA
CCAAGGGTTGAAAAGAAAAAGCTTGATGGCTTTATGGGTAGAATTCGATCTGTCTATCCAGTTGCGTCAC
CAAATGAATGCAACCAAATGTGCCTTTCAACTCTCATGAAGTGTGATCATTGTGGTGAAACTTCATGGCA

➢ Primer no 2 GGCATACACTCGCTATGTC

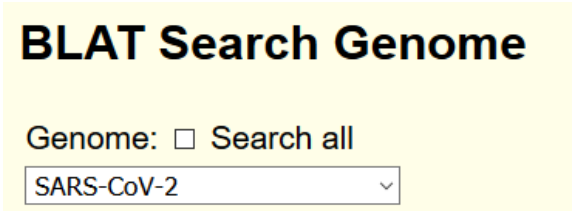
Test the specificity of the primers (2)

Examples of primers used for PCR testing:

- Primer no 1: CTCGAACTGCACCTCATGG
- Primer no 2: GGCATACACTCGCTATGTC

2. How often do you find the sequence of a primer in the **SARS-CoV-2 genome** (with a bioinformatics tool)?

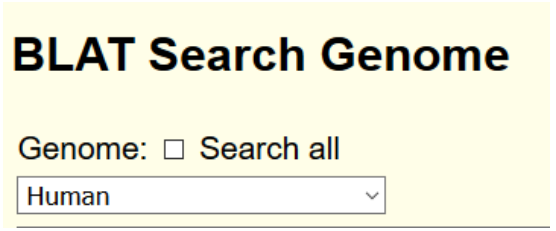
- Use **BLAT@UCSC**
- Copy/paste the primer's sequence
- Select 'genome' : SARS-CoV-2
- Click 'Submit'



3. How often do you find the sequence of a primer in the **human genome**?

Human genome consists of 3 billion nucleotides…

- Use **BLAT@UCSC**
- Copy/paste the primer's sequence
- Select 'genome' : human
- Click 'Submit'



BLAT Search Genome

Genome: ☐ Search all

SARS-CoV-2

Assembly: Dec. 2013 (GRCh38/hg38)

Query type: BLAT's guess

Sort output: query,score

Output type: hyperlink

CTCGAACTGCACCTCATGG

☐ All Results (no minimum matches)

Submit | I'm feeling lucky | Clear

BLAT Search Results

Go back to NC_045512v2:492-510 on the Genome Browser.

Custom track name: blat YourSeq

Custom track description: blat on YourSeq

Build a custom track with these results

ACTIONS	QUERY	SCORE	START	END	QSIZE	IDENTITY	CHROM	STRAND	START	END	SPAN
browser details	YourSeq	19	1	19	19	100.0%	NC_045512v2	+	492	510	19

The primer (19 nucleotides) is found only once in the genome of the virus (100 % identity) at positions 492-510 in the reference genome sequence.

BLAT Search Genome

Genome: ☐ Search all

Human

Assembly: Dec. 2013 (GRCh38/hg38)

Query type: BLAT's guess

Sort output: query,score

Output type: hyperlink

CTCGAACTGCACCTCATGG

☐ All Results (no minimum matches)

Submit | I'm feeling lucky | Clear

Genomes | Genome Browser | Tools | Mirrors | Downloads | My Data | Projects | Help | About Us

Human (hg38) BLAT Results

Sorry, no matches found (with score at least 20)

The sequence of the primer is not found in the human genome (3 billion letters!)

Test the specificity of the primers (3)

Examples of primers used for PCR testing:

- Primer no 1 CTCGAACTGCACCTCATGG
- Primer no 2 GGCATACACTCGCTATGTC

4. Try to type a random sequence (20 letters): can you find it in the genome of the SARS-CoV-2 virus?

- Look for your random sequence in the SARS-CoV-2 genome (Use Ctrl F or commandF for mac)

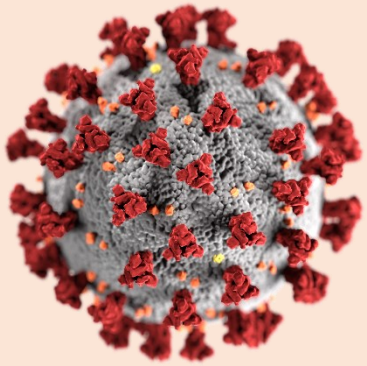
5. Do you find the primer sequences in the **genome of another coronavirus**?

- Look for the primer sequence in SARS CoV genome (Use Ctrl F or commandF for mac)

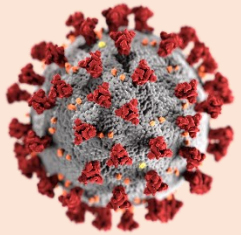
- Alignment of the sequences of the 2 coronavirus genomes in the region 'matching' the primers (top: SARS-CoV-2):

```
NC_045512.2  SARS-CoV-2  419 GGCTTAGTAGAAGTTGAAAAAGGCGTTTTGCCTCAACTTGAACAGCCCTATGTGTTCATC
AY362699.1   SARS-CoV    418 GGTCTAGTAGAGCTGGAAAAAGGCGTACTGCCCCAGCTTGAACAGCCCTATGTGTTCATT
                             **  *******. * ***********: **** **.**********************

NC_045512.2  SARS-CoV-2  479 AAACGTTCGGATGCTCGAACTGCACCTCATGGTCATGTTATGGTTGAGCTGGTAGCAGAA
AY362699.1   SARS-CoV    478 AAACGTTCTGATGCCTTAAGCACCAATCACGGCCACAAGGTCGTTGAGCTGGTTGCAGAA
                             ******** *****   **  .*...*** ** ** .: .* *******

NC_045512.2  SARS-CoV-2  539 CTCGAAGGCATTCAGTACGGTCGTAGTGGTGAGACACTTGGTGTCCTTC
AY362699.1   SARS-CoV    538 ATGGACGGCATTCAGTACGGTCGTAGCGGTATAACACTGGGAGTACTCC
                             .* **.******************** ***.:.***** **:**.** *
```

```
NC_045512.2  SARS-CoV-2  719 GATCCTTATGAAGATTTTCAAGAAAACTGGAACACTAAACATAGCAGTGGTGTTACCCGT
AY362699.1   SARS-CoV    718 GATCCCATTGAAGATTATGAACAAAACTGGAACACTAAGCATGGCAGTGGTGCACTCCGT
                             ***** ::*********:* ** ****************.***.********* :. ****

NC_045512.2  SARS-CoV-2  779 GAACTCATGCGTGAGCTTAACGGAGGGGCATACACTCGCTATGTCGATAACAACTTCTGT
AY362699.1   SARS-CoV    778 GAACTCACTCGTGAGCTCAATGGAGGTGCAGTCACTCGCTATGTCGACAACAATTTCTGT
                             *******  ******** ** ***** *** :*************** ***** ******

NC_045512.2  SARS-CoV-2  839 GGCCCTGATGGCTACCCTCTTGAGTGCATTAAAGACCTTCTAGCACGTGCTGGTAAAGCT
AY362699.1   SARS-CoV    838 GGCCCAGATGGGTACCCTCTTGATTGCATCAAAGATTTTCTCGCACGCGCGGGCAAGTCA
                             *****:***** ********** ***** *****  ****.***** ** ** **. *:
```

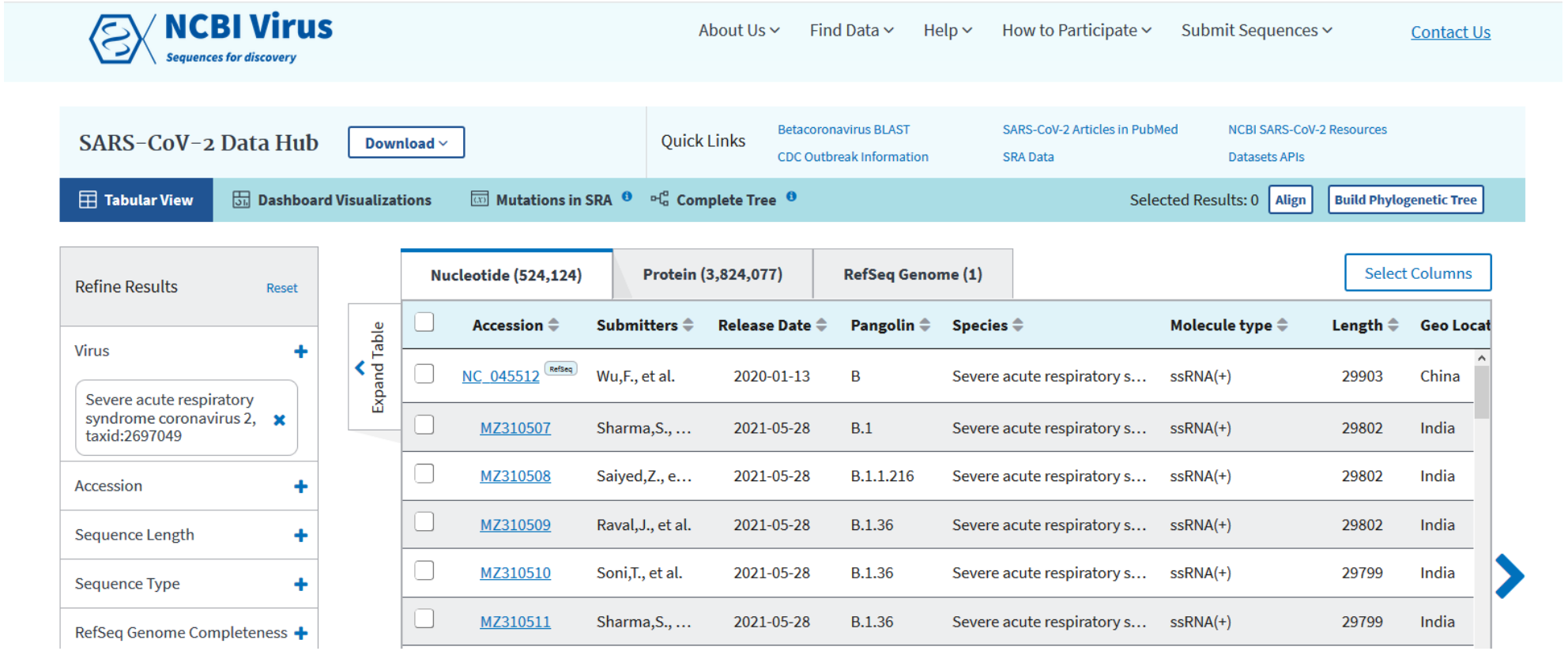


4 - SARS-CoV-2 and its genome(s)



Since the first ‘reference’ genome, several thousand SARS-CoV-2 genomes have been sequenced in different countries. Several research centers give free access to this data. This is essential!

Example (NCBI): [Severe acute respiratory syndrome coronavirus 2 data hub]



Access virus genomes sequenced in different countries

Each genome sequence has its own accession number. As an example, the following URL gives access to the GenBank entry (in **Fasta format**) of the reference genome (NC_045512.2): www.ncbi.nlm.nih.gov/nuccore/**NC_045512.2**?report=fasta**.**

1. Replacing the AC number of your genome sequence of interest in this URL gives access to its GenBank entry: *e.g.*, **www.ncbi.nlm.nih.gov/nuccore/LR991698.2?report=fasta** gives the GenBank entry for LR991698.2 (the UK variant B 1.1.7).

2. Use the AC numbers listed below to explore the SARS-CoV-2 genome sequences from different countries, sequenced and submitted at different times:
    1. MT612198.1: 23-JUN-2020 (Australia);
    2. MT911538.1: 21-AUG-2020 (Minnesota);
    3. MW079825.1: 07-OCT-2020 (Egypt);
    4. MW592707.1: 11-FEB-2021 (Brazil);
    5. MZ026889.1: 26-APR-2021(Bahrain).

As an example, here are the genomes of different virus variants:

SARS-CoV-2 sequence, China (10-JAN-2020 (Wuhan))

- Variant alpha (B1.1.7, UK), alternative: LR991698)
- Variant beta (B.1.351, South Africa)
- Variant gamma (B.1.1.28.1, Brazil, P1)
- Variant delta (B.1.617.2, India)
- Variant epsilon (B.1.427, California, US)
- Variant omicron (B.1.1.529)

Compare the genome sequences 2 by 2: align@UniProt …*it may take a few minutes....*
Once the job is finished, click on 'Similarity' in the left-hand column to see the differences (mutations) more clearly.
Find the different mutations and in particular the one located in position 23,063 (A->T):

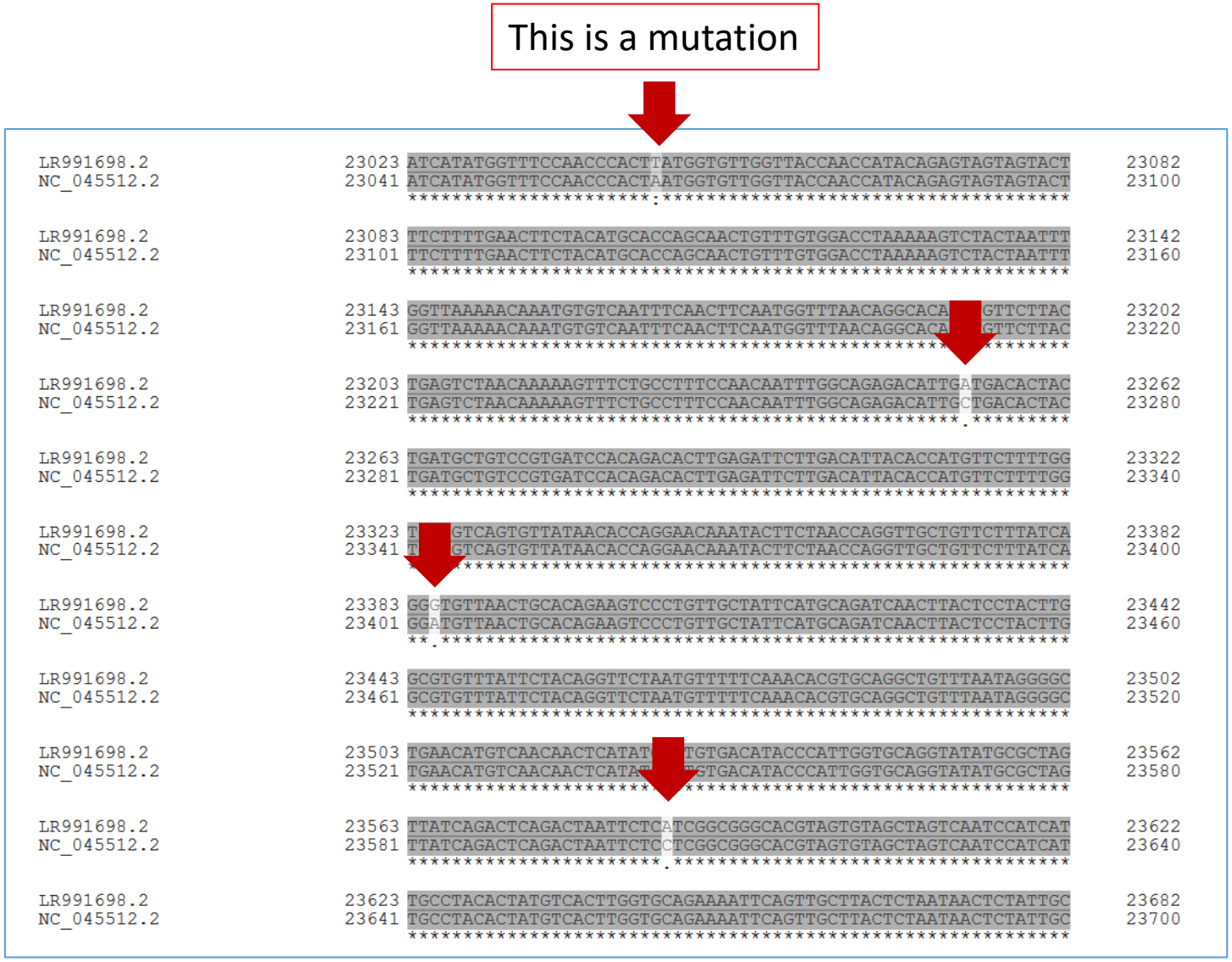
A mutation at position 23,063


```
NC_045512.2    23041 ATCATATGGTTTCCAACCCACTAATGGTGTTGGTTACCAACCATACAGAGTAGTAGTACT    23100
LR991698.2     23023 ATCATATGGTTTCCAACCCACTTATGGTGTTGGTTACCAACCATACAGAGTAGTAGTACT    23082
                     **********************:************************************
```

Example of alignment: reference genome (Wuhan) and genome Alpha, B.1.1.7 (UK): file txt

The discovery of the variant Alpha (UK B1.1.7) in December 2020 had a massive impact. The variant Alpha of SARS-CoV-2 genome has apparently acquired 17 mutations.

The differences are highlighted in white. These differences correspond to the new mutations present in the variant Alpha of SARS-CoV-2 genome.



Alignment of genomes

- Reference genome (NC_045512.2)
- Alpha, B.1.1.7 genome (LR9911698)

Align@UniProt

Copy/paste the fasta format of the sequences

Identify your viral genome: to which variant does it correspond?

The particular combination of differences/mutations found in a given genome allows identification of the virus variant (also called lineage).

Pangolin COVID-19 Lineage Assigner is a tool which allows to identify a virus variant according to its genome sequence.

1. Go to the Pangolin COVID-19 Lineage Assigner of PANGO lineages: pangolin.cog-uk.io
2. Copy & paste one of your previous sequences (Fasta format) into a '.Fasta' file, instead of a .txt file.
3. Import your '.Fasta' file
4. Click on 'Start your analysis' button on the top left.
5. The name of your lineage/variant will appear, together with additional information. Click on the 'i' button. 

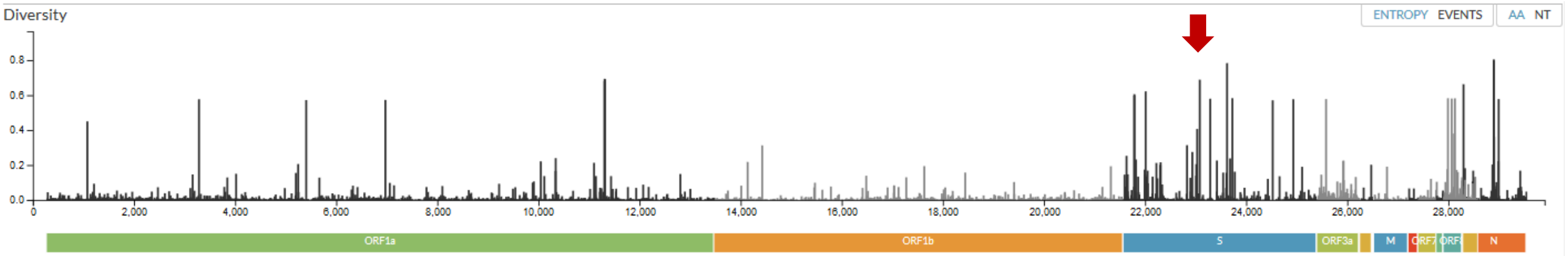
The actual distribution of your variant, if known, will appear on the planet icon.

Note that each virus variant (also called lineage) can have different names (UK variant = B.1.1.7, 20I, 501Y V1, **alpha**, etc)

“The coronavirus mutates relatively rarely. In any case less than a flu, gastroenteritis or hepatitis. (…) But it mutates enough for us to be able to recognise it and identify its 'ancestors'," Pr Didier Trono.

Monitoring the epidemic in real time : https://nextstrain.org/ncov/global

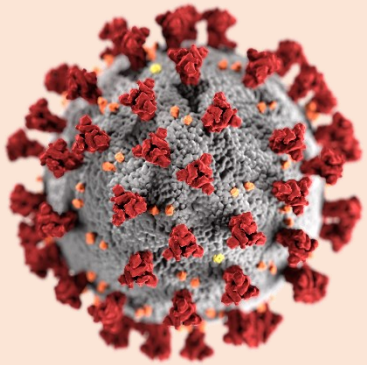
Vertical bars represent the most frequently observed mutations among the approximately 29,000 nucleotides (comparison of thousands of genomes).



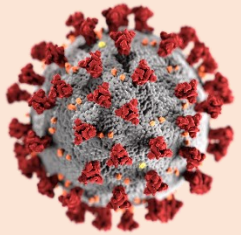
➡ **"Making sense of mutations: what 23,063 A->T means for the COVID-19 pandemic ?“**

How to evaluate the impact of such mutations on the pandemic ?

In order to evaluate the impact of these mutations, we need to look at the SARS-CoV-2 proteins….

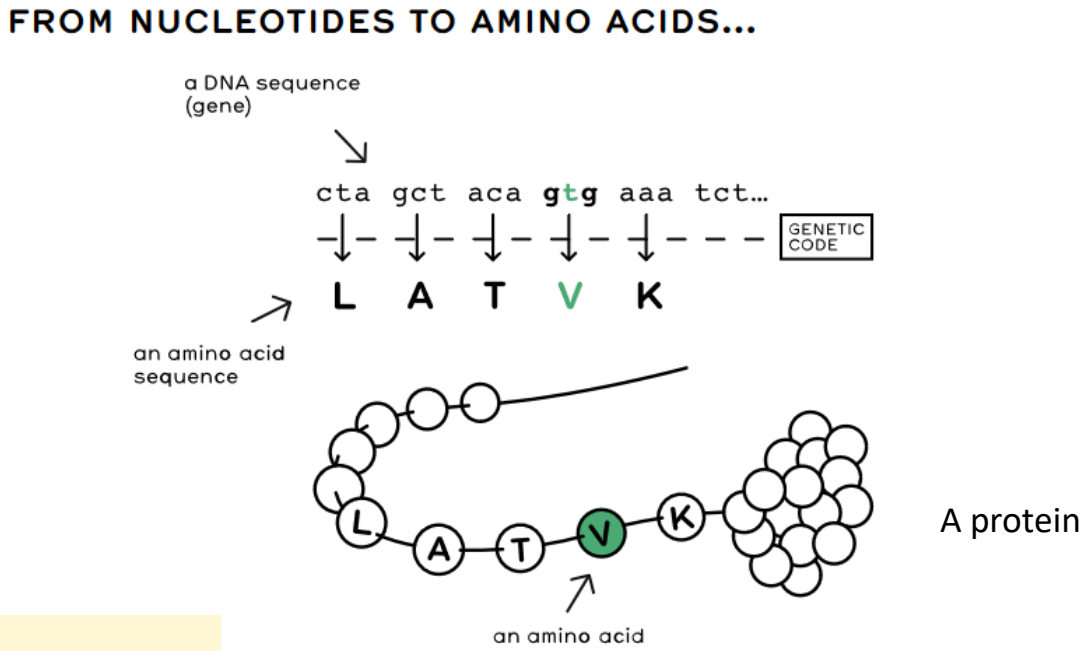


5 - SARS-CoV-2 and its proteins

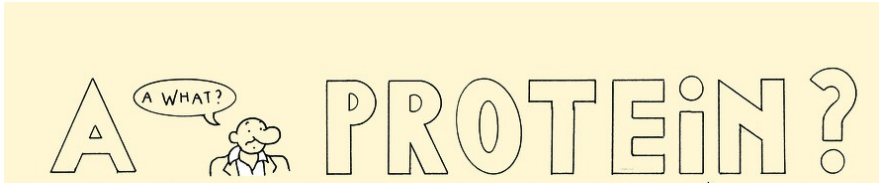


What is a protein?

Analysis of the virus's genome sequence allowed to find the amino acid sequences of the virus's proteins.



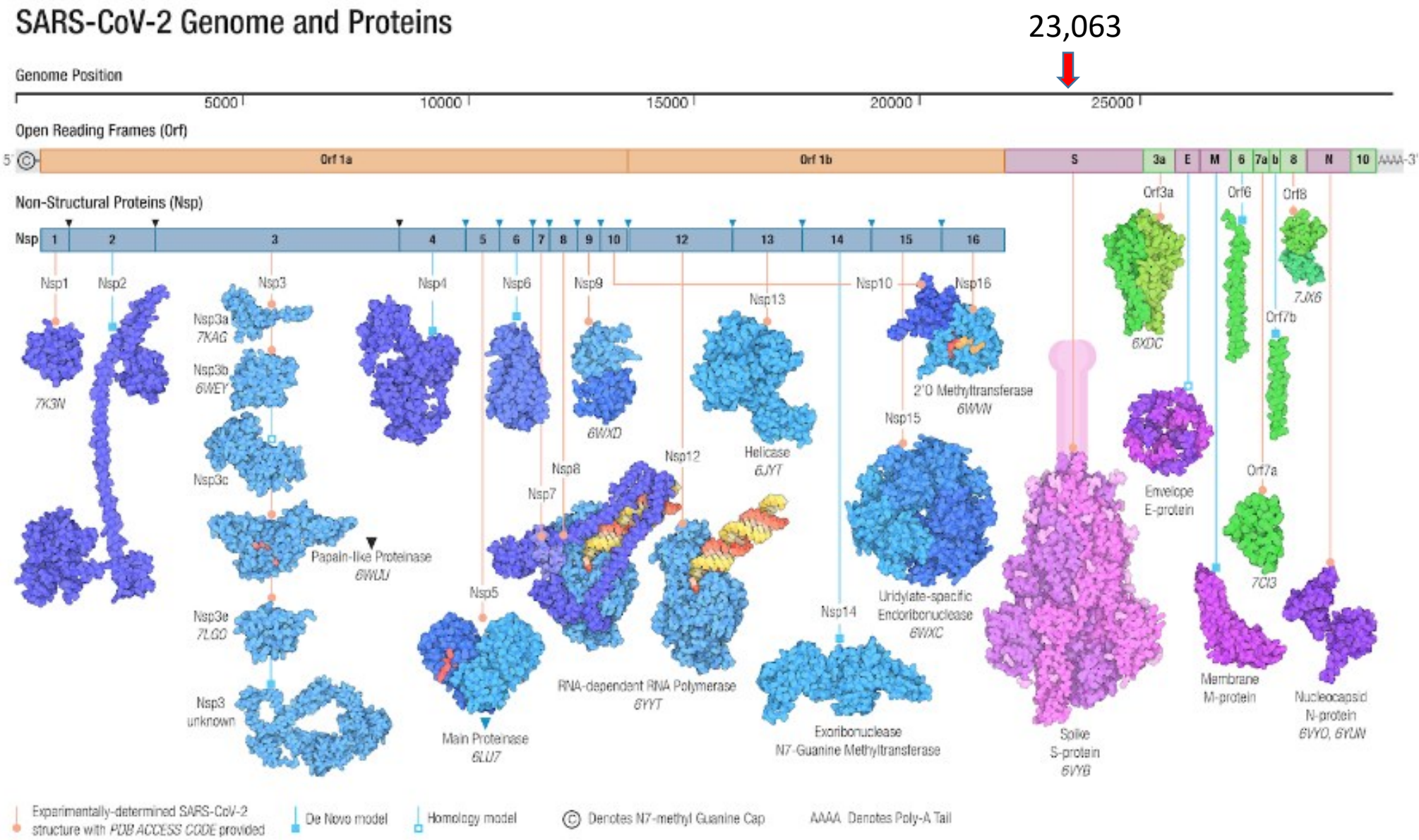
A protein



A protein?

What's a protein?

www.precisionmed.ch

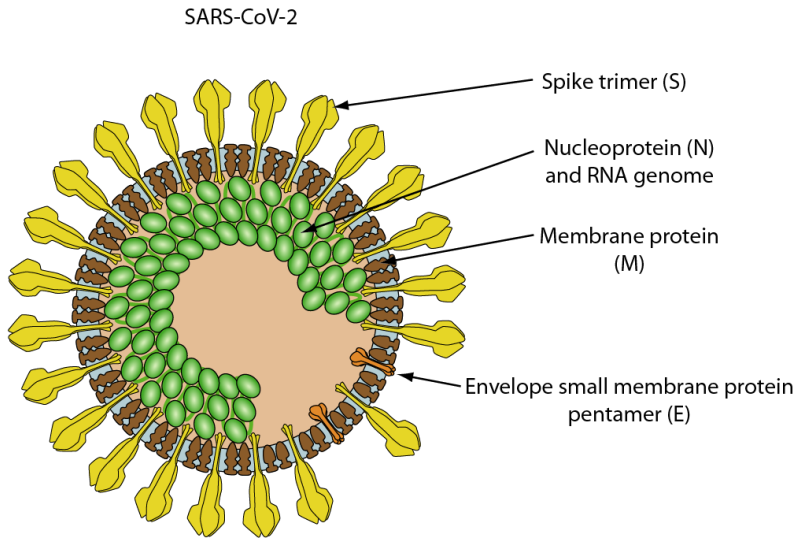


Architecture of the SARS-CoV-2 genome and proteome from *Proteins: Structure, Function, and Bioinformatics* 2021; doi: 10.1002/prot.26250

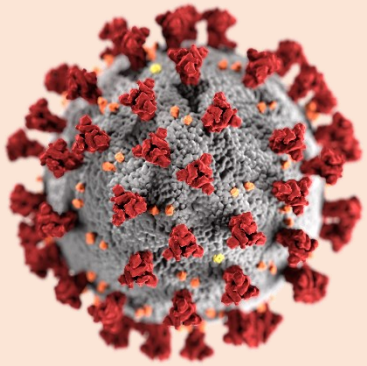
https://www.rcsb.org/news?year=2020&article=5e74d55d2d410731e9944f52&feature=true

Approximately 17 genes coding for proteins have been identified (S, N, M, E, ….) in the SARS-CoV-2 genome.

List of the SARS-CoV-2 proteins and their function in UniProtKB: list

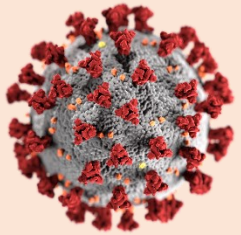


We will now focus on the gene coding for **the Spike protein: gene S**

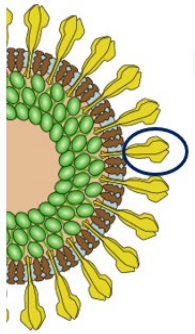


6 - SARS-CoV-2 and its spike protein

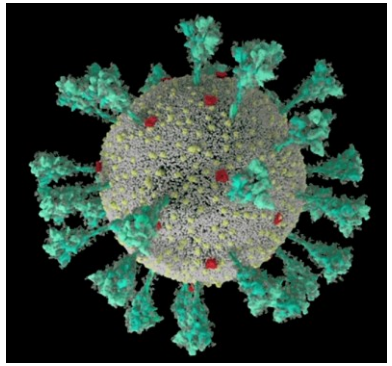
«Only one protein on the surface of the new coronavirus is responsible for its transmission to humans and the resulting pandemic: the Spike protein. Driven by major technological developments in recent years, scientists have rapidly determined its composition and 3D structure, which has greatly aided vaccine development». *Radio Canada*



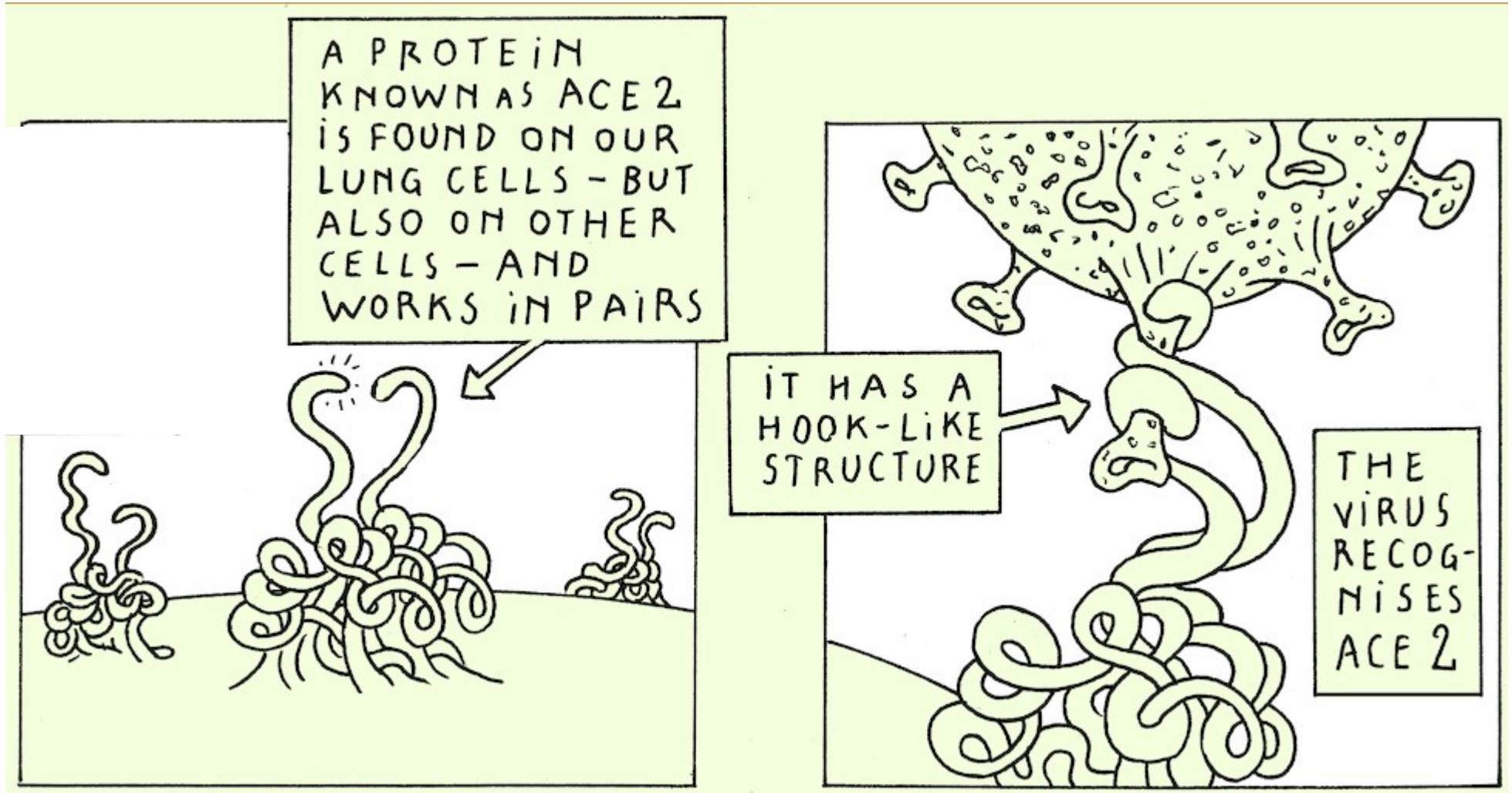
The Spike protein function



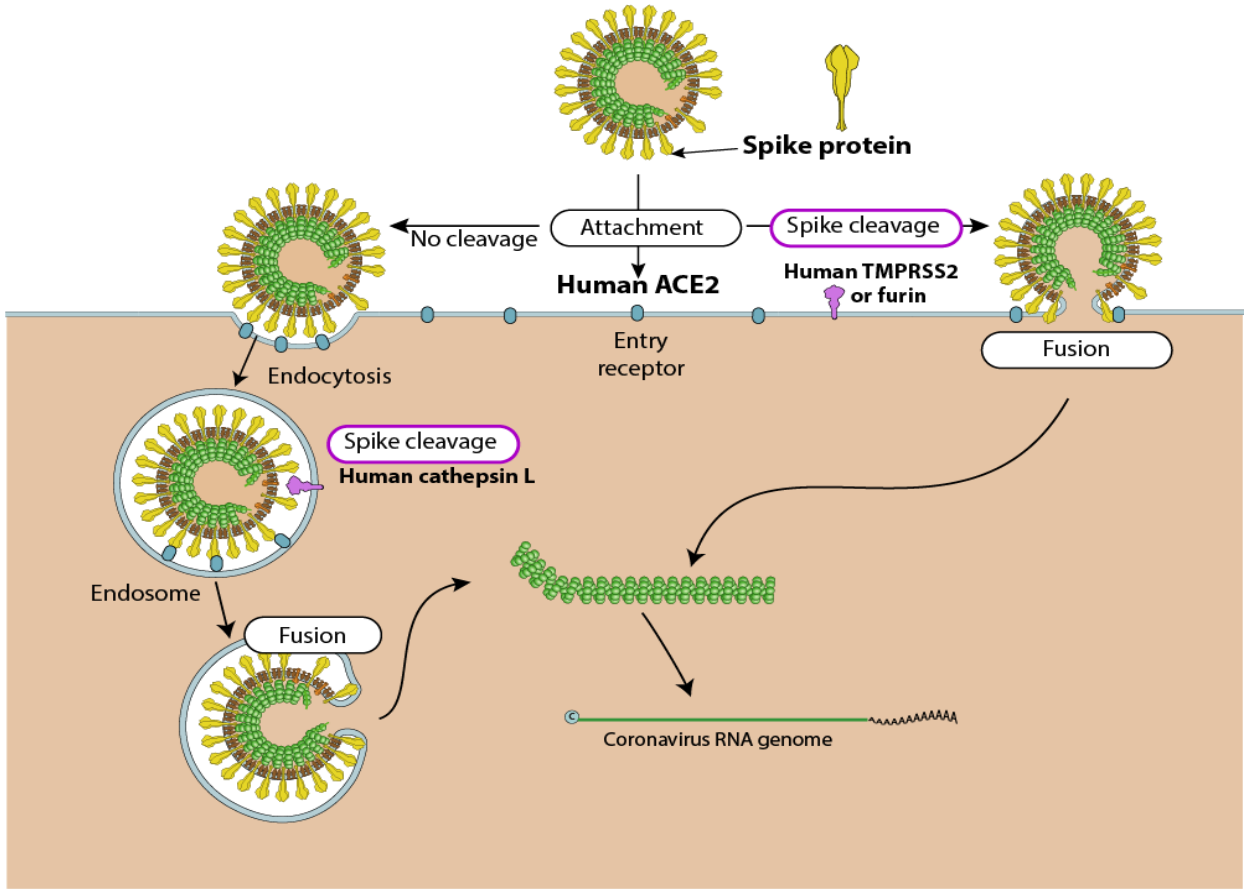
The **Spike** protein allows the virus to enter our cells by interacting, among other things, with a human protein called **ACE2**, which is present on the surface of some human cells (lung, small intestine, …).



How the coronavirus infects cells — and why Delta is so dangerous



[Protein Spotlight comics ACE2](): A way in!

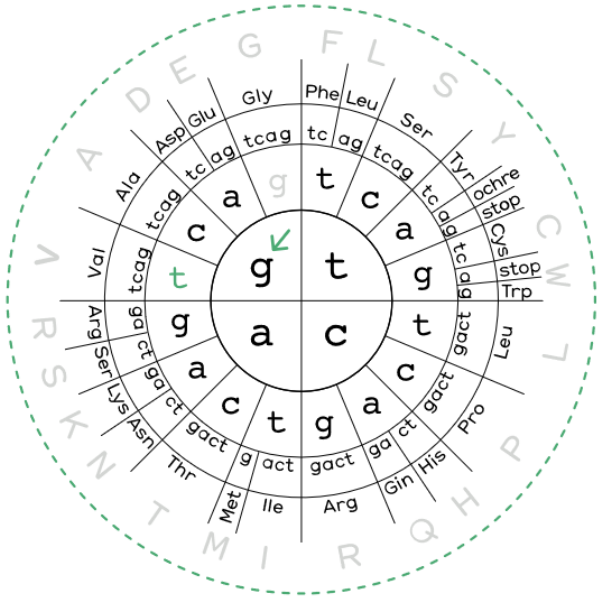


How SARS-CoV-2 enters human cells *(after* ***https://viralzone.expasy.org/9096****): the spike protein located on the surface of the virion interacts with the ACE2 protein (entry receptor) located on some human cells, allowing the virion to enter via two different mechanisms. When another human protein, the protease TMPRSS2 is present at the cell surface, the virion membrane can fuse directly and release the viral RNA genome into the cell. If no human protease activates the spike protein, the virion can be taken up into an endosome in which the human protein called cathepsin L activates the spike to allow fusion and release of the viral RNA genome into the cell.*

The spike gene & protein sequence (1)

Gene S coding for the Spike protein (beginning of the sequence):

```
  atgtttgt ttttcttgtt ttattgccac tagtctctag tcagtgtgtt
aatcttacaa ccagaactca attacccccct gcatacacta attctttcac
acgtggtgtt tattaccctg acaaagtttt cagatcctca gttttacatt
caactcagga cttgttctta cctttctttt ccaatgttac ttggttccat
gctatacatg tctctgggac caatggtact aagaggtttg ataaccctgt
cctaccattt aatgatggtg tttattttgc ttccactgag aagtctaaca
taataagagg ctggattttt ggtactactt tagattcgaa gacccagtcc
ctacttattg ttaataacgc tactaatgtt …
```



(Genetic code)

What are the first amino acids in the Spike protein?

atg ttt gtt ttt ctt gtt tta ttg …

M F V … … … … …

The spike gene & protein sequence (2)

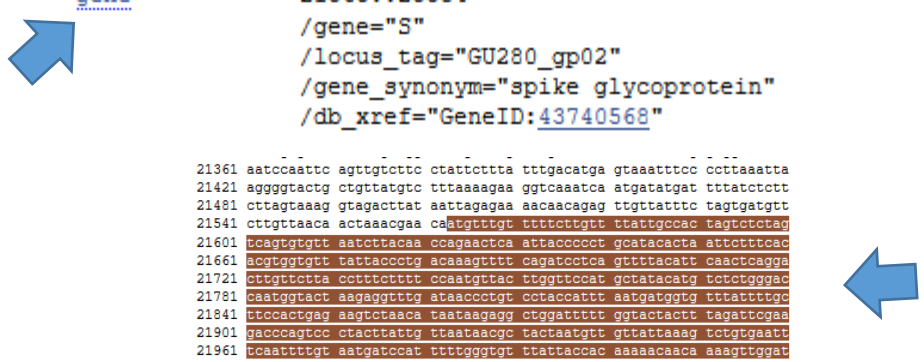
(1) Here is the complete sequence of the gene S coding for the Spike protein of SARS-CoV-2.

To translate this nucleotide sequence in an amino acid sequence, use Translate@Expasy .

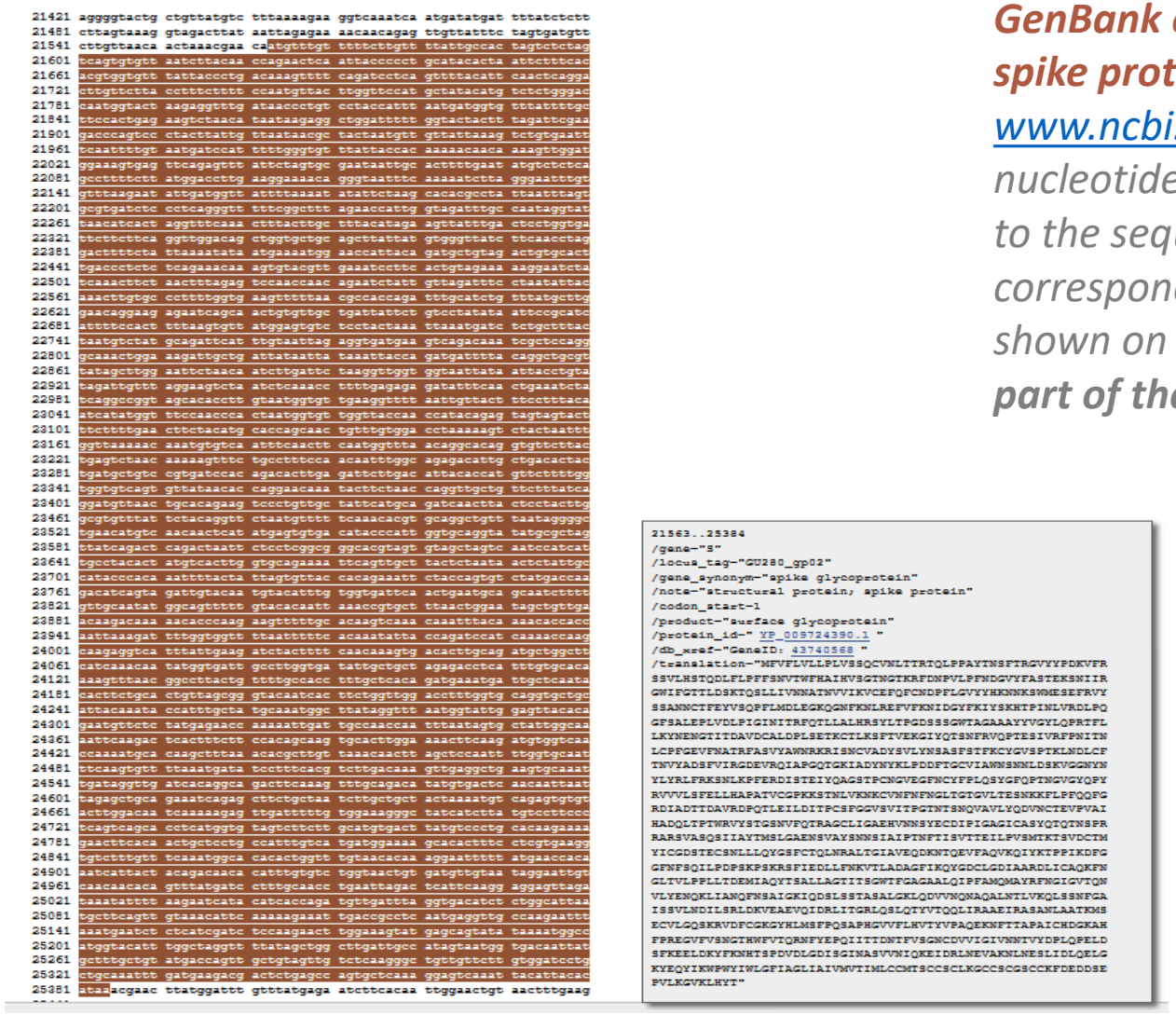
The amino acid sequence of the Spike protein is found in 'Frame 1'.

(2) Locate the gene coding for the Spike protein in the SARS-CoV-2 genome:

- From: https://www.ncbi.nlm.nih.gov/nuccore/NC_045512.2
- Look for 'Spike' (use Ctrl F or command F on mac)
- Click on 'gene'



About the RNA vaccine



```
21563..25384
/gene="S"
/locus_tag="GU280_gp02"
/gene_synonym="spike glycoprotein"
/note="structural protein; spike protein"
/codon_start=1
/product="surface glycoprotein"
/protein_id=" YP_009724390.1 "
/db_xref="GeneID: 43740568 "
/translation="MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFR
SSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFDNPVLPFNDGVYFASTEKSNIIR
GWIFGTTLDSKTQSLLIVNNATNVVIKVCEFQFCNDPFLGVYYHKNNKSWMESEFRVY
SSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRDLPQ
GFSALEPLVDLPIGINITRFQTLLALHRSYLTPGDSSSGWTAGAAAYYVGYLQPRTFL
LKYNENGTITDAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITN
LCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCF
TNVYADSFVIRGDEVRQIAPGQTGKIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYN
YLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVEGFNCYFPLQSYGFQPTNGVGYQPY
RVVVLSFELLHAPATVCGPKKSTNLVKNKCVNFNFNGLTGTGVLTESNKKFLPFQQFG
RDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQDVNCTEVPVAI
HADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICASYQTQTNSPR
RARSVASQSIIAYTMSLGAENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTM
YICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFG
GFNFSQILPDPSKPSKRSFIEDLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFN
GLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQN
VLYENQKLIANQFNSAIGKIQDSLSSTASALGKLQDVVNQNAQALNTLVKQLSSNFGA
ISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMS
ECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPAICHDGKAH
FPREGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELD
SFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELG
KYEQYIKWPWYIWLGFIAGLIAIVMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSE
PVLKGVKLHYT"
```

Part of the reference genome sequence of SARS-CoV-2 in GenBank entry corresponding to the gene S coding for the spike protein

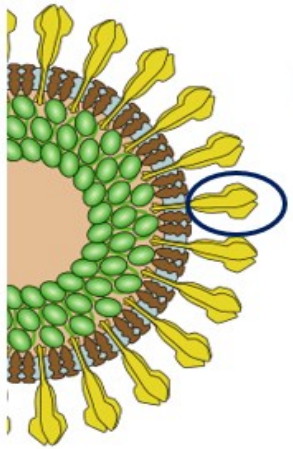
www.ncbi.nlm.nih.gov/nuccore/NC_045512.2?report=genbank: nucleotides 21,563 to 25,384, highlighted in brown, correspond to the sequence of gene S, which encodes the spike protein. The corresponding amino acid sequence of the spike protein is shown on the right. ***The nucleotides highlighted in brown are part of the RNA sequence found in the RNA vaccines.***

Reverse Engineering the source code of the BioNTech/Pfizer SARS-CoV-2 Vaccine

Dec 25 2020 *20 mins read*

«the source code of the BioNTech/Pfizer SARS-CoV-2 vaccine»

The spike protein sequence

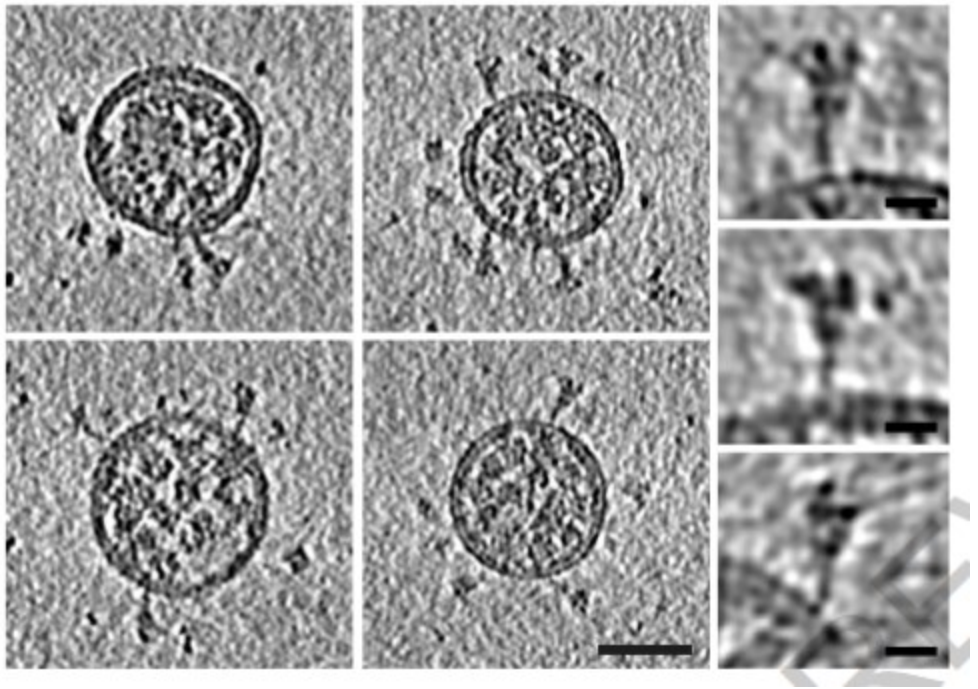


Discover SARS-CoV-2's amino acid sequence of the Spike protein in the UniProtKB/Swiss-Prot database.

What is the length (number of amino acids) of the Spike protein?

The spike protein 3D structure

First Cryo-EM images of the Spike protein on the surface of the virion.



https://www.nature.com/articles/s41586-020-2665-2_reference.pdf

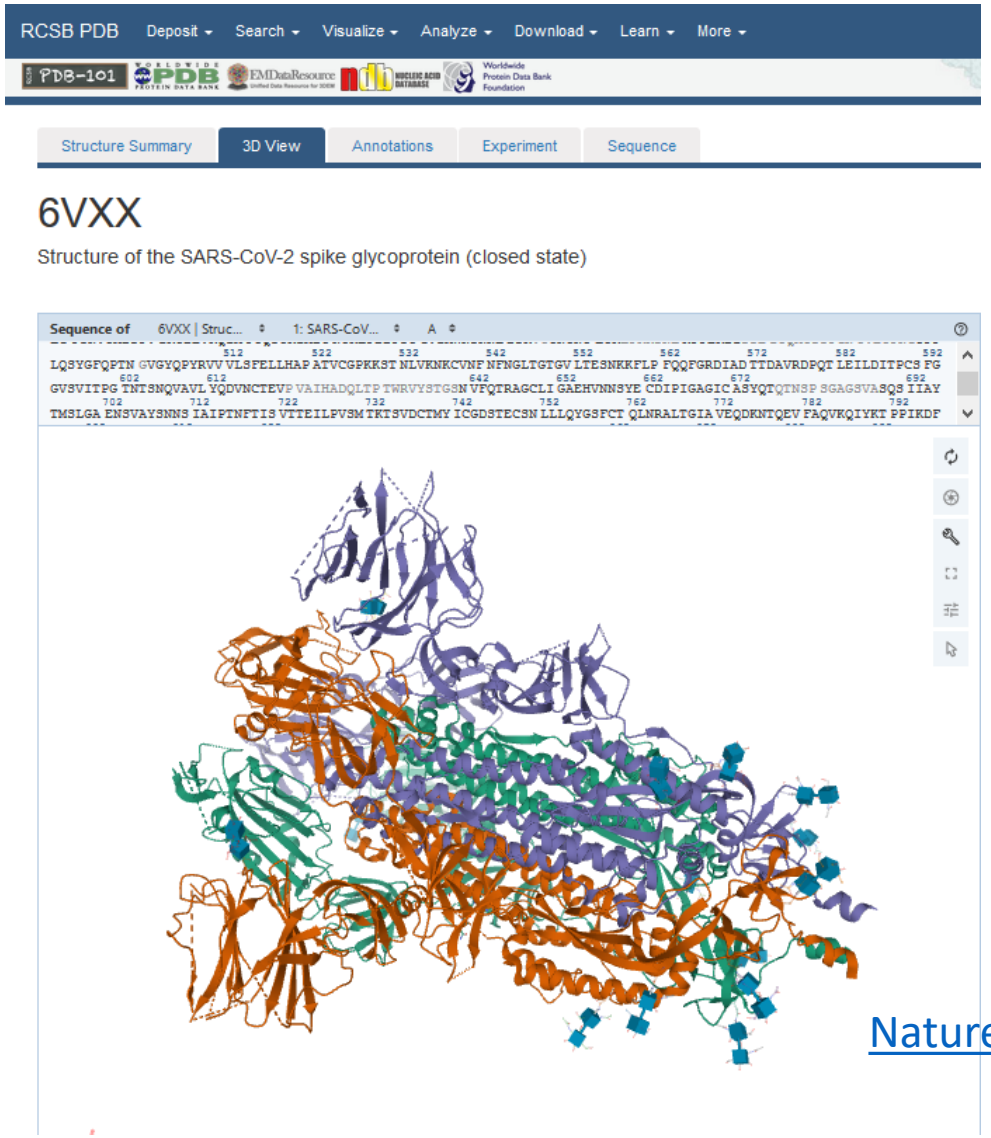
The spike protein 3D structure (PDB) (1)

The spike protein forms a trimer. This is the 3D structure of the spike protein trimer (PDB 6VXX). Each of the 3 chains has a different color (green, blue, red).

Follow this link:
https://www.rcsb.org/3d-view/6vxx

Click on an amino acid to see where it is located in the 3D structure.

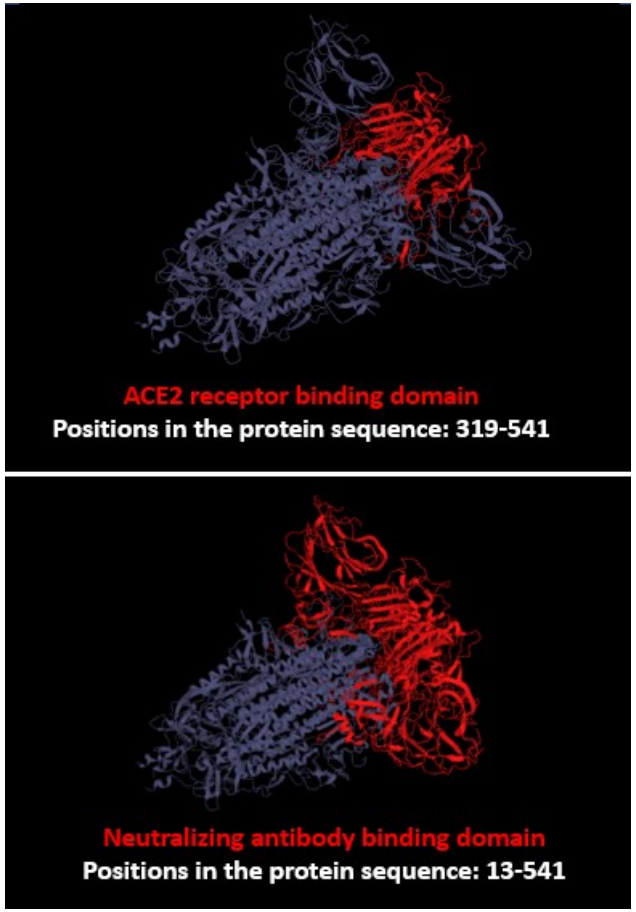
- Expert: compare the 2 structures of Spike
  - 6vxx: closed
  - 6vsb: open



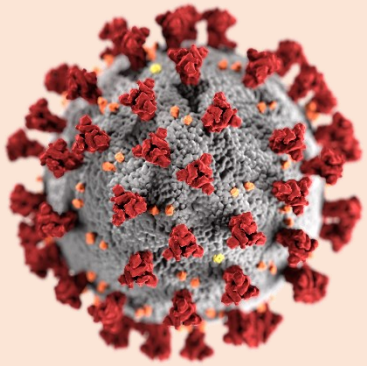
Nature, 2020

The spike protein 3D structure (2)

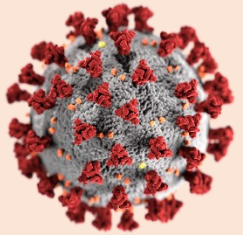
Knowledge of the 3D structure of the protein allows us to study the domain involved in the interaction of spike with the human ACE2 receptor. And also to study the part of the spike protein which is recognized by neutralizing antibodies.



(Source: viralzone.expasy.org/9556)



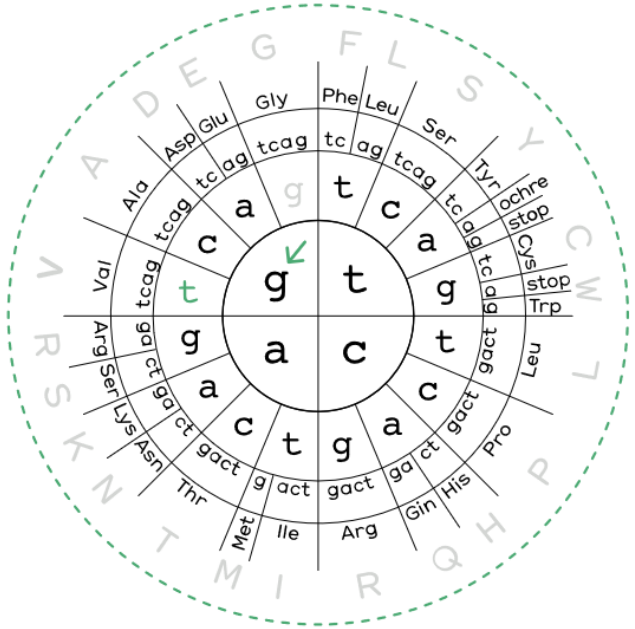
7 – The spike protein: impacts of mutations



Certain mutations in the genome sequence can alter the amino acid sequence of the corresponding protein.
Example with the gene coding for the Spike protein:

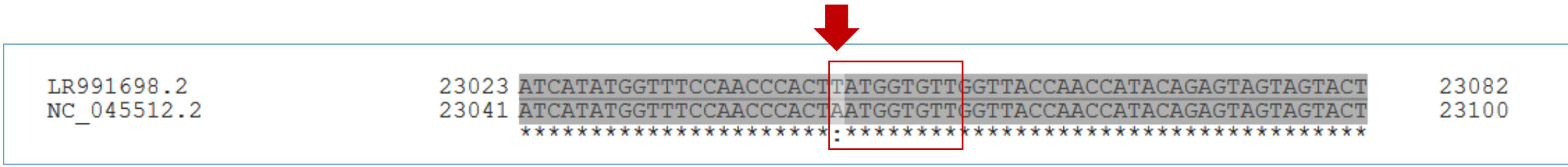
```
LR991698.2        23023 ATCATATGGTTTCCAACCCACTTATGGTGTTGGTTACCAACCATACAGAGTAGTAGTACT 23082
NC_045512.2       23041 ATCATATGGTTTCCAACCCACTAATGGTGTTGGTTACCAACCATACAGAGTAGTAGTACT 23100
                        **********************:**********************************
```

```
Genome 'UK' (LR991698.2)          tat ggt gtt
Protein Spike variant alpha        Y   G   V

Genome reference (NC_045512.2)    aat ggt gtt
Protein Spike reference            …   G   V

Genome 'XX'                       aac ggt gtt
Protein Spike XX                   …   G   V

                                … 501 502 503 …
```



Genetic code

Certain mutations in the genome sequence can alter the amino acid sequence of the corresponding protein.
Example with the gene coding for the Spike protein:



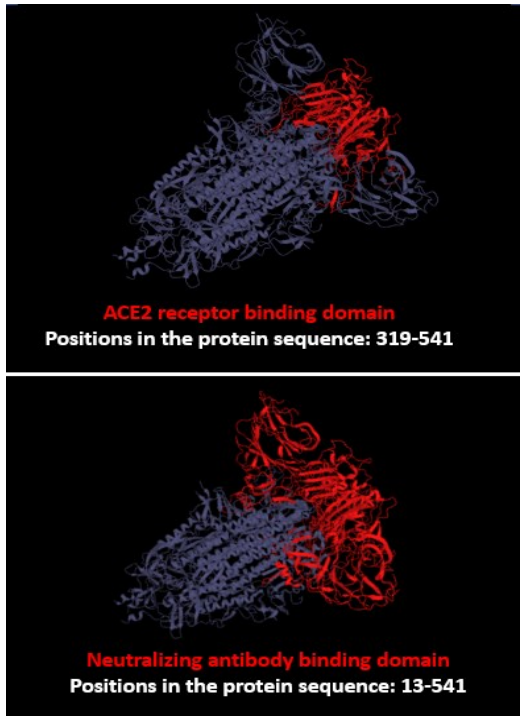
Genome 'UK' (LR991698.2)	tat	ggt	gtt
Protein Spike variant alpha	**Y**	**G**	**V**
Genome reference (NC_045512.2)	aat	ggt	gtt
Protein Spike reference	**N**	**G**	**V**
Genome 'XX'	aac	ggt	gtt
Protein Spike XX	**Y**	**G**	**V**
	… 501	502	503 …

The mutation tat -> aat (at position 23,063 in the virus genome) is located in the spike gene and changes the protein sequence.
The Spike protein is composed of 1273 amino acids.
The mutation is in position 501 in the protein sequence: it is called **N501Y** *(Nelly).*
...and the variant Alpha is also called B.1.1.7 (501Y.V1).

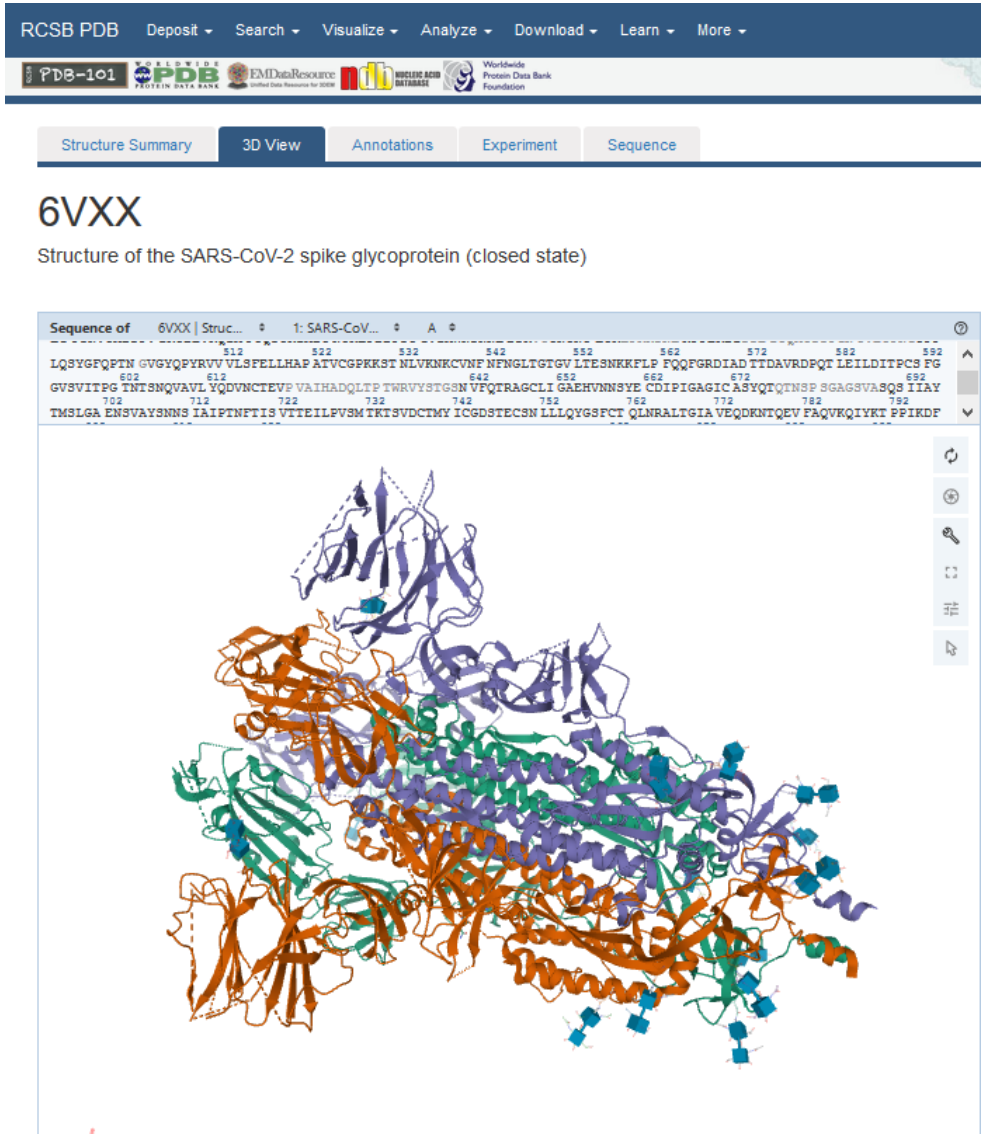
Note that the mutation aat -> aac does not change the protein sequence because of the genetic code redundancy.

Follow this link:
https://www.rcsb.org/3d-view/6vxx

Locate the mutation: N501Y



(Source: viralzone.expasy.org/9556)



Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1, complete genome

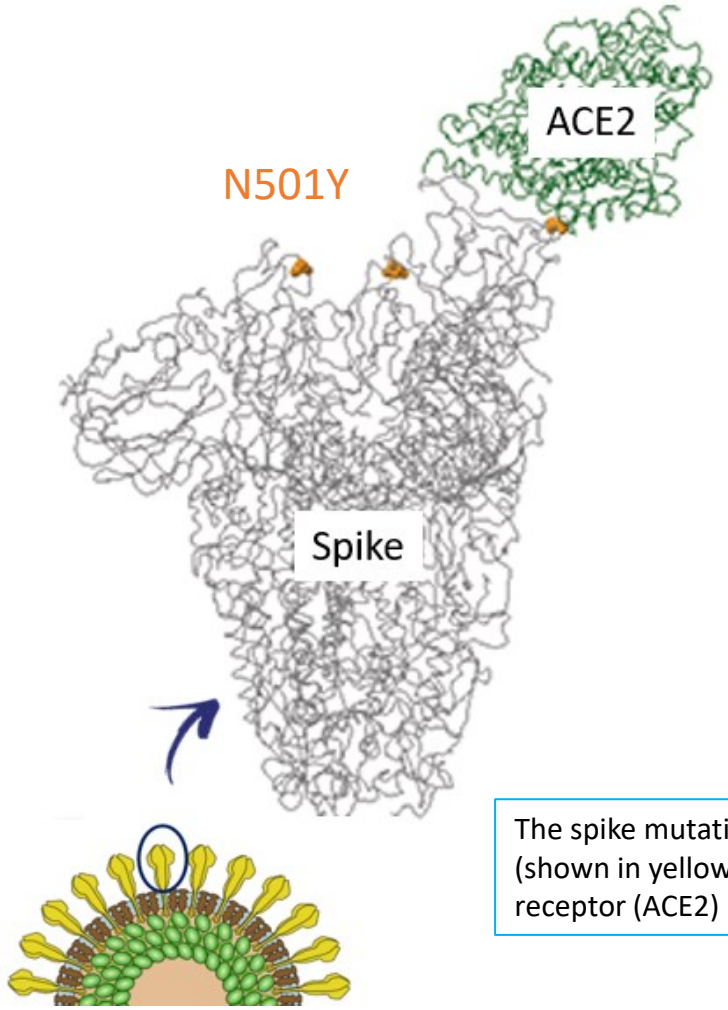
NCBI Reference Sequence: NC_045512.2

GenBank Graphics

>NC_045512.2 Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1, complete genome

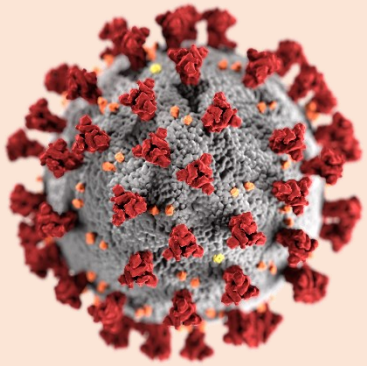
ATTAAAGGTTTATACCTTCCCAGGTAACAAACCAACCAACTTTCGATCTCTTGTAGATCTGTTCTCTAAA
CGAACTTTAAAATCTGTGTGGCTGTCACTCGGCTGCATGCTTAGTGCACTCACGCAGTATAATTAATAAC
TAATTACTGTCGTTGACAGGACACGAGTAACTCGTCTATCTTCTGCAGGCTGCTTACGGTTTCGTCCGTG
TTGCAGCCGATCATCAGCACATCTAGGTTTCGTCCGGGTGTGACCGAAAGGTAAGATGGAGAGCCTTGTC
CCTGGTTTCAACGAGAAAACACACGTCCAACTCAGTTTGCCTGTTTTACAGGTTCGCGACGTGCTCGTAC
GTGGCTTTGGAGACTCCGTGGAGGAGGTCTTATCAGAGGCACGTCAACATCTTAAAGATGGCACTTGTGG
CTTAGTAGAAGTTGAAAAAGGCGTTTTGCCTCAACTTGAACAGCCCTATGTGTTCATCAAACGTTCGGAT
GCTCGAACTGCACCTCATGGTCATGTTATGGTTGAGCTGGTAGCAGAACTCGAAGGCATTCAGTACGGTC
GTAGTGGTGAG
TCTTCGTAAGA
GGCGACGAGCT
TTACCCGTGAA
CCCTGATGGCT
TCCGAACAACT
CTTGGTACACG
ATTTGACACCT
CCAAGGGTTGA
CAAATGAATGC
GACGGGCGATT
ACTTGTGGTTA
GACCTGAGCAT
CACTATTGCCT
CGTGCTAGCGC
ACCTTCTTGAA
GATCGCCATTA
TATAAAGCATT

```
NC_045512.2       TAGATTGTTTAGGAAGTCTAATCTCAAACCTTTTGAGAGAGATATTTCAACTGAAATCTA   22980
LR991698.2        TAGATTGTTTAGGAAGTCTAATCTCAAACCTTTTGAGAGAGATATTTCAACTGAAATCTA   22962
                  ************************************************************

NC_045512.2       TCAGGCCGGTAGCACACCTTGTAATGGTGTTGAAGGTTTTAATTGTTACTTTCCTTTACA   23040
LR991698.2        TCAGGCCGGTAGCACACCTTGTAATGGTGTTGAAGGTTTTAATTGTTACTTTCCTTTACA   23022
                  ************************************************************

NC_045512.2       ATCATATGGTTTCCAACCCACTAATGCTGTTGGTTACCAACCATACAGAGTAGTAGTACT   23100
LR991698.2        ATCATATGGTTTCCAACCCACTTATGGTGTTGGTTACCAACCATACAGAGTAGTAGTACT   23082
                  ********************* ************************************

NC_045512.2       TTCTTTTGAACTTCTACATGCACCAGCAACTGTTTGTGGACCTAAAAAGTCTACTAATTT   23160
LR991698.2        TTCTTTTGAACTTCTACATGCACCAGCAACTGTTTGTGGACCTAAAAAGTCTACTAATTT   23142
                  ************************************************************

NC_045512.2       GGTTAAAAACAAATGTGTCAATTTCAACTTCAATGGTTTAACAGGCACAGGTGTTCTTAC   23220
LR991698.2        GGTTAAAAACAAATGTGTCAATTTCAACTTCAATGGTTTAACAGGCACAGGTGTTCTTAC   23202
                  ************************************************************

NC_045512.2       TGAGTCTAACAAAAAGTTTCTGCCTTTCCAACAATTTGGCAGAGACATTGCTGACACTAC   23280
LR991698.2        TGAGTCTAACAAAAAGTTTCTGCCTTTCCAACAATTTGGCAGAGACATTGATGACACTAC   23262
                  ************************************************** **********

NC_045512.2       TGATGCTGTCCGTGATCCACAGACACTTGAGATTCTTGACATTACACCATGTTCTTTTGG   23340
LR991698.2        TGATGCTGTCCGTGATCCACAGACACTTGAGATTCTTGACATTACACCATGTTCTTTTGG   23322
                  ************************************************************
```

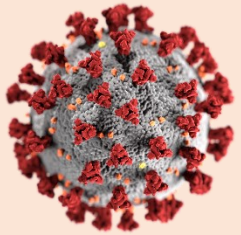


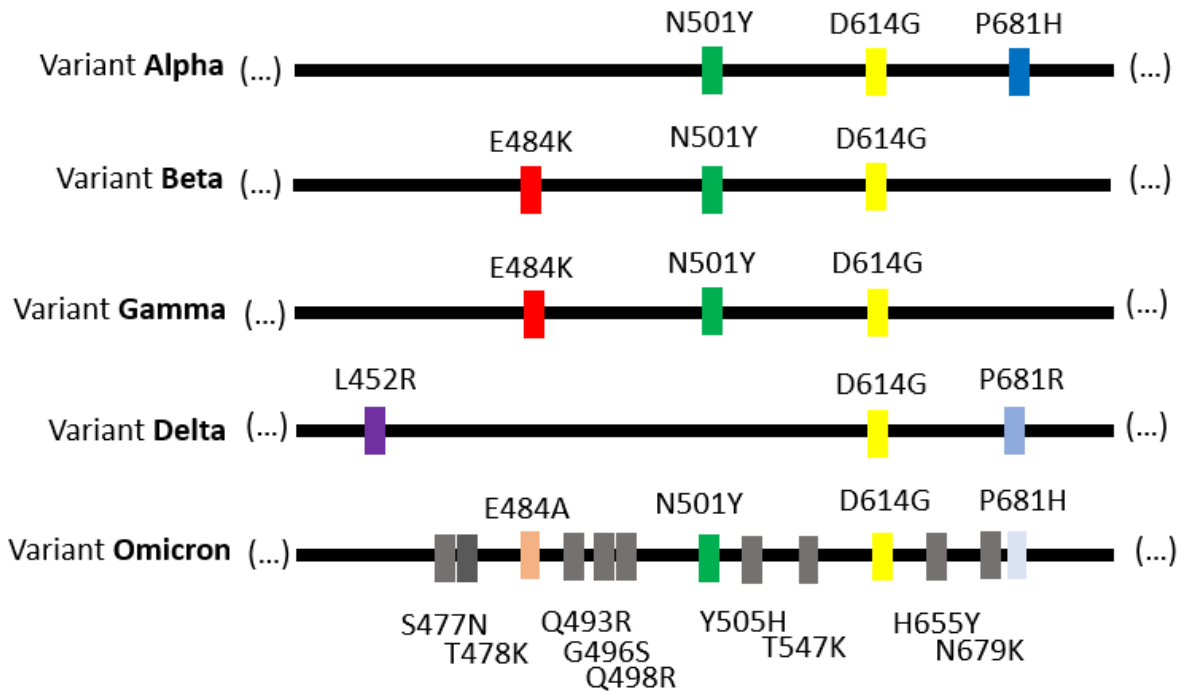
The spike mutation **N501Y** affects amino acids (shown in yellow), which binds to the human receptor (ACE2) (green).

https://www.science.org/doi/full/10.1126/science.371.6524.9



8 – The spike mutations & the SARS-CoV-2 variants





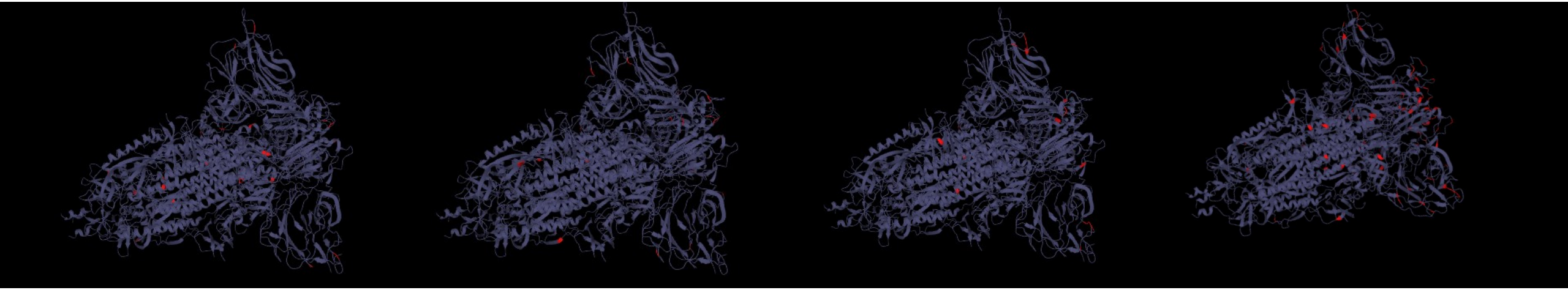
Source

Representation of the combination of some spike mutations *found in different well-known virus variants. This combination of mutations can be seen as a code bar, which allows to identify the different virus variants (Source: viralzone.expasy.org/9556).*

The N501Y mutation is found in several variants. This mutation may help the virus spread more easily. The E484K mutation, also found in several variants, may affect the antibody response.

Positions of the different mutations (in red) in the Spike 3D structure

https://viralzone.expasy.org/9556

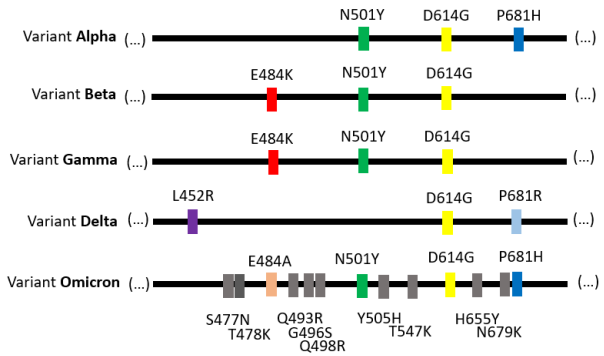


Alpha

Beta
~10 mutations
February 2021

Delta
~10 mutations
June 2021

Omicron
~30 mutations
November 2021



1’273 acides aminés

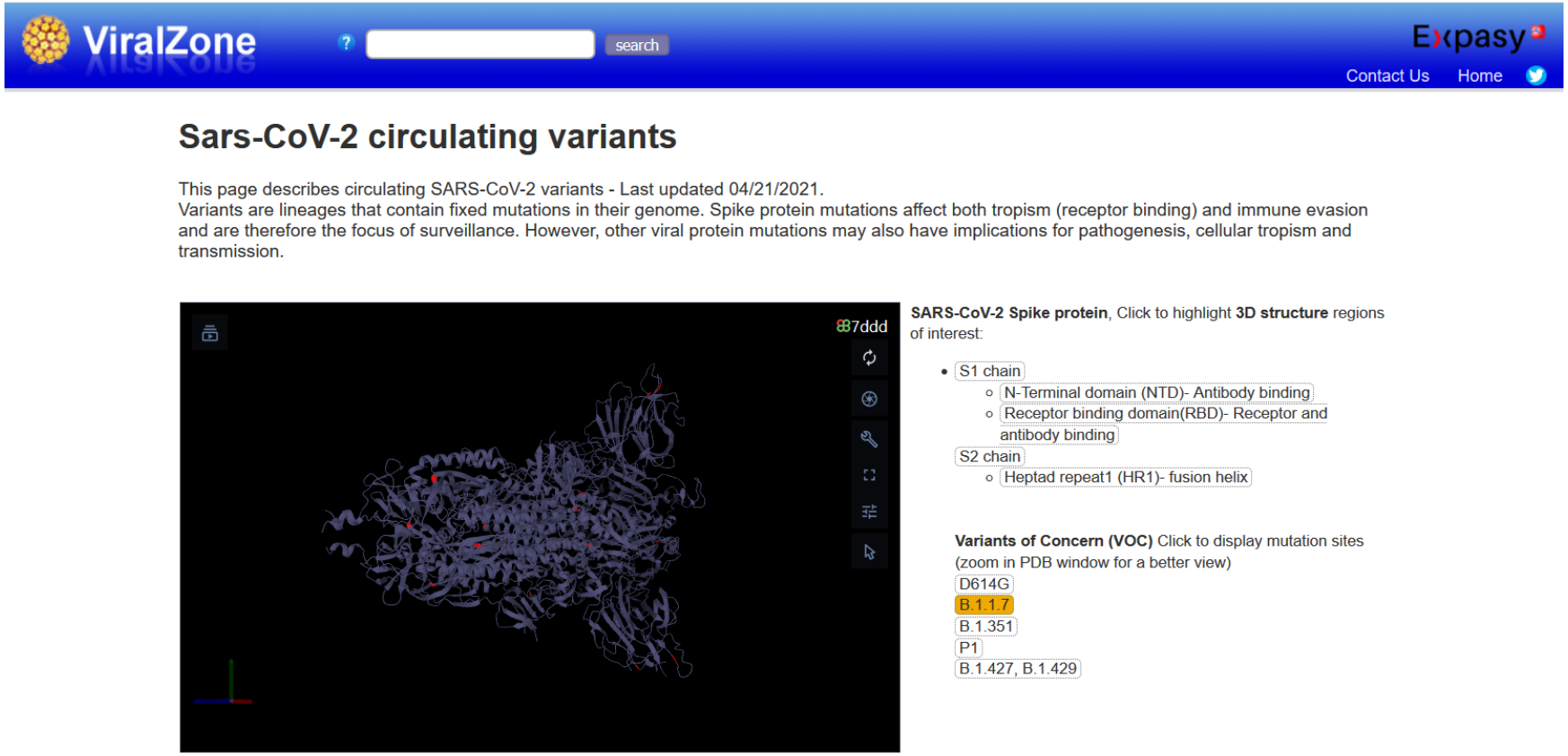
Part of a table listing mutations found in the spike protein sequence for different well-known virus variants *(Source: viralzone.expasy.org/9556). The biologically important regions of the spike protein (such as neutralizing antibody binding and ACE2 receptor binding) are indicated with the respective purple (NTD) & blue boxes (RBD).*

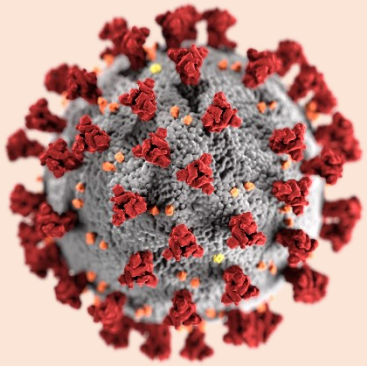
Localize the mutations in the spike 3D structure found in some important virus variants

https://viralzone.expasy.org/9556

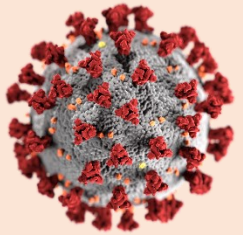
Click on the virus variant you are interested in (B.1.1.7 (Alpha), B.1.351 (Beta), P1 (Gamma), etc…).

The position of the mutations in the protein is highlighted in red. Note: spike forms a trimer…





9 – Inferring the origin of SARS-CoV-2

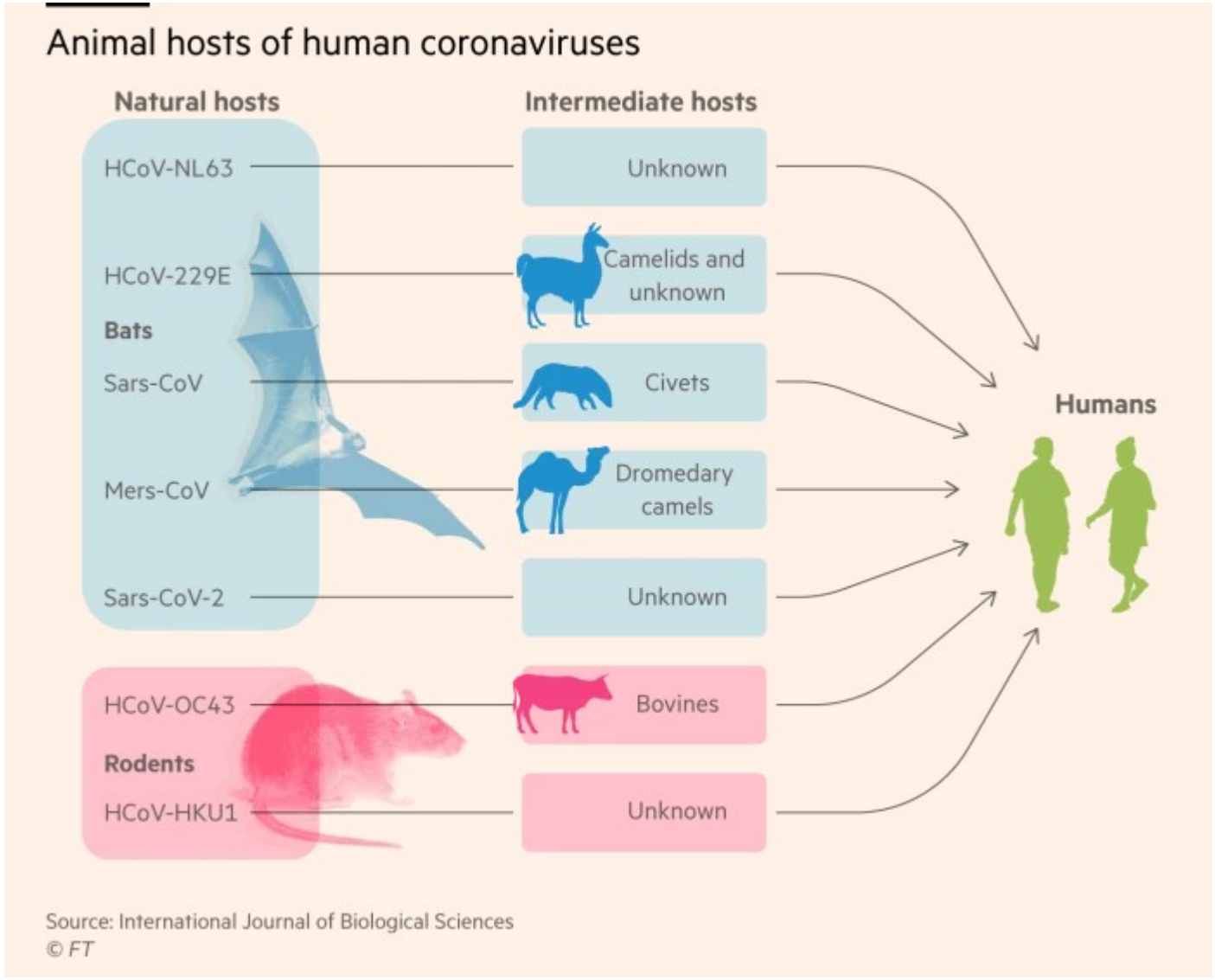


Nature examines arguments that the coronavirus SARS-CoV-2 escaped from a lab in China, and the science behind them.

Coronaviruses infect many mammalian species

SARS coronavirus (*human*)
Civet coronavirus
Hedgehog coronavirus
Bat coronavirus
Rabit coronavirus
Camel coronavirus
Dog coronavirus
Rat coronavirus
Bovine coronavirus
Equine coronavirus
Yak coronavirus
Pangolin coronavirus
Porcine coronavirus

The classification of viruses is particularly complex ...
It is not always easy to find your way around...



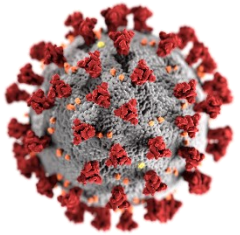
What bats can teach us about developing immunity to Covid-19 | Free to read

Efforts to develop effective drugs or vaccines depend on understanding how the virus outwits the immune system

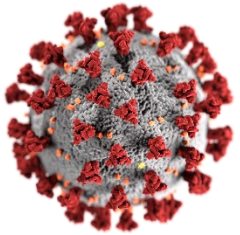
Financial Time

"Viruses love bats. (…)The big difference is that bats' remarkable immune system tames and tolerates many viruses that cause havoc when they spread to humans, including the coronavirus responsible for Covid-19.
(…) Coronaviruses have been evolving in bats for thousands or millions of years.
(…)Viruses are much more virulent when they spread to humans from bats than from other mammals," says Prof Crespi. "Yet they seem to do little harm to the bats themselves." "

Which coronavirus is most similar to SARS-CoV-2?



Compare the Spike protein sequence of different coronaviruses (bat, civet, pangolin, …)



Spike protein in different coronaviruses(1)

Here are partial sequences of the Spike protein from different coronaviruses infecting different species, at different times.

```
>Human_SARS2020
FSTFKCYGVSPTKLNDLCFTNVYADSFVIRGDEVRQIAPGQTGKIAD
>Civet_2003_coronavirus
FSTFKCYGVSATKLNDLCFSNVYADSFVVKGDDVRQIAPGQTGVIAD
>Pangolin_2020_coronavirus
FSTFKCYGVSPTKLNDLCFTNVYADSFVVRGDEVRQIAPGQTGRIAD
>Human_SARS2003
FSTFKCYGVSATKLNDLCFSNVYADSFVVKGDDVRQIAPGQTGVIAD
>Bat_2020_coronavirus
FSTFKCYGVSPTKLNDLCFTNVYADSFVITGDEVRQIAPGQTGKIAD
>Human_MERS2012
VNDFTCSQISPAAIASNCYSSLILDYFSYPLSMKSDLSVSSAGPISQ
>Bat_2007_coronavirus
VDEFSCNGISPDSIARGCYSTLTVDYFAYPLSMKSYIRPGSAGNIPL
```

Spike protein in different coronaviruses(2)

To compare the partial Spike protein sequences of different coronaviruses make alignments by pairs using Align@UniProt.

Fill in the following table :

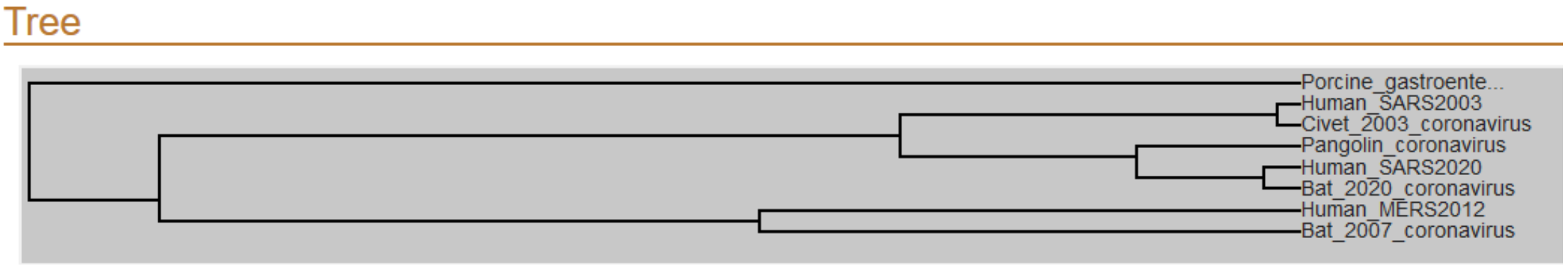
% identité	Human_SARS2020
Human_SARS2020	100
Human_SARS2003	
Civet_2003_coronavirus	
Pangolin_2020_coronavirus	
Bat_2020_coronavirus	
Human_MERS2012	
Bat_2007_coronavirus	

Answer:

Spike protein in different coronaviruses(3)

Copy/paste the complete Spike protein sequences from different coronavirus in Align@UniProt

Here is a very simplified representation of the evolutionary relationships existing between different coronaviruses infecting different mammalian species



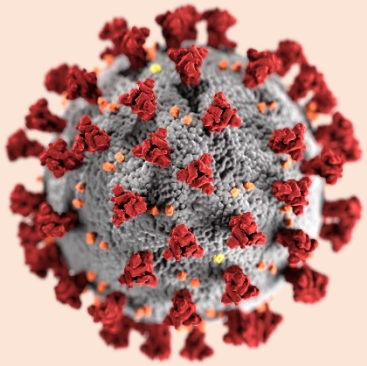
Which SARS-CoV-2 transmission chain(s) could be considered according to these results?

Human -> Human	1
Pig -> Human	2
Civet -> Human	3
Bat -> Human	4
Pangolin -> Human	5
Bat -> Pangolin -> Bat -> Human	6

Warning: these are hypotheses and not conclusions!
Conclusions on transmission chains are impossible to make with so little data!

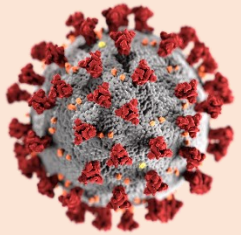
Sampling (presence or absence of a sequence (camel coronavirus 2012)) and sequencing errors can significantly influence the interpretation.

Answer: 1, 4, 5, 6



10 – Looking for a treatment …

Biology:
What’s a drug?



https://viralzone.expasy.org/9078

An illustration of the virus's infection cycle in a human cell and the main treatments under investigation: vaccine (neutralizing antibody) and drug molecules targeting different coronavirus biological pathways and proteins.

ViralZone

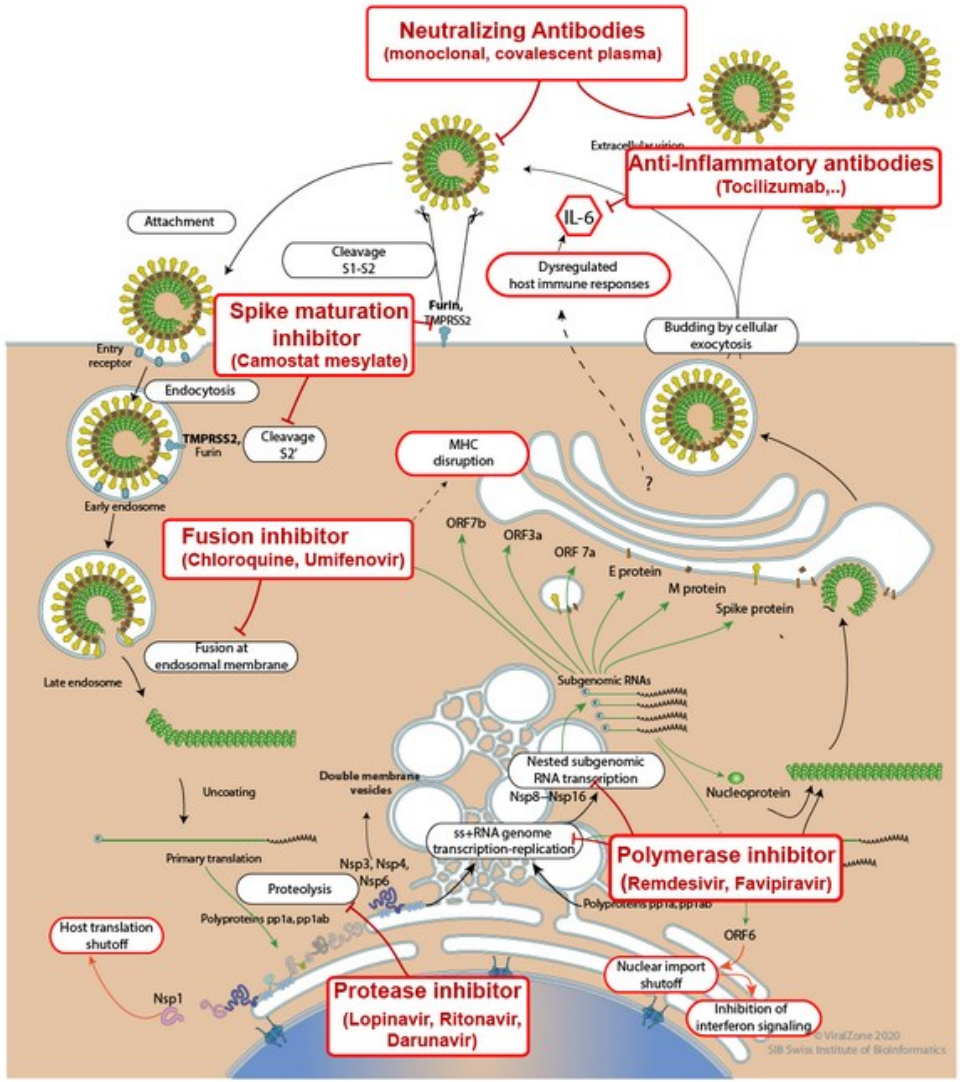
External resource

Antiviral drugs

There is no antiviral drug globally accepted for treatment, but many drugs are under investigations. About 21 trials of antivirals are under development (5th May 2020).
Antiviral drugs are difficult to find, because of the cell-parasitic nature of viruses. It has been long to get efficient drugs against HIV or Hepatitis C.
COVID-19 vaccine may arrive before we get efficient antivirals.

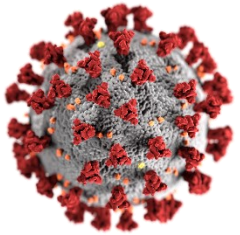
COVID-19 Treatment and Vaccine Tracker

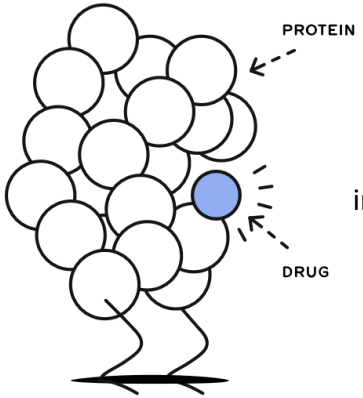
COVID19 Internet Book of Crit Care



ANTIVIRAL DRUGS This table displays the main drugs under investigation

Drugs ?

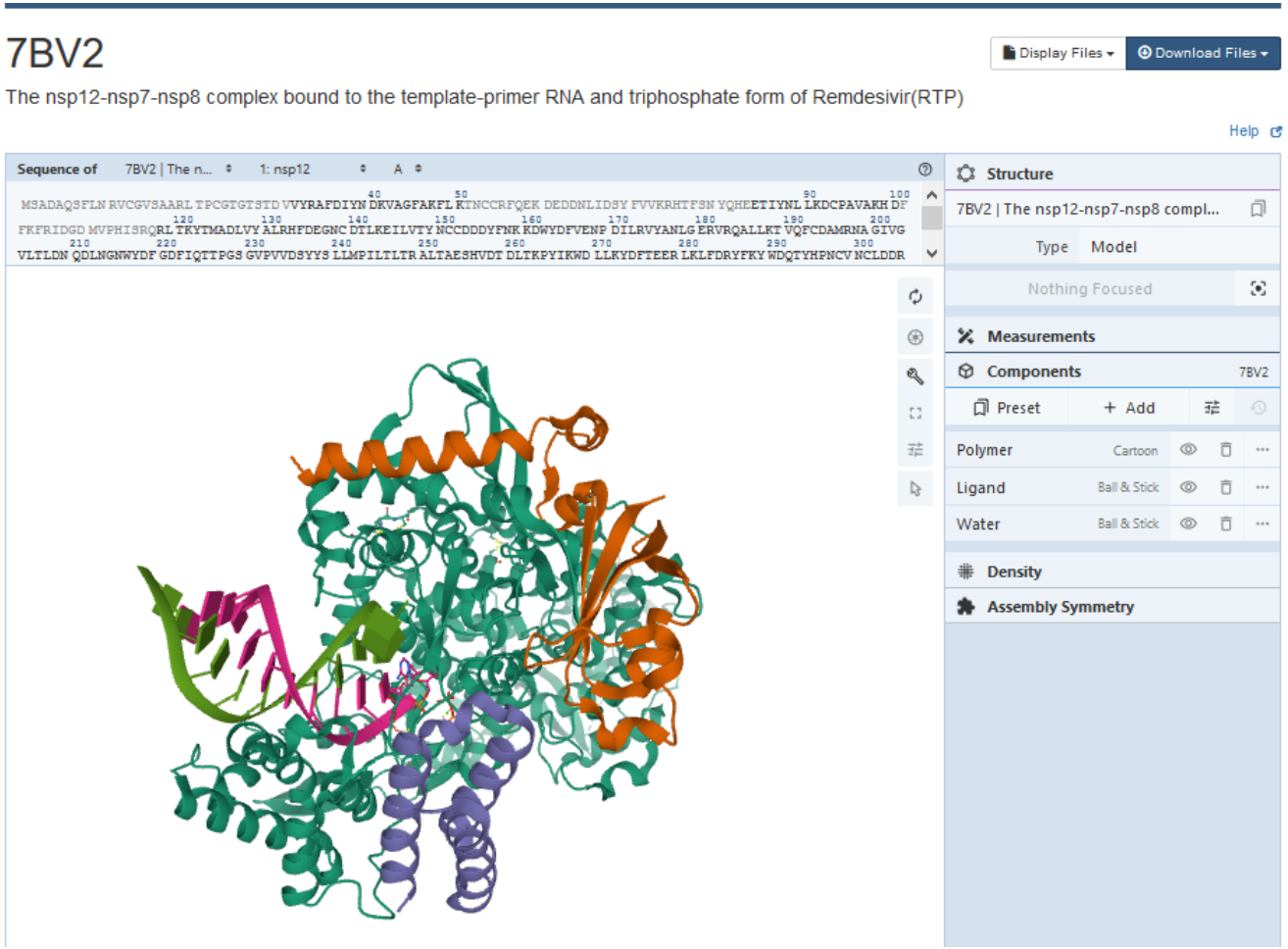


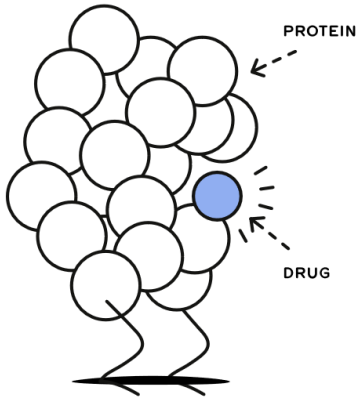


The protein-drug interaction is a bit like the interaction of a key with a lock: it depends very much on the shape of both the drug and the protein.

[3D Structure of the coronavirus polymerase] in presence of the Remdesivir drug ([Publication])

Observe the ‘duplicated’ RNA (‘double strand’) and the drug molecule

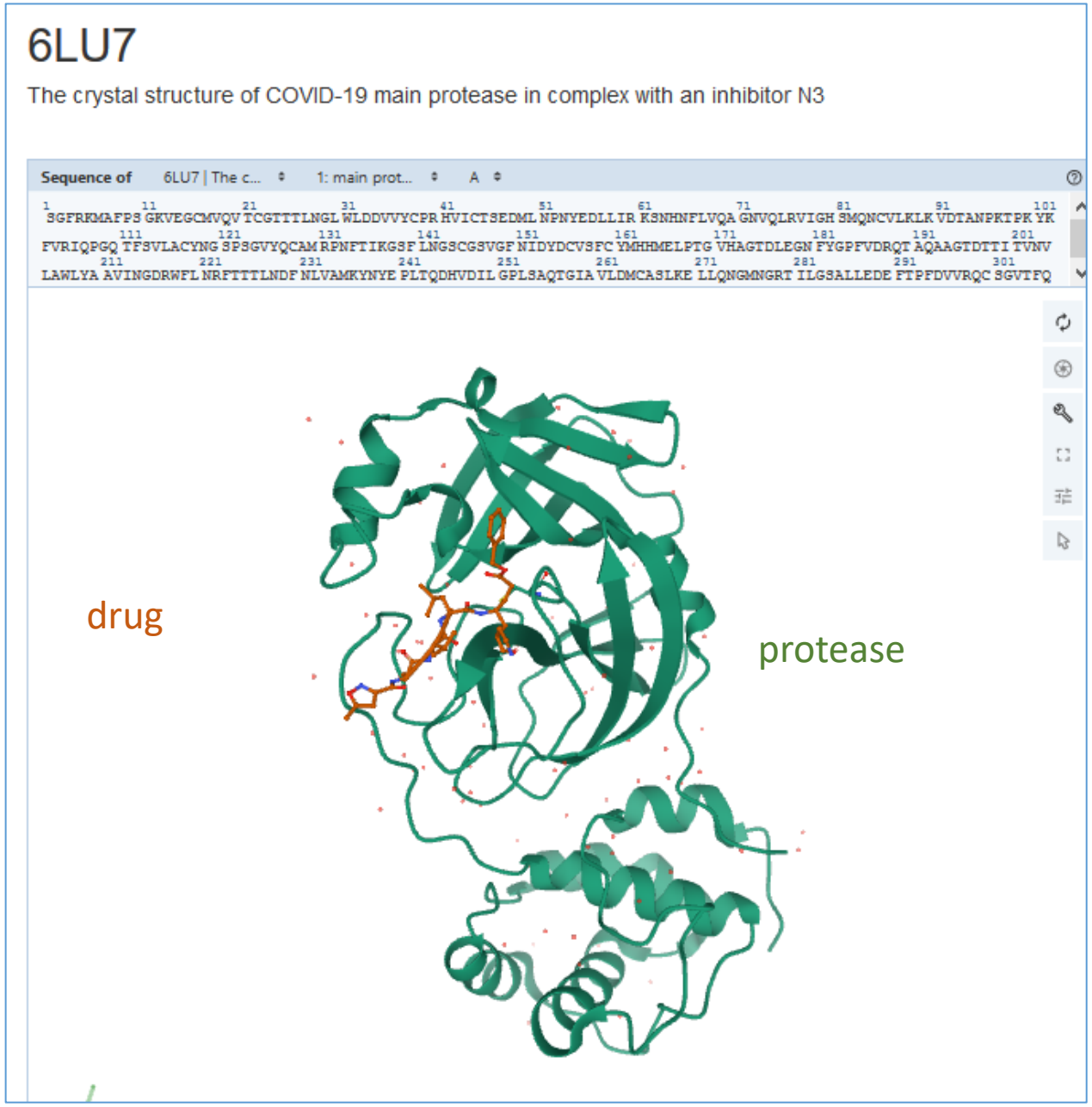




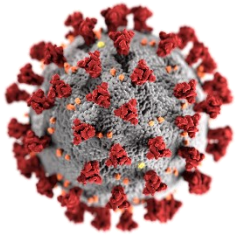
The protein-drug interaction is a bit like the teraction of a key with a lock: it depends very much on the shape of both the drug and the protein.

[3D Structure 3D of the coronavirus protease] in presence of a potential inhibitor([Publication])

- [The protease in UniProtKB/Swiss-Prot] (3C-like proteinase)



Vaccine ?



SARS-CoV-2 and Neutralizing Antibodies, 2020

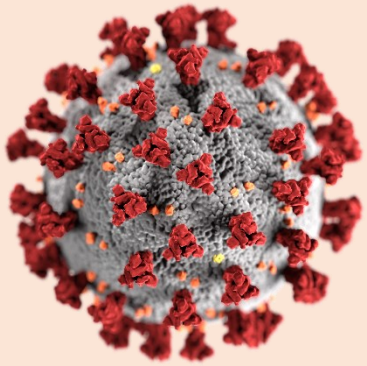
Download high quality TIFF image

http://pdb101.rcsb.org/sci-art/goodsell-gallery/sars-cov-2-and-neutralizing-antibodies

Acknowledgement: David S. Goodsell, RCSB Protein Data Bank and Springer Nature; doi: 10.2210/rcsb_pdb/goodsell-gallery-025

This painting shows a cross section through SARS-CoV-2 surrounded by blood plasma, with neutralizing antibodies in bright yellow. The painting was commissioned for the cover of a special COVID-19 issue of Nature, presented 20 August 2020.

It incorporates information from two cryoelectron microscopy studies that explore the shape and distribution of spikes and the nucleoprotein:
Yao H et al. (2020) Molecular architecture of the SARS-CoV-2 virus. bioRxiv preprint DOI: 10.1101/2020.07.08.192104
Ke Z et al. (2020) Structures, conformations and distributions of SARS-CoV-2 spike protein trimers on intact virions. bioRxiv preprint DOI: 10.1101/2020.06.27.174979

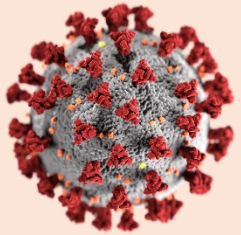


11 - An example of research carried out at the SIB Swiss Institute of Bioinformatics

Référence:

Christian Sigrist, Alan Bridge, Philippe Le Mercier

DOI:https://doi.org/10.1016/j.antiviral.2020.104759 (pdf)



The SARS-CoV-2 virus is a very close cousin of the SARS-CoV virus,
responsible for the 2003 epidemic,
which infected more than 8,000 people in 30 different countries.

Swiss Institute of Bioinformatics

Researchers at SIB came up with the idea of **comparing** the amino acid sequence of the Spike protein of different coronaviruses.

They compared the Spike protein of the SARS-CoV-2 coronavirus with the Spike protein of the SARS-CoV coronavirus, the virus that infected humans in 2003.

Compare 2 amino acid sequences:

```
>Spike_SARS-CoV-2
RISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFVIRGDEVRQIAPGQTG

>Spike_SARS-CoV
KISNCVADYSVLYNSTFFSTFKCYGVSATKLNDLCFSNVYADSFVVKGDDVRQIAPGQTG
```

You can do this manually or with the help of the bioinformatics tool Align@UniProt

How many differences ?

Do you find 3 consecutive 'RGD' amino acids in one of the sequences? Which one?

Compare the complete Spike protein sequence of SARS-CoV-2 and SARS-CoV

Spike_SARS-CoV-2: link
Spike_SARS-CoV: link

Copy / Paste the 2 sequences in Align@Uniprot

Look for the 3 consecutive amino acids RGD in the alignment (Crtl F)

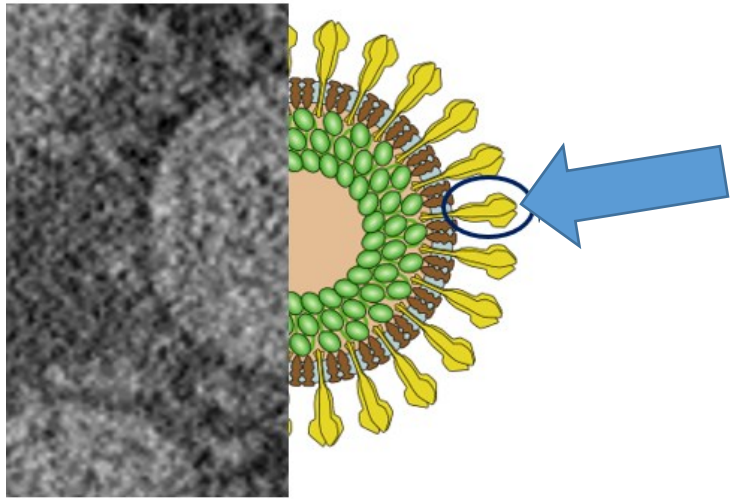
At which position in the SARS-CoV-2 sequence do you find it?

Swiss Institute of Bioinformatics

The researchers discovered the presence of an RGD motif in the sequence of the SARS-CoV-2 protein, a motif that is absent in the sequence of the SARS-CoV Spike protein.

These 3 consecutive amino acids are known to play an important role in virus biology: this motif, when located in the right place in the 3D structure of the protein, could allow the virus to enter human cells using not only ACE2, but also other proteins called **integrins**!

For the RGD motif to play a role in interaction with human cells, it must be on the surface of the Spike protein.

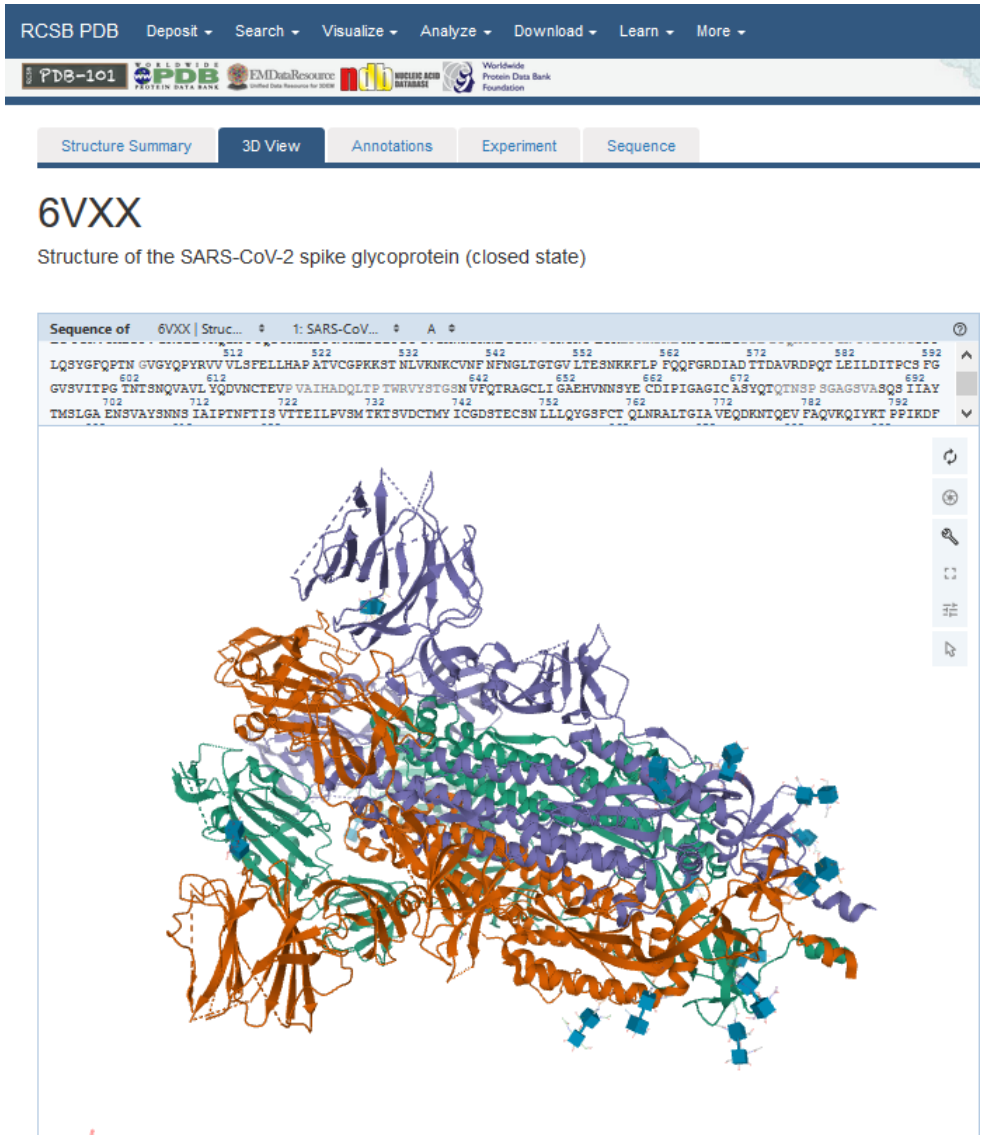


Is the RGD motif located on the surface of the Spike protein?

To answer this question, the 3D structure of the Spike protein must be studied.

Here is a representation of the 3D structure of the Spike protein and its amino acid sequence (databank PDB):

https://www.rcsb.org/3d-view/6vxx



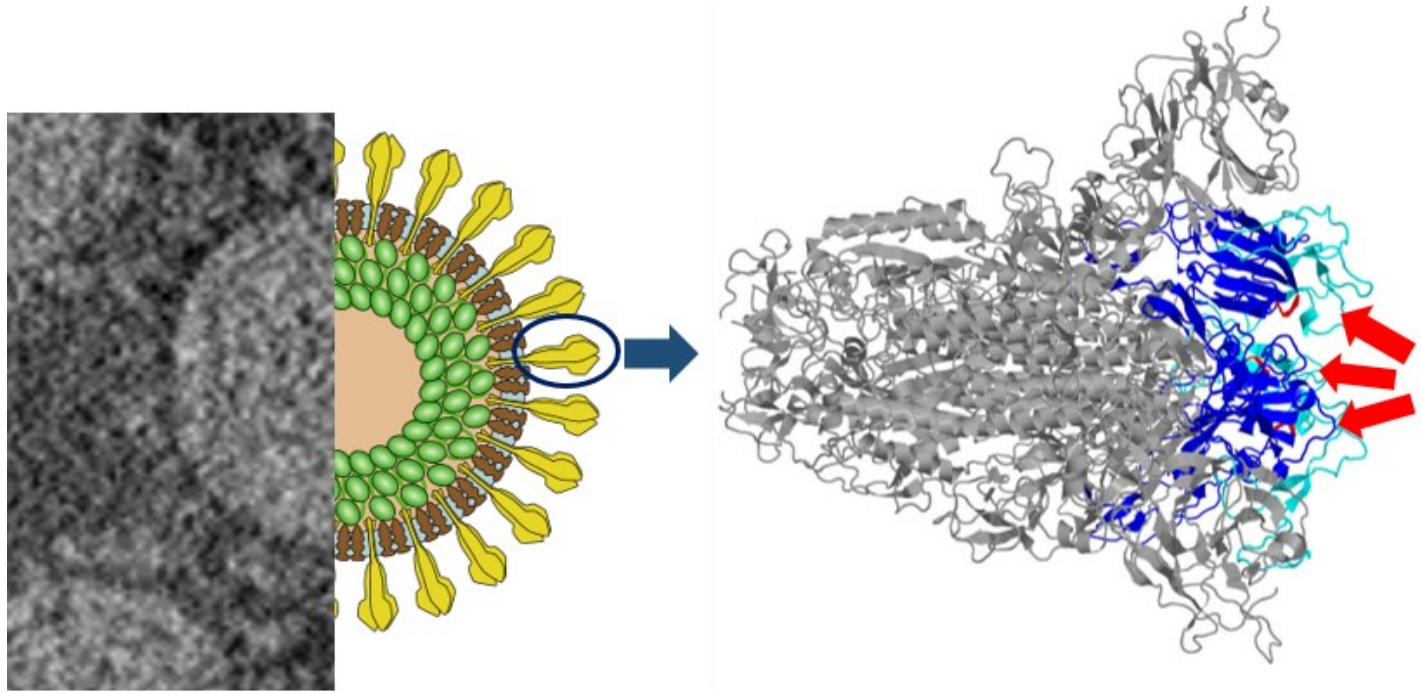
Look for the RGD motif in the 3D structure of the Spike protein

Note: the RGD motif is at position 403-405 in the protein sequence

Nature, 2020

The RGD motif in the 3D structure of the Spike protein

Using the data available for the 3D structure of the Spike S protein of SARS CoV-2, SIB researchers have shown that the RGD motif (in red) is located on the surface of the Spike S protein, close to the region involved in the interaction with the human receptor proteins (in blue).



SIB
Swiss Institute of Bioinformatics

Spike, RGD motif RGD and integrins

The RGD motif, found on the surface of the Spike S protein of the SARS CoV-2 virus, is known to promote interaction with other human proteins called integrins.

This RGD motif has been found in all the Spike S proteins of SARS-CoV-2 viruses that have been sequenced to date. It is possible that the coronavirus acquired this motif during its evolution and thus gained the ability to bind integrins to promote entry into host cells, but this remains to be proven.

Integrins are not expressed by the same cells as the ACE2 protein. Binding to integrins could therefore allow the virus to infect other cells and organs in addition to those expressing the ACE2 protein.

There are currently few antiviral molecules effective against SARS-CoV-2. Agents that block binding to integrins may be a promising avenue of investigation. Known integrin-binding blockers include the antibody natalizumab used for the treatment of multiple sclerosis/Crohn's disease or the small molecule tirofiban used for the treatment of acute coronary syndrome.

ELSEVIER

Antiviral Research

Volume 177, May 2020, 104759

A potential role for integrins in host cell entry by SARS-CoV-2

Christian JA Sigrist ✉, Alan Bridge ✉, Philippe Le Mercier ✉

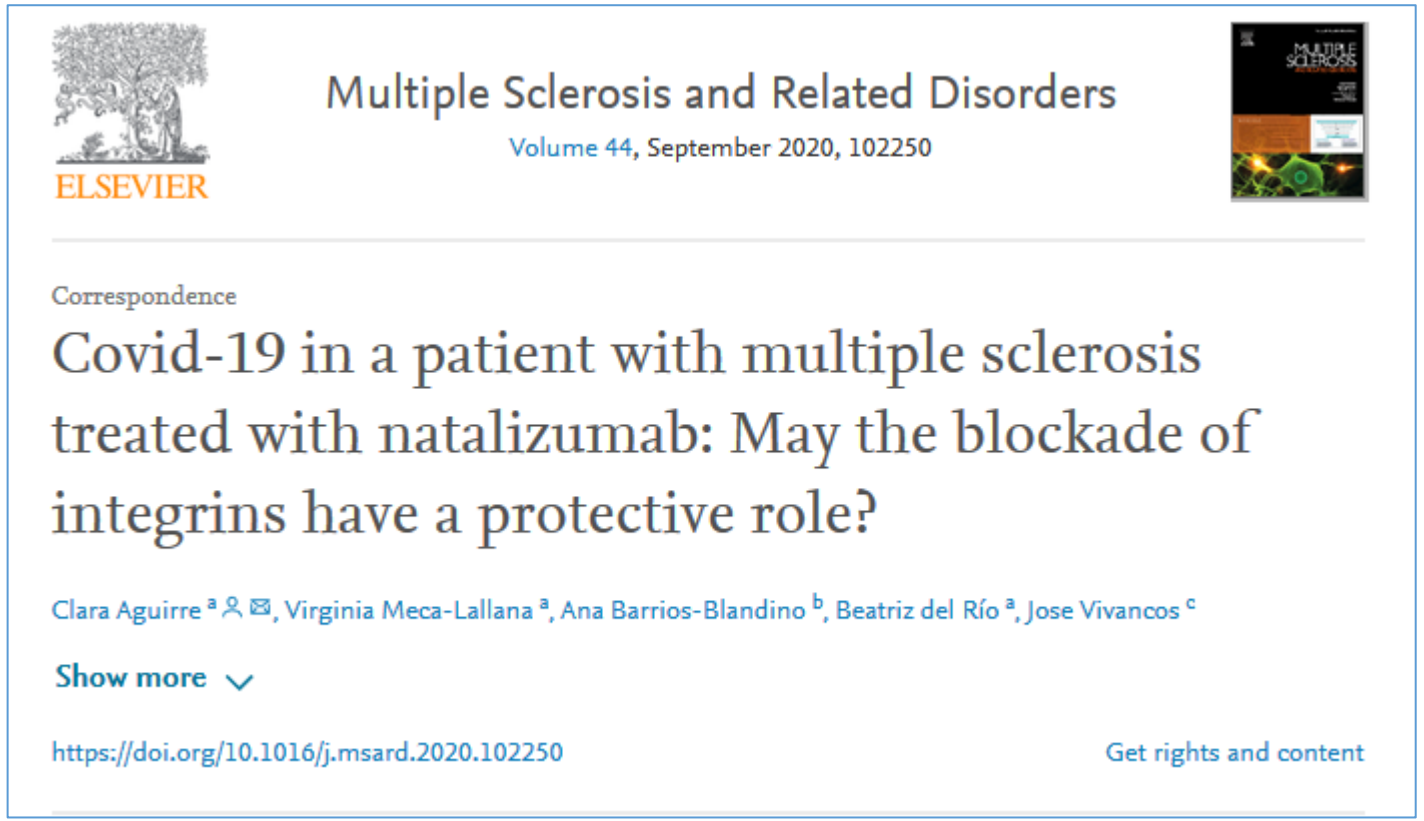
Swiss-Prot Group, SIB Swiss Institute of Bioinformatics, Switzerland

Received 20 February 2020, Revised 26 February 2020, Accepted 28 February 2020, Available online 1 March 2020.

https://doi.org/10.1016/j.antiviral.2020.104759

SIB
Swiss Institute of Bioinformatics

Thanks to this information, Spanish doctors continued to treat a Covid-19 patient with a drug targeting integrins.



ELSEVIER

Multiple Sclerosis and Related Disorders

Volume 44, September 2020, 102250

Correspondence

Covid-19 in a patient with multiple sclerosis treated with natalizumab: May the blockade of integrins have a protective role?

Clara Aguirre [a], Virginia Meca-Lallana [a], Ana Barrios-Blandino [b], Beatriz del Río [a], Jose Vivancos [c]

Show more

https://doi.org/10.1016/j.msard.2020.102250

Get rights and content

https://doi.org/10.1016/j.msard.2020.102250

Researchers are looking for a drug that blocks the interaction between Spike and integrins.

The Integrin Binding Peptide, ATN-161, as a Novel Therapy for SARS-CoV-2 Infection

Brandon J Beddingfield [1 2], Naoki Iwanaga [3], Prem P Chapagain [4 5], Wenshu Zheng [6], Chad J Roy [1 2], Tony Y Hu [6], Jay K Kolls [3], Gregory J Bix [7 8 9 10 11]

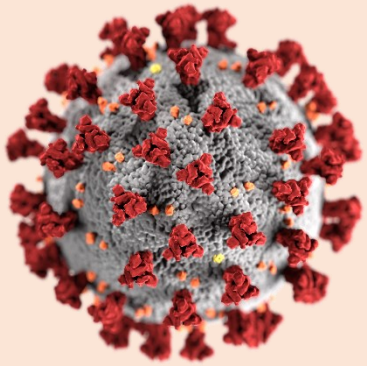
Affiliations + expand

PMID: 33102950 PMCID: PMC7566794 DOI: 10.1016/j.jacbts.2020.10.003

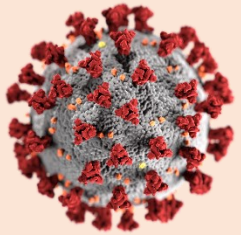
Free PMC article

Abstract

Many efforts to design and screen therapeutics for the current severe acute respiratory syndrome coronavirus (SARS-CoV-2) pandemic have focused on inhibiting viral host cell entry by disrupting ACE2 binding with the SARS-CoV-2 spike protein. This work focuses on the potential to inhibit SARS-CoV-2 entry through a hypothesized α5β1 integrin-based mechanism, and indicates that inhibiting the spike protein interaction with α5β1 integrin (+/- ACE2), and the interaction between α5β1 integrin and ACE2 using a novel molecule ATN-161 represents a promising approach to treat COVID-19.



12 - SARS-CoV-2 and HIV



SARS-CoV-2: a man-made virus?

FAKE NEWS **FACTS**

- SARS-Cov-2 is very similar to many strains of coronavirus circulating in nature in Asia before and after the pandemic.

- These viruses are known to jump from one species to another without any problem, so this is not surprising.

- The genetic analysis of SARS-CoV-2 shows a genomic organisation and proteins that are similar in all respects to other wild viruses.

The hypothesis that part of HIV was inserted into the virus is a misinterpretation of similarity searches (BLAST)

Uncanny similarity of unique inserts in the 2019-nCoV spike protein to HIV-1 gp120 and Gag

Prashant Pradhan, Ashutosh Kumar Pandey, Akhilesh Mishra, Parul Gupta, Praveen Kumar Tripathi, Manoj Balakrishnan Menon, James Gomes, Perumal Vivekanandan, Bishwajit Kundu

doi: https://doi.org/10.1101/2020.01.30.927871

Abstract | Info/History | Metrics | Preview PDF

Abstract

This paper has been withdrawn by its authors. They intend to revise it in response to comments received from the research community on their technical approach and their interpretation of the results. If you have any questions, please contact the corresponding author.

FAKE NEWS

FACTS

This paper has been withdrawn.
The authors used fragments of 6 to 10 amino acids to perform similarity searches (BLAST) (looking for SARS-CoV-2 spike protein sequence similar to HIV proteins): there are nearly 1 million HIV sequences located in hypervariable regions, so it is inevitable to find similarities just by chance. This paper, although it was never published, made a buzz ...

For experts (1):

Emerg Microbes Infect. 2020; 9(1): 378–381.
Published online 2020 Feb 14. doi: 10.1080/22221751.2020.1727299
PMCID: PMC7033698
PMID: 32056509

HIV-1 did not contribute to the 2019-nCoV genome

Chuan Xiao,[a,CONTACT] Xiaojun Li,[b] Shuying Liu,[c] Yongming Sang,[d] Shou-Jiang Gao,[e] and Feng Gao[b,f]

The following are examples of SARS-CoV-2 amino acid sequences used to supposedly 'demonstrate' that SARS-CoV-2 contains pieces of the HIV genome:

TNGTKR
HKNNKS
RSYLTPGDSSSG
QTNSPRRA

Do a ‘protein’ Blast @ NCBI: https://blast.ncbi.nlm.nih.gov/Blast.cgi
Select the organism: “Human immunodeficiency virus (taxid:12721)”
Click on BLAST 

Check the ‘E value’

The 'E value' is a probability of finding the same result by chance. The smaller this value (< 0), the more 'valid’ is the match.

Create a random sequence with the same 'letters' ([edit sequence]) and redo a BLAST.
What can you conclude?

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7033698/

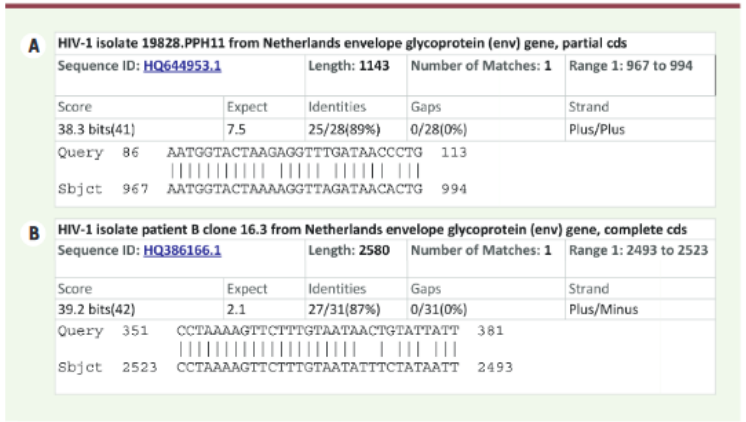
For experts (2):
This is a part of the gene coding for the Spike protein:
AATGGTACTAAGAGGTTTGATAACCCTG

Do a ‘nucleotide’ Blast @ NCBI: https://blast.ncbi.nlm.nih.gov/Blast.cgi.
Select the organism: “Human immunodeficiency virus (taxid:12721)”
Click on BLAST 

The 'E value' is a probability of finding the same result by chance. The smaller this value (< 0), the more 'valid’ is the match.

Create a random sequence with the same 'letters' ([edit sequence]) and redo a BLAST.

What can you conclude?



A HIV-1 isolate 19828.PPH11 from Netherlands envelope glycoprotein (env) gene, partial cds

Sequence ID: HQ644953.1	Length: 1143	Number of Matches: 1	Range 1: 967 to 994

Score	Expect	Identities	Gaps	Strand
38.3 bits(41)	7.5	25/28(89%)	0/28(0%)	Plus/Plus

```
Query  86   AATGGTACTAAGAGGTTTGATAACCCTG  113
            ||||||||||| ||||| |||||| |||
Sbjct  967  AATGGTACTAAAAGGTTAGATAACACTG  994
```

B HIV-1 isolate patient B clone 16.3 from Netherlands envelope glycoprotein (env) gene, complete cds

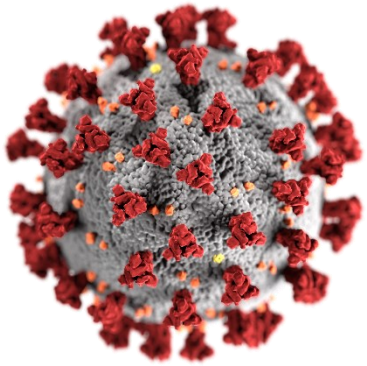
Sequence ID: HQ386166.1	Length: 2580	Number of Matches: 1	Range 1: 2493 to 2523

Score	Expect	Identities	Gaps	Strand
39.2 bits(42)	2.1	27/31(87%)	0/31(0%)	Plus/Minus

```
Query  351   CCTAAAAGTTCTTTGTAATAACTGTATTATT  381
              |||||||||||||||||||  | ||| |||
Sbjct  2523  CCTAAAAGTTCTTTGTAATATTTCTATAATT  2493
```

Figure 6. ***Recherche de similarités entre les séquences codant pour la protéine spike de CoV2 et le génome de HIV. A.*** Alignement le plus significatif entre la séquence codant pour la protéine S de SARS-CoV-2 (query) et le génome du VIH (subject). ***B.*** Contrôle négatif : alignement le plus significatif entre une séquence aléatoire, obtenue en mélangeant les nucléotides de la séquence précédente, et le génome du VIH. Noter la valeur du score *expect*, qui indique le nombre de faux-positifs attendus au hasard. Ce score présente pour les deux alignements des valeurs supérieures à 1, et est même plus élevé pour l'alignement de la séquence de CoV que pour la séquence aléatoire. On peut en conclure que la similarité entre la séquence codante de la protéine S et le génome du VIH n'est pas significative. Les alignements ont été réalisés sur le site BLAST du NCBI (https://blast.ncbi.nlm.nih.gov/Blast.cgi).

https://www.medecinesciences.org/fr/articles/medsci/pdf/first/msc200195.pdf



This workshop is the result of a collaboration between

L'éprouvette, University of Lausanne's public laboratory
(Service Culture et Médiation Scientifique)

and

the Swiss-Prot group - SIB Swiss Institute of Bioinformatics

UNIL | Université de Lausanne
Service Culture et Médiation scientifique

SIB Swiss Institute of Bioinformatics