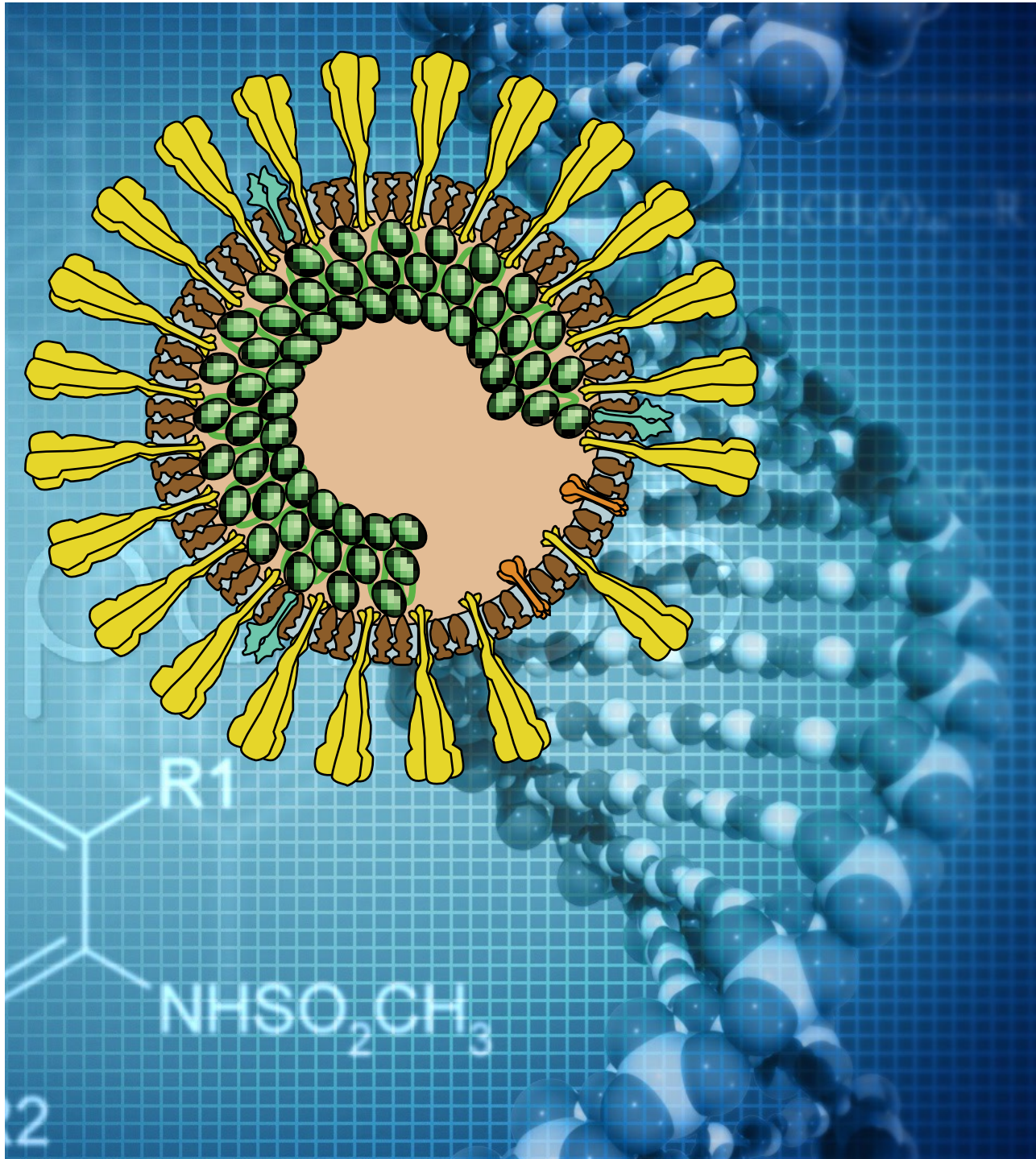
Swiss Institute of Bioinformatics

VIRAL SUBSPECIES IN

ViralZone and UniProtKB

Philippe Le Mercier, Swiss-Prot group

9th April 2024



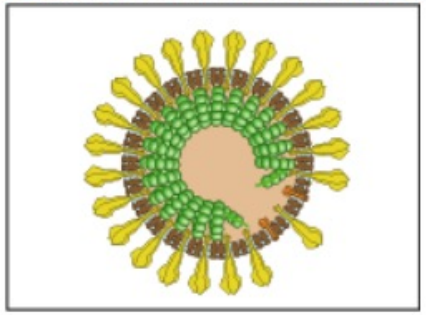
Viral subspecies in ViralZone and UniProtKB

- **SARS-CoV-2 and Variants in ViralZone**
- **Variants in UniProt**
- **Variants annotation/ metrics**

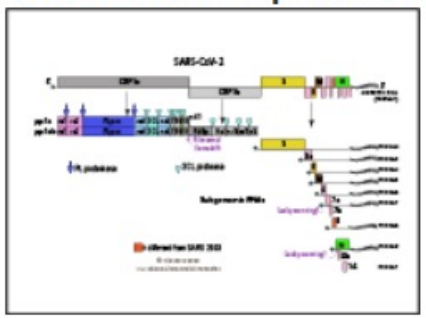
SARS-Cov-2, COVID-19 Coronavirus Resource

ViralZone resources

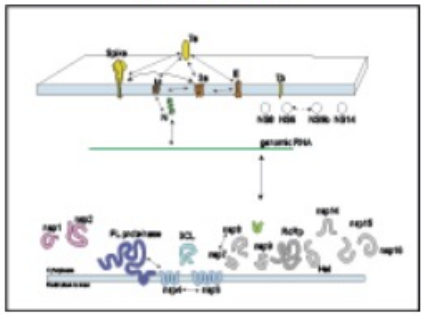
Betacoronavirus fact sheet



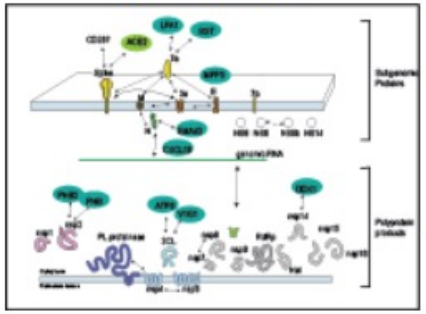
Genome and expression



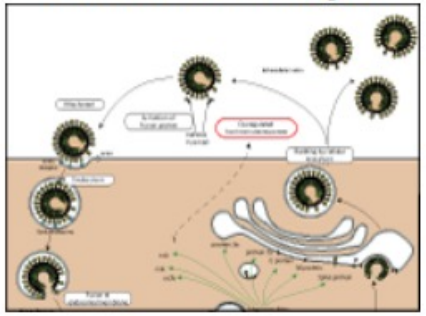
Proteome



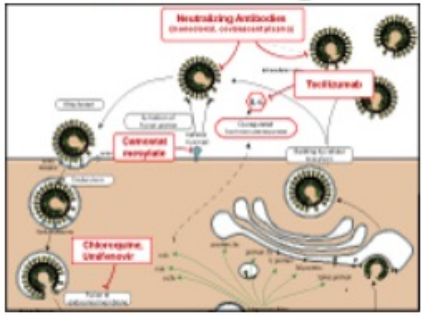
Interactome



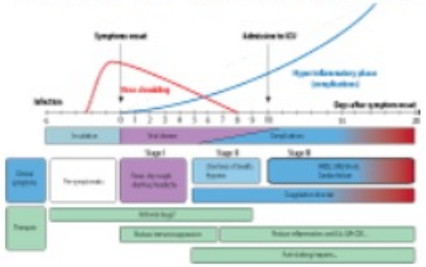
Coronavirus life cycle



Antiviral drugs



COVID-19 and treatment

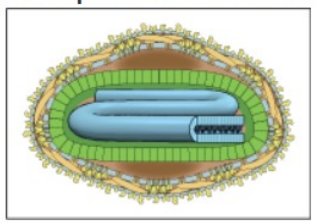


https://viralzone.expasy.org/9056

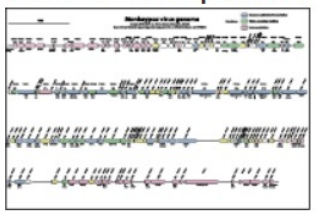
Monkeypox resource

MonkeyPox is a misnomer, **the natural reservoir hosts are rodents**. The virus can spillover to monkeys and human but it is a rodent virus.
The model organism for MonkeyPox is vaccinia virus.

Orthopoxvirus fact sheet

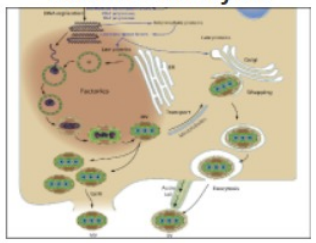


Genome and expression

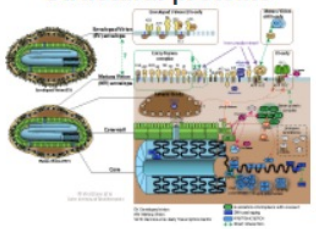


Host-virus interactions

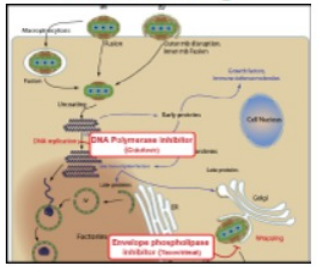
Poxvirus life cycle



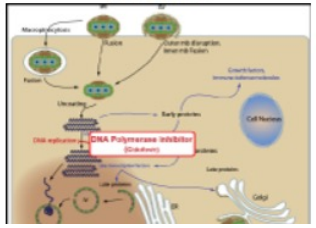
Structural proteins



Antiviral drugs



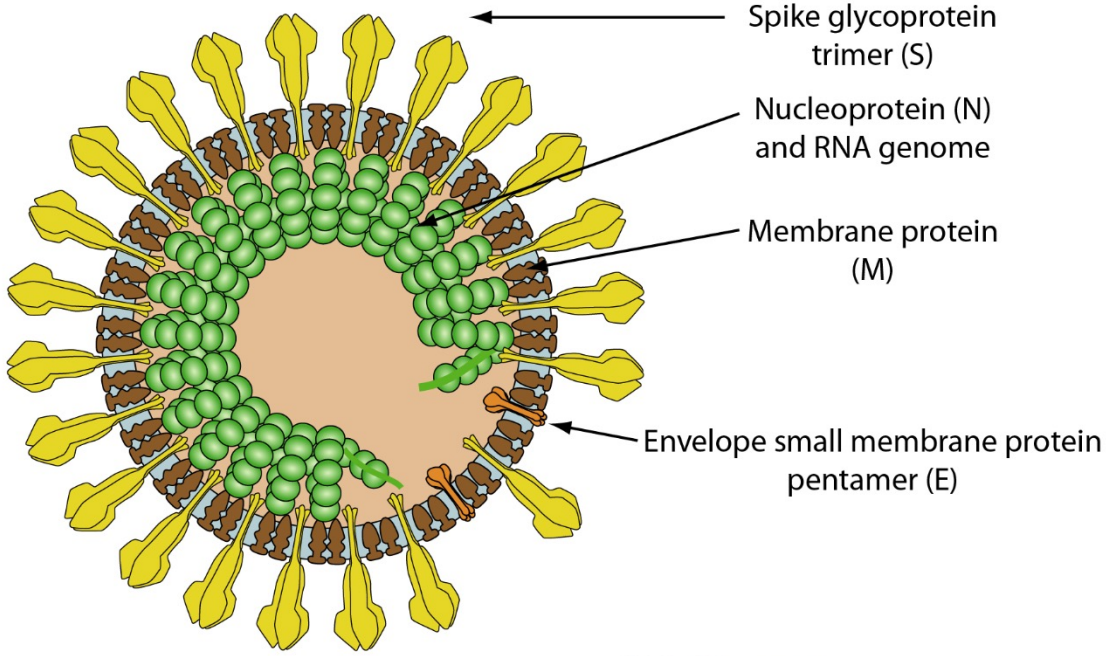
Vaccines



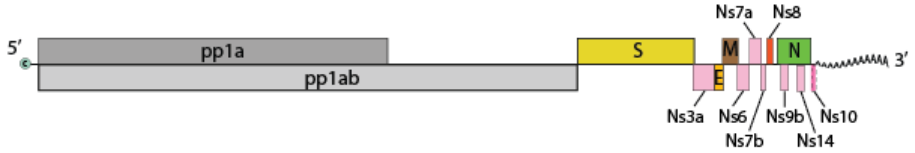
Nextstrain: Monkeypox epidemiology

SARS coronavirus 2

Virion about 120 nm in diameter

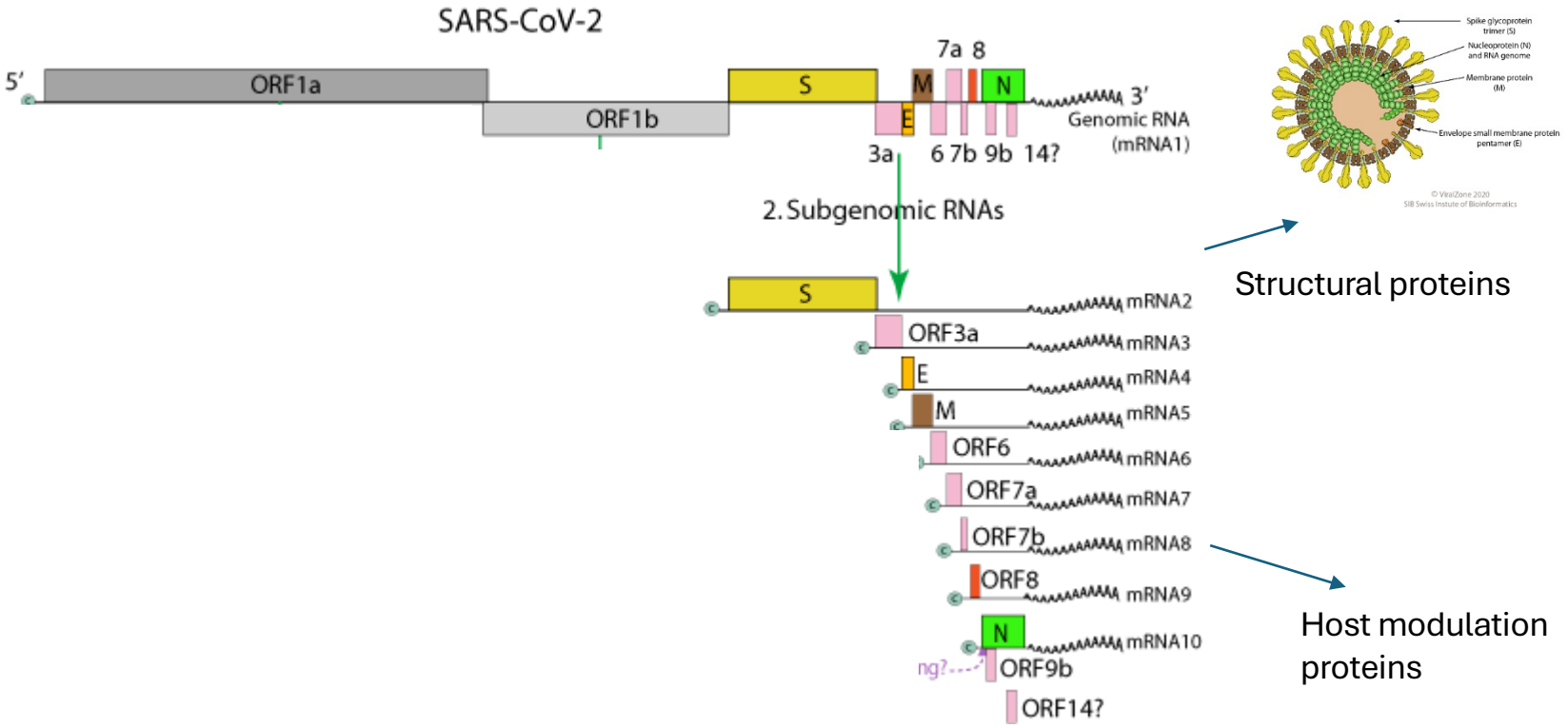


Genome: single stranded RNA messenger 29.9kb long, encoding 13 ORFs



SARS-Cov-2 genome

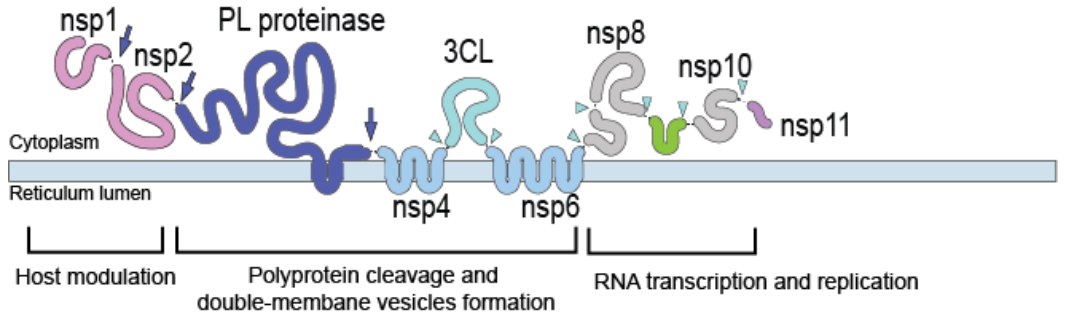
- 9 SgRNA are produced during viral replication by discontinuous transcription



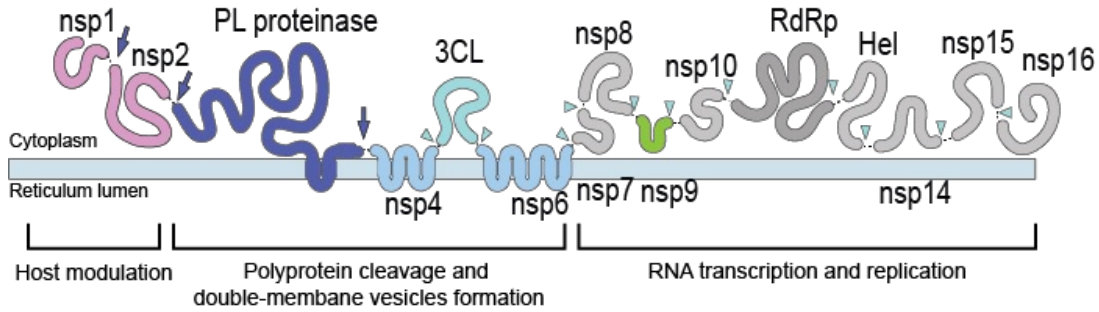
Coronavirus polyproteins

Cleaved in 17 differerent chains, necessary for virus RNA transcription and replication.

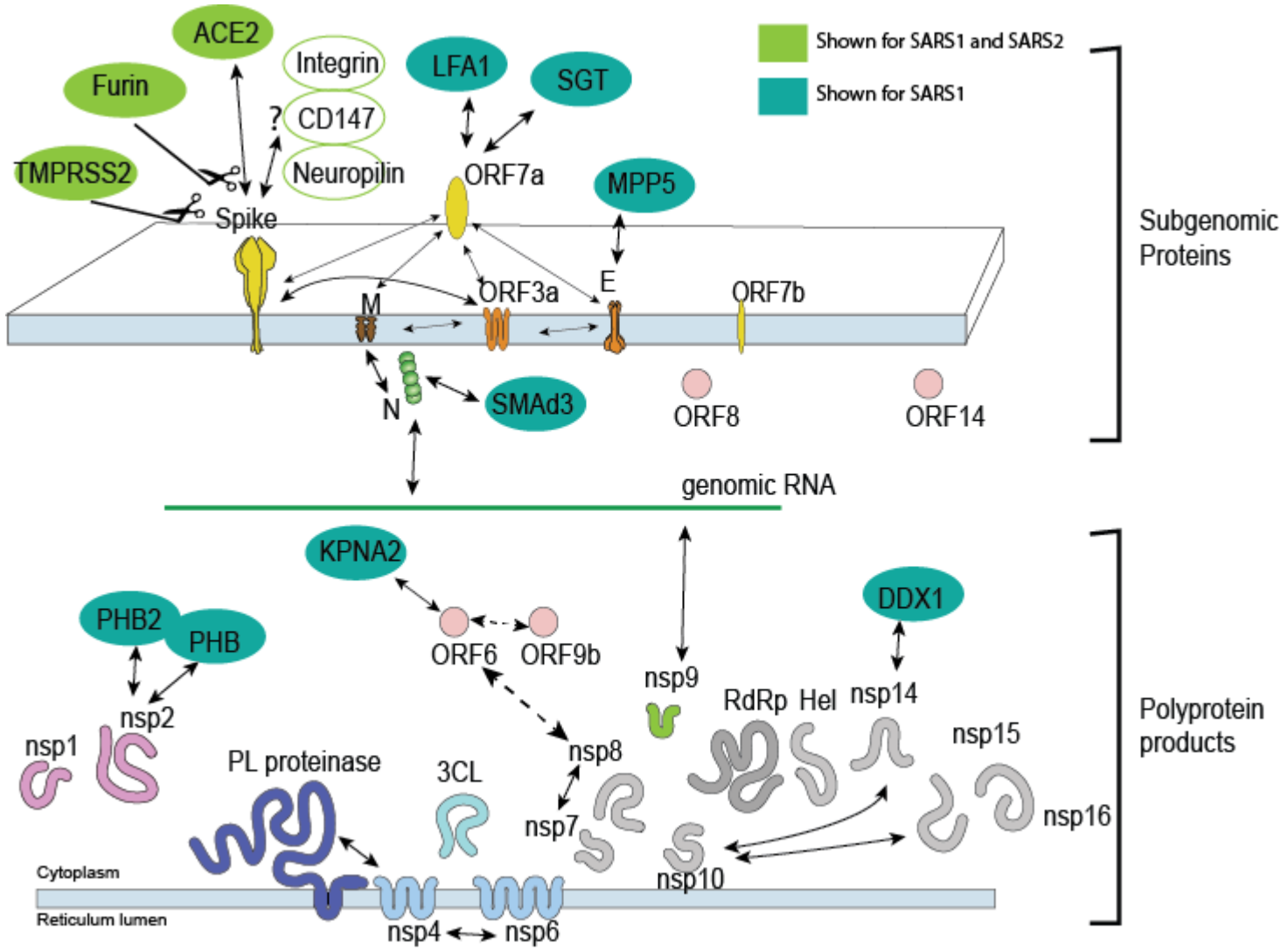
ORF1a =
polyprotein 1a (pp1a)



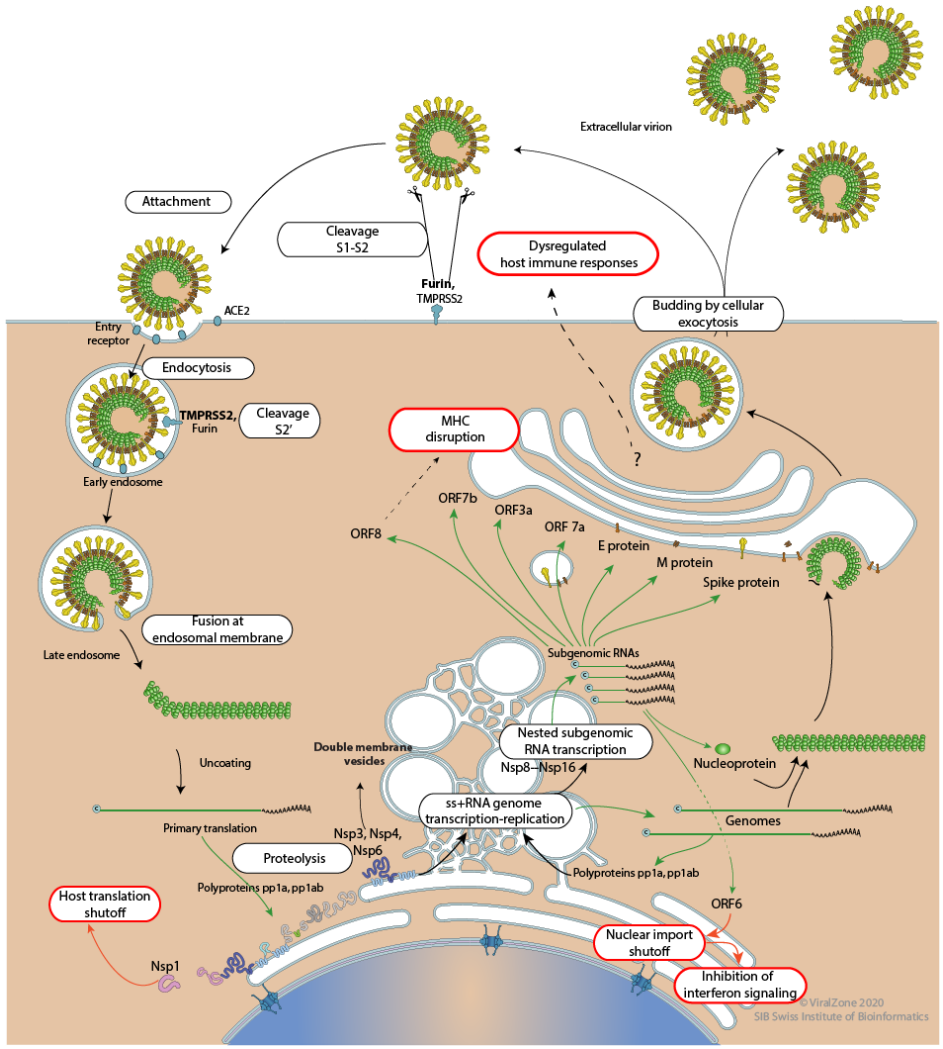
ORF1a frameshifted to ORFb =
polyprotein 1ab (pp1ab)



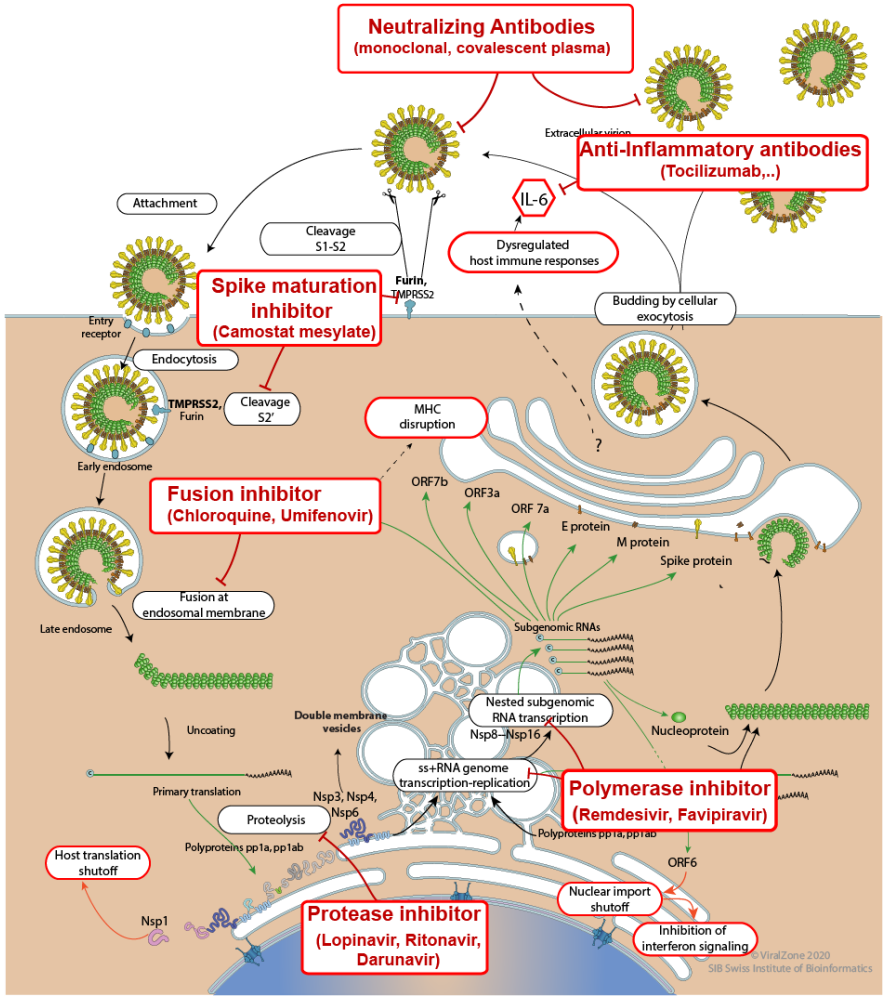
SARS-CoV-2 curated interactome



Coronavirus life-cycle

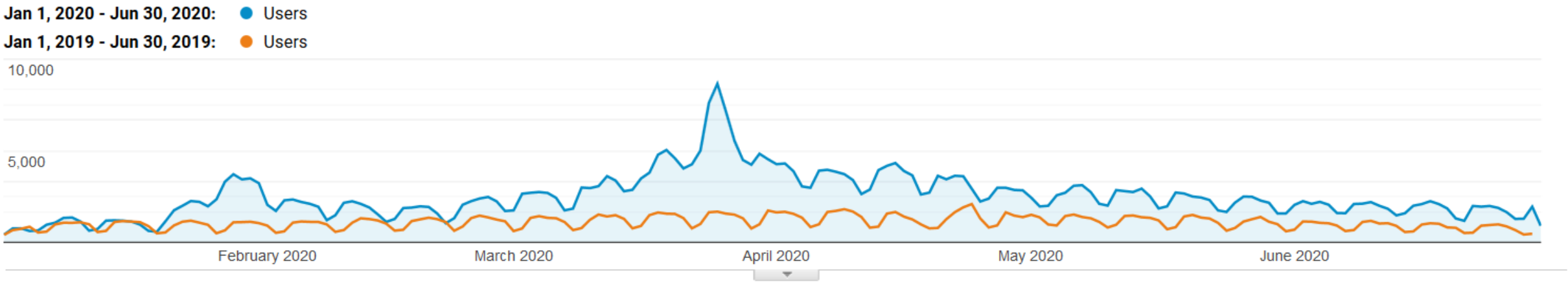


Antivirals in the context of life cycle



Compound	Ref	Status for Covid-19	Notes
Camostat		Under investigation	TMPRSS2 inhibitor that may prevent Spike cleavage and activation. Spike could still be activated by furin
Hydroxychloroquine		No-benefit in clinical trials	Non-specific anti-malaria drug. May prevent virus entry by modifying endosomal pH.
Umifenovir/Arbidol		No-benefit in clinical trials	Anti influenza drug
Ritonavir, Lopinavir		No-benefit in clinical trials	Molecules targeting HIV protease which is a ASP protease, but SARS-Cov proteases are THIOL proteases.
Darunavir		No evidence of SARS-CoV-2 inactivation	
Remdesivir		Weakly effective in clinical trials	Molecule under investigation against MERS and SARS coronaviruses.
Favipiravir		No evidence of SARS-CoV-2 inactivation, under investigation	Drug developed against influenza, redirected against ebola and now SARS-CoV-2

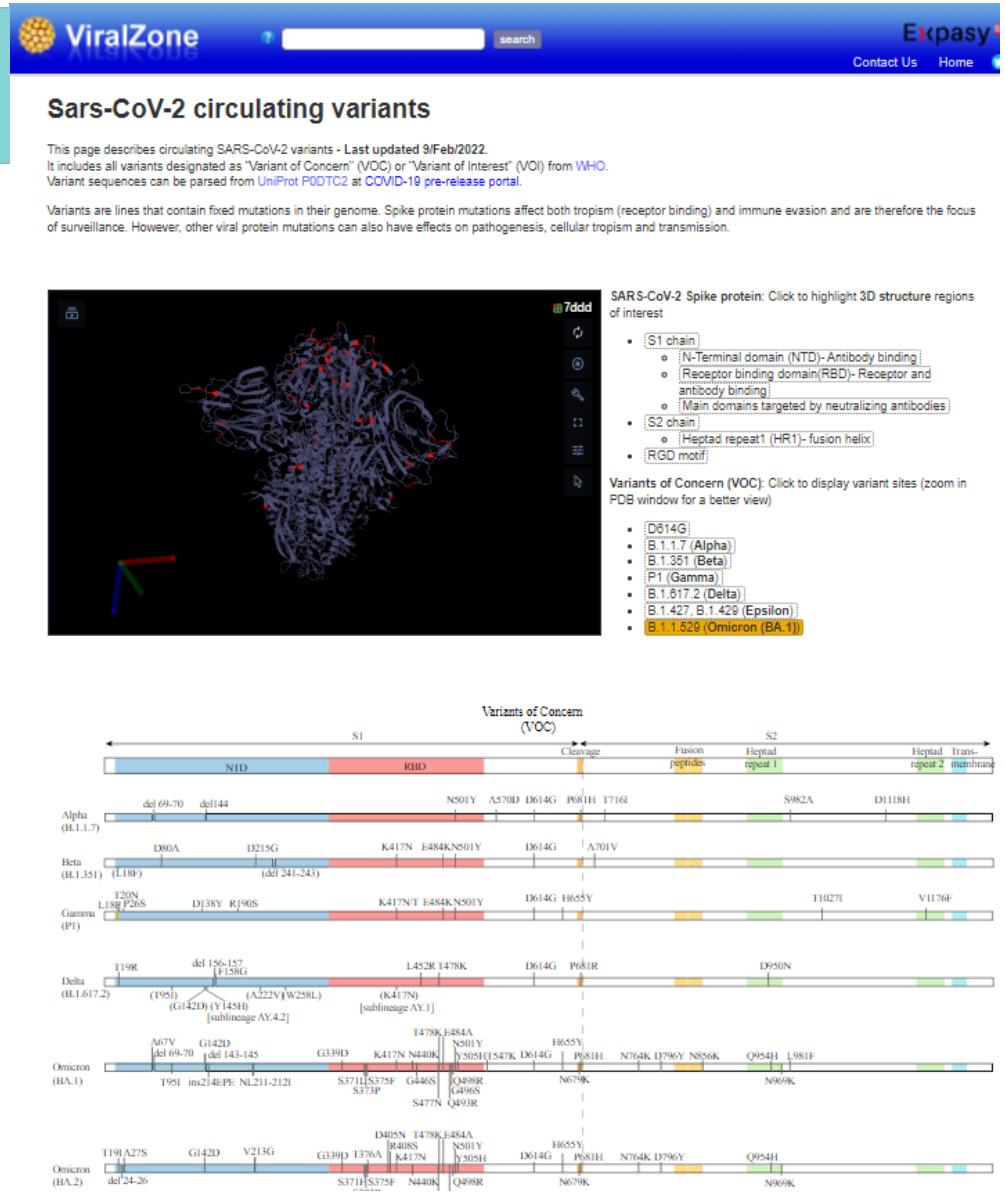
Users found the coronavirus resource usefull



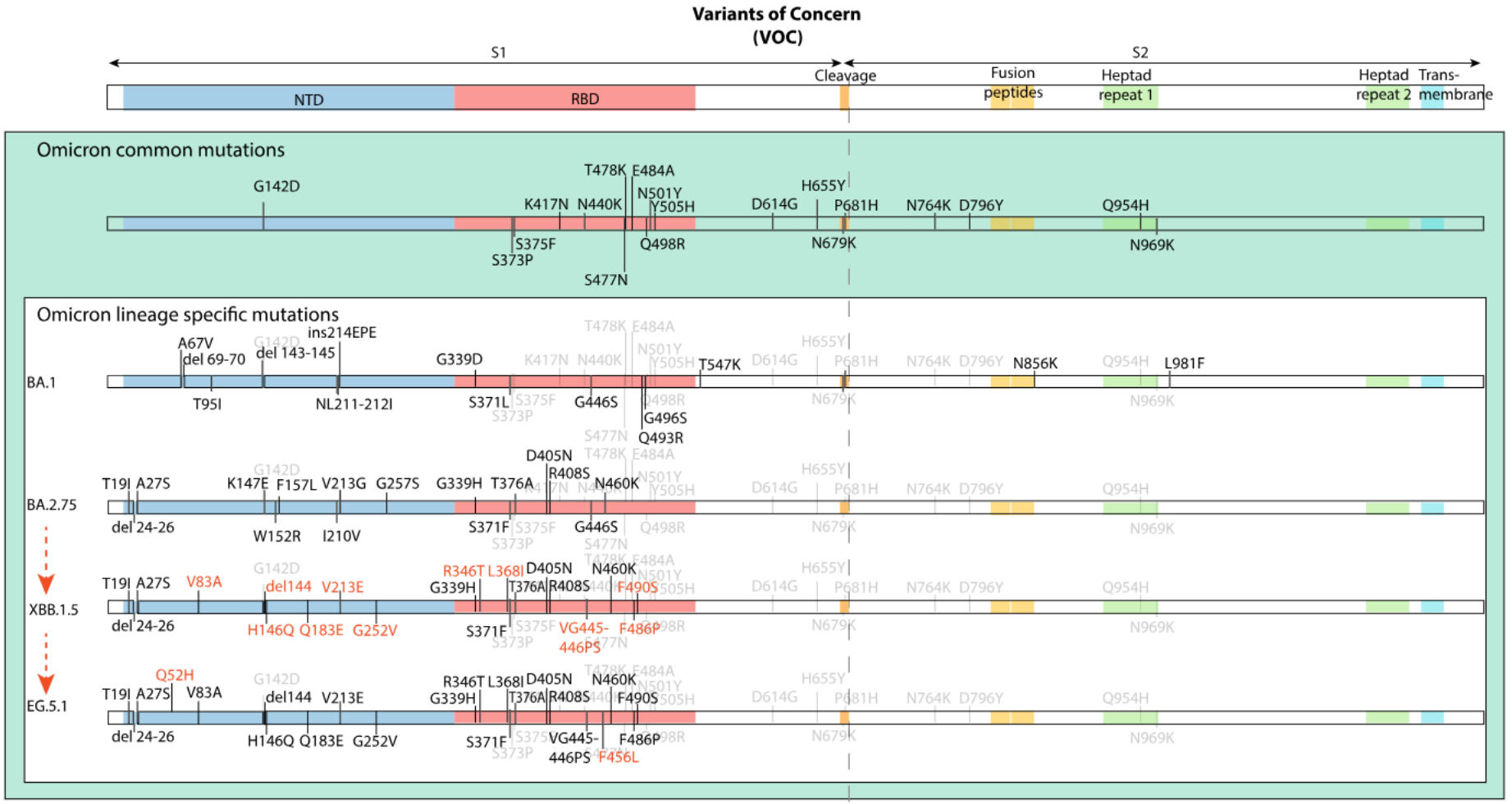
Source Google analytics

Variants in ViralZone

- Description of amino acid variations
- Links to
  - PANGO,
  - CoVariants.org,
  - outbreak.info,
  - BV-BRC
- Links to reference genomes /proteins



Overview of circulating variants (2023)



Variants data and links to reference sequences

Reference sequence

Lineage	Synonyms	Emergence	Spike protein	Other proteins	Consensus sequences	Phenotypes
19A	Wuhan-Hu-1, nCoV	China, Dec 2019	None	(ORF8: L84S in outbreak.info reference)	Genome Spike protein All proteins	Reference for variants

Variant of concern (VOC): A variant for which there is evidence of an increase in transmissibility, more severe disease and or escape of neutralization by antibodies

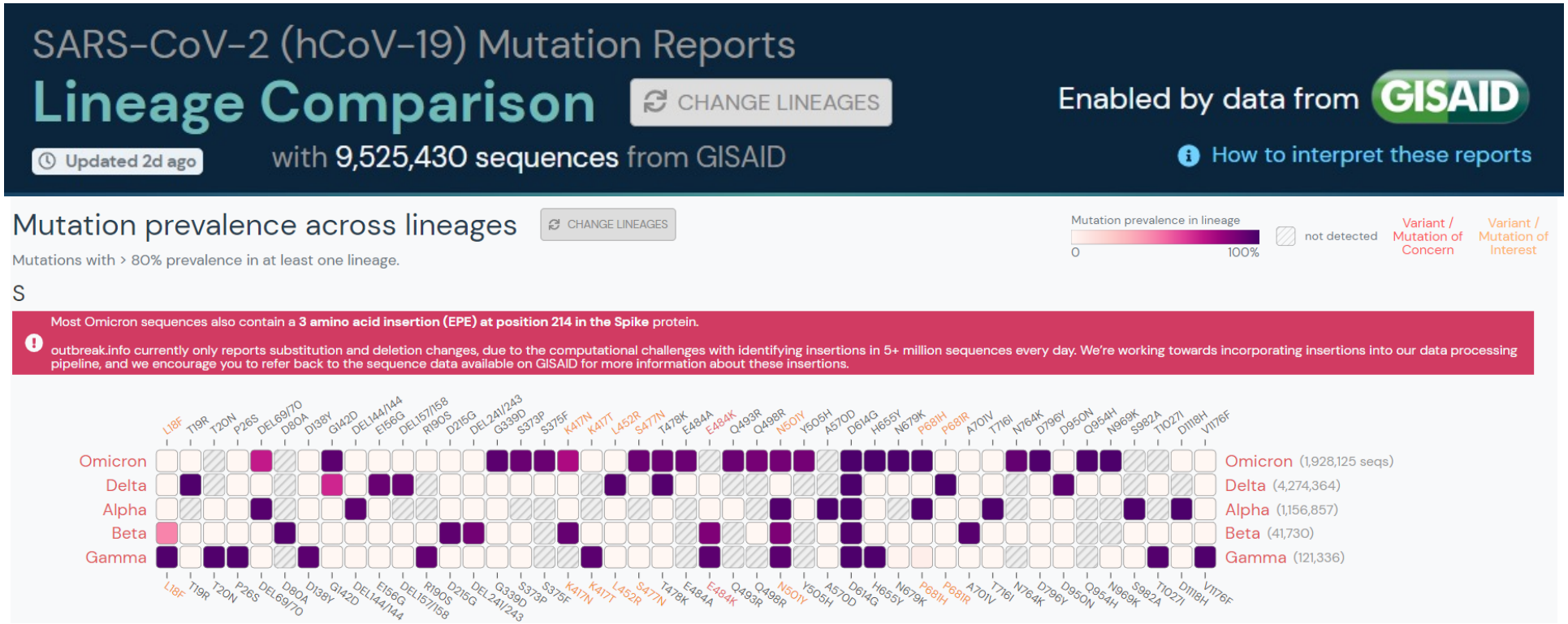
Show 150 entries Search:

Lineage	Synonyms	Emergence	Spike protein	Other proteins	Consensus sequences	Phenotypes
Omicron BA.1 >PANGO CoVariants.org oubtbreak.info: */world prevalence* */mutations* BV-BRC	BA.1 (previously B.1.1.529), 21K	South Africa, Dec 2021	show mutations	show mutations	Genome Spike Protein NCBI	Less pathogenic than previous variants because of S and NSP6 mutations
Omicron BA.2 >PANGO CoVariants.org oubtbreak.info: */world prevalence* */mutations* BV-BRC	BA.2 21L	South Africa, Dec 2021	show mutations	show mutations	Genome Spike protein NCBI	Omicrons are less pathogenic than previous variants because of S and NSP6 mutations
Omicron BA.2.12.1 >PANGO CoVariants.org BV-BRC oubtbreak.info: */world prevalence* */mutations*	BA.2.12.1 22C	North-America, Dec 2021	show mutations	show mutations	Genome Spike protein NCBI	Omicrons are less pathogenic than previous variants because of S and NSP6 mutations
Omicron BA.2.75 >PANGO CoVariants.org	BA.2.75 22D	India, spring 2022	show mutations	show mutations	Genome Spike protein NCBI	Omicrons are less pathogenic than previous variants

Choosing the right mutations!

WHO gives names, but no sequence.

Many mutations occur within a lineage. We choose to consider those that are observed in >80% of sequences (https://outbreak.info/).



CoV3D allowed to find an early reference sequence

Below are polymorphisms in SARS-CoV-2 spike glycoprotein sequences, with respect to original sequence used in full spike glycoprotein structures.

Search: L212I

Download Data as CSV

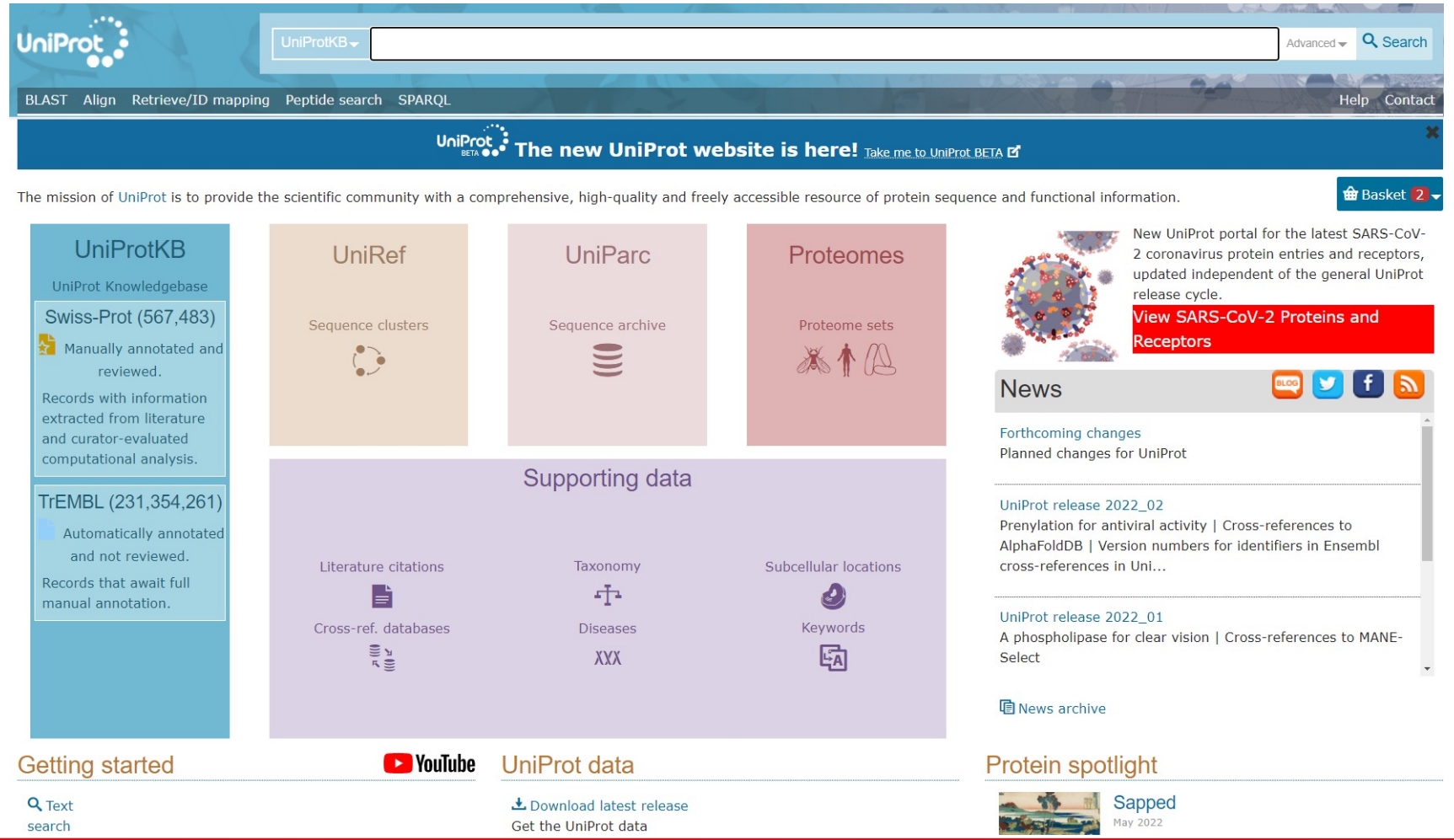
Mutations	Count	Example sequence	View mutation
A67V/H69del/V70del/T95I...	102836	UGW32340	View
A67V/H69del/V70del/T95I...	43589	UFP04971	View
A67V/H69del/V70del/T95I...	4937	UHV59178	View
A67V/H69del/V70del/T95I...	2374	UIX13664	View
A67V/H69del/V70del/T95I...	2069	UHV86077	View
A67V/H69del/V70del/T95I...	1536	UGO98052	View
A67V/H69del/V70del/T95I...	1034	UHV74268	View
L5F/A67V/H69del/V70del/...	960	UHT27752	View
A67V/H69del/V70del/T95I...	959	UHW31637	View
A67V/H69del/V70del/T95I...	555	UHW31532	View
A67V/H69del/V70del/T95I...	394	UHW40387	View
A67V/H69del/V70del/T95I...	386	UHV86291	View
A67V/H69del/V70del/T95I...	385	UHW38802	View

Viral subspecies in ViralZone and UniProtKB

- SARS-CoV-2 and Variants in ViralZone
- **Variants in UniProt**
- Variants annotation/ metrics

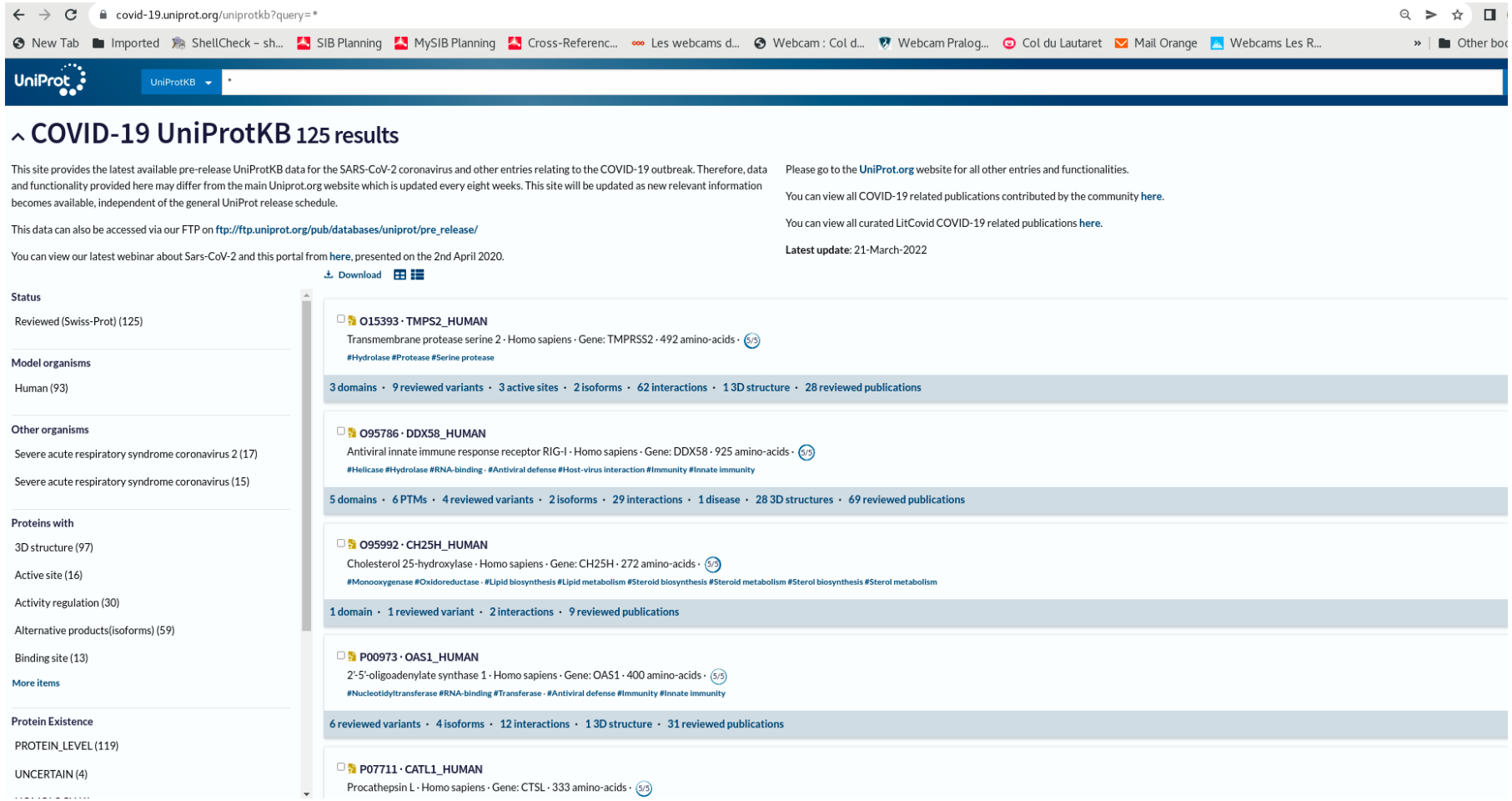
Early access to SARS-CoV-2 annotation in UniProt

UniProt releases are published every 8 weeks.
A pre-release portal has been established to provide users with quick access to the SARS-CoV-2 annotation.



Covid-19 UniProt website

https://covid-19.uniprot.org/



Variant annotation in Swis-Prot

We annotate the WHO named variants in Swiss-Prot with FT VARIANT in the Wuhan 2020 reference proteome.

```
FT VARIANT      796     796 D -> Y (in strain: 19B/501Y,
                            Omicron/BA.1, Omicron/BA.2)
                            {*CUR}
FT VARIANT      856     856 N -> K (in strain: Omicron/BA.1)
                            {*CUR}
FT VARIANT      859     859 T -> N (in strain: Lambda/C.37)
                            {*CUR}
FT VARIANT      888     888 F -> L (in strain: Eta/B.1.525)
                            {*CUR}
```

Nomenclature for deletions is not consistent

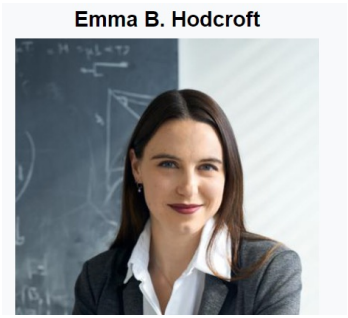
Outbreak.info: N211I, del212

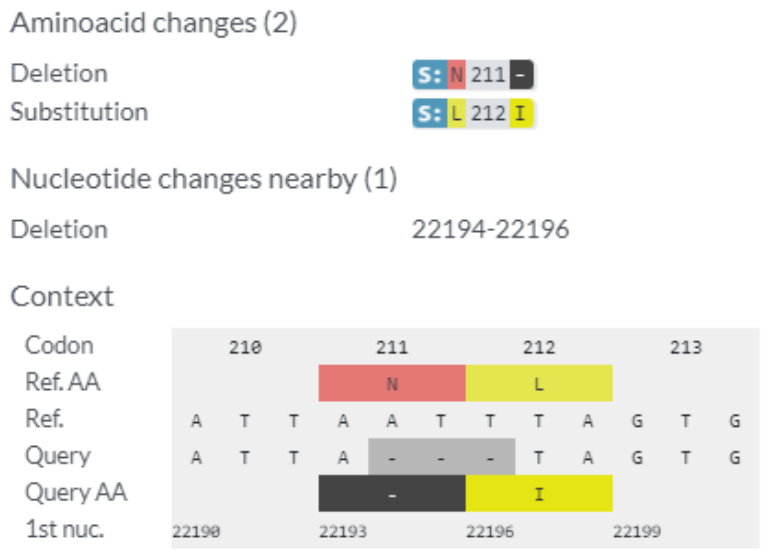
CoVariant.org

Del211, L212I

BV-BRC

N211-, L212I





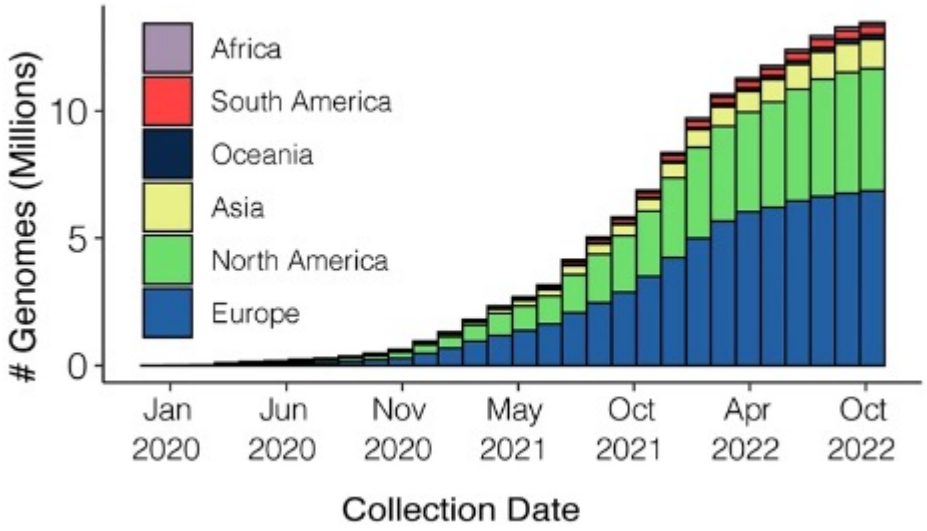
FT VARIANT 211..212 FT /note="NL -> I (in strain: Omicron/BA.1)"
FT /evidence="ECO:0000305"

Viral subspecies in ViralZone and UniProtKB

- SARS-CoV-2 and Variants in ViralZone
- Variants in UniProt
- **Variants annotation/ metrics**

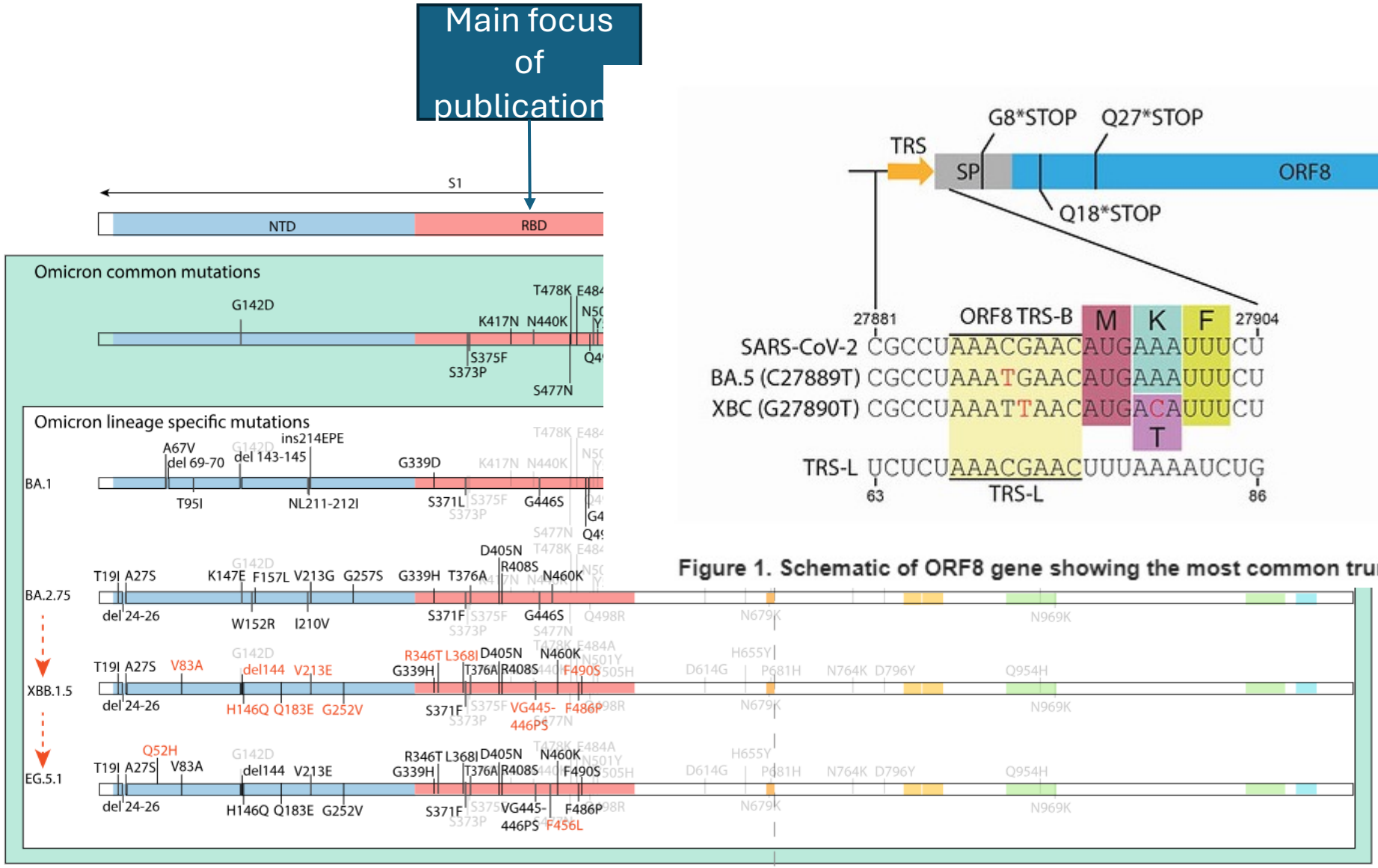
Variants mutations: how to make sense of it

- Virus surveillance produces a lot of sequences



- But there are many more power for the production of sequences than for their analysis

Mutations considered in the context of domains



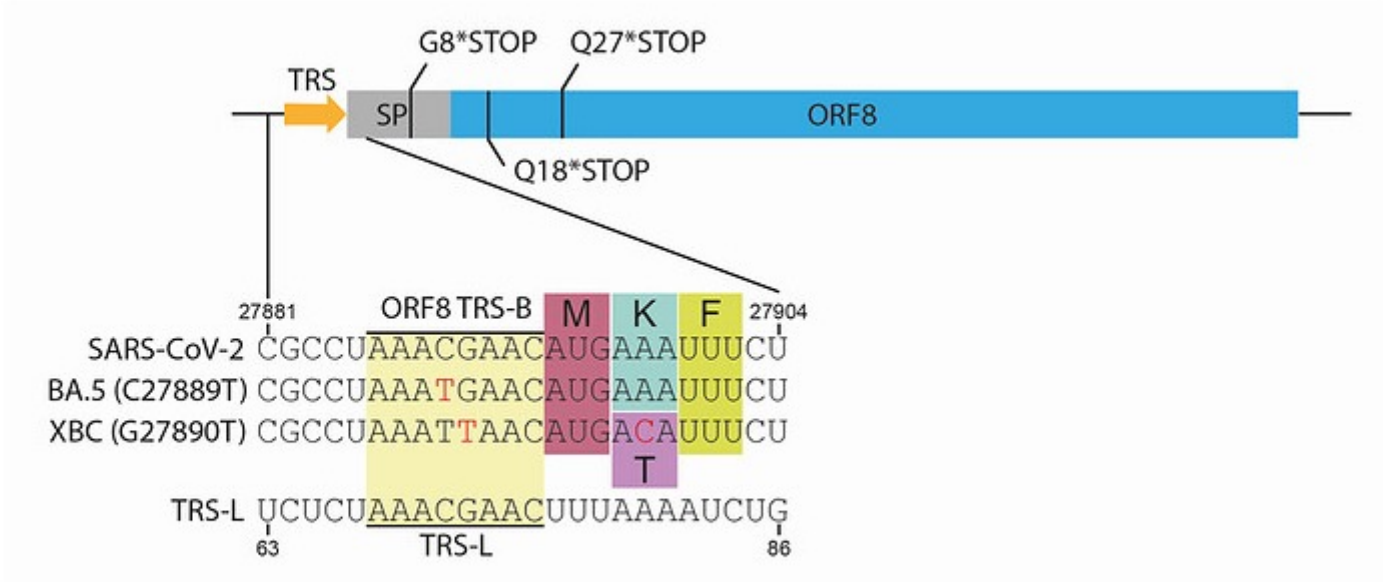
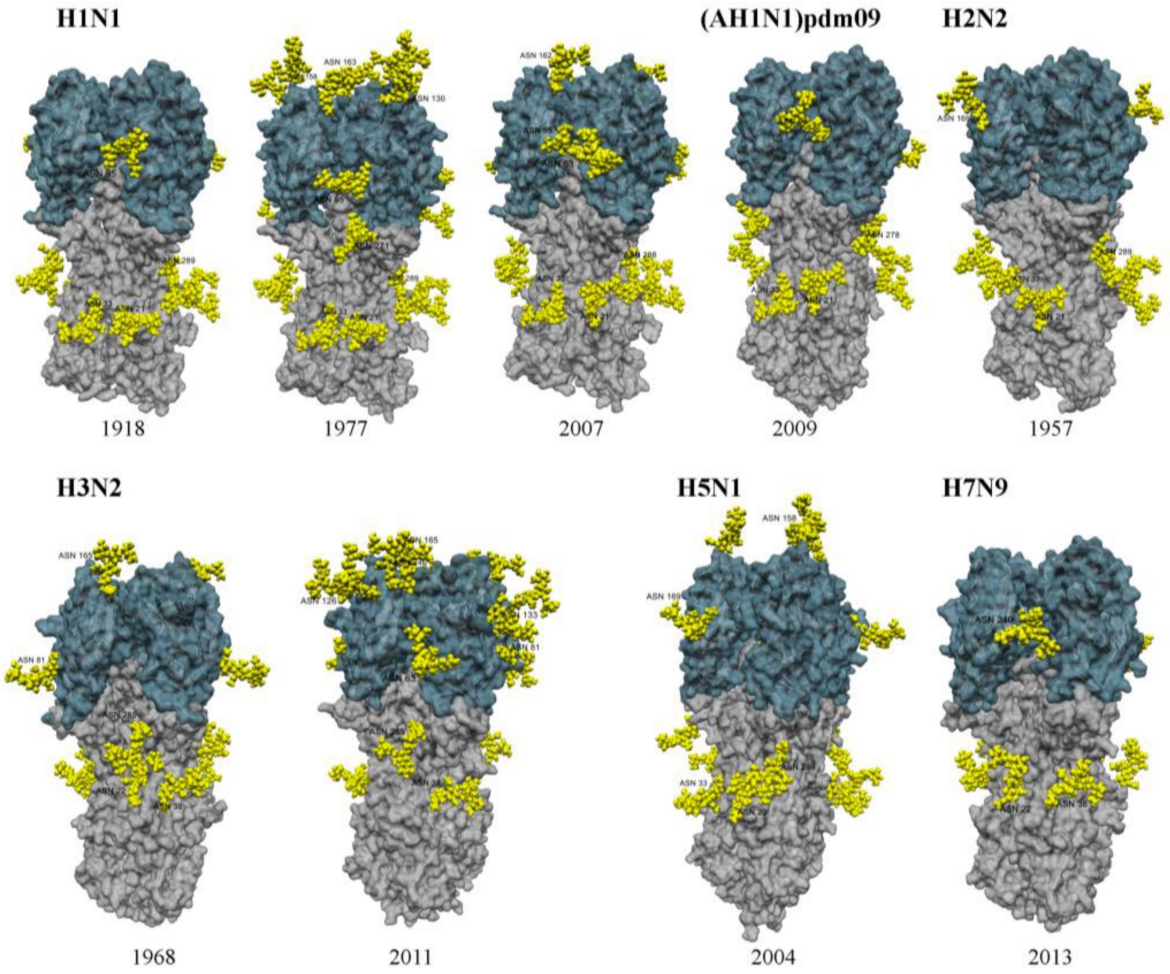


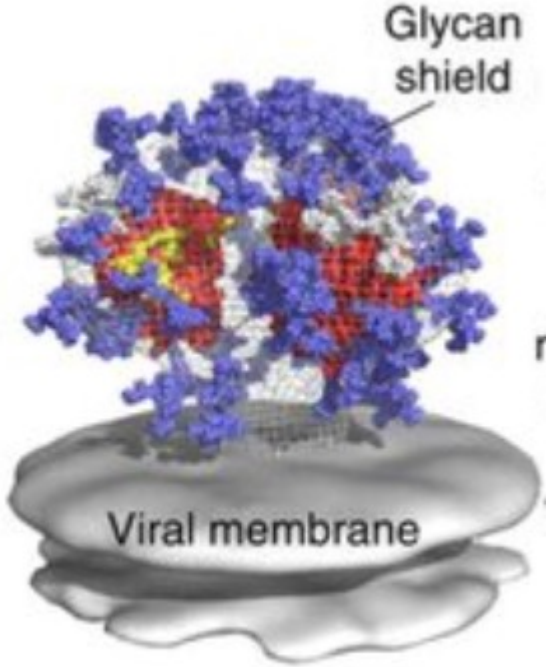
Figure 1. Schematic of ORF8 gene showing the most common truncations and TRS mutations.

Mutations considered in the context of glycosylation

Influenza glycosylation repositioning: immunity escape

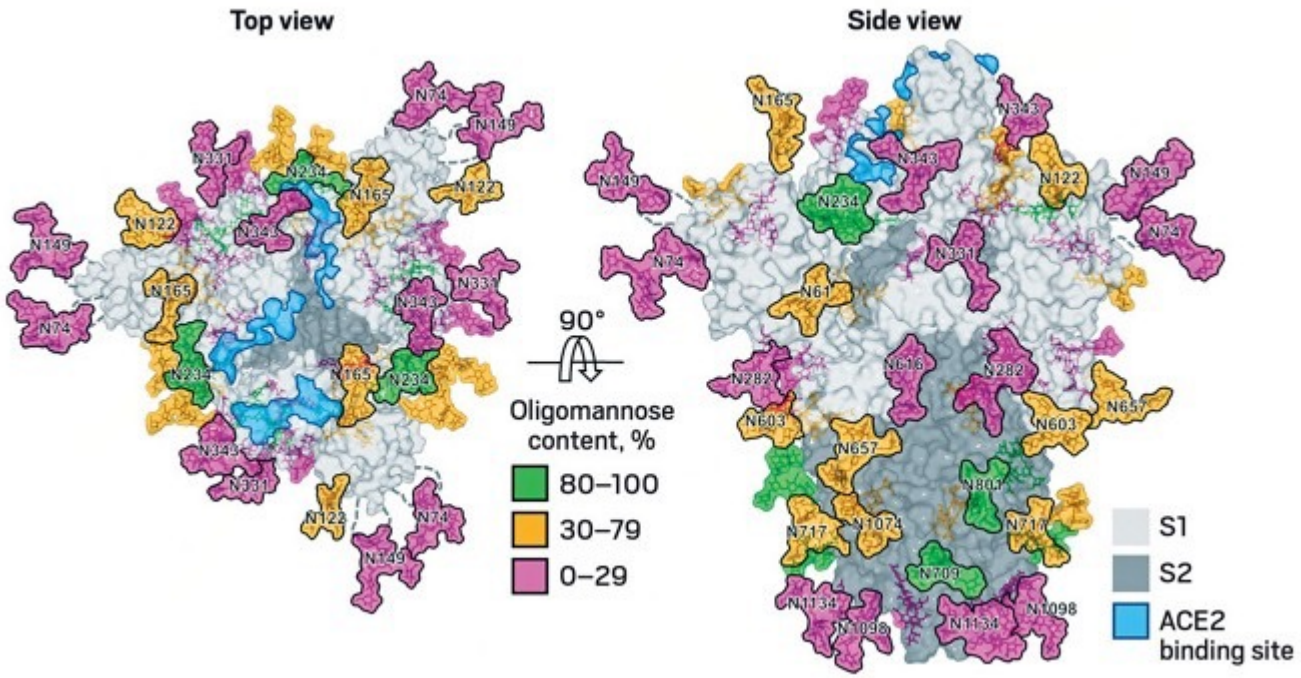


HIV glycan shield



Do sars-cov-2 variants reposition glycosylation sites?

Sars-CoV-2 glycosylations



Glycan prediction using NetNGlyc-1.0

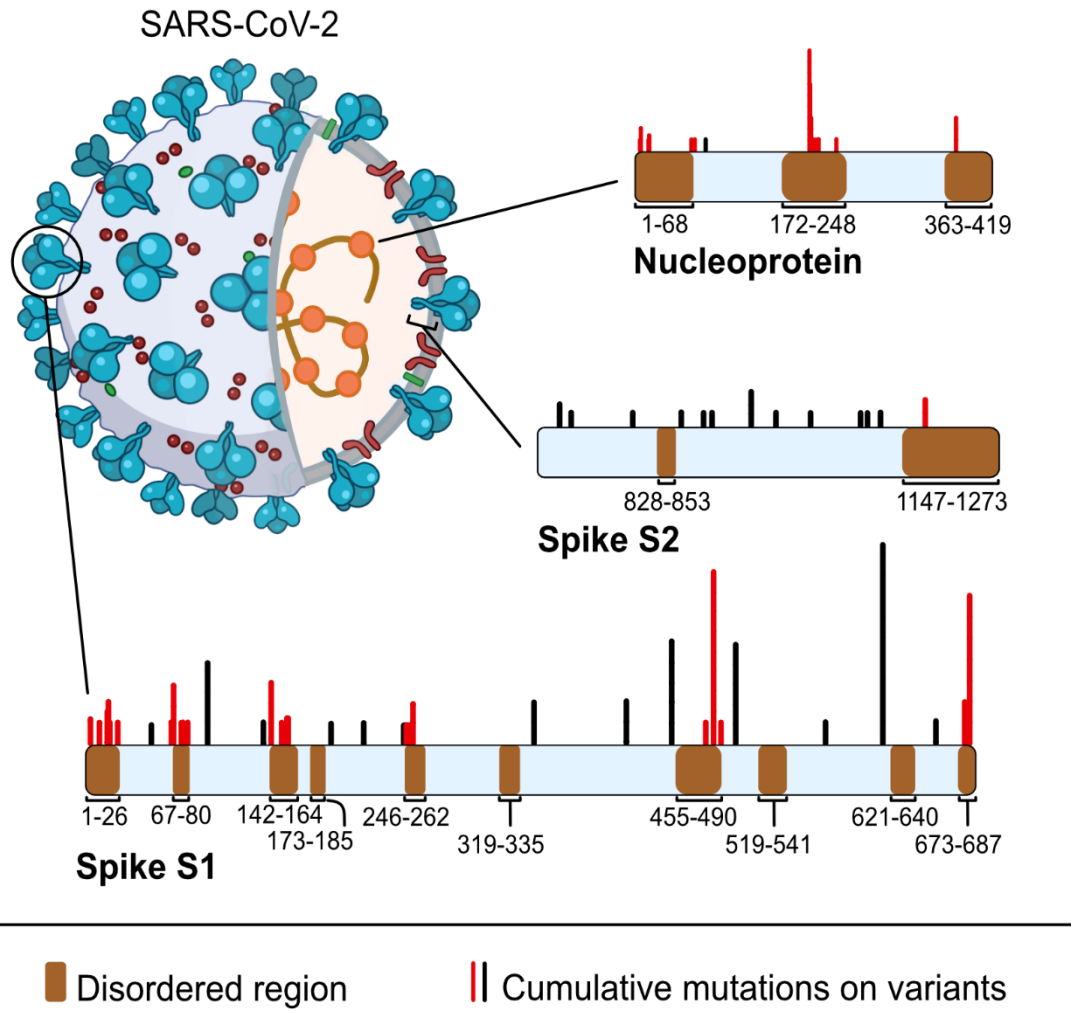
Output for 'UFO69279.1'

```
Name:  UFO69279.1      Length:  1270
MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHVISGTNGTKRFDNP      80
VLPFNDGVYFASIEKSNIIRGWIFGTTLDSKTQSLLIVNNATNVVIKVCEFQFCNDPFLDHKNNKSWMESEFRVYSSANN     160
CTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPIIVREPEDLPQGFSALEPLVDLPIGINITRFQTLLA     240
LHRSYLTPGDSSSGWTAGAAAYYVGYLQPRTFLLKYNENGTITDAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPT     320
ESIVRFPNITNLCPFDEVFNATRFASVYAWNRKRISNCVADYSVLYNLAPFFTFKCYGVSPTKLNDLCFTNVYADSFVIR     400
GDEVRQIAPGQTGNIADYNYKLPDDFTGCVIAWNSNKLDSKVSGNYNYLYRLFRKSNLKPFERDISTEIYQAGNKPCNGV     480
AGFNCYFPLRSYSFRPTYGVGHQPYRVVVLSFELLHAPATVCGPKKSTNLVKNKCVNFNFNGLKGTGVLTESNKKFLPFQ     560
QFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQGVNCTEVPVAIHADQLTPTWRVYSTGSNVF     640
QTRAGCLIGAEYVNNSYECDIPIGAGICASYQTQTKSHRRARSVASQSIIAYTMSLGAENSVAYSNNSIAIPTNFTISVT     720
TEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLKRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKYFGGFNFS     800
QILPDPSKPSKRSFIEDLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFKGLTVLPPLLTDEMIAQYTSALLAGTIT     880
SGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASALGKLQDVVNHNAQALNTLV     960
KQLSSKFGAISSVLNDIFSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFC    1040
GKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPAICHDGKAHFPREGVFVSNGTHWFVTQRNFYEPQIITTDNTFVS    1120
GNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQEL    1200
GKYEQYIKWPWYIWLGFIAGLIAIVMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
.............N.....................N.....................N.....N.....      80
.....................N.....................N.....................N     160
.....................N.....................N.....N.....N.....N.....      240
.....................N.....................N.....N.....N.....N.....      320
.....N.....N.....N.....N.....N.....N.....N.....N.....N.....N.....      400
.....N.....N.....N.....N.....N.....N.....N.....N.....N.....N.....      480
.....N.....N.....N.....N.....N.....N.....N.....N.....N.....N.....      560
.....N.....N.....N.....N.....N.....N.....N.....N.....N.....N.....      640
.....N.....N.....N.....N.....N.....N.....N.....N.....N.....N.....      720
.....N.....N.....N.....N.....N.....N.....N.....N.....N.....N.....      800
.....N.....N.....N.....N.....N.....N.....N.....N.....N.....N.....      880
.....N.....N.....N.....N.....N.....N.....N.....N.....N.....N.....      960
.....N.....N.....N.....N.....N.....N.....N.....N.....N.....N.....     1040
.....N.....N.....N.....N.....N.....N.....N.....N.....N.....N.....     1120
.....N.....N.....N.....N.....N.....N.....N.....N.....N.....N.....     1200
.....N.....N.....N.....N.....N.....N.....N.....N.....N.....N.....     1280

(Threshold=0.5)
--------------------------------------------------------------------------------
SeqName      Position  Potential    Jury    N-Glyc
                                  agreement result
--------------------------------------------------------------------------------
UFO69279.1     17 NLTT   0.6607     (8/9)   +
UFO69279.1     61 NVTW   0.7863     (9/9)   +++
UFO69279.1     72 NGTK   0.7056     (9/9)   ++
UFO69279.1    120 NATN   0.6784     (8/9)   +
UFO69279.1    144 NKSW   0.5437     (6/9)   +
UFO69279.1    160 NCTF   0.6226     (8/9)   +
UFO69279.1    231 NITR   0.7616     (9/9)   +++
UFO69279.1    279 NGTI   0.7381     (9/9)   ++
UFO69279.1    328 NITN   0.5888     (8/9)   +
UFO69279.1    340 NATR   0.5704     (8/9)   +
UFO69279.1    600 NTSN   0.5787     (6/9)   +
UFO69279.1    613 NCTE   0.7281     (9/9)   ++
UFO69279.1    654 NNSY   0.5141     (5/9)   +
UFO69279.1    706 NNSI   0.3533     (8/9)   -
UFO69279.1    714 NFTI   0.6427     (9/9)   ++
UFO69279.1    798 NFSQ   0.6246     (9/9)   ++
```

-> There was no significant changes in any variant spike glycosylation

SARS-CoV-2 variants preferentially emerge at intrinsically disordered protein sites



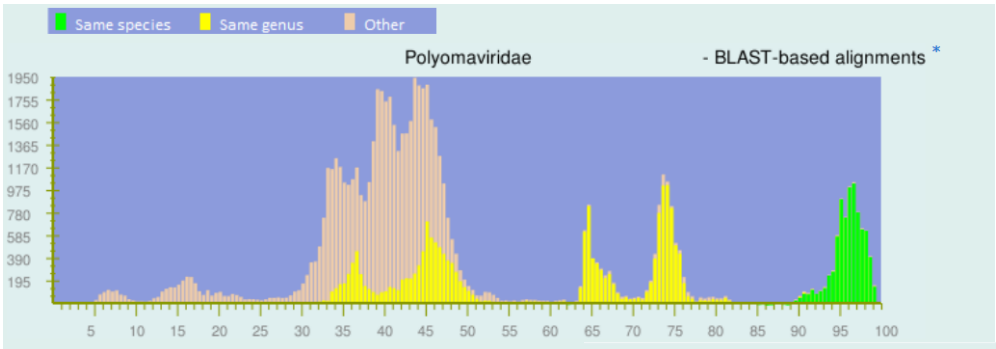
With the disprot group,

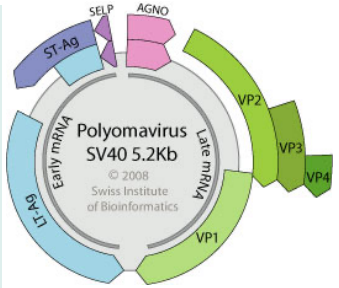
We have established that variants occur preferentially at disordered regions in Spike S1 and Nucleoprotein

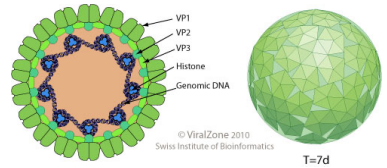
From:SARS-CoV-2 variants preferentially emerge at intrinsically disordered protein sites helping immune evasion.
FEBS J. 2022 Feb 2

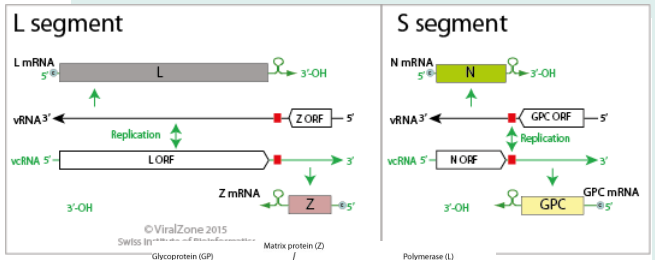
Genomic heterogeneity

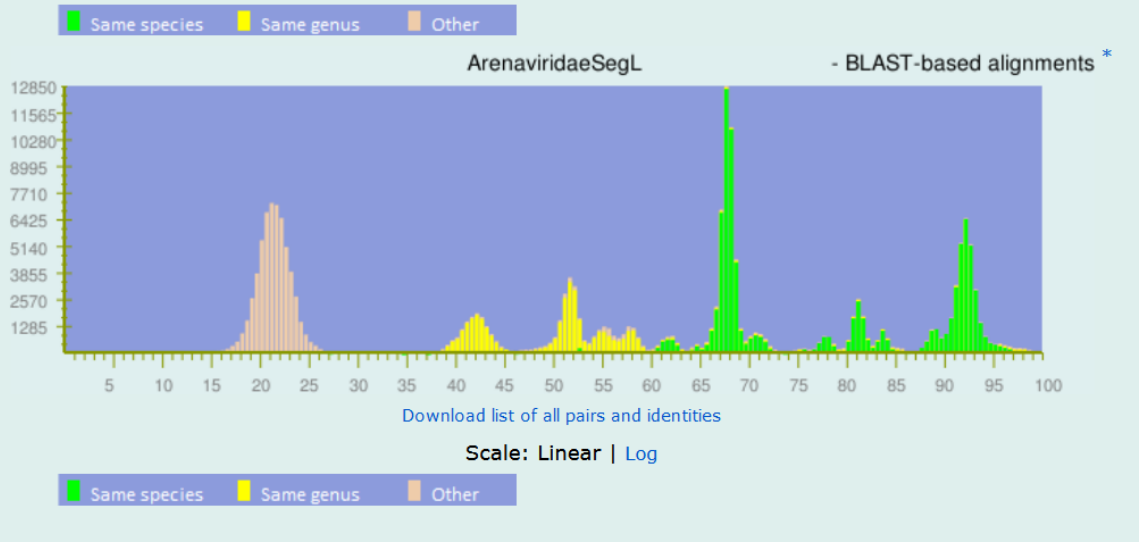
- **PASC (PAirwise Sequence Comparison)** is a web tool for analysis of pairwise identity distribution within viral families.



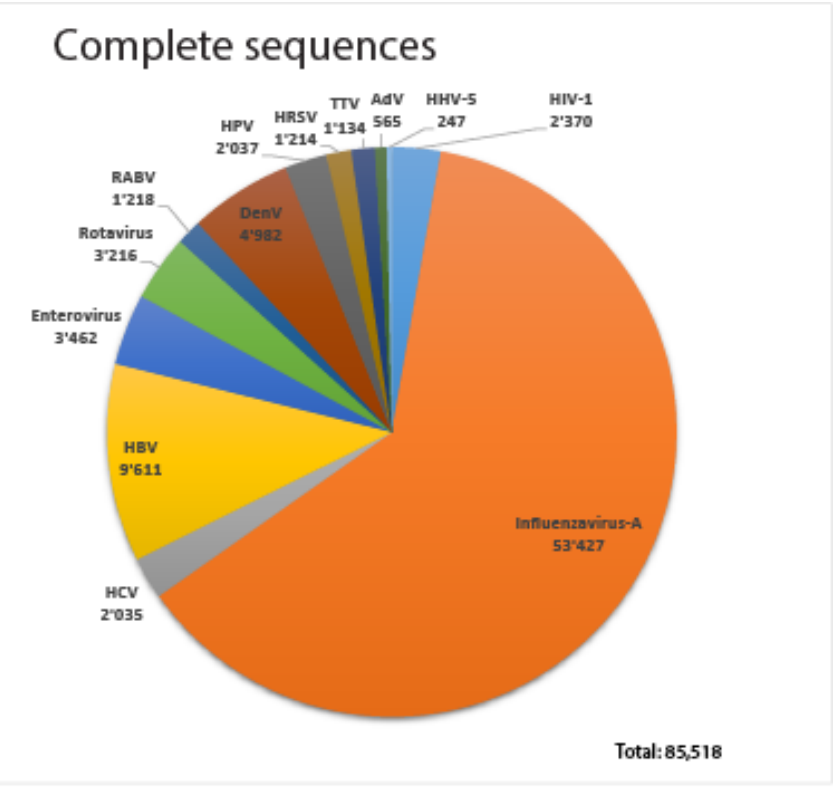


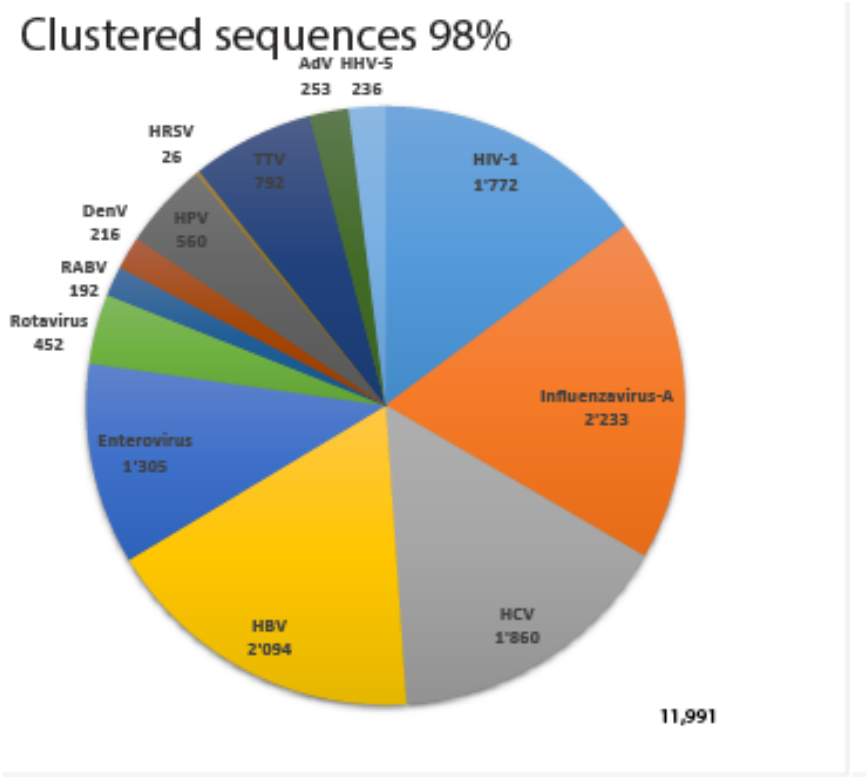


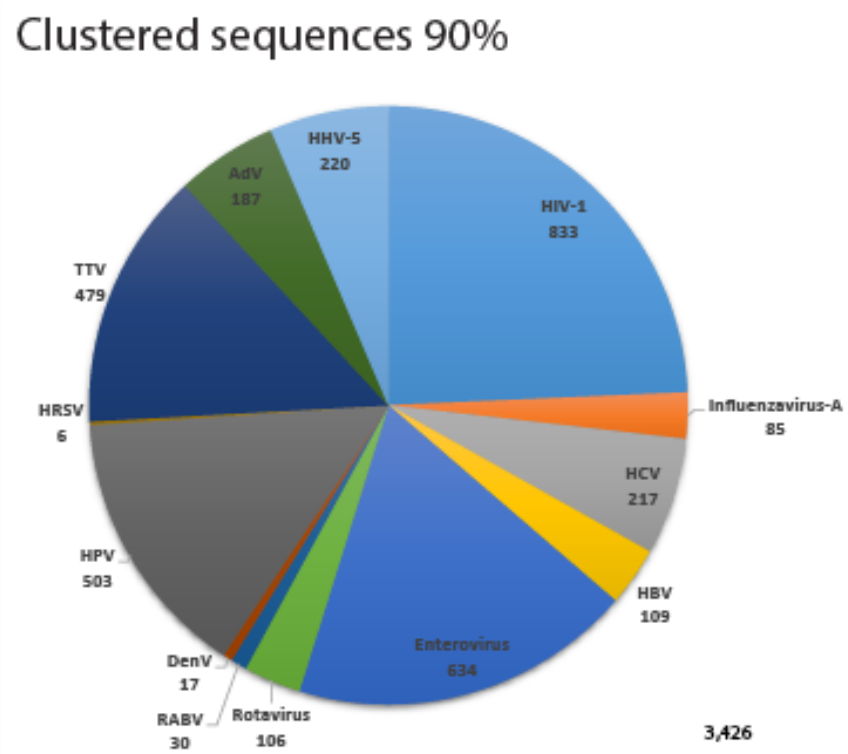




Developing a metric for the diversity of variants?

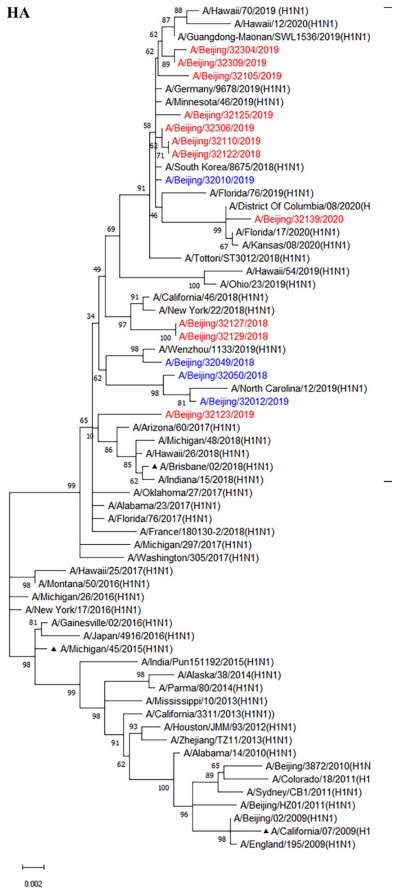




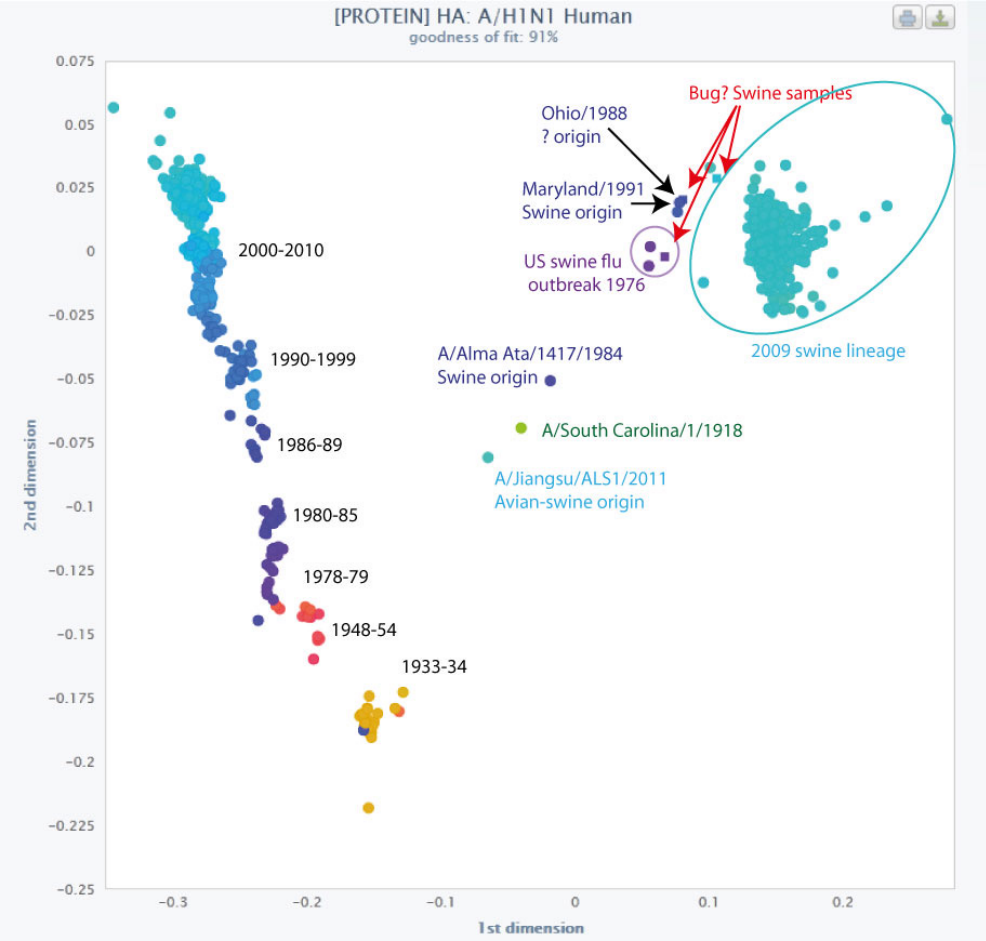


Genome landscape representation

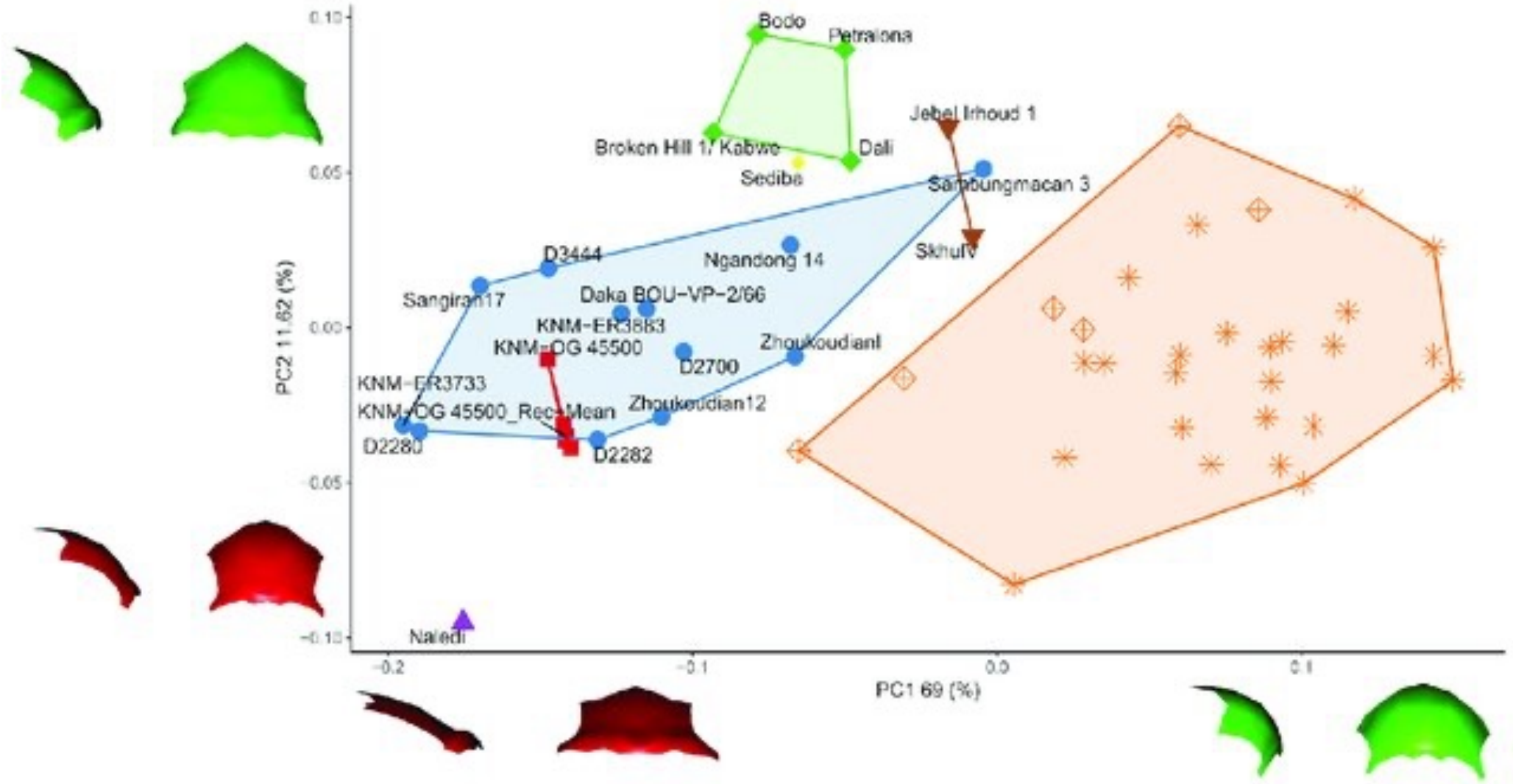
- Phylogenetic trees



- PCA analysis



PCA plot of fossil specimens and recent modern humans



Conclusion

Monitoring below-species sequence: users needs

Classification

Repository database → Classification system → Automatic classification tool

INSDC, GISAID,..

Influenza
HIV
Ebola
SARS
Rotavirus
Monkeypox
...

PANGOLIN, Nextclade,..

Genotyping: HIV, HTLV, HBV, HCV, CMV, HPV, Dengue,...

Pathology

Drug resistance

Tools for HIV, Influenza,..

Experimental

Antigenicity: Vaccine, antibody therapy escape

Experimental

3D modeling

glycosylation

Antigenicity Predictions

tropism

Experimental

3D modeling

Receptor/tropism Predictions

Analysis

Epidemiology

Nextstrain, Usher, Beast, BV-BRC, ...

Epidemic models

Functional analysis

Experimental

predictions

Knowledge

Pubmed, UniProt, ViralZone, BV-BRC...

Thanks to

swissprot

- Arnaud Kerhornou,
- Edouard de Castro,
- Cristina Casals-Casas,
- Patrick Masson,
- Chantal Hulo,
- Sylvain Poux,
- Ivo Pedruzzi,
- Nicole Redaschi,
- Alan Bridge

Collaborations:

HUG Hôpitaux Universitaires Genève

Samuel Cordey,

Pauline Vetter,

Laurent Kaiser

University of Basel

Richard Neher

UNIVERSITÄT BERN

Emma Hodcroft

JCVI J. CRAIG VENTER INSTITUTE™

Anna Maria Niewiadomska

Christian Zmasek

Richard H. Scheuermann

Los Angeles County Health Laboratories

Alejandro R. Gener

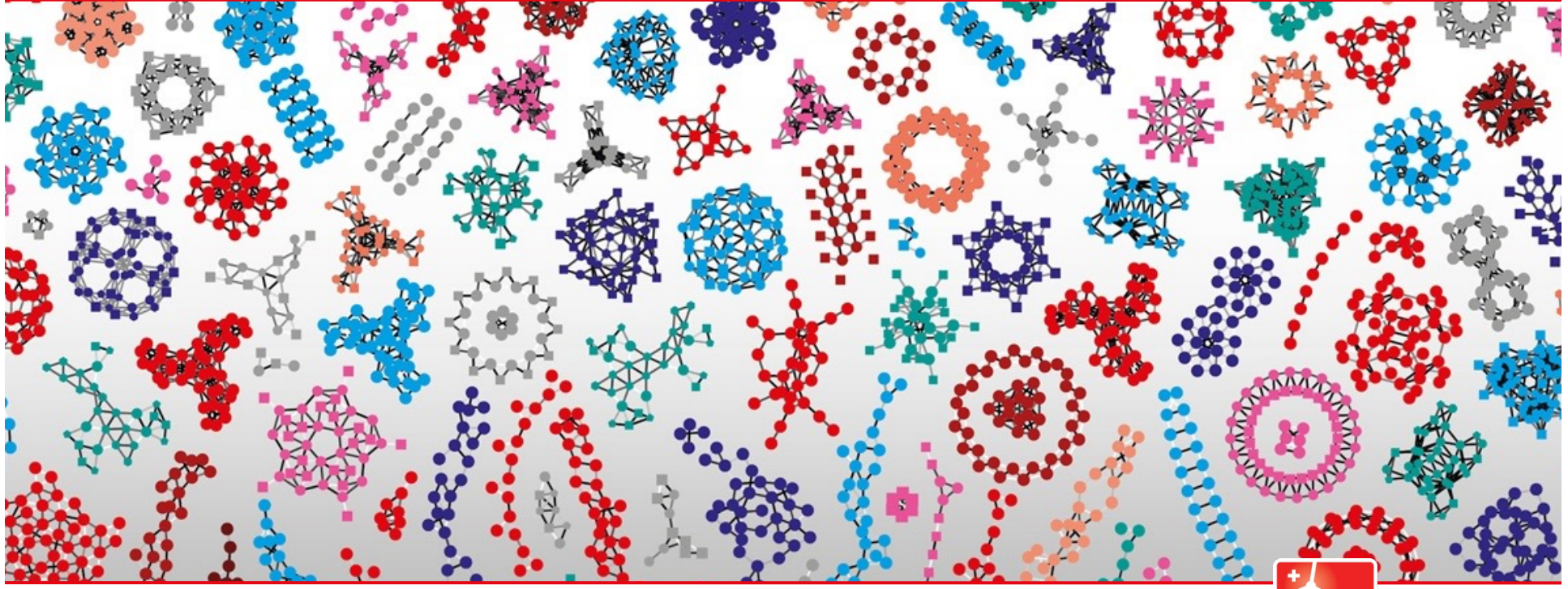
NIH

Bernard Moss

Ryan Connor

UF UNIVERSITY of FLORIDA

Rich Condit



SIB
Swiss Institute of
Bioinformatics

Thank you

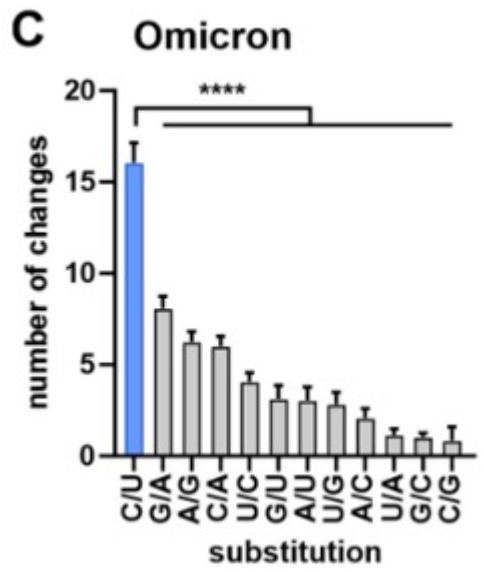
elixir SWITZERLAND

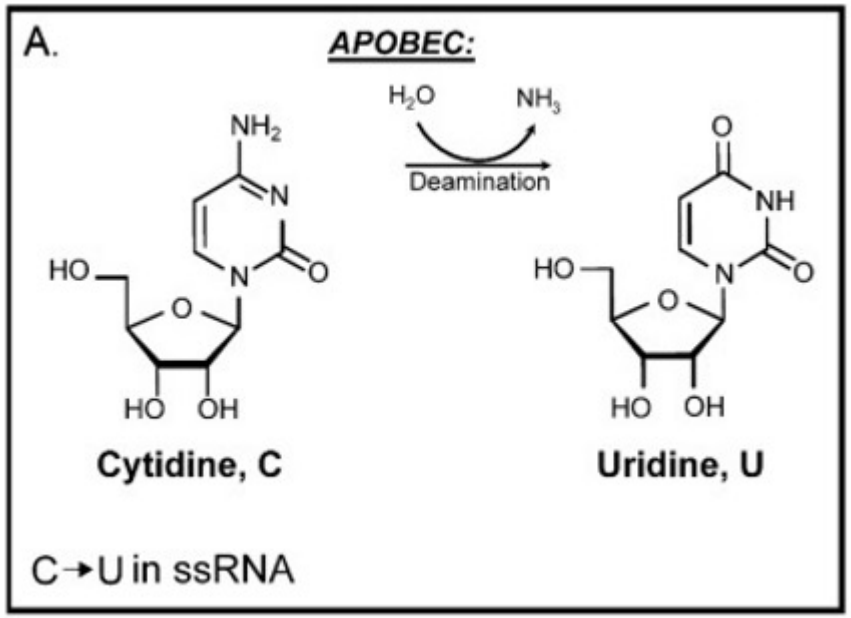
www.sib.swiss

The dynamic of SARS-CoV-2 mutations

Mutations occur due to:

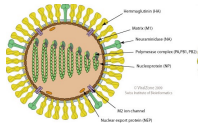
- Replication errors of the viral polymerase
- Cellular editing enzymes in the cytoplasm (immune cells)





Isolate nomenclature

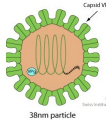
Few isolates nomenclature exist today, unfortunately there is no standart
= no way to manage virus data with common bioinformmatics tools



Influenza (1953):

Antigenic type/ Isolation host/Location of sampling/Isolate ID/Year of sampling

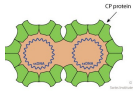
Ex: Influenza A/Puerto Rico/8/1934 H1N1



Caliciviridae (2000):

Isolation host/Genus/Species/Isolate ID/Year of sampling/Location of sampling

Ex: Hu/NLV/I/Norwalk/1968/US



Geminiviridae (2005):

Species [Location of sampling: Isolate ID: Isolation host: Year of sampling]

Ex: Maize streak virus, A [Uganda:Kas42:2005]