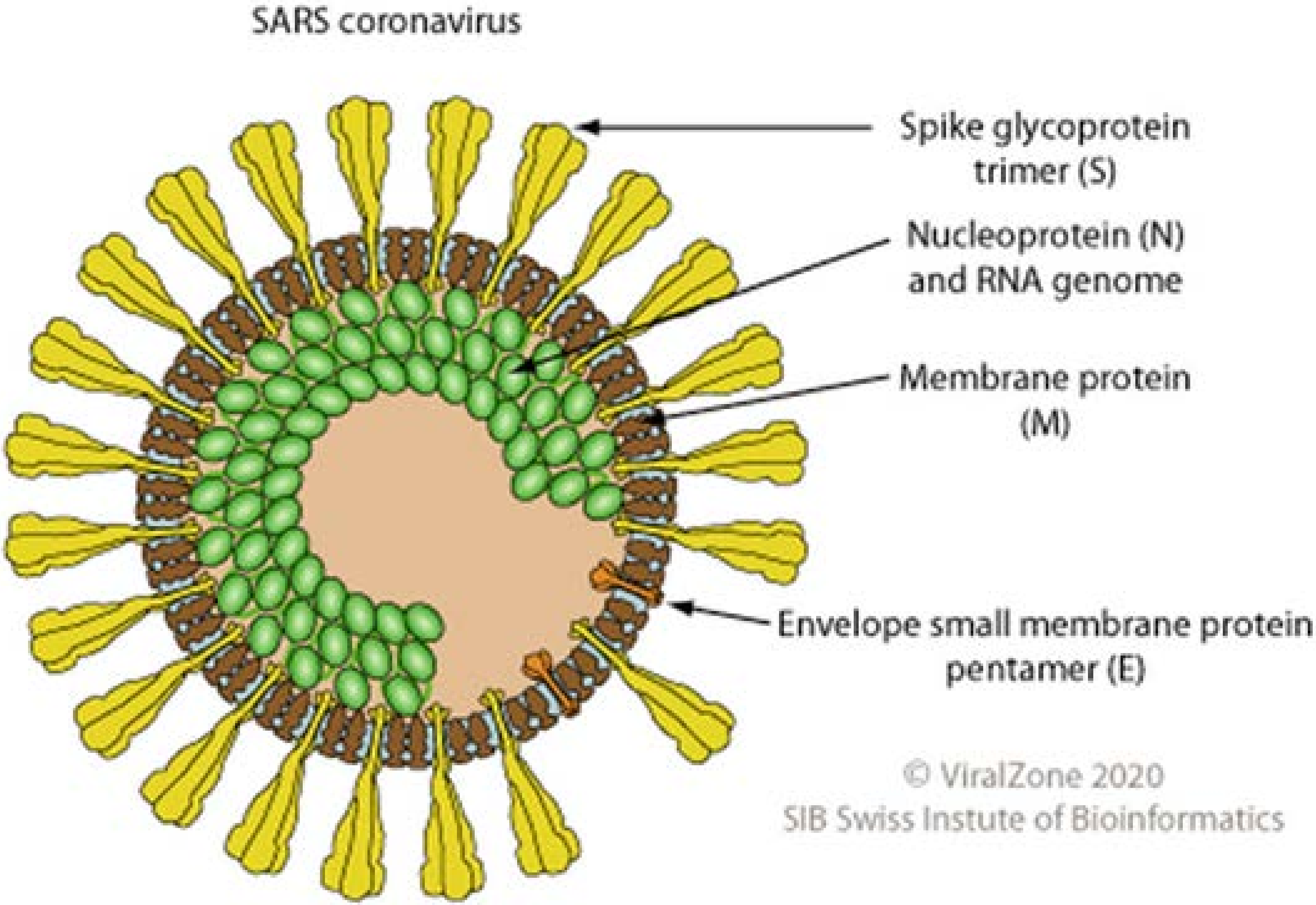
Universidade de São Paulo
Instituto de Química

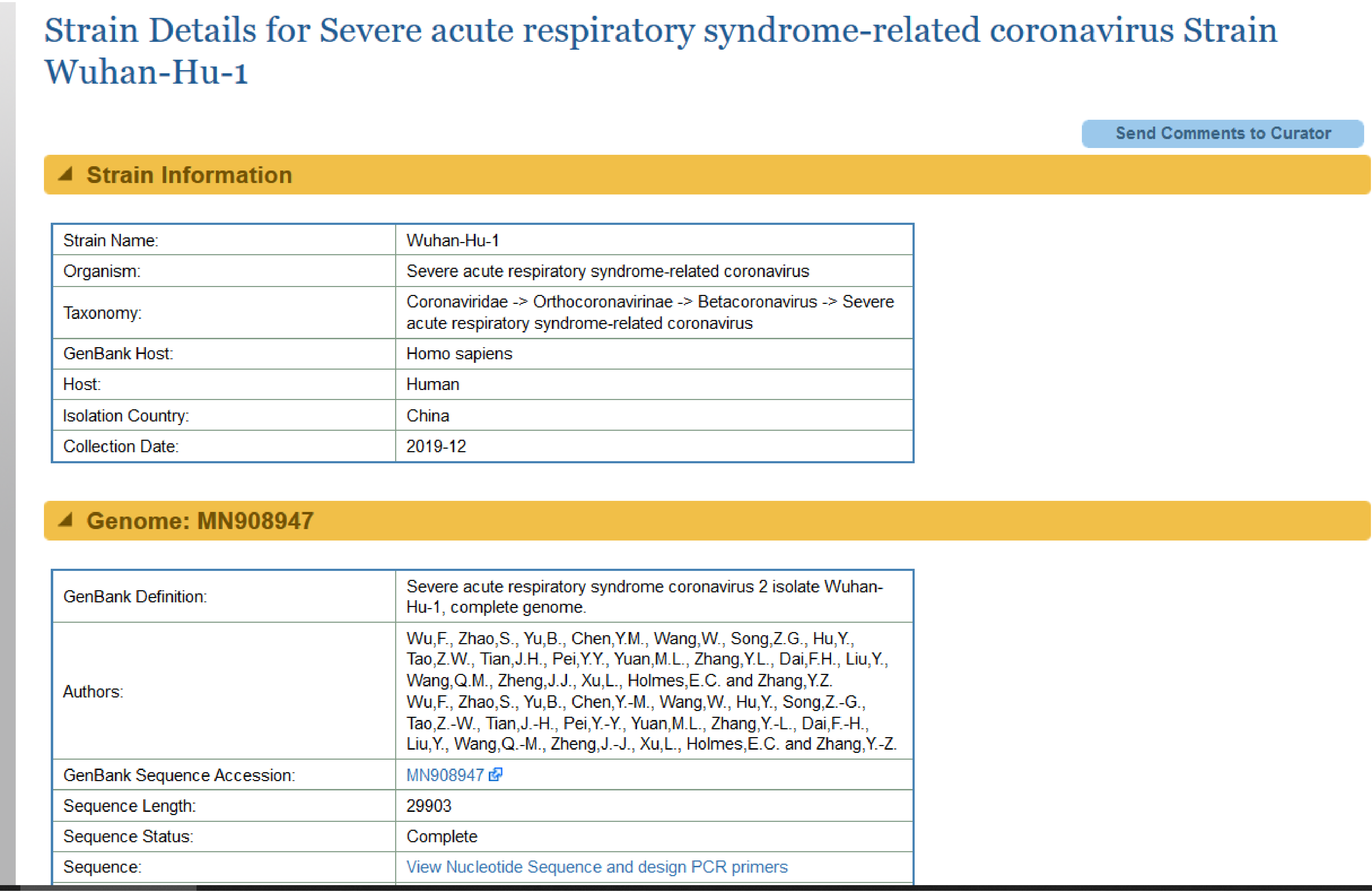
Genômica e Bioinformática no combate à pandemia de COVID-19

João Carlos Setubal

2021



© ViralZone 2020
SIB Swiss Instute of Bioinformatics



Strain Details for Severe acute respiratory syndrome-related coronavirus Strain Wuhan-Hu-1

Strain Information

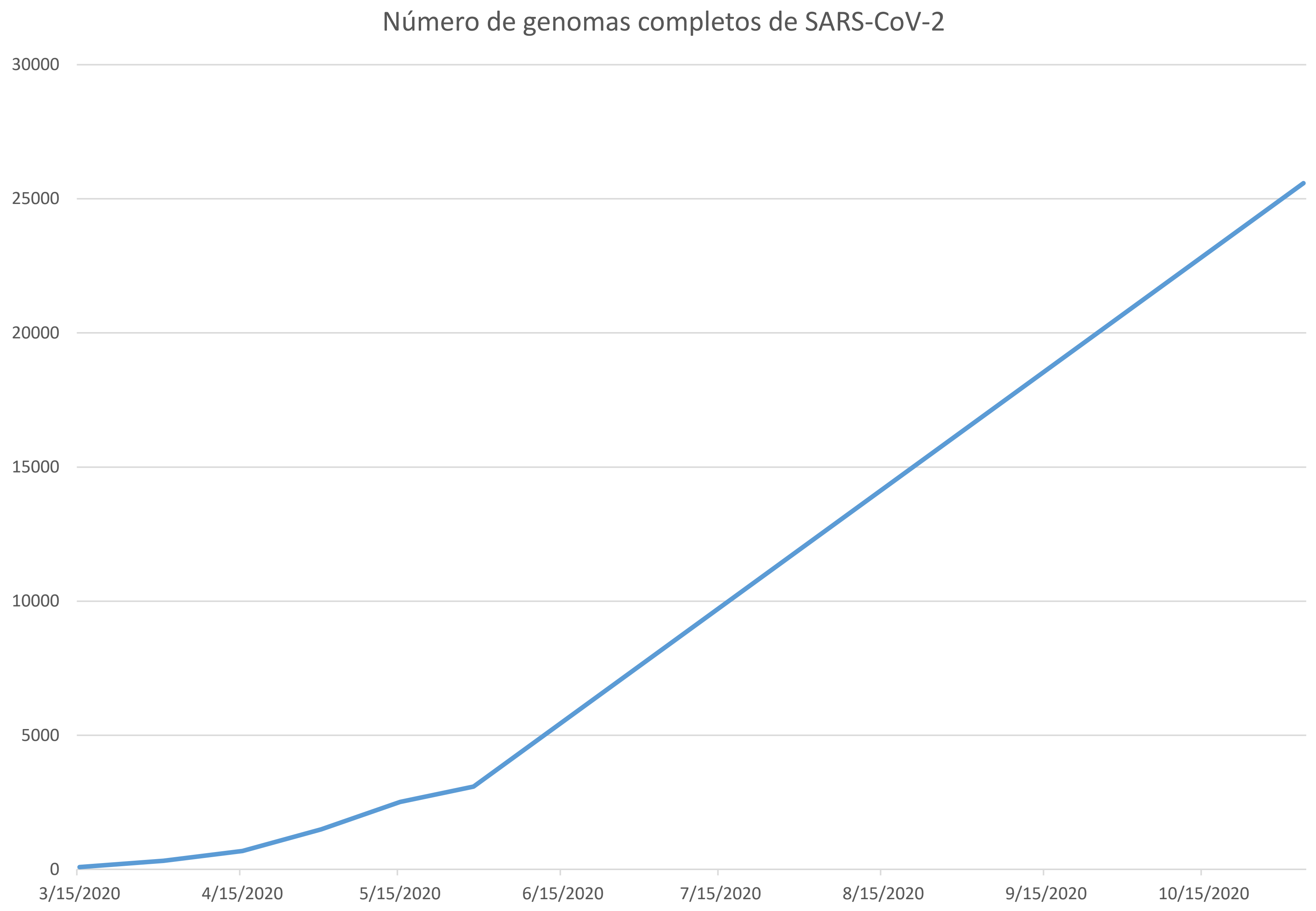
Strain Name:	Wuhan-Hu-1
Organism:	Severe acute respiratory syndrome-related coronavirus
Taxonomy:	Coronaviridae -> Orthocoronavirinae -> Betacoronavirus -> Severe acute respiratory syndrome-related coronavirus
GenBank Host:	Homo sapiens
Host:	Human
Isolation Country:	China
Collection Date:	2019-12

Genome: MN908947

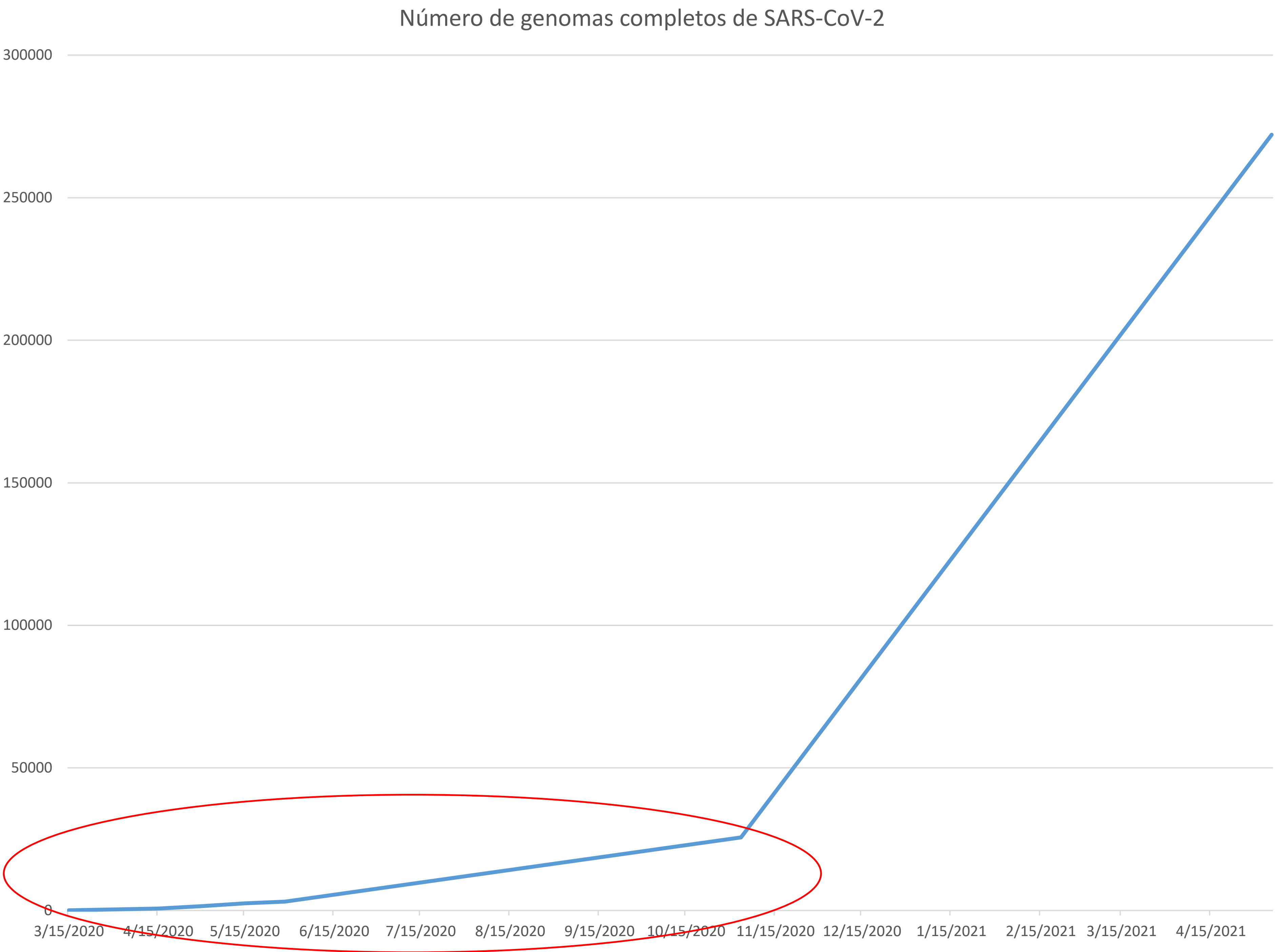
GenBank Definition:	Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1, complete genome.
Authors:	Wu,F., Zhao,S., Yu,B., Chen,Y.M., Wang,W., Song,Z.G., Hu,Y., Tao,Z.W., Tian,J.H., Pei,Y.Y., Yuan,M.L., Zhang,Y.L., Dai,F.H., Liu,Y., Wang,Q.M., Zheng,J.J., Xu,L., Holmes,E.C. and Zhang,Y.Z. Wu,F., Zhao,S., Yu,B., Chen,Y.-M., Wang,W., Hu,Y., Song,Z.-G., Tao,Z.-W., Tian,J.-H., Pei,Y.-Y., Yuan,M.L., Zhang,Y.-L., Dai,F.-H., Liu,Y., Wang,Q.-M., Zheng,J.-J., Xu,L., Holmes,E.C. and Zhang,Y.-Z.
GenBank Sequence Accession:	MN908947
Sequence Length:	29903
Sequence Status:	Complete
Sequence:	View Nucleotide Sequence and design PCR primers

A empresa MODERNA...

- ...desenvolveu sua vacina (pelo menos no início) sem sequer ter o SARS-CoV-2 em seus laboratórios...
- ...apenas a partir da sequência genômica!


Número de genomas completos de SARS-CoV-2
30000
25000
20000
15000
10000
5000
0
3/15/2020
4/15/2020
5/15/2020
6/15/2020
7/15/2020
8/15/2020
9/15/2020
10/15/2020

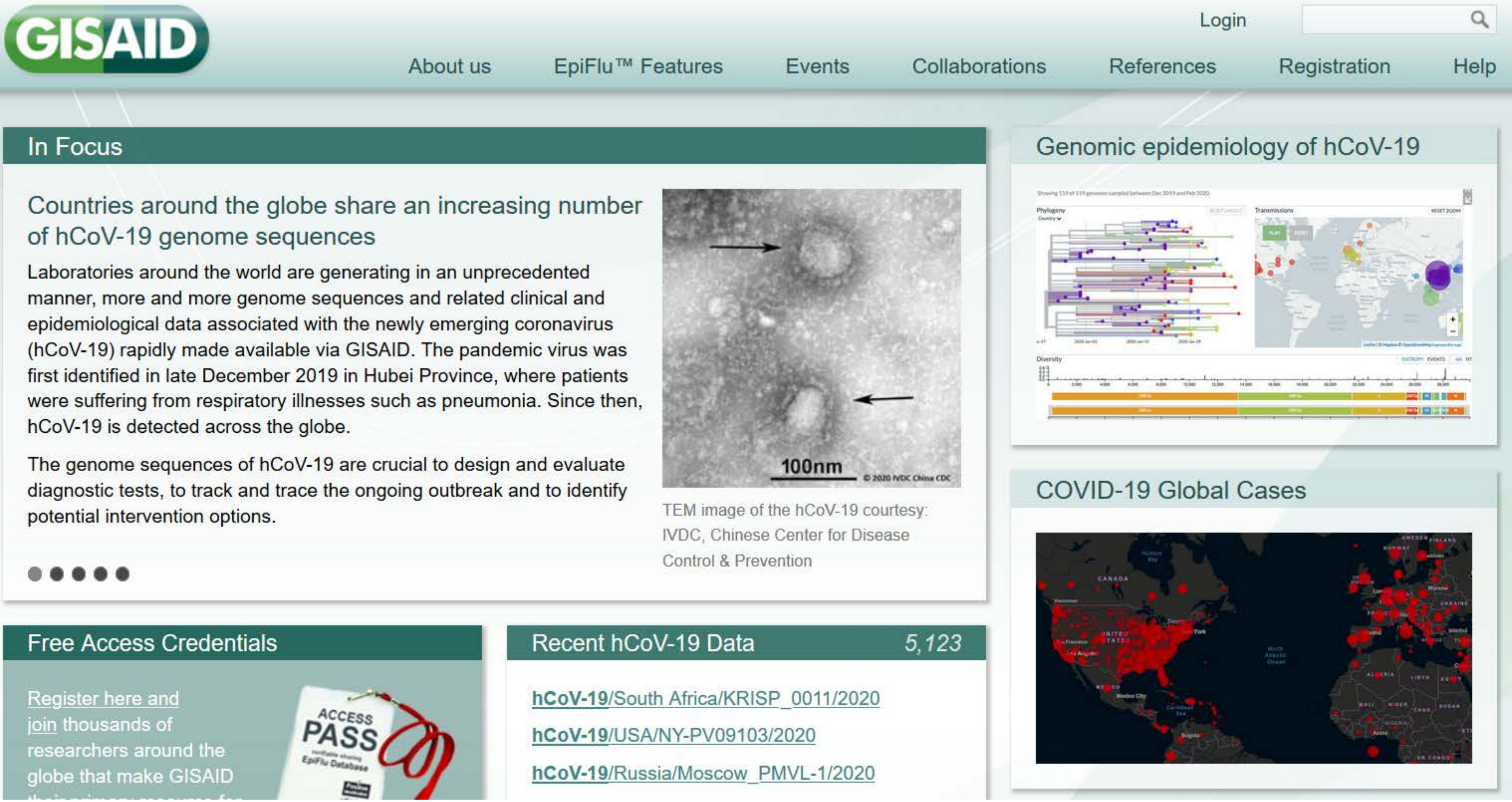
Do Brasil: 14


Número de genomas completos de SARS-CoV-2
300000
250000
200000
150000
100000
50000
0
3/15/2020
4/15/2020
5/15/2020
6/15/2020
7/15/2020
8/15/2020
9/15/2020
10/15/2020
11/15/2020
12/15/2020
1/15/2021
2/15/2021
3/15/2021
4/15/2021

Epidemiologia (ou vigilância) genômica

- qual é a geografia e o histórico da disseminação do virus e de suas variantes?
- depende crucialmente de sequenciamento massivo, contínuo, e sistemático
- comparação entre os mais de 400 mil genomas de SARS-CoV-2 que já foram sequenciados

https://www.gisaid.org/



In Focus

Countries around the globe share an increasing number of hCoV-19 genome sequences

Laboratories around the world are generating in an unprecedented manner, more and more genome sequences and related clinical and epidemiological data associated with the newly emerging coronavirus (hCoV-19) rapidly made available via GISAID. The pandemic virus was first identified in late December 2019 in Hubei Province, where patients were suffering from respiratory illnesses such as pneumonia. Since then, hCoV-19 is detected across the globe.

The genome sequences of hCoV-19 are crucial to design and evaluate diagnostic tests, to track and trace the ongoing outbreak and to identify potential intervention options.

TEM image of the hCoV-19 courtesy: IVDC, Chinese Center for Disease Control & Prevention

Genomic epidemiology of hCoV-19

COVID-19 Global Cases

Free Access Credentials

Register here and join thousands of researchers around the globe that make GISAID

Recent hCoV-19 Data *5,123*

hCoV-19/South Africa/KRISP_0011/2020

hCoV-19/USA/NY-PV09103/2020

hCoV-19/Russia/Moscow_PMVL-1/2020

https://nextstrain.org/

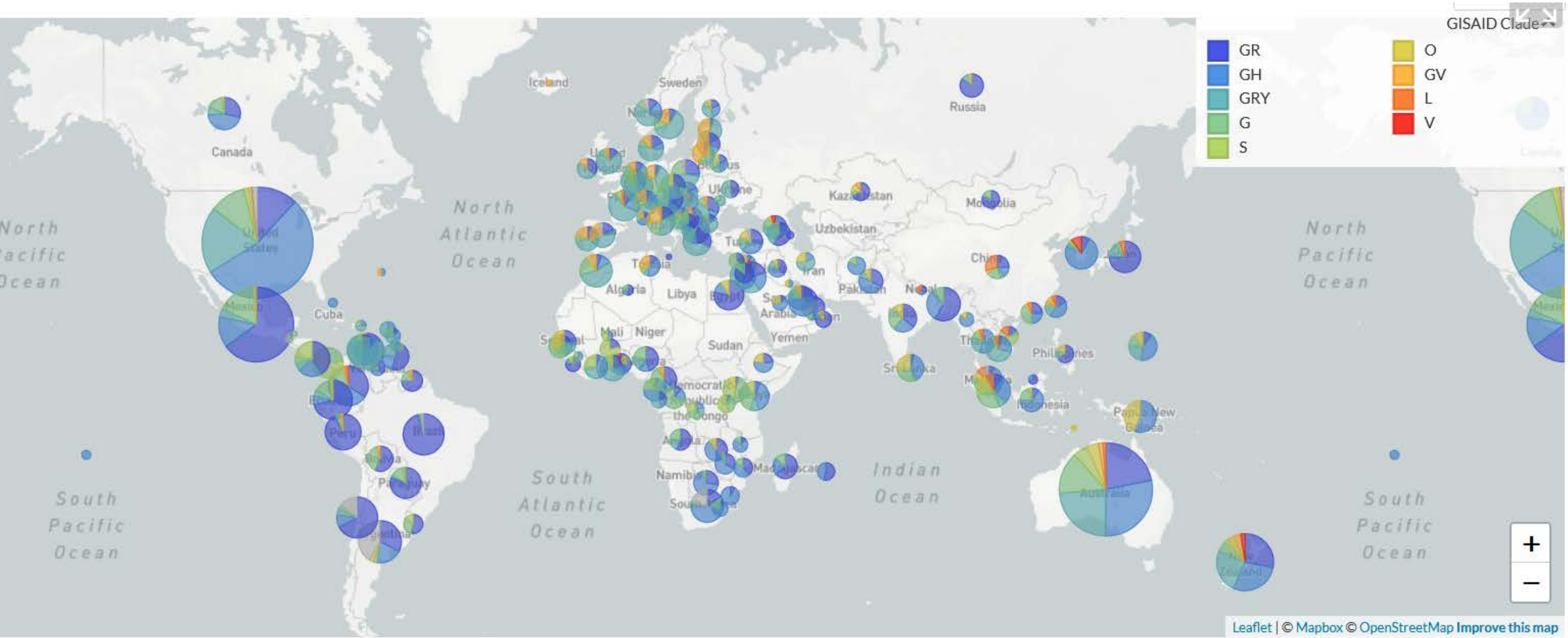
Nextstrain

Real-time tracking of pathogen evolution

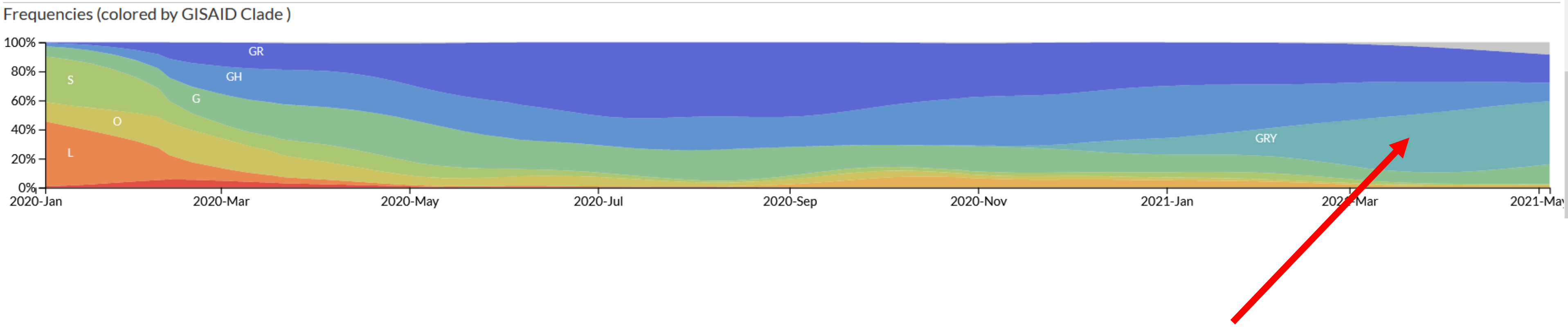
Nextstrain is an open-source project to harness the scientific and public health potential of pathogen genome data. We provide a continually-updated view of publicly available data alongside powerful analytic and visualization tools for use by the community. Our goal is to aid epidemiological understanding and improve outbreak response. If you have any questions, or simply want to say hi, please give us a shout at hello@nextstrain.org.

Novel coronavirus (2019-nCoV)

We are incorporating nCoV genomes as soon as they are shared and providing analyses and situation reports. Please see below for the latest updates.


GISAID Clade
GR
GH
GRY
G
S
O
GV
L
V
North Pacific Ocean
North Atlantic Ocean
South Pacific Ocean
South Atlantic Ocean
Indian Ocean
Canada
United States
Mexico
Cuba
Peru
Brazil
Bolivia
Paraguay
Argentina
Iceland
Sweden
Norway
United Kingdom
Belarus
Ukraine
Tunisia
Algeria
Libya
Egypt
Mali
Niger
Sudan
Senegal
Yemen
Saudi Arabia
Turkey
Iran
Kazakhstan
Uzbekistan
Russia
Mongolia
China
Japan
Pakistan
Nepal
India
Sri Lanka
Thailand
Malaysia
Philippines
Indonesia
Papua New Guinea
Australia
New Zealand
Democratic Republic of the Congo
Angola
Namibia
South Africa
Madagascar
Leaflet | © Mapbox © OpenStreetMap Improve this map
+
−

Linha do tempo: frequência dos clados



O clado GRY é o dominante atualmente

Latest *SARS-CoV-2* Variants and Lineages of Concern

Search

Search our comprehensive database for:

- Sequences & strains
- Immune epitopes
- 3D protein structures
- Host Factor Data
- Antiviral Drugs
- Plasmid Data

Analyze

Analyze data online:

- Sequence Alignment
- Phylogenetic Tree
- Sequence Variation (SNP)
- Metadata-driven Comparative Analysis
- BLAST
- VIGOR4 Genome Annotator

Save to Workbench

Sign up for a workbench to:

- Store and share data
- Combine working sets
- Integrate your data with ViPR data
- Store and share analyses
- Custom search alert

Supported Programs

Click on a supported program of interest to go to program description page.

NIAID Functional Genomics

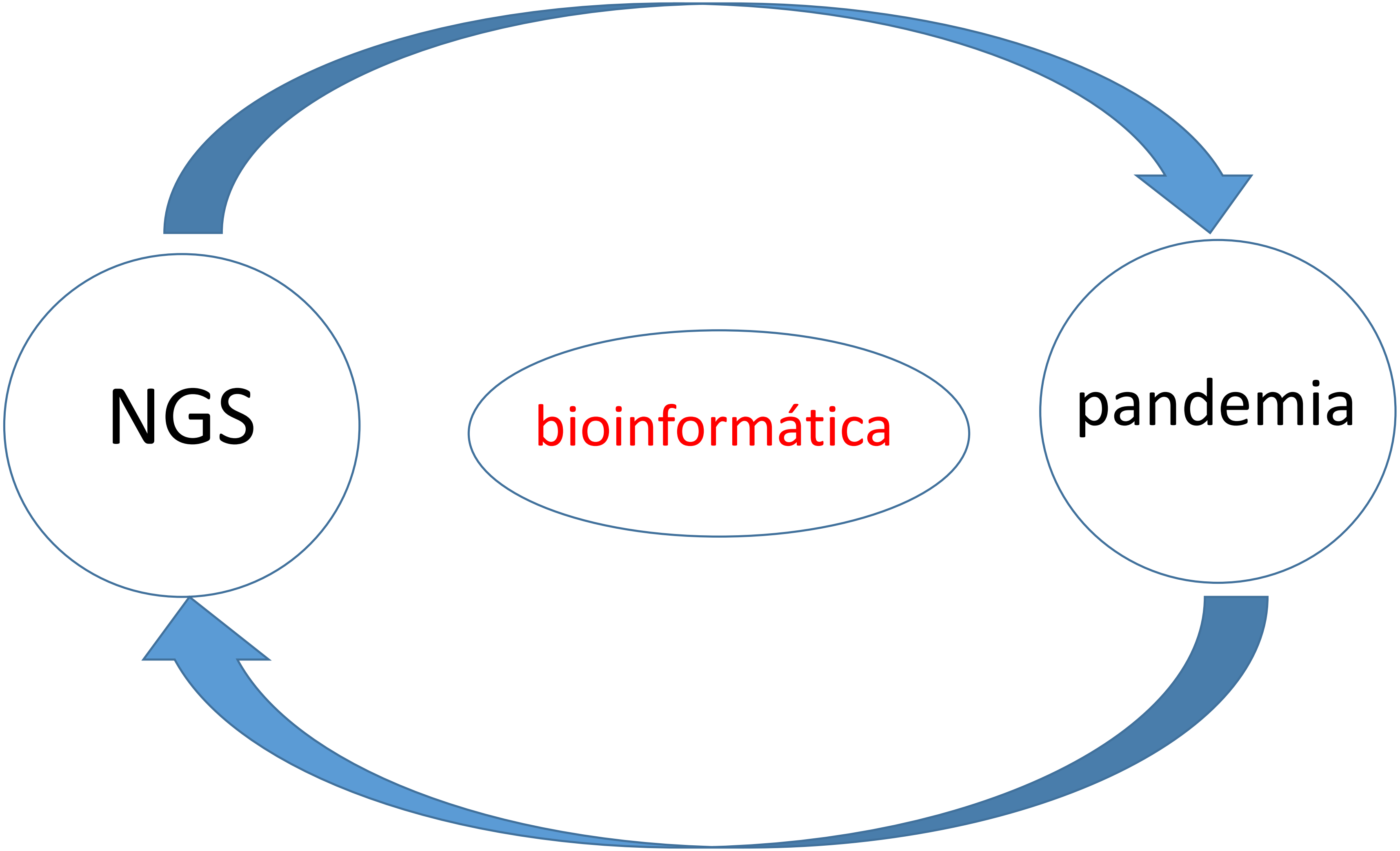
NIAID Systems Biology

Virus Families

Click on icon of family or species of interest. Click here to to view all families and species in list format. Don't know family of species? Provide species name

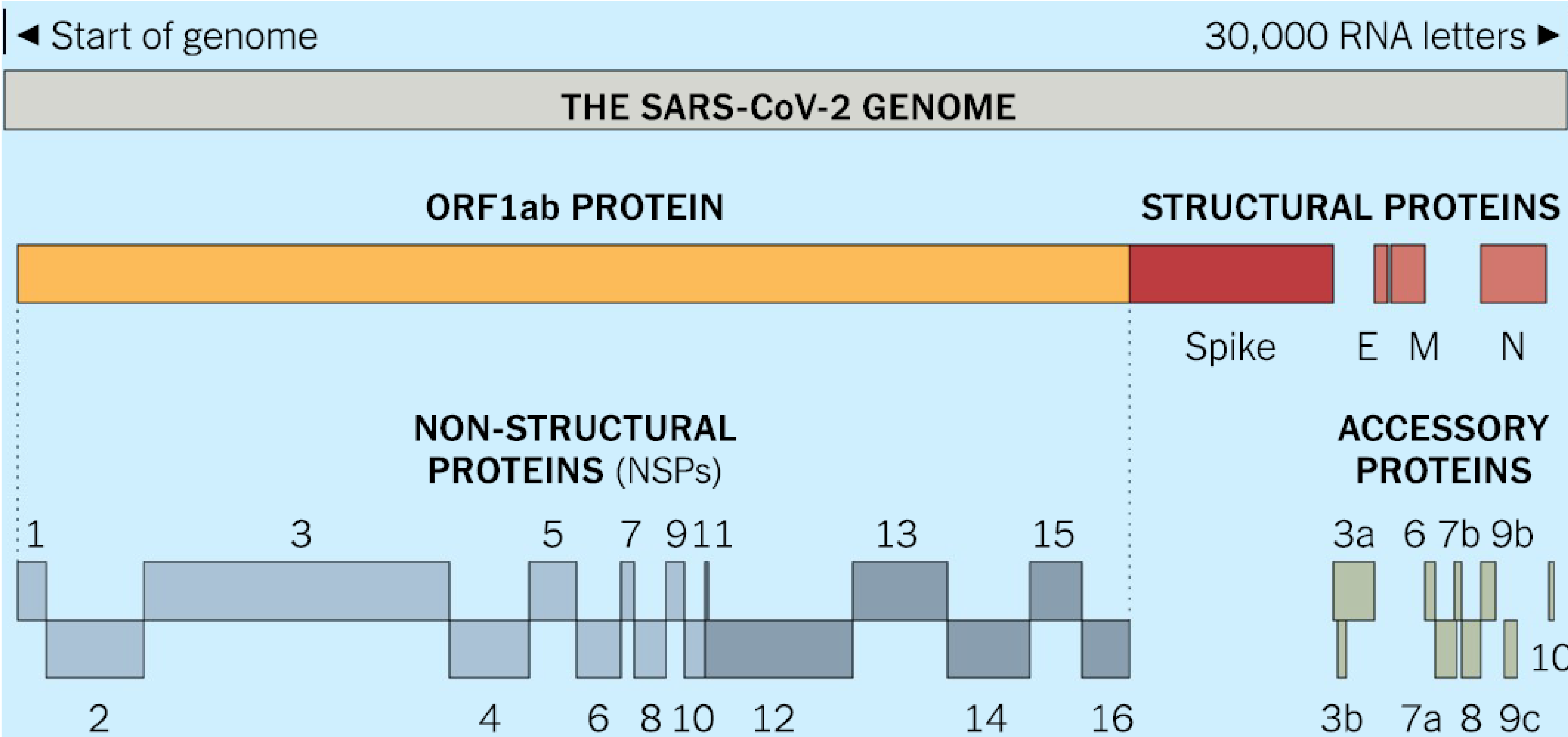
https://www.viprbrc.org/

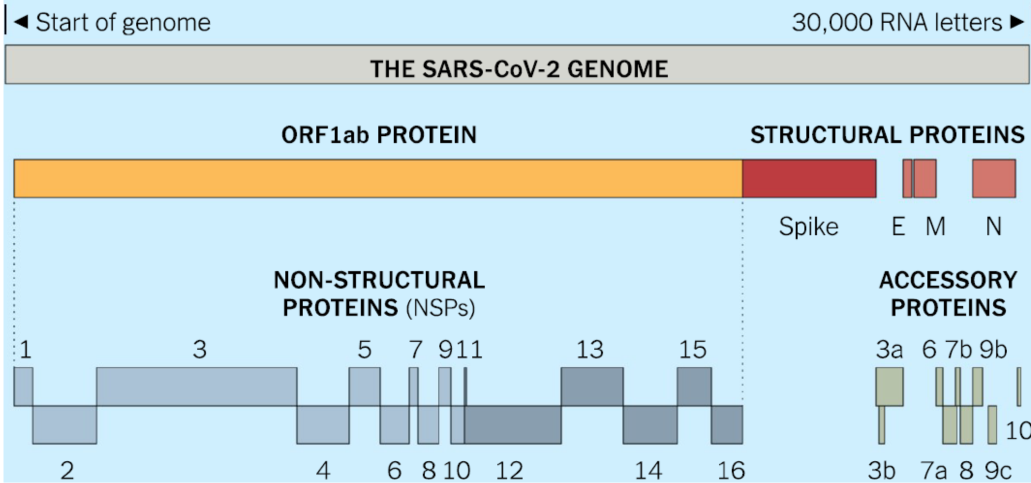
Sinergia!



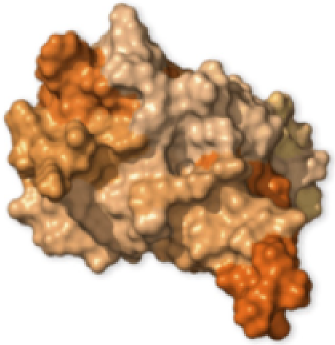
Anatomia do genoma de SARS-CoV-2

- artigo do New York Times **3 de abril de 2020**
- autores: Jonathan Corum and Carl Zimmer
- https://www.nytimes.com/interactive/2020/04/03/science/coronavirus-genome-bad-news-wrapped-in-protein.html?action=click&module=Top%20Stories&pgtype=Homepage


◄ Start of genome
30,000 RNA letters ►
THE SARS-CoV-2 GENOME
ORF1ab PROTEIN
STRUCTURAL PROTEINS
Spike
E
M
N
NON-STRUCTURAL PROTEINS (NSPs)
ACCESSORY PROTEINS
1
3
5
7
9
11
13
15
2
4
6
8
10
12
14
16
3a
6
7b
9b
1
3b
7a
8
9c

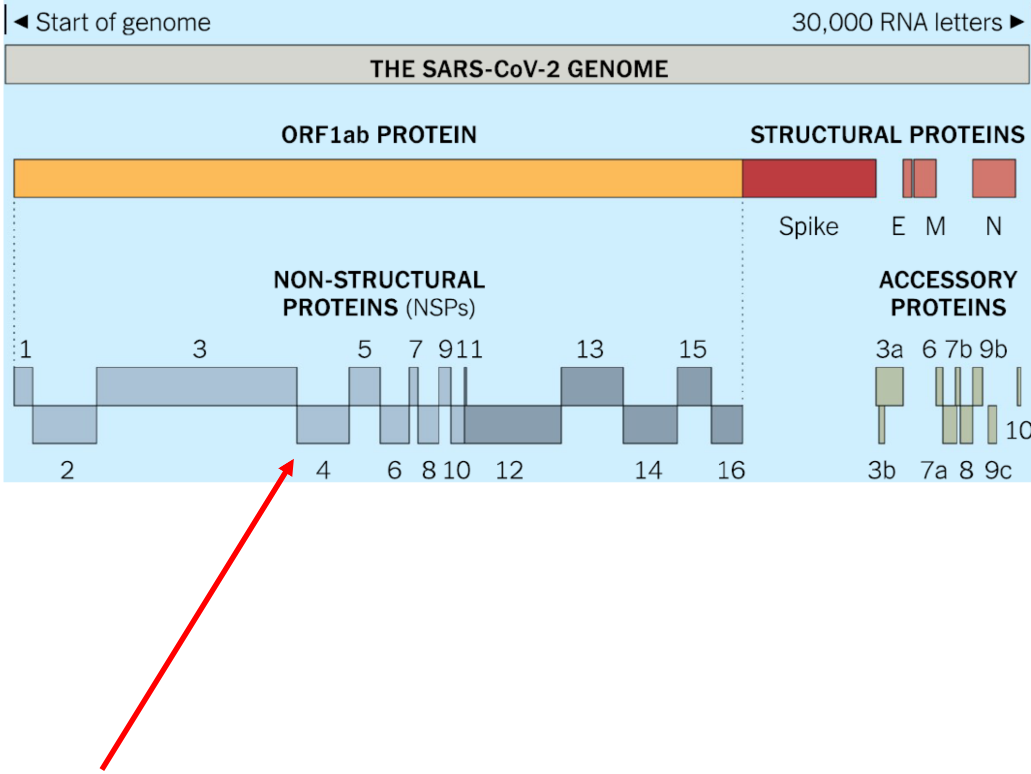


Cellular Saboteur · NSP1

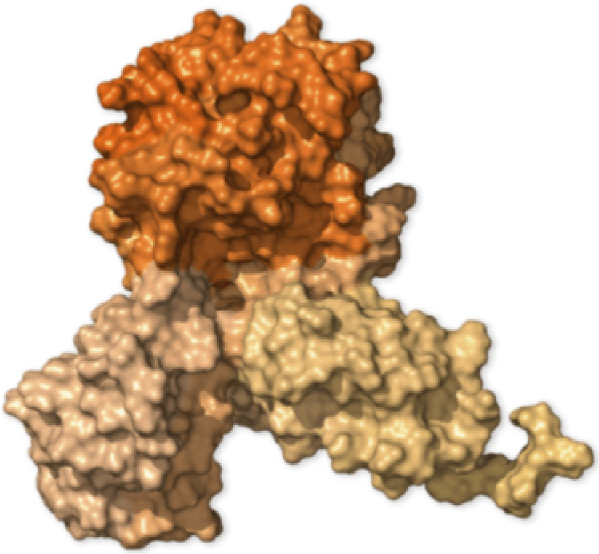


This protein slows down the infected cell's production of its own proteins. This sabotage forces the cell to make more virus proteins and prevents it from assembling antiviral proteins that could stop the virus.

```
auggagagccuugucccugguuucaacgagaaaacacacguccaacucaguuugccuguuuuacagguu
cgcgacgugcucguacguggcuuuggagacuccguggaggaggucuuaucagaggcacgucaacaucuu
aaagauggcacuuguggcuuaguagaaguugaaaaaggcguuuugccucaacuugaacagcccuaugug
uucaucaaacguucggaugcucgaacugcaccucauggucauguuaugguugagcugguagcagaacuc
gaaggcauucaguacggucguagugguggagacacuugguguccuugucccucaugugggcgaaauacca
guggcuuaccgcaagguucuucuucguaagaacgguaauaaaggagcugguggccauaguuacggcgcc
gaucuaaagucauuugacuuaggcgacgagcuuggcacugauccuuaugaagauuuucaagaaaacugg
aacacuaaacauagcagugguguuacccgugaacucaugcgugagcuuaacggaggg
```

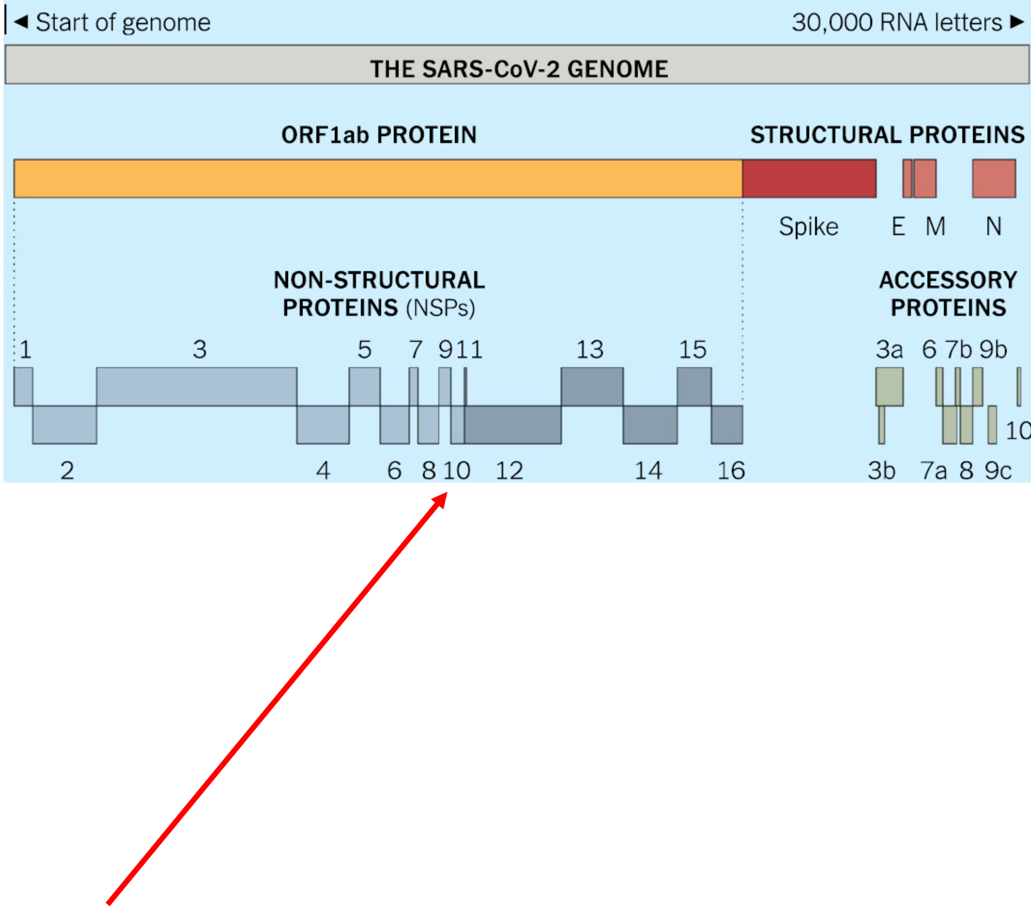


Bubble Maker · NSP4

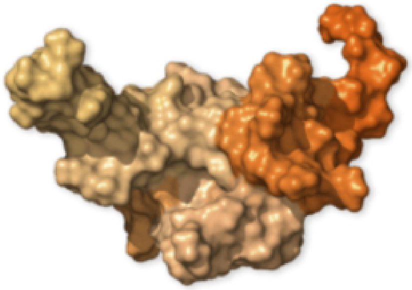


Combining with other proteins, NSP4 helps build fluid-filled bubbles within infected cells. Inside these bubbles, parts for new copies of the virus are constructed.

aaaauuguuaauaauugguugaagcaguuaauuaaaguuacacuuguguuccuuuuuguugcuauu
uucuauuuaauaacaccuguucaugucaugucuaaacauacugacuuuucaagugaaaucauaggauac
aaggcuauugauggugguguucacucgugacauagcaucuacagauacuuguuuugcuaacaaacaugcu
gauuuugacacaugguuuagccagcgugguggua guuauacuaaugacaaagcuugcccauugauugcu
gcagucauaacaagagaagugggguuuugucgugccugguuugccuggcacgauauuacgcacaacuaau
ggugacuuuuugcauuucuuaccuagaguuuuagugcaguugguaacaucuguuacacaccaucaaaa
cuuauagaguacacugacuuugcaacaucagcuuguguuuuggcugcugaauguacaauuuuuaaagau
gcuucugguaagccaguaccauauuguuaugauaccaauguacuagaagguucuguugcuuaugaaagu



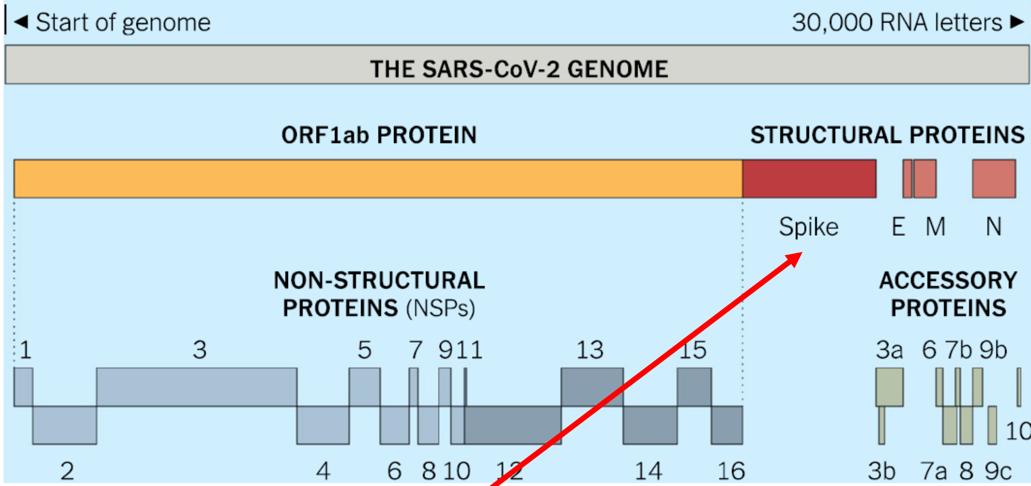
Genetic Camouflage · NSP10

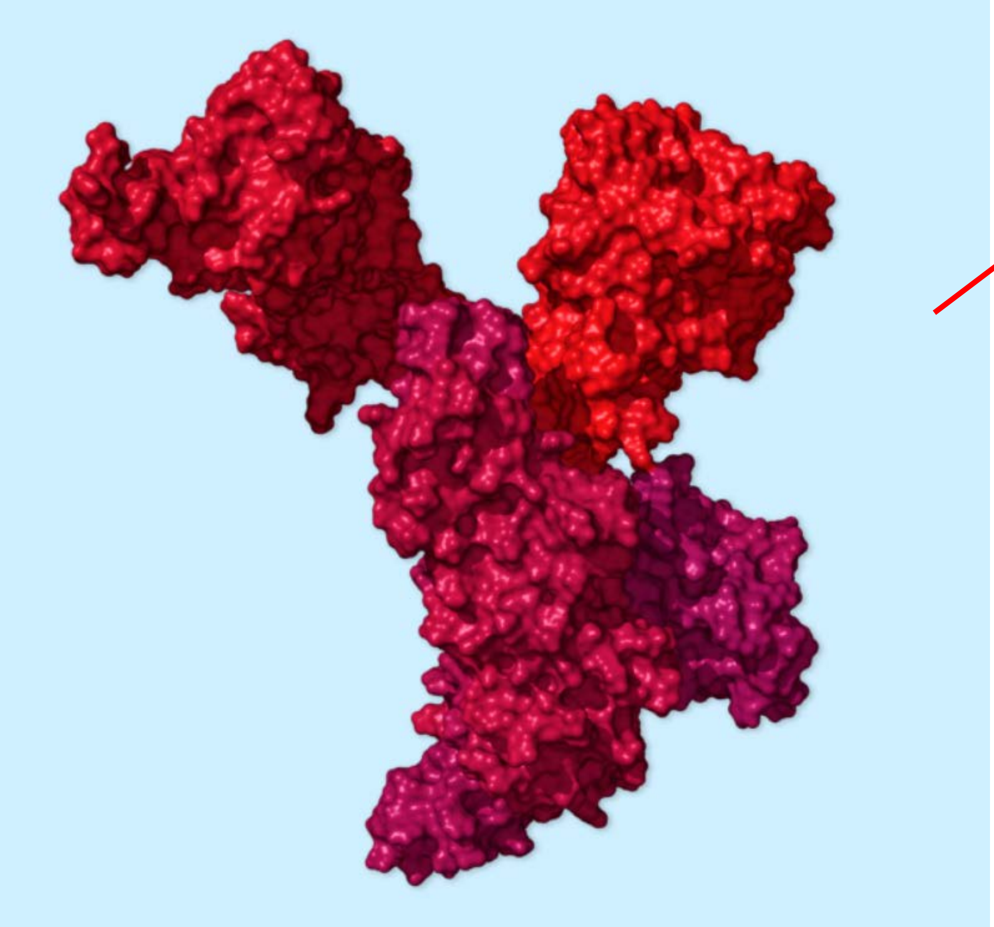


Human cells have antiviral proteins that find viral RNA and shred it. This protein works with NSP16 to camouflage the virus's genes so that they don't get attacked.

```
gcugguaaugcaacagaagugccugccaauucaacuguauuaucuuucugugcuuuugcuguagaugcu
gcuaaagcuuacaaagauuaucuagcuagugggggacaaccaaucacuaauuguguuaagauguugugu
acacacacugguacuggucaggcaauaacaguuacaccggaagccaauauggaucaagaauccuuuggu
ggugcaucguguugucuguacugccguugccacauagaucauccaaauccuaaaggauuuugugacuua
aaagguaaguauguacaaauaccuacaacuugugcuaaugacccugugguuuuacacuuaaaaacaca
gucuguaccgucugcgguauguggaaagguuauggcuguaguugugaucaacuccgcgaacccaugcuu
cag
```

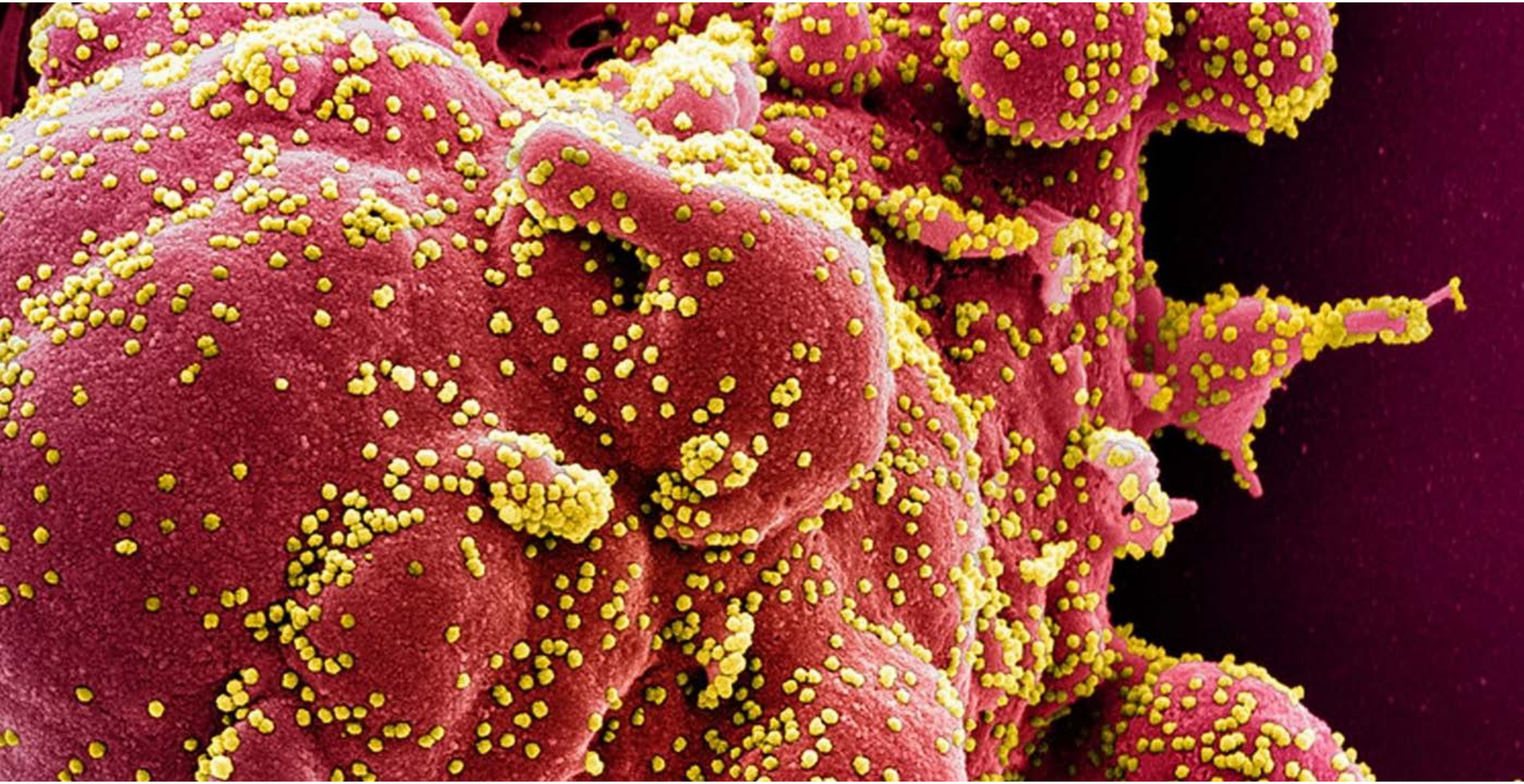
Spike protein S



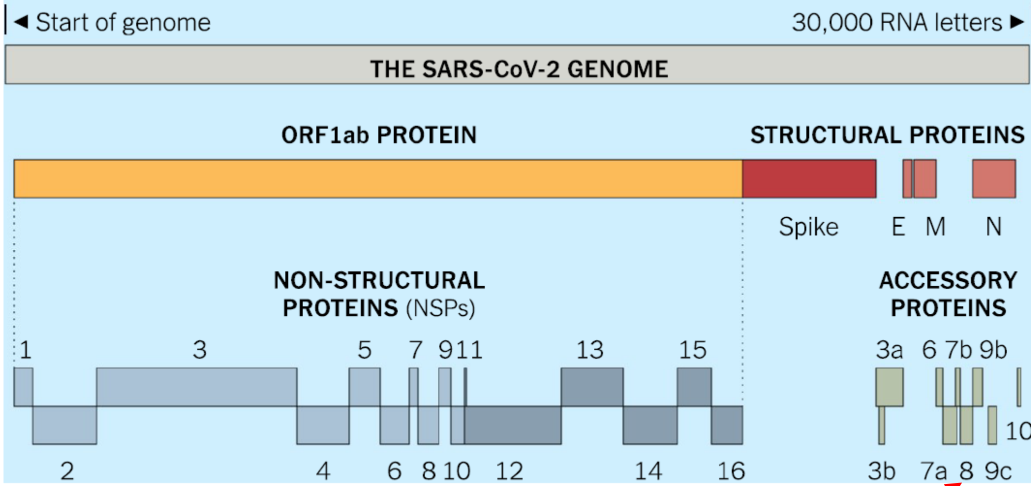


ACE2

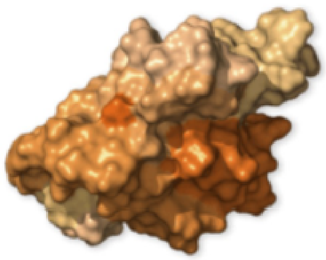
Angiotensin Converting Enzyme



National Institutes of Health/EPA, via Shutterstock



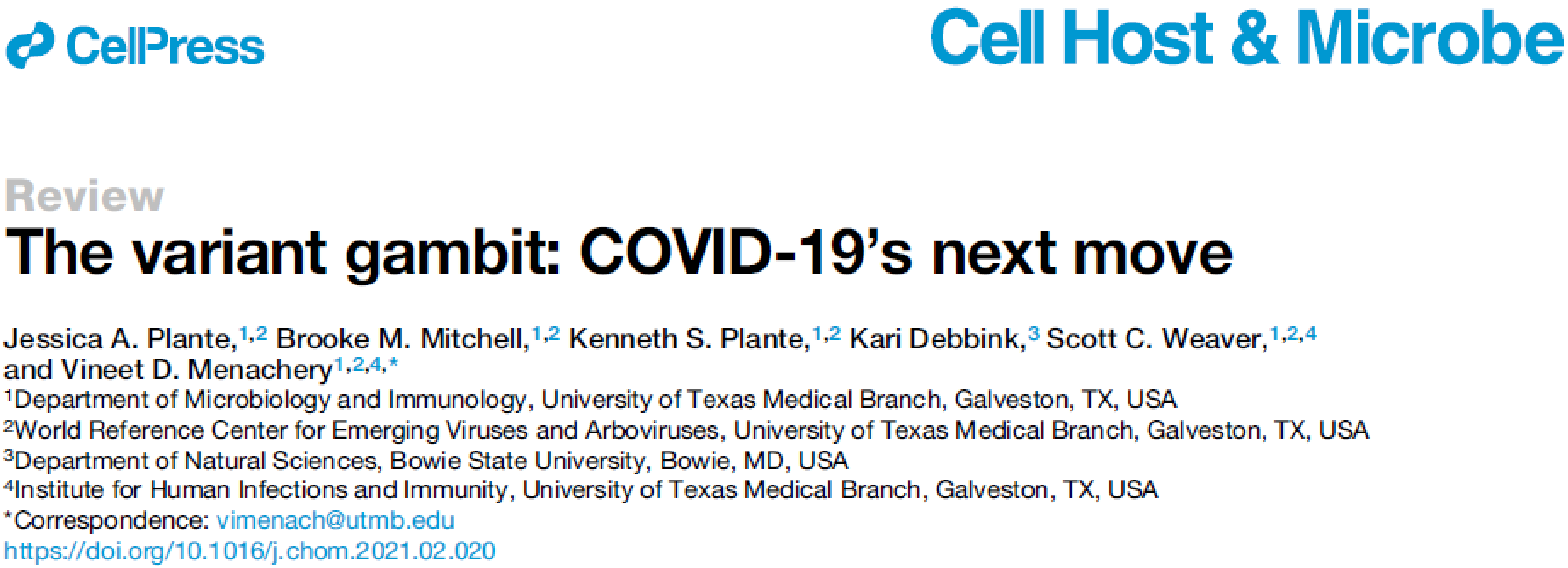
Mystery Protein · ORF8



The gene for this accessory protein is dramatically different in SARS-CoV-2 than in other coronaviruses. Researchers are debating what it does.

```
augaaauuucuuguuuucuuaggaaucaucacaacuguagcugcauuucaccaagaauguaguuuacag
ucauguacucaacaucaaccauauguaguugaugacccguguccuauucacuucuauucuaaaugguau
auuagaguaggagcuagaaaaucagcaccuuuaauugaauugugcguggaugaggcugguucuaaauca
cccauucaguacaucgauaucgguaauuauacaguuuccuguuuaccuuuuacaauuaauugccaggaa
ccuaaauuggguagucuuguagugcguuguucguucuaugaagacuuuuuagaguaucaugacguucgu
guuguuuuagauuucaucuaaacgaacaaacuaaa
```

Variantes



CellPress

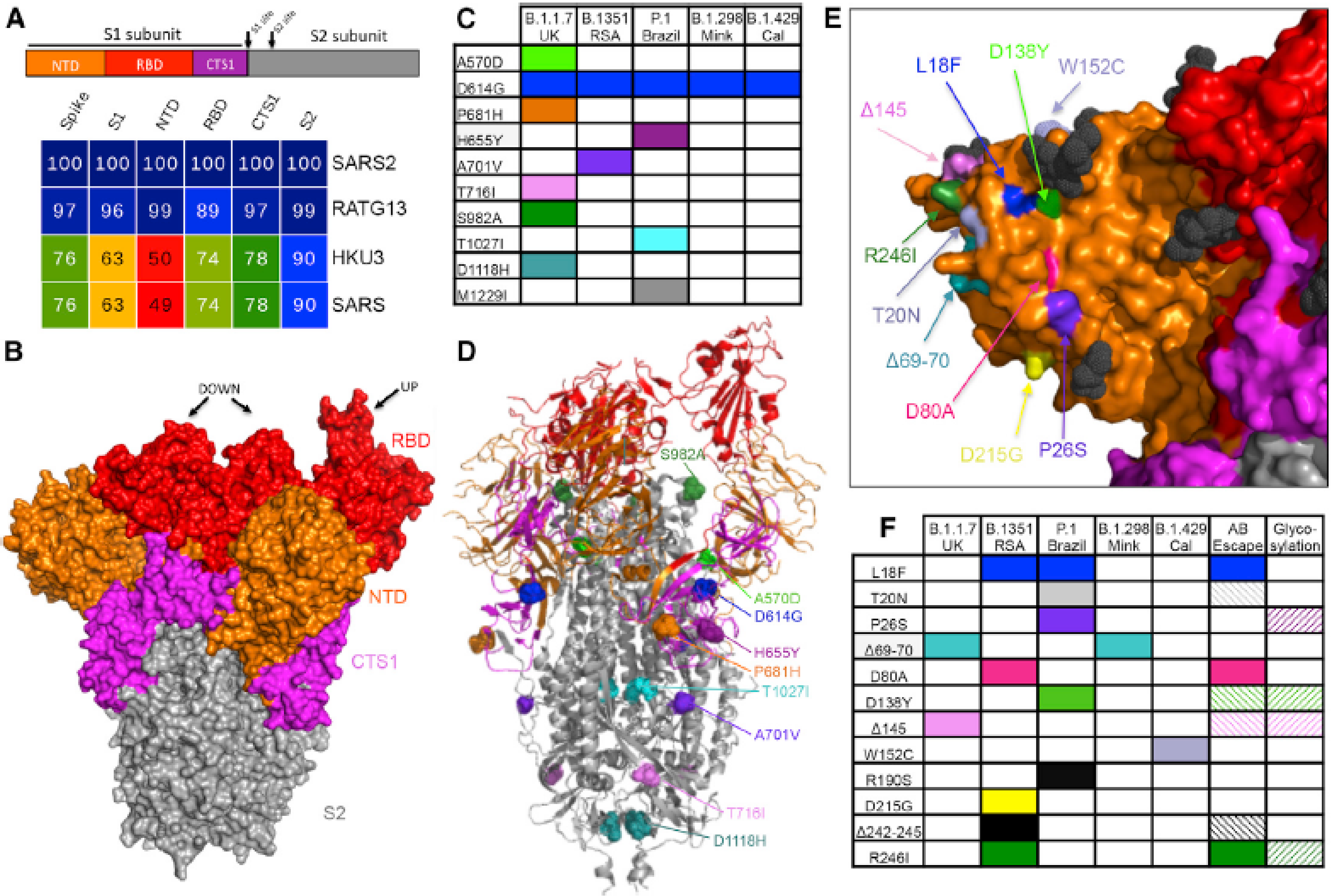
Cell Host & Microbe

Review

The variant gambit: COVID-19's next move

Jessica A. Plante,[1,2] Brooke M. Mitchell,[1,2] Kenneth S. Plante,[1,2] Kari Debbink,[3] Scott C. Weaver,[1,2,4] and Vineet D. Menachery[1,2,4,*]
[1]Department of Microbiology and Immunology, University of Texas Medical Branch, Galveston, TX, USA
[2]World Reference Center for Emerging Viruses and Arboviruses, University of Texas Medical Branch, Galveston, TX, USA
[3]Department of Natural Sciences, Bowie State University, Bowie, MD, USA
[4]Institute for Human Infections and Immunity, University of Texas Medical Branch, Galveston, TX, USA
*Correspondence: vimenach@utmb.edu

https://doi.org/10.1016/j.chom.2021.02.020



A

	Spike	S1	NTD	RBD	CTS1	S2
SARS2	100	100	100	100	100	100
RATG13	97	96	99	89	97	99
HKU3	76	63	50	74	78	90
SARS	76	63	49	74	78	90

C

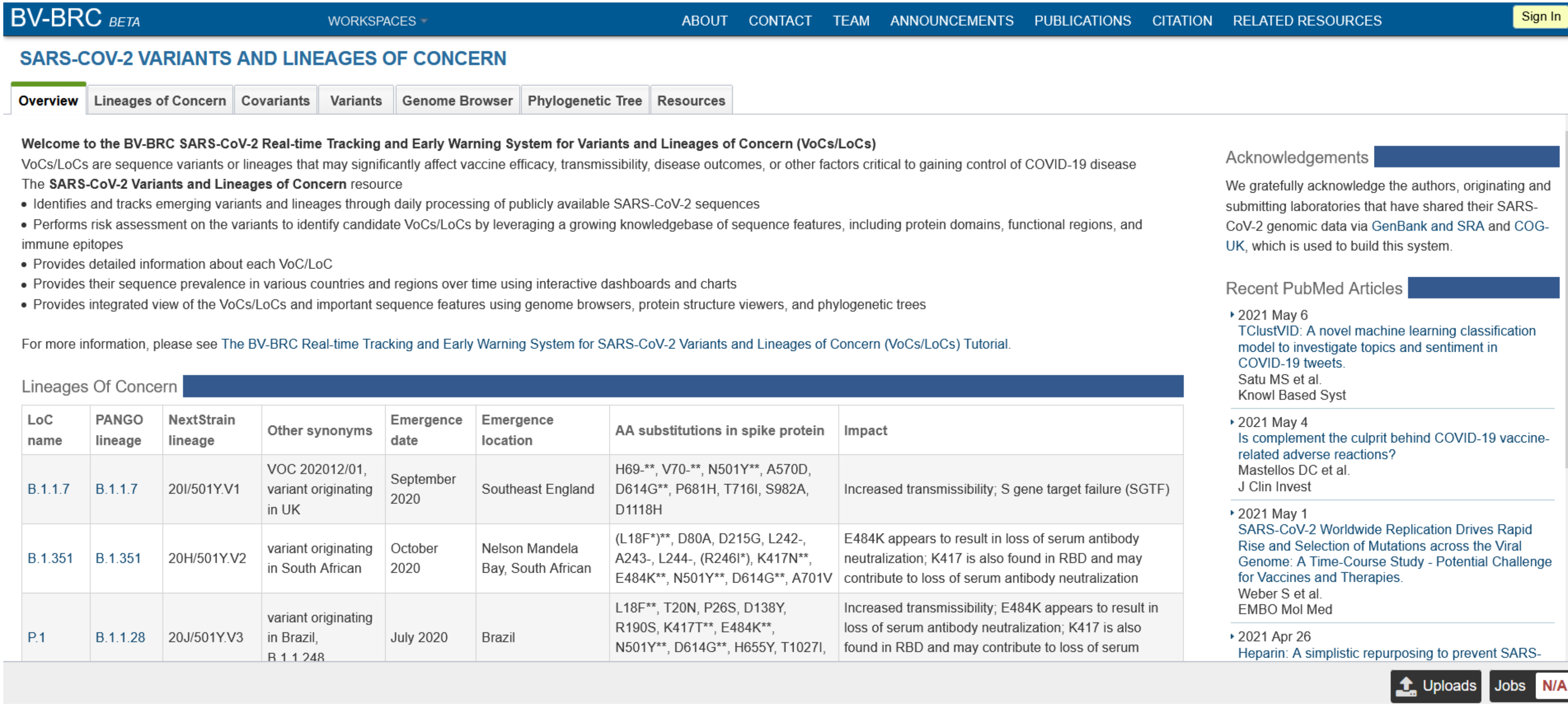
	B.1.1.7 UK	B.1351 RSA	P.1 Brazil	B.1.298 Mink	B.1.429 Cal
A570D	■				
D614G	■	■	■	■	■
P681H	■				
H655Y			■		
A701V		■			
T716I	■				
S982A	■				
T1027I			■		
D1118H	■				
M1229I			■		

F

	B.1.1.7 UK	B.1351 RSA	P.1 Brazil	B.1.298 Mink	B.1.429 Cal	AB Escape	Glyco-sylation
L18F		■	■			■	
T20N			■			▨	
P26S			■				▨
Δ69-70	■			■			
D80A		■				■	
D138Y			■			▨	▨
Δ145	■					▨	▨
W152C					■		
R190S			■				
D215G		■					
Δ242-245		■				▨	
R246I		■				■	▨

Figure 1. Variant substitutions in the conserved and NTD portions of the spike

The Bacterial and Viral Bioinformatics Resource Center (BV-BRC) https://bv-brc.org



SARS-COV-2 VARIANTS AND LINEAGES OF CONCERN

Welcome to the BV-BRC SARS-CoV-2 Real-time Tracking and Early Warning System for Variants and Lineages of Concern (VoCs/LoCs)

VoCs/LoCs are sequence variants or lineages that may significantly affect vaccine efficacy, transmissibility, disease outcomes, or other factors critical to gaining control of COVID-19 disease

The **SARS-CoV-2 Variants and Lineages of Concern** resource

- Identifies and tracks emerging variants and lineages through daily processing of publicly available SARS-CoV-2 sequences
- Performs risk assessment on the variants to identify candidate VoCs/LoCs by leveraging a growing knowledgebase of sequence features, including protein domains, functional regions, and immune epitopes
- Provides detailed information about each VoC/LoC
- Provides their sequence prevalence in various countries and regions over time using interactive dashboards and charts
- Provides integrated view of the VoCs/LoCs and important sequence features using genome browsers, protein structure viewers, and phylogenetic trees

For more information, please see The BV-BRC Real-time Tracking and Early Warning System for SARS-CoV-2 Variants and Lineages of Concern (VoCs/LoCs) Tutorial.

Lineages Of Concern

LoC name	PANGO lineage	NextStrain lineage	Other synonyms	Emergence date	Emergence location	AA substitutions in spike protein	Impact
B.1.1.7	B.1.1.7	20I/501Y.V1	VOC 202012/01, variant originating in UK	September 2020	Southeast England	H69-**, V70-**, N501Y**, A570D, D614G**, P681H, T716I, S982A, D1118H	Increased transmissibility; S gene target failure (SGTF)
B.1.351	B.1.351	20H/501Y.V2	variant originating in South African	October 2020	Nelson Mandela Bay, South African	(L18F*)**, D80A, D215G, L242-, A243-, L244-, (R246I*), K417N**, E484K**, N501Y**, D614G**, A701V	E484K appears to result in loss of serum antibody neutralization; K417 is also found in RBD and may contribute to loss of serum antibody neutralization
P.1	B.1.1.28	20J/501Y.V3	variant originating in Brazil, B.1.1.248	July 2020	Brazil	L18F**, T20N, P26S, D138Y, R190S, K417T**, E484K**, N501Y**, D614G**, H655Y, T1027I,	Increased transmissibility; E484K appears to result in loss of serum antibody neutralization; K417 is also found in RBD and may contribute to loss of serum

Acknowledgements

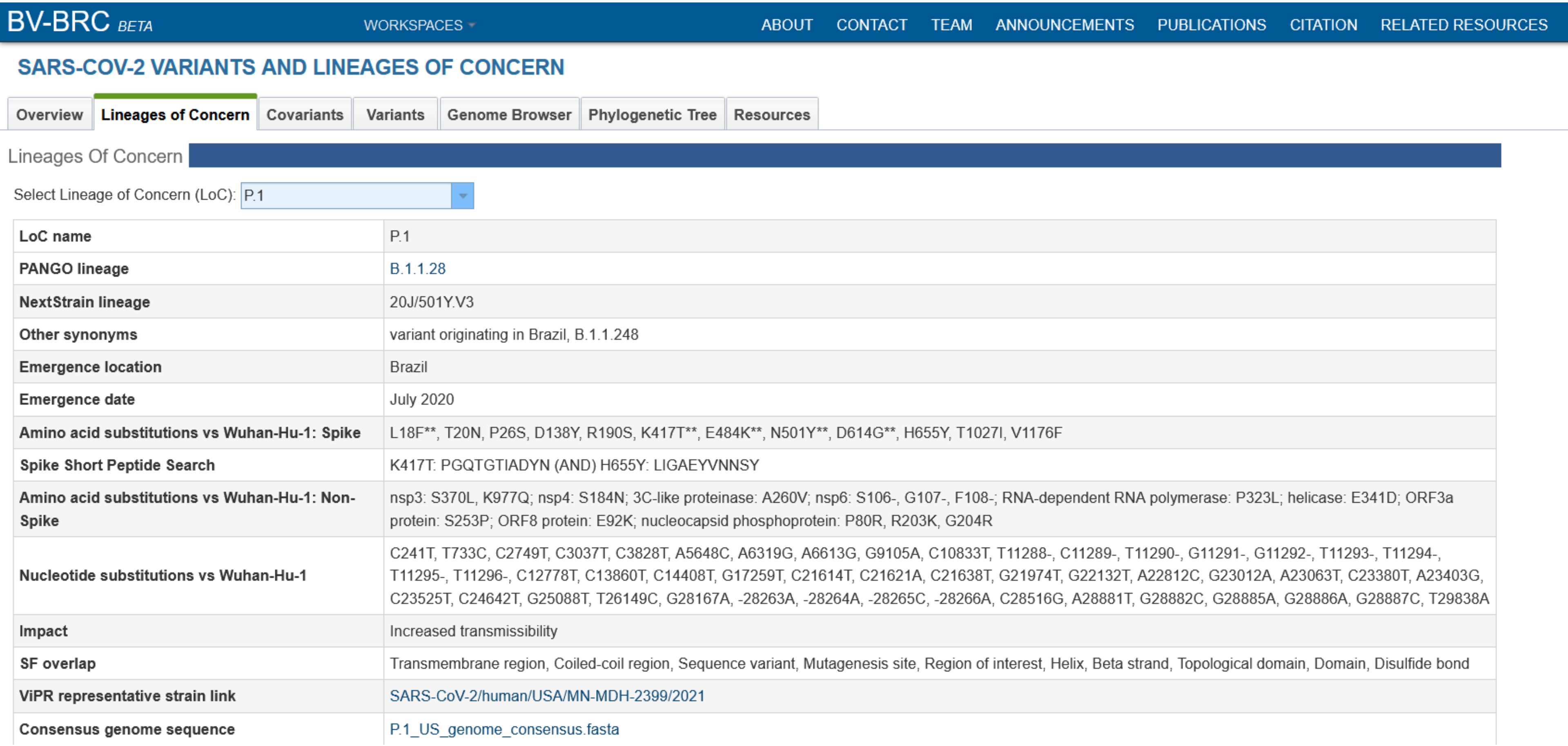
We gratefully acknowledge the authors, originating and submitting laboratories that have shared their SARS-CoV-2 genomic data via GenBank and SRA and COG-UK, which is used to build this system.

Recent PubMed Articles

- 2021 May 6
  TClustVID: A novel machine learning classification model to investigate topics and sentiment in COVID-19 tweets.
  Satu MS et al.
  Knowl Based Syst
- 2021 May 4
  Is complement the culprit behind COVID-19 vaccine-related adverse reactions?
  Mastellos DC et al.
  J Clin Invest
- 2021 May 1
  SARS-CoV-2 Worldwide Replication Drives Rapid Rise and Selection of Mutations across the Viral Genome: A Time-Course Study - Potential Challenge for Vaccines and Therapies.
  Weber S et al.
  EMBO Mol Med
- 2021 Apr 26
  Heparin: A simplistic repurposing to prevent SARS-

The **SARS-CoV-2 Variants and Lineages of Concern** resource

- Identifies and tracks emerging variants and lineages through daily processing of publicly available SARS-CoV-2 sequences
- Performs risk assessment on the variants to identify candidate VoCs/LoCs by leveraging a growing knowledgebase of sequence features, including protein domains, functional regions, and immune epitopes
- Provides detailed information about each VoC/LoC
- Provides their sequence prevalence in various countries and regions over time using interactive dashboards and charts
- Provides integrated view of the VoCs/LoCs and important sequence features using genome browsers, protein structure viewers, and phylogenetic trees
- **Tudo isto depende de sequenciamento contínuo (vigilância genômica)**

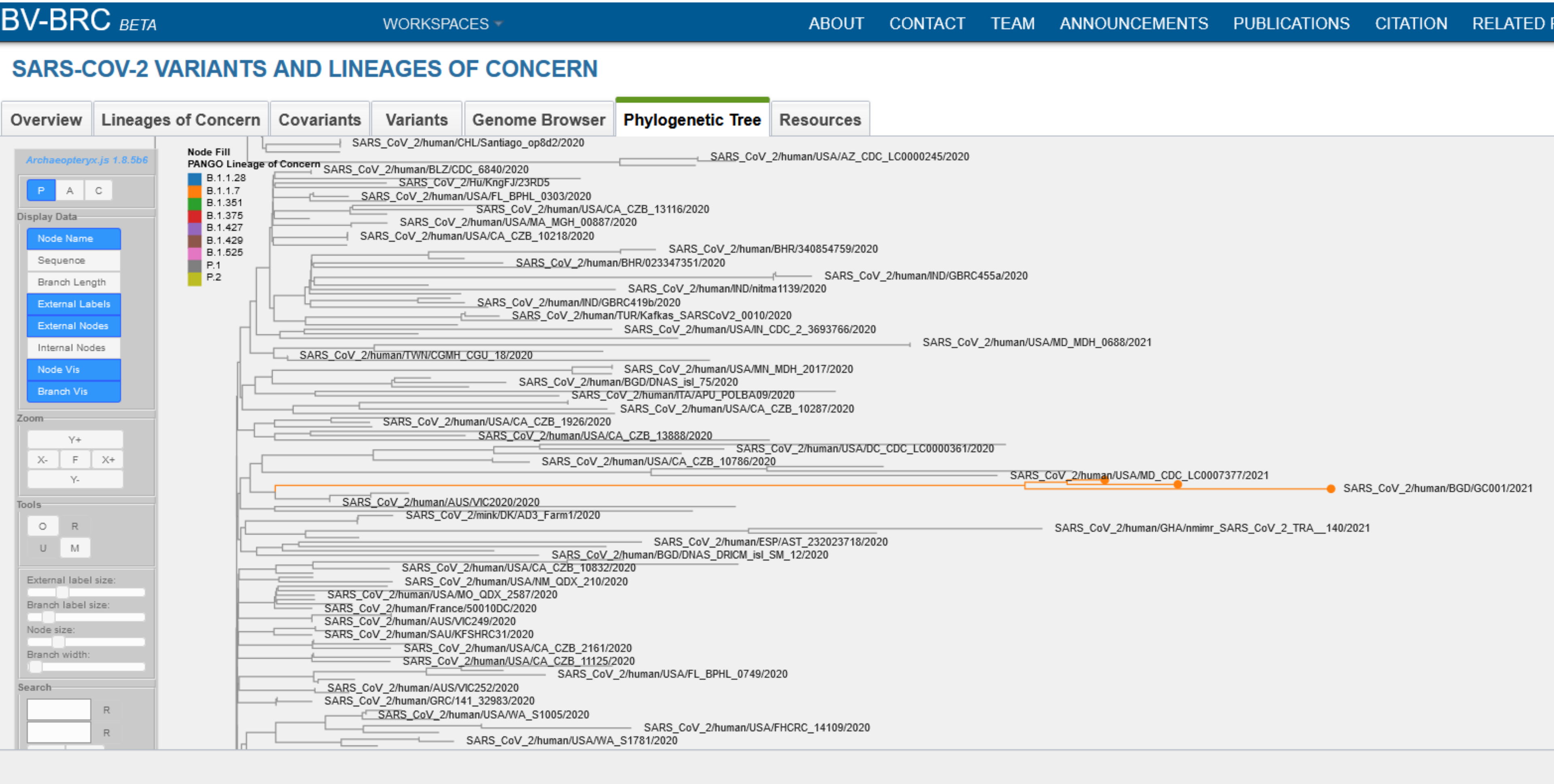


SARS-COV-2 VARIANTS AND LINEAGES OF CONCERN

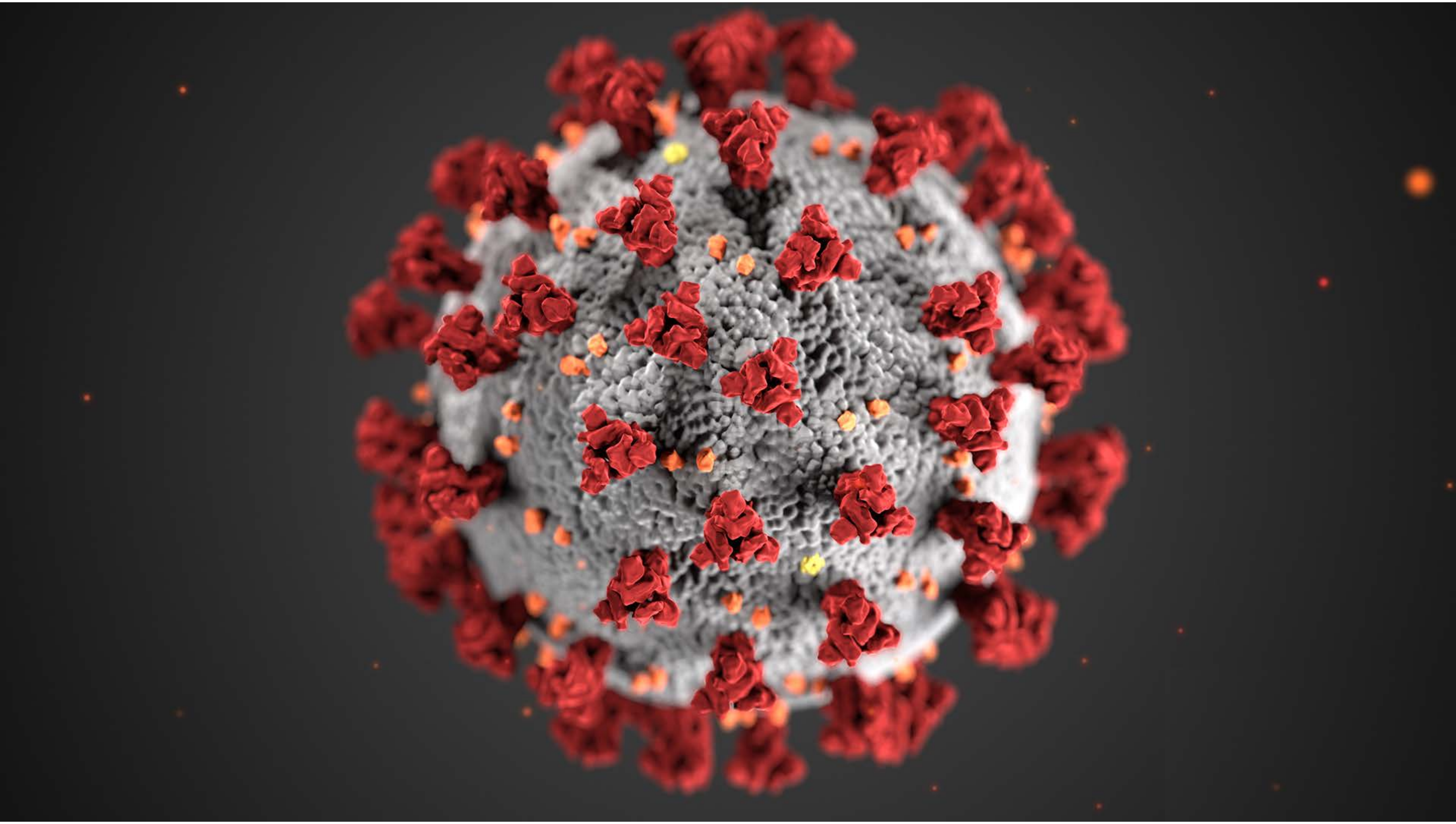
Lineages Of Concern

Select Lineage of Concern (LoC): P.1

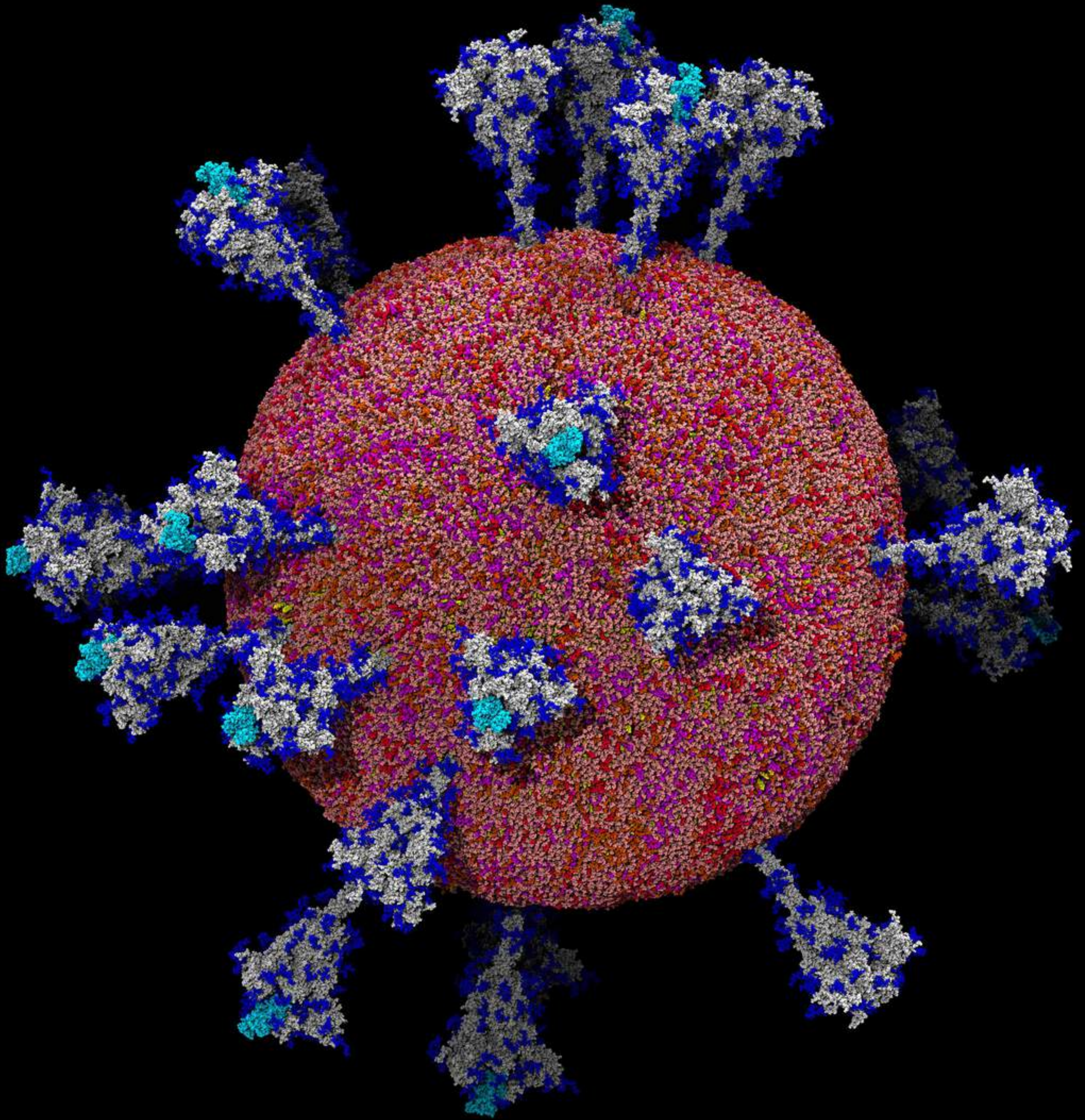
LoC name	P.1
PANGO lineage	B.1.1.28
NextStrain lineage	20J/501Y.V3
Other synonyms	variant originating in Brazil, B.1.1.248
Emergence location	Brazil
Emergence date	July 2020
Amino acid substitutions vs Wuhan-Hu-1: Spike	L18F**, T20N, P26S, D138Y, R190S, K417T**, E484K**, N501Y**, D614G**, H655Y, T1027I, V1176F
Spike Short Peptide Search	K417T: PGQTGTIADYN (AND) H655Y: LIGAEYVNNSY
Amino acid substitutions vs Wuhan-Hu-1: Non-Spike	nsp3: S370L, K977Q; nsp4: S184N; 3C-like proteinase: A260V; nsp6: S106-, G107-, F108-; RNA-dependent RNA polymerase: P323L; helicase: E341D; ORF3a protein: S253P; ORF8 protein: E92K; nucleocapsid phosphoprotein: P80R, R203K, G204R
Nucleotide substitutions vs Wuhan-Hu-1	C241T, T733C, C2749T, C3037T, C3828T, A5648C, A6319G, A6613G, G9105A, C10833T, T11288-, C11289-, T11290-, G11291-, G11292-, T11293-, T11294-, T11295-, T11296-, C12778T, C13860T, C14408T, G17259T, C21614T, C21621A, C21638T, G21974T, G22132T, A22812C, G23012A, A23063T, C23380T, A23403G, C23525T, C24642T, G25088T, T26149C, G28167A, -28263A, -28264A, -28265C, -28266A, C28516G, A28881T, G28882C, G28885A, G28886A, G28887C, T29838A
Impact	Increased transmissibility
SF overlap	Transmembrane region, Coiled-coil region, Sequence variant, Mutagenesis site, Region of interest, Helix, Beta strand, Topological domain, Domain, Disulfide bond
ViPR representative strain link	SARS-CoV-2/human/USA/MN-MDH-2399/2021
Consensus genome sequence	P.1_US_genome_consensus.fasta

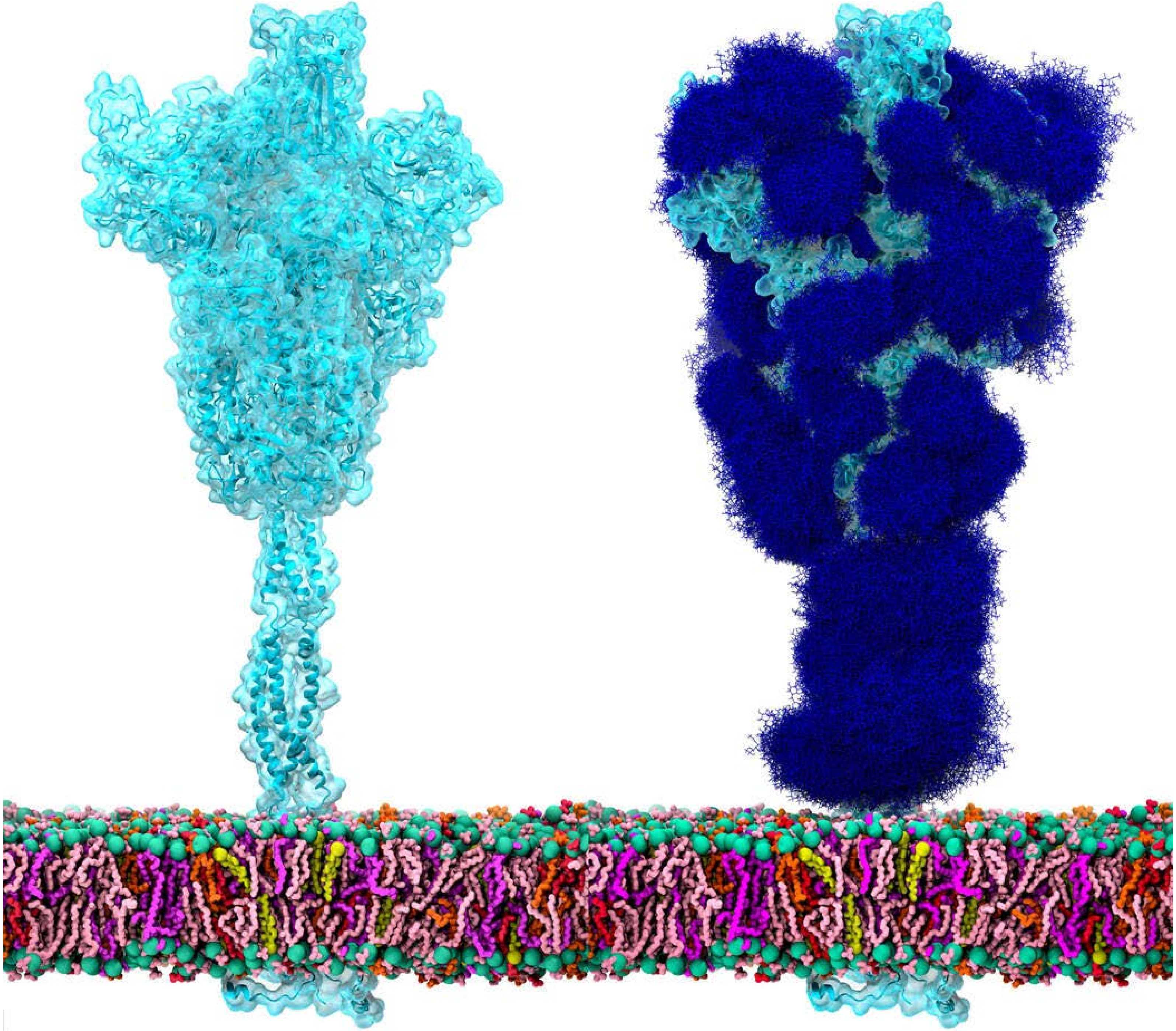

BV-BRC BETA
WORKSPACES
ABOUT
CONTACT
TEAM
ANNOUNCEMENTS
PUBLICATIONS
CITATION
RELATED
SARS-COV-2 VARIANTS AND LINEAGES OF CONCERN
Overview
Lineages of Concern
Covariants
Variants
Genome Browser
Phylogenetic Tree
Resources
Archaeopteryx.js 1.8.5b6
P
A
C
Display Data
Node Name
Sequence
Branch Length
External Labels
External Nodes
Internal Nodes
Node Vis
Branch Vis
Zoom
Y+
X-
F
X+
Y-
Tools
O
R
U
M
External label size:
Branch label size:
Node size:
Branch width:
Search
R
R
Node Fill
PANGO Lineage of Concern
B.1.1.28
B.1.1.7
B.1.351
B.1.375
B.1.427
B.1.429
B.1.525
P.1
P.2
SARS_CoV_2/human/CHL/Santiago_op8d2/2020
SARS_CoV_2/human/USA/AZ_CDC_LC0000245/2020
SARS_CoV_2/human/BLZ/CDC_6840/2020
SARS_CoV_2/Hu/KngFJ/23RD5
SARS_CoV_2/human/USA/FL_BPHL_0303/2020
SARS_CoV_2/human/USA/CA_CZB_13116/2020
SARS_CoV_2/human/USA/MA_MGH_00887/2020
SARS_CoV_2/human/USA/CA_CZB_10218/2020
SARS_CoV_2/human/BHR/340854759/2020
SARS_CoV_2/human/BHR/023347351/2020
SARS_CoV_2/human/IND/GBRC455a/2020
SARS_CoV_2/human/IND/nitma1139/2020
SARS_CoV_2/human/IND/GBRC419b/2020
SARS_CoV_2/human/TUR/Kafkas_SARSCoV2_0010/2020
SARS_CoV_2/human/USA/IN_CDC_2_3693766/2020
SARS_CoV_2/human/USA/MD_MDH_0688/2021
SARS_CoV_2/human/TWN/CGMH_CGU_18/2020
SARS_CoV_2/human/USA/MN_MDH_2017/2020
SARS_CoV_2/human/BGD/DNAS_isl_75/2020
SARS_CoV_2/human/ITA/APU_POLBA09/2020
SARS_CoV_2/human/USA/CA_CZB_10287/2020
SARS_CoV_2/human/USA/CA_CZB_1926/2020
SARS_CoV_2/human/USA/CA_CZB_13888/2020
SARS_CoV_2/human/USA/DC_CDC_LC0000361/2020
SARS_CoV_2/human/USA/CA_CZB_10786/2020
SARS_CoV_2/human/USA/MD_CDC_LC0007377/2021
SARS_CoV_2/human/BGD/GC001/2021
SARS_CoV_2/human/AUS/VIC2020/2020
SARS_CoV_2/mink/DK/AD3_Farm1/2020
SARS_CoV_2/human/GHA/nmimr_SARS_CoV_2_TRA__140/2021
SARS_CoV_2/human/ESP/AST_232023718/2020
SARS_CoV_2/human/BGD/DNAS_DRICM_isl_SM_12/2020
SARS_CoV_2/human/USA/CA_CZB_10832/2020
SARS_CoV_2/human/USA/NM_QDX_210/2020
SARS_CoV_2/human/USA/MO_QDX_2587/2020
SARS_CoV_2/human/France/50010DC/2020
SARS_CoV_2/human/AUS/VIC249/2020
SARS_CoV_2/human/SAU/KFSHRC31/2020
SARS_CoV_2/human/USA/CA_CZB_2161/2020
SARS_CoV_2/human/USA/CA_CZB_11125/2020
SARS_CoV_2/human/USA/FL_BPHL_0749/2020
SARS_CoV_2/human/AUS/VIC252/2020
SARS_CoV_2/human/GRC/141_32983/2020
SARS_CoV_2/human/USA/WA_S1005/2020
SARS_CoV_2/human/USA/FHCRC_14109/2020
SARS_CoV_2/human/USA/WA_S1781/2020

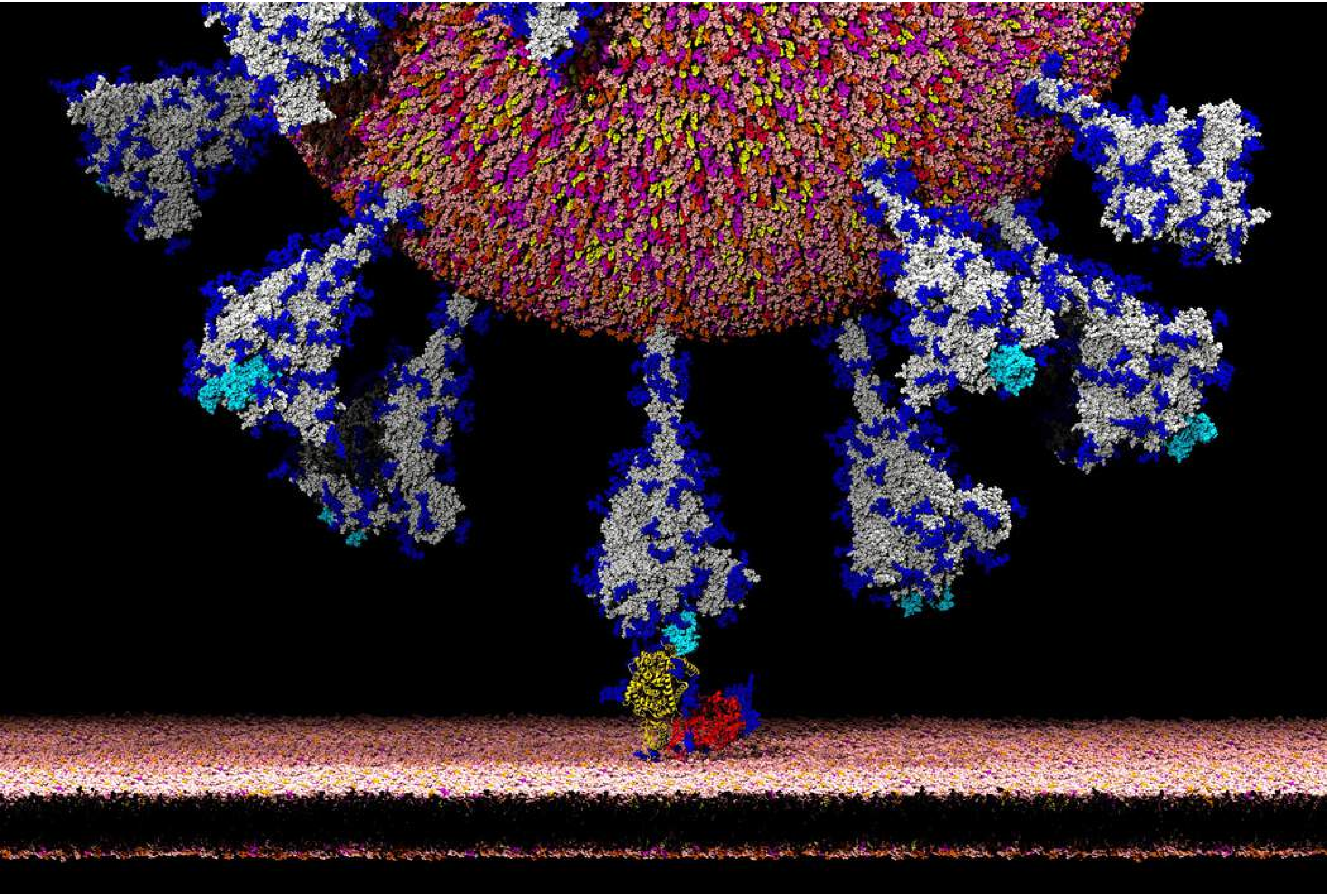
Coronavirus também é arte

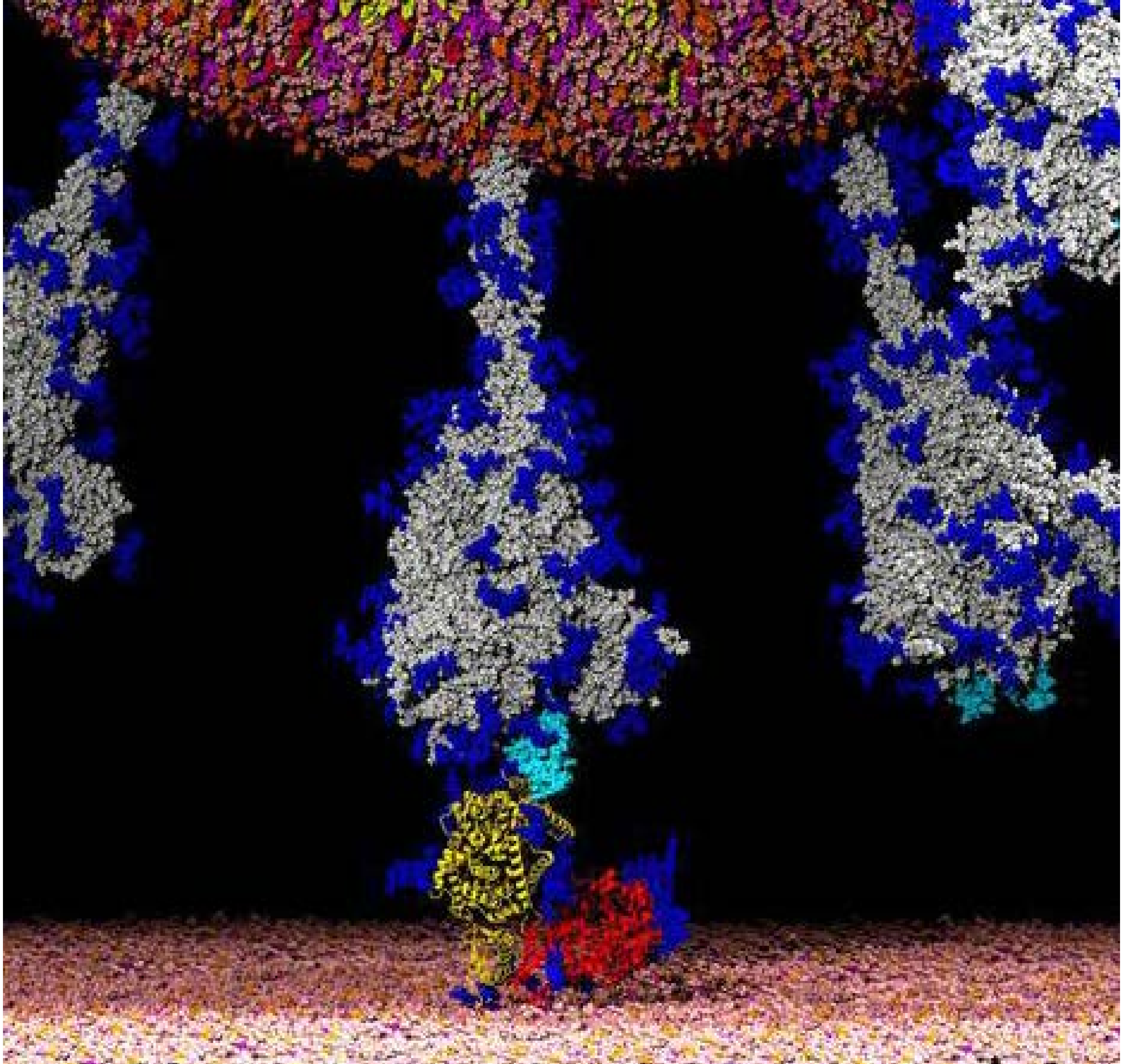


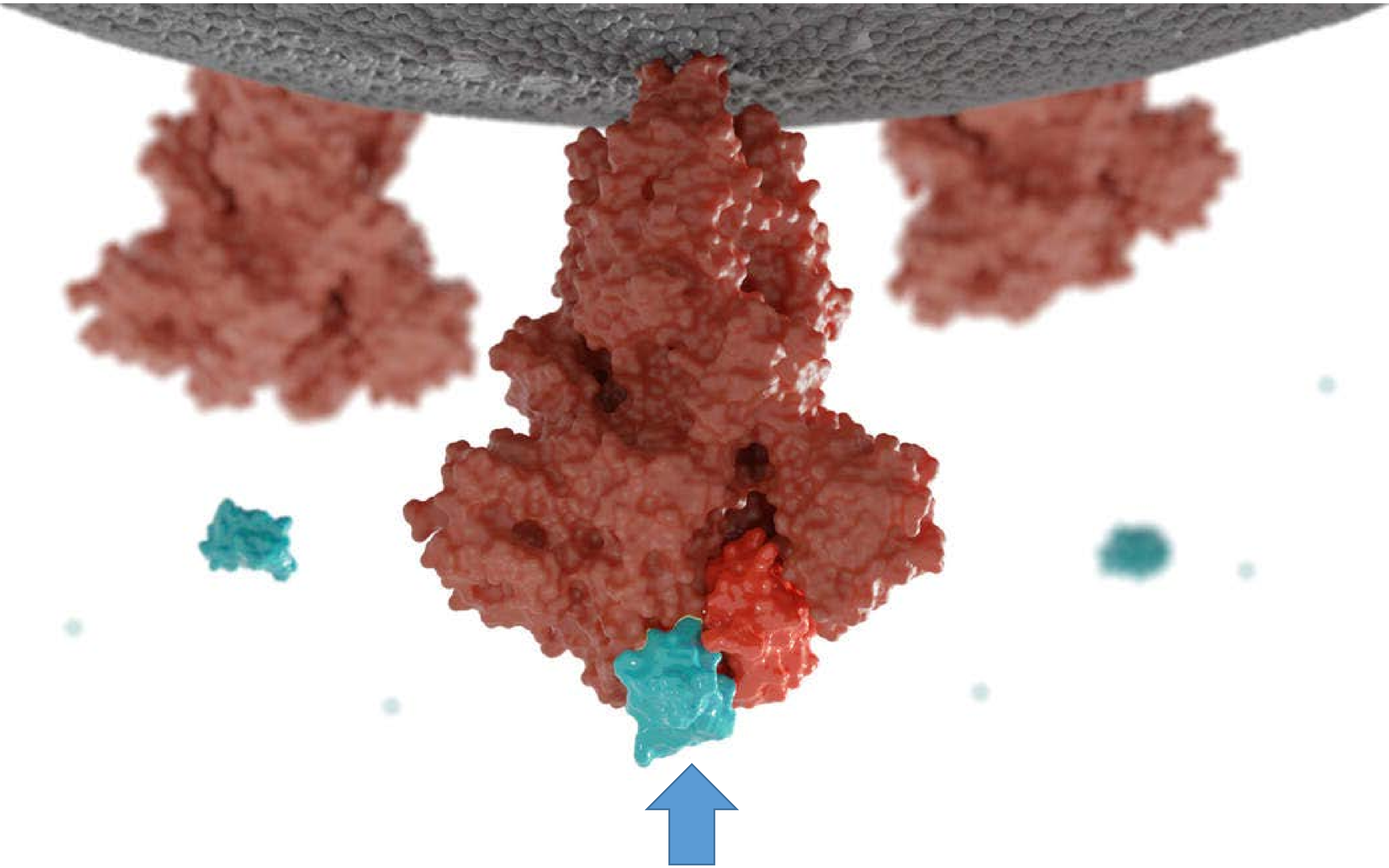
Alissa Eckert, Dan Higgins/CDC











molécula bloqueadora do sítio de ligação na ACE2

Créditos das imagens

- artigo de Carl Zimmer no NY Times (9/10/2020)
- proteômica + dinâmica molecular + computação gráfica
- https://www.nytimes.com/interactive/2020/health/coronavirus-unveiled.html?action=click&module=Top%20Stories&pgtype=Homepage