41e
RÉUNION INTERDISCIPLINAIRE DE CHIMIOTHÉRAPIE ANTI-INFECTIEUSE

LUNDI 13 & MARDI 14 DÉCEMBRE 2021

PALAIS DES CONGRÈS • PARIS

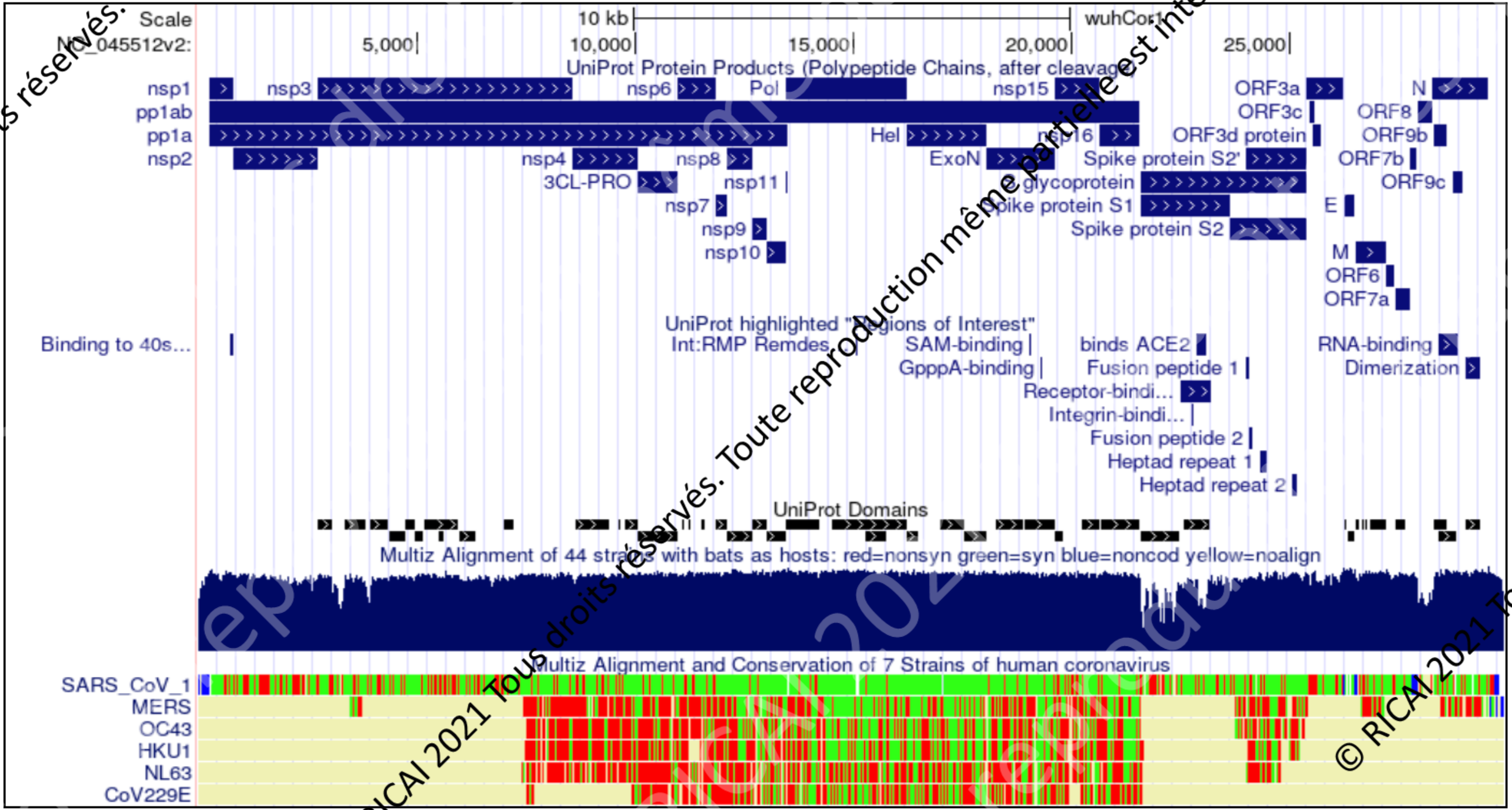
Data sharing of SARS-CoV-2 sequences

Max Haeussler, PhD
Angie Hinrichs, Russ Corbett-Detig,
Linda Rosewood, Daniel Wilson
Jim Kent, David Haussler

University of California at Santa Cruz
Genomics Institute - UCSC Genome Browser Group
maxh@ucsc.edu

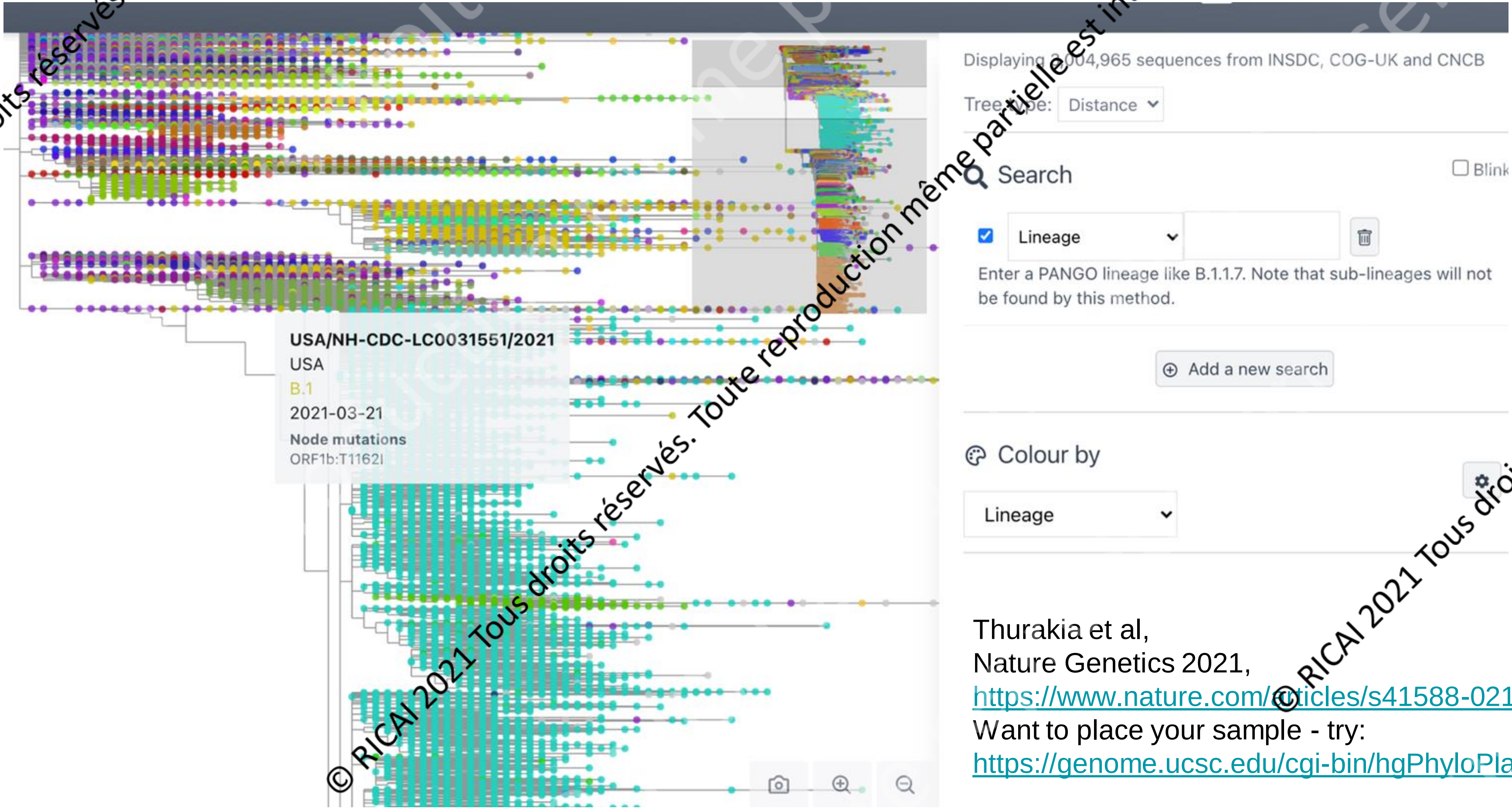
UNIVERSITY OF CALIFORNIA SANTA CRUZ Genomics Institute

UCSC SARS-CoV-2 Genome Browser: genome.ucsc.edu



Fernandes et al, Nature Genetics 2020, https://www.nature.com/articles/s41588-020-0700-8

Taxonium.org, Cov2Tree.org and Pangolin --usher



USA/NH-CDC-LC0031551/2021
USA
B.1
2021-03-21
Node mutations
ORF1b:T1162I

Displaying 2,004,965 sequences from INSDC, COG-UK and CNCB

Tree type: Distance

Search ☐ Blink

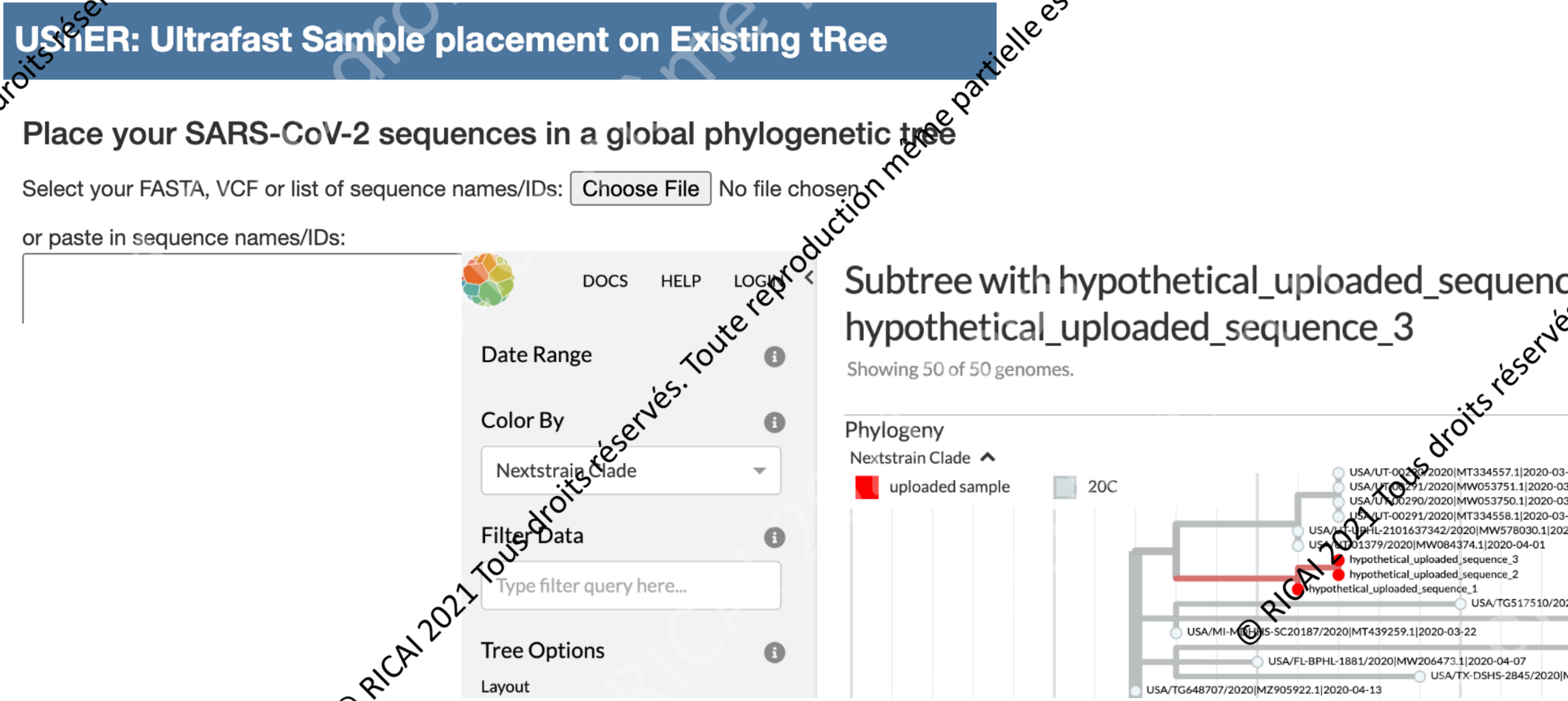
☑ Lineage

Enter a PANGO lineage like B.1.1.7. Note that sub-lineages will not be found by this method.

Colour by: Lineage

Thurakia et al,
Nature Genetics 2021,
https://www.nature.com/articles/s41588-021-00862-7
Want to place your sample - try:
https://genome.ucsc.edu/cgi-bin/hgPhyloPlace

Usher: upload sequence - show neighbors on Nextstrain



UShER: Ultrafast Sample placement on Existing tRee

Place your SARS-CoV-2 sequences in a global phylogenetic tree

Select your FASTA, VCF or list of sequence names/IDs: Choose File No file chosen

or paste in sequence names/IDs:

Subtree with hypothetical_uploaded_sequenc
hypothetical_uploaded_sequence_3

Showing 50 of 50 genomes.

Date Range

Color By: Nextstrain Clade

Filter Data: Type filter query here...

Tree Options

Layout

Phylogeny

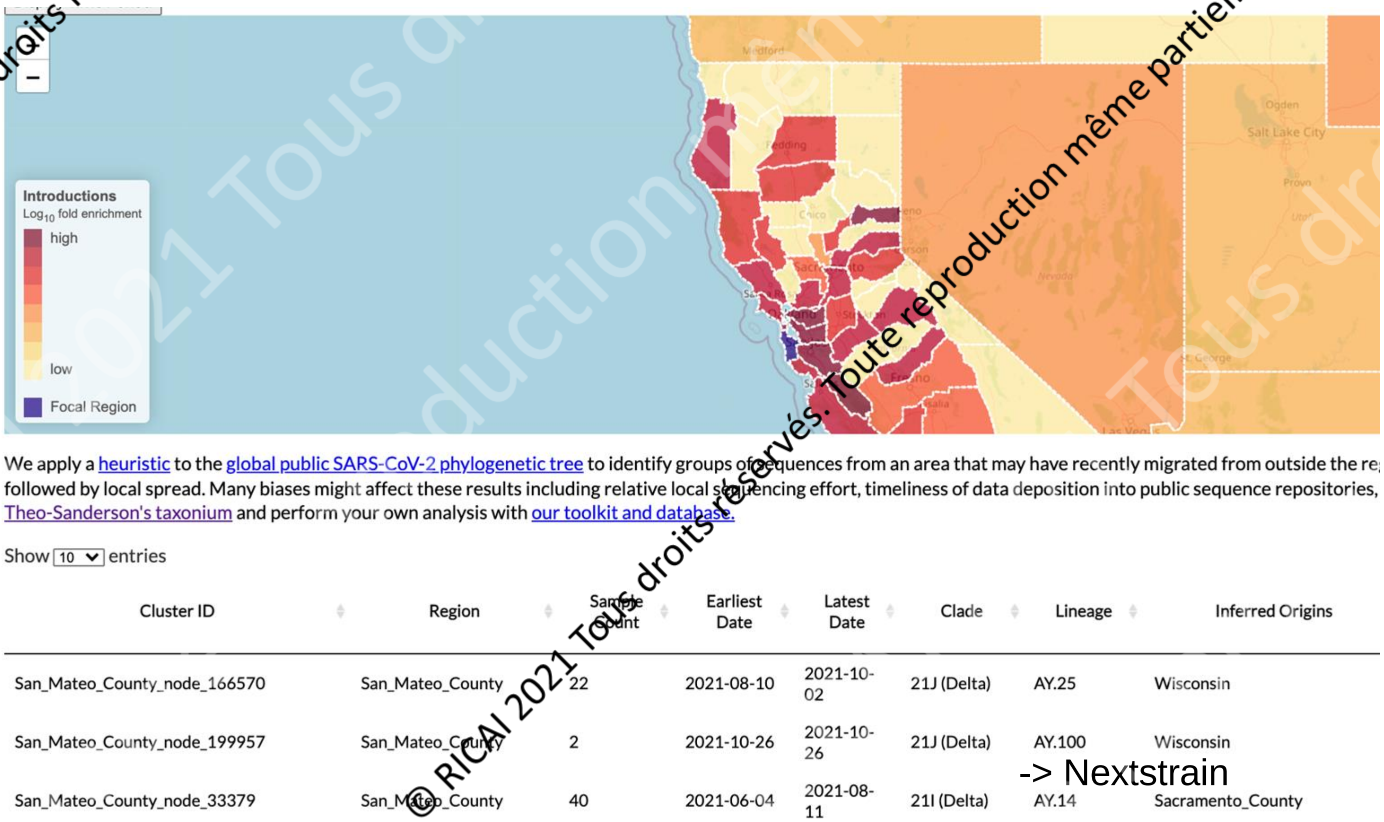
Nextstrain Clade

🟥 uploaded sample

▢ 20C

https://genome.ucsc.edu/cgi-bin/hgPhyloPlace

California Big Tree Cluster Tracker



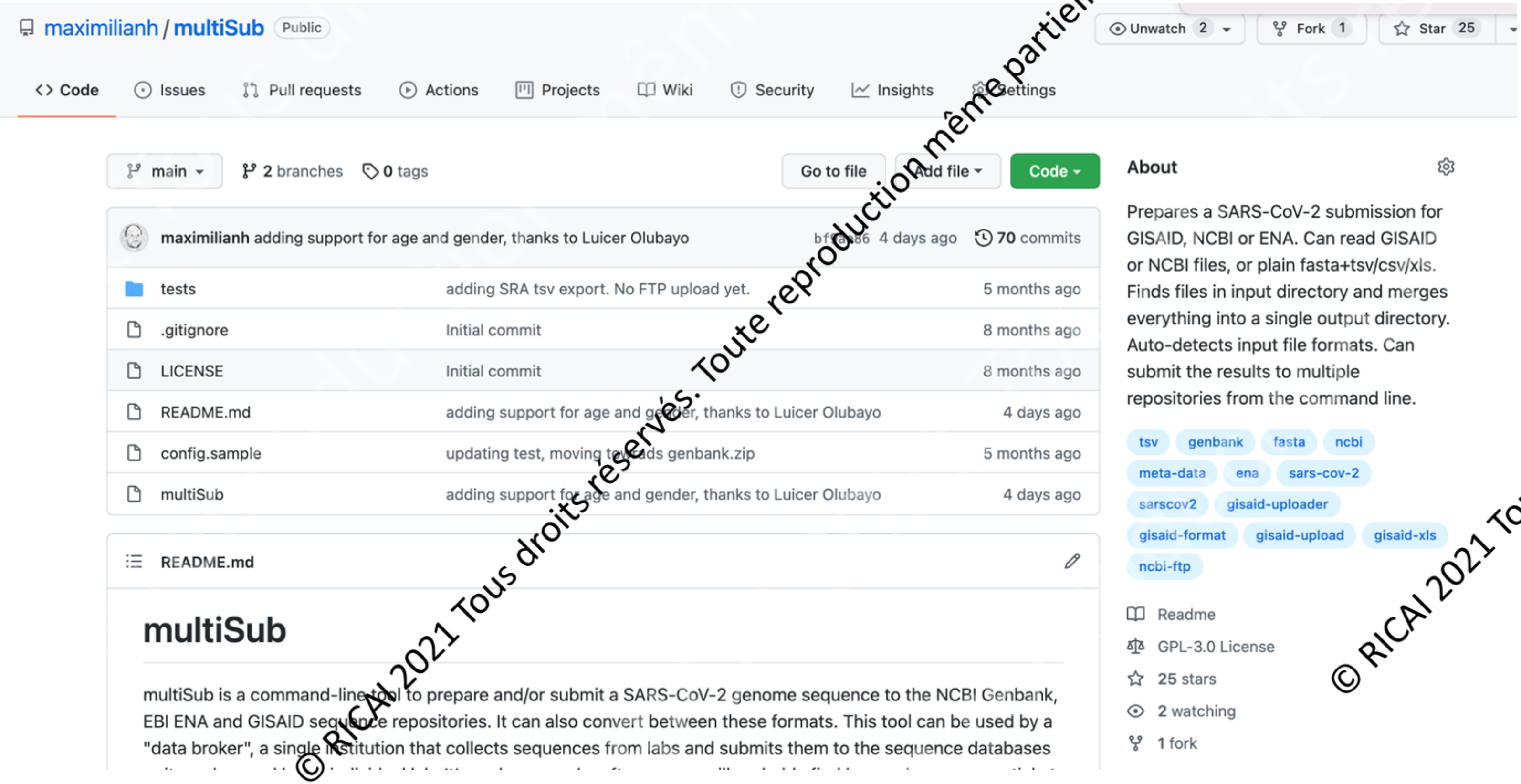
We apply a heuristic to the global public SARS-CoV-2 phylogenetic tree to identify groups of sequences from an area that may have recently migrated from outside the re followed by local spread. Many biases might affect these results including relative local sequencing effort, timeliness of data deposition into public sequence repositories, Theo-Sanderson's taxonium and perform your own analysis with our toolkit and database.

Show 10 entries

Cluster ID	Region	Sample Count	Earliest Date	Latest Date	Clade	Lineage	Inferred Origins
San_Mateo_County_node_166570	San_Mateo_County	22	2021-08-10	2021-10-02	21J (Delta)	AY.25	Wisconsin
San_Mateo_County_node_199957	San_Mateo_County	2	2021-10-26	2021-10-26	21J (Delta)	AY.100	Wisconsin
San_Mateo_County_node_33379	San_Mateo_County	40	2021-06-04	2021-08-11	21I (Delta)	AY.14	Sacramento_County

-> Nextstrain

MultiSub: Submit all your sequences to NCBI or ENA & GISAID with a single command.`



maximilianh adding support for age and gender, thanks to Luicer Olubayo bf9ae86 4 days ago 70 commits

tests	adding SRA tsv export. No FTP upload yet.	5 months ago
.gitignore	Initial commit	8 months ago
LICENSE	Initial commit	8 months ago
README.md	adding support for age and gender, thanks to Luicer Olubayo	4 days ago
config.sample	updating test, moving towards genbank.zip	5 months ago
multiSub	adding support for age and gender, thanks to Luicer Olubayo	4 days ago

About

Prepares a SARS-CoV-2 submission for GISAID, NCBI or ENA. Can read GISAID or NCBI files, or plain fasta+tsv/csv/xls. Finds files in input directory and merges everything into a single output directory. Auto-detects input file formats. Can submit the results to multiple repositories from the command line.

README.md

multiSub

multiSub is a command-line tool to prepare and/or submit a SARS-CoV-2 genome sequence to the NCBI Genbank, EBI ENA and GISAID sequence repositories. It can also convert between these formats. This tool can be used by a "data broker", a single institution that collects sequences from labs and submits them to the sequence databases

Why international sequencing databases matter

- Sequences not useful without an archive
- Get an idea of global distribution of variants
- Decisions on travel restrictions
- Molecular contact tracing
- Historical dataset for future analyses

- Must be: open, fast, reliable

Three nucleotide databases

INSDC

- National Center for Biotechnology Information NCBI
  - US government/NIH-funded, outside of Washington
  - Many databases, since 1980s
- ENA: European Nucleotide Archive
  - Funded by EMBL/EU, outside of Cambridge UK
  - Many databases, since 1980s
  - Exchanges data with NCBI and Japan every x days
- GISAID: Global Initiative on Sharing All Influenza Data
  - Independent non-profit organisation
  - not clear who funds it, not hosted by Germany (commercial hoster in Bavaria)
  - Relies on volunteers in many countries

GISAID ?

(Japan, soon China, run their own INSDC databases)

Worldwide: Dec 10, 2021

- 5.6 M sequences: biggest sequencing project in history
- 1/3 of sequences from the United States
- 55% from US + UK
- Others:
    - Germany, Denmark, Canada 5%
    - Japan, France, Sweden 2-3%
    - All others smaller, “long tail”
- Of all samples, OECD Countries sequence 3% (US) to 13% (UK)
    - up to 45% in Denmark, (an outlier)

Worldwide: Dec 10, 2021

- 6 M sequences in GISAID
- 2.7M sequences in NCBI/ENA
- 1/3 of sequences from the United States
- 55% from **US** + UK
- Others:
  - **Germany**, Denmark, Canada 5%
  - Japan, France, Sweden 2-3%
  - All others smaller, “long tail”
- Of all samples, OECD Countries sequence 3% (US) to 13% (UK)
  - up to 45% in Denmark, (an outlier)

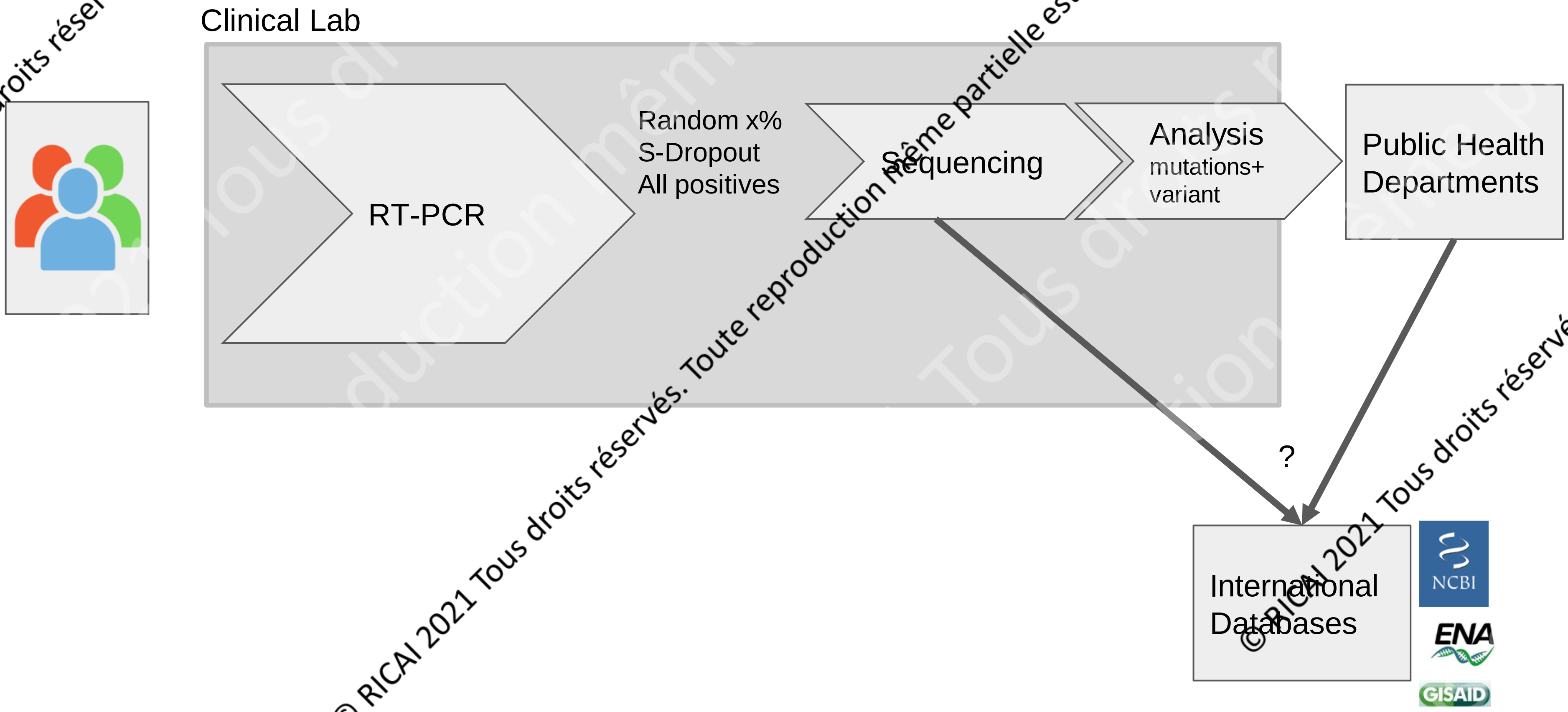
Rest of the talk

1. USA, CDC: “SPHERES”

    Copied from SPHERES, Duncan MacCannell, CDC, Dec 1st

1. Germany, Robert Koch Institut: “DESH”
2. A discussion of data licenses for SARS-CoV-2

How do sequences land in databases?



Clinical Lab

RT-PCR

Random x%
S-Dropout
All positives

Sequencing

Analysis
mutations+
variant

Public Health Departments

?

International Databases

Example 1: USA

The **SPHERES** consortium brings together over **1,250** scientists from more than **200** organizations in a network of federal, state, private, academic, and non-profit organizations in the **largest public health sequencing effort** in US history.

https://www.cdc.gov/coronavirus/2019-ncov/variants/spheres.html
https://github.com/CDCgov/SARS-CoV-2_Sequencing

SPHERES

- Participatory mode.
- Engages academia, public health and private sector. CONNECTS NEW PARTNERS with public health.
- Share capacity & expertise, improve the utility, quality and timeliness of data.

CDC NS3 & Baseline Contracts

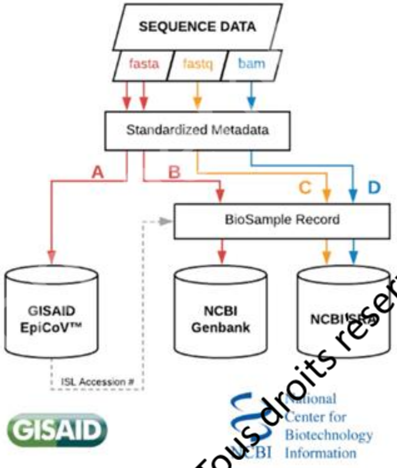
- Based on established influenza public health surveillance models.
- Surveillance enhanced and expanded using commercial labs to extend the depth and reach of coverage.
- PRIMARY GOAL: establish consistent national baseline monitoring for variant surveillance and risk assessment; contribute to global picture of circulating viruses.
- Supports interagency risk assessment (SIG), academic research, federal decisionmaking.

State and Local Public Health

- Build flexible, scalable, resilient and responsive capacity for genomic surveillance that can be adapted to many scenarios and is responsive to local public health priorities.
- Leverage state and regional capacity to strengthen sequencing/bioinformatics/epi capabilities.
- PRIMARY GOAL: local surveillance and monitoring for unusual variants, patterns or clinical outcomes, but often focused on outbreak response, infection control, contact tracing.

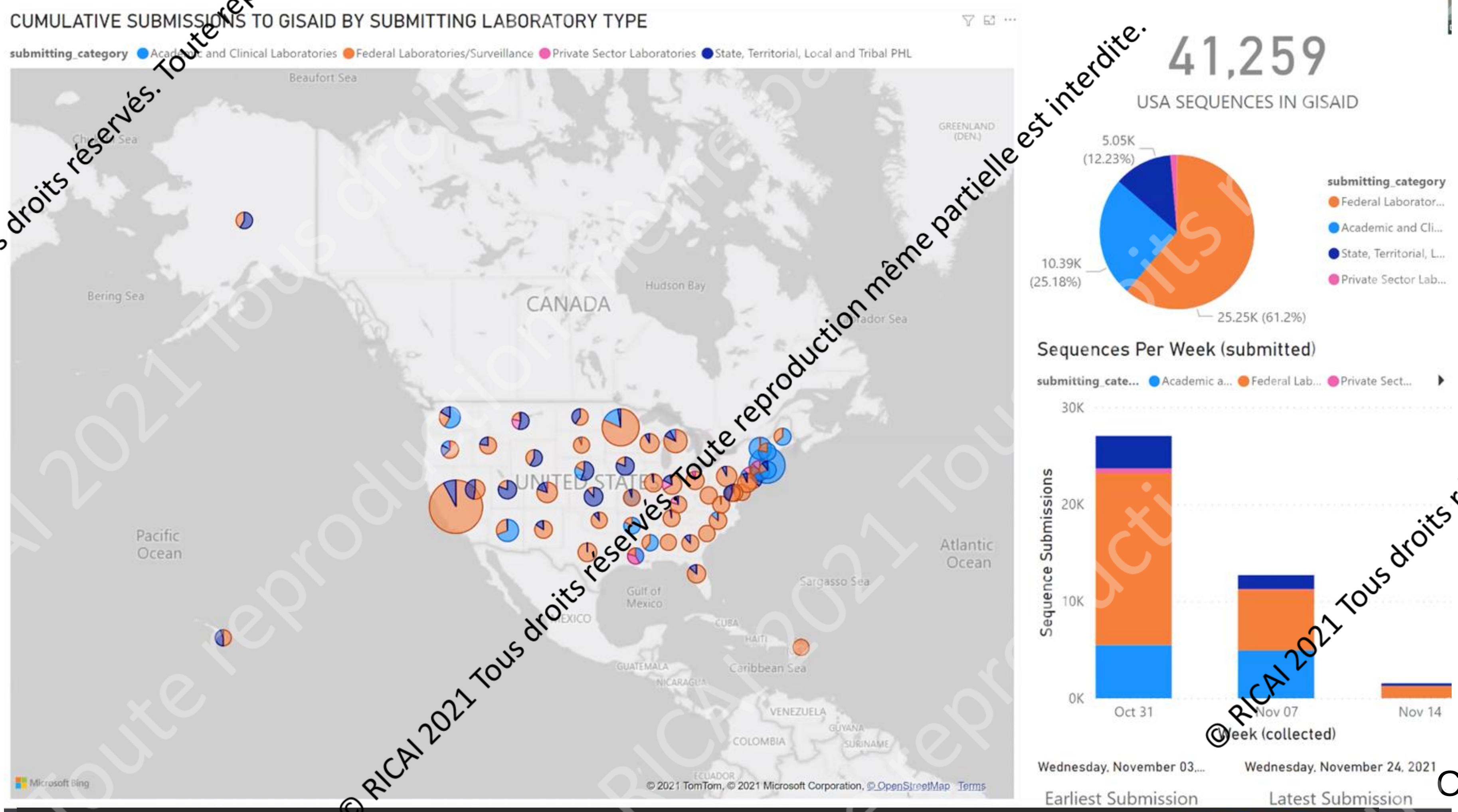
- State Consortia: AZ, CA, MN, Others.

- Long term, these are the critical systems to build.



CALIFORNIA

CDC



CUMULATIVE SUBMISSIONS TO GISAID BY SUBMITTING LABORATORY TYPE

submitting_category ● Academic and Clinical Laboratories ● Federal Laboratories/Surveillance ● Private Sector Laboratories ● State, Territorial, Local and Tribal PHL

41,259

USA SEQUENCES IN GISAID

Sequences Per Week (submitted)

Wednesday, November 03,...

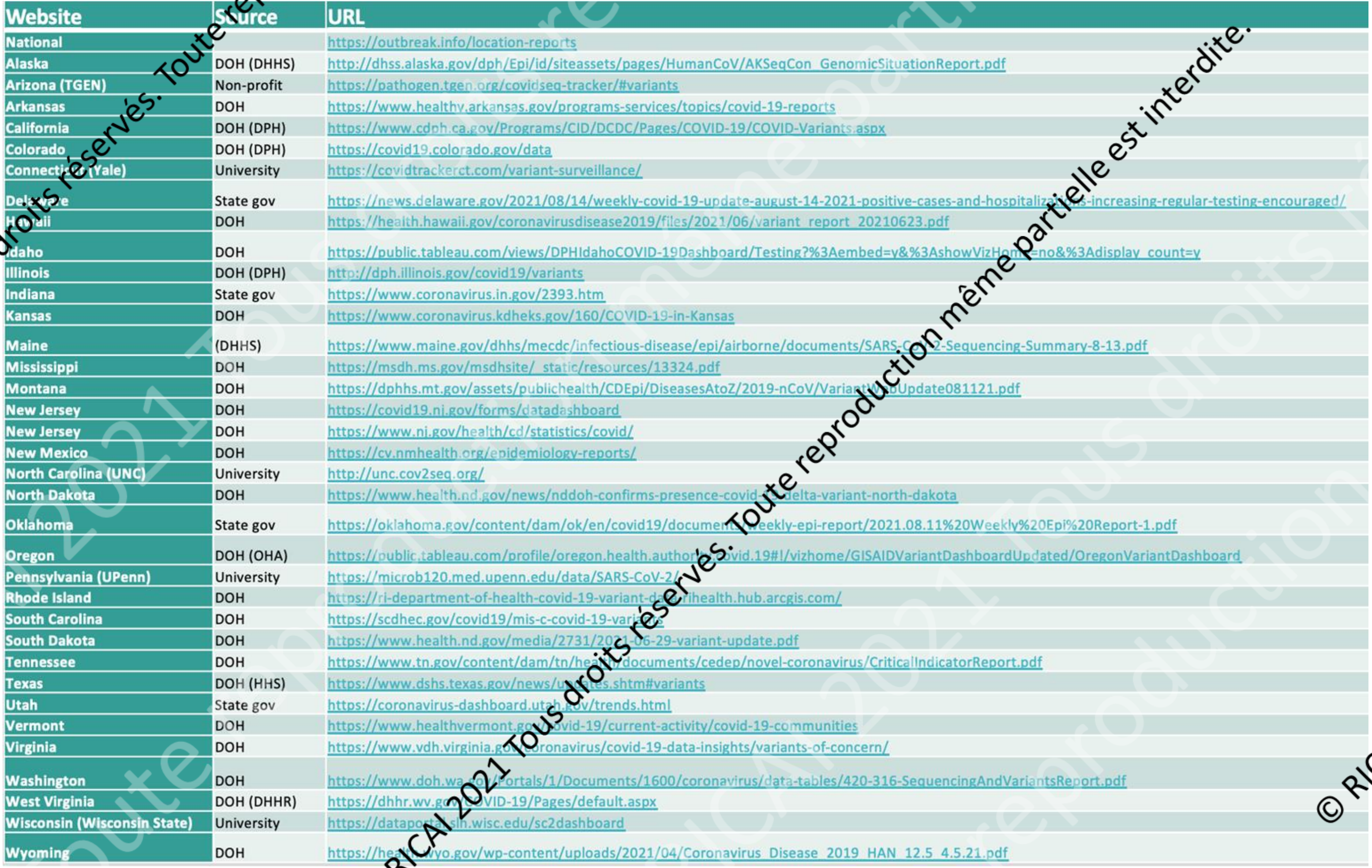
Earliest Submission

Wednesday, November 24, 2021

Latest Submission

CDC

Nextmeta 2021-11-05 / GISAID - manual classification of submitting laboratories - COLLECTION AFTER NOVEMBER 1

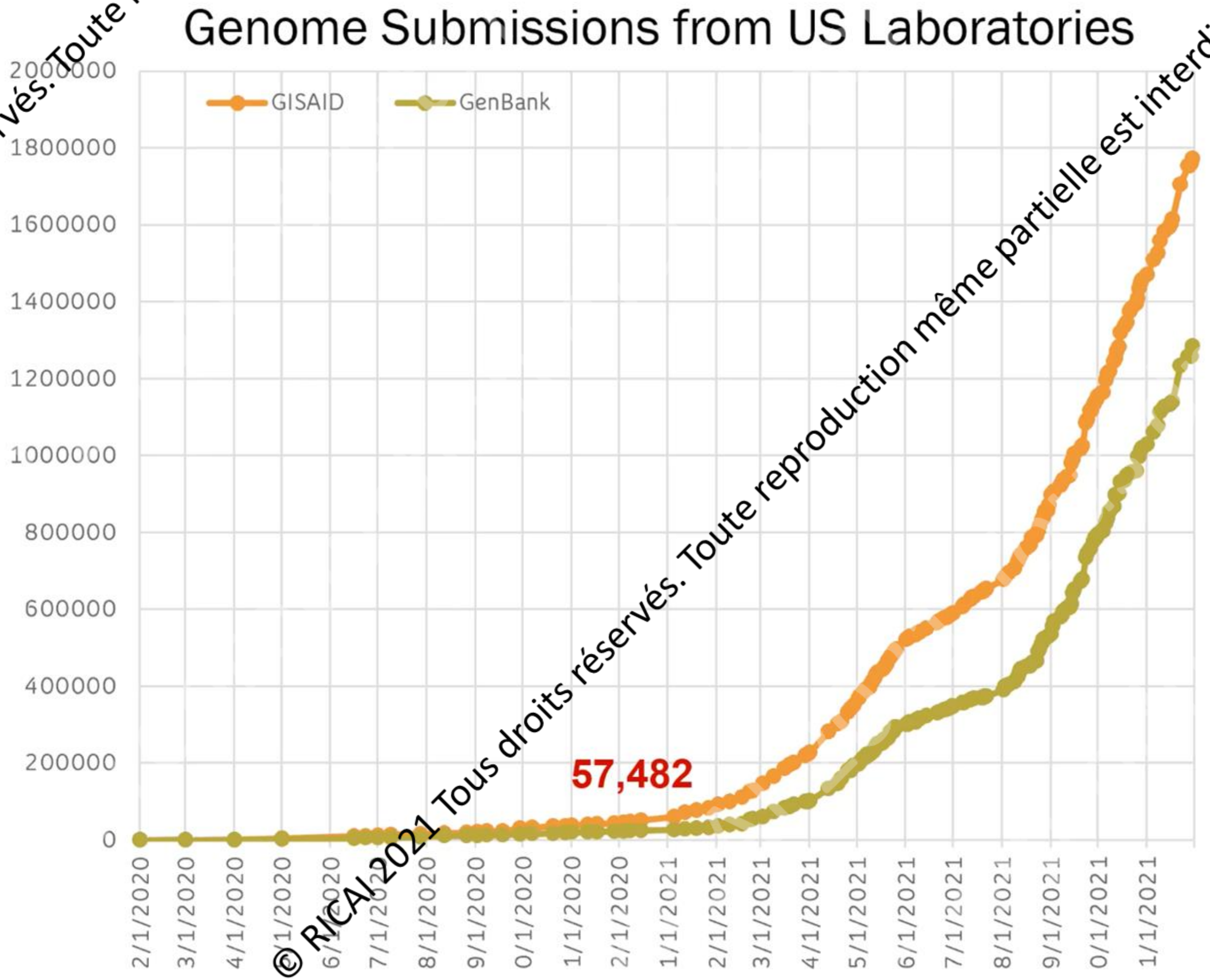


Website	Source	URL
National		https://outbreak.info/location-reports
Alaska	DOH (DHHS)	http://dhss.alaska.gov/dph/Epi/id/siteassets/pages/HumanCoV/AKSeqCon_GenomicSituationReport.pdf
Arizona (TGEN)	Non-profit	https://pathogen.tgen.org/covidseq-tracker/#variants
Arkansas	DOH	https://www.healthy.arkansas.gov/programs-services/topics/covid-19-reports
California	DOH (DPH)	https://www.cdph.ca.gov/Programs/CID/DCDC/Pages/COVID-19/COVID-Variants.aspx
Colorado	DOH (DPH)	https://covid19.colorado.gov/data
Connecticut (Yale)	University	https://covidtrackerct.com/variant-surveillance/
Delaware	State gov	https://news.delaware.gov/2021/08/14/weekly-covid-19-update-august-14-2021-positive-cases-and-hospitalizations-increasing-regular-testing-encouraged/
Hawaii	DOH	https://health.hawaii.gov/coronavirusdisease2019/files/2021/06/variant_report_20210623.pdf
Idaho	DOH	https://public.tableau.com/views/DPHIdahoCOVID-19Dashboard/Testing?%3Aembed=y&%3AshowVizHome=no&%3Adisplay_count=y
Illinois	DOH (DPH)	http://dph.illinois.gov/covid19/variants
Indiana	State gov	https://www.coronavirus.in.gov/2393.htm
Kansas	DOH	https://www.coronavirus.kdheks.gov/160/COVID-19-in-Kansas
Maine	(DHHS)	https://www.maine.gov/dhhs/mecdc/infectious-disease/epi/airborne/documents/SARS-CoV-2-Sequencing-Summary-8-13.pdf
Mississippi	DOH	https://msdh.ms.gov/msdhsite/_static/resources/13324.pdf
Montana	DOH	https://dphhs.mt.gov/assets/publichealth/CDEpi/DiseasesAtoZ/2019-nCoV/VariantWebUpdate081121.pdf
New Jersey	DOH	https://covid19.nj.gov/forms/datadashboard
New Jersey	DOH	https://www.nj.gov/health/cd/statistics/covid/
New Mexico	DOH	https://cv.nmhealth.org/epidemiology-reports/
North Carolina (UNC)	University	http://unc.cov2seq.org/
North Dakota	DOH	https://www.health.nd.gov/news/nddoh-confirms-presence-covid-19-delta-variant-north-dakota
Oklahoma	State gov	https://oklahoma.gov/content/dam/ok/en/covid19/documents/weekly-epi-report/2021.08.11%20Weekly%20Epi%20Report-1.pdf
Oregon	DOH (OHA)	https://public.tableau.com/profile/oregon.health.authority.covid.19#!/vizhome/GISAIDVariantDashboardUpdated/OregonVariantDashboard
Pennsylvania (UPenn)	University	https://microb120.med.upenn.edu/data/SARS-CoV-2/
Rhode Island	DOH	https://ri-department-of-health-covid-19-variant-data-rihealth.hub.arcgis.com/
South Carolina	DOH	https://scdhec.gov/covid19/mis-c-covid-19-variants
South Dakota	DOH	https://www.health.nd.gov/media/2731/2021-06-29-variant-update.pdf
Tennessee	DOH	https://www.tn.gov/content/dam/tn/health/documents/cedep/novel-coronavirus/CriticalIndicatorReport.pdf
Texas	DOH (HHS)	https://www.dshs.texas.gov/news/updates.shtm#variants
Utah	State gov	https://coronavirus-dashboard.utah.gov/trends.html
Vermont	DOH	https://www.healthvermont.gov/covid-19/current-activity/covid-19-communities
Virginia	DOH	https://www.vdh.virginia.gov/coronavirus/covid-19-data-insights/variants-of-concern/
Washington	DOH	https://www.doh.wa.gov/Portals/1/Documents/1600/coronavirus/data-tables/420-316-SequencingAndVariantsReport.pdf
West Virginia	DOH (DHHR)	https://dhhr.wv.gov/COVID-19/Pages/default.aspx
Wisconsin (Wisconsin State)	University	https://dataportal.slh.wisc.edu/sc2dashboard
Wyoming	DOH	https://health.wyo.gov/wp-content/uploads/2021/04/Coronavirus_Disease_2019_HAN_12.5_4.5.21.pdf

https://cspheres.slack.com/archives/C010Q9WCZJT/p1633648887116100

Data flow

- CDC uploads to GISAID and NCBI (50%)
- Academic (25%) + Public Health (25%): uploads to public database(s)
- Some raw sequencing data, often only genomes
- Some states have their own system (CA, AZ)
- No federal database for patient, personal identifiable information
    - except N3C = research cohorts and HealthVerity in the future



1,772,741
(+214k)

1,283,704
(+170k)

72% in NCBI
Average turnaround
16-17 days,
Major share: 10 days
Local labs faster

CDC

Metadata standards

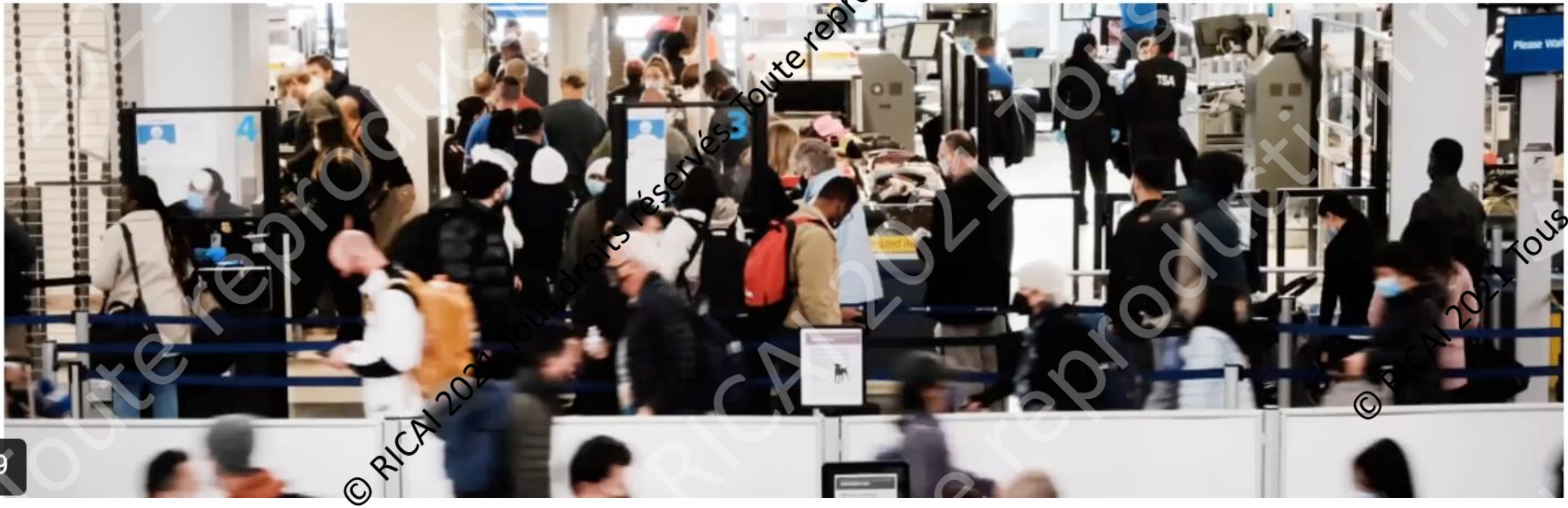
https://github.com/pha4ge/SARS-CoV-2-Contextual-Data-Specification

NS3 meta data requirements:

https://www.aphl.org/programs/preparedness/Crisis-Management/COVID-19-Response/Pages/Sequence-Based-Surveillance-Submission.aspx

CDC expanding surveillance at 4 major US airports to look for Omicron

CNN Staff • Updated 30th November 2021



A field guide to CDC NS3 and Baseline Data

National SARS-CoV-2 Strain Surveillance (NS3):

- ST-CDC-2-######### → NS3 DATA and EVS (targets variants, small proportion)
  - CDC/State NS3 samples – random submission from state and local public health laboratories, submitted to CDC, run through a standard sequencing and bioinformatic workflow.
- ST-CDC-[A-z0-9]{4}-[0-9]{4} → Usually NS3 Enhanced or Special Studies

Baseline Surveillance:

- ST-CDC-LC######### → LabCorp – Random selection
- ST-CDC-######### & ST-CDC-QDX######## → Quest – Random selection
- ST-CDC-STM-######### → Helix/Illumina – SGTF prior to mid-Feb – Retail Pharmacy
- ST-CDC-ASC######### → Aegis – Random selection – Retail Pharmacy
- ST-CDC-FG######### → Fulgent – Random selection – Community Testing
- ST-CDC-IBX######### → Infinity Biologix – Random selection – Community Testing
- ST-CDCBI-######### → Broad Institute – Random selection – Broad CLIA Laboratory

>> SAME IDENTIFIERS ACROSS GISAID AND SRA.

Summary USA

- Federal/CDC (50%), local/state (25%), academic (25%)
- Big regional differences
- Total cost: around 250 mil $ (though now < 100$/sample)
- **Players tied together by the databases (GISAID/NCBI)**

For questions on SPHERES, please contact:
Duncan MacCannell, fms2@cdc.gov

Bi-weekly Zoom/Slack meetings are open to anyone, contact Duncan or me maxh@ucsc.edu for the links.

German Robert Koch Institut (RKI): DESH

- Any lab that sequences SARS-CoV-2 is **required** to submit sequence to the RKI (this regulation will laps on 30.06.2022)
- Rate: 200 €/sample, mostly random 5%
- > 100 labs across Germany have DESH accounts
- DESH accepts genome sequence (consensus) and basic metadata:
  - Sequencer, date, lab, unique sample ID, reason for sequencing
  - Personal data (name, etc) is transmitted via another, older system, linked via sample ID
- GISAID/NCBI/ENA not allowed as a replacement, labs must use DESH
- DESH then submits data to GISAID and ENA
- In practice: 283k in GISAID, 180k in ENA, 160k in NCBI
- Only 63% in ENA, 56% in NCBI ?

https://www.rki.de/DE/Content/InfAZ/N/Neuartiges_Coronavirus/DESH/DESH.html

Summary: Germany

- One single central, national database, like COG-UK
- No raw sequencing data: small, quick
- Exports to both public databases
- Only 56% in NCBI as of Dec 11

For questions, contact:
Torsten Semmler, SemmlerT@rki.de, Stephan Fuchs, FuchsS@rki.de

The SARS-CoV-2 databases:
GISAID versus NCBI/ENA

What is the meaning of “open data” ?

Open Data crucial for biomedical research

- 1971 Protein Data Bank - computer tapes, shipped for free
- 1981 Genbank - free and unrestricted access. no use restrictions or licensing requirements can be shown (but can exist - see patents)
- 2000 Human Genome - Free, unrestricted access
- Large ecosystem of services, products and software
- “Human genome-enabled industry” 100-250 billion revenue in 2019, millions of jobs, taxes
- See: Internet, Linux, Wikipedia, OpenStreetMaps, etc
- Open data requirement of most governments. Taxpayer funded data.
- Most of taxpayer funded genomics data in open databases

The Economic Impact and Functional Applications of Human Genetics and Genomics
https://www.ashg.org/wp-content/uploads/2021/05/ASHG-TEConomy-Impact-Report-Final.pdf

Licenses: National Center of Biotechnology Information and European Nucleotide Archive

- “Free and unrestricted access”
- “no use restrictions or licensing requirements will be included in any sequence data records, and no restrictions or licensing fees will be placed on the redistribution or use of the database by any party”
- “Credit is given by citing the original submission, following the practices of scientists utilizing published scientific literature.”

https://www.insdc.org/policy.html

GISAID Data Access agreement: Closed access

Flu sequence labs were exploited: companies patent sequences, make vaccines without crediting the labs that publish sequences

Poor countries sequence -> Western countries publish in Nature

2009: Labs would not publish flu sequences before publication

GISAID’s data access agreement tried to address this problem. It says:

- *You may use Data to author, co-author or publish results obtained from your analyses of relevant Data, provided that any such published results acknowledge, as the original source of the Data, the laboratory where the clinical specimen(s) and/or virus isolate(s) were first obtained ("Originating Laboratory")*

A database tries to enforce scientific rules through a data access license
Poor countries feel somewhat protected

Elbe et al, 2017 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6607375/

GISAID License rules, #1: No redistribution

- Access only to registered users
- No redistribution of sequences or meta data or derived data to third parties

Means:

- Everyone in this field must have a GISAID account
- No public backups - Website down = all SARS-CoV-2 data gone
- GISAID updated? - older analyses not reproducible anymore
- No public third party data analysis tools or visualisation tools without GISAID's OK
- You're not allowed to share anything derived from GISAID (mutations!)
- Must give credit to 20,000 labs if you work on 3 million SARS-CoV-2 sequences

"Coronavirus Trove sparks Frustration", https://www.science.org/doi/epdf/10.1126/science.371.6534.1086

GISAID License rules, #2: No linking of GISAID data

No linking of the data to other databases

Means:

- Splits the SARS-CoV-2 data ecosystem into GISAID and non-GISAID data.
- Difficult to create phylogenetic trees or conduct global analyses across this split.
- Some % of the data is only in GISAID, so all new data also go into it
- Infective license: once data is in GISAID, other data must also be in GISAID

https://bit.ly/sc2licenses

GISAID License rules, #3: No reverse engineering

No reverse engineering of the data platform. No automated queries.

Means:

The GISAID software is explicitly not open source and researchers are not legally allowed access to the code to find bugs in it or automate tedious manual steps.

You are not allowed to automate tedious work with programs.

See: https://bit.ly/sc2licenses

Effects of this license

- No public backup of GISAID data
- No data browsers, data tools for SARS-CoV-2 data without GISAID agreement
- Variant assignment software cannot include the GISAID training sequences, it must encrypt them
- And all of this is valid for eternity
- Detailed description: https://bit.ly/sc2licenses

- Violates all open data rules
- CDC recommends double submission, COG-UK, Germany, France, submit to both
- WHO recommends the researchers submit to Genbank/ENA/DDBJ, and to GISAID “for situations where data providers seek retention of ownership of their data” (WHO code of conduct 2019)

- Result: Open letter signed by 800 scientists to submit all SARS-CoV-2 to open databases
    - https://www.covid19dataportal.org/support-data-sharing-covid19/sign (see Nature Feb 03 2021)

Summary:

- Today, 50% of global SARS-CoV-2 data not in an open database
- GISAID is technically a very good database
- But its data access agreement is harmful
- Slows down data exchange, analysis, error correction, reproducibility
- And unlike book/music copyright, this license will never expire

See: https://bit.ly/sc2licenses

Credits

Genome Browser: Jason Fernandes, Max Haeussler, Angie Hinrichs, Hiram Clawson and Genome Browser Group
Usher, Taxodium: Russ Corbett-Detig, Yatish Thurakhya, Angie Hinrichs
Big Tree Explorer: Russ Corbett-Detig, Linda Rosewood, Daniel Wilson and the UCSC CDPH team,
Motivation for all these projects: David Haussler

Funding:
Genome Browser: NHGRI, NIAID, UC/CITRIS, P&R Rebele, Schmidt Futures Foundation
Usher, CovTree.org, multiSub: CDC
Taxodium, Big Tree Explorer: California Department of Public Health

GISAID data displayed in the Genome Browser are subject to GISAID's Terms and Conditions. SARS-CoV-2 genome sequences and metadata are available for download from GISAID EpiCoV™

Logos below do not imply that these institutions endorse this presentation or any statements in it, but are only used to illustrate funding sources

Conflict of interest statement: Genome Browser staff get a percentage of license sales to commercial companies. All academic and non-profit use is always free. All specific SARS-CoV-2 display and analysis software in this talk has been released as open-source software (GPL or MIT)

National Human Genome Research Institute

NIH National Institute of Allergy and Infectious Diseases

California Department of Public Health

CDC

See: https://bit.ly/sc2licenses

- *I. “**You may not access and use the GISAID EpiFlu™ Database or collect, store, reproduce, access, modify, display, distribute, coordinate, arrange, and otherwise use the Data in connection with any other database related to influenza gene sequences, including, without limitation, by enabling others to access or use the Data through a separate portal or across a network of institutions, except for operators duly authorized by GISAID.”***
-

  Research groups cannot use GISAID data, but the GISAID website can and does integrate their tools.

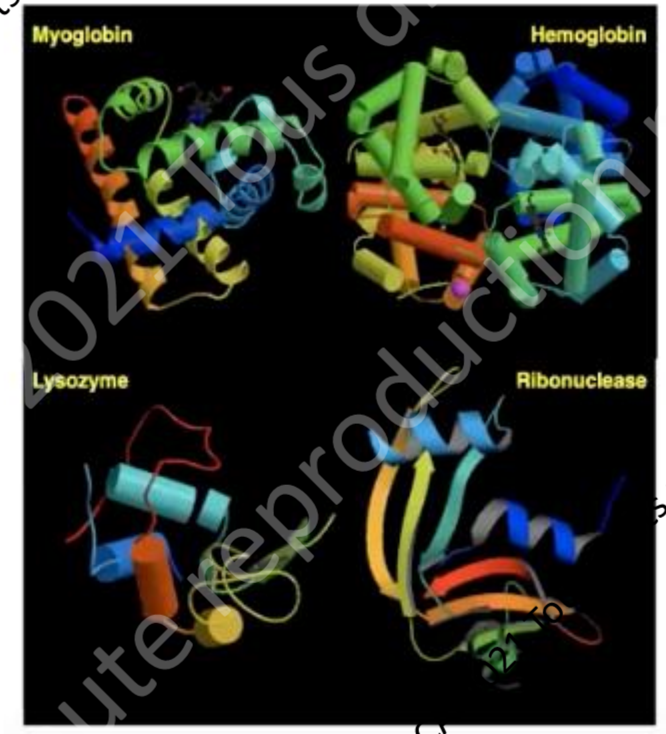
- *II. “****You may use Data to author, co-author or publish results obtained from your analyses of relevant Data, provided that any such published results acknowledge, as the original source of the Data, the laboratory where the clinical specimen(s) and/or virus isolate(s) were first obtained ("Originating Laboratory") and if applicable, the laboratory where Data have been generated from the isolate(s) and/or specimen(s) received and submitted to the GISAID EpiFlu™ Database. ("Submitting Laboratory"). “***
-

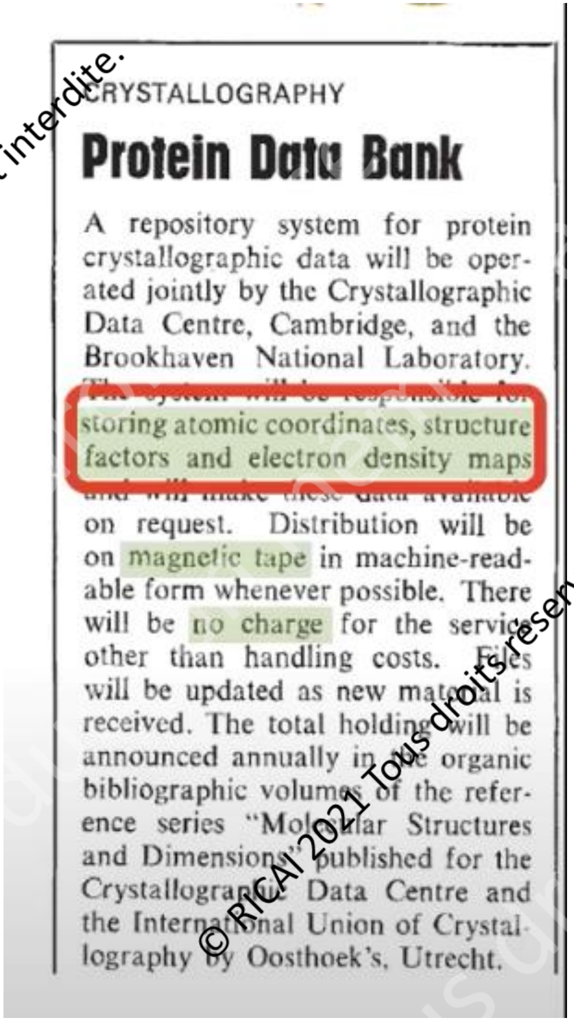
- “***You agree not to distribute Data to any thirdparty other than Authorized Users as contemplated by this Agreement.”***
- If you publish a software tool that assigns sequences to SARS-CoV-2 lineages, you cannot use GISAID sequences as references in your tool, as you’re not allowed to redistribute them with your software. The only possible way to share your software would be to obfuscate (encrypt) the sequences, which is contrary to all open-source / transparency rules put in place by scientific funding bodies for software publications over the last 20 years. At least one major software tool for SARS-CoV-2 analysis had to encrypt its reference sequence set to comply with this GISAID regulation.

- *IV. “**You acknowledge and agree that GISAID may, subject to any applicable laws, suspend access to all or any part of the GISAID EpiFlu™ Database and/or Data without any prior notice or liability to You.** “*

If someone at GISAID, an unnamed person that you cannot reach, decides that you somehow have infringed this license, your access can be blocked. This has happened in the past and is still happening, has been confirmed by statements from numerous institutes across the world, often anonymously for fear of retribution from GISAID, as reported in a Science article: https://science.sciencemag.org/content/371/6534/1086

Crystallography: Protein Data Bank, 1971





CRYSTALLOGRAPHY

Protein Data Bank

A repository system for protein crystallographic data will be operated jointly by the Crystallographic Data Centre, Cambridge, and the Brookhaven National Laboratory. The system will be responsible for storing atomic coordinates, structure factors and electron density maps and will make these data available on request. Distribution will be on magnetic tape in machine-readable form whenever possible. There will be no charge for the service other than handling costs. Files will be updated as new material is received. The total holding will be announced annually in the organic bibliographic volumes of the reference series "Molecular Structures and Dimensions" published for the Crystallographic Data Centre and the International Union of Crystallography by Oosthoek's, Utrecht.