Open databases and FAIR standards for SARS-CoV-2 genomic data

Matteo Chiara
matteo.chiara@unimi.it

www.elixir-europe.org

Aims

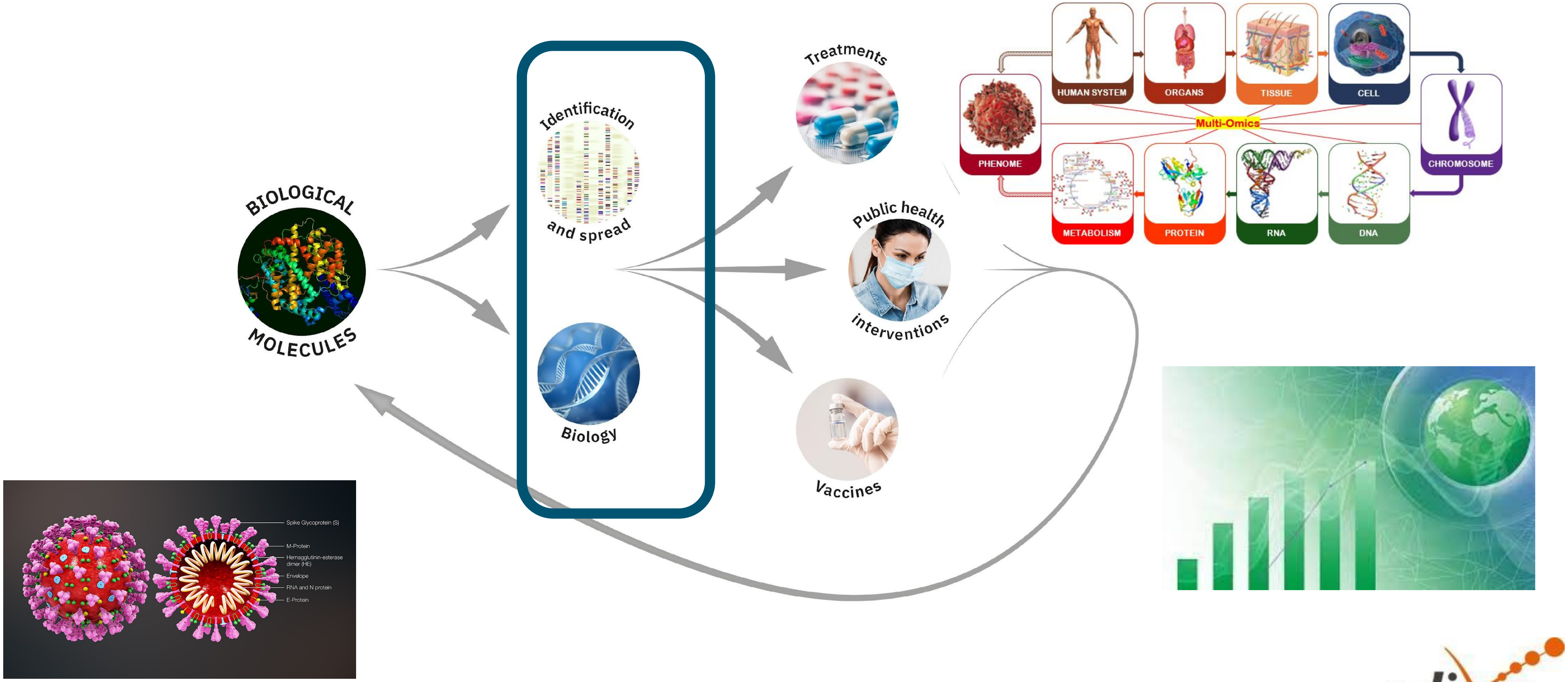
In this tutorial:

- guidelines and pointers for handling SARS-CoV-2 genome sequencing data
- links to useful tools and methods
- an overview (yet incomplete) of the main issues/problems
- a brief intro to genome data quality check
- a (hopefully) useful session of Q&A

This tutorial does not cover

- guidelines and methods for handling any other type of COVID-19 data

From data to action in infectious disease



What we do

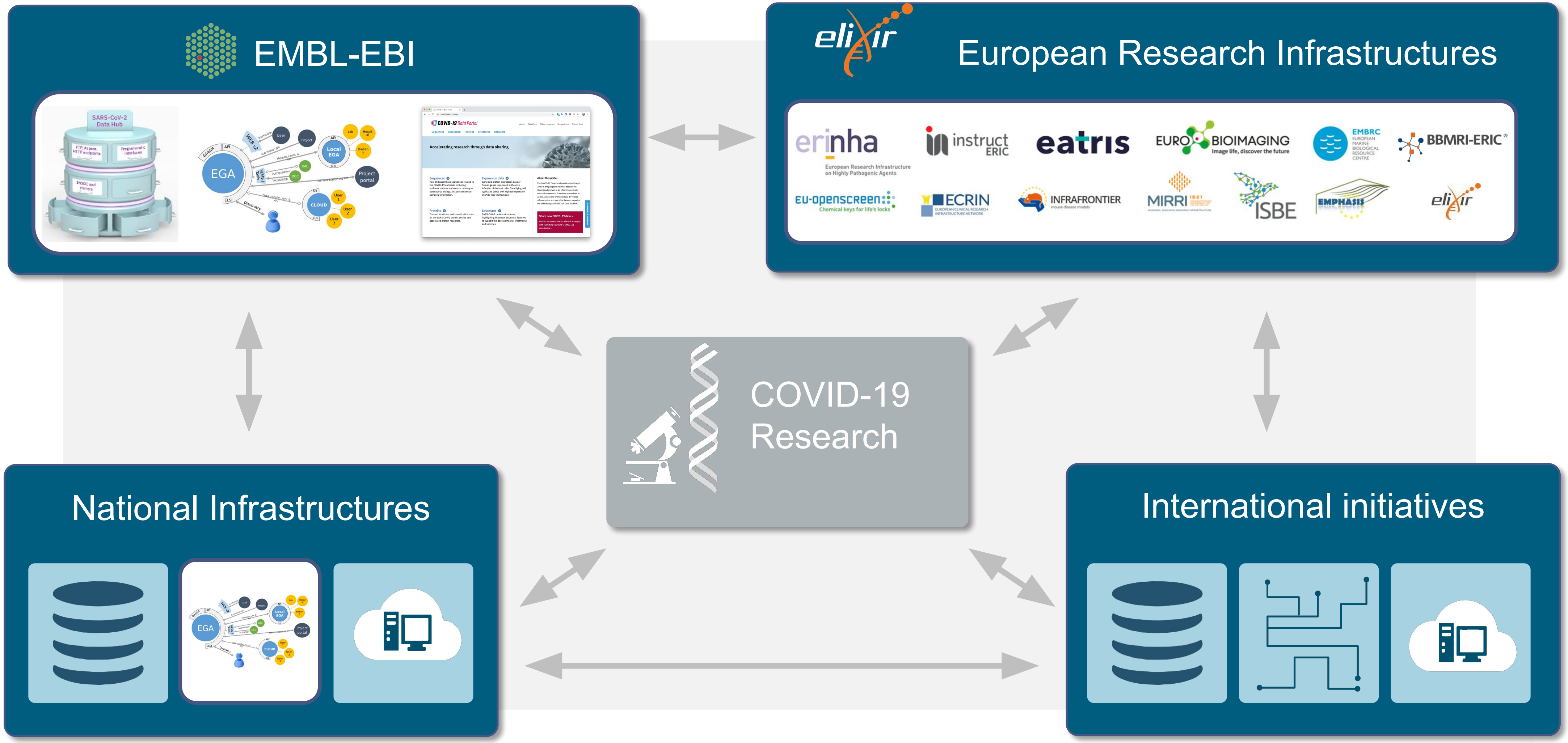
As ELIXIR-IT we are engaged in several activities and projects to

1. Make COVID-19 data FAIR
2. Develop/port services and tools for COVID-19 data analysis
3. Engage with stakeholders and colleagues

The COVID-19 data portal is the main one stop shop to check out the state of the art, or contribute to any of the above

COVID-19 Data Portal ITALY

European Federated COVID-19 Data Platform



EMBL-EBI

European Research Infrastructures

COVID-19 Research

National Infrastructures

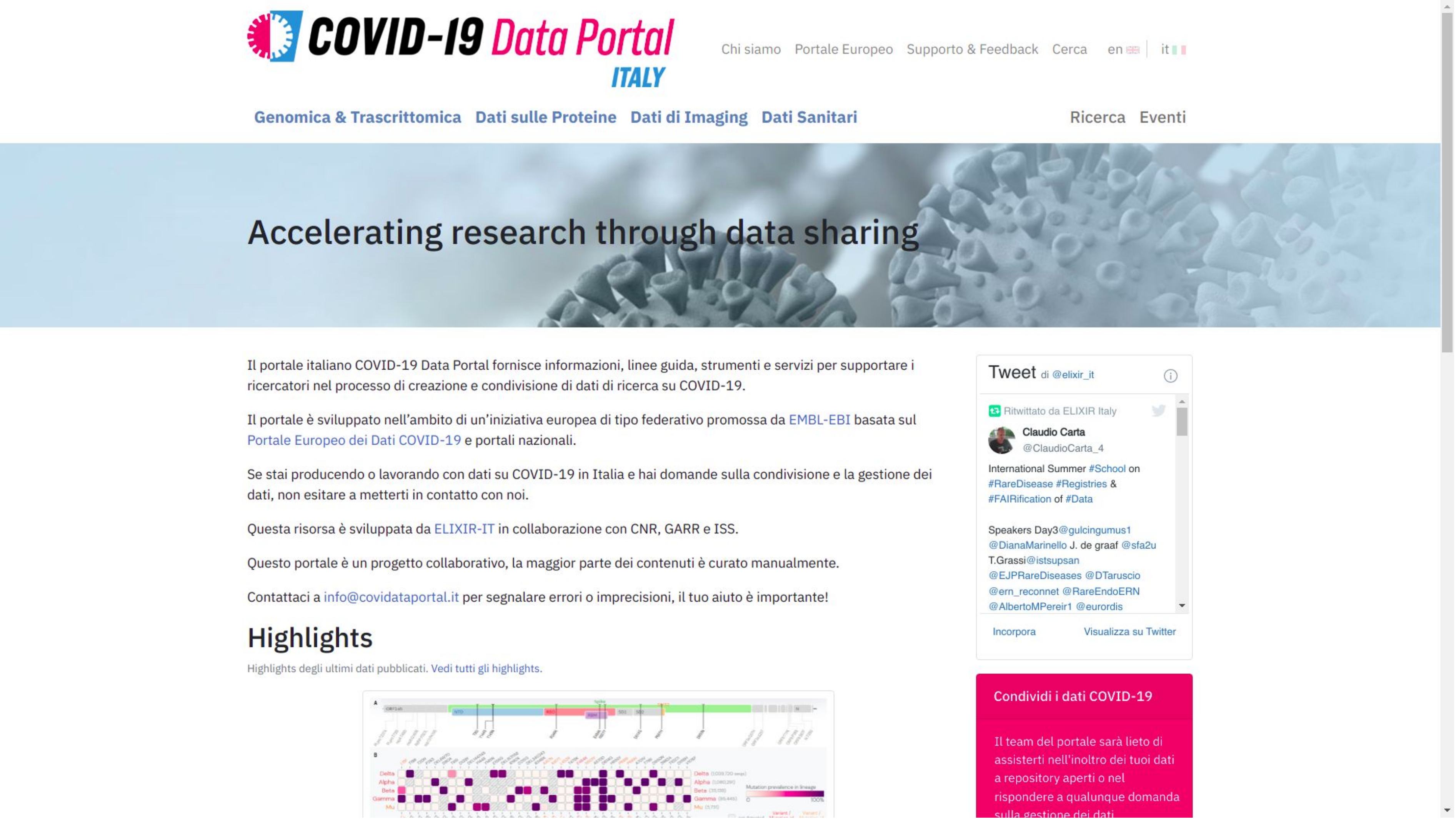
International initiatives

A one stop shop for COVID-19 data: the COVID-19 data-portal

- Unprecedented amounts of **data** were produced during the COVID-19 pandemic
- Making this data **available** and accessible is a fundamental prerequisite to advance our knowledge
- The EU has launched and international initiative to pro promote best practices for data sharing and curation: **COVID-19 data portal**
- We currently run the Italian instance

Findable
Accessible
Interoperable
Re-Usable

European Commission



COVID-19 Data Portal ITALY

Chi siamo Portale Europeo Supporto & Feedback Cerca en | it

Genomica & Trascrittomica Dati sulle Proteine Dati di Imaging Dati Sanitari Ricerca Eventi

Accelerating research through data sharing

Il portale italiano COVID-19 Data Portal fornisce informazioni, linee guida, strumenti e servizi per supportare i ricercatori nel processo di creazione e condivisione di dati di ricerca su COVID-19.

Il portale è sviluppato nell'ambito di un'iniziativa europea di tipo federativo promossa da EMBL-EBI basata sul Portale Europeo dei Dati COVID-19 e portali nazionali.

Se stai producendo o lavorando con dati su COVID-19 in Italia e hai domande sulla condivisione e la gestione dei dati, non esitare a metterti in contatto con noi.

Questa risorsa è sviluppata da ELIXIR-IT in collaborazione con CNR, GARR e ISS.

Questo portale è un progetto collaborativo, la maggior parte dei contenuti è curato manualmente.

Contattaci a info@covidataportal.it per segnalare errori o imprecisioni, il tuo aiuto è importante!

Highlights

Highlights degli ultimi dati pubblicati. Vedi tutti gli highlights.

Tweet di @elixir_it

Ritwittato da ELIXIR Italy

Claudio Carta
@ClaudioCarta_4

International Summer #School on #RareDisease #Registries & #FAIRification of #Data

Speakers Day3@gulcingumus1 @DianaMarinello J. de graaf @sfa2u T.Grassi@istsupsan @EJPRareDiseases @DTaruscio @ern_reconnet @RareEndoERN @AlbertoMPereir1 @eurordis

Incorpora Visualizza su Twitter

Condividi i dati COVID-19

Il team del portale sarà lieto di assisterti nell'inoltro dei tuoi dati a repository aperti o nel rispondere a qualunque domanda sulla gestione dei dati

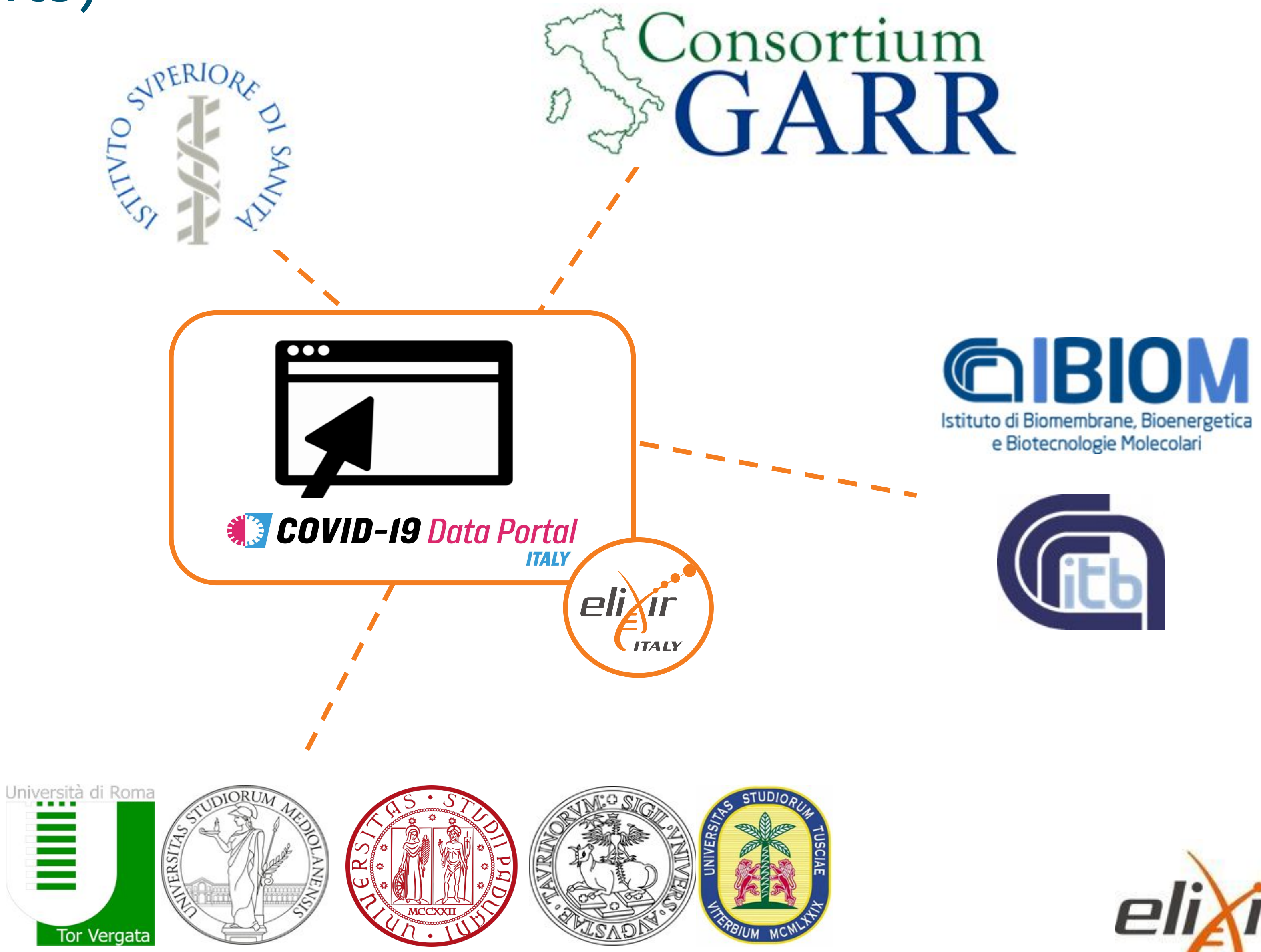
EMBL-EBI

elixir

elixir

https://www.covid19dataportal.it/

Institutions (credits)

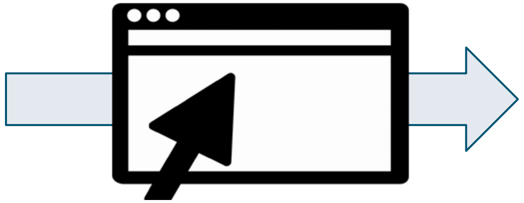

ISTITVTO SVPERIORE DI SANITÀ
Consortium GARR
COVID-19 Data Portal
ITALY
elixir
ITALY
IBIOM
Istituto di Biomembrane, Bioenergetica e Biotecnologie Molecolari
itb
Università di Roma
Tor Vergata
UNIVERSITAS STUDIORUM MEDIOLANENSIS
UNIVERSITAS STUDII PADUANI
MCCXXII
TAVRINORVM SIGIL VNIVERS AVGVSTAE
STUDIORUM UNIVERSITAS TUSCIAE
VITERBIUM MCMLXXIX

elixir

Why do we need a data portal in Italy (everywhere)

Research coordination issues within the Country

- Clinical research coordinated within region/institution
- Lack of national research facilities
- Limited Open Science, DM/DS practices awareness
- Inefficient efforts duplications
- Lack of dedicated funding

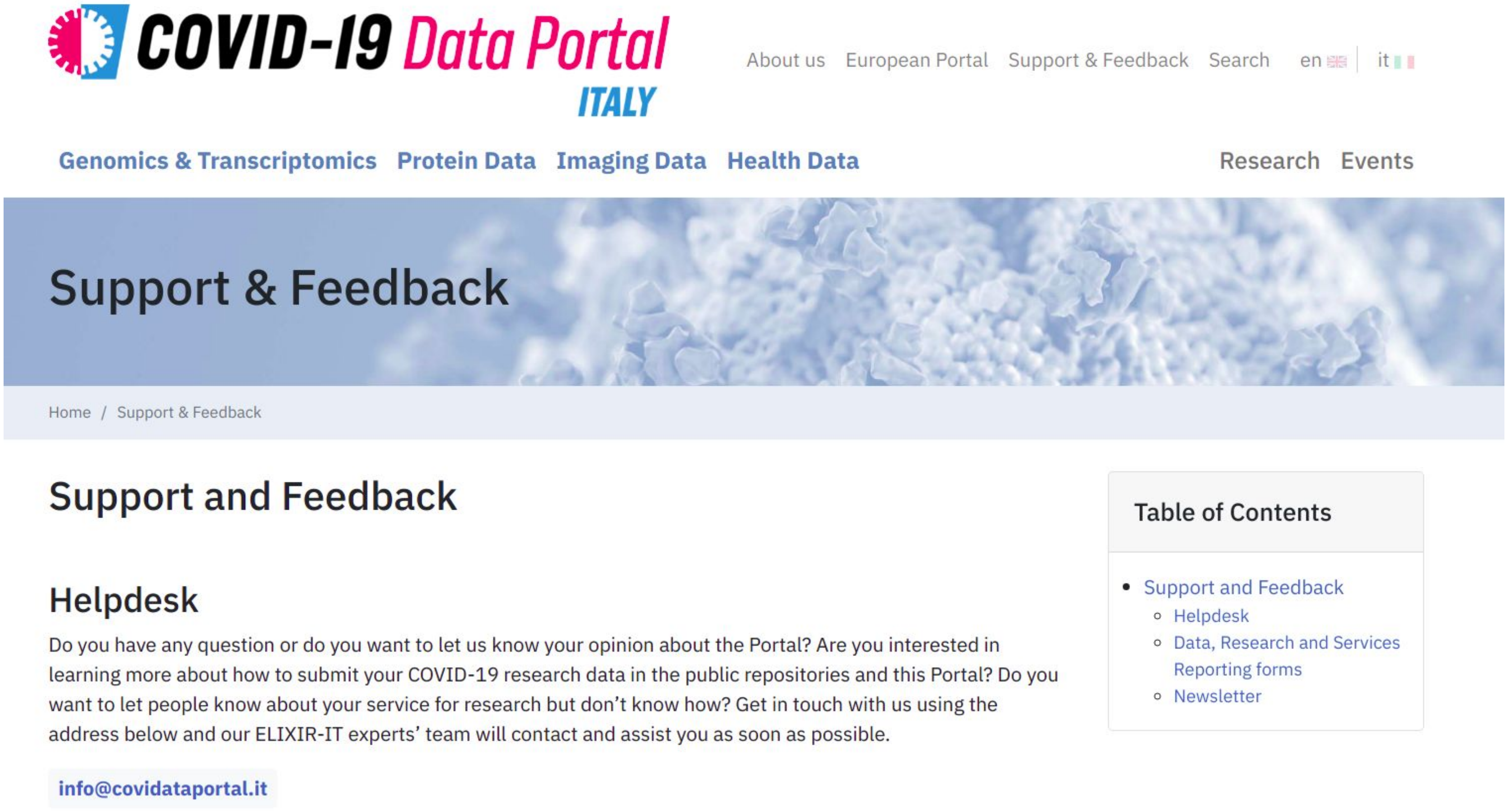


COVID-19 Data Portal ITALY

- Stimulate coordination between institutions
- Increase best practice implementation
- Rise Open Science, DM/DS practices awareness
- Increase use of national available resources
- Support joint grant applications

Better coordination

Data portal: how to contribute: LINK



COVID-19 Data Portal ITALY

About us European Portal Support & Feedback Search en it

Genomics & Transcriptomics Protein Data Imaging Data Health Data Research Events

Support & Feedback

Home / Support & Feedback

Support and Feedback

Helpdesk

Do you have any question or do you want to let us know your opinion about the Portal? Are you interested in learning more about how to submit your COVID-19 research data in the public repositories and this Portal? Do you want to let people know about your service for research but don’t know how? Get in touch with us using the address below and our ELIXIR-IT experts’ team will contact and assist you as soon as possible.

info@covidataportal.it

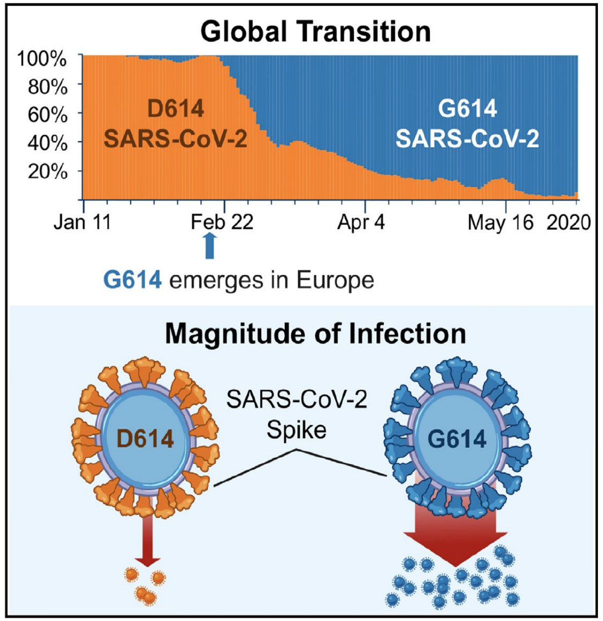
Table of Contents

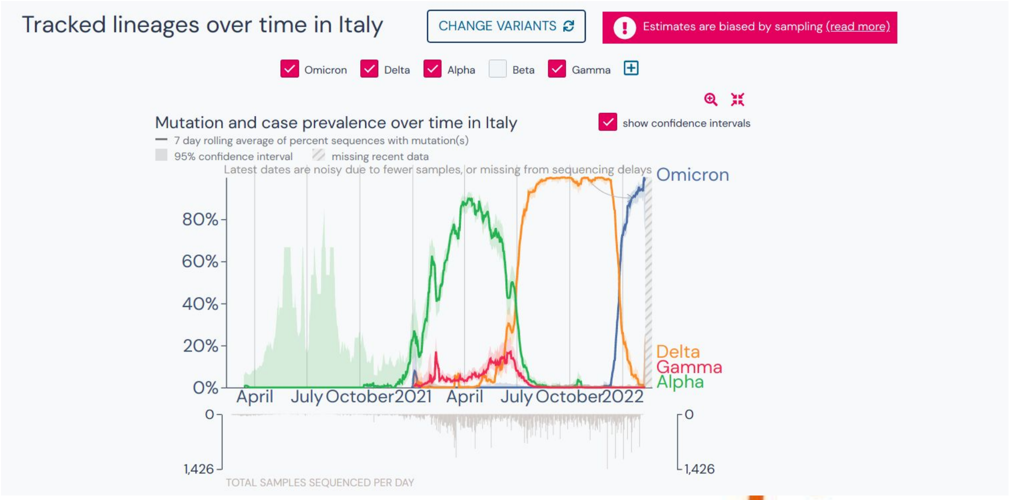
elixir

Why do we need SARS-CoV-2 genomes?

- Genomic surveillance: **to find and track viral variants**
- To **compare** data in space and time
- Identify dangerous variants
- More than 7.5M genome sequences form Jan 2020

These data are fundamental to fight COVID-19




Tracked lineages over time in Italy

CHANGE VARIANTS

Estimates are biased by sampling (read more)

☑ Omicron ☑ Delta ☑ Alpha □ Beta ☑ Gamma

Mutation and case prevalence over time in Italy

☑ show confidence intervals

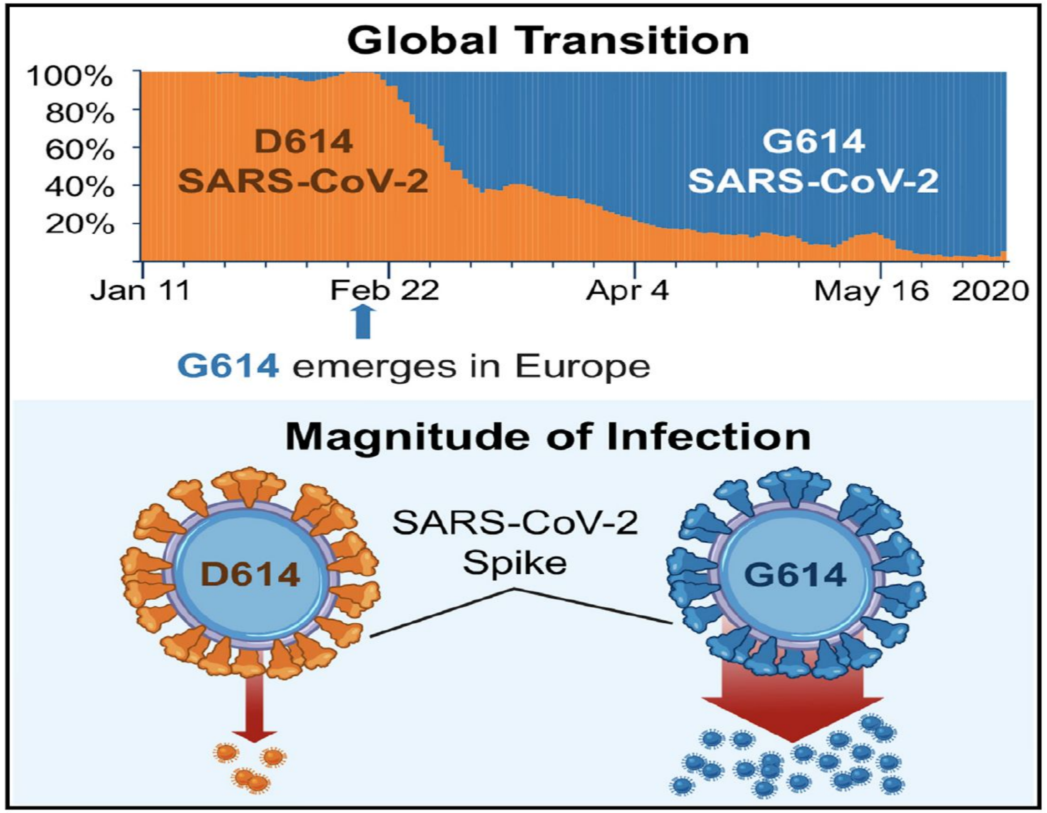
7 day rolling average of percent sequences with mutation(s)

95% confidence interval

missing recent data

Latest dates are noisy due to fewer samples, or missing from sequencing delays

Monitoring SARS-Cov-2 genome evolution



- At the end of March 2021 a novel allele variant of the spike protein (D614G) became highly prevalent worldwide. In different "geographic" areas.
- **Korber et al. (Cell, 2020):** Viruses carrying this allele variant have an increased capacity to infect cell lines (2x to 9x)
- All current variants of SARS-CoV-2 do now carry this mutation

- **Novel variants of the virus emerge by "selecting" advantageous mutations**

In 2021 D614G prevalently observed **outside China**, although analyses of genomes sequenced in January/February suggest that this variant originated in China. But not in Wuhan!

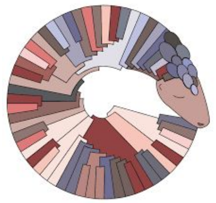
Korber et al, Cell 2020

Variant “hunting” starts with genomics

- Normally a **single** mutation **does not** significantly **change** the property of a virus
- To identify and track novel variants of the virus we need to observe and **track “combinations” of mutations**
- i.e. Do viruses that have specific combinations in their genome get better?

We need computational tools for this task: we currently have thousands of variants of SARS-CoV-2 (>1.5K). Only **a few are considered dangerous**

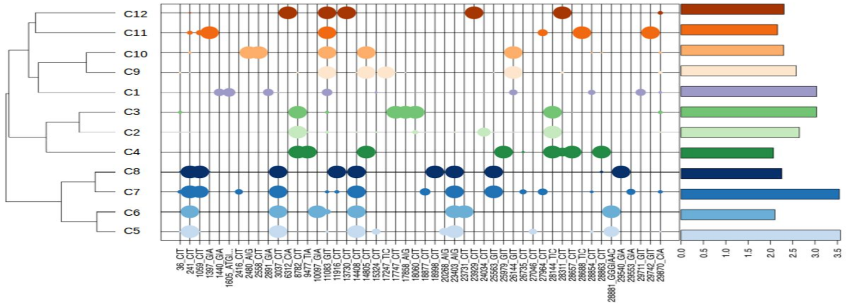
Different tools/methods to name/track variants



Pango. Rambaut et al

Nextstrain

Nextstrain. Hadfield et al



HaploCoV. Chiara et al

How do WE Identify “dangerous” variants?

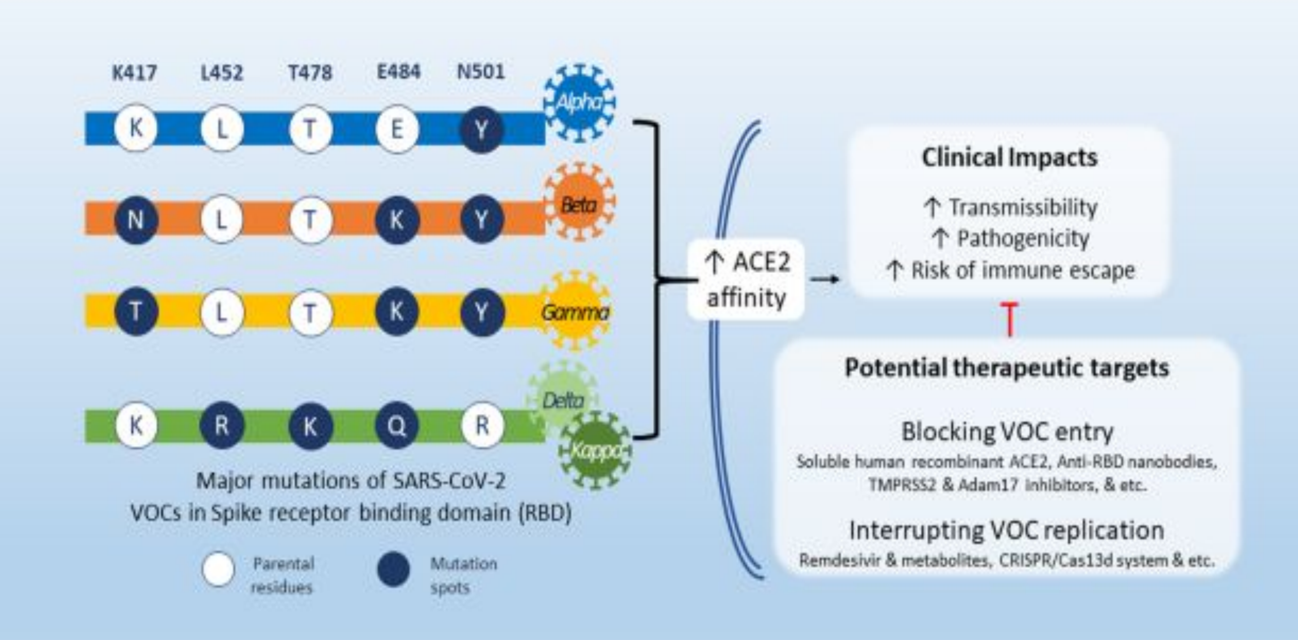
- International health Authorities define/identify novel variants based on epidemiological data (retrospectively)
- 3 (4 main classes)

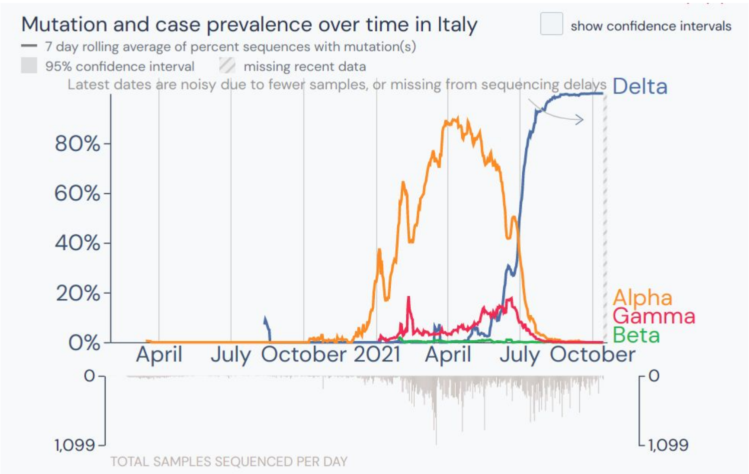
  - **VOC**: significant impact on transmissibility, severity and/or immunity. (total **5**, currently **4**)
  - **VOI**: potential impact on transmissibility, severity and/or immunity (based on genomic, not epidemiological data. (total 5, currently **3)**
  - **VUM**: weak evidence of a potential epidemiological impact (monitored since they could potentially evolve into more dangerous variants). Total **27**, currently **9**

- **Others**: the majority of the currently known variant. No advantage compared to the “Wuhan” strain of the virus. More than **1500** “variants”

How do WE Identify “dangerous” variants?

Mutations

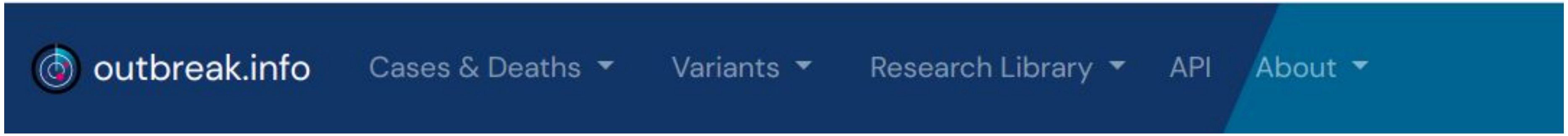

K417
L452
T478
E484
N501
K L T E Y
Alpha
N L T K Y
Beta
T L T K Y
Gamma
Delta
K R K Q R
Kappa
Major mutations of SARS-CoV-2
VOCs in Spike receptor binding domain (RBD)
Parental residues
Mutation spots
↑ ACE2 affinity
Clinical Impacts
↑ Transmissibility
↑ Pathogenicity
↑ Risk of immune escape
Potential therapeutic targets
Blocking VOC entry
Soluble human recombinant ACE2, Anti-RBD nanobodies, TMPRSS2 & Adam17 inhibitors, & etc.
Interrupting VOC replication
Remdesivir & metabolites, CRISPR/Cas13d system & etc.

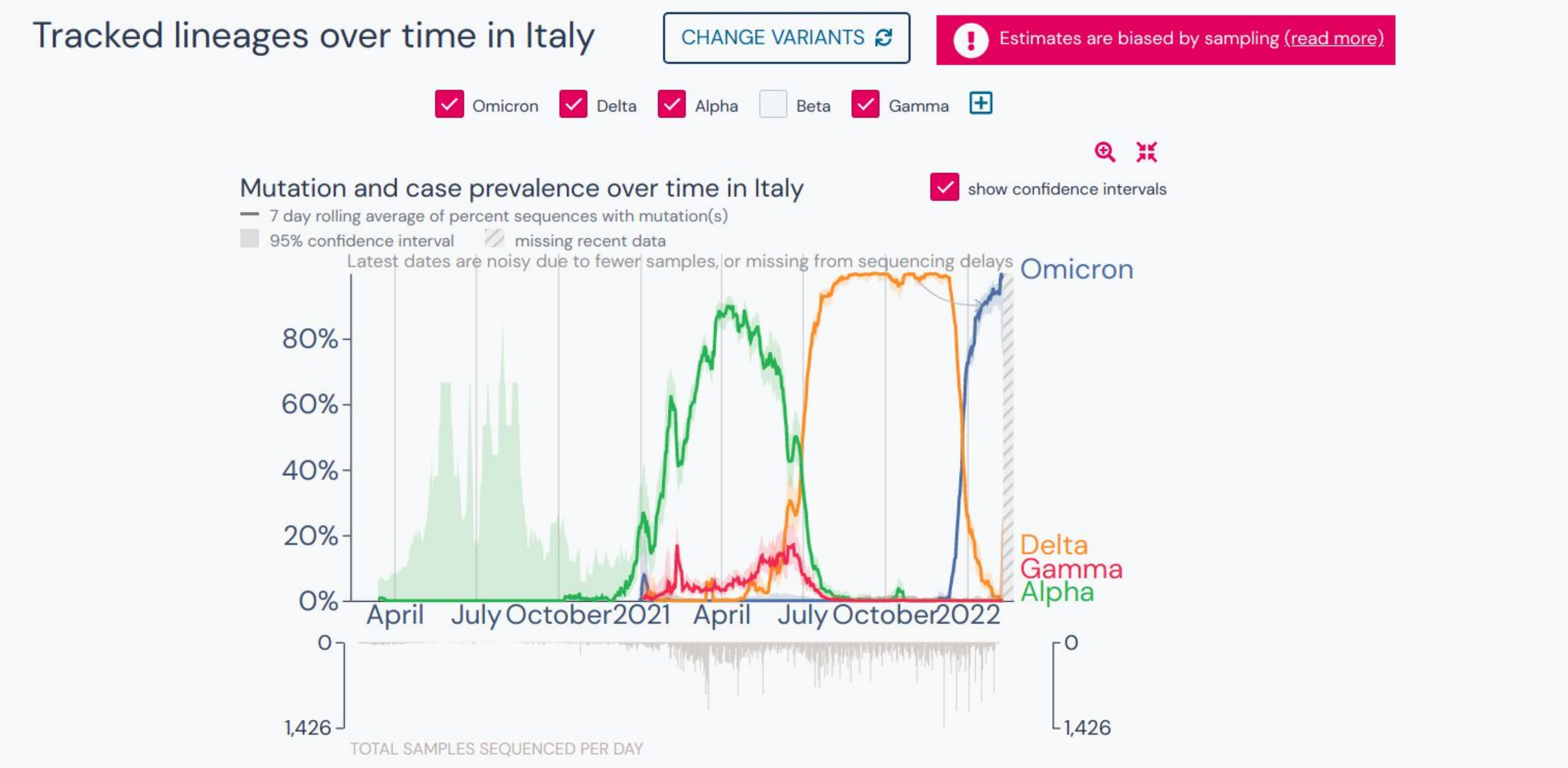


- Dangerous variants have an advantage over other variants, **hence they spread more rapidly**
- **This happened repeatedly for the 5* current variants of concern (VOC)**
- Right now we can only “spot” dangerous variants retrospectively: i.e track the variant, see what happens
- **Advantage, 3 VOCs (Alpha, Delta and Omicron) account for more than 60% of the total number of genome sequences**

How do we identify dangerous variants

In Italy, different lineages prevalent during different “waves”




Tracked lineages over time in Italy

CHANGE VARIANTS

Estimates are biased by sampling (read more)

☑ Omicron ☑ Delta ☑ Alpha □ Beta ☑ Gamma

Mutation and case prevalence over time in Italy

☑ show confidence intervals

— 7 day rolling average of percent sequences with mutation(s)

▨ 95% confidence interval ▨ missing recent data

Latest dates are noisy due to fewer samples, or missing from sequencing delays

From:
https://outbreak.info/

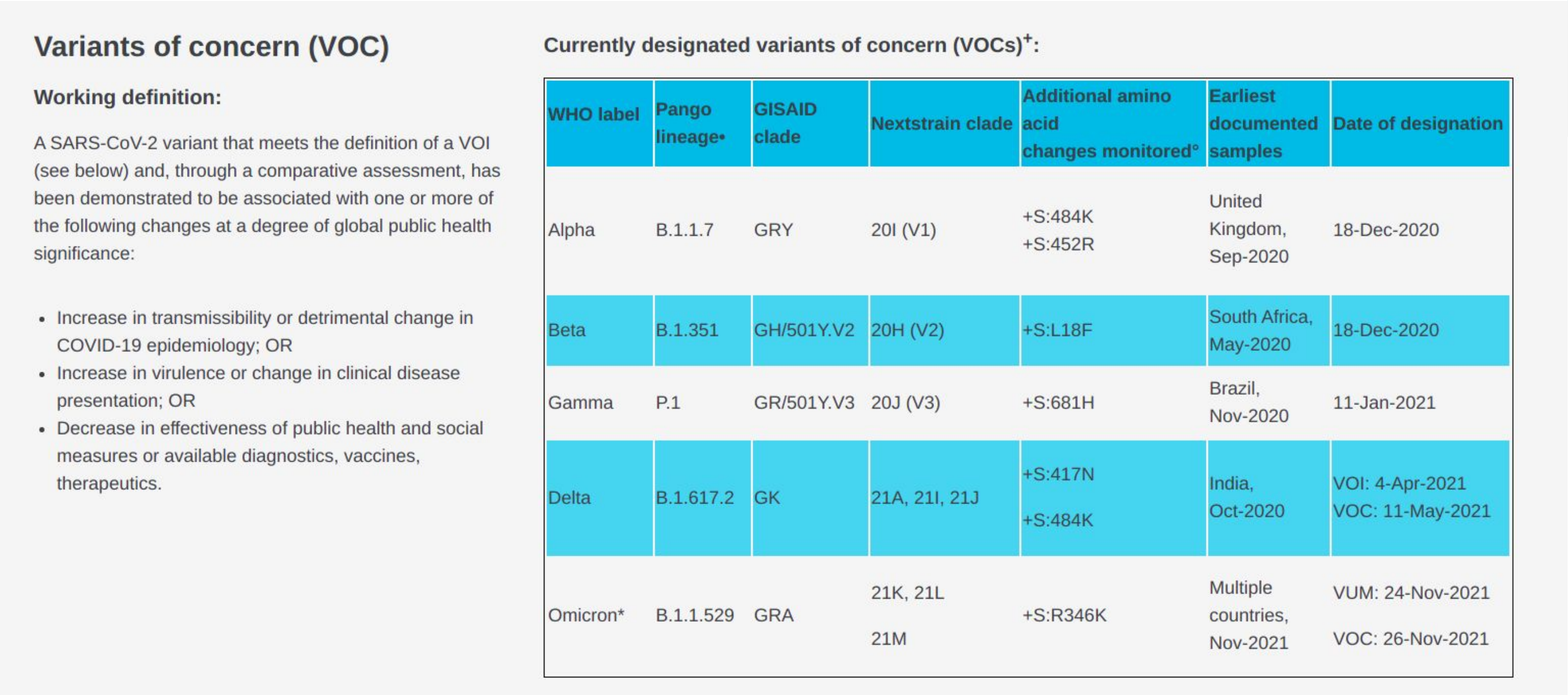
Tracking SARS-CoV-2 variants: WHO

LINK

31 May 2021

| Departmental news

WHO announces simple, easy-to-say labels for SARS-CoV-2 Variants of Interest and Concern



Variants of concern (VOC)

Working definition:

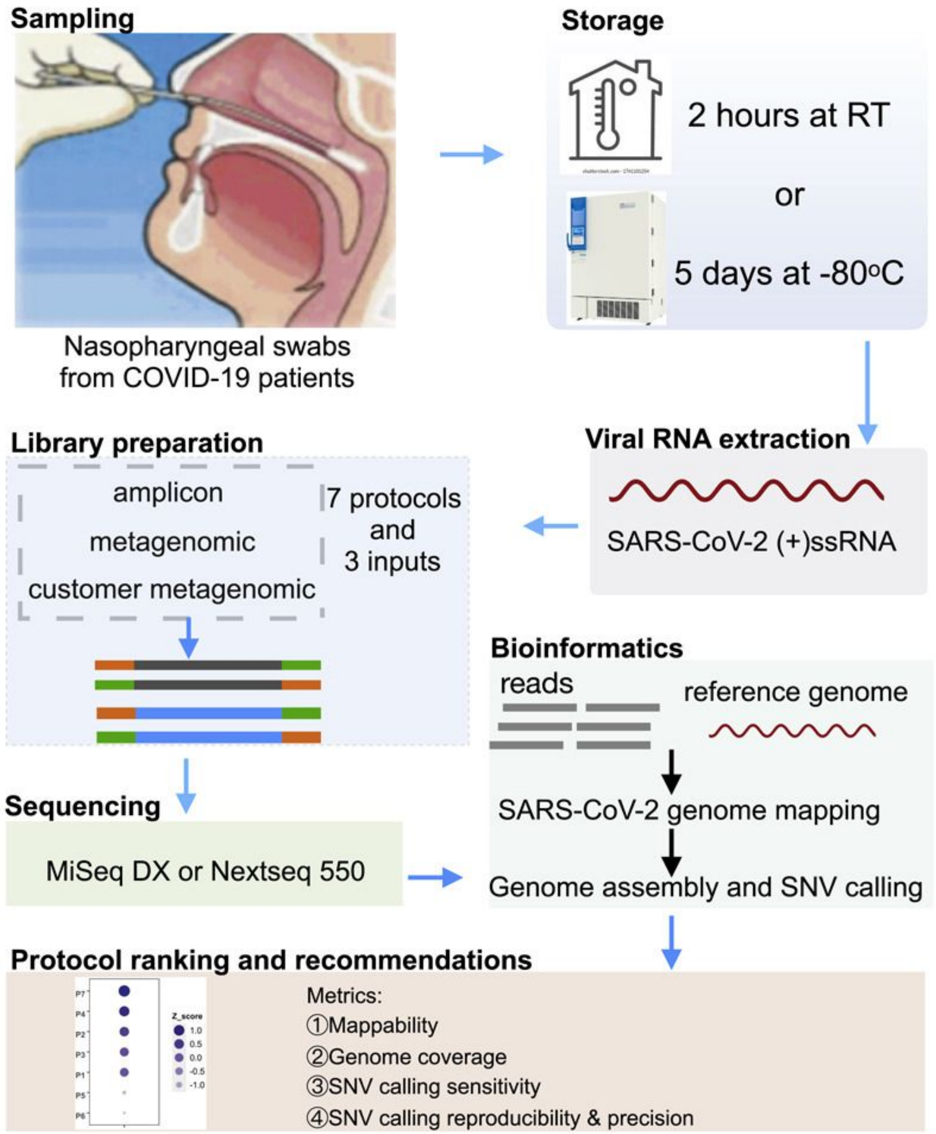
A SARS-CoV-2 variant that meets the definition of a VOI (see below) and, through a comparative assessment, has been demonstrated to be associated with one or more of the following changes at a degree of global public health significance:

- Increase in transmissibility or detrimental change in COVID-19 epidemiology; OR
- Increase in virulence or change in clinical disease presentation; OR
- Decrease in effectiveness of public health and social measures or available diagnostics, vaccines, therapeutics.

Currently designated variants of concern (VOCs)$^{+}$:

WHO label	Pango lineage•	GISAID clade	Nextstrain clade	Additional amino acid changes monitored°	Earliest documented samples	Date of designation
Alpha	B.1.1.7	GRY	20I (V1)	+S:484K +S:452R	United Kingdom, Sep-2020	18-Dec-2020
Beta	B.1.351	GH/501Y.V2	20H (V2)	+S:L18F	South Africa, May-2020	18-Dec-2020
Gamma	P.1	GR/501Y.V3	20J (V3)	+S:681H	Brazil, Nov-2020	11-Jan-2021
Delta	B.1.617.2	GK	21A, 21I, 21J	+S:417N +S:484K	India, Oct-2020	VOI: 4-Apr-2021 VOC: 11-May-2021
Omicron*	B.1.1.529	GRA	21K, 21L 21M	+S:R346K	Multiple countries, Nov-2021	VUM: 24-Nov-2021 VOC: 26-Nov-2021

Where do SARS-CoV-2 genomes come from?



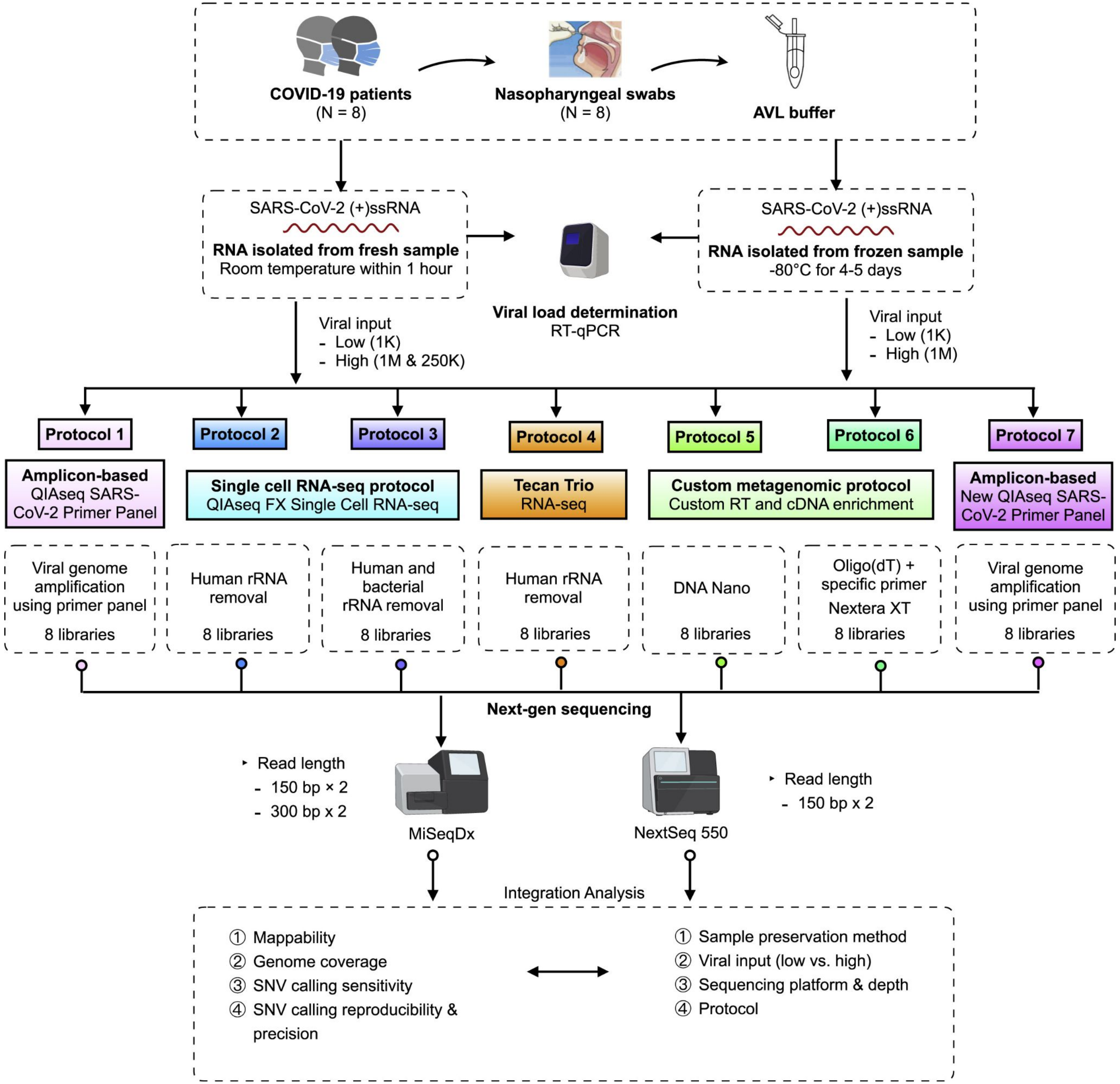
From:
https://doi.org/10.1016/j.isci.2021.102892

- People who got COVID-19 (mostly)

- RNAs extracted from swabs are sequenced with different methods

- A plethora of protocols do exist!

The hollow truth



COVID-19 patients (N = 8)

Nasopharyngeal swabs (N = 8)

AVL buffer

SARS-CoV-2 (+)ssRNA

RNA isolated from fresh sample
Room temperature within 1 hour

SARS-CoV-2 (+)ssRNA

RNA isolated from frozen sample
-80°C for 4-5 days

Viral load determination
RT-qPCR

Viral input
- Low (1K)
- High (1M & 250K)

Viral input
- Low (1K)
- High (1M)

Protocol 1 – **Amplicon-based** QIAseq SARS-CoV-2 Primer Panel – Viral genome amplification using primer panel – 8 libraries

Protocol 2 – **Single cell RNA-seq protocol** QIAseq FX Single Cell RNA-seq – Human rRNA removal – 8 libraries

Protocol 3 – Human and bacterial rRNA removal – 8 libraries

Protocol 4 – **Tecan Trio** RNA-seq – Human rRNA removal – 8 libraries

Protocol 5 – **Custom metagenomic protocol** Custom RT and cDNA enrichment – DNA Nano – 8 libraries

Protocol 6 – Oligo(dT) + specific primer Nextera XT – 8 libraries

Protocol 7 – **Amplicon-based** New QIAseq SARS-CoV-2 Primer Panel – Viral genome amplification using primer panel – 8 libraries

Next-gen sequencing

- Read length
  - 150 bp × 2
  - 300 bp x 2

MiSeqDx

NextSeq 550

- Read length
  - 150 bp x 2

Integration Analysis

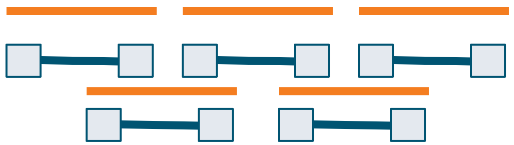
① Mappability
② Genome coverage
③ SNV calling sensitivity
④ SNV calling reproducibility & precision

① Sample preservation method
② Viral input (low vs. high)
③ Sequencing platform & depth
④ Protocol

- Bioinformatics and lab protocols can be complex!
- Sometimes need to tailor adjust things for specific protocols!
- If you want “FAIR” data you need to keep track of everything you do
- Which is a **remarkable effort**

How do we get SARS-CoV-2 genomes

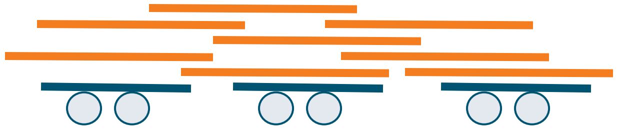
Amplicon (PCR)



- Need reference genome (**bias**)
- PCR drop-out
- Reference guided
- Little or no "contamination"

$$

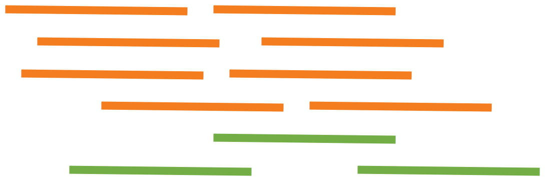
Hybrid capture



- Need reference genome (**bias**)
- Robust to variation
- Reference guided
- Contaminant sequences?

$$$

Shotgun(meta)

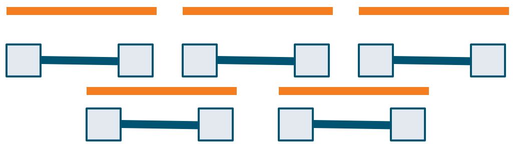


- Reference genome not strictly needed
- Not affected by variation
- *de-novo* assembly possible
- Contaminant sequences (human?)

$$$$

Bioinformatics analyses

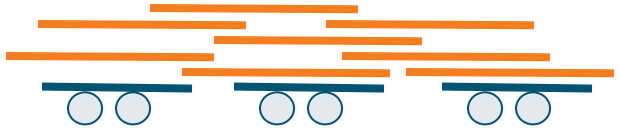
Amplicon (PCR)



- Carefully check primers
- Minimum coverage?
- Co-infections?

$$

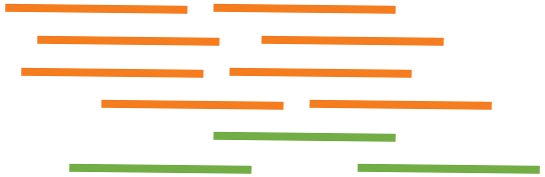
Hybrid capture



- Minimum coverage
- Co-infections?

$$$

Shotgun(meta)

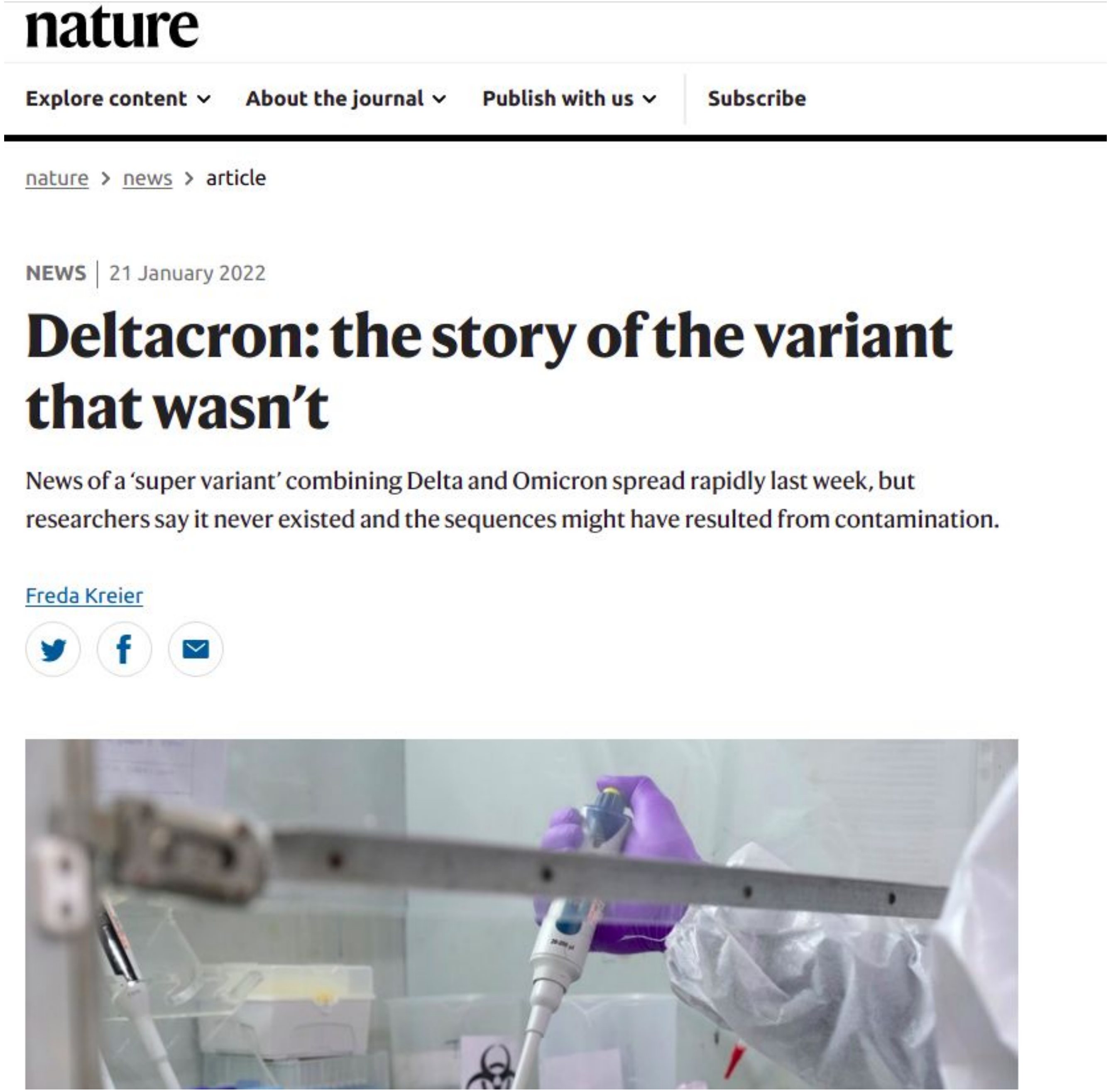


- Need to remove human contaminants
- Uniform coverage
- Co-infections

$$$$

Different sequencing methods require different workflows:

- Bioinformatics required to get the “final” consensus sequence



NEWS | 21 January 2022

Deltacron: the story of the variant that wasn't

News of a 'super variant' combining Delta and Omicron spread rapidly last week, but researchers say it never existed and the sequences might have resulted from contamination.

Freda Kreier

- Bioinformatics analysis is an integral part of SARS-CoV-2 genomics
  - Can introduce errors/biases
  - Need to be reproducible

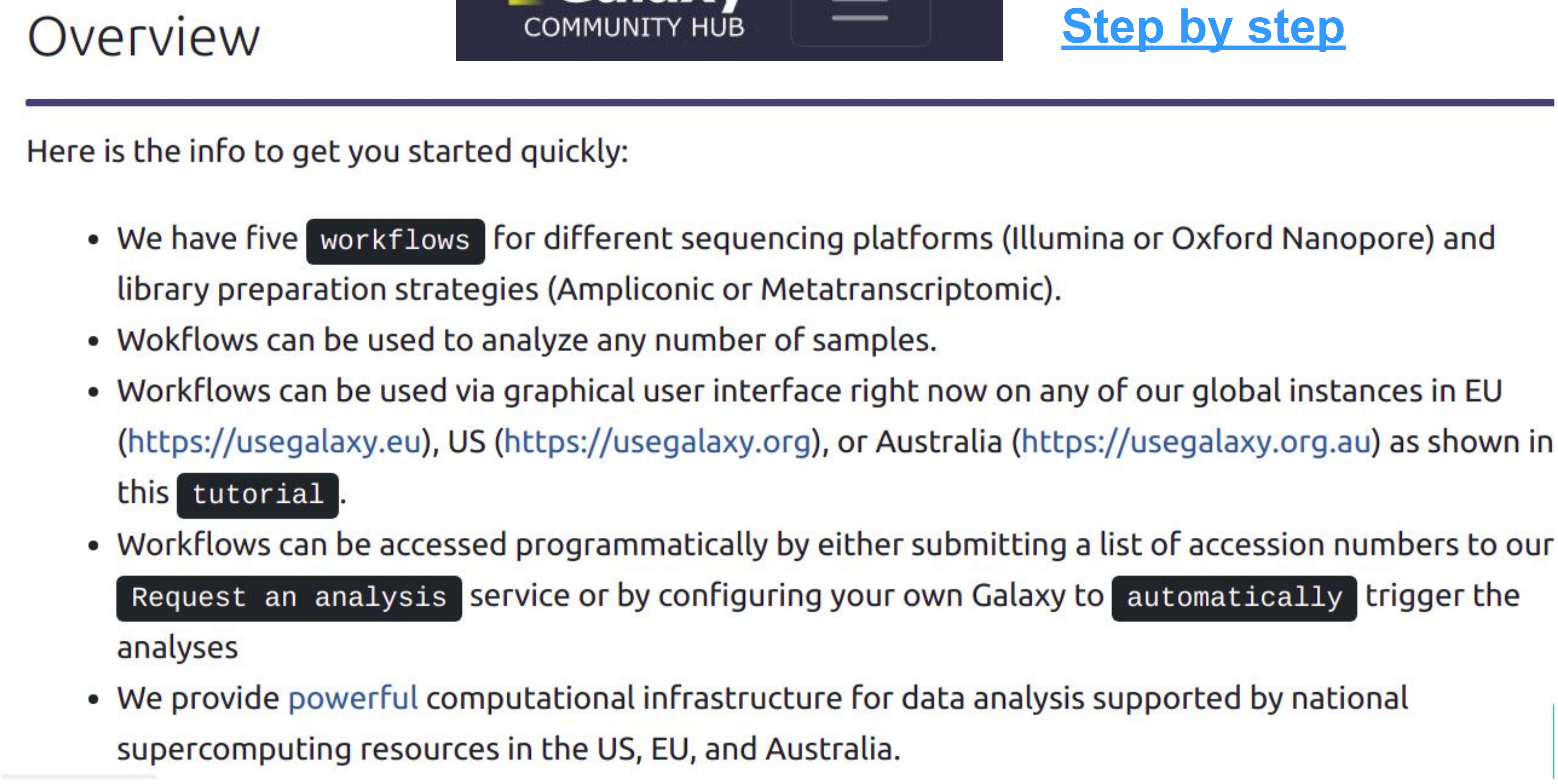
- If/when possible it would be highly advisable to
  - 1 check results carefully
  - 2 use high quality, reproducible workflows
  - Or Alternatively, to publish yours somewhere

https://workflowhub.eu/

Workflows

For more information: https://pubmed.ncbi.nlm.nih.gov/32790776/

Galaxy COVID-19 Step by step

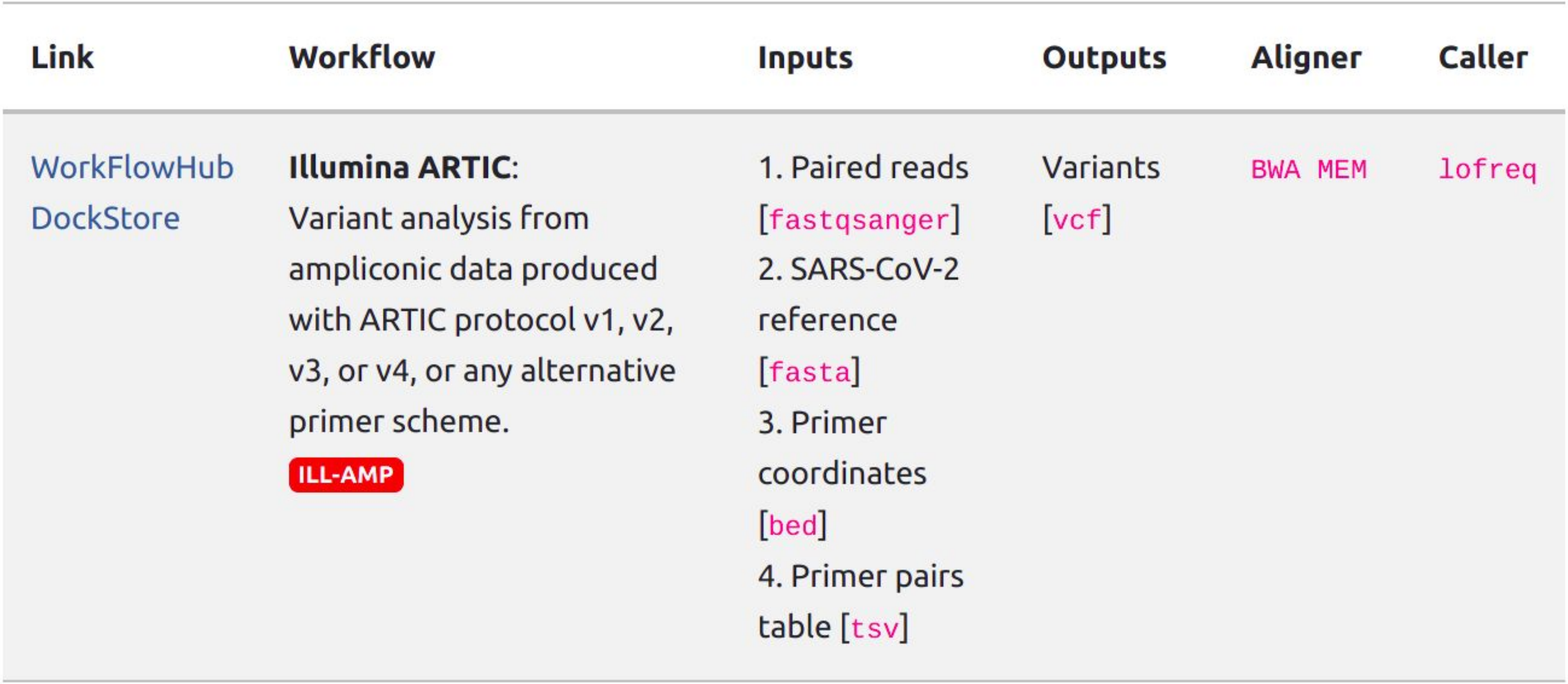


Overview

Here is the info to get you started quickly:

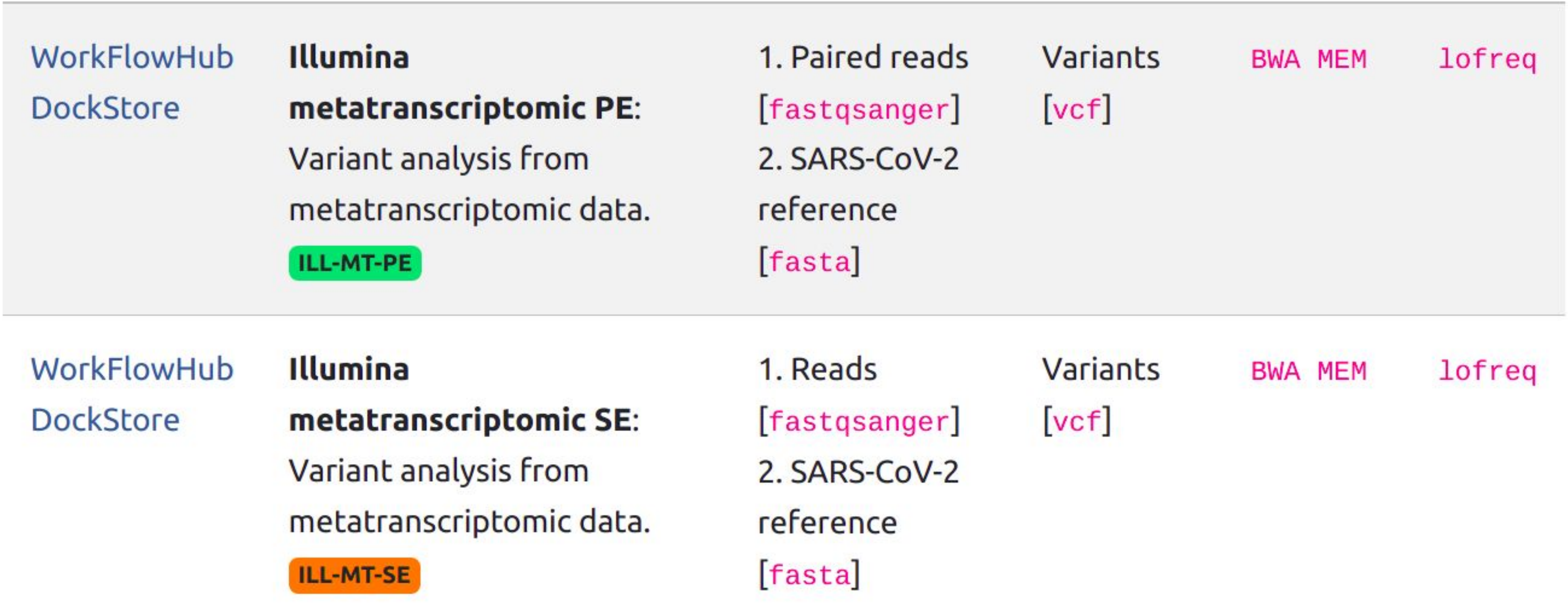
- We have five `workflows` for different sequencing platforms (Illumina or Oxford Nanopore) and library preparation strategies (Ampliconic or Metatranscriptomic).
- Wokflows can be used to analyze any number of samples.
- Workflows can be used via graphical user interface right now on any of our global instances in EU (https://usegalaxy.eu), US (https://usegalaxy.org), or Australia (https://usegalaxy.org.au) as shown in this `tutorial`.
- Workflows can be accessed programmatically by either submitting a list of accession numbers to our `Request an analysis` service or by configuring your own Galaxy to `automatically` trigger the analyses
- We provide powerful computational infrastructure for data analysis supported by national supercomputing resources in the US, EU, and Australia.

Galaxy COVID-19

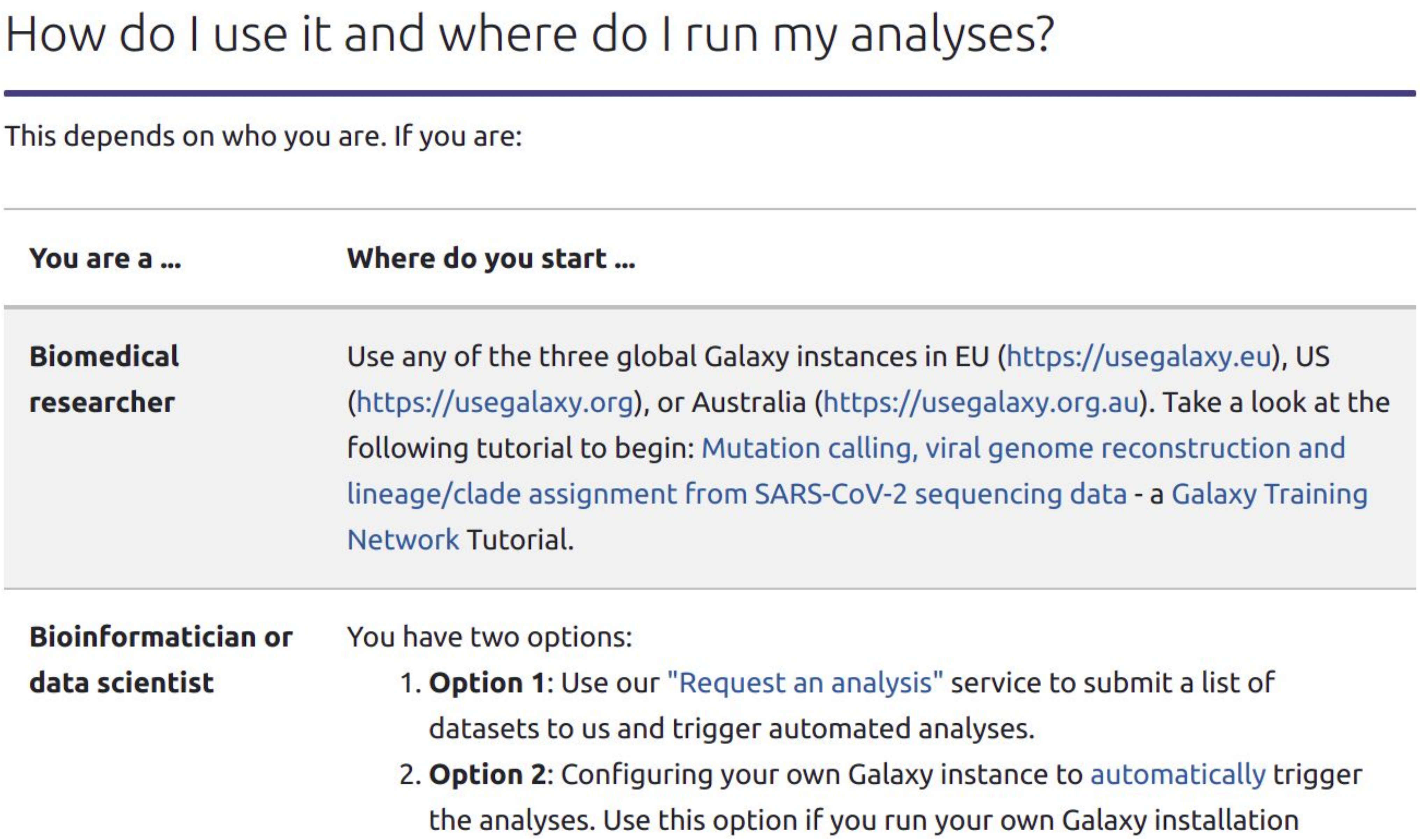


Link	Workflow	Inputs	Outputs	Aligner	Caller
WorkFlowHub DockStore	**Illumina ARTIC**: Variant analysis from ampliconic data produced with ARTIC protocol v1, v2, v3, or v4, or any alternative primer scheme. ILL-AMP	1. Paired reads [`fastqsanger`] 2. SARS-CoV-2 reference [`fasta`] 3. Primer coordinates [`bed`] 4. Primer pairs table [`tsv`]	Variants [`vcf`]	`BWA MEM`	`lofreq`

Galaxy COVID-19



WorkFlowHub DockStore	**Illumina metatranscriptomic PE**: Variant analysis from metatranscriptomic data. ILL-MT-PE	1. Paired reads [fastqsanger] 2. SARS-CoV-2 reference [fasta]	Variants [vcf]	BWA MEM	lofreq
WorkFlowHub DockStore	**Illumina metatranscriptomic SE**: Variant analysis from metatranscriptomic data. ILL-MT-SE	1. Reads [fastqsanger] 2. SARS-CoV-2 reference [fasta]	Variants [vcf]	BWA MEM	lofreq

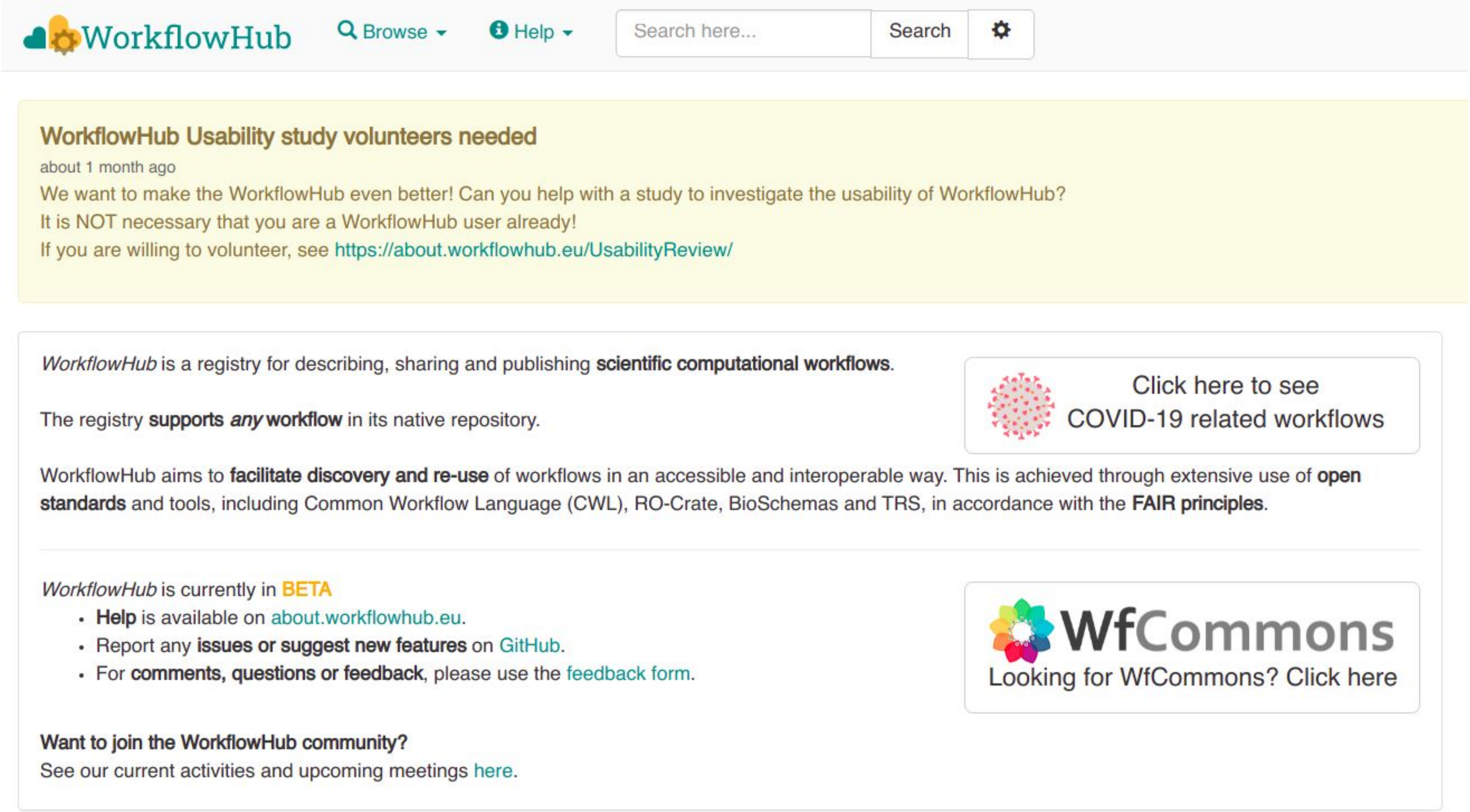


How do I use it and where do I run my analyses?

This depends on who you are. If you are:

You are a ...	Where do you start ...
Biomedical researcher	Use any of the three global Galaxy instances in EU (https://usegalaxy.eu), US (https://usegalaxy.org), or Australia (https://usegalaxy.org.au). Take a look at the following tutorial to begin: Mutation calling, viral genome reconstruction and lineage/clade assignment from SARS-CoV-2 sequencing data - a Galaxy Training Network Tutorial.
Bioinformatician or data scientist	You have two options: 1. **Option 1**: Use our "Request an analysis" service to submit a list of datasets to us and trigger automated analyses. 2. **Option 2**: Configuring your own Galaxy instance to automatically trigger the analyses. Use this option if you run your own Galaxy installation

Where can I publish my WFs?



WorkflowHub Usability study volunteers needed

about 1 month ago

We want to make the WorkflowHub even better! Can you help with a study to investigate the usability of WorkflowHub?
It is NOT necessary that you are a WorkflowHub user already!
If you are willing to volunteer, see https://about.workflowhub.eu/UsabilityReview/

WorkflowHub is a registry for describing, sharing and publishing **scientific computational workflows**.

The registry **supports *any* workflow** in its native repository.

WorkflowHub aims to **facilitate discovery and re-use** of workflows in an accessible and interoperable way. This is achieved through extensive use of **open standards** and tools, including Common Workflow Language (CWL), RO-Crate, BioSchemas and TRS, in accordance with the **FAIR principles**.

Click here to see COVID-19 related workflows

WorkflowHub is currently in BETA

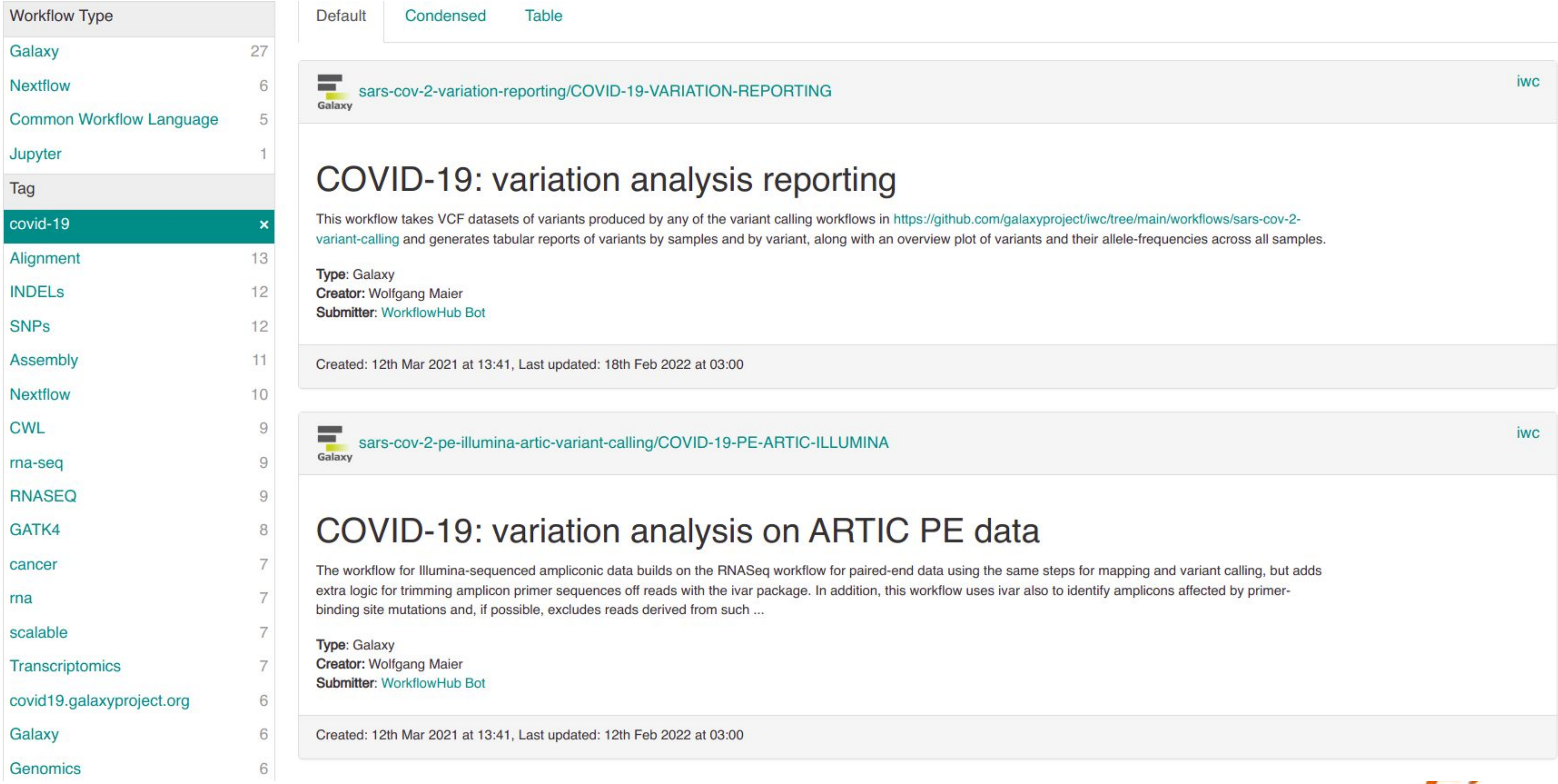
- **Help** is available on about.workflowhub.eu.
- Report any **issues or suggest new features** on GitHub.
- For **comments, questions or feedback**, please use the feedback form.

WfCommons
Looking for WfCommons? Click here

Want to join the WorkflowHub community?
See our current activities and upcoming meetings here.

Click here to see
COVID-19 related workflows

https://workflowhub.eu/



Workflow Type	
Galaxy	27
Nextflow	6
Common Workflow Language	5
Jupyter	1

Tag	
covid-19	×
Alignment	13
INDELs	12
SNPs	12
Assembly	11
Nextflow	10
CWL	9
rna-seq	9
RNASEQ	9
GATK4	8
cancer	7
rna	7
scalable	7
Transcriptomics	7
covid19.galaxyproject.org	6
Galaxy	6
Genomics	6

sars-cov-2-variation-reporting/COVID-19-VARIATION-REPORTING

iwc

COVID-19: variation analysis reporting

This workflow takes VCF datasets of variants produced by any of the variant calling workflows in https://github.com/galaxyproject/iwc/tree/main/workflows/sars-cov-2-variant-calling and generates tabular reports of variants by samples and by variant, along with an overview plot of variants and their allele-frequencies across all samples.

Type: Galaxy
Creator: Wolfgang Maier
Submitter: WorkflowHub Bot

Created: 12th Mar 2021 at 13:41, Last updated: 18th Feb 2022 at 03:00

sars-cov-2-pe-illumina-artic-variant-calling/COVID-19-PE-ARTIC-ILLUMINA

iwc

COVID-19: variation analysis on ARTIC PE data

The workflow for Illumina-sequenced ampliconic data builds on the RNASeq workflow for paired-end data using the same steps for mapping and variant calling, but adds extra logic for trimming amplicon primer sequences off reads with the ivar package. In addition, this workflow uses ivar also to identify amplicons affected by primer-binding site mutations and, if possible, excludes reads derived from such ...

Type: Galaxy
Creator: Wolfgang Maier
Submitter: WorkflowHub Bot

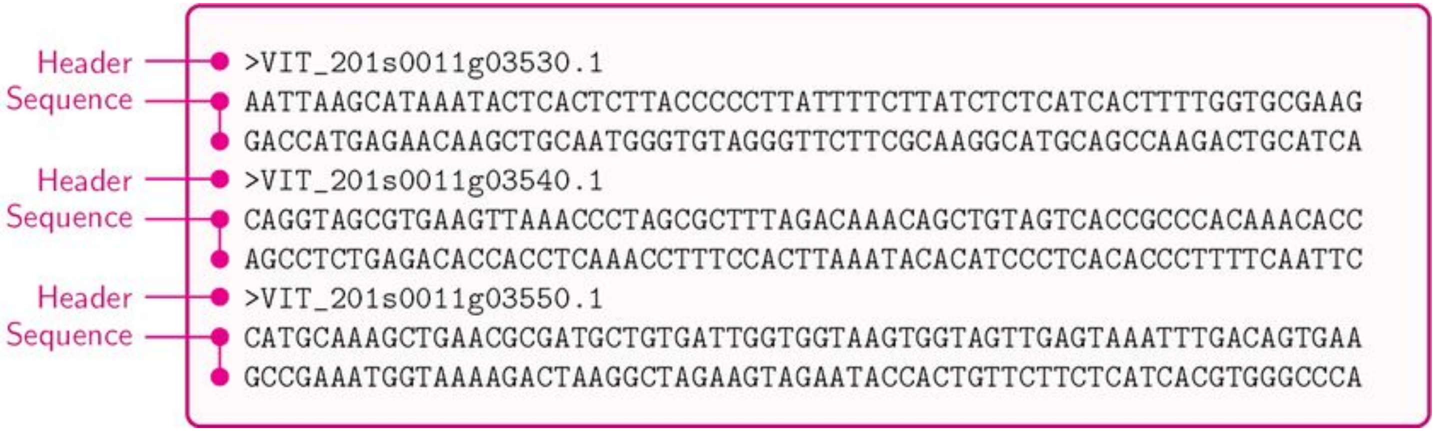
Created: 12th Mar 2021 at 13:41, Last updated: 12th Feb 2022 at 03:00

What do I get in the end?

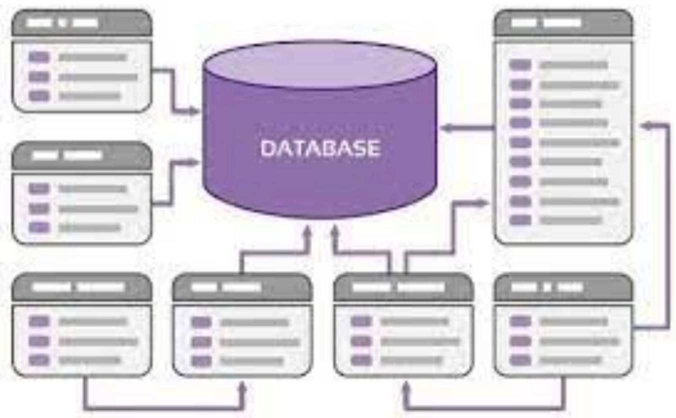
- A (consensus) genome sequence

- In fasta format

- Data stewards: make the sequence data, and metadata available to the scientific community*



```
>VIT_201s0011g03530.1
AATTAAGCATAAATACTCACTCTTACCCCCTTATTTTCTTATCTCTCATCACTTTTGGTGCGAAG
GACCATGAGAACAAGCTGCAATGGGTGTAGGGTTCTTCGCAAGGCATGCAGCCAAGACTGCATCA
>VIT_201s0011g03540.1
CAGGTAGCGTGAAGTTAAACCCTAGCGCTTTAGACAAACAGCTGTAGTCACCGCCCACAAACACC
AGCCTCTGAGACACCACCTCAAACCTTTCCACTTAAATACACATCCCTCACACCCTTTTCAATTC
>VIT_201s0011g03550.1
CATGCAAAGCTGAACGCGATGCTGTGATTGGTGGTAAGTGGTAGTTGAGTAAATTTGACAGTGAA
GCCGAAATGGTAAAAGACTAAGGCTAGAAGTAGAATACCACTGTTCTTCTCATCACGTGGGCCCA
```



* in accordance with GDPR/ELSI

Where to submit genome data?

INSDC

ENA

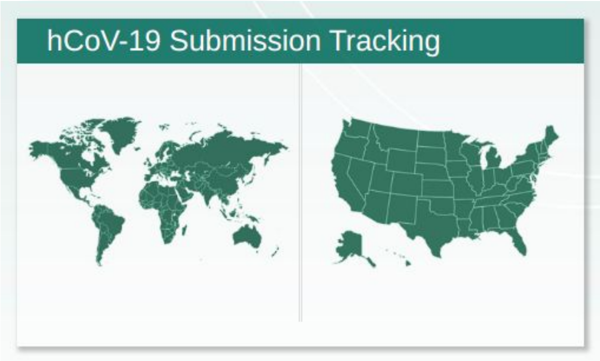
NCBI

DDBJ
DNA Data Bank of Japan

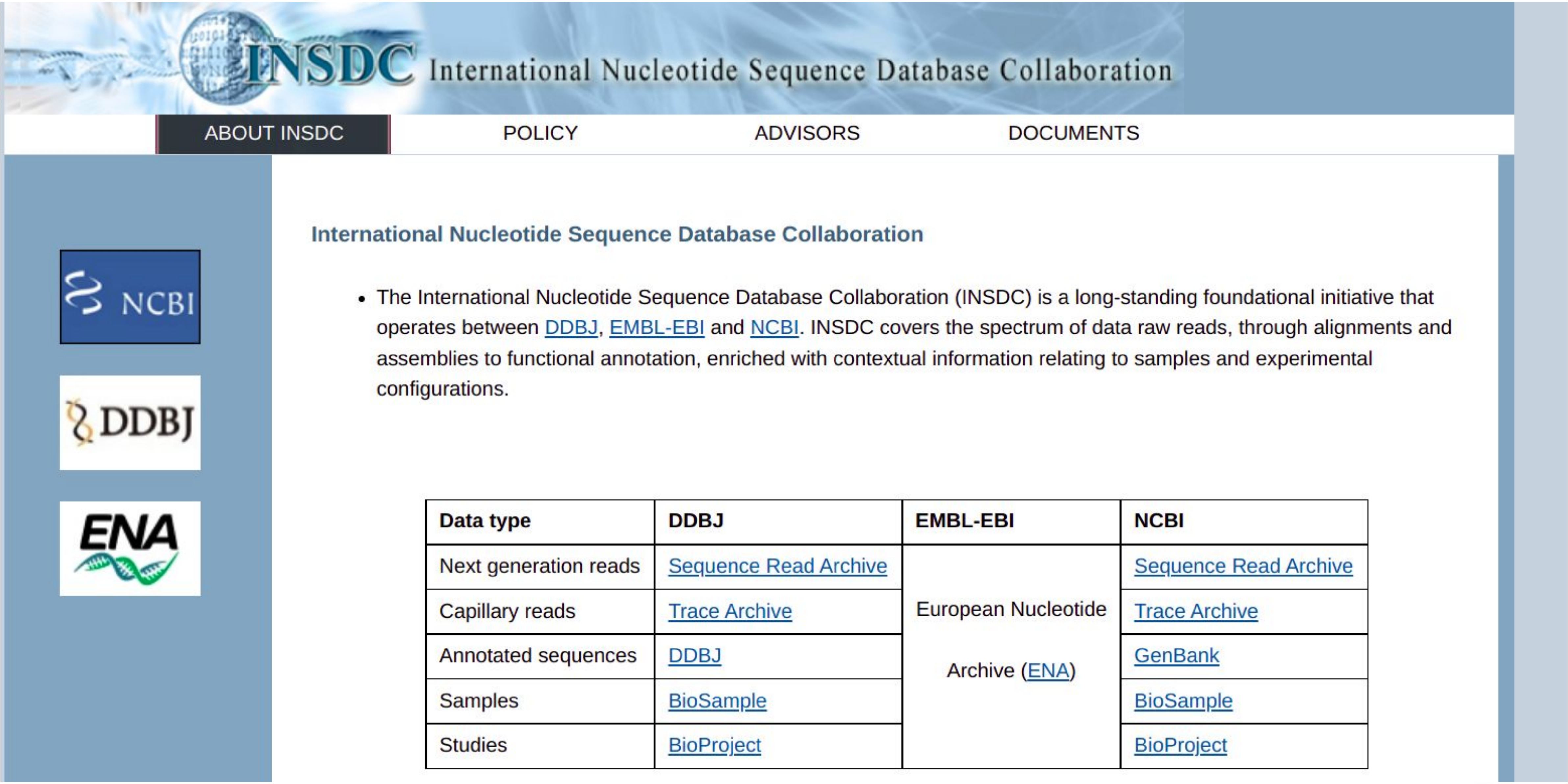
- Open access
- Handle different data types
    - raw sequencing data
- Embargo: can set a release date
- Multi-purpose: can link with other data
    - i.e from the host

GISAID

GISAID



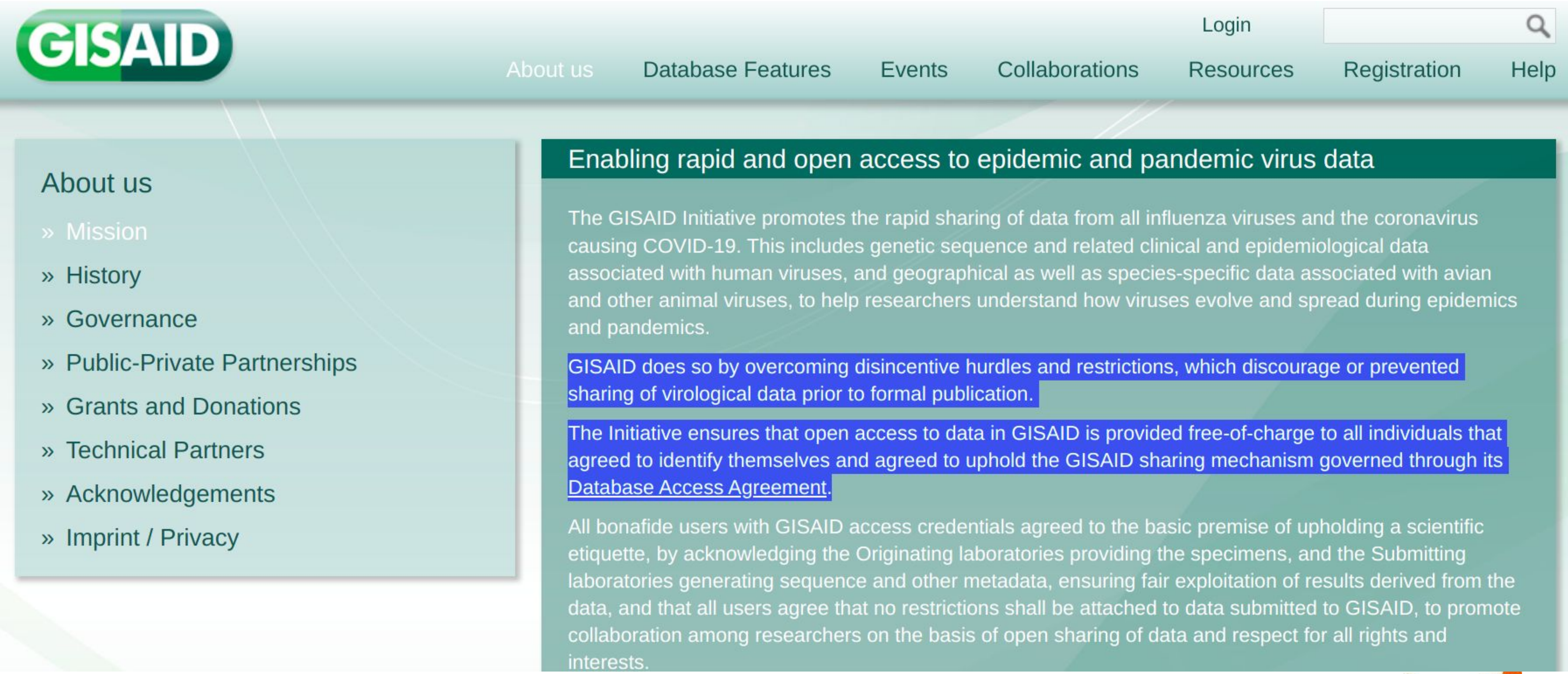
- Restricted access
- Only viral data
- Only consensus genomes
- No embargo


INSDC International Nucleotide Sequence Database Collaboration
ABOUT INSDC
POLICY
ADVISORS
DOCUMENTS
NCBI
DDBJ
ENA

International Nucleotide Sequence Database Collaboration

- The International Nucleotide Sequence Database Collaboration (INSDC) is a long-standing foundational initiative that operates between DDBJ, EMBL-EBI and NCBI. INSDC covers the spectrum of data raw reads, through alignments and assemblies to functional annotation, enriched with contextual information relating to samples and experimental configurations.

Data type	DDBJ	EMBL-EBI	NCBI
Next generation reads	Sequence Read Archive	European Nucleotide Archive (ENA)	Sequence Read Archive
Capillary reads	Trace Archive		Trace Archive
Annotated sequences	DDBJ		GenBank
Samples	BioSample		BioSample
Studies	BioProject		BioProject



Enabling rapid and open access to epidemic and pandemic virus data

The GISAID Initiative promotes the rapid sharing of data from all influenza viruses and the coronavirus causing COVID-19. This includes genetic sequence and related clinical and epidemiological data associated with human viruses, and geographical as well as species-specific data associated with avian and other animal viruses, to help researchers understand how viruses evolve and spread during epidemics and pandemics.

GISAID does so by overcoming disincentive hurdles and restrictions, which discourage or prevented sharing of virological data prior to formal publication.

The Initiative ensures that open access to data in GISAID is provided free-of-charge to all individuals that agreed to identify themselves and agreed to uphold the GISAID sharing mechanism governed through its Database Access Agreement.

All bonafide users with GISAID access credentials agreed to the basic premise of upholding a scientific etiquette, by acknowledging the Originating laboratories providing the specimens, and the Submitting laboratories generating sequence and other metadata, ensuring fair exploitation of results derived from the data, and that all users agree that no restrictions shall be attached to data submitted to GISAID, to promote collaboration among researchers on the basis of open sharing of data and respect for all rights and interests.

Where are genome data submitted?

INSDC

ENA

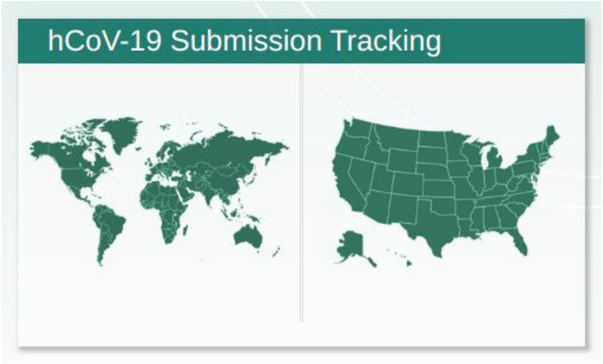
NCBI

DDBJ
DNA Data Bank of Japan

- ~ 4M viral sequences

GISAID

GISAID



hCoV-19 Submission Tracking

- ~ 8M viral sequences

Why?

elixir

Data flow, GDPR and issues

The following data/metadata are considered sensitive personal data in Italy*

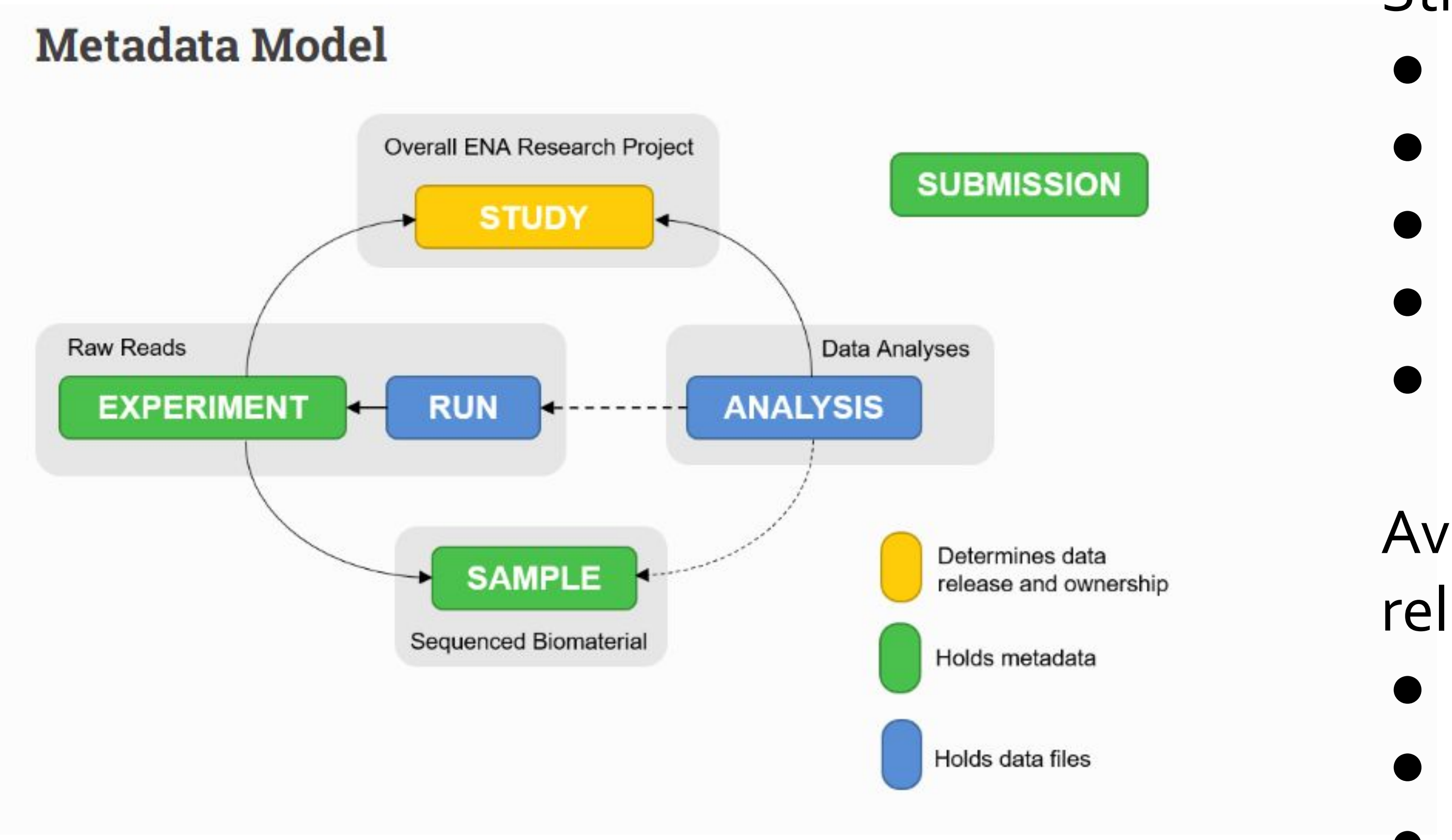
Date test taken
Place test taken
Age
Sex
Disease severity
Comorbidities

Collection date -> **seq date**
Place test taken -> **address seq center**
Age -> **only 65% of the samples**
Sex -> **only 78% of the samples**
Disease severity -> **12% of the samples**
Comorbidities -> **less than 1%**

*But not by all of the 20 administrative regions **
** and different DPOs provide different indications in the same regions

So controlled access seems a more viable option

ENA/INDSC: data model



Metadata model ENA: LINK

Structured, hierarchical

- Study
- Sample
- Experiment
- Run
- Submission

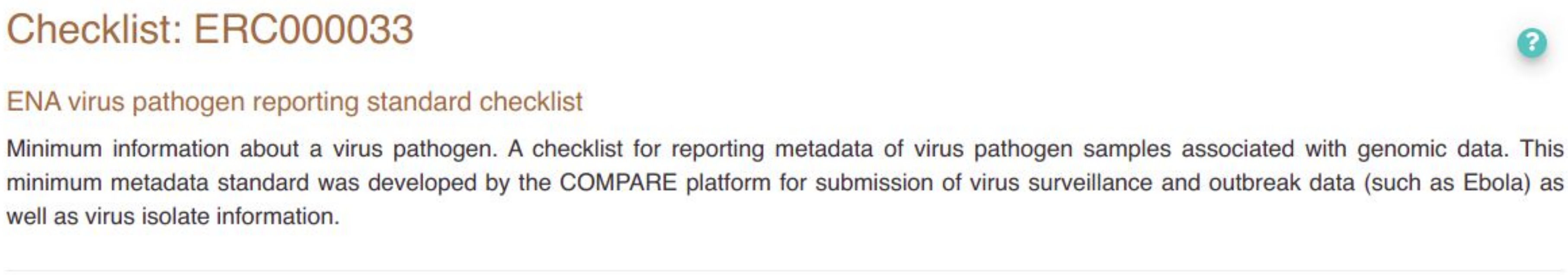
Average time submission to release:

- ~2 days
- can set release date (embargo)
- can link to external resources

ENA metadata model

- **Study**: groups together data submitted to the archive and controls its release date.
- **Sample**: contains information about the sequenced source material.
- **Experiment**: sequencing experiment, library and instrument details.
- **Run**: data files containing sequence reads
- **Submission**: contains submission actions to be performed by the archive. A submission can add more objects to the archive, update already submitted objects or make objects publicly available.

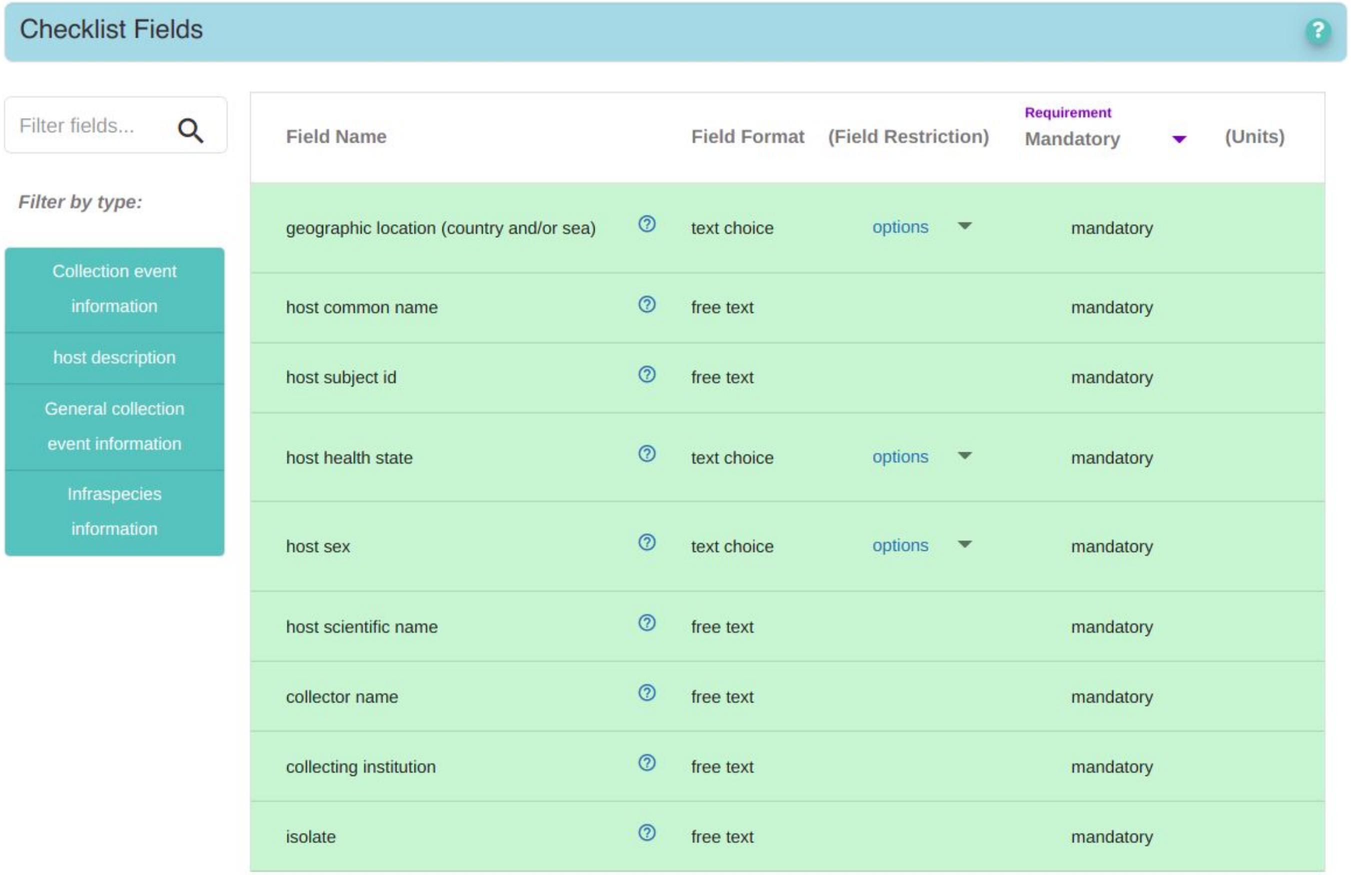
ENA metadata, samples (ERC000033)



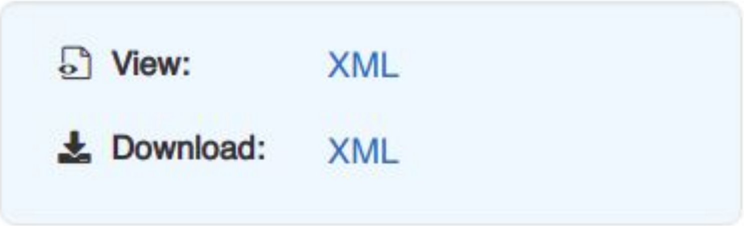
Checklist: ERC000033

ENA virus pathogen reporting standard checklist

Minimum information about a virus pathogen. A checklist for reporting metadata of virus pathogen samples associated with genomic data. This minimum metadata standard was developed by the COMPARE platform for submission of virus surveillance and outbreak data (such as Ebola) as well as virus isolate information.

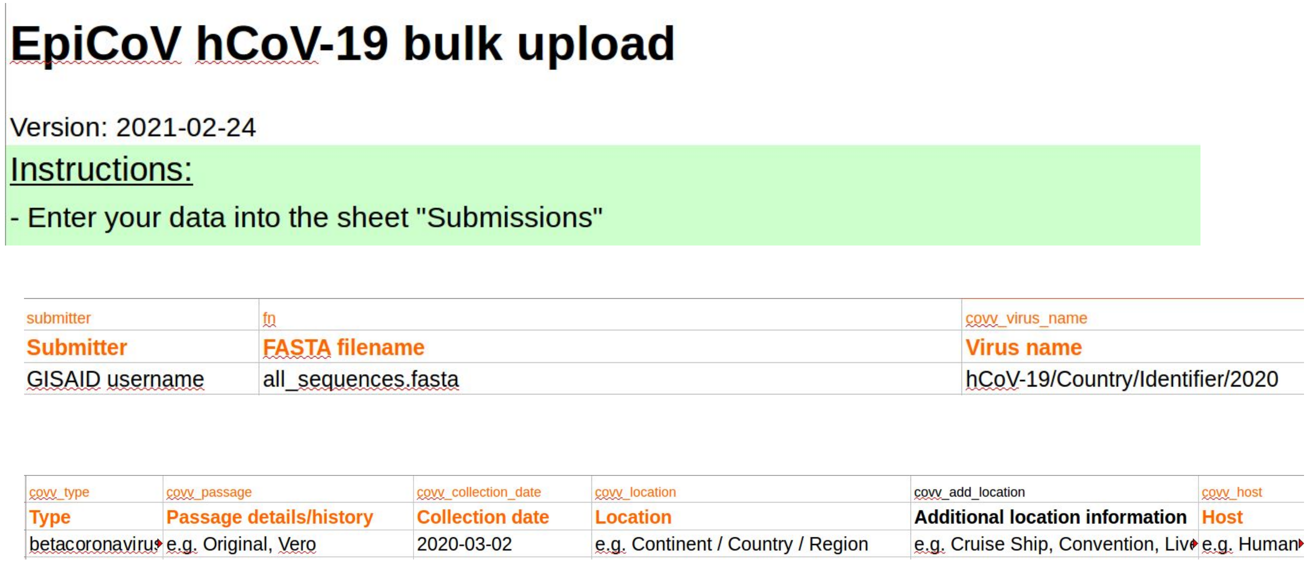


Field Name	Field Format	(Field Restriction)	Requirement Mandatory	(Units)
geographic location (country and/or sea)	text choice	options	mandatory	
host common name	free text		mandatory	
host subject id	free text		mandatory	
host health state	text choice	options	mandatory	
host sex	text choice	options	mandatory	
host scientific name	free text		mandatory	
collector name	free text		mandatory	
collecting institution	free text		mandatory	
isolate	free text		mandatory	



LINK

GISAID: data model



EpiCoV hCoV-19 bulk upload

Version: 2021-02-24

Instructions:

- Enter your data into the sheet "Submissions"

submitter	fn	covv_virus_name
Submitter	**FASTA filename**	**Virus name**
GISAID username	all_sequences.fasta	hCoV-19/Country/Identifier/2020

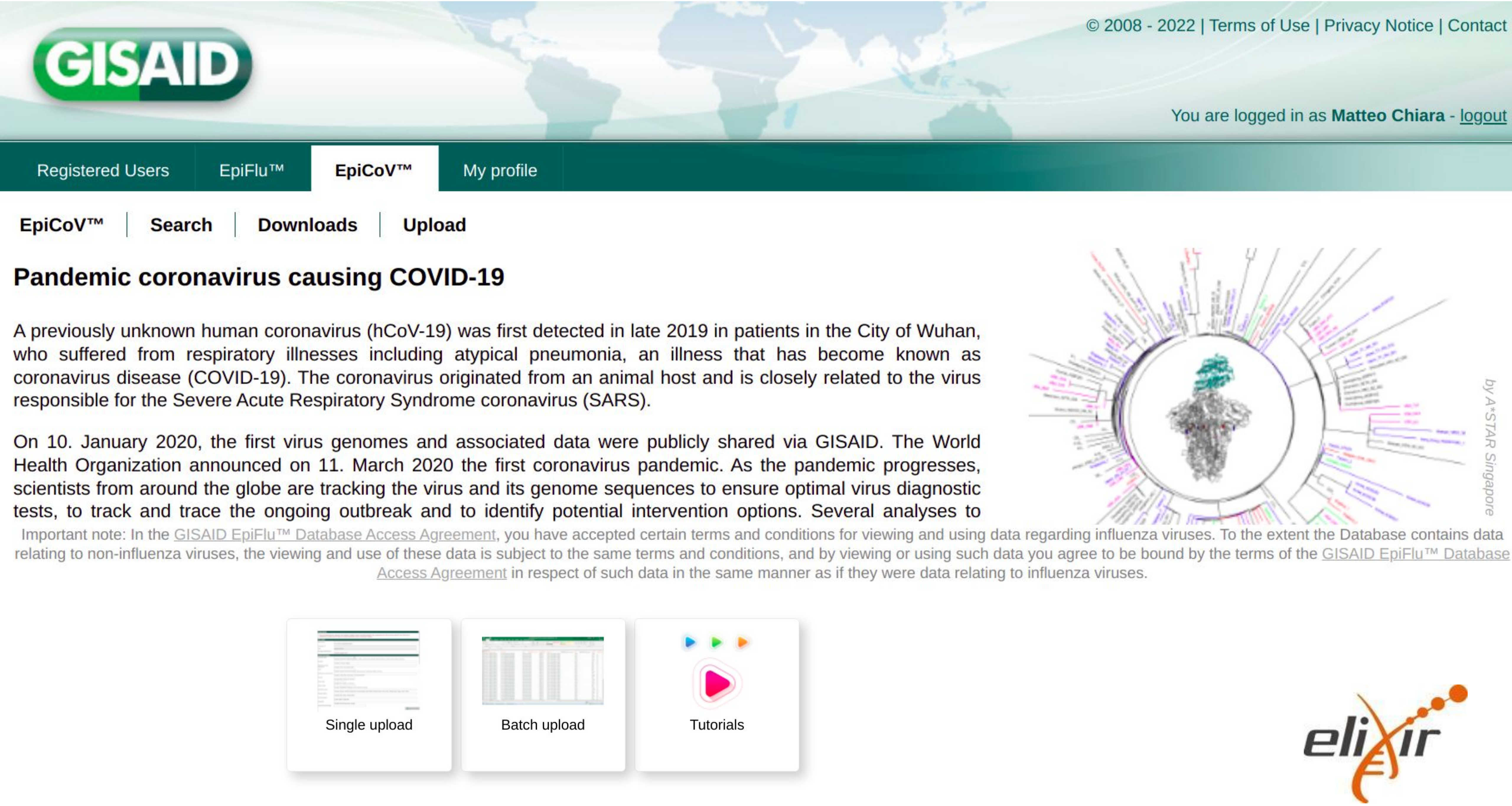
covv_type	covv_passage	covv_collection_date	covv_location	covv_add_location	covv_host
Type	**Passage details/history**	**Collection date**	**Location**	**Additional location information**	**Host**
betacoronaviru	e.g. Original, Vero	2020-03-02	e.g. Continent / Country / Region	e.g. Cruise Ship, Convention, Liv	e.g. Human

Bulk submission: large spreadsheet

- with some mandatory fields
  - vocabulary is limited, not controlled
- metadata are limited.
- No “ancillary” data

Time from submission to release

- ~1 dd
- release date can not be set
- can not (easily) link to external resources

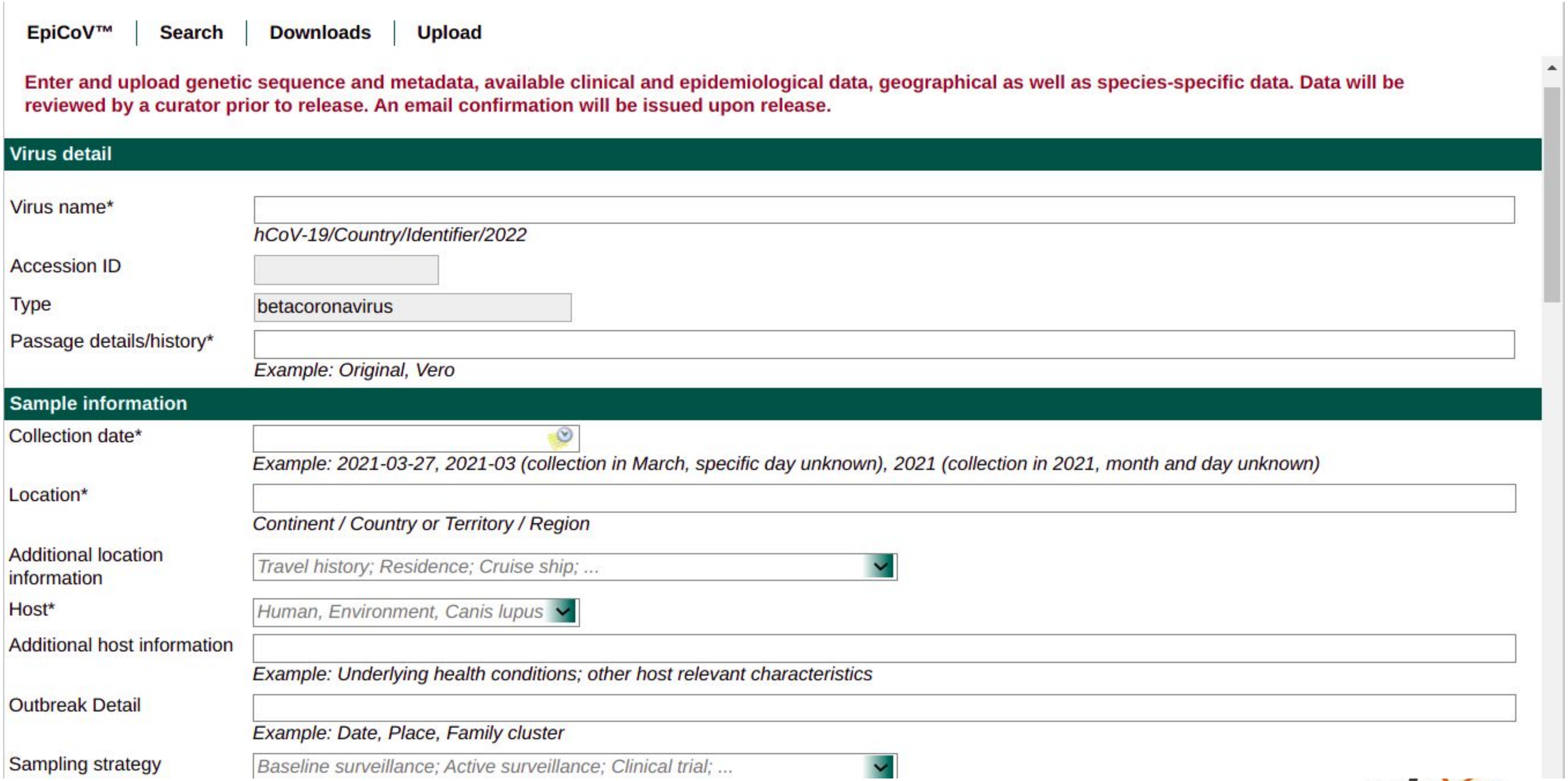

GISAID
© 2008 - 2022 | Terms of Use | Privacy Notice | Contact
You are logged in as Matteo Chiara - logout
Registered Users
EpiFlu™
EpiCoV™
My profile
EpiCoV™
Search
Downloads
Upload

Pandemic coronavirus causing COVID-19

A previously unknown human coronavirus (hCoV-19) was first detected in late 2019 in patients in the City of Wuhan, who suffered from respiratory illnesses including atypical pneumonia, an illness that has become known as coronavirus disease (COVID-19). The coronavirus originated from an animal host and is closely related to the virus responsible for the Severe Acute Respiratory Syndrome coronavirus (SARS).

On 10. January 2020, the first virus genomes and associated data were publicly shared via GISAID. The World Health Organization announced on 11. March 2020 the first coronavirus pandemic. As the pandemic progresses, scientists from around the globe are tracking the virus and its genome sequences to ensure optimal virus diagnostic tests, to track and trace the ongoing outbreak and to identify potential intervention options. Several analyses to

Important note: In the GISAID EpiFlu™ Database Access Agreement, you have accepted certain terms and conditions for viewing and using data regarding influenza viruses. To the extent the Database contains data relating to non-influenza viruses, the viewing and use of these data is subject to the same terms and conditions, and by viewing or using such data you agree to be bound by the terms of the GISAID EpiFlu™ Database Access Agreement in respect of such data in the same manner as if they were data relating to influenza viruses.



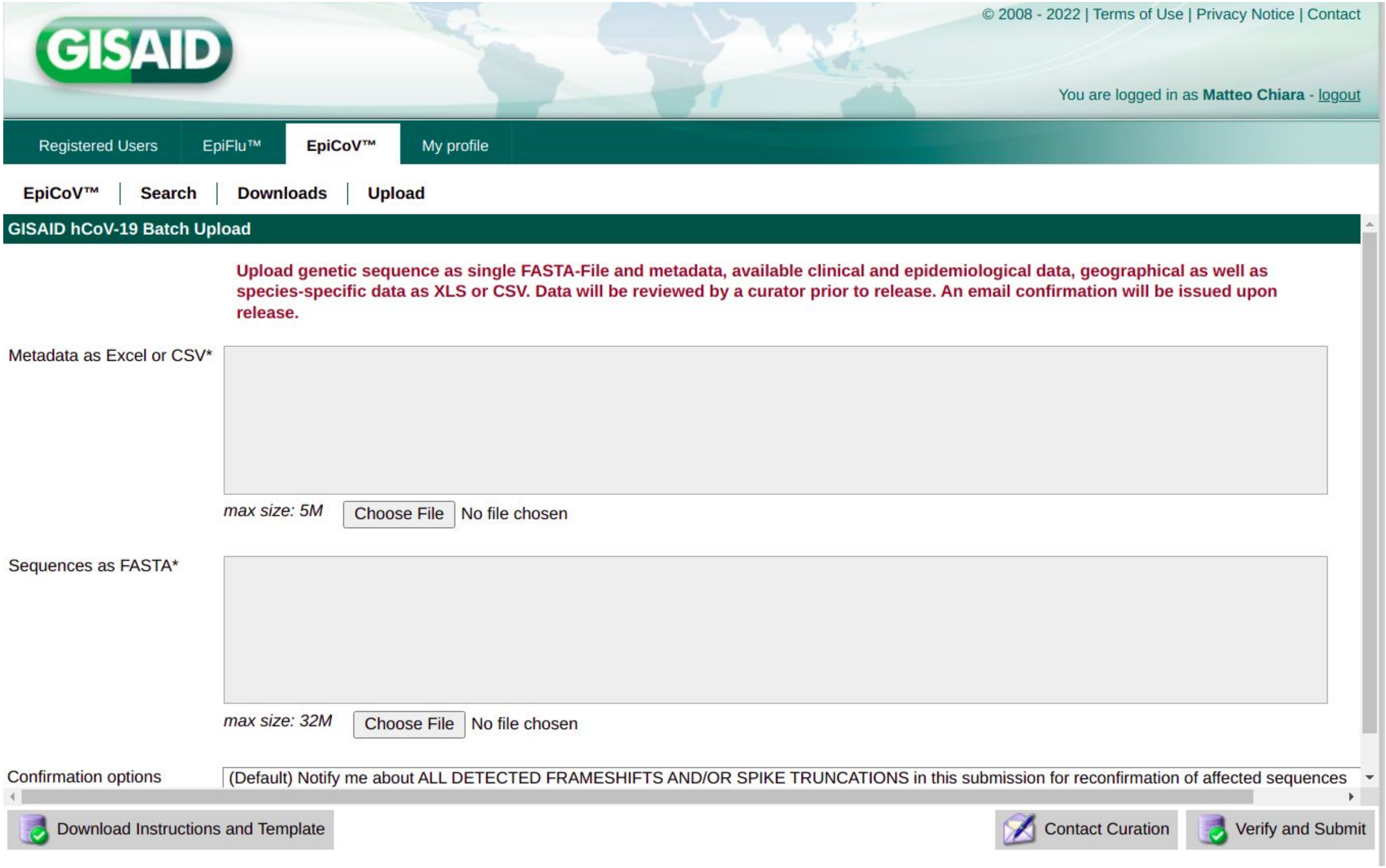
Enter and upload genetic sequence and metadata, available clinical and epidemiological data, geographical as well as species-specific data. Data will be reviewed by a curator prior to release. An email confirmation will be issued upon release.

Virus detail

Virus name*	*hCoV-19/Country/Identifier/2022*
Accession ID	
Type	betacoronavirus
Passage details/history*	*Example: Original, Vero*

Sample information

Collection date*	*Example: 2021-03-27, 2021-03 (collection in March, specific day unknown), 2021 (collection in 2021, month and day unknown)*
Location*	*Continent / Country or Territory / Region*
Additional location information	*Travel history; Residence; Cruise ship; ...*
Host*	*Human, Environment, Canis lupus*
Additional host information	*Example: Underlying health conditions; other host relevant characteristics*
Outbreak Detail	*Example: Date, Place, Family cluster*
Sampling strategy	*Baseline surveillance; Active surveillance; Clinical trial; ...*


GISAID
© 2008 - 2022 | Terms of Use | Privacy Notice | Contact
You are logged in as Matteo Chiara - logout
Registered Users
EpiFlu™
EpiCoV™
My profile
EpiCoV™
Search
Downloads
Upload
Choose File
Choose File
Download Instructions and Template
Contact Curation
Verify and Submit

GISAID hCoV-19 Batch Upload

Upload genetic sequence as single FASTA-File and metadata, available clinical and epidemiological data, geographical as well as species-specific data as XLS or CSV. Data will be reviewed by a curator prior to release. An email confirmation will be issued upon release.

Metadata as Excel or CSV*

max size: 5M No file chosen

Sequences as FASTA*

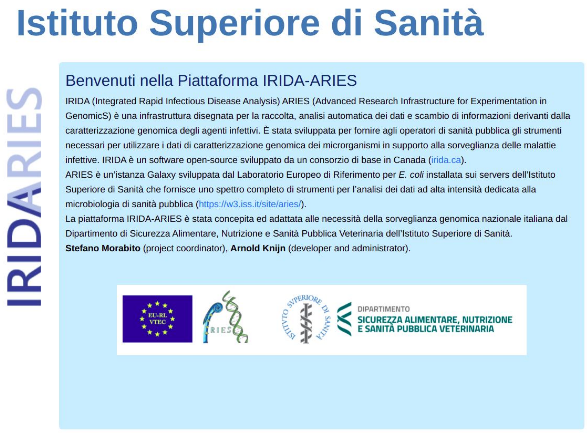
max size: 32M No file chosen

Confirmation options: (Default) Notify me about ALL DETECTED FRAMESHIFTS AND/OR SPIKE TRUNCATIONS in this submission for reconfirmation of affected sequences

Genomic surveillance in Italy

~100 sequencing/testing centers (4.8 per region)

I.Z.S. - Istituti zooprofilattici sperimentali



Istituto Superiore di Sanità

Benvenuti nella Piattaforma IRIDA-ARIES

IRIDA (Integrated Rapid Infectious Disease Analysis) ARIES (Advanced Research Infrastructure for Experimentation in GenomicS) è una infrastruttura disegnata per la raccolta, analisi automatica dei dati e scambio di informazioni derivanti dalla caratterizzazione genomica degli agenti infettivi. È stata sviluppata per fornire agli operatori di sanità pubblica gli strumenti necessari per utilizzare i dati di caratterizzazione genomica dei microrganismi in supporto alla sorveglianza delle malattie infettive. IRIDA è un software open-source sviluppato da un consorzio di base in Canada (irida.ca).

ARIES è un'istanza Galaxy sviluppata dal Laboratorio Europeo di Riferimento per *E. coli* installata sui servers dell'Istituto Superiore di Sanità che fornisce uno spettro completo di strumenti per l'analisi dei dati ad alta intensità dedicata alla microbiologia di sanità pubblica (https://w3.iss.it/site/aries/).

La piattaforma IRIDA-ARIES è stata concepita ed adattata alle necessità della sorveglianza genomica nazionale italiana dal Dipartimento di Sicurezza Alimentare, Nutrizione e Sanità Pubblica Veterinaria dell'Istituto Superiore di Sanità.

Stefano Morabito (project coordinator), **Arnold Knijn** (developer and administrator).

EU-RL VTEC
DIPARTIMENTO
SICUREZZA ALIMENTARE, NUTRIZIONE E SANITÀ PUBBLICA VETERINARIA

From April 2021

GISAID

91.353 genome sequences

(94% through I-CoGen)

WORK IN PROGRESS

ENA European Nucleotide Archive

only 344 sequences

elixir

Where should we put our data?

INSDC

ENA

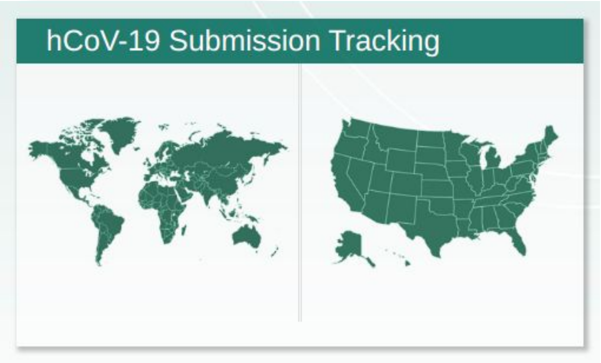
DDBJ
DNA Data Bank of Japan

NCBI

- More structured
  - More effort
- Different data types
  - (quality check/reanalyses)
- Link with "host data"

GISAID

GISAID



hCoV-19 Submission Tracking

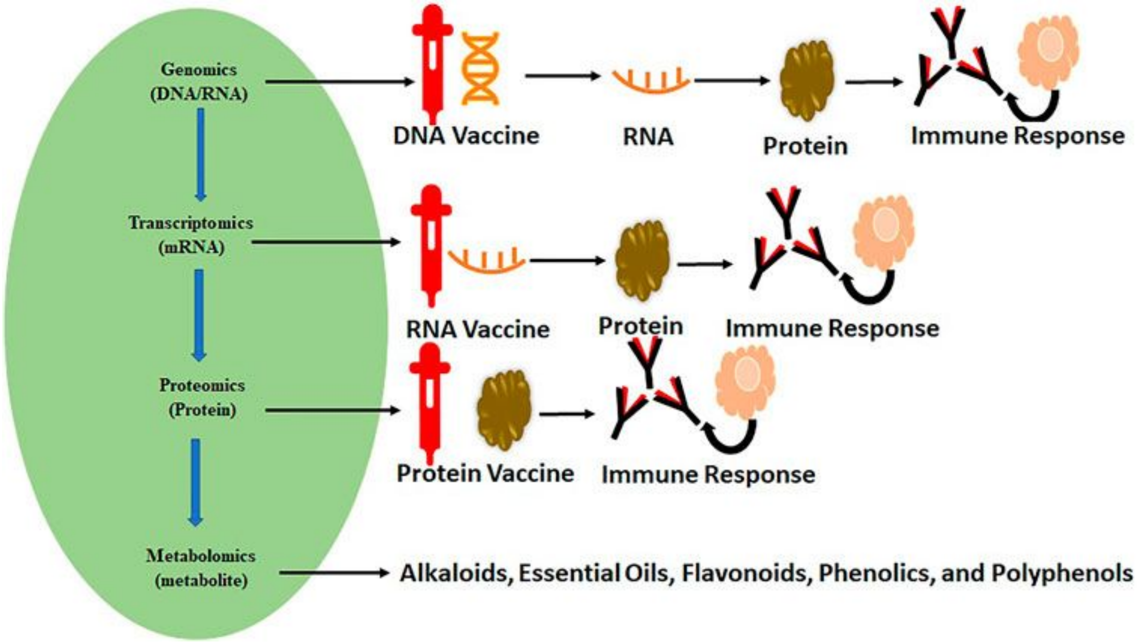
- Easier, quicker
- Only genome assemblies
- Reference db "worldwide"
- Difficult to link with external resources

What are we (scientific community) giving up?

- Data integration:
  - genome sequences with host data:
  - Serological data
  - Transcriptomic data
  - Host genome

- Data reanalysis
  - co-infection
  - within-host evolution
  - benchmarks for comparing tools

- Data re-use



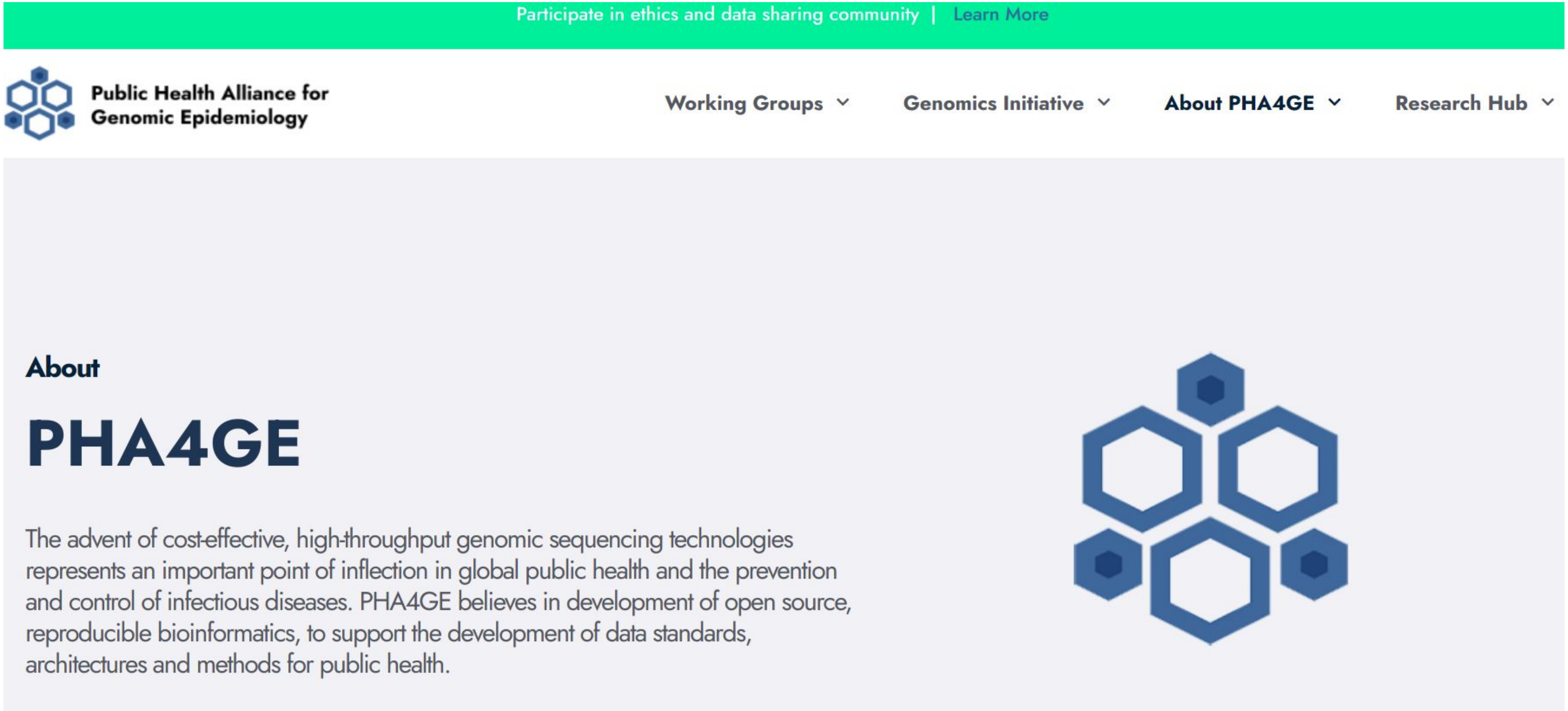
HOW tos

How to submit to ENA: LINK (please contact info@covidataportal.it in case of issues)

How to submit to GISAID: LINK + a couple of videos in the "restricted access" area of the db

Can I migrate data from GISAID to ENA: likely so. Please see: Roncoroni et al. and LINK

more HOW tos



Participate in ethics and data sharing community | Learn More

About

PHA4GE

The advent of cost-effective, high-throughput genomic sequencing technologies represents an important point of inflection in global public health and the prevention and control of infectious diseases. PHA4GE believes in development of open source, reproducible bioinformatics, to support the development of data standards, architectures and methods for public health.

https://pha4ge.org/

Conclusions

- Handling SARS-CoV-2 data might be a complex task
- There is a hell of work behind one genome sequence
- Data stewards needed to correctly handle all this data…
  - But not just the data itsef:
    - Bioinoformatics
    - Lab protocols
    - Sequencid data

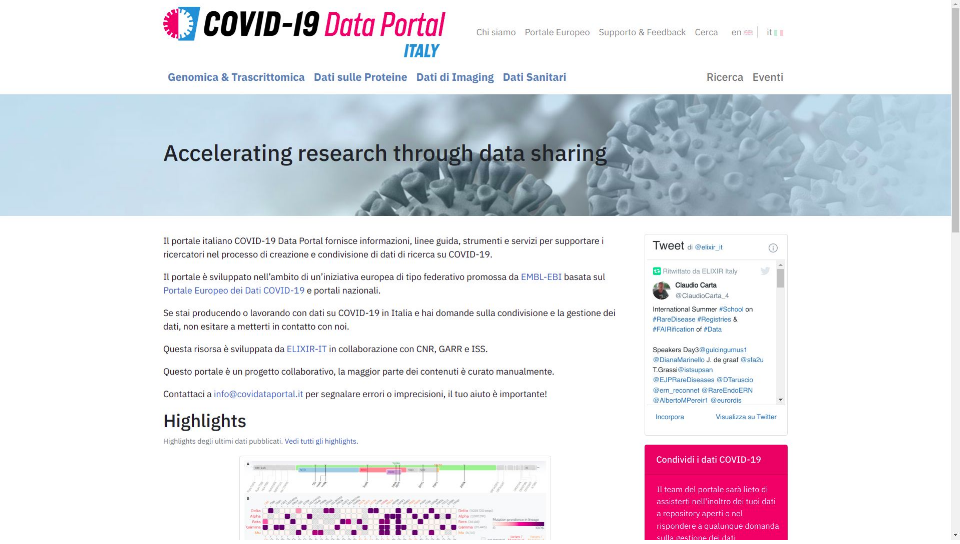
- At the moment, GISAID the resource used by most does not comply completely with open and FAIR
  - consider INSDC where possible

Open questions and future perspectives

- Currently the majority of SARS-CoV-2 genomes from Italian institutions is at GISAID
  - restricted access
  - **only genomic assemblies no raw data**

- Working with ISS to
  - migrate to INDSC databases (ENA)
  - deposit also raw data if available
  - tools already in place but. **Ethical/legal (GDPR) constraints are slowing us down**

- HelpDesk:
  - we help people migrate seqs from GISAID to ENA



Accelerating research through data sharing

Il portale italiano COVID-19 Data Portal fornisce informazioni, linee guida, strumenti e servizi per supportare i ricercatori nel processo di creazione e condivisione di dati di ricerca su COVID-19.

Il portale è sviluppato nell'ambito di un'iniziativa europea di tipo federativo promossa da EMBL-EBI basata sul Portale Europeo dei Dati COVID-19 e portali nazionali.

Se stai producendo o lavorando con dati su COVID-19 in Italia e hai domande sulla condivisione e la gestione dei dati, non esitare a metterti in contatto con noi.

Questa risorsa è sviluppata da ELIXIR-IT in collaborazione con CNR, GARR e ISS.

Questo portale è un progetto collaborativo, la maggior parte dei contenuti è curato manualmente.

Contattaci a info@covidataportal.it per segnalare errori o imprecisioni, il tuo aiuto è importante!

Highlights

Highlights degli ultimi dati pubblicati. Vedi tutti gli highlights.

Tweet di @elixir_it

Ritwittato da ELIXIR Italy

Claudio Carta
@ClaudioCarta_4

International Summer #School on #RareDisease #Registries & #FAIRification of #Data

Speakers Day3@gulcingumus1 @DianaMarinello J. de graaf @sfa2u T.Grassi@istsupsan @EJPRareDiseases @DTaruscio @ern_reconnet @RareEndoERN @AlbertoMPereir1 @eurordis

Condividi i dati COVID-19

Il team del portale sarà lieto di assisterti nell'inoltro dei tuoi dati a repository aperti o nel rispondere a qualunque domanda sulla gestione dei dati

What about other types of data

- TBH, in Italy (or Europe) viral genomes is still the **<hot topic>**

- Host genome sequences -> **see B1MG**
  - Beacon, Federated EGA
  - GDPR!

- Imaging/Patients data -> **see 1+MG/B1MG**
  - see above. Ontologies

- Serological data -> **converge+ data portal**
  - ongoing discussion
  - help wanted!

Thanks!

@elixir_ita

www.elixir-europe.org

CNR-IBIOM

CNR-ITB

Università di Milano

Università di Padova

Università di Roma Tor Vergata

Università di Torino

Università della Tuscia

Istituto Superiore di Sanità

Consortium GARR

Open databases and FAIR standards for SARS-CoV-2: the quick "tutorial"

Matteo Chiara

www.elixir-europe.org

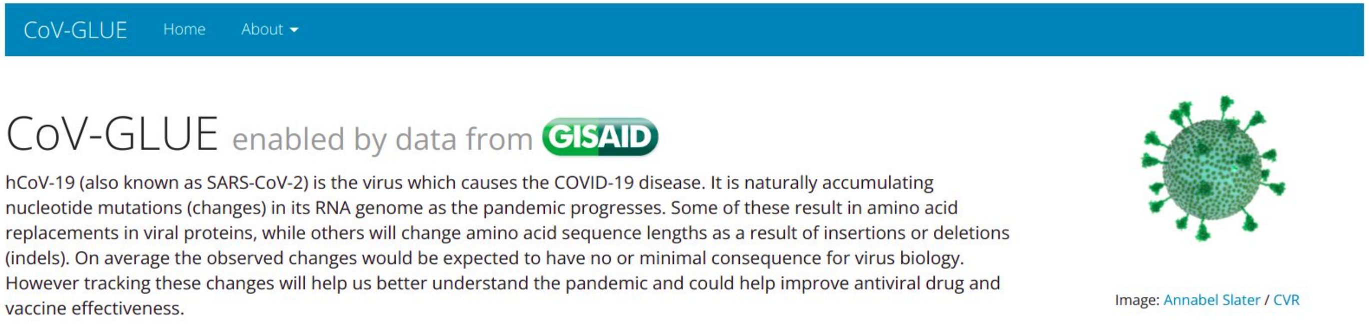
How can we (double) check data quality? SARS-CoV-2 use case

- We get one or more genome sequences

- We want to check/know if they might have issues

- Can we use tools/methods to check (without being hardcore bioinformaticians)?

MOSTLY SO



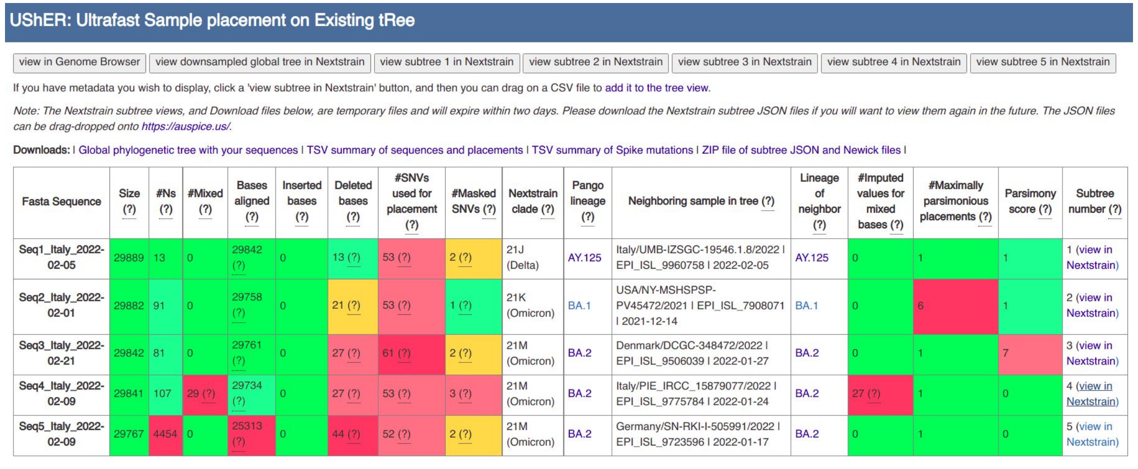
CoV-GLUE enabled by data from GISAID

hCoV-19 (also known as SARS-CoV-2) is the virus which causes the COVID-19 disease. It is naturally accumulating nucleotide mutations (changes) in its RNA genome as the pandemic progresses. Some of these result in amino acid replacements in viral proteins, while others will change amino acid sequence lengths as a result of insertions or deletions (indels). On average the observed changes would be expected to have no or minimal consequence for virus biology. However tracking these changes will help us better understand the pandemic and could help improve antiviral drug and vaccine effectiveness.

Image: Annabel Slater / CVR

Amino acid variation database

The dataset of amino acid replacements, insertions and deletions which have been observed in GISAID hCoV-19/SARS-CoV-2 sequences sampled from the pandemic is available at Cov-GLUE-Viz



UShER: Ultrafast Sample placement on Existing tRee

view in Genome Browser
view downsampled global tree in Nextstrain
view subtree 1 in Nextstrain
view subtree 2 in Nextstrain
view subtree 3 in Nextstrain
view subtree 4 in Nextstrain
view subtree 5 in Nextstrain

If you have metadata you wish to display, click a 'view subtree in Nextstrain' button, and then you can drag on a CSV file to add it to the tree view.

Note: The Nextstrain subtree views, and Download files below, are temporary files and will expire within two days. Please download the Nextstrain subtree JSON files if you will want to view them again in the future. The JSON files can be drag-dropped onto https://auspice.us/.

Downloads: | Global phylogenetic tree with your sequences | TSV summary of sequences and placements | TSV summary of Spike mutations | ZIP file of subtree JSON and Newick files |

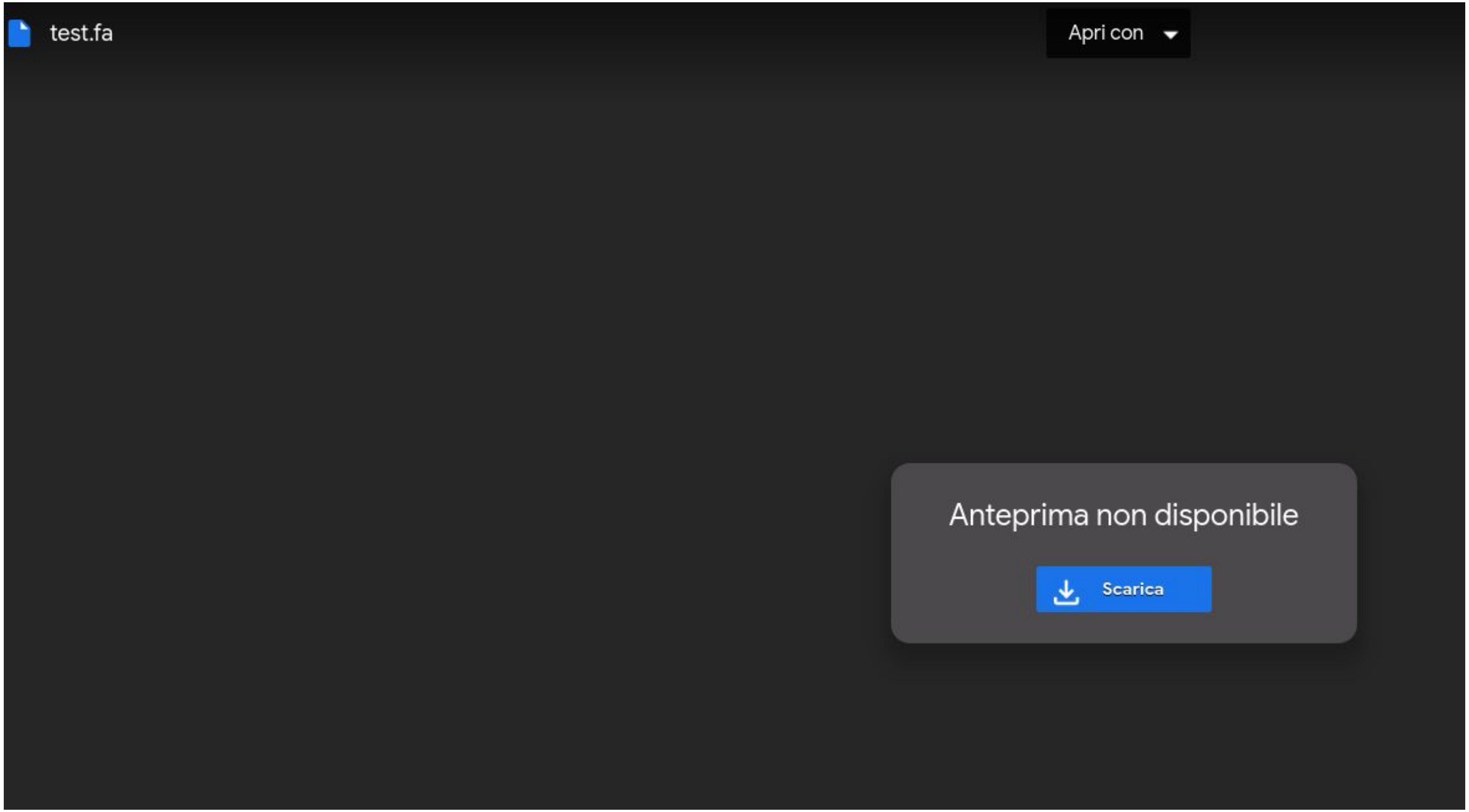
Fasta Sequence	Size (?)	#Ns (?)	#Mixed (?)	Bases aligned (?)	Inserted bases (?)	Deleted bases (?)	#SNVs used for placement (?)	#Masked SNVs (?)	Nextstrain clade (?)	Pango lineage (?)	Neighboring sample in tree (?)	Lineage of neighbor (?)	#Imputed values for mixed bases (?)	#Maximally parsimonious placements (?)	Parsimony score (?)	Subtree number (?)
Seq1_Italy_2022-02-05	29889	13	0	29842 (?)	0	13 (?)	53 (?)	2 (?)	21J (Delta)	AY.125	Italy/UMB-IZSGC-19546.1.8/2022 I EPI_ISL_9960758 I 2022-02-05	AY.125	0	1	1	1 (view in Nextstrain)
Seq2_Italy_2022-02-01	29882	91	0	29758 (?)	0	21 (?)	53 (?)	1 (?)	21K (Omicron)	BA.1	USA/NY-MSHSPSP-PV45472/2021 I EPI_ISL_7908071 I 2021-12-14	BA.1	0	6	1	2 (view in Nextstrain)
Seq3_Italy_2022-02-21	29842	81	0	29761 (?)	0	27 (?)	61 (?)	2 (?)	21M (Omicron)	BA.2	Denmark/DCGC-348472/2022 I EPI_ISL_9506039 I 2022-01-27	BA.2	0	1	7	3 (view in Nextstrain)
Seq4_Italy_2022-02-09	29841	107	29 (?)	29734 (?)	0	27 (?)	53 (?)	3 (?)	21M (Omicron)	BA.2	Italy/PIE_IRCC_15879077/2022 I EPI_ISL_9775784 I 2022-01-24	BA.2	27 (?)	1	0	4 (view in Nextstrain)
Seq5_Italy_2022-02-09	29767	4454	0	25313 (?)	0	44 (?)	52 (?)	2 (?)	21M (Omicron)	BA.2	Germany/SN-RKI-I-505991/2022 I EPI_ISL_9723596 I 2022-01-17	BA.2	0	1	0	5 (view in Nextstrain)

The data

- 5 randomly picked and **anonymised** genome sequences
- In fasta format: see here

- To check if sequences have issues we can
  - see if they have strange “bits” (Ns, sequences that resemble sequencing primers, an excess of “genetic variants”)
  - see if they are similar to other known sequences (SARS-CoV-2 is not “fast evolving”)
  - see if they “match” known variants and if they got the right mutations

- In fasta format: see here



test.fa

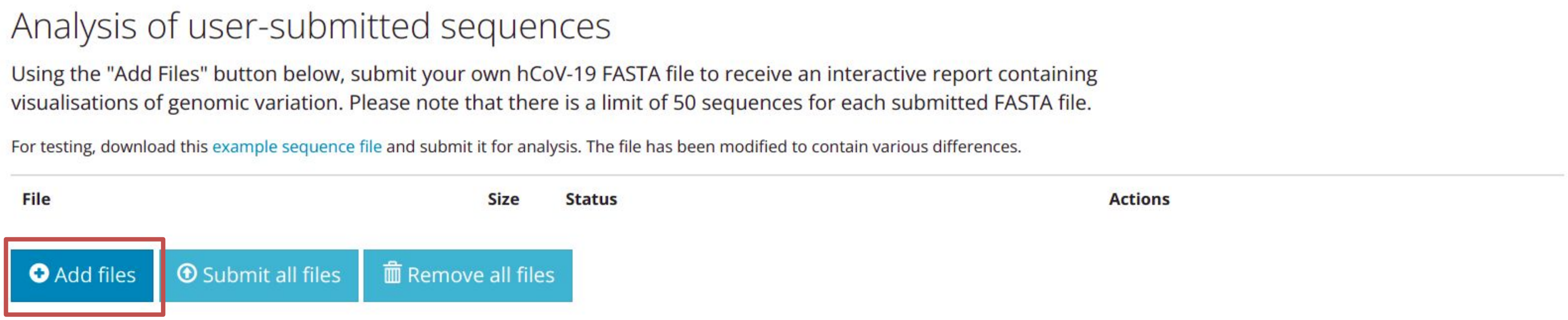
Anteprima non disponibile

CoV-GLUE: quality check, step#1

- Quick and highly curated “web service” for getting a quality check report of SARS-CoV-2 assemblies
- CoV-GLUE web application http://cov-glue.cvr.gla.ac.uk
- Detailed report of
  - completeness of the genome sequence
  - mutations (complete list)
  - impact (of mutations) on sequencing and diagnostics

By Singer et al, University of Glasgow. See here for the preprint

P.S. □ = click on from here onward



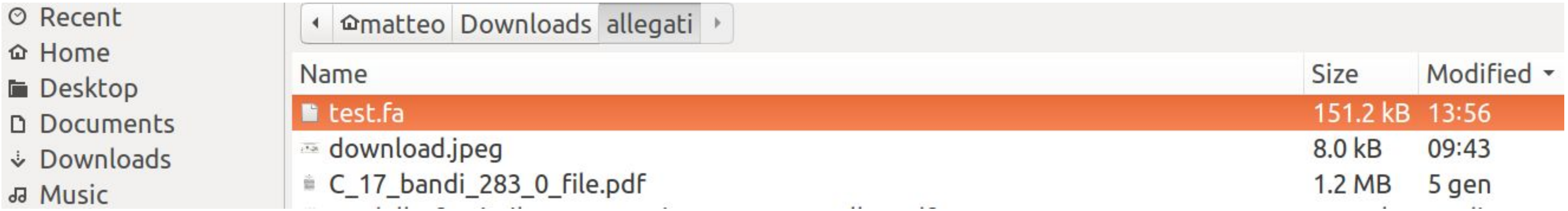
Analysis of user-submitted sequences

Using the "Add Files" button below, submit your own hCoV-19 FASTA file to receive an interactive report containing visualisations of genomic variation. Please note that there is a limit of 50 sequences for each submitted FASTA file.

For testing, download this example sequence file and submit it for analysis. The file has been modified to contain various differences.

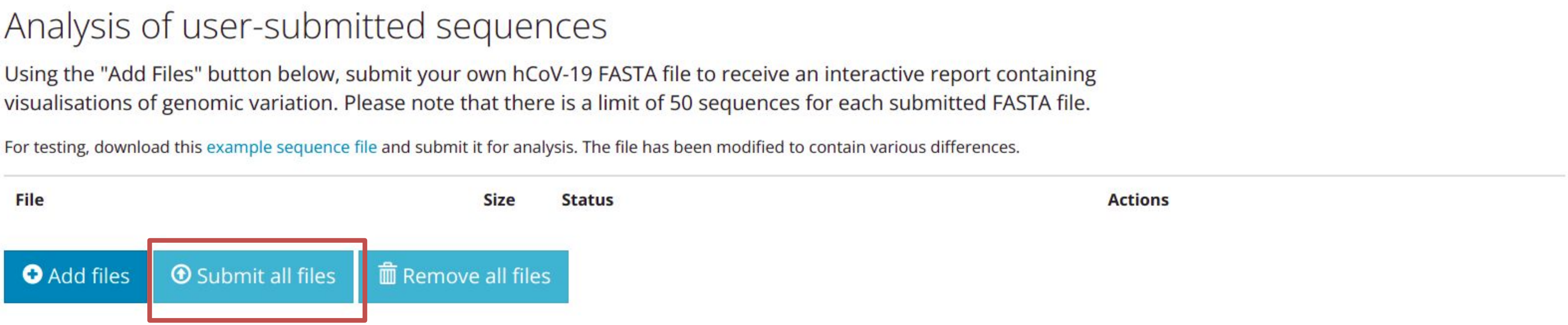
File	Size	Status	Actions

1



Name	Size	Modified
test.fa	151.2 kB	13:56
download.jpeg	8.0 kB	09:43
C_17_bandi_283_0_file.pdf	1.2 MB	5 gen

2



Analysis of user-submitted sequences

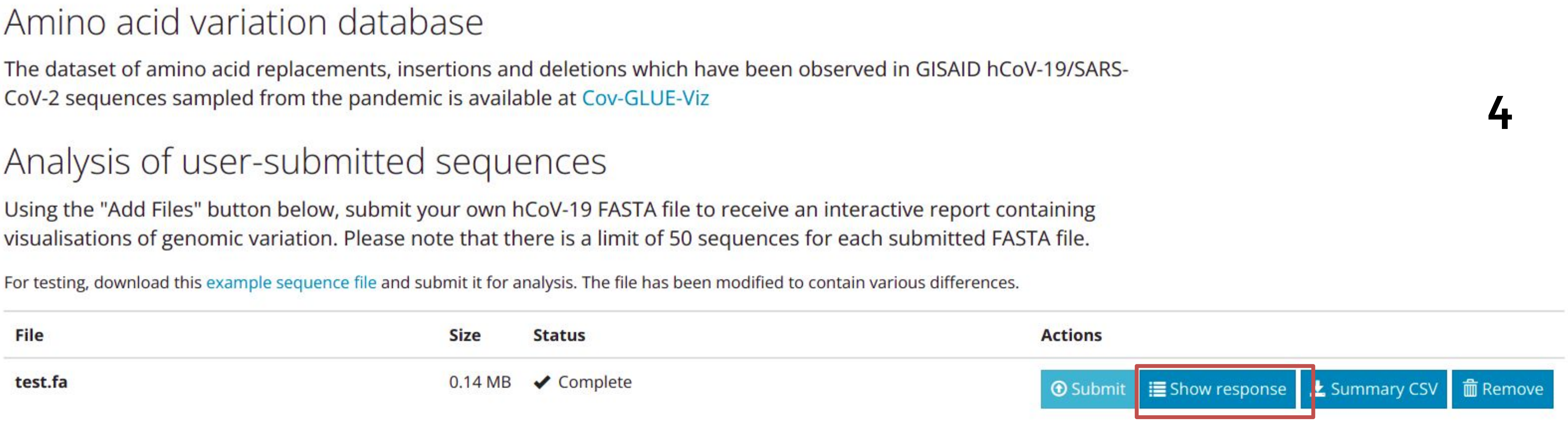
Using the "Add Files" button below, submit your own hCoV-19 FASTA file to receive an interactive report containing visualisations of genomic variation. Please note that there is a limit of 50 sequences for each submitted FASTA file.

For testing, download this example sequence file and submit it for analysis. The file has been modified to contain various differences.

File	Size	Status	Actions

3

4



Amino acid variation database

The dataset of amino acid replacements, insertions and deletions which have been observed in GISAID hCoV-19/SARS-CoV-2 sequences sampled from the pandemic is available at Cov-GLUE-Viz

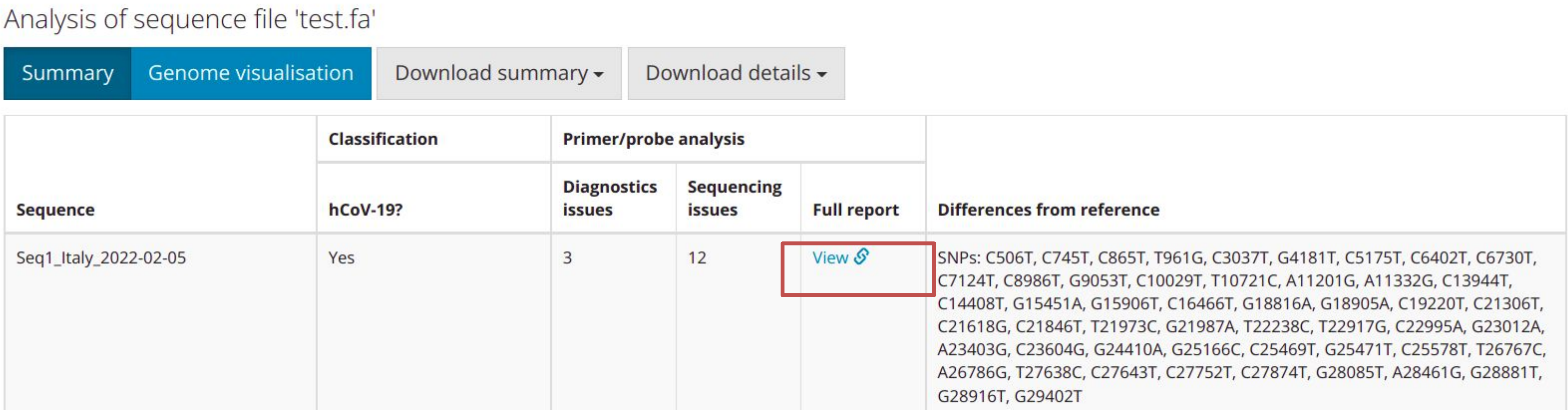
Analysis of user-submitted sequences

Using the "Add Files" button below, submit your own hCoV-19 FASTA file to receive an interactive report containing visualisations of genomic variation. Please note that there is a limit of 50 sequences for each submitted FASTA file.

For testing, download this example sequence file and submit it for analysis. The file has been modified to contain various differences.

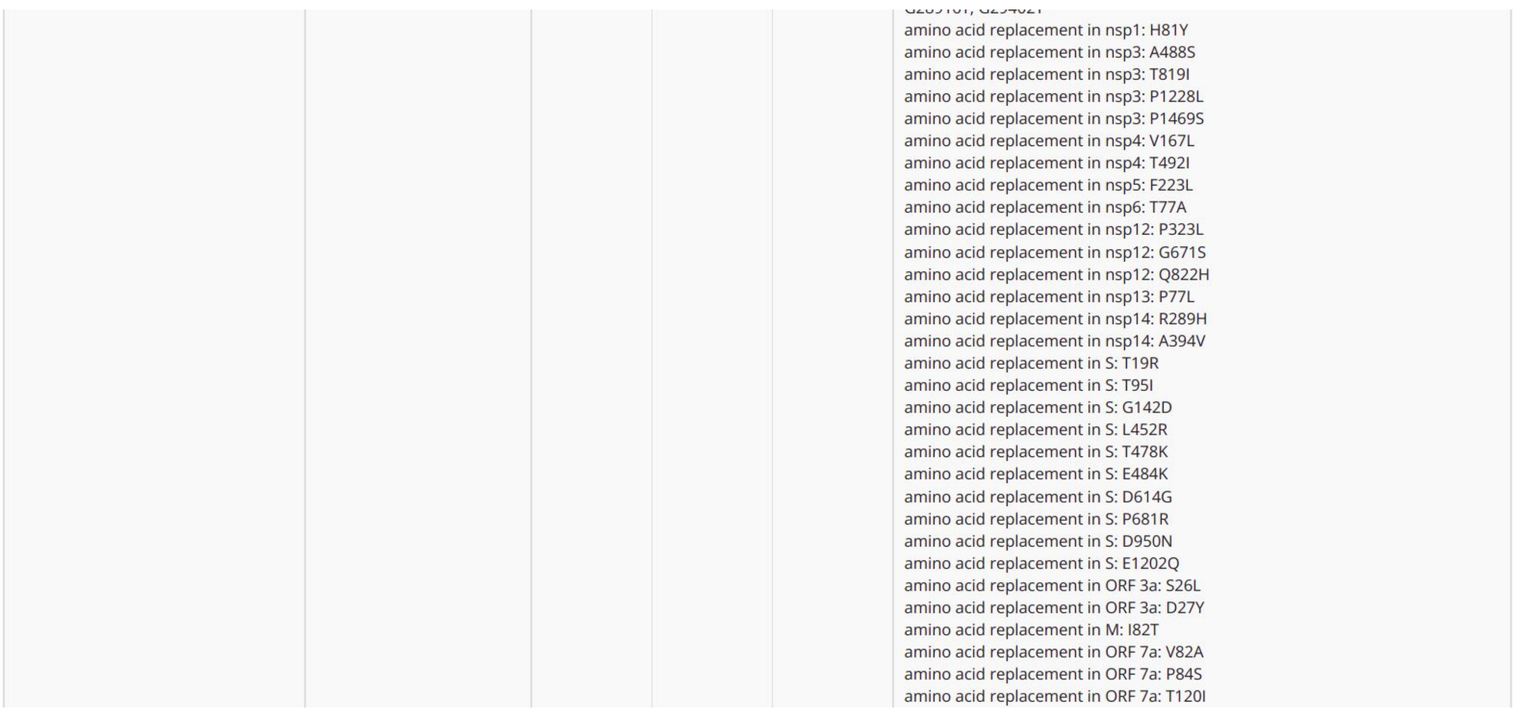
File	Size	Status	Actions
test.fa	0.14 MB	✓ Complete	Submit Show response Summary CSV Remove

In 1/2 minutes …

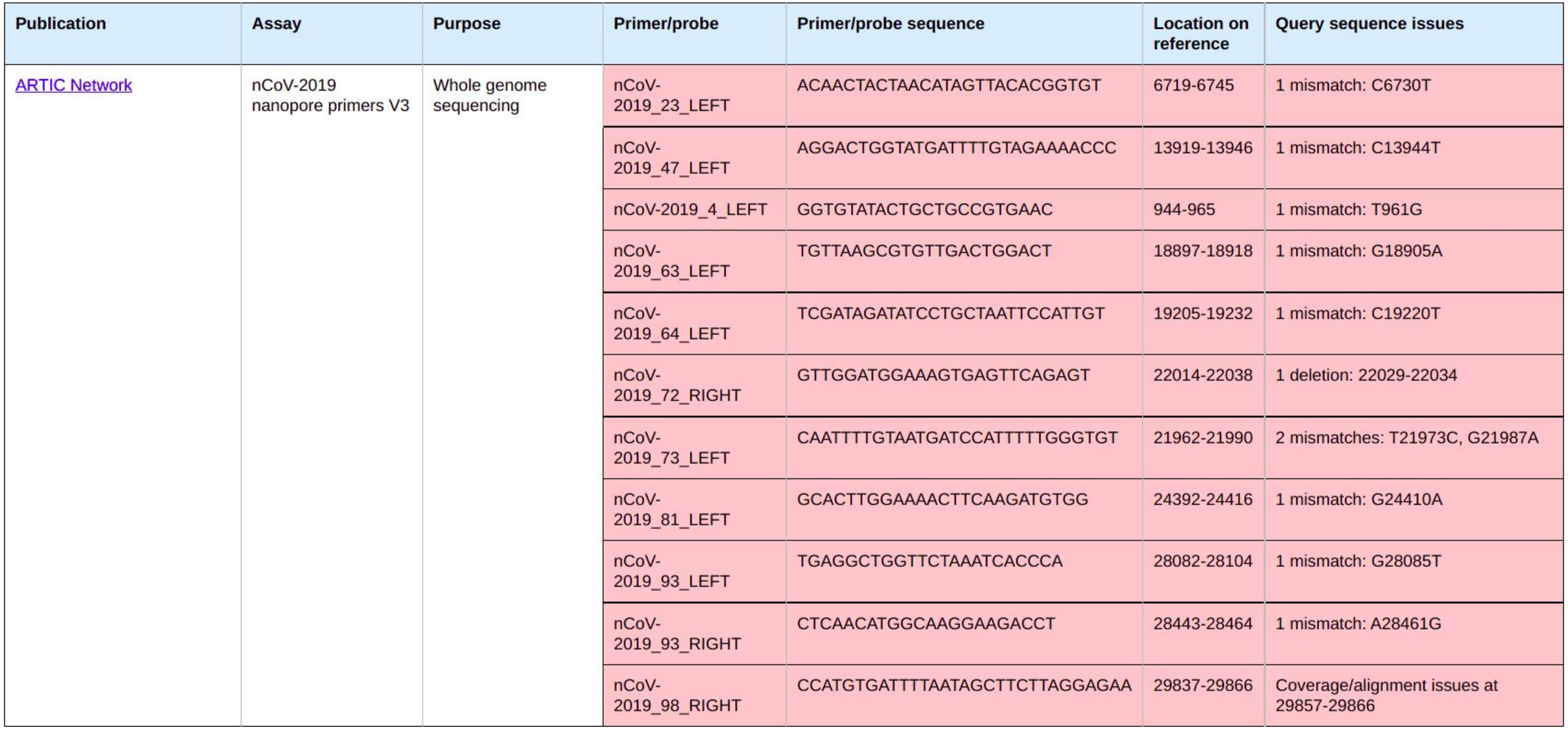


Analysis of sequence file 'test.fa'

Sequence	Classification	Primer/probe analysis			Differences from reference
	hCoV-19?	Diagnostics issues	Sequencing issues	Full report	
Seq1_Italy_2022-02-05	Yes	3	12	View	SNPs: C506T, C745T, C865T, T961G, C3037T, G4181T, C5175T, C6402T, C6730T, C7124T, C8986T, G9053T, C10029T, T10721C, A11201G, A11332G, C13944T, C14408T, G15451A, G15906T, C16466T, G18816A, G18905A, C19220T, C21306T, C21618G, C21846T, T21973C, G21987A, T22238C, T22917G, C22995A, G23012A, A23403G, C23604G, G24410A, G25166C, C25469T, G25471T, C25578T, T26767C, A26786G, T27638C, C27643T, C27752T, C27874T, G28085T, A28461G, G28881T, G28916T, G29402T

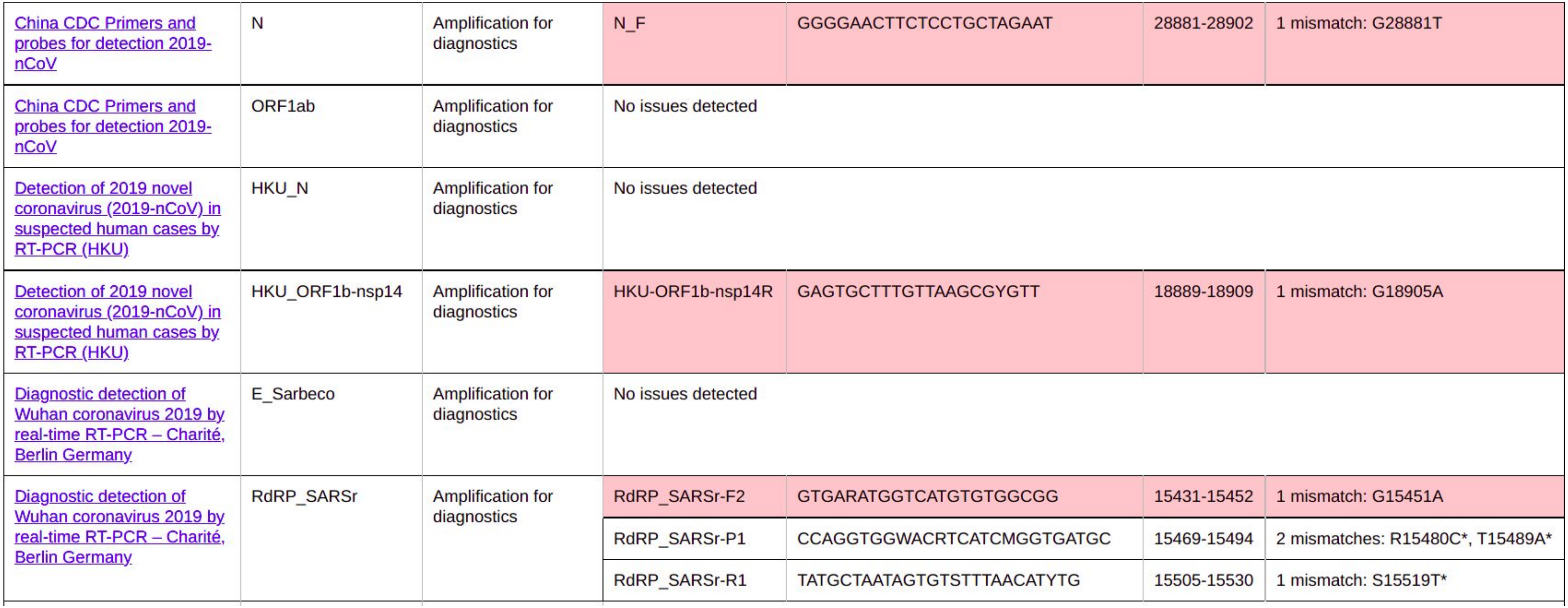


amino acid replacement in nsp1: H81Y
amino acid replacement in nsp3: A488S
amino acid replacement in nsp3: T819I
amino acid replacement in nsp3: P1228L
amino acid replacement in nsp3: P1469S
amino acid replacement in nsp4: V167L
amino acid replacement in nsp4: T492I
amino acid replacement in nsp5: F223L
amino acid replacement in nsp6: T77A
amino acid replacement in nsp12: P323L
amino acid replacement in nsp12: G671S
amino acid replacement in nsp12: Q822H
amino acid replacement in nsp13: P77L
amino acid replacement in nsp14: R289H
amino acid replacement in nsp14: A394V
amino acid replacement in S: T19R
amino acid replacement in S: T95I
amino acid replacement in S: G142D
amino acid replacement in S: L452R
amino acid replacement in S: T478K
amino acid replacement in S: E484K
amino acid replacement in S: D614G
amino acid replacement in S: P681R
amino acid replacement in S: D950N
amino acid replacement in S: E1202Q
amino acid replacement in ORF 3a: S26L
amino acid replacement in ORF 3a: D27Y
amino acid replacement in M: I82T
amino acid replacement in ORF 7a: V82A
amino acid replacement in ORF 7a: P84S
amino acid replacement in ORF 7a: T120I



Publication	Assay	Purpose	Primer/probe	Primer/probe sequence	Location on reference	Query sequence issues
ARTIC Network	nCoV-2019 nanopore primers V3	Whole genome sequencing	nCoV-2019_23_LEFT	ACAACTACTAACATAGTTACACGGTGT	6719-6745	1 mismatch: C6730T
			nCoV-2019_47_LEFT	AGGACTGGTATGATTTTGTAGAAAACCC	13919-13946	1 mismatch: C13944T
			nCoV-2019_4_LEFT	GGTGTATACTGCTGCCGTGAAC	944-965	1 mismatch: T961G
			nCoV-2019_63_LEFT	TGTTAAGCGTGTTGACTGGACT	18897-18918	1 mismatch: G18905A
			nCoV-2019_64_LEFT	TCGATAGATATCCTGCTAATTCCATTGT	19205-19232	1 mismatch: C19220T
			nCoV-2019_72_RIGHT	GTTGGATGGAAAGTGAGTTCAGAGT	22014-22038	1 deletion: 22029-22034
			nCoV-2019_73_LEFT	CAATTTTGTAATGATCCATTTTTGGGTGT	21962-21990	2 mismatches: T21973C, G21987A
			nCoV-2019_81_LEFT	GCACTTGGAAAACTTCAAGATGTGG	24392-24416	1 mismatch: G24410A
			nCoV-2019_93_LEFT	TGAGGCTGGTTCTAAATCACCCA	28082-28104	1 mismatch: G28085T
			nCoV-2019_93_RIGHT	CTCAACATGGCAAGGAAGACCT	28443-28464	1 mismatch: A28461G
			nCoV-2019_98_RIGHT	CCATGTGATTTTAATAGCTTCTTAGGAGAA	29837-29866	Coverage/alignment issues at 29857-29866

- mismatches at primer sequences: can introduce errors (but not necessarily so)
- coverage/alignment issues: the sequence is incomplete!

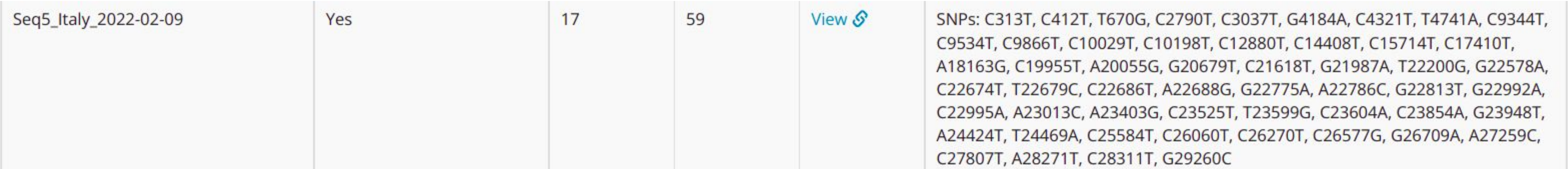


China CDC Primers and probes for detection 2019-nCoV	N	Amplification for diagnostics	N_F	GGGGAACTTCTCCTGCTAGAAT	28881-28902	1 mismatch: G28881T
China CDC Primers and probes for detection 2019-nCoV	ORF1ab	Amplification for diagnostics	No issues detected			
Detection of 2019 novel coronavirus (2019-nCoV) in suspected human cases by RT-PCR (HKU)	HKU_N	Amplification for diagnostics	No issues detected			
Detection of 2019 novel coronavirus (2019-nCoV) in suspected human cases by RT-PCR (HKU)	HKU_ORF1b-nsp14	Amplification for diagnostics	HKU-ORF1b-nsp14R	GAGTGCTTTGTTAAGCGYGTT	18889-18909	1 mismatch: G18905A
Diagnostic detection of Wuhan coronavirus 2019 by real-time RT-PCR – Charité, Berlin Germany	E_Sarbeco	Amplification for diagnostics	No issues detected			
Diagnostic detection of Wuhan coronavirus 2019 by real-time RT-PCR – Charité, Berlin Germany	RdRP_SARSr	Amplification for diagnostics	RdRP_SARSr-F2	GTGARATGGTCATGTGTGGCGG	15431-15452	1 mismatch: G15451A
			RdRP_SARSr-P1	CCAGGTGGWACRTCATCMGGTGATGC	15469-15494	2 mismatches: R15480C*, T15489A*
			RdRP_SARSr-R1	TATGCTAATAGTGTSTTTAACATYTG	15505-15530	1 mismatch: S15519T*

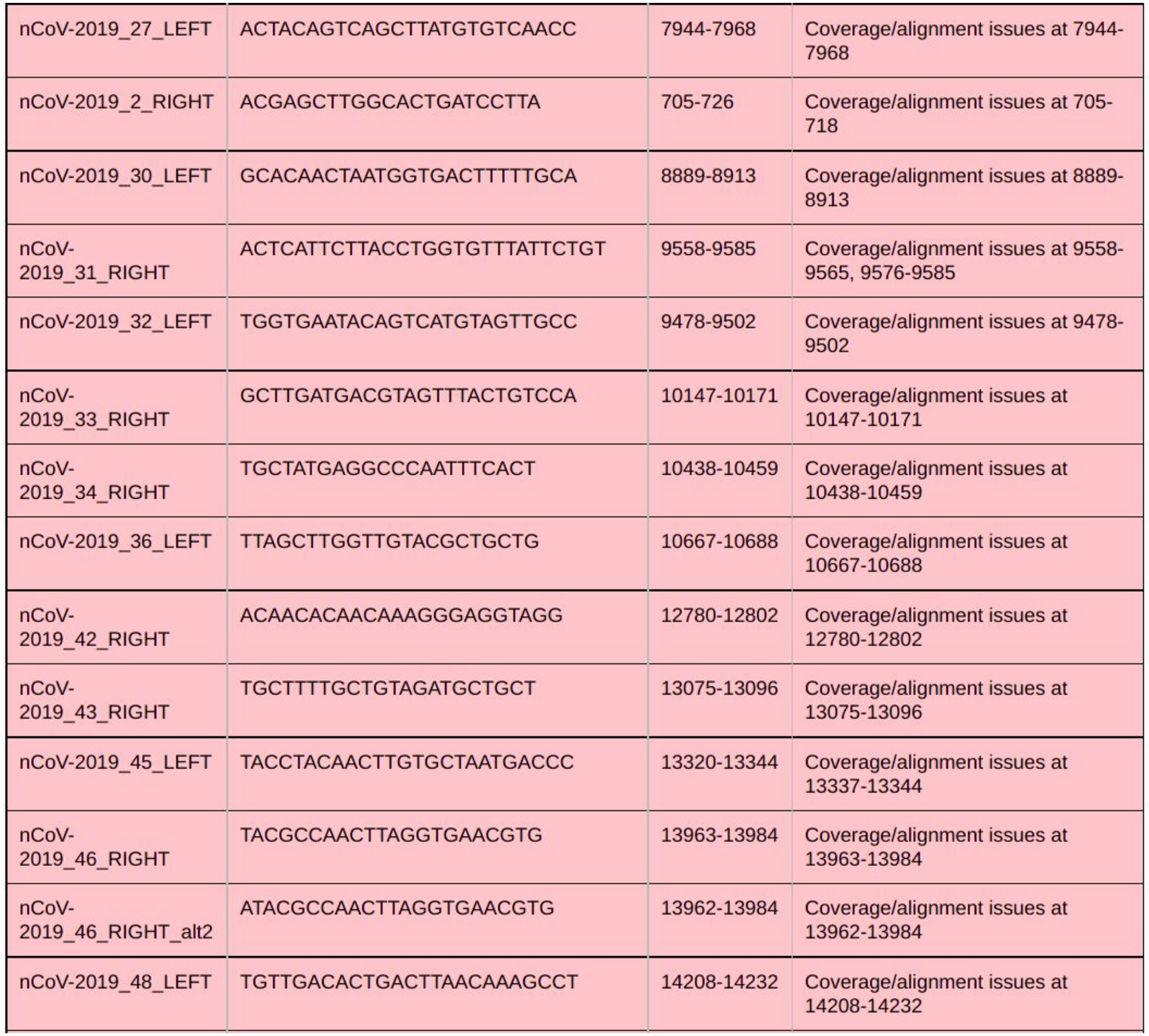
- alerts on “diagnostic tests”. In pink: might fail detection of one or more targets

CoV-GLUE: quality check

- If we scroll down and check the sequences, does any have more “issues” compared with the others?



Seq5_Italy_2022-02-09	Yes	17	59	View	SNPs: C313T, C412T, T670G, C2790T, C3037T, G4184A, C4321T, T4741A, C9344T, C9534T, C9866T, C10029T, C10198T, C12880T, C14408T, C15714T, C17410T, A18163G, C19955T, A20055G, G20679T, C21618T, G21987A, T22200G, G22578A, C22674T, T22679C, C22686T, A22688G, G22775A, A22786C, G22813T, G22992A, C22995A, A23013C, A23403G, C23525T, T23599G, C23604A, C23854A, G23948T, A24424T, T24469A, C25584T, C26060T, C26270T, C26577G, G26709A, A27259C, C27807T, A28271T, C28311T, G29260C



nCoV-2019_27_LEFT	ACTACAGTCAGCTTATGTGTCAACC	7944-7968	Coverage/alignment issues at 7944-7968
nCoV-2019_2_RIGHT	ACGAGCTTGGCACTGATCCTTA	705-726	Coverage/alignment issues at 705-718
nCoV-2019_30_LEFT	GCACAACTAATGGTGACTTTTTGCA	8889-8913	Coverage/alignment issues at 8889-8913
nCoV-2019_31_RIGHT	ACTCATTCTTACCTGGTGTTTATTCTGT	9558-9585	Coverage/alignment issues at 9558-9565, 9576-9585
nCoV-2019_32_LEFT	TGGTGAATACAGTCATGTAGTTGCC	9478-9502	Coverage/alignment issues at 9478-9502
nCoV-2019_33_RIGHT	GCTTGATGACGTAGTTTACTGTCCA	10147-10171	Coverage/alignment issues at 10147-10171
nCoV-2019_34_RIGHT	TGCTATGAGGCCCAATTTCACT	10438-10459	Coverage/alignment issues at 10438-10459
nCoV-2019_36_LEFT	TTAGCTTGGTTGTACGCTGCTG	10667-10688	Coverage/alignment issues at 10667-10688
nCoV-2019_42_RIGHT	ACAACACAACAAAGGGAGGTAGG	12780-12802	Coverage/alignment issues at 12780-12802
nCoV-2019_43_RIGHT	TGCTTTTGCTGTAGATGCTGCT	13075-13096	Coverage/alignment issues at 13075-13096
nCoV-2019_45_LEFT	TACCTACAACTTGTGCTAATGACCC	13320-13344	Coverage/alignment issues at 13337-13344
nCoV-2019_46_RIGHT	TACGCCAACTTAGGTGAACGTG	13963-13984	Coverage/alignment issues at 13963-13984
nCoV-2019_46_RIGHT_alt2	ATACGCCAACTTAGGTGAACGTG	13962-13984	Coverage/alignment issues at 13962-13984
nCoV-2019_48_LEFT	TGTTGACACTGACTTAACAAAGCCT	14208-14232	Coverage/alignment issues at 14208-14232

- coverage/alignment issues: the sequence is incomplete!
- at several “loci”

UShER: quality check, step#2

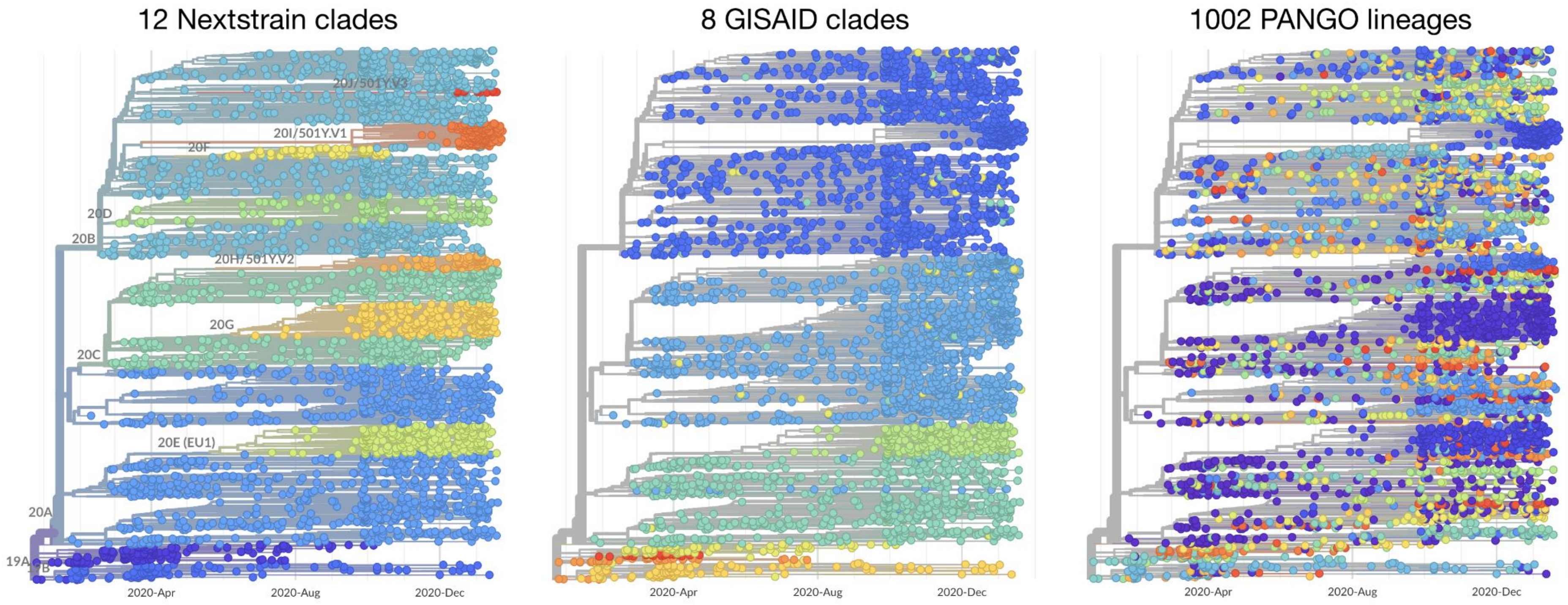
- Rapid and effective method to compare to other genome sequences (in GISAID or INSDC)
- web application https://genome.ucsc.edu/cgi-bin/hgPhyloPlace
- Detailed report of
  - completeness of the genome sequence
  - mutations (complete list)
  - similarity/dissimilarity with other sequences in dbs
  - phylogeny

By Turakhia et al, UCSC. See here for the paper

Tutorial: here

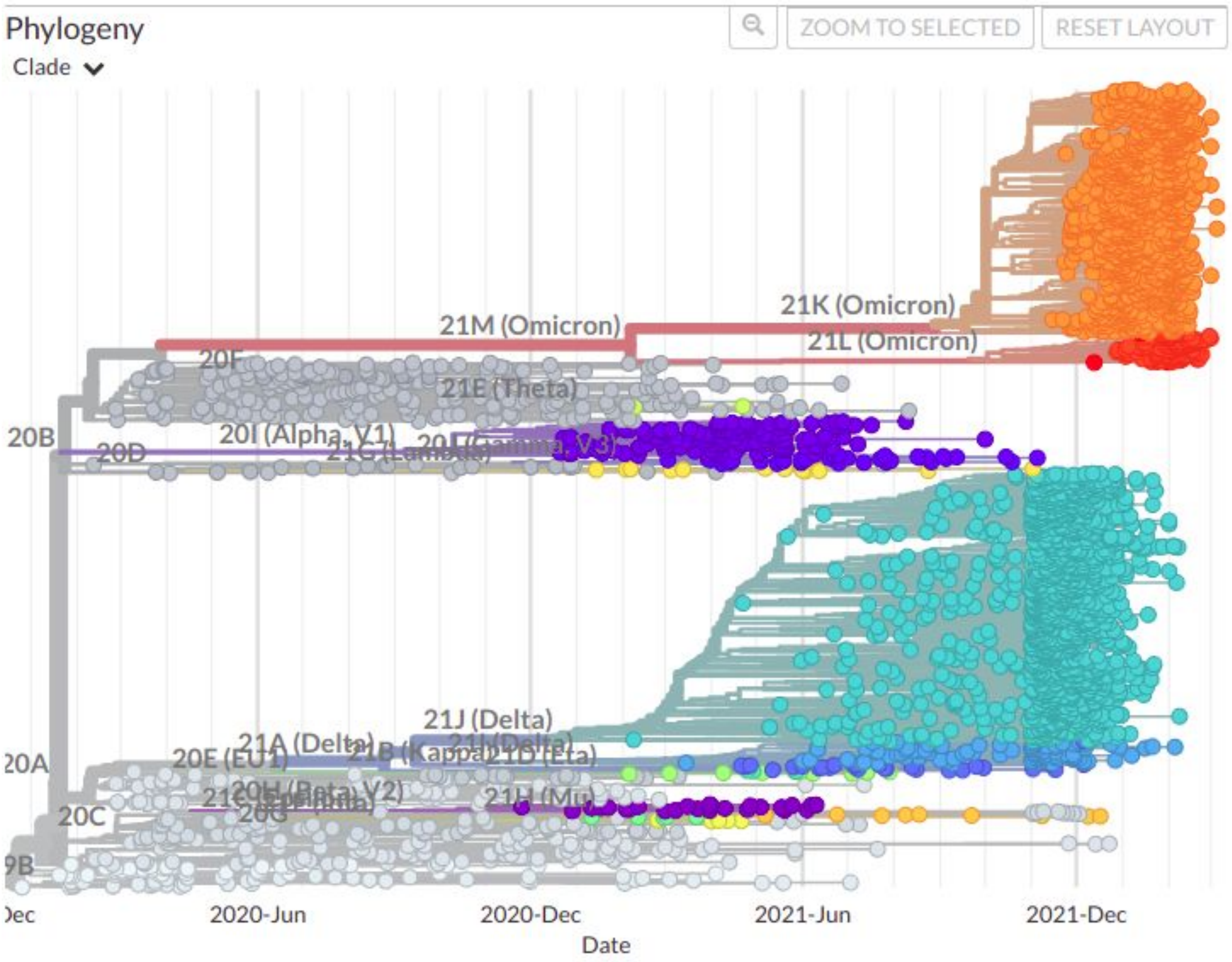
VideoTutorial: here

SARS-CoV-2: nomenclature



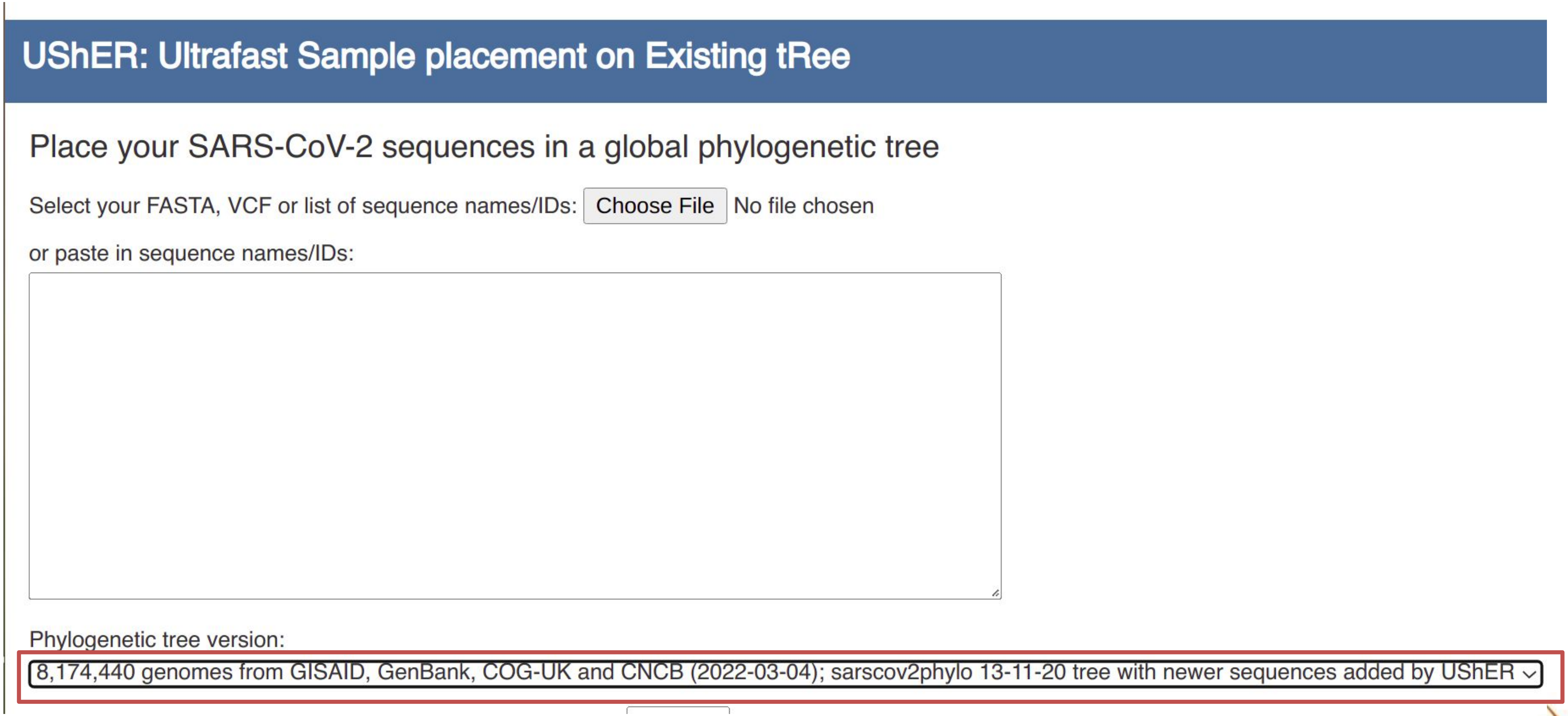
- Groups/variants are defined based on the evolutionary history of the virus
- Pango: currently the gold standard method
  - more granularity (groups) than Nextstrain and GISAID
    - better at tracking
    - less robust to noise

UShER, in brief



- Take your sequence(s)
- Fit them on the global SARS-CoV-2 phylogeny
- Compare with similar sequences in the tree
  - “classify” your sequence (variant)
  - check if potential sequencing issues (similar to other sequences of the same type?)

UShER, hands on



UShER: Ultrafast Sample placement on Existing tRee

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

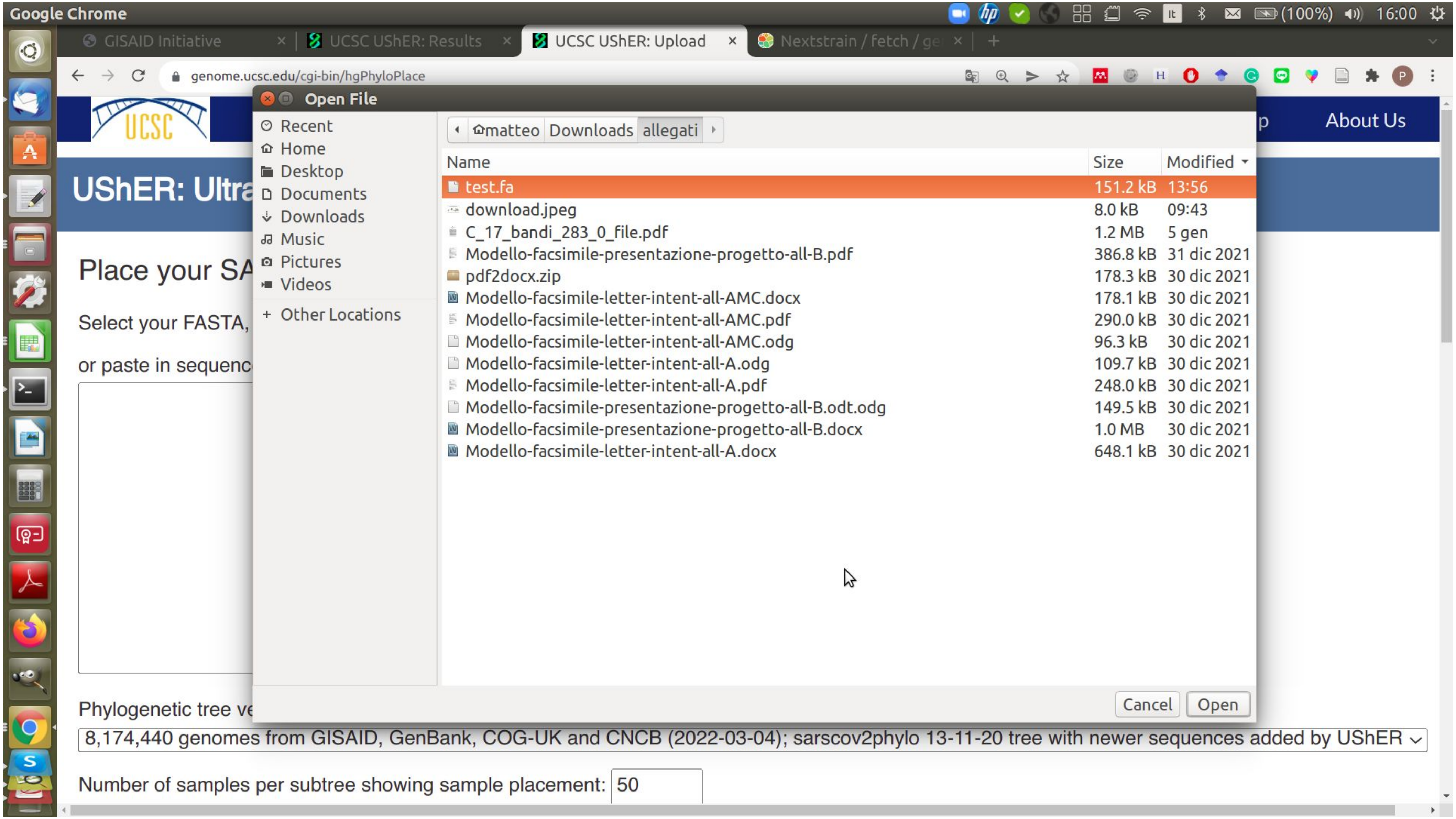
Select your FASTA, VCF or list of sequence names/IDs: No file chosen

or paste in sequence names/IDs:

Phylogenetic tree version:

8,174,440 genomes from GISAID, GenBank, COG-UK and CNCB (2022-03-04); sarscov2phylo 13-11-20 tree with newer sequences added by UShER

UShER, hands on



genome.ucsc.edu/cgi-bin/hgPhyloPlace

UShER: Ultra

Place your SA

Select your FASTA,

or paste in sequenc

matteo › Downloads › allegati

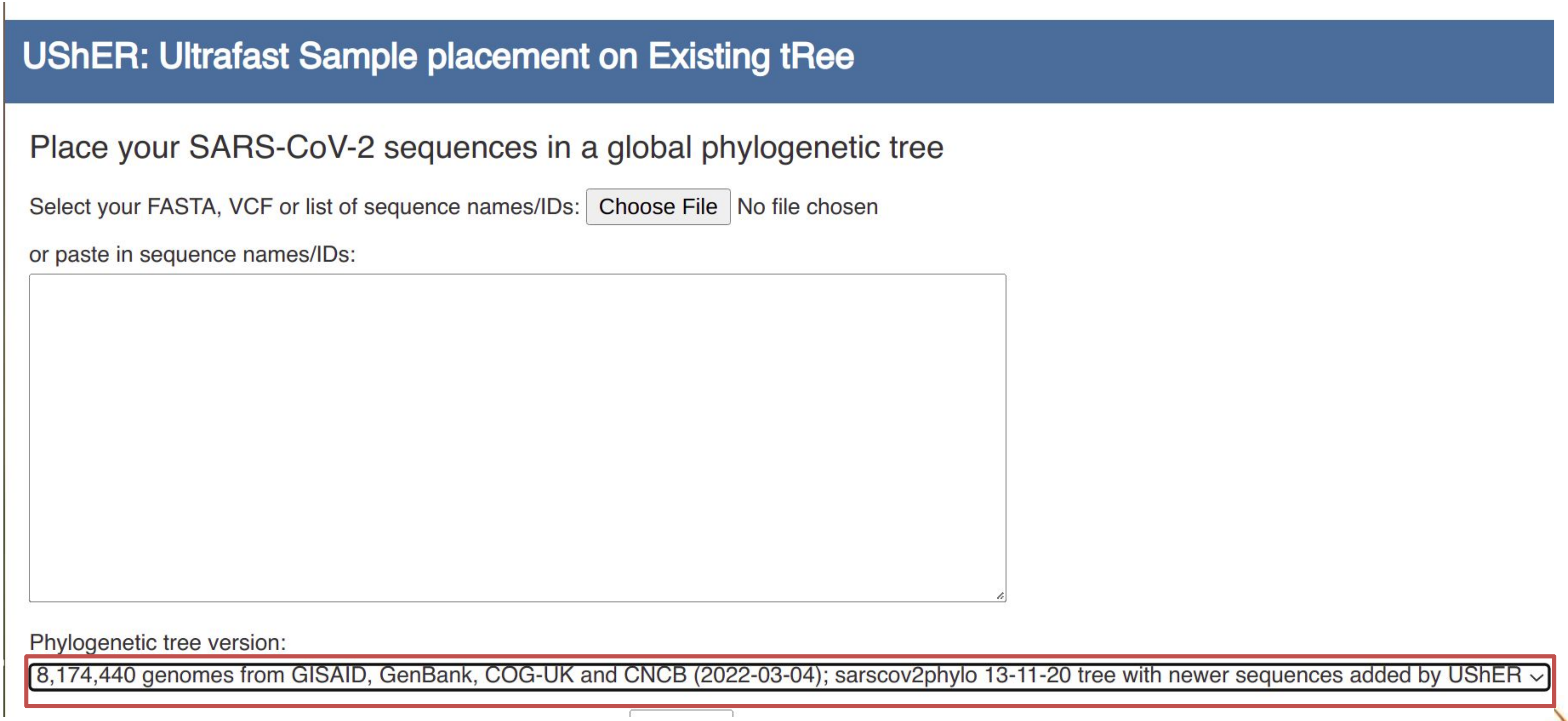
Name	Size	Modified
test.fa	151.2 kB	13:56
download.jpeg	8.0 kB	09:43
C_17_bandi_283_0_file.pdf	1.2 MB	5 gen
Modello-facsimile-presentazione-progetto-all-B.pdf	386.8 kB	31 dic 2021
pdf2docx.zip	178.3 kB	30 dic 2021
Modello-facsimile-letter-intent-all-AMC.docx	178.1 kB	30 dic 2021
Modello-facsimile-letter-intent-all-AMC.pdf	290.0 kB	30 dic 2021
Modello-facsimile-letter-intent-all-AMC.odg	96.3 kB	30 dic 2021
Modello-facsimile-letter-intent-all-A.odg	109.7 kB	30 dic 2021
Modello-facsimile-letter-intent-all-A.pdf	248.0 kB	30 dic 2021
Modello-facsimile-presentazione-progetto-all-B.odt.odg	149.5 kB	30 dic 2021
Modello-facsimile-presentazione-progetto-all-B.docx	1.0 MB	30 dic 2021
Modello-facsimile-letter-intent-all-A.docx	648.1 kB	30 dic 2021

Phylogenetic tree ve

8,174,440 genomes from GISAID, GenBank, COG-UK and CNCB (2022-03-04); sarscov2phylo 13-11-20 tree with newer sequences added by UShER

Number of samples per subtree showing sample placement: 50

UShER, hands on



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

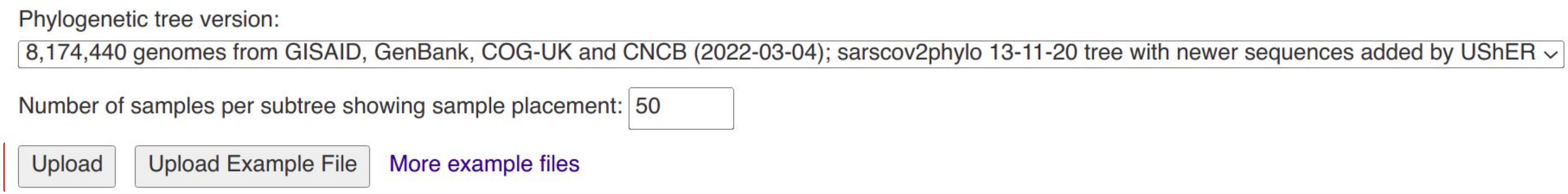
or paste in sequence names/IDs:

Phylogenetic tree version:

8,174,440 genomes from GISAID, GenBank, COG-UK and CNCB (2022-03-04); sarscov2phylo 13-11-20 tree with newer sequences added by UShER

- Select: GISAID or INSDC -> GISAID

UShER, hands on



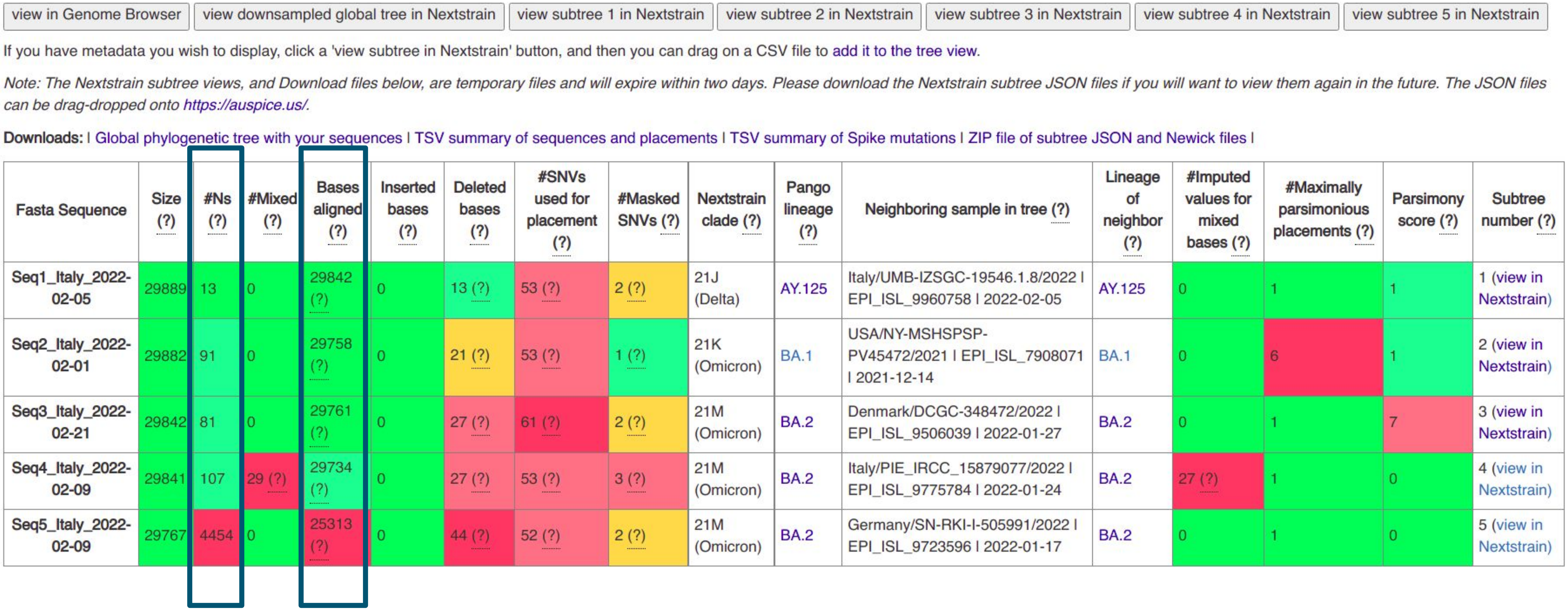
Phylogenetic tree version:

8,174,440 genomes from GISAID, GenBank, COG-UK and CNCB (2022-03-04); sarscov2phylo 13-11-20 tree with newer sequences added by UShER

Number of samples per subtree showing sample placement: 50

- Hit:UPLOAD
  - Results in approx ~ 5 minutes

UShER, main results



view in Genome Browser
view downsampled global tree in Nextstrain
view subtree 1 in Nextstrain
view subtree 2 in Nextstrain
view subtree 3 in Nextstrain
view subtree 4 in Nextstrain
view subtree 5 in Nextstrain

If you have metadata you wish to display, click a 'view subtree in Nextstrain' button, and then you can drag on a CSV file to add it to the tree view.

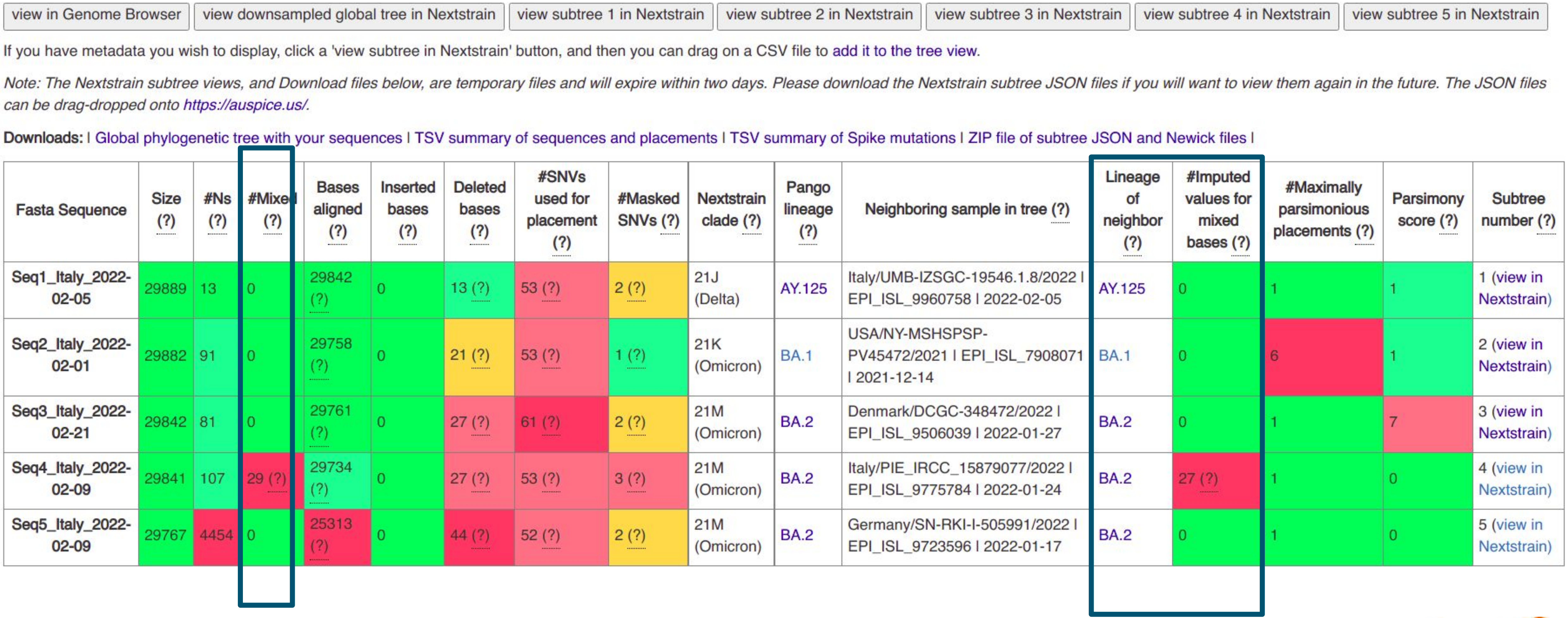
Note: The Nextstrain subtree views, and Download files below, are temporary files and will expire within two days. Please download the Nextstrain subtree JSON files if you will want to view them again in the future. The JSON files can be drag-dropped onto https://auspice.us/.

Downloads: | Global phylogenetic tree with your sequences | TSV summary of sequences and placements | TSV summary of Spike mutations | ZIP file of subtree JSON and Newick files |

Fasta Sequence	Size (?)	#Ns (?)	#Mixed (?)	Bases aligned (?)	Inserted bases (?)	Deleted bases (?)	#SNVs used for placement (?)	#Masked SNVs (?)	Nextstrain clade (?)	Pango lineage (?)	Neighboring sample in tree (?)	Lineage of neighbor (?)	#Imputed values for mixed bases (?)	#Maximally parsimonious placements (?)	Parsimony score (?)	Subtree number (?)
Seq1_Italy_2022-02-05	29889	13	0	29842 (?)	0	13 (?)	53 (?)	2 (?)	21J (Delta)	AY.125	Italy/UMB-IZSGC-19546.1.8/2022 | EPI_ISL_9960758 | 2022-02-05	AY.125	0	1	1	1 (view in Nextstrain)
Seq2_Italy_2022-02-01	29882	91	0	29758 (?)	0	21 (?)	53 (?)	1 (?)	21K (Omicron)	BA.1	USA/NY-MSHSPSP-PV45472/2021 | EPI_ISL_7908071 | 2021-12-14	BA.1	0	6	1	2 (view in Nextstrain)
Seq3_Italy_2022-02-21	29842	81	0	29761 (?)	0	27 (?)	61 (?)	2 (?)	21M (Omicron)	BA.2	Denmark/DCGC-348472/2022 | EPI_ISL_9506039 | 2022-01-27	BA.2	0	1	7	3 (view in Nextstrain)
Seq4_Italy_2022-02-09	29841	107	29 (?)	29734 (?)	0	27 (?)	53 (?)	3 (?)	21M (Omicron)	BA.2	Italy/PIE_IRCC_15879077/2022 | EPI_ISL_9775784 | 2022-01-24	BA.2	27 (?)	1	0	4 (view in Nextstrain)
Seq5_Italy_2022-02-09	29767	4454	0	25313 (?)	0	44 (?)	52 (?)	2 (?)	21M (Omicron)	BA.2	Germany/SN-RKI-I-505991/2022 | EPI_ISL_9723596 | 2022-01-17	BA.2	0	1	0	5 (view in Nextstrain)

- S5: many masked bases. We were already aware of

UShER, main results



view in Genome Browser
view downsampled global tree in Nextstrain
view subtree 1 in Nextstrain
view subtree 2 in Nextstrain
view subtree 3 in Nextstrain
view subtree 4 in Nextstrain
view subtree 5 in Nextstrain

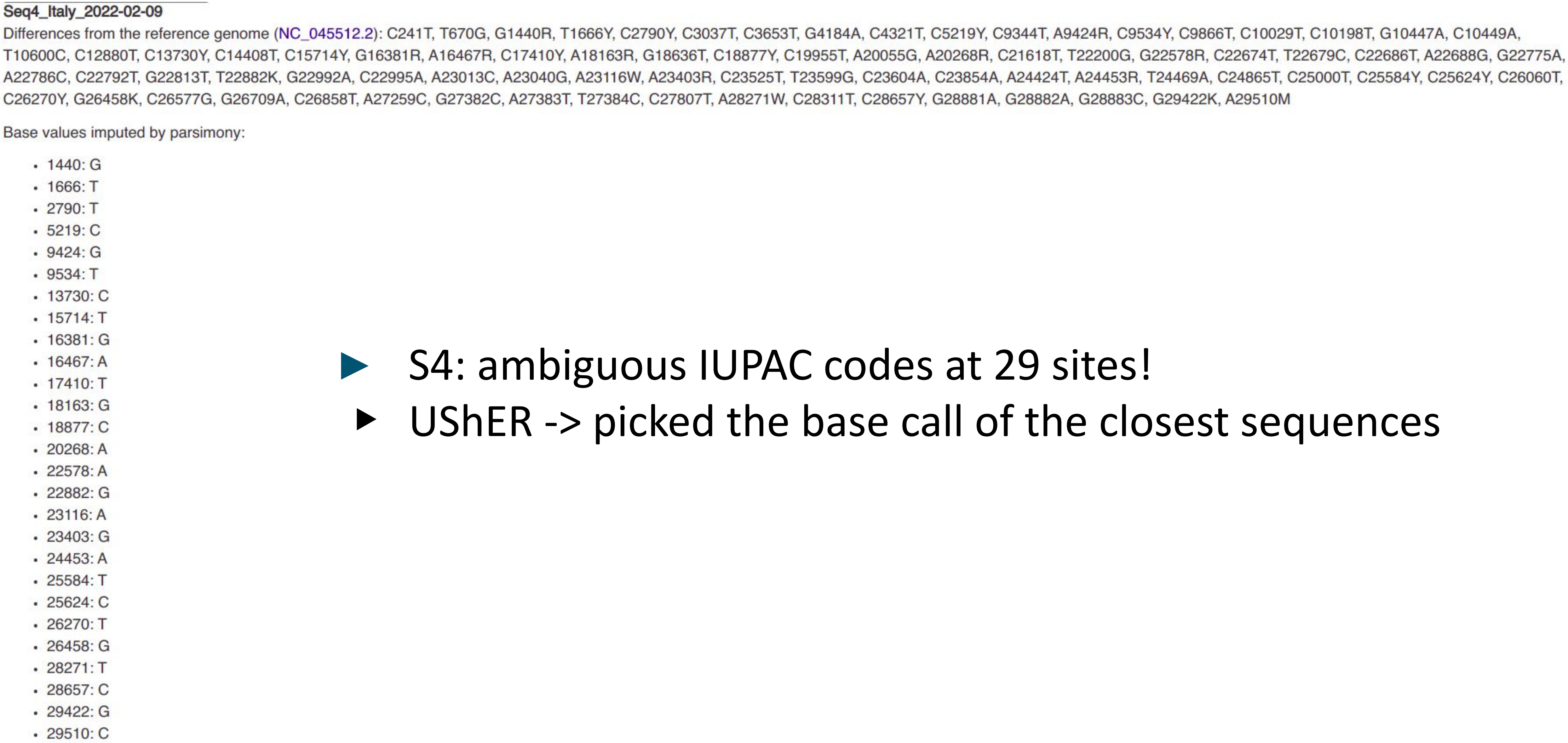
If you have metadata you wish to display, click a 'view subtree in Nextstrain' button, and then you can drag on a CSV file to add it to the tree view.

Note: The Nextstrain subtree views, and Download files below, are temporary files and will expire within two days. Please download the Nextstrain subtree JSON files if you will want to view them again in the future. The JSON files can be drag-dropped onto https://auspice.us/.

Downloads: | Global phylogenetic tree with your sequences | TSV summary of sequences and placements | TSV summary of Spike mutations | ZIP file of subtree JSON and Newick files |

Fasta Sequence	Size (?)	#Ns (?)	#Mixed (?)	Bases aligned (?)	Inserted bases (?)	Deleted bases (?)	#SNVs used for placement (?)	#Masked SNVs (?)	Nextstrain clade (?)	Pango lineage (?)	Neighboring sample in tree (?)	Lineage of neighbor (?)	#Imputed values for mixed bases (?)	#Maximally parsimonious placements (?)	Parsimony score (?)	Subtree number (?)
Seq1_Italy_2022-02-05	29889	13	0	29842 (?)	0	13 (?)	53 (?)	2 (?)	21J (Delta)	AY.125	Italy/UMB-IZSGC-19546.1.8/2022 | EPI_ISL_9960758 | 2022-02-05	AY.125	0	1	1	1 (view in Nextstrain)
Seq2_Italy_2022-02-01	29882	91	0	29758 (?)	0	21 (?)	53 (?)	1 (?)	21K (Omicron)	BA.1	USA/NY-MSHSPSP-PV45472/2021 | EPI_ISL_7908071 | 2021-12-14	BA.1	0	6	1	2 (view in Nextstrain)
Seq3_Italy_2022-02-21	29842	81	0	29761 (?)	0	27 (?)	61 (?)	2 (?)	21M (Omicron)	BA.2	Denmark/DCGC-348472/2022 | EPI_ISL_9506039 | 2022-01-27	BA.2	0	1	7	3 (view in Nextstrain)
Seq4_Italy_2022-02-09	29841	107	29 (?)	29734 (?)	0	27 (?)	53 (?)	3 (?)	21M (Omicron)	BA.2	Italy/PIE_IRCC_15879077/2022 | EPI_ISL_9775784 | 2022-01-24	BA.2	27 (?)	1	0	4 (view in Nextstrain)
Seq5_Italy_2022-02-09	29767	4454	0	25313 (?)	0	44 (?)	52 (?)	2 (?)	21M (Omicron)	BA.2	Germany/SN-RKI-I-505991/2022 | EPI_ISL_9723596 | 2022-01-17	BA.2	0	1	0	5 (view in Nextstrain)

- S4: ambiguous IUPAC codes at 29 sites!
  - UShER -> picked the base call of the closest sequences



Seq4_Italy_2022-02-09

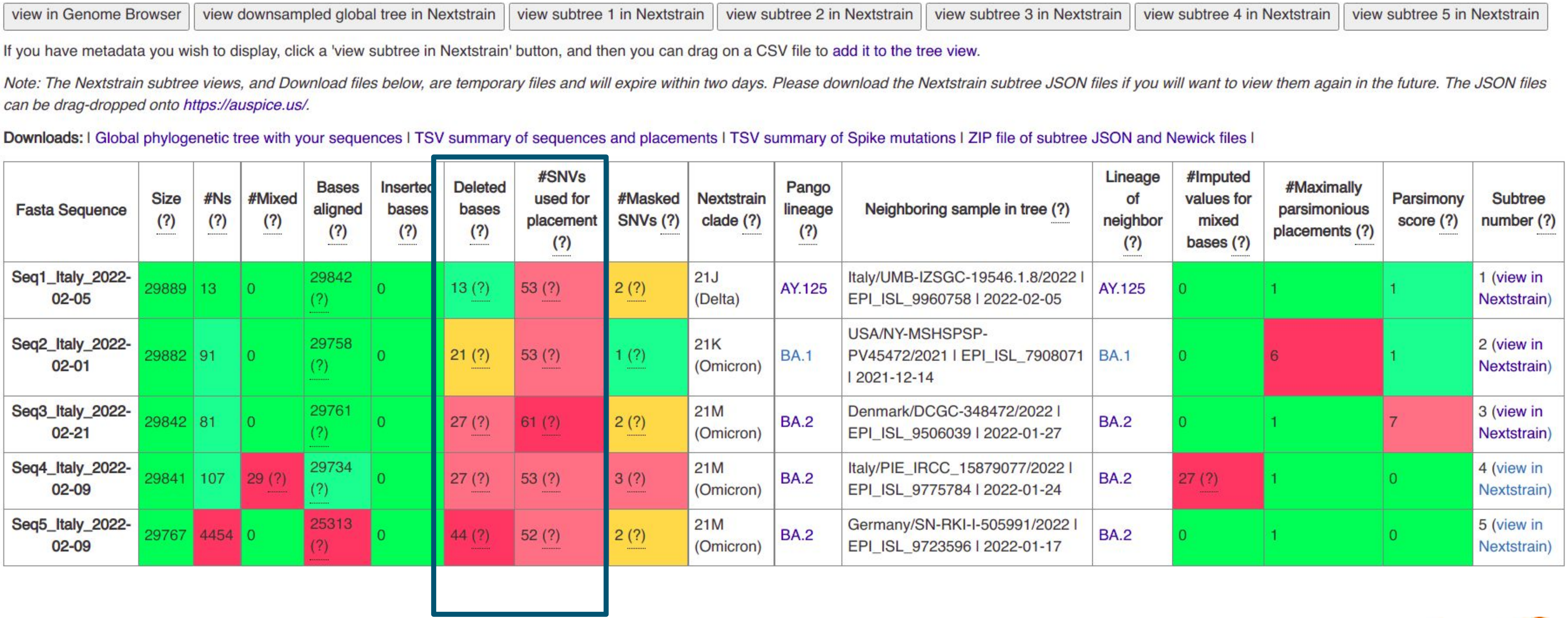
Differences from the reference genome (NC_045512.2): C241T, T670G, G1440R, T1666Y, C2790Y, C3037T, C3653T, G4184A, C4321T, C5219Y, C9344T, A9424R, C9534Y, C9866T, C10029T, C10198T, G10447A, C10449A, T10600C, C12880T, C13730Y, C14408T, C15714Y, G16381R, A16467R, C17410Y, A18163R, G18636T, C18877Y, C19955T, A20055G, A20268R, C21618T, T22200G, G22578R, C22674T, T22679C, C22686T, A22688G, G22775A, A22786C, C22792T, G22813T, T22882K, G22992A, C22995A, A23013C, A23040G, A23116W, A23403R, C23525T, T23599G, C23604A, C23854A, A24424T, A24453R, T24469A, C24865T, C25000T, C25584Y, C25624Y, C26060T, C26270Y, G26458K, C26577G, G26709A, C26858T, A27259C, G27382C, A27383T, T27384C, C27807T, A28271W, C28311T, C28657Y, G28881A, G28882A, G28883C, G29422K, A29510M

Base values imputed by parsimony:

- 1440: G
- 1666: T
- 2790: T
- 5219: C
- 9424: G
- 9534: T
- 13730: C
- 15714: T
- 16381: G
- 16467: A
- 17410: T
- 18163: G
- 18877: C
- 20268: A
- 22578: A
- 22882: G
- 23116: A
- 23403: G
- 24453: A
- 25584: T
- 25624: C
- 26270: T
- 26458: G
- 28271: T
- 28657: C
- 29422: G
- 29510: C

- S4: ambiguous IUPAC codes at 29 sites!
  - UShER -> picked the base call of the closest sequences

UShER, main results



view in Genome Browser | view downsampled global tree in Nextstrain | view subtree 1 in Nextstrain | view subtree 2 in Nextstrain | view subtree 3 in Nextstrain | view subtree 4 in Nextstrain | view subtree 5 in Nextstrain

If you have metadata you wish to display, click a 'view subtree in Nextstrain' button, and then you can drag on a CSV file to add it to the tree view.

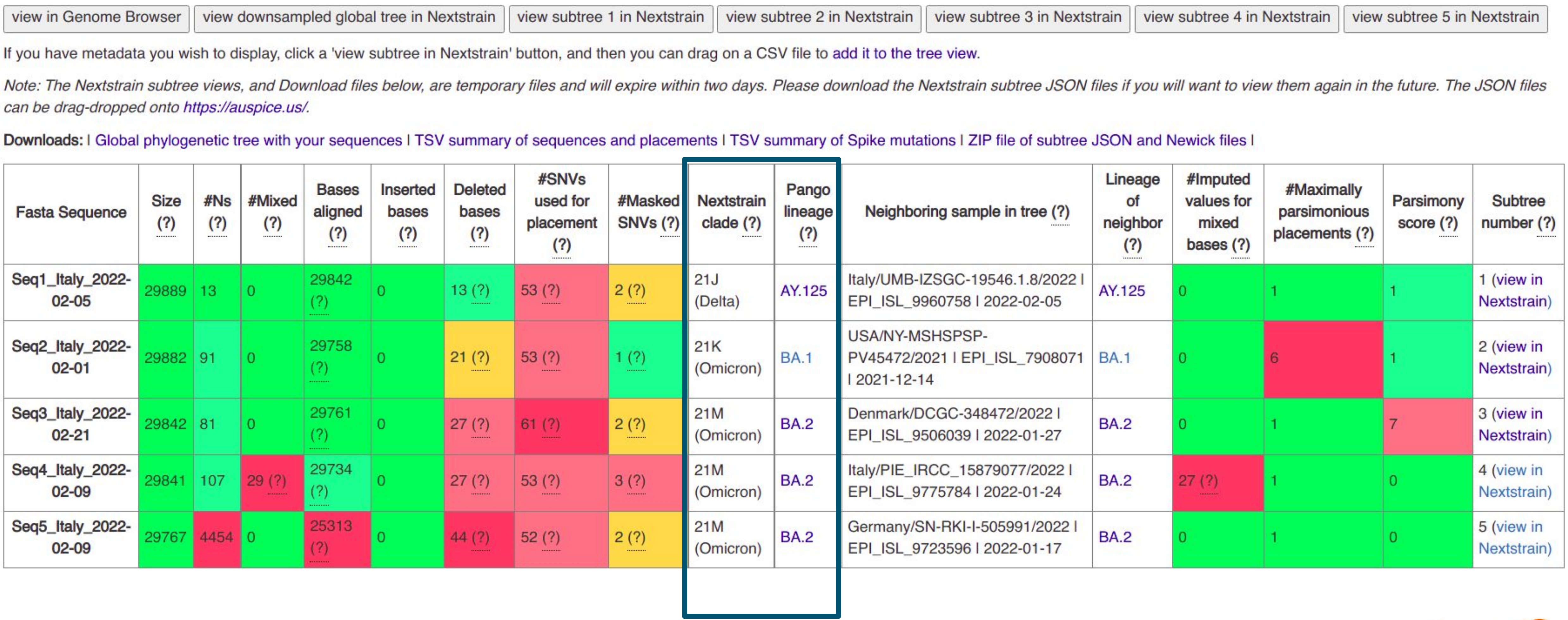
Note: The Nextstrain subtree views, and Download files below, are temporary files and will expire within two days. Please download the Nextstrain subtree JSON files if you will want to view them again in the future. The JSON files can be drag-dropped onto https://auspice.us/.

Downloads: | Global phylogenetic tree with your sequences | TSV summary of sequences and placements | TSV summary of Spike mutations | ZIP file of subtree JSON and Newick files |

Fasta Sequence	Size (?)	#Ns (?)	#Mixed (?)	Bases aligned (?)	Inserted bases (?)	Deleted bases (?)	#SNVs used for placement (?)	#Masked SNVs (?)	Nextstrain clade (?)	Pango lineage (?)	Neighboring sample in tree (?)	Lineage of neighbor (?)	#Imputed values for mixed bases (?)	#Maximally parsimonious placements (?)	Parsimony score (?)	Subtree number (?)
Seq1_Italy_2022-02-05	29889	13	0	29842 (?)	0	13 (?)	53 (?)	2 (?)	21J (Delta)	AY.125	Italy/UMB-IZSGC-19546.1.8/2022 | EPI_ISL_9960758 | 2022-02-05	AY.125	0	1	1	1 (view in Nextstrain)
Seq2_Italy_2022-02-01	29882	91	0	29758 (?)	0	21 (?)	53 (?)	1 (?)	21K (Omicron)	BA.1	USA/NY-MSHSPSP-PV45472/2021 | EPI_ISL_7908071 | 2021-12-14	BA.1	0	6	1	2 (view in Nextstrain)
Seq3_Italy_2022-02-21	29842	81	0	29761 (?)	0	27 (?)	61 (?)	2 (?)	21M (Omicron)	BA.2	Denmark/DCGC-348472/2022 | EPI_ISL_9506039 | 2022-01-27	BA.2	0	1	7	3 (view in Nextstrain)
Seq4_Italy_2022-02-09	29841	107	29 (?)	29734 (?)	0	27 (?)	53 (?)	3 (?)	21M (Omicron)	BA.2	Italy/PIE_IRCC_15879077/2022 | EPI_ISL_9775784 | 2022-01-24	BA.2	27 (?)	1	0	4 (view in Nextstrain)
Seq5_Italy_2022-02-09	29767	4454	0	25313 (?)	0	44 (?)	52 (?)	2 (?)	21M (Omicron)	BA.2	Germany/SN-RKI-I-505991/2022 | EPI_ISL_9723596 | 2022-01-17	BA.2	0	1	0	5 (view in Nextstrain)

- All sequences have many “mutations”
  - Marked in red. But not an issue: see next slide

UShER, main results



view in Genome Browser
view downsampled global tree in Nextstrain
view subtree 1 in Nextstrain
view subtree 2 in Nextstrain
view subtree 3 in Nextstrain
view subtree 4 in Nextstrain
view subtree 5 in Nextstrain

If you have metadata you wish to display, click a 'view subtree in Nextstrain' button, and then you can drag on a CSV file to add it to the tree view.

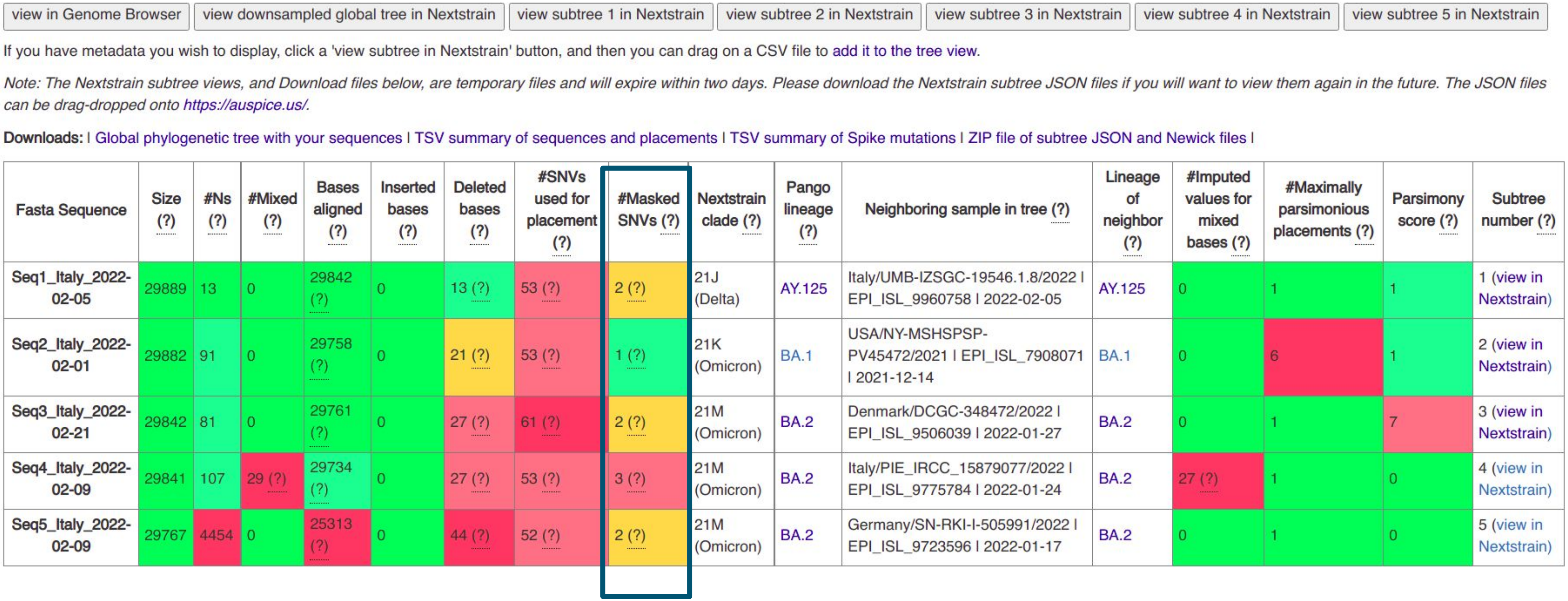
Note: The Nextstrain subtree views, and Download files below, are temporary files and will expire within two days. Please download the Nextstrain subtree JSON files if you will want to view them again in the future. The JSON files can be drag-dropped onto https://auspice.us/.

Downloads: | Global phylogenetic tree with your sequences | TSV summary of sequences and placements | TSV summary of Spike mutations | ZIP file of subtree JSON and Newick files |

Fasta Sequence	Size (?)	#Ns (?)	#Mixed (?)	Bases aligned (?)	Inserted bases (?)	Deleted bases (?)	#SNVs used for placement (?)	#Masked SNVs (?)	Nextstrain clade (?)	Pango lineage (?)	Neighboring sample in tree (?)	Lineage of neighbor (?)	#Imputed values for mixed bases (?)	#Maximally parsimonious placements (?)	Parsimony score (?)	Subtree number (?)
Seq1_Italy_2022-02-05	29889	13	0	29842 (?)	0	13 (?)	53 (?)	2 (?)	21J (Delta)	AY.125	Italy/UMB-IZSGC-19546.1.8/2022 | EPI_ISL_9960758 | 2022-02-05	AY.125	0	1	1	1 (view in Nextstrain)
Seq2_Italy_2022-02-01	29882	91	0	29758 (?)	0	21 (?)	53 (?)	1 (?)	21K (Omicron)	BA.1	USA/NY-MSHSPSP-PV45472/2021 | EPI_ISL_7908071 | 2021-12-14	BA.1	0	6	1	2 (view in Nextstrain)
Seq3_Italy_2022-02-21	29842	81	0	29761 (?)	0	27 (?)	61 (?)	2 (?)	21M (Omicron)	BA.2	Denmark/DCGC-348472/2022 | EPI_ISL_9506039 | 2022-01-27	BA.2	0	1	7	3 (view in Nextstrain)
Seq4_Italy_2022-02-09	29841	107	29 (?)	29734 (?)	0	27 (?)	53 (?)	3 (?)	21M (Omicron)	BA.2	Italy/PIE_IRCC_15879077/2022 | EPI_ISL_9775784 | 2022-01-24	BA.2	27 (?)	1	0	4 (view in Nextstrain)
Seq5_Italy_2022-02-09	29767	4454	0	25313 (?)	0	44 (?)	52 (?)	2 (?)	21M (Omicron)	BA.2	Germany/SN-RKI-I-505991/2022 | EPI_ISL_9723596 | 2022-01-17	BA.2	0	1	0	5 (view in Nextstrain)

- We have 1 Delta and 4 Omicron genomes
  - Omicron and Delta have many mutations. No issue here!

UShER, main results



view in Genome Browser | view downsampled global tree in Nextstrain | view subtree 1 in Nextstrain | view subtree 2 in Nextstrain | view subtree 3 in Nextstrain | view subtree 4 in Nextstrain | view subtree 5 in Nextstrain

If you have metadata you wish to display, click a 'view subtree in Nextstrain' button, and then you can drag on a CSV file to add it to the tree view.

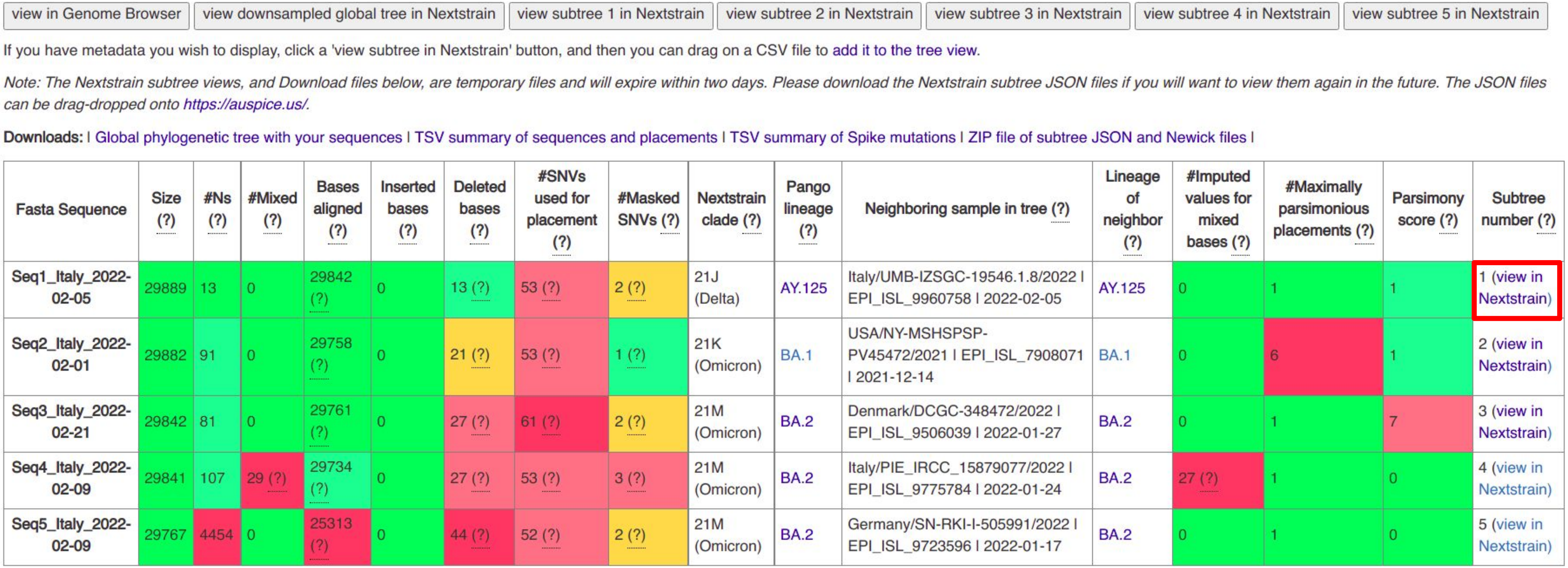
Note: The Nextstrain subtree views, and Download files below, are temporary files and will expire within two days. Please download the Nextstrain subtree JSON files if you will want to view them again in the future. The JSON files can be drag-dropped onto https://auspice.us/.

Downloads: | Global phylogenetic tree with your sequences | TSV summary of sequences and placements | TSV summary of Spike mutations | ZIP file of subtree JSON and Newick files |

Fasta Sequence	Size (?)	#Ns (?)	#Mixed (?)	Bases aligned (?)	Inserted bases (?)	Deleted bases (?)	#SNVs used for placement (?)	#Masked SNVs (?)	Nextstrain clade (?)	Pango lineage (?)	Neighboring sample in tree (?)	Lineage of neighbor (?)	#Imputed values for mixed bases (?)	#Maximally parsimonious placements (?)	Parsimony score (?)	Subtree number (?)
Seq1_Italy_2022-02-05	29889	13	0	29842 (?)	0	13 (?)	53 (?)	2 (?)	21J (Delta)	AY.125	Italy/UMB-IZSGC-19546.1.8/2022 | EPI_ISL_9960758 | 2022-02-05	AY.125	0	1	1	1 (view in Nextstrain)
Seq2_Italy_2022-02-01	29882	91	0	29758 (?)	0	21 (?)	53 (?)	1 (?)	21K (Omicron)	BA.1	USA/NY-MSHSPSP-PV45472/2021 | EPI_ISL_7908071 | 2021-12-14	BA.1	0	6	1	2 (view in Nextstrain)
Seq3_Italy_2022-02-21	29842	81	0	29761 (?)	0	27 (?)	61 (?)	2 (?)	21M (Omicron)	BA.2	Denmark/DCGC-348472/2022 | EPI_ISL_9506039 | 2022-01-27	BA.2	0	1	7	3 (view in Nextstrain)
Seq4_Italy_2022-02-09	29841	107	29 (?)	29734 (?)	0	27 (?)	53 (?)	3 (?)	21M (Omicron)	BA.2	Italy/PIE_IRCC_15879077/2022 | EPI_ISL_9775784 | 2022-01-24	BA.2	27 (?)	1	0	4 (view in Nextstrain)
Seq5_Italy_2022-02-09	29767	4454	0	25313 (?)	0	44 (?)	52 (?)	2 (?)	21M (Omicron)	BA.2	Germany/SN-RKI-I-505991/2022 | EPI_ISL_9723596 | 2022-01-17	BA.2	0	1	0	5 (view in Nextstrain)

- We have 1 Delta and 4 Omicron genomes
  - Omicron and Delta have many mutations.

To see the phylogeny



view in Genome Browser | view downsampled global tree in Nextstrain | view subtree 1 in Nextstrain | view subtree 2 in Nextstrain | view subtree 3 in Nextstrain | view subtree 4 in Nextstrain | view subtree 5 in Nextstrain

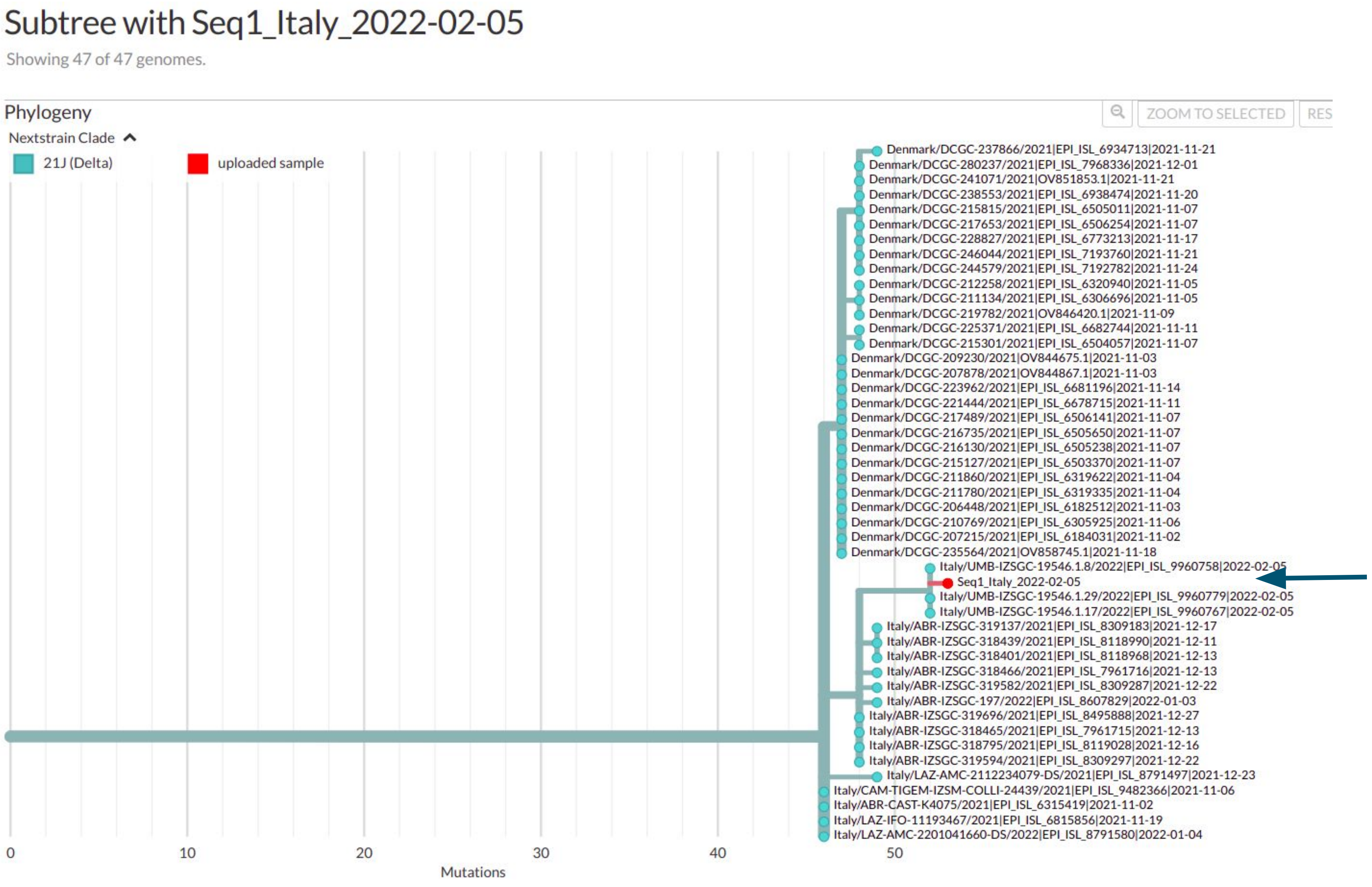
If you have metadata you wish to display, click a 'view subtree in Nextstrain' button, and then you can drag on a CSV file to add it to the tree view.

Note: The Nextstrain subtree views, and Download files below, are temporary files and will expire within two days. Please download the Nextstrain subtree JSON files if you will want to view them again in the future. The JSON files can be drag-dropped onto https://auspice.us/.

Downloads: | Global phylogenetic tree with your sequences | TSV summary of sequences and placements | TSV summary of Spike mutations | ZIP file of subtree JSON and Newick files |

Fasta Sequence	Size (?)	#Ns (?)	#Mixed (?)	Bases aligned (?)	Inserted bases (?)	Deleted bases (?)	#SNVs used for placement (?)	#Masked SNVs (?)	Nextstrain clade (?)	Pango lineage (?)	Neighboring sample in tree (?)	Lineage of neighbor (?)	#Imputed values for mixed bases (?)	#Maximally parsimonious placements (?)	Parsimony score (?)	Subtree number (?)
Seq1_Italy_2022-02-05	29889	13	0	29842 (?)	0	13 (?)	53 (?)	2 (?)	21J (Delta)	AY.125	Italy/UMB-IZSGC-19546.1.8/2022 | EPI_ISL_9960758 | 2022-02-05	AY.125	0	1	1	1 (view in Nextstrain)
Seq2_Italy_2022-02-01	29882	91	0	29758 (?)	0	21 (?)	53 (?)	1 (?)	21K (Omicron)	BA.1	USA/NY-MSHSPSP-PV45472/2021 | EPI_ISL_7908071 | 2021-12-14	BA.1	0	6	1	2 (view in Nextstrain)
Seq3_Italy_2022-02-21	29842	81	0	29761 (?)	0	27 (?)	61 (?)	2 (?)	21M (Omicron)	BA.2	Denmark/DCGC-348472/2022 | EPI_ISL_9506039 | 2022-01-27	BA.2	0	1	7	3 (view in Nextstrain)
Seq4_Italy_2022-02-09	29841	107	29 (?)	29734 (?)	0	27 (?)	53 (?)	3 (?)	21M (Omicron)	BA.2	Italy/PIE_IRCC_15879077/2022 | EPI_ISL_9775784 | 2022-01-24	BA.2	27 (?)	1	0	4 (view in Nextstrain)
Seq5_Italy_2022-02-09	29767	4454	0	25313 (?)	0	44 (?)	52 (?)	2 (?)	21M (Omicron)	BA.2	Germany/SN-RKI-I-505991/2022 | EPI_ISL_9723596 | 2022-01-17	BA.2	0	1	0	5 (view in Nextstrain)

UShER, main results



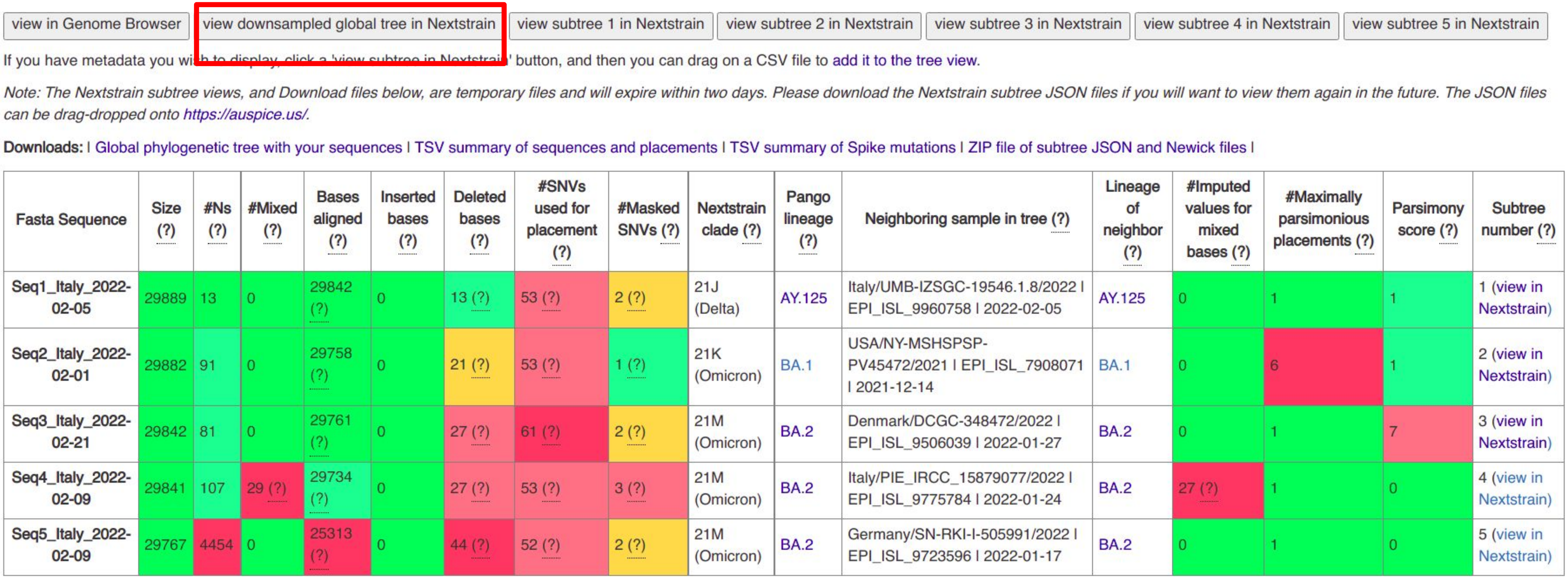
Subtree with Seq1_Italy_2022-02-05

Showing 47 of 47 genomes.

Phylogeny
ZOOM TO SELECTED
RES
Nextstrain Clade
🟩 21J (Delta)
🟥 uploaded sample
Denmark/DCGC-237866/2021|EPI_ISL_6934713|2021-11-21
Denmark/DCGC-280237/2021|EPI_ISL_7968336|2021-12-01
Denmark/DCGC-241071/2021|OV851853.1|2021-11-21
Denmark/DCGC-238553/2021|EPI_ISL_6938474|2021-11-20
Denmark/DCGC-215815/2021|EPI_ISL_6505011|2021-11-07
Denmark/DCGC-217653/2021|EPI_ISL_6506254|2021-11-07
Denmark/DCGC-228827/2021|EPI_ISL_6773213|2021-11-17
Denmark/DCGC-246044/2021|EPI_ISL_7193760|2021-11-21
Denmark/DCGC-244579/2021|EPI_ISL_7192782|2021-11-24
Denmark/DCGC-212258/2021|EPI_ISL_6320940|2021-11-05
Denmark/DCGC-211134/2021|EPI_ISL_6306696|2021-11-05
Denmark/DCGC-219782/2021|OV846420.1|2021-11-09
Denmark/DCGC-225371/2021|EPI_ISL_6682744|2021-11-11
Denmark/DCGC-215301/2021|EPI_ISL_6504057|2021-11-07
Denmark/DCGC-209230/2021|OV844675.1|2021-11-03
Denmark/DCGC-207878/2021|OV844867.1|2021-11-03
Denmark/DCGC-223962/2021|EPI_ISL_6681196|2021-11-14
Denmark/DCGC-221444/2021|EPI_ISL_6678715|2021-11-11
Denmark/DCGC-217489/2021|EPI_ISL_6506141|2021-11-07
Denmark/DCGC-216735/2021|EPI_ISL_6505650|2021-11-07
Denmark/DCGC-216130/2021|EPI_ISL_6505238|2021-11-07
Denmark/DCGC-215127/2021|EPI_ISL_6503370|2021-11-07
Denmark/DCGC-211860/2021|EPI_ISL_6319622|2021-11-04
Denmark/DCGC-211780/2021|EPI_ISL_6319335|2021-11-04
Denmark/DCGC-206448/2021|EPI_ISL_6182512|2021-11-03
Denmark/DCGC-210769/2021|EPI_ISL_6305925|2021-11-06
Denmark/DCGC-207215/2021|EPI_ISL_6184031|2021-11-02
Denmark/DCGC-235564/2021|OV858745.1|2021-11-18
Italy/UMB-IZSGC-19546.1.8/2022|EPI_ISL_9960758|2022-02-05
Seq1_Italy_2022-02-05
Italy/UMB-IZSGC-19546.1.29/2022|EPI_ISL_9960779|2022-02-05
Italy/UMB-IZSGC-19546.1.17/2022|EPI_ISL_9960767|2022-02-05
Italy/ABR-IZSGC-319137/2021|EPI_ISL_8309183|2021-12-17
Italy/ABR-IZSGC-318439/2021|EPI_ISL_8118990|2021-12-11
Italy/ABR-IZSGC-318401/2021|EPI_ISL_8118968|2021-12-13
Italy/ABR-IZSGC-318466/2021|EPI_ISL_7961716|2021-12-13
Italy/ABR-IZSGC-319582/2021|EPI_ISL_8309287|2021-12-22
Italy/ABR-IZSGC-197/2022|EPI_ISL_8607829|2022-01-03
Italy/ABR-IZSGC-319696/2021|EPI_ISL_8495888|2021-12-27
Italy/ABR-IZSGC-318465/2021|EPI_ISL_7961715|2021-12-13
Italy/ABR-IZSGC-318795/2021|EPI_ISL_8119028|2021-12-16
Italy/ABR-IZSGC-319594/2021|EPI_ISL_8309297|2021-12-22
Italy/LAZ-AMC-2112234079-DS/2021|EPI_ISL_8791497|2021-12-23
Italy/CAM-TIGEM-IZSM-COLLI-24439/2021|EPI_ISL_9482366|2021-11-06
Italy/ABR-CAST-K4075/2021|EPI_ISL_6315419|2021-11-02
Italy/LAZ-IFO-11193467/2021|EPI_ISL_6815856|2021-11-19
Italy/LAZ-AMC-2201041660-DS/2022|EPI_ISL_8791580|2022-01-04
0
10
20
30
40
50
Mutations

- Neighbor=Delta
- Your isolate= Delta

To see the phylogeny



view in Genome Browser
view downsampled global tree in Nextstrain
view subtree 1 in Nextstrain
view subtree 2 in Nextstrain
view subtree 3 in Nextstrain
view subtree 4 in Nextstrain
view subtree 5 in Nextstrain

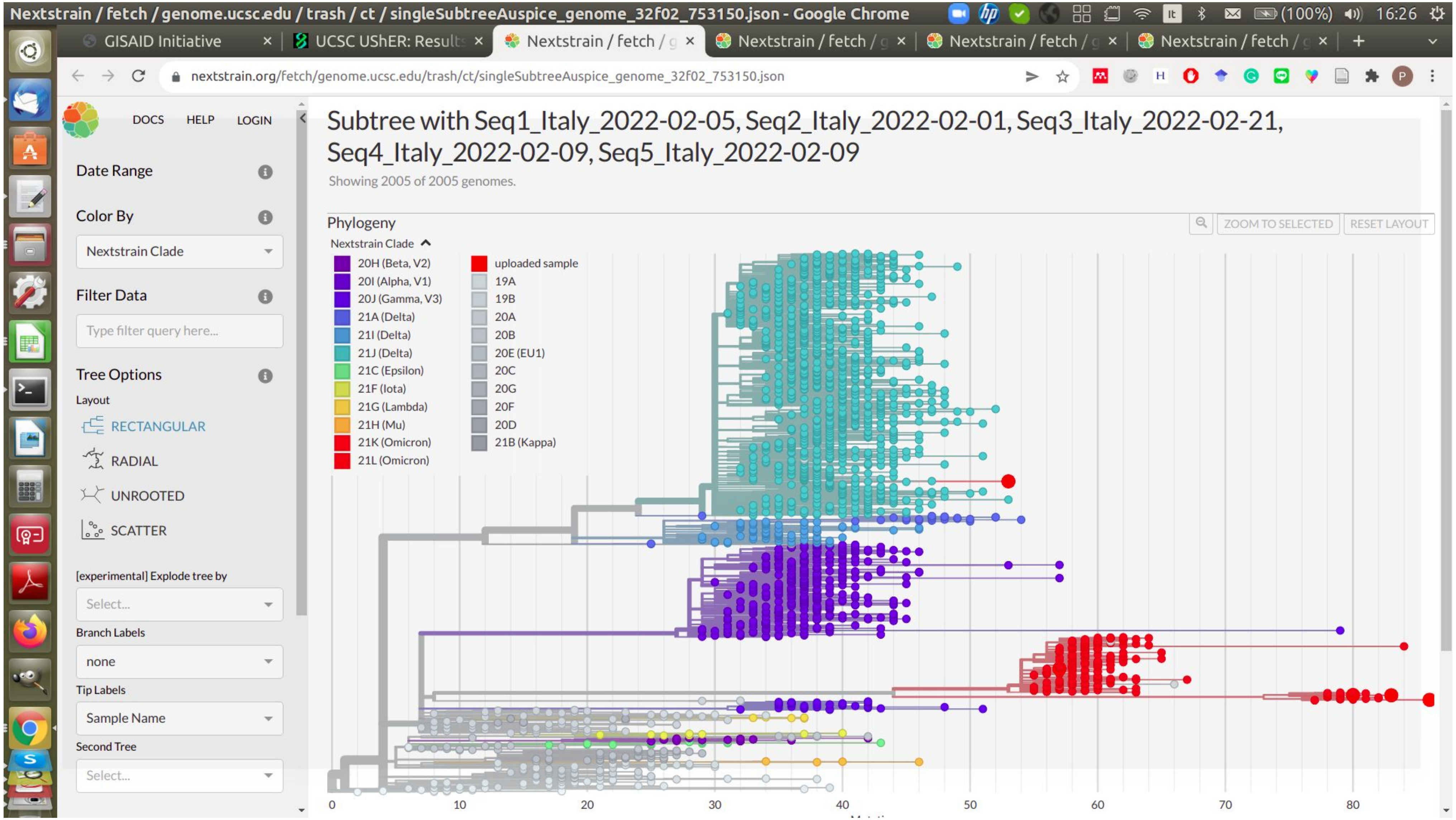
If you have metadata you wish to display, click a 'view subtree in Nextstrain' button, and then you can drag on a CSV file to add it to the tree view.

Note: The Nextstrain subtree views, and Download files below, are temporary files and will expire within two days. Please download the Nextstrain subtree JSON files if you will want to view them again in the future. The JSON files can be drag-dropped onto https://auspice.us/.

Downloads: | Global phylogenetic tree with your sequences | TSV summary of sequences and placements | TSV summary of Spike mutations | ZIP file of subtree JSON and Newick files |

Fasta Sequence	Size (?)	#Ns (?)	#Mixed (?)	Bases aligned (?)	Inserted bases (?)	Deleted bases (?)	#SNVs used for placement (?)	#Masked SNVs (?)	Nextstrain clade (?)	Pango lineage (?)	Neighboring sample in tree (?)	Lineage of neighbor (?)	#Imputed values for mixed bases (?)	#Maximally parsimonious placements (?)	Parsimony score (?)	Subtree number (?)
Seq1_Italy_2022-02-05	29889	13	0	29842 (?)	0	13 (?)	53 (?)	2 (?)	21J (Delta)	AY.125	Italy/UMB-IZSGC-19546.1.8/2022 | EPI_ISL_9960758 | 2022-02-05	AY.125	0	1	1	1 (view in Nextstrain)
Seq2_Italy_2022-02-01	29882	91	0	29758 (?)	0	21 (?)	53 (?)	1 (?)	21K (Omicron)	BA.1	USA/NY-MSHSPSP-PV45472/2021 | EPI_ISL_7908071 | 2021-12-14	BA.1	0	6	1	2 (view in Nextstrain)
Seq3_Italy_2022-02-21	29842	81	0	29761 (?)	0	27 (?)	61 (?)	2 (?)	21M (Omicron)	BA.2	Denmark/DCGC-348472/2022 | EPI_ISL_9506039 | 2022-01-27	BA.2	0	1	7	3 (view in Nextstrain)
Seq4_Italy_2022-02-09	29841	107	29 (?)	29734 (?)	0	27 (?)	53 (?)	3 (?)	21M (Omicron)	BA.2	Italy/PIE_IRCC_15879077/2022 | EPI_ISL_9775784 | 2022-01-24	BA.2	27 (?)	1	0	4 (view in Nextstrain)
Seq5_Italy_2022-02-09	29767	4454	0	25313 (?)	0	44 (?)	52 (?)	2 (?)	21M (Omicron)	BA.2	Germany/SN-RKI-I-505991/2022 | EPI_ISL_9723596 | 2022-01-17	BA.2	0	1	0	5 (view in Nextstrain)

UShER, global phylogeny



Nextstrain / fetch / genome.ucsc.edu / trash / ct / singleSubtreeAuspice_genome_32f02_753150.json - Google Chrome

GISAID Initiative | UCSC UShER: Results | Nextstrain / fetch / g | Nextstrain / fetch / g | Nextstrain / fetch / g | Nextstrain / fetch / g

nextstrain.org/fetch/genome.ucsc.edu/trash/ct/singleSubtreeAuspice_genome_32f02_753150.json

DOCS HELP LOGIN

Date Range

Color By: Nextstrain Clade

Filter Data: Type filter query here...

Tree Options

Layout: RECTANGULAR | RADIAL | UNROOTED | SCATTER

[experimental] Explode tree by: Select...

Branch Labels: none

Tip Labels: Sample Name

Second Tree: Select...

Subtree with Seq1_Italy_2022-02-05, Seq2_Italy_2022-02-01, Seq3_Italy_2022-02-21, Seq4_Italy_2022-02-09, Seq5_Italy_2022-02-09

Showing 2005 of 2005 genomes.

Phylogeny

ZOOM TO SELECTED | RESET LAYOUT

Conclusions part #2

- We can easily perform some quality assessment of SARS-CoV-2 genome sequences
- If we have a “reasonable” number, web interface based tools can be used
- In our case of study
  - all the sequences fit well within the global phylogeny
  - S4 had some ambiguous base calls. **Could be solved by UShER!**
    - /we can tell the IT guys
  - S5 has 5 Kb missing. But no errors
    - /again we can check with the IT guys
    - sequence is however informative. Resequencing an option?

CNR-IBIOM

CNR-ITB

Università di Milano

Università di Padova

Università di Roma Tor Vergata

Università di Torino

Università della Tuscia

Istituto Superiore di Sanità

Consortium GARR

Thanks!

@elixir_ita

www.elixir-europe.org

This presentation is part of a project that has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement Nos 857650 (EOSC-Pillar), 824087 (EOSC-Life), and 676559 (ELIXIR-EXCELERATE), and was supported by the Italian Computing and Data Infrastructure (ICDI).

Pavone, Gina, Lazzeri, Emma, Carta, Claudio, Chiara, Matteo, & Quaglia, Federica. (2022, March 11). Practical course on FAIR Data Stewardship in Life Science. Zenodo. https://doi.org/10.5281/zenodo.6323375.