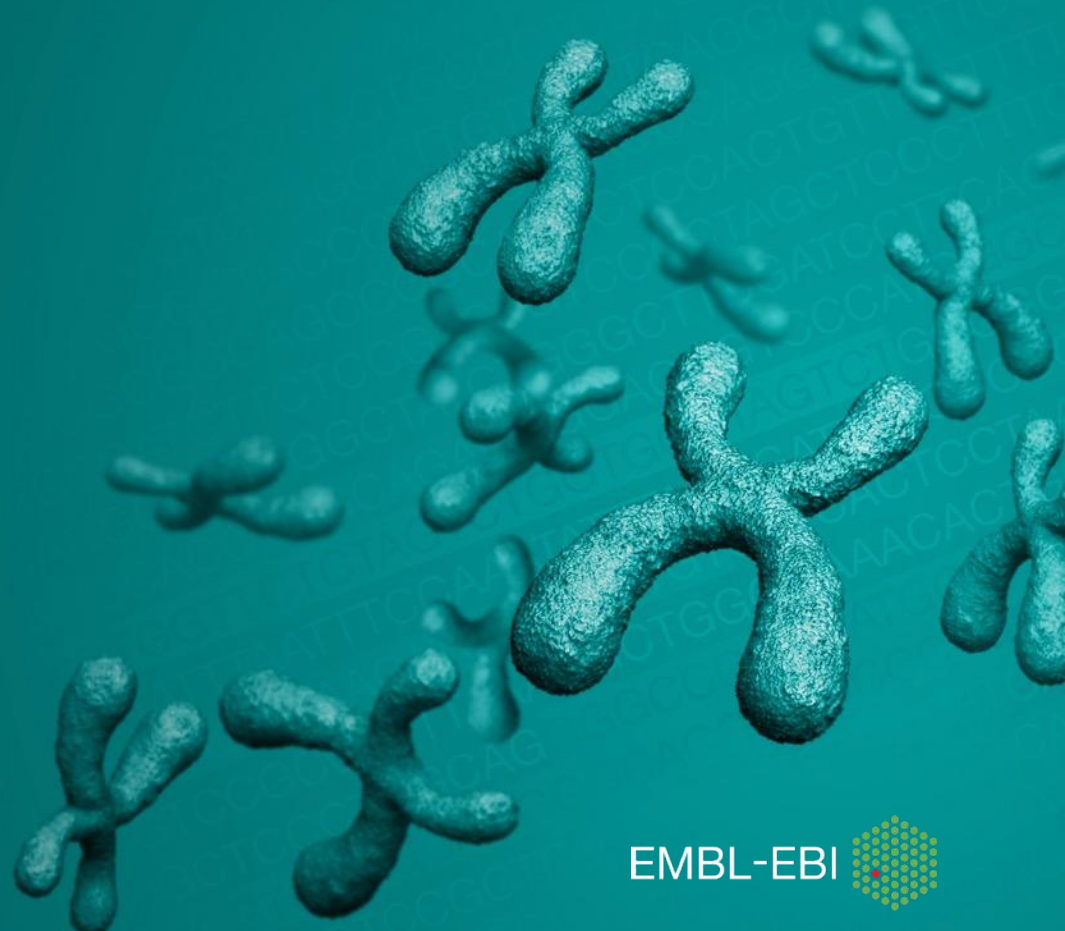
SARS-CoV-2 at the ENA (COVID-19 Data Platform)

PHA4GE Webinar 21/09/21

Colman O'Cathail

Bioinformatician, European Nucleotide Archive



ENA European Nucleotide Archive

EMBL-EBI

What We Do

- Submitting to the ENA
  - How we relate to the INSDC
  - Our metadata model
  - Submission routes
  - Helpdesk
- The European COVID-19 Data Platform (VEO partners)
  - The COVID-19 Data Portal tour
  - Systematic analyses of public data
  - Data Hubs

Submitting to ENA

What, why, how

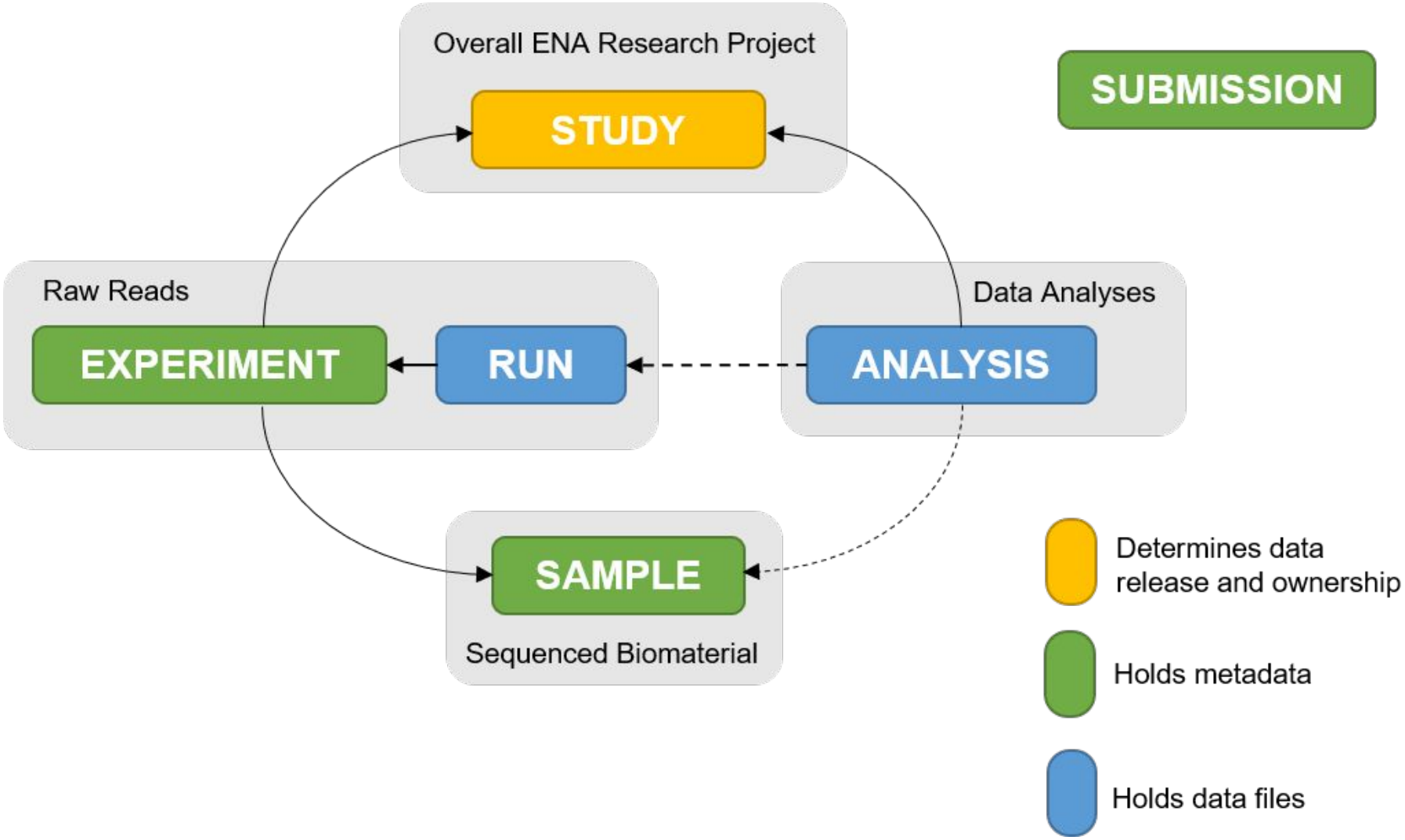
EMBL-EBI

The INSDC

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

www.insdc.org

The ENA Metadata Model

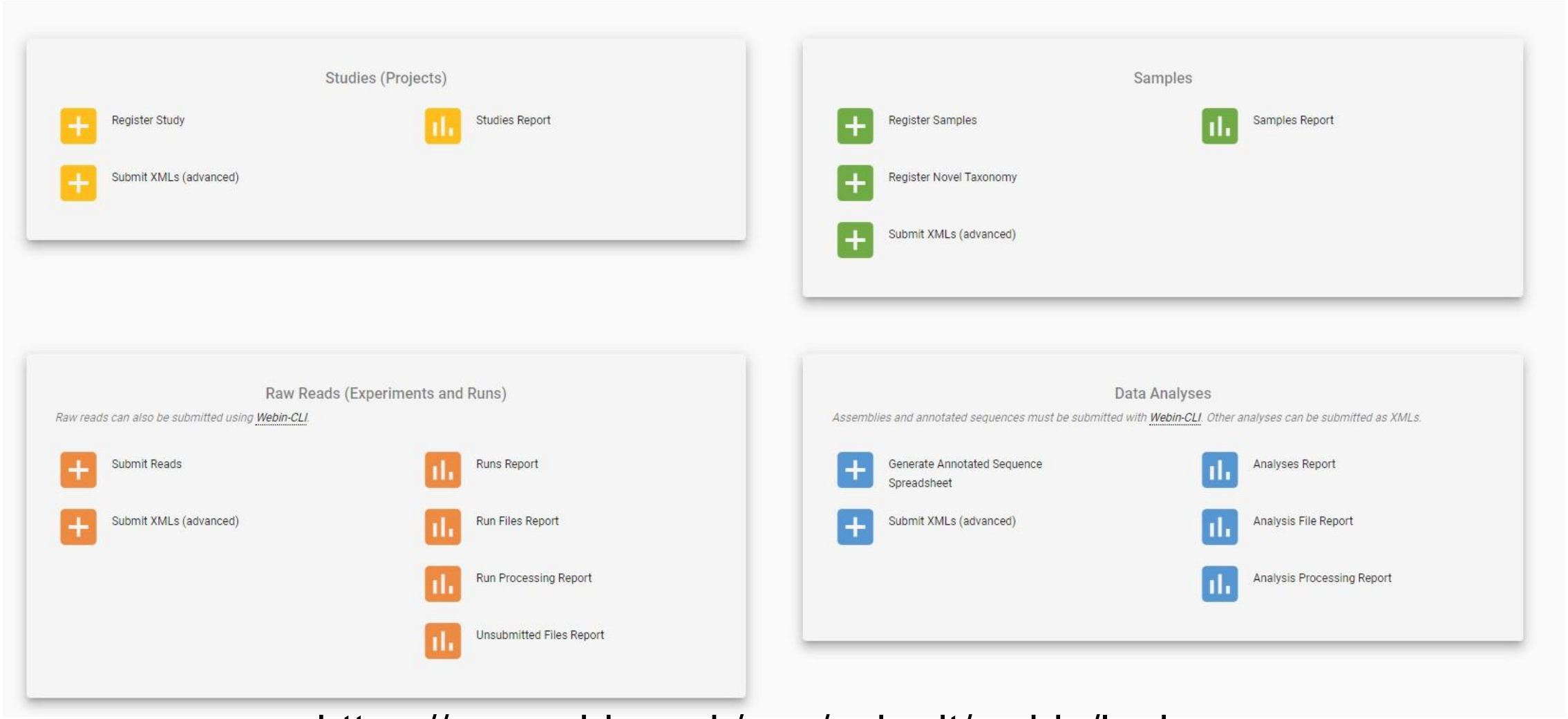


Submission methods

	Interactive	Webin-CLI	Programmatic
Study	Y	N	Y
Sample	Y	N	Y
Read data	Y	Y*	Y
Genome Assembly	N	Y	N
Transcriptome Assembly	N	Y	N
Template Sequence	Y	Y	N
Other Analyses	N	N	Y

https://ena-docs.readthedocs.io/en/latest/index.html

Interactive Mode



https://www.ebi.ac.uk/ena/submit/webin/login

Webin-CLI

	Webin-CLI
Study	N
Sample	N
Read data	Y*
Genome Assembly	Y
Transcriptome Assembly	Y
Template Sequence	Y
Other Analyses	N

https://ena-docs.readthedocs.io/en/latest/index.html

- Webin-CLI is a ja ENA

```
(base) [ocathail@hh-yoda-08-01 ocathail]$ java -jar webin-cli-3.7.0.jar -help

Usage: java -jar webin-cli-3.7.0.jar [-ascp] [-fields] [-help] [-quick]
                                     [-submit] [-test] [-validate] [-version]
                                     [-centerName=CENTER] -context=TYPE
                                     [-inputDir=DIRECTORY] -manifest=FILE
                                     [-outputDir=DIRECTORY]
                                     [-password=PASSWORD] [-passwordEnv=VAR]
                                     [-passwordFile=FILE] -userName=USER

Description:

Validate and submit files to ENA using the Webin submission service. Use the
-fields option to see supported manifest fields for all contexts or for a
specific -context. Detailed instructions are available from:
https://ena-docs.readthedocs.io/en/latest/cli.html

Options:
      -context=TYPE        Submission type: genome, transcriptome, sequence, reads,
                             taxrefset
      -manifest=FILE       Manifest text file containing file and metadata fields.
      -userName, -username=USER
                           Webin submission account name or e-mail address.
      -password=PASSWORD   Webin submission account password.
      -passwordFile=FILE   File containing the Webin submission account password.
      -passwordEnv=VAR     Environment variable containing the Webin submission
                             account password.
      -inputDir, -inputdir=DIRECTORY
                           Root directory for the files declared in the manifest
                             file. By default the current working directory is used
                             as the input directory.
      -outputDir, -outputdir=DIRECTORY
                           Root directory for any output files written in
                             <context>/<name>/<validate,process,submit> directory
                             structure. By default the manifest file directory is
                             used as the output directory. The <name> is the unique
                             name from the manifest file. The validation reports are
                             written in the <validate> sub-directory.
      -centerName, -centername=CENTER
                           Mandatory center name for broker accounts.
      -validate            Validate files without uploading or submitting them.
      -quick               Validates submitted read files (BAM, CRAM, Fastq) within
                             a fixed time period (5 minutes). All CRAM reference
                             sequence md5 checksums are always validated. When this
                             option is used files may only be partially validated
                             and may fail post-submission processing.
      -submit              Validate, upload and submit files.
      -test                Use the test submission service.
      -ascp                Use Aspera (if Aspera Cli is available) instead of FTP
                             when uploading files. The path to the installed "ascp"
                             program must be in the PATH variable.
      -help                Show this help message and exit.
      -fields              Show manifest fields for all contexts or for the given
                             -context.
      -version             Print version information and exit.
 Exit codes: 0=SUCCESS, 1=INTERNAL ERROR, 2=USER ERROR, 3=VALIDATION ERROR
(base) [ocathail@hh-yoda-08-01 ocathail]$
```

Webin-CLI – Genome Manifest Example

STUDY PRJEBXXXX

SAMPLE SAMEAXXXXX

ASSEMBLYNAME example_assembly

COVERAGE 60

PROGRAM TODO

PLATFORM TODO

MINGAPLENGTH TODO

MOLECULETYPE genomic DNA

FASTA genome.fasta.gz

Programmatic Submission

	Programmatic
Study	Y
Sample	Y
Read data	Y
Genome Assembly	N
Transcriptome Assembly	N
Template Sequence	N
Other Analyses	Y

https://ena-docs.readthedocs.io/en/latest/index.html

Programmatic Submission

- Most advanced submission method
- Requires users to create and submit XMLs directly
- Good idea to utilise support if going this route
- Can be very powerful when done right

Submitting to SARS-CoV-2 Data

Recommendations, tools & support

EMBL-EBI

SARS-CoV-2 Submissions

- Dedicated read the docs page
  - https://ena-browser-docs.readthedocs.io/en/latest/help_and_guides/sars-cov-2-submissions.html
- Dedicated support – virus-dataflow@ebi.ac.uk

SARS-CoV-2 Submissions

- Study registration as normal
- Sample registration – please select the right checklist
  - ERC000033 – ENA Virus Pathogen Reporting Standard Checklist
- We offer a tool to convert GISAID metadata into ENA metadata

  (https://github.com/enasequence/ena-content-dataflow/blob/master/scripts/gisaid_to_ena.py)

ERC000033 Sample Checklist

- Mandatory fields:

geographic location (country and/or sea)

host common name

host subject id

host health state

host sex

host scientific name

collector name

collecting institution

isolate

- Recommended fields:
  - Organism name = *Severe acute respiratory syndrome coronavirus 2*
  - Taxon ID = 2697049
  - Collection date = <*insert_here*>
  - Sample capture status = “active surveillance in response to outbreak”

Submitting reads and assemblies

- Reads can be submitted by the users preferred routes
- Assembly submission is recommended via Webin-CLI or the new SARS-CoV-2 Web API
- Don’t forget you can reference any archived runs!

SARS-CoV-2 Web API

- https://ena-browser-docs.readthedocs.io/en/latest/help_and_guides/Webin-Cli_SARS-CoV-2_Genome_Submission_REST_API.html

- Two service endpoints; Validate & Submit (test and production for both)

```
{
  *"name": "string", # Unique name for the assembly within the Webin submission account
  *"study": "string", # Study accession number or unique name (alias)
  *"sample": "string", # Sample accession number or unique name (alias)
  *"coverage": "number", # The estimated depth of sequencing coverage
  *"program": "string", # The assembly program
  *"platform": "string", # The sequencing platform
  *"sequence": "string", # The assembled genome sequence
  "description": "string", # Free text description of the genome assembly
  "minGapLength": "integer", # Minimum length of consecutive Ns to be considered a gap
  "moleculeType": "genomic DNA",
  "runRef": "string", # Run accession number containing the raw reads associated with this genome assembly
  "tpa": boolean, # Set to true for third party assemblies (by default false)
  "authors": "string" # List of authors associated with this genome assembly (by default authors of the submission account will be used)
  "address": "string", # Address where this genome was assembled (by default addrss of the submission account will be used)
  "submissionTool": "string", # Submission tool that called this endpoint
  "submissionToolVersion": "string"  # Submission tool version that called this endpoint
}
```

SARS-CoV-2 Web API

Curl Example

```
curl -X 'POST' -u Webin-N:password   \
  'https://wwwdev.ebi.ac.uk/ena/submit/webin-cli/api/v1/genome/covid-19' \
  -H 'accept: application/json' \
  -H 'Content-Type: application/json' \
  -d '{
  "name": "test_1",
  "study": "PRJEB46468",
  "sample": "ERS6670887",
  "coverage": 100,
  "program": "Ilumina",
  "platform": "Ilumina",
  "sequence": "CTCTCGATCGATCAAATTTGGGTTTAAGGCCCTTGGAATT",
  "description": "test",
  "minGapLength": 1,
  "moleculeType": "genomic DNA",
  "authors": "test",
  "address": "test"
}'
```

Python Example

```
import sys
import requests
import json

data = [
    {
      "name": "test_1", "study": "PRJEB46811", "sample": "ERS7306048",
      "coverage": 100, "program": "Ilumina", "platform": "Ilumina",
      "sequence": "CTCTCGATCGATCAAATTTGGGTTTAAGGCCCTTGGAATT",
      "description": "test", "minGapLength": 1, "moleculeType": "genomic DNA",
      "tpa": False, "authors": "test", "address": "test"
    },
    {
      "name": "test_2", "study": "PRJEB46811", "sample": "ERS7306049",
      "coverage": 100, "program": "Ilumina", "platform": "Ilumina",
      "sequence": "CTCTCGATCGATCAAATTTGGGTTTAAGGCCCTTGGAATT",
      "description": "test", "minGapLength": 1, "moleculeType": "genomic DNA",
       "authors": "test", "address": "test"
    }
]

## Please remove /validate from the URL to submit the genome instead of just validating it
server = "https://wwwdev.ebi.ac.uk/ena/submit/webin-cli/api/v1/genome/covid-19/validate"

for sample in data:
    sample_json = json.dumps(sample)
    response = requests.post(
        server, headers={"accept":"application/json", "Content-Type":"application/json"},
        data=sample_json, auth=('Webin-XXXXXX', 'password')
    )
    status = response.status_code
    if status != 200:
        print("Bad REST call : {}".format(status))
        sys.exit(1)
    else:
        receipt = json.loads(response.content)
        print("{} : {}".format(sample['name'], receipt))
```

Bulk Webin-CLI tool

- https://github.com/enasequence/ena-bulk-webincli
- Python wrapper that allows users to submit many data with one input spreadsheet
- Cotainered using Docker and Singularity
- Create manifests automatically, uploads data and handles submission
- Can be used to validate submissions also before submission

Bulk Webin-CLI tool

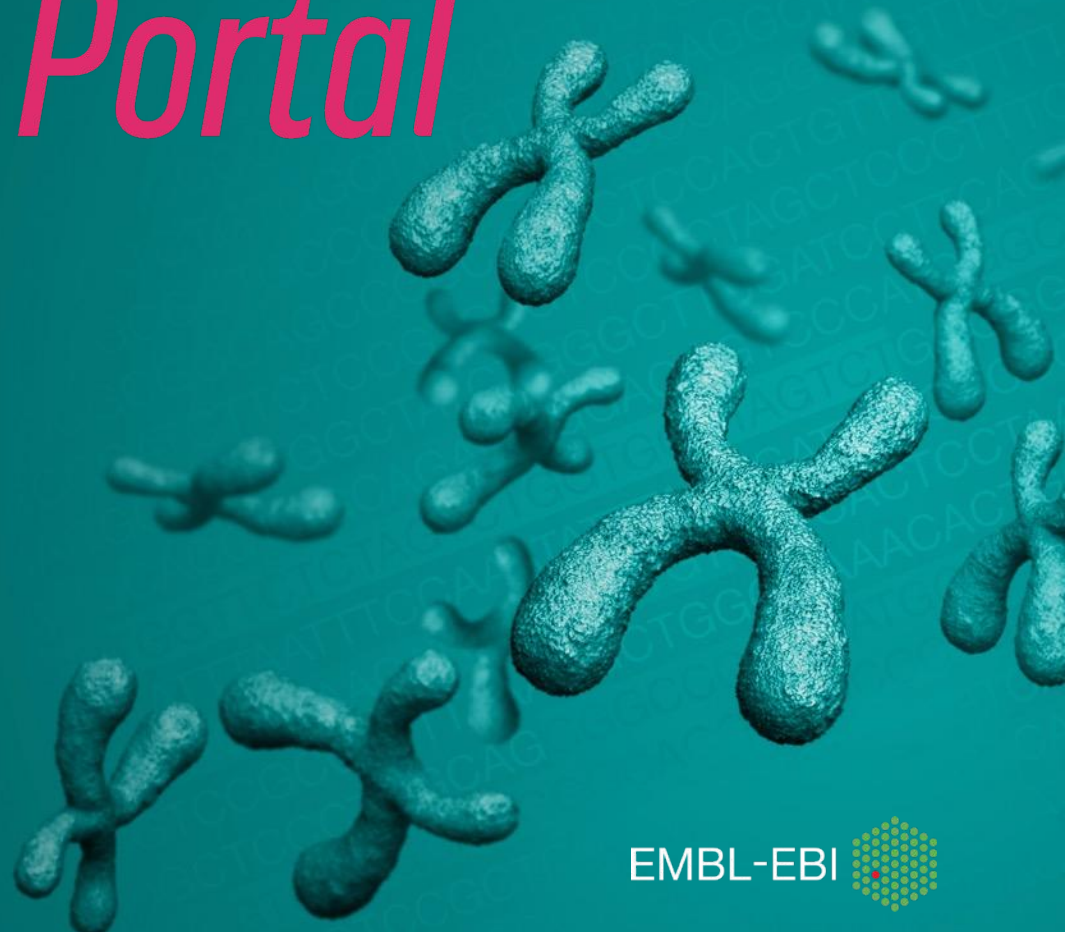
- https://github.com/enasequence/ena-bulk-webincli
- Python wrapper that allows users to submit many data with one input spreadsheet
- Cotainered using Docker and Singularity
- Create manifests automatically, uploads data and handles submission
- Can be used to validate submissions also before submission

ORCID Data Claiming

- https://www.ebi.ac.uk/ebisearch/orcidclaimdocumentation.ebi

- Allows researchers to claim credit for your deposited public data records

COVID-19 Data Portal



EMBL-EBI

The who behind the portal

DTU

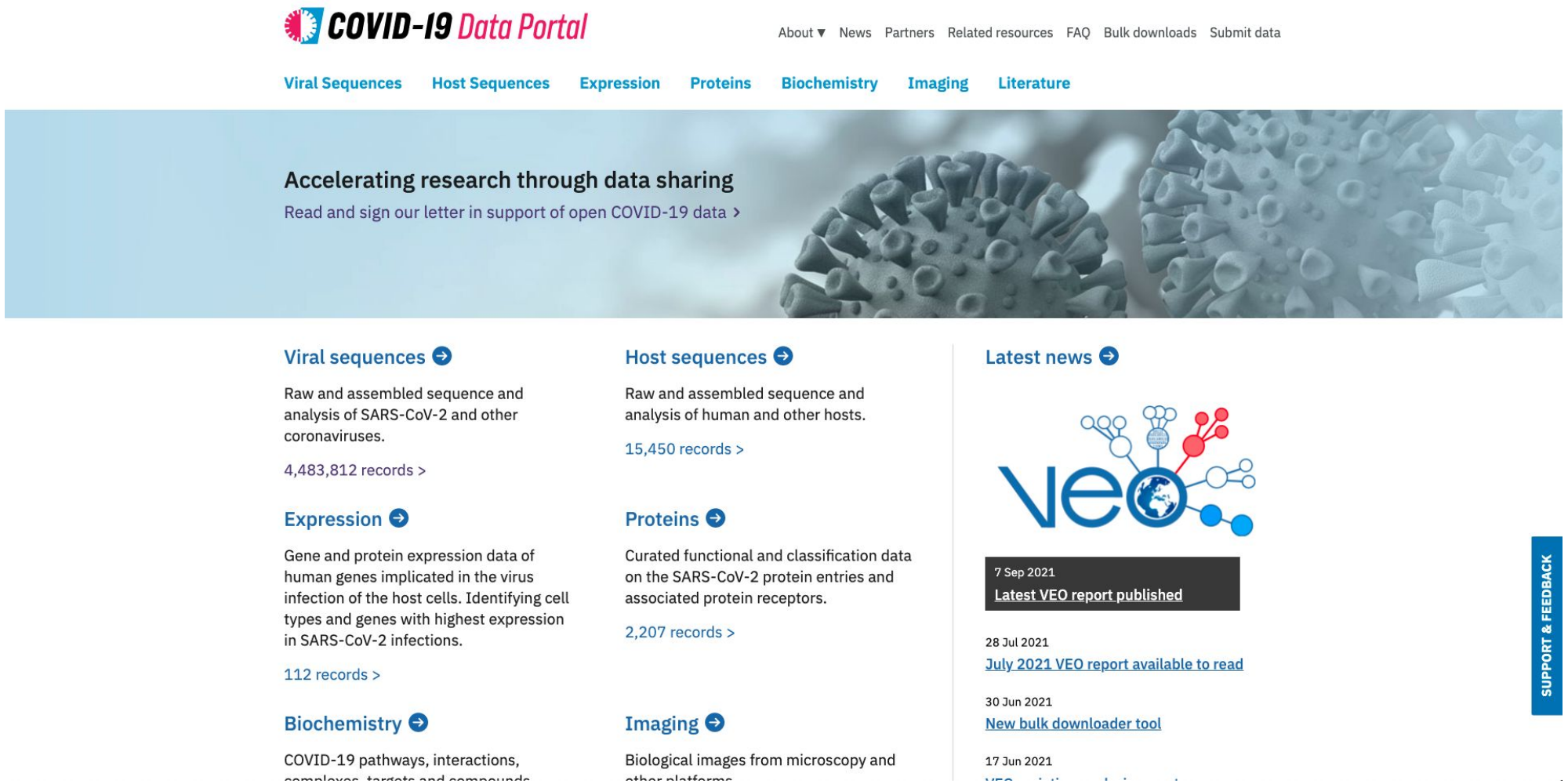
Erasmus MC

Eötvös Loránd University

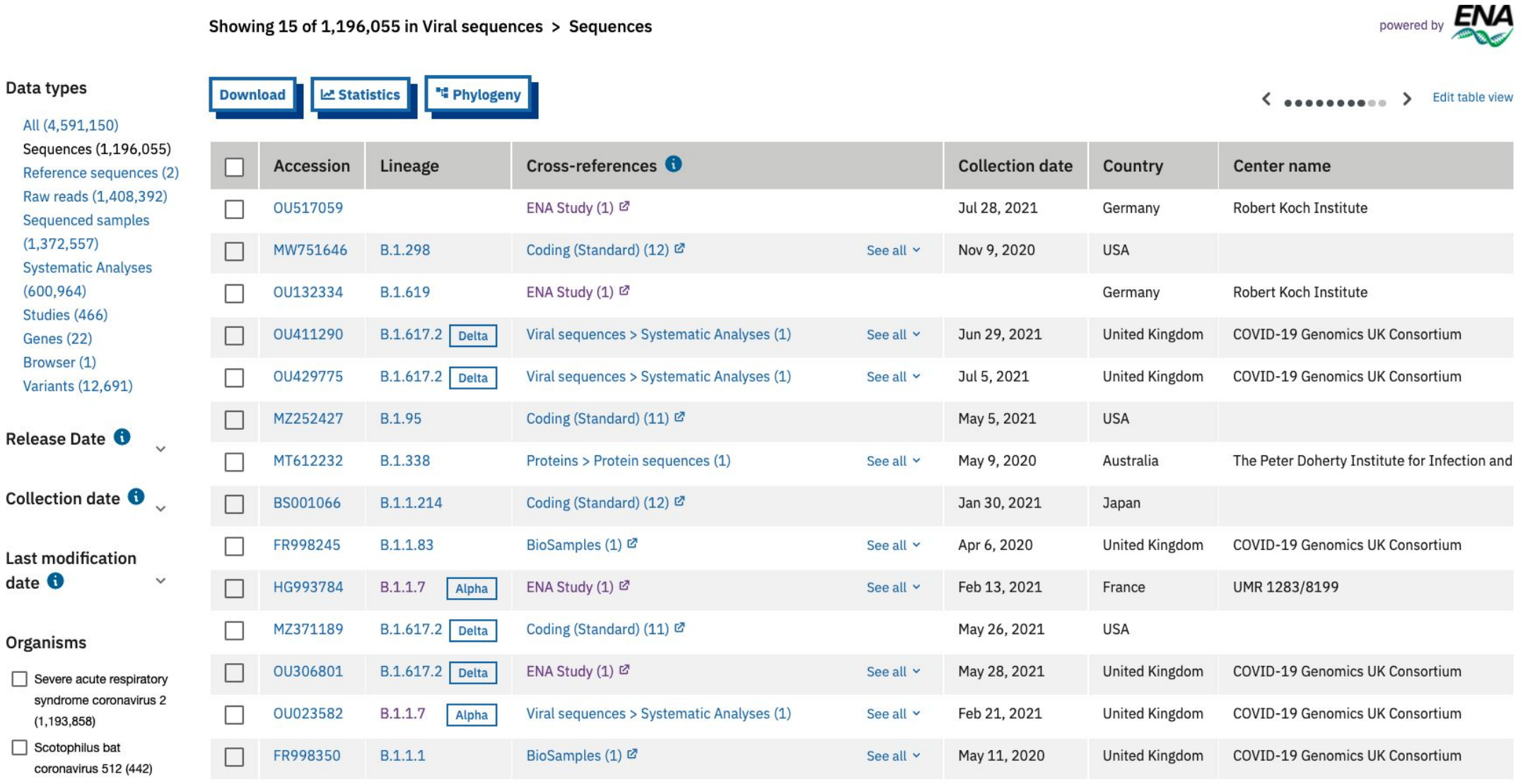
EMBL-EBI

elixir

Portal Home


COVID-19 Data Portal
About ▾ News Partners Related resources FAQ Bulk downloads Submit data
Viral Sequences Host Sequences Expression Proteins Biochemistry Imaging Literature
Accelerating research through data sharing
Read and sign our letter in support of open COVID-19 data >
Viral sequences
Raw and assembled sequence and analysis of SARS-CoV-2 and other coronaviruses.
4,483,812 records >
Host sequences
Raw and assembled sequence and analysis of human and other hosts.
15,450 records >
Expression
Gene and protein expression data of human genes implicated in the virus infection of the host cells. Identifying cell types and genes with highest expression in SARS-CoV-2 infections.
112 records >
Proteins
Curated functional and classification data on the SARS-CoV-2 protein entries and associated protein receptors.
2,207 records >
Biochemistry
COVID-19 pathways, interactions,
Imaging
Biological images from microscopy and
Latest news
VEO
7 Sep 2021
Latest VEO report published
28 Jul 2021
July 2021 VEO report available to read
30 Jun 2021
New bulk downloader tool
17 Jun 2021
SUPPORT & FEEDBACK

Data exploration



Showing 15 of 1,196,055 in Viral sequences > Sequences

powered by ENA

Download | Statistics | Phylogeny

Edit table view

Data types

- All (4,591,150)
- Sequences (1,196,055)
- Reference sequences (2)
- Raw reads (1,408,392)
- Sequenced samples (1,372,557)
- Systematic Analyses (600,964)
- Studies (466)
- Genes (22)
- Browser (1)
- Variants (12,691)

Release Date

Collection date

Last modification date

Organisms

- Severe acute respiratory syndrome coronavirus 2 (1,193,858)
- Scotophilus bat coronavirus 512 (442)

Accession	Lineage	Cross-references		Collection date	Country	Center name
OU517059		ENA Study (1)		Jul 28, 2021	Germany	Robert Koch Institute
MW751646	B.1.298	Coding (Standard) (12)	See all	Nov 9, 2020	USA	
OU132334	B.1.619	ENA Study (1)			Germany	Robert Koch Institute
OU411290	B.1.617.2 Delta	Viral sequences > Systematic Analyses (1)	See all	Jun 29, 2021	United Kingdom	COVID-19 Genomics UK Consortium
OU429775	B.1.617.2 Delta	Viral sequences > Systematic Analyses (1)	See all	Jul 5, 2021	United Kingdom	COVID-19 Genomics UK Consortium
MZ252427	B.1.95	Coding (Standard) (11)		May 5, 2021	USA	
MT612232	B.1.338	Proteins > Protein sequences (1)	See all	May 9, 2020	Australia	The Peter Doherty Institute for Infection and
BS001066	B.1.1.214	Coding (Standard) (12)		Jan 30, 2021	Japan	
FR998245	B.1.1.83	BioSamples (1)	See all	Apr 6, 2020	United Kingdom	COVID-19 Genomics UK Consortium
HG993784	B.1.1.7 Alpha	ENA Study (1)	See all	Feb 13, 2021	France	UMR 1283/8199
MZ371189	B.1.617.2 Delta	Coding (Standard) (11)		May 26, 2021	USA	
OU306801	B.1.617.2 Delta	ENA Study (1)	See all	May 28, 2021	United Kingdom	COVID-19 Genomics UK Consortium
OU023582	B.1.1.7 Alpha	Viral sequences > Systematic Analyses (1)	See all	Feb 21, 2021	United Kingdom	COVID-19 Genomics UK Consortium
FR998350	B.1.1.1	BioSamples (1)	See all	May 11, 2020	United Kingdom	COVID-19 Genomics UK Consortium

SUPPORT & FEEDBACK

EMBL-EBI

Data exploration

Organisms

- Severe acute respiratory syndrome coronavirus 2 (1,193,819)
- Scotophilus bat coronavirus 512 (442)
- Murine coronavirus (407)

More... >

Center name

- COVID-19 Genomics UK Consortium (525,489)
- Robert Koch Institute (126,189)
- CDC-OAMD (47,470)

More... >

Country

- United Kingdom (527,776)
- USA (458,557)
- Germany (126,361)

More... >

Coverage %

1195801 in this range

Apply

☐	FR995092	B.1.1.29	BioSamples (1)	See all	Apr 4, 2020	United Kingdom	COVID-19 Genomics UK Consortium
☐	FR990449	B.1.160	BioSamples (2)	See all	Jan 29, 2021	Switzerland	Swiss Pathogen Surveillance Platform (S
☐	OD944682	B.1.1.7 Alpha	Viral sequences > Systematic Analyses (1)	See all	Jan 10, 2021	United Kingdom	COVID-19 Genomics UK Consortium
☐	OU377872	B.1.617.2 Delta	BioSamples (1)	See all	Jun 26, 2021	United Kingdom	COVID-19 Genomics UK Consortium
☐	OU113468	B.1.1.7 Alpha	BioSamples (2)	See all	Mar 24, 2021	Germany	Robert Koch Institute
☐	MZ187102	B.1.1.7 Alpha	Coding (Standard) (12)		Apr 15, 2021	USA	
☐	FR994316	B.1.1.220	BioSamples (1)	See all	Apr 5, 2020	United Kingdom	COVID-19 Genomics UK Consortium
☐	FR997260	B.1.1.279	BioSamples (1)	See all	Apr 18, 2020	United Kingdom	COVID-19 Genomics UK Consortium
☐	FR995150	B.1	BioSamples (1)	See all	Apr 2, 2020	United Kingdom	COVID-19 Genomics UK Consortium
☐	MZ232211	B.1.429			Mar 22, 2021	USA	
☐	OU556889				Aug 10, 2021	United Kingdom	COVID-19 Genomics UK Consortium
☐	OD991475	B.1.1.7 Alpha			Feb 15, 2021	United Kingdom	COVID-19 Genomics UK Consortium
☐	OU103567	B.1.1.7 Alpha			Mar 16, 2021	United Kingdom	COVID-19 Genomics UK Consortium
☐	MZ342566	B.1.1.7 Alpha			May 24, 2021	USA	
☐	MZ152854	B.1.1.7 Alpha			May 1, 2021	USA	Minnesota Department of Health
☐	MW495237	B.1			Jan 5, 2021	USA	Quest Diagnostics
☐	MW749952	B.1.517			Jan 6, 2021	USA	Broad Institute of MIT and Harvard
☐	MW838286	B.1			Jan 18, 2021	USA	
☐	OU263732	B.1.1.7 Alpha			Apr 21, 2021	United Kingdom	COVID-19 Genomics UK Consortium

Data exploration

Instrument platform

- ILLUMINA (1,187,210)
- OXFORD_NANOPORE (147,740)
- PACBIO_SMRT (140,255)

Library selection

- PCR (1,097,326)
- RT-PCR (323,582)
- unspecified (32,282)

More... >

Library strategy

- AMPLICON (1,421,244)
- WGA (24,197)
- WGS (17,582)

Instrument model

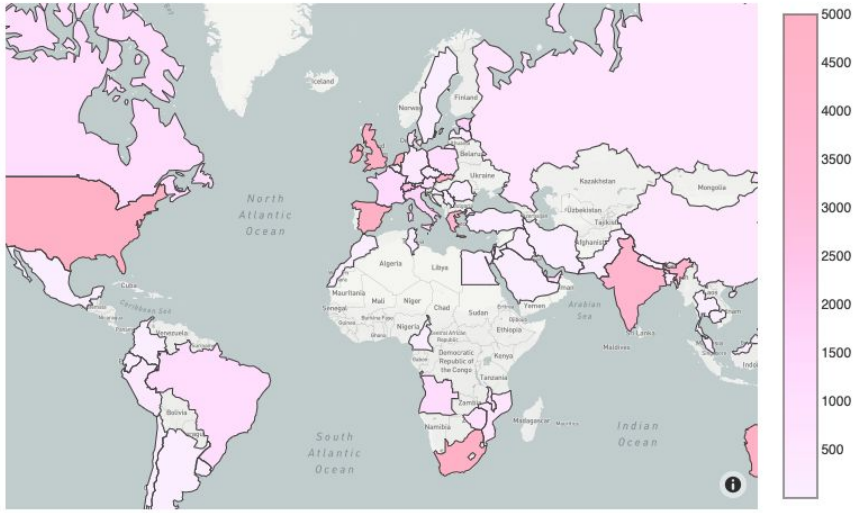
- Illumina NovaSeq 6000 (831,870)
- Sequel II (140,251)
- Illumina MiSeq (138,722)

More... >

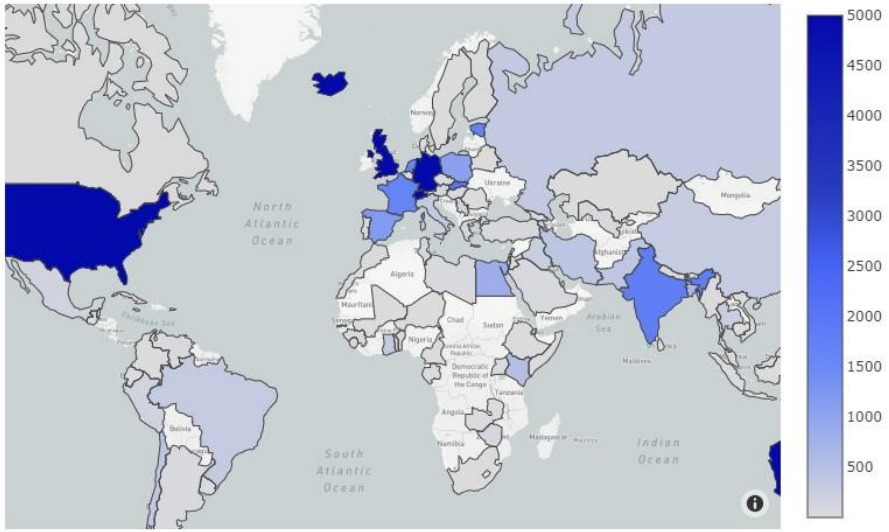
☐	DRX276349	BioSamples (1)	See all	NextSeq 550 paired end sequencing of SAMD00280768	Japa
☐	DRX276352	BioSamples (1)	See all	NextSeq 550 paired end sequencing of SAMD00280771	Japa
☐	ERX5962095	BioSamples (1)	See all	Illumina NovaSeq 6000 paired end sequencing; COG-UK/ALDP-18D5E87/SANG:210714_A00495_0594_AHFTYKDRXY/1t47	Unit
☐	DRX276354	BioSamples (1)	See all	NextSeq 550 paired end sequencing of SAMD00280773	Japa
☐	ERX5962097	BioSamples (1)	See all	Illumina NovaSeq 6000 paired end sequencing; COG-UK/QEUH-18D71B9/SANG:210714_A00495_0594_AHFTYKDRXY/1t50	Unit
☐	DRX276356	BioSamples (1)	See all	NextSeq 550 paired end sequencing of SAMD00280775	Japa
☐	ERX5962098	BioSamples (1)	See all	Illumina NovaSeq 6000 paired end sequencing; COG-UK/QEUH-18D708F/SANG:210714_A00495_0594_AHFTYKDRXY/1t52	Unit
☐	DRX276358	BioSamples (1)	See all	NextSeq 550 paired end sequencing of SAMD00280777	Japa
☐	DRX276359	BioSamples (1)	See all	NextSeq 550 paired end sequencing of SAMD00280778	Japa
☐	ERX5962105	BioSamples (1)	See all	Illumina NovaSeq 6000 paired end sequencing; COG-UK/PLYM-18D784B/SANG:210714_A00495_0594_AHFTYKDRXY/1t68	Unit
☐	DRX276362	BioSamples (1)	See all	NextSeq 550 paired end sequencing of SAMD00280781	Japa
☐	DRX276363	BioSamples (1)	See all	NextSeq 550 paired end sequencing of SAMD00280782	Japa
☐	ERX5962107	BioSamples (1)	See all	Illumina NovaSeq 6000 paired end sequencing; COG-UK/PLYM-18D76AB/SANG:210714_A00495_0594_AHFTYKDRXY/1t70	Unit
☐	DRX276365	BioSamples (1)	See all	NextSeq 550 paired end sequencing of SAMD00280784	Japa
☐	ERX5962108	BioSamples (1)	See all	Illumina NovaSeq 6000 paired end sequencing; COG-UK/ALDP-18D5DF3/SANG:210714_A00495_0594_AHFTYKDRXY/1t71	Unit
☐	DRX276366	BioSamples (1)	See all	NextSeq 550 paired end sequencing of SAMD00280785	Japa

Improved Stats Page

Raw sequences submitted by Country



Nucleotide sequences submitted by Country



Country	Sequences submitted	Raw sequences submitted
United Kingdom	527,776	886,601
USA	458,557	385,978
Germany	126,362	320
Switzerland	41,050	2,473
Australia	13,311	15,629
Iceland	5,365	N/A

Download summary data (.csv)

Summary data columns: Sequences submitted, Raw Sequences submitted

SUPPORT & FEEDBACK

Systematic Analysis

- Workflows automatically pick up SARS-CoV-2 raw sequencing reads
- Variant calling, filtered variant calls & consensus sequences

This data is produced as part of the VEO project. For more information, see the ENA project.

Showing 4 of 64 in Viral sequences > Systematic Analyses

powered by ENA

Data types

All (4,591,150)
Sequences (1,196,055)
Reference sequences (2)
Raw reads (1,408,392)
Sequenced samples (1,372,557)
Systematic Analyses (600,964)
Studies (466)
Genes (22)
Browser (1)
Variants (12,691)

Download | Statistics | Phylogeny

Edit table view

Run accession	Sample accession	Collection date	Country	Full VCF	Filtered VCF	Consensus Sequences
ERR6134035	ERS7006985	Jan 19, 2021	Germany	Complete	Complete	Pending
ERR6134036	ERS7006986	Jan 18, 2021	Germany	Complete	Complete	Pending
ERR6134037	ERS7015866	Jan 22, 2021	Germany	Complete	Complete	Pending
SRR13649823	SRS8211031	Jun 1, 2020	Germany	Complete	Pending	Pending

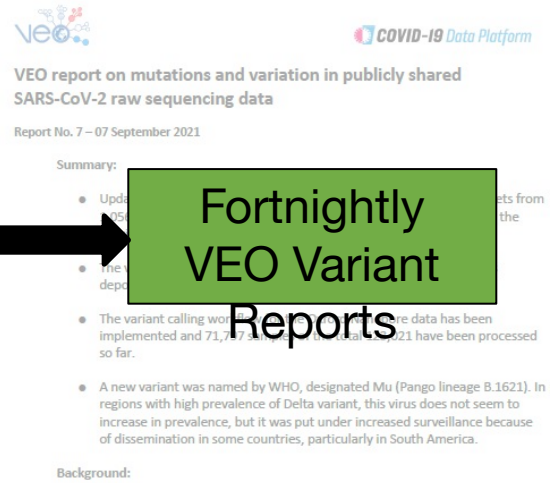
Showing 15 results

Previous | Page 5 of 5 | Next

Pipelines

- **COVID-19 Sequence Analysis Workflow (ELTE)**
  - Processing Illumina raw reads
  - Output (per sample):
    - Unfiltered VCF
    - Filtered VCF (cutoff AF=0.25)
    - Consensus sequence
- **Nanopore Analysis Workflow (EMC)**
  - Processing Nanopore raw reads
  - Output (per sample):
    - Unfiltered VCF

Fortnightly Snapshots → Fortnightly VEO Variant Reports


VEO report on mutations and variation in publicly shared SARS-CoV-2 raw sequencing data

Report No. 7 – 07 September 2021

Summary:

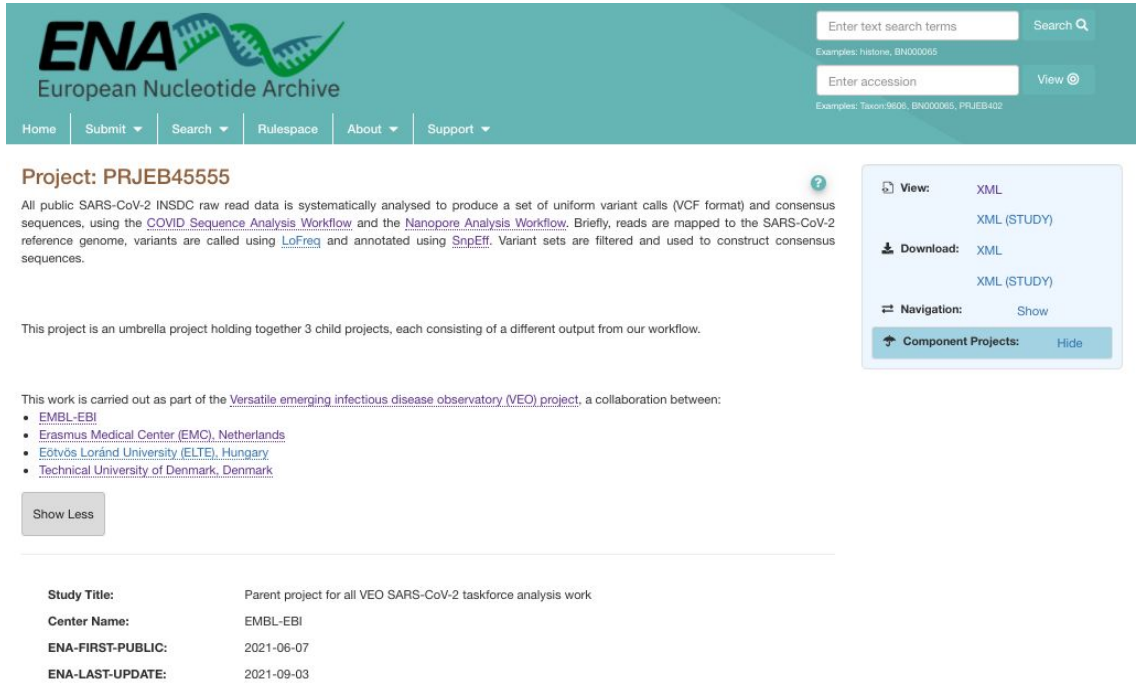
- A new variant was named by WHO, designated Mu (Pango lineage B.1621). In regions with high prevalence of Delta variant, this virus does not seem to increase in prevalence, but it was put under increased surveillance because of dissemination in some countries, particularly in South America.

Background:

Fortnightly Snapshots and Archival

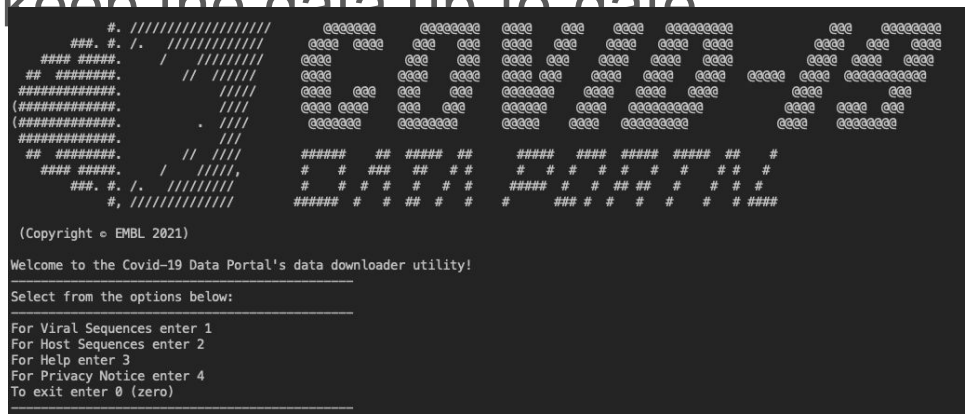
- Latest: 14th Snapshot
  - Analysed: 601,136 samples

- Archival:
  - Analysis umbrella project: PRJEB45555
    - Unfiltered VCF analysis - PRJEB43947
    - Filtered variant calls – PRJEB45554
    - Consensus sequences – PRJEB45619

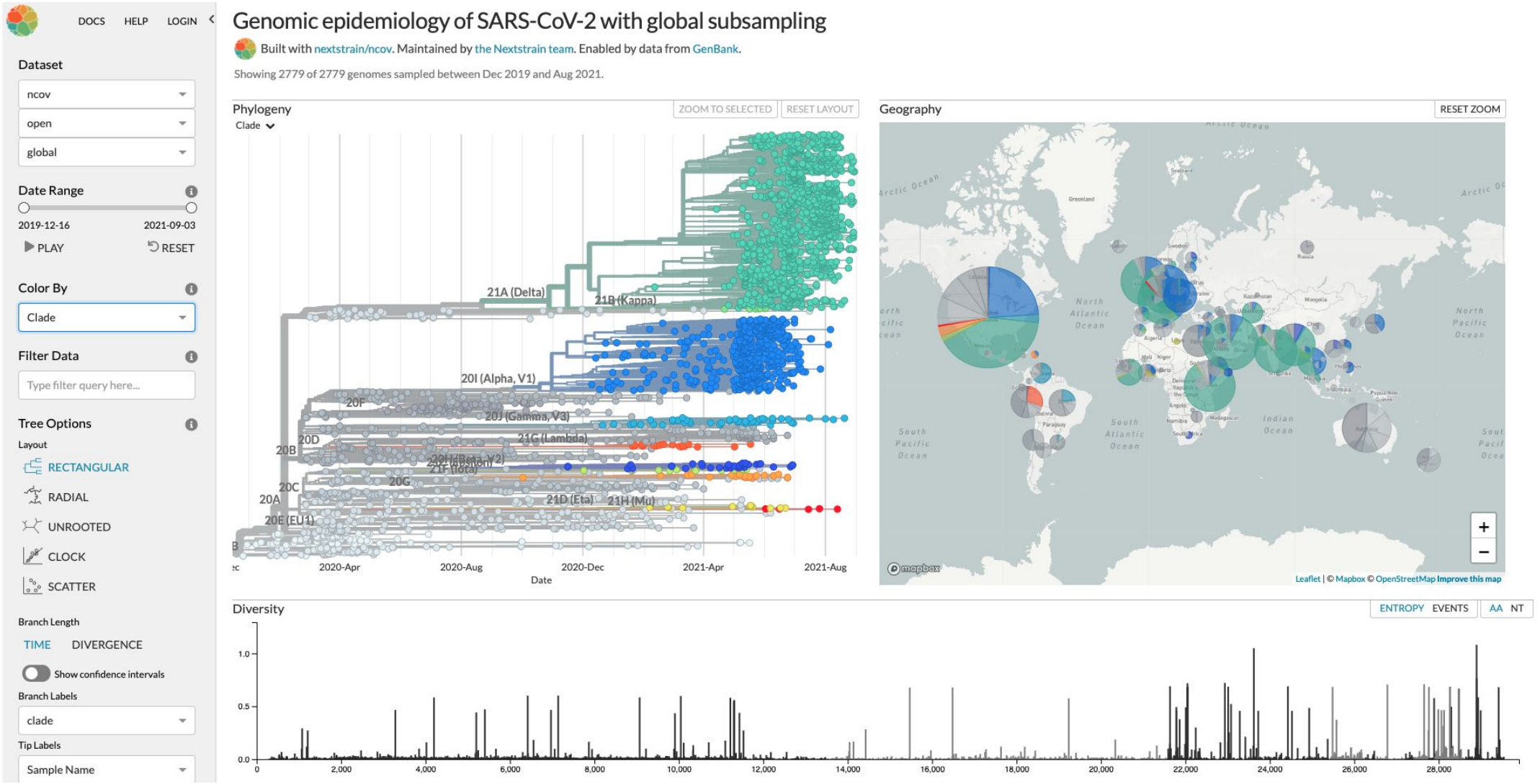


Bulk Downloader

- Utility based on Java8 to download directly from ENA FTP
- Downloads data in the following formats:
  - XML, FASTA, EMBL, FASTQ, SUBMITTED
- Download accessions
- Download data using FTP or Aspera
- Create scripts to easily download data and keep the data up to date



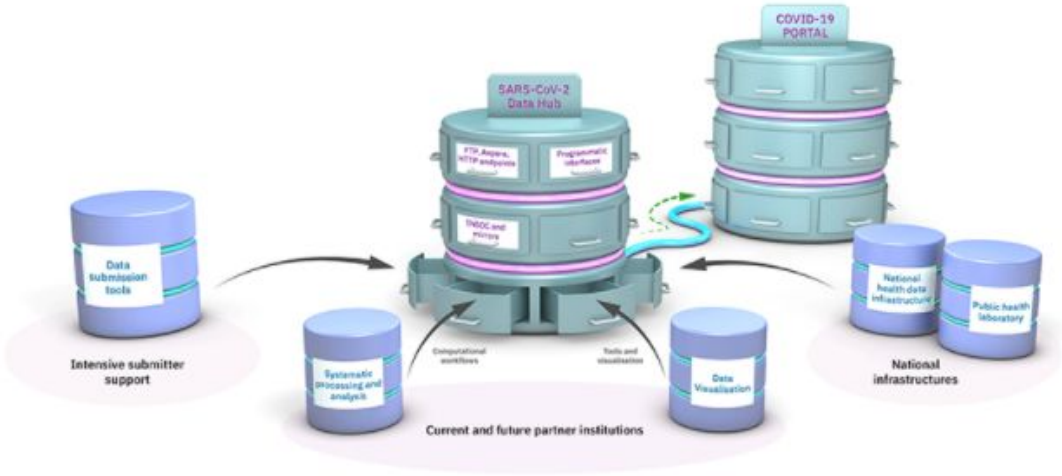
NextStrain Visualisation



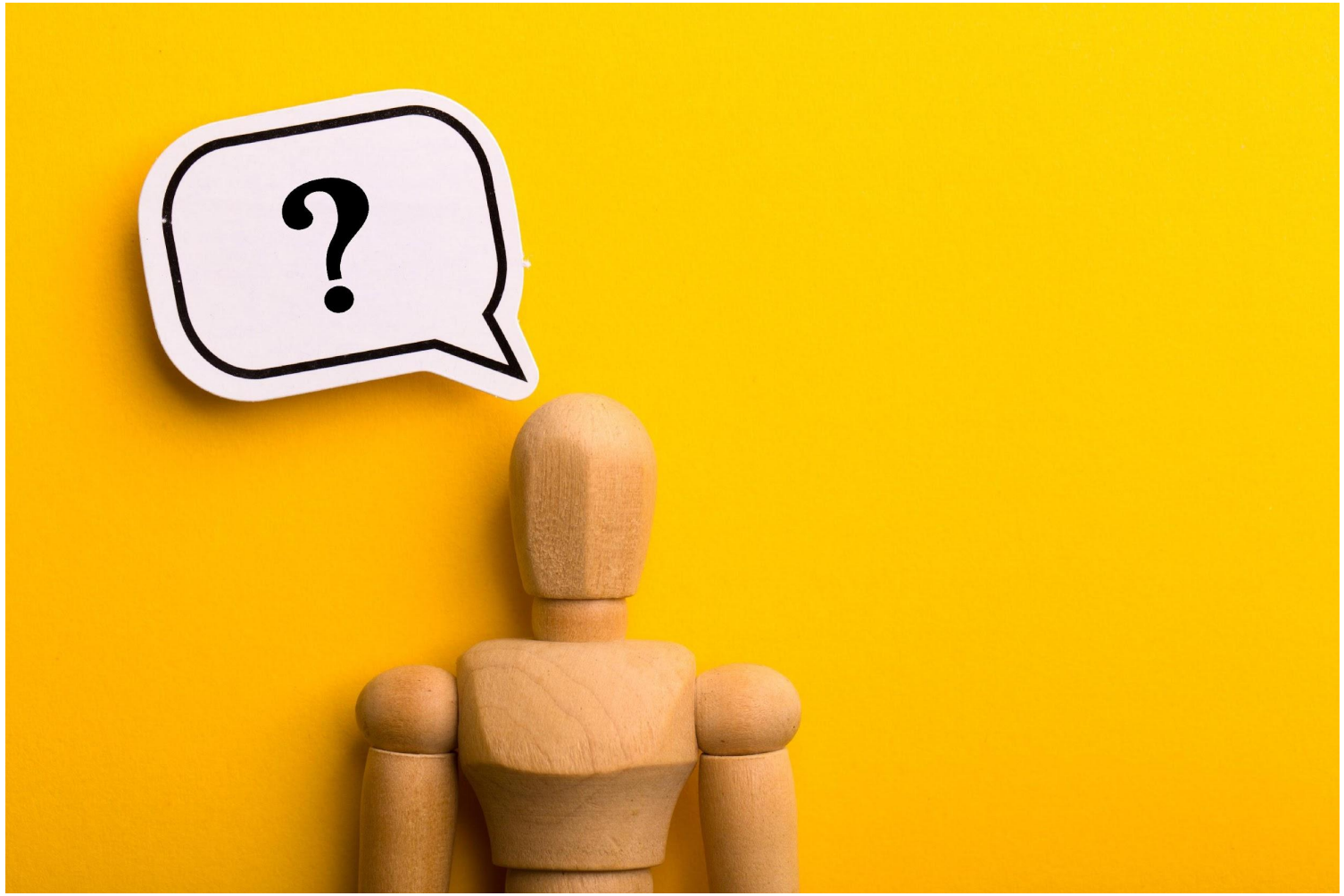
EMBL-EBI

Private Data Hubs

- Reference-based mapping workflows (CSAW, NAW)
- Evergreen phylogenetic tree (without visualisation)
- Nextstrain reports (including phylogeny and other plots)
- Lineage classifications



Questions?



ocathail@ebi.ac.uk

virus-dataflow@ebi.ac.uk