Training in genomic Epidemiology and Public Health Bioinformatics
"Bridging the Gap"

Day 9: Introduction to sequence databases and data sharing

2025-01-22

Objectives

Specific objectives of this session:

1. Understand the importance of real-time sharing of sequence data nationally, internationally and across sectors
2. Know the principles of fair and sustainable data sharing (Nagoya protocol)
3. Understand the specificities of restricted-access (e.g. GISAID) and public access sequence databases (e.g. ENA, SRA)
4. Learn about how data are structured at the INSDC (ENA) and learn about the different steps involved in data preparation and submission to the ENA database
5. Learn how to explore and retrieve data from ENA and associated data portals

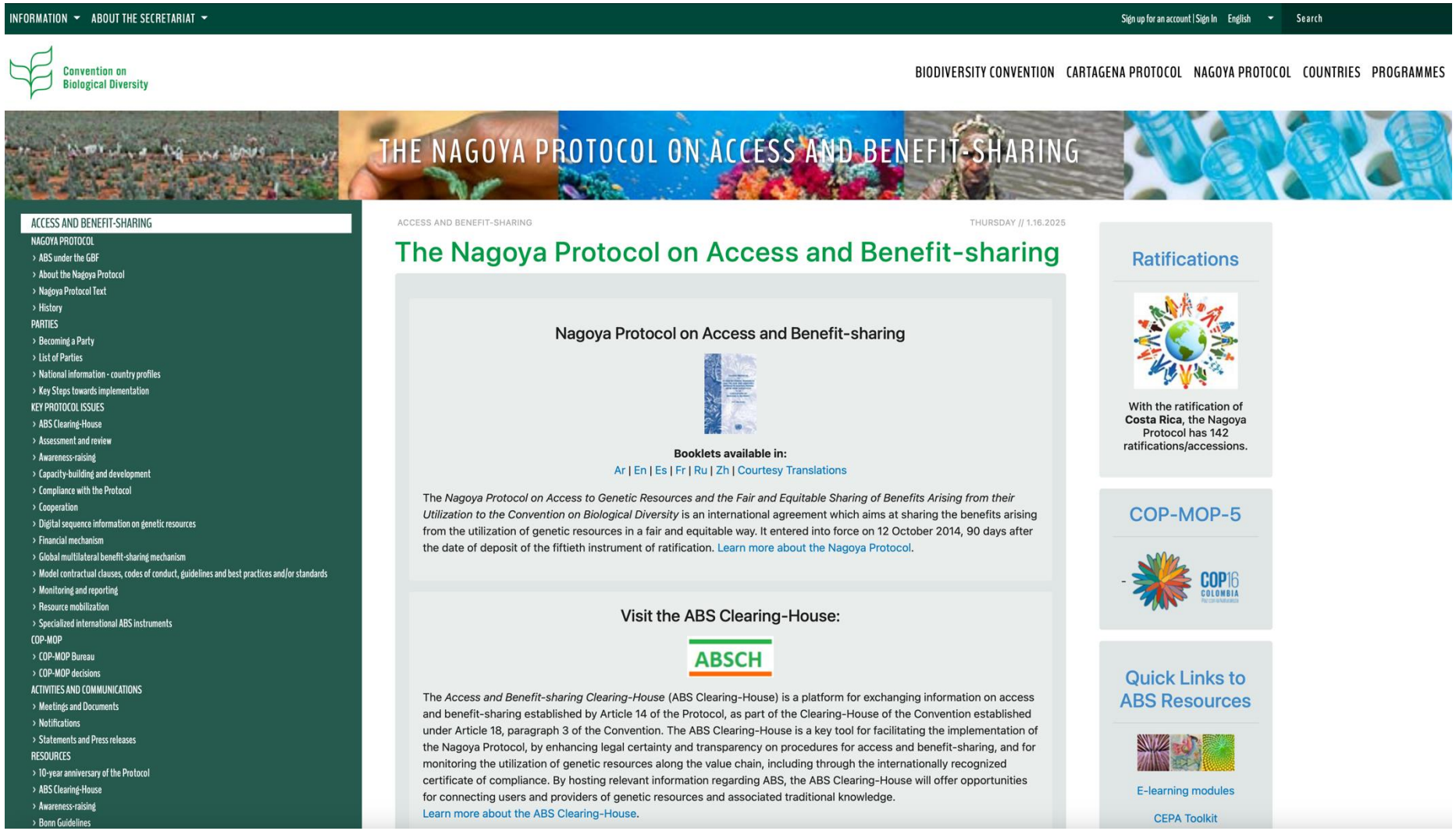
Global sharing of digital sequence information (DSI)

What does the term Digital Sequence Information include?

- Sequences and chemical structures (DNA, RNA, protein, epigenetic modifications, metabolites and other macromolecules)
- Associated information (annotation, traditional knowledge, place of origin, etc.)

Sharing of data implies guidelines and legal framework to protect the providers of these data and their stakeholders.

The Nagoya protocol on access and benefit-sharing



Entered into force on 12 October 2014

142 ratifiations

https://www.cbd.int/

Recommendation issued after the Fourth meeting Working Group on the Post-2020 Global Biodiversity Framework (21-26 June 2022)

[*Agrees* as follows:][1]

(a) [Take measures to] encourage more deposits of data;

(b) Use of tags indicating the [country [or region][or place] of] origin of [and providing] the genetic resources from which digital sequence information was generated for new submissions to [and existing digital sequence information in] [public][all] databases;

(c) Provide legal certainty and clarity for providers [of genetic resources from which digital sequence information on genetic resources is [obtained][generated]] and [for] users of [that] digital sequence information on genetic resources;

(d) Be efficient, feasible and practical [[, be][and] effective in [ensuring][enabling appropriate access to and] fair and equitable sharing of benefits arising out of the use of digital sequence information on genetic resources] and generate more benefits, including both monetary and non-monetary, than costs;

(e) Be adaptable to future technology changes;

(f) [A solution on fair and equitable sharing of benefits from the [utilization][use] of digital sequence information should] be mutually supportive of [and adaptable to] other [relevant] access and benefit-sharing instruments;

(g) [Urge Parties to take actions to promote][Promotion of] research and innovation and technical and scientific cooperation, capacity-building and technology transfer [to developing countries [under fair and most favourable terms][as specified in Article[s] 16 [and 18] of the Convention][upon mutually agreed terms]] [and increased mobilization of resources] for the purpose of conservation and sustainable use of biodiversity;

(h) [Respect and protect] the rights of indigenous peoples and local communities over their traditional knowledge associated with genetic resources [and take into account their role as stewards of biocultural, biological and genetic diversity;]

(i) [Recognizes that] the monetary and non-monetary benefits arising from the use of digital sequence information on genetic resources should be used to support conservation and sustainable use of biodiversity and [inter alia] benefit indigenous peoples and local communities;

(j) [Recognizes that] the monetary [and][or] non-monetary benefits arising from the use of digital sequence information on genetic resources [must be shared in a fair and equitable way and][that are shared] should be used to support conservation and sustainable use of biodiversity [as well as sustainable development] and [inter alia] benefit indigenous peoples and local communities[, as applicable];

(k) [[Agrees that] "digital sequence information [on genetic resources]" [is constituted of] [information on][sequences and chemical structures on][annotated sequences of] [DNA, RNA, [proteins, epigenetic modifications,[2] metabolites,] [and other macromolecules, [derivatives]] and recognizes the relevance of associated information [particularly traditional knowledge]];]

(l) ["Digital sequence information" is any information in [electronic][any] format that results from "utilization of genetic resources";]

(m) [Any solution on digital sequence information on genetic resources needs in principle to lie within the legal framework of the Convention. Solutions which lie outside the scope of the Convention on Biological Diversity would first require revision of the Convention;]

(n) [Access to [pooled][pools of diverse] digital sequence information on genetic [resources][diversity] in public databases supports research and innovation and therefore remains open [and unrestricted] [as per current [best [available] scientific] practices [and international standards]], [subject to provisions to ensure][while addressing challenges related to] benefit-sharing and the protection of traditional knowledge associated with genetic resources, as necessary and appropriate [in order to not hinder [responsible] research and innovation [and fair and equitable sharing of outcomes of such research and innovation] [, [inter alia for] public health and food security] and be consistent with open [science principles][access to data]];]

(o) [Digital sequence information on genetic resources is made publicly available;]

(p) [The pooling of data [from different databases] benefits research and innovation and brings mutual benefits to the research and database communities [although open data in itself is not a means to ensure benefit-sharing];]

(q) [Relevance of [tracking and] tracing may depend on the approach taken to address digital sequence information [, for example, for hybrid approaches];]

(r) [Users of digital sequence information must inform the country [of origin or providing country] prior to accessing in case of both commercial and non-commercial use;]

(s) [Tracking and tracing could be [used][useful] for limited specialized subsets of digital sequence information on genetic resources [but is currently not feasible technically or financially at a large scale [and could also lead to a significant environmental footprint]];]

(t) [Be consistent with international human rights and obligations;]

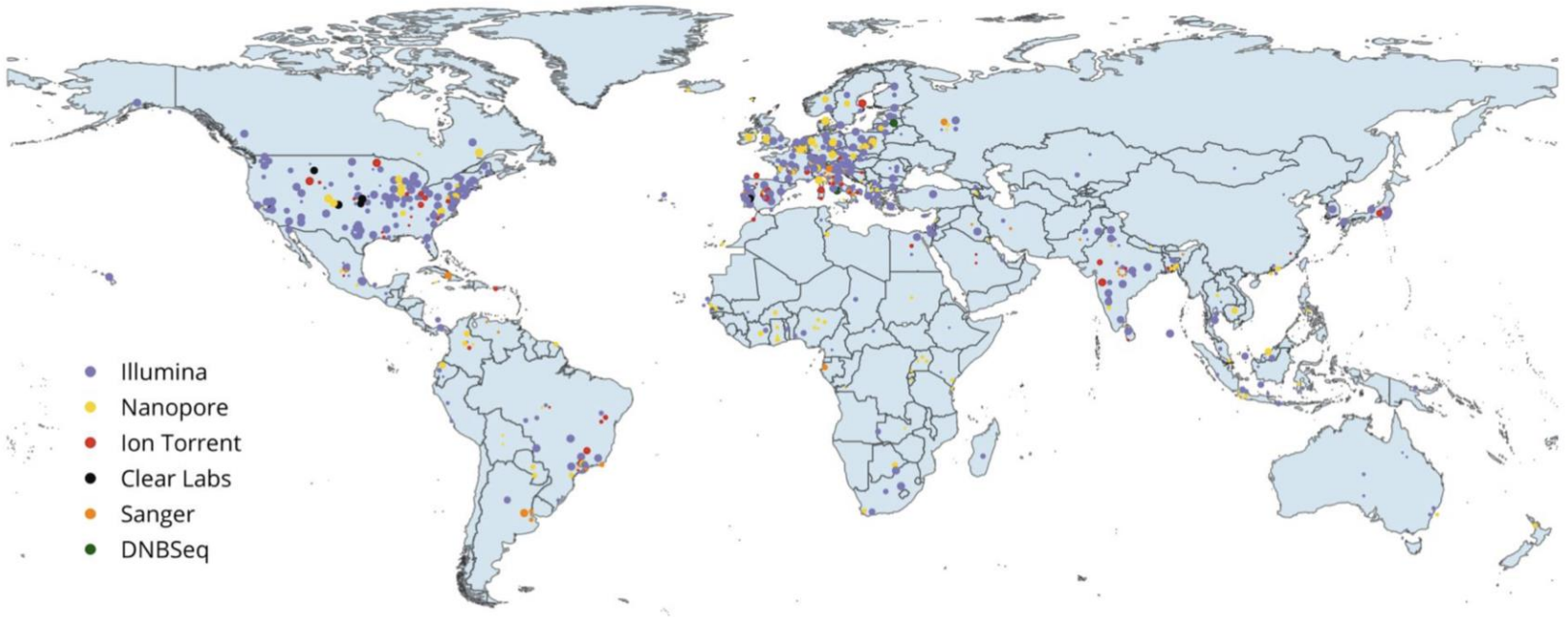
(u) [The Convention on Biological Diversity [could] provide a framework for a solution on the fair and equitable sharing of benefits from the [utilization of genetic resources in the form][use of] of digital sequence information;]

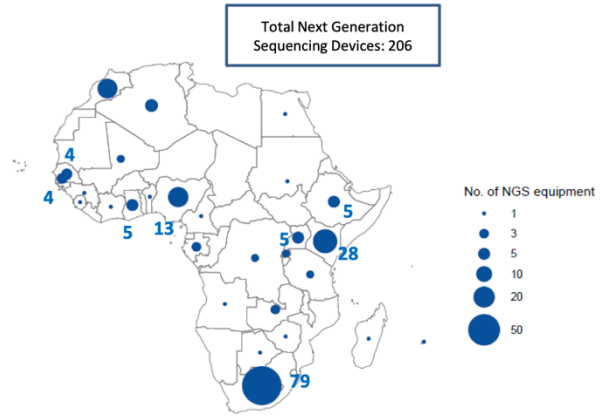
(v) [A solution on digital sequence information on genetic resources is likely to include a multilateral mechanism (for example, a multilateral fund). There are various views regarding the benefits of a solely multilateral system versus a hybrid system (i.e. multilateral with limited bilateral exceptions) and regarding the need for mixed models of funding or governance for such systems;]

Need for ethical and equitable data sharing to achieve timeliness and accuracy in epidemic surveillance and research

https://www.cbd.int/doc/recommendations/wg2020-04/wg2020-04-rec-02-en.pdf

Geographic disparities of sequencing technologies availability during the COVID-19 pandemic





Next Generation Sequencing Devices in Africa

https://doi.org/10.1016/S1473-3099(20)30939-7

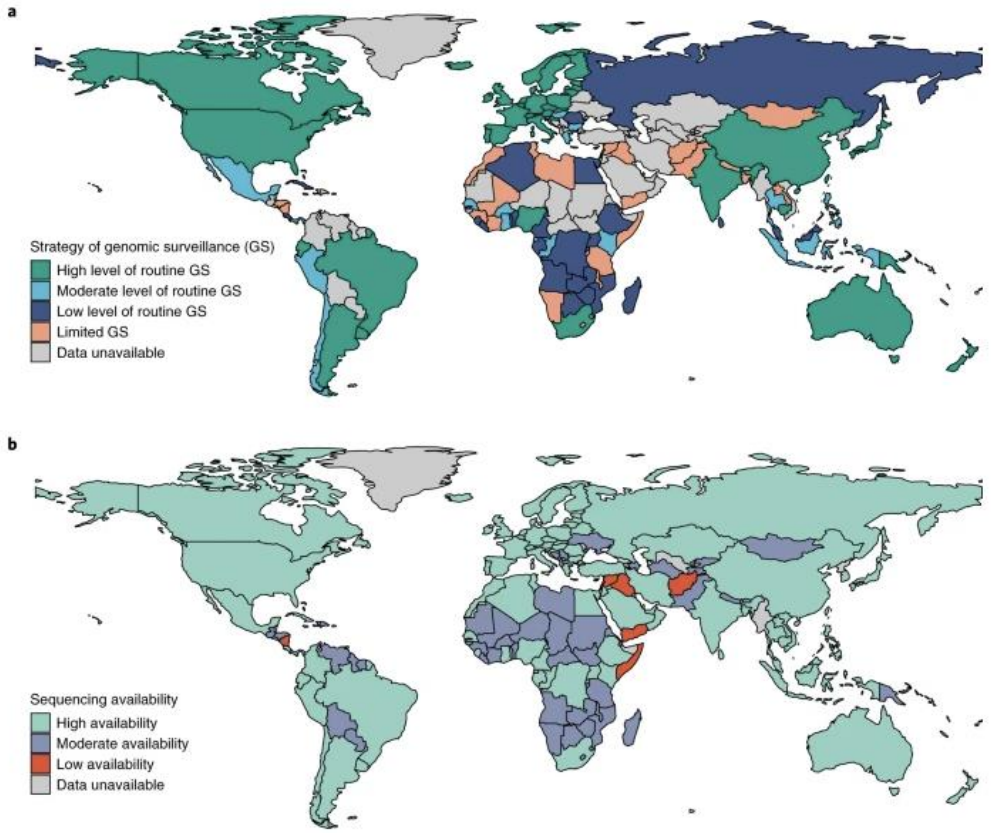
Map of sequencing tools available across global regions

Source: worldbank.org publication
https://documents1.worldbank.org/curated/en/099716212222240993/pdf/P17618002a6bfd00c09e39087a18be85684.pdf

based on data submitted to GISAID

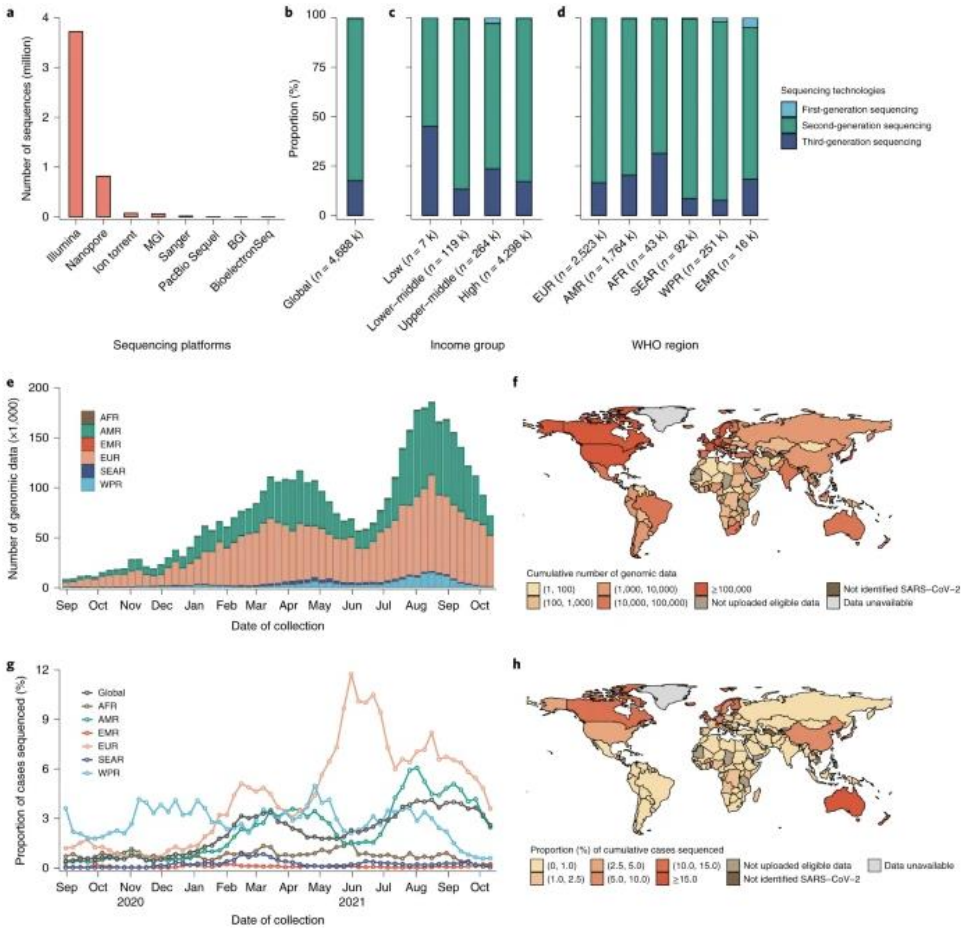
Geographic disparities and gaps in data sharing during the COVID-19 pandemic

Fig. 1: Global SARS-CoV-2 genomic surveillance and sequencing availability.



a, The global distribution of four strategies for SARS-CoV-2 genomic surveillance. **b**, The global availability of SARS-CoV-2 sequencing. 'Data unavailable' include locations that do not belong to the 194 Member States or do not have applicable data. Data shown here are as of 31 October 2021. Administrative boundaries were adapted from the database of Global Administrative Areas (GADM).

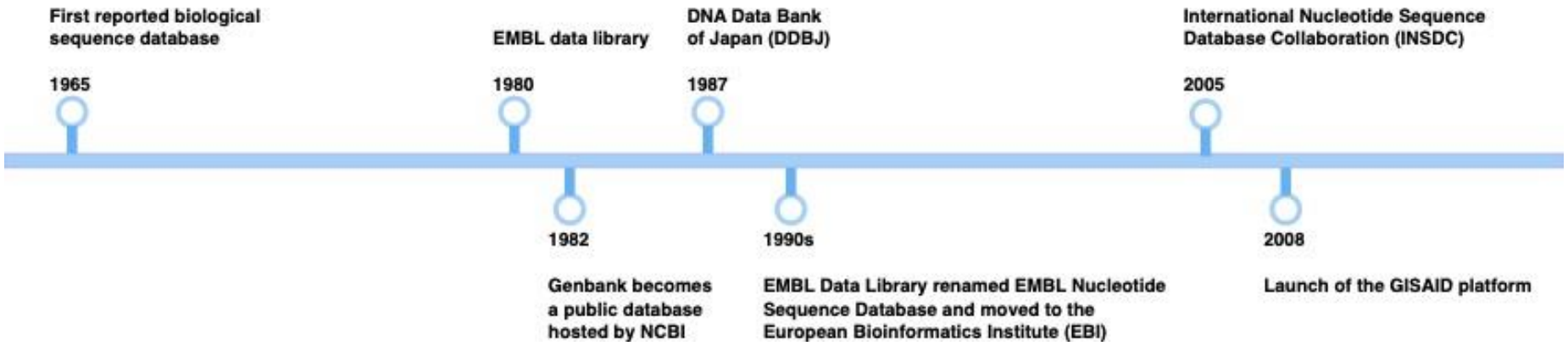
Fig. 2: Sequencing technologies and distribution of global publicly deposited genomic data.



https://doi.org/10.1038/s41588-022-01033-y

Biological sequences databases, a brief overview

Historical timeline



The INSDC (International Nucleotide Sequence Database Collaboration)

-Free and unrestricted access with appropriate citation following literature publishing standards

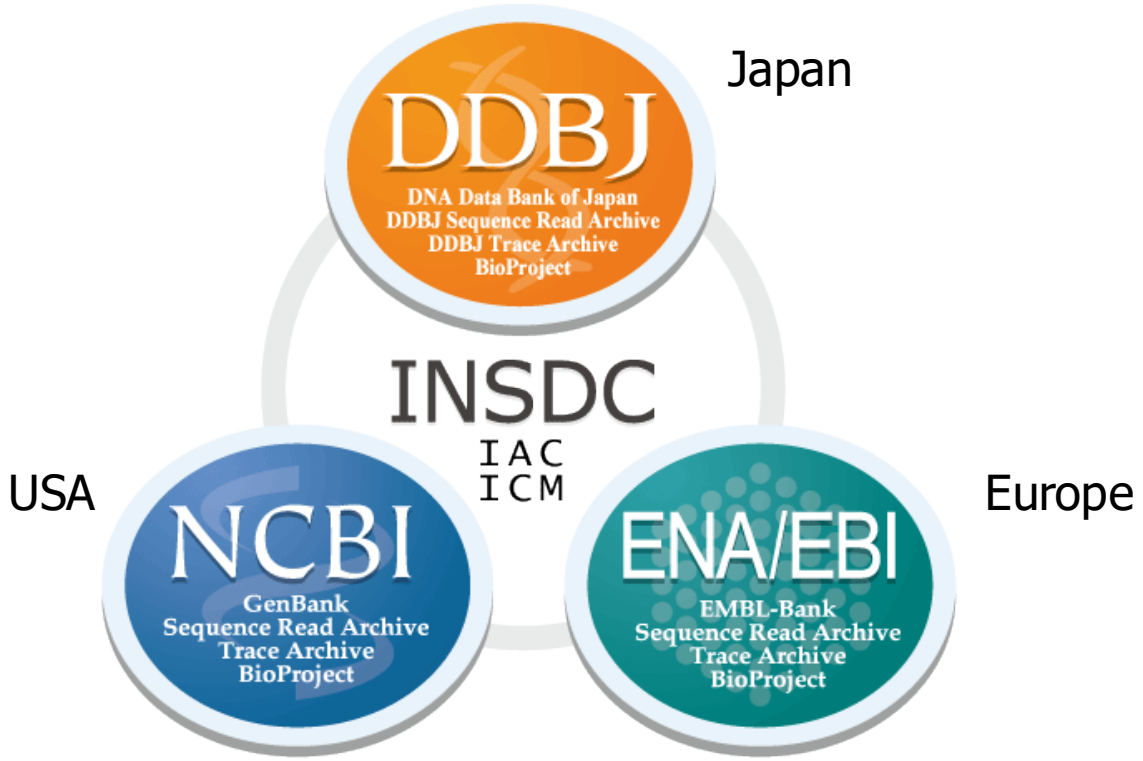
-No restriction or licencing fees attached to records

-All submitted records remain permanently accessible

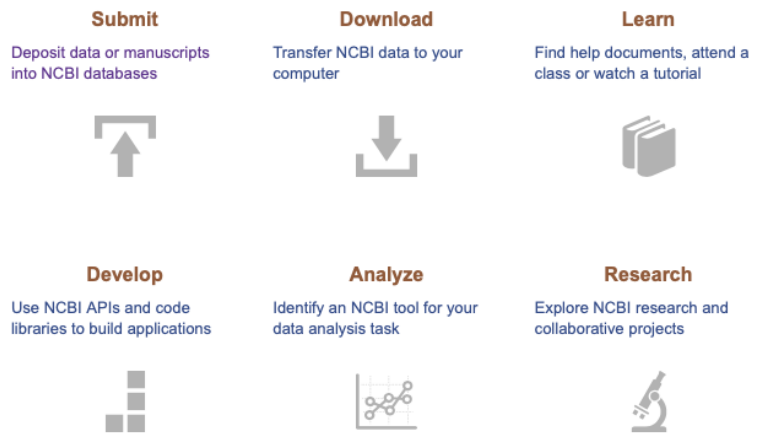
-Information disclosed to the web is fully public

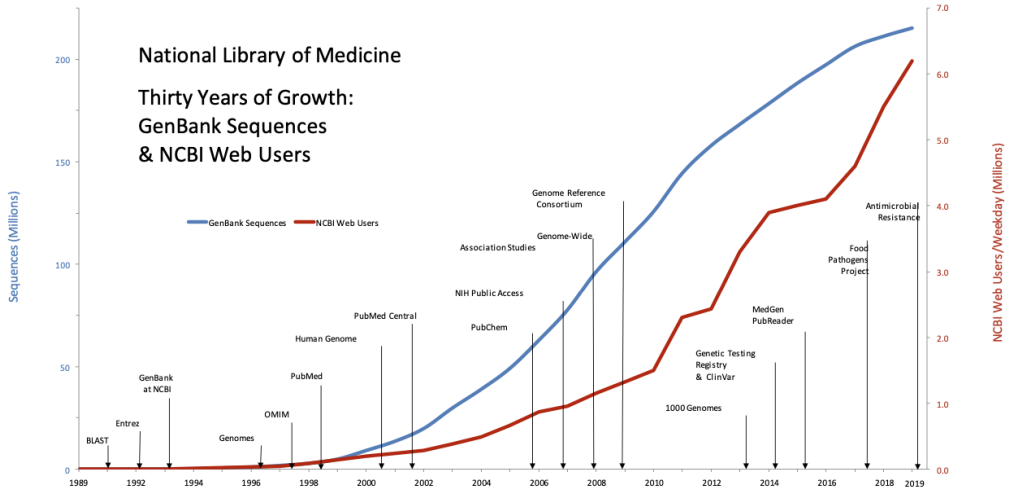
-Data integrity and accuracy are the responsibility of the submitting author, not the database

Manifest published in
Science 298 (5597): 1333 15 Nov 2002



The NCBI portal





https://www.nlm.nih.gov/about/2021CJ_NLM.pdf

The NCBI web portal is a free-access comprehensive resource for education and research and an entry point to all NCBI tools and databases.

Digital Sequence Information is organized in multiple databases by data types (e.g. SRA, Genbank, BioSample)

Or by dedicated organism-specific databases (e.g. SARS-CoV-2, Influenza Virus Resource)

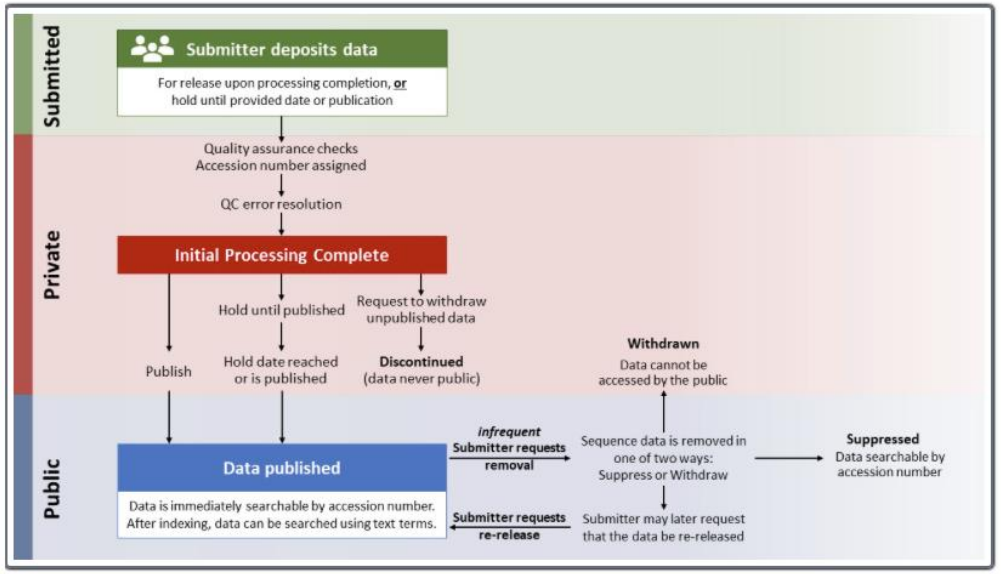
NCBI Genbank

Example of a Genbank record
https://www.ncbi.nlm.nih.gov/nuccore/NC_045512

Genbank is an annotated database of all publicly available genomic sequences.

Submissions to Genbank are pre-processed and checked for quality and integrity and submitter can withdraw data upon request.

As a part of the INSDC, Genbank exchanges data with the other sequences databases at ENA and DDBJ daily.



Submission workflow at Genbank and the Sequence Read Archive SRA

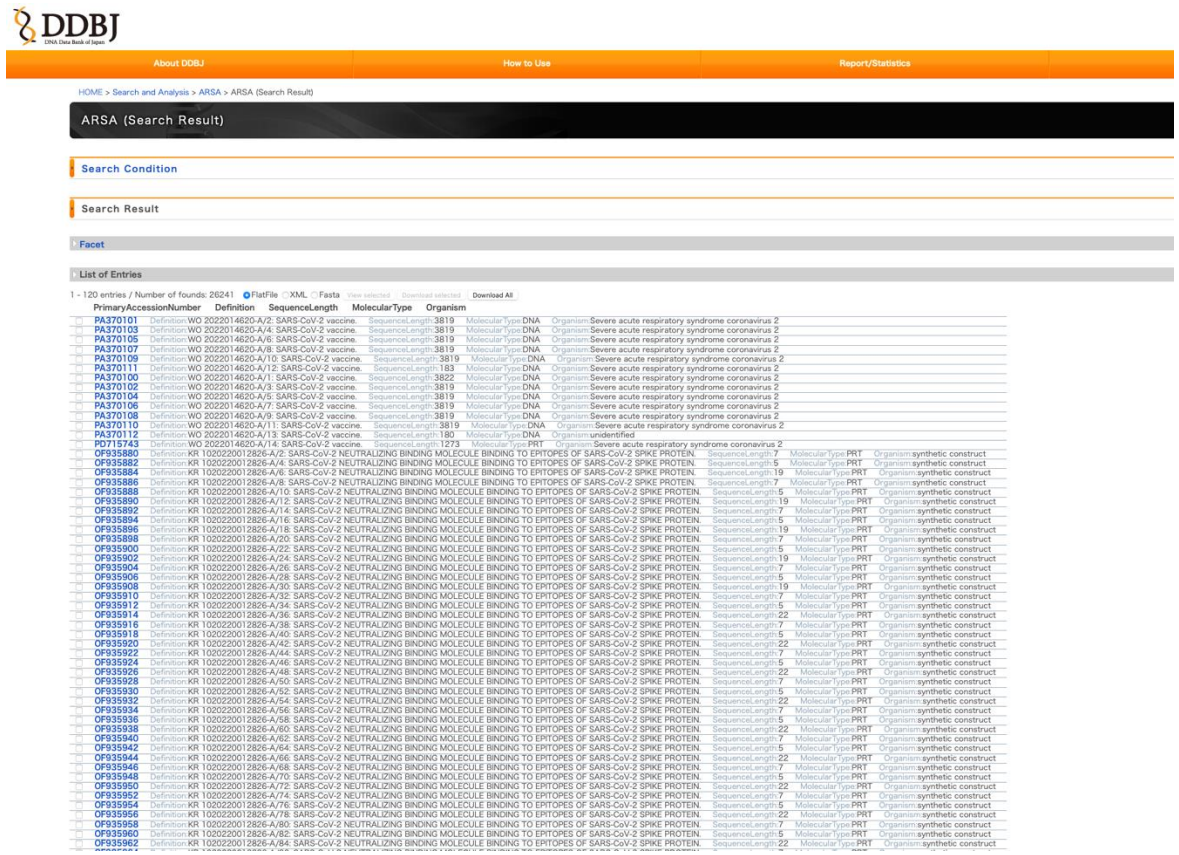
https://www.ncbi.nlm.nih.gov/sra/docs/sequence-data-processing/

The DDBJ (DNA Data Bank of Japan)

Provides freely available nucleotide sequence data and supercomputer resource to support life science research.

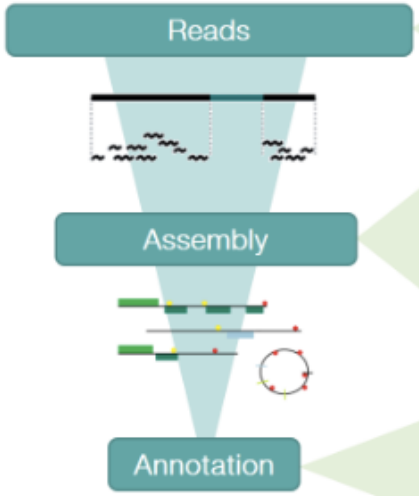
Provides all the sequence data information available in other INSDC databases

Provides nucleotide sequence data to INSDC related to patent applications in Japan, Korea, Europe and USA and amino acid sequence data related to patent applications in Japan, Korea.



https://www.ddbj.nig.ac.jp/index-e.html

The European Nucleotide Archive (ENA)



Run	Sequencing data files
Experiment	Sequencing methods

WGS set	Whole Genome Shotgun contig set
TSA set	Transcriptome assembly contig set
Scaffold	Assembled scaffold sequences
Chromosome	Fully assembled chromosomes (including organelles, plasmids and viral segments)

Coding sequences	Sequence regions reported as being protein-coding regions
Non-coding sequences	Sequence regions reported as representing non-protein-coding (RNA) genes

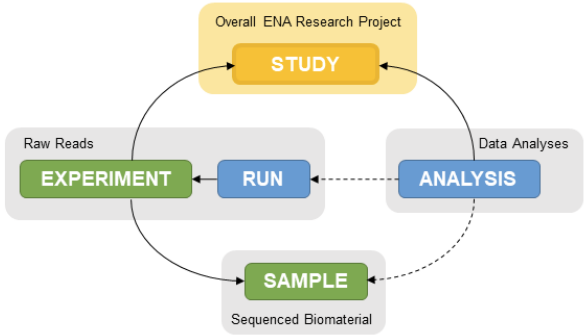
Contains data relating to experimental workflows based around nucleotide sequencing

Covers raw data, assemblies, annotations and information on machine configuration and outputs

https://www.ebi.ac.uk/ena/browser/about

Data types shared by the INSDC

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



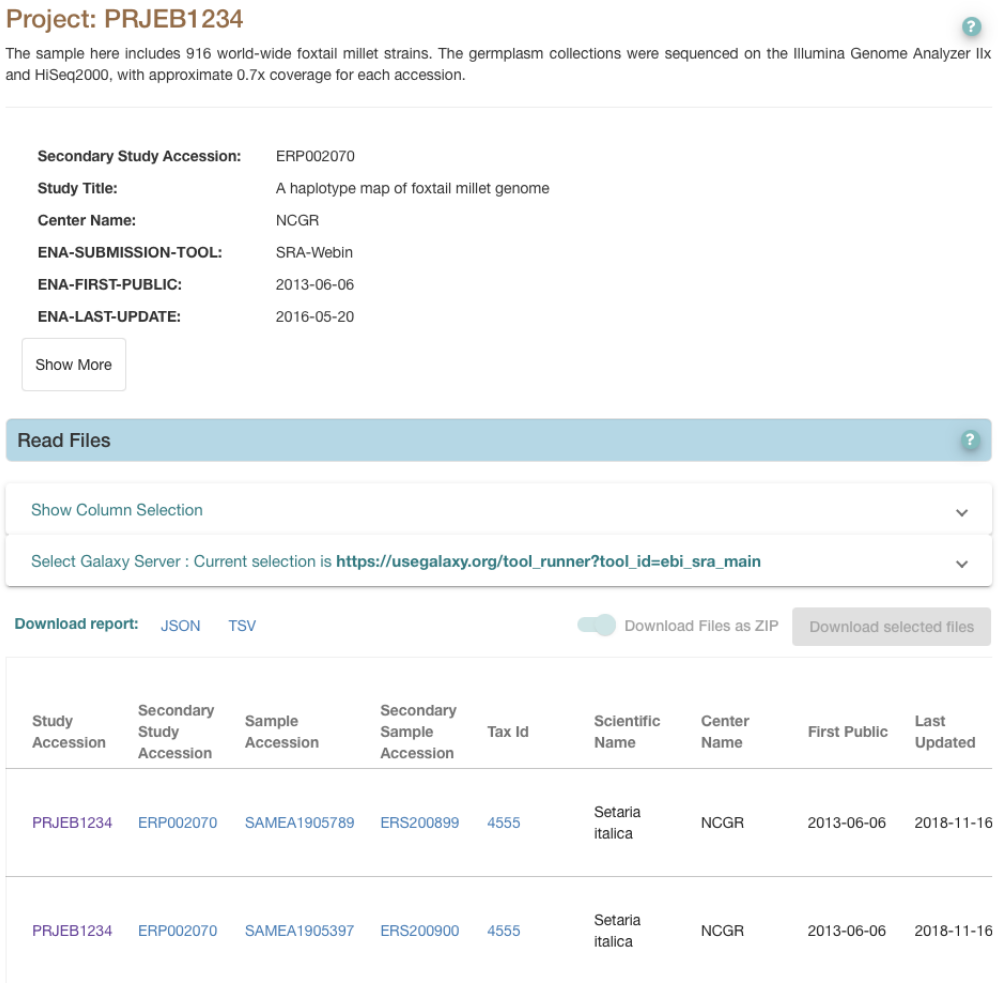
The Study object

Every submission requires Study

Contains the study name, short description and release date

Groups all other objects together (Sample, Experiment, Analysis, Runs)

The study and its associated data will not become public until the release date has expired.



Project: PRJEB1234

The sample here includes 916 world-wide foxtail millet strains. The germplasm collections were sequenced on the Illumina Genome Analyzer IIx and HiSeq2000, with approximate 0.7x coverage for each accession.

Secondary Study Accession:	ERP002070
Study Title:	A haplotype map of foxtail millet genome
Center Name:	NCGR
ENA-SUBMISSION-TOOL:	SRA-Webin
ENA-FIRST-PUBLIC:	2013-06-06
ENA-LAST-UPDATE:	2016-05-20

Show More

Read Files

Show Column Selection

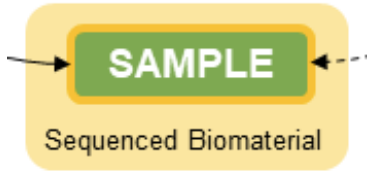
Select Galaxy Server : Current selection is **https://usegalaxy.org/tool_runner?tool_id=ebi_sra_main**

Download report: JSON TSV Download Files as ZIP Download selected files

Study Accession	Secondary Study Accession	Sample Accession	Secondary Sample Accession	Tax Id	Scientific Name	Center Name	First Public	Last Updated
PRJEB1234	ERP002070	SAMEA1905789	ERS200899	4555	Setaria italica	NCGR	2013-06-06	2018-11-16
PRJEB1234	ERP002070	SAMEA1905397	ERS200900	4555	Setaria italica	NCGR	2013-06-06	2018-11-16

https://www.ebi.ac.uk/ena/browser/view/PRJEB1234

The Sample object



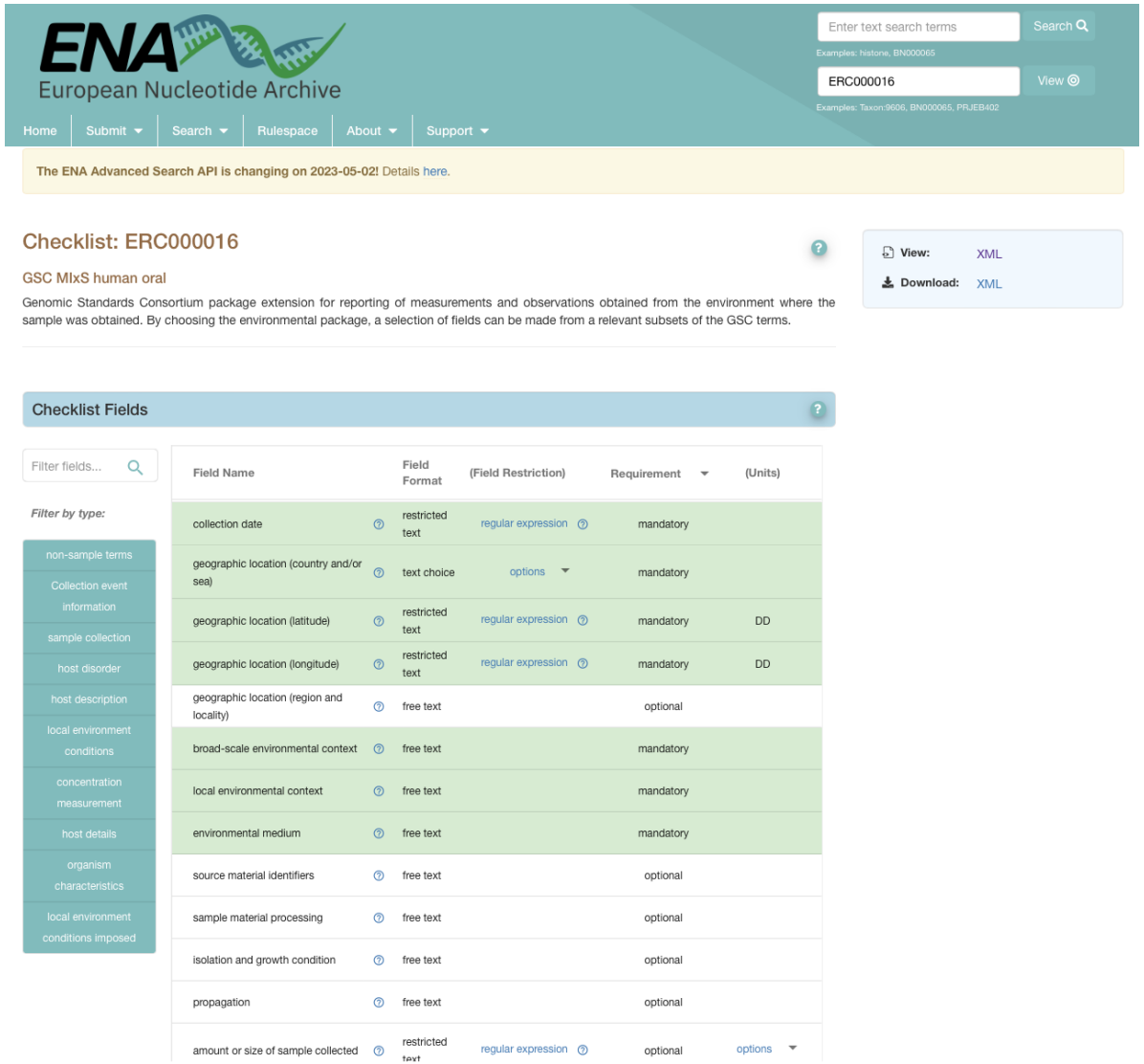
A sample object represents a unit of biomaterial associated to a single collection event

Required prior to submitting most types of data

Provides context to the data (location, collection date, milieu, etc.)

Different ENA checklists of which metadata to provide for different types of samples
https://www.ebi.ac.uk/ena/browser/checklists

Example of a ENA sample checklist



Structured data collection at ENA is implemented using sample checklists.

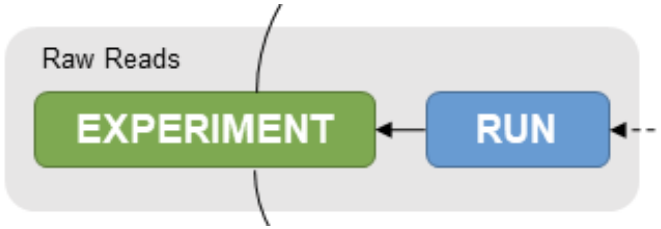
Sample data must be collected in a csv or XML format.

Green: mandatory field

Recommended or mandatory fields cannot be left blank.

https://www.ebi.ac.uk/ena/browser/view/ERC000016

The Experiment and Runs objects



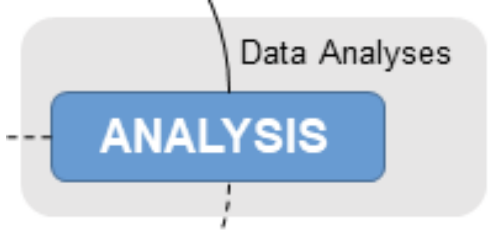
Raw reads are described on ENA by the 'Experiment' and 'Run' objects

An 'Experiment' object contains the metadata describing the method used to sequence the sample (platform, library selection strategy, etc)

A 'Run' object contains the information about the raw read files themselves (filename, md5 checksum)

Experiment and Runs objects are attached to the 'Sample' and 'Study' objects and therefore both 'Sample's and 'Studie's need to be created prior to submitting raw reads

The Analysis object



Genome assemblies are described by the object ‘Analysis’ within ENA

Different level of assemblies can be submitted: contig, scaffold, chromosome, requiring different files to include

SARS-CoV-2 assembly submissions are processed by a dedicated submission system.

To submit an assembly to ENA, a study and a sample need to be created prior to submission.

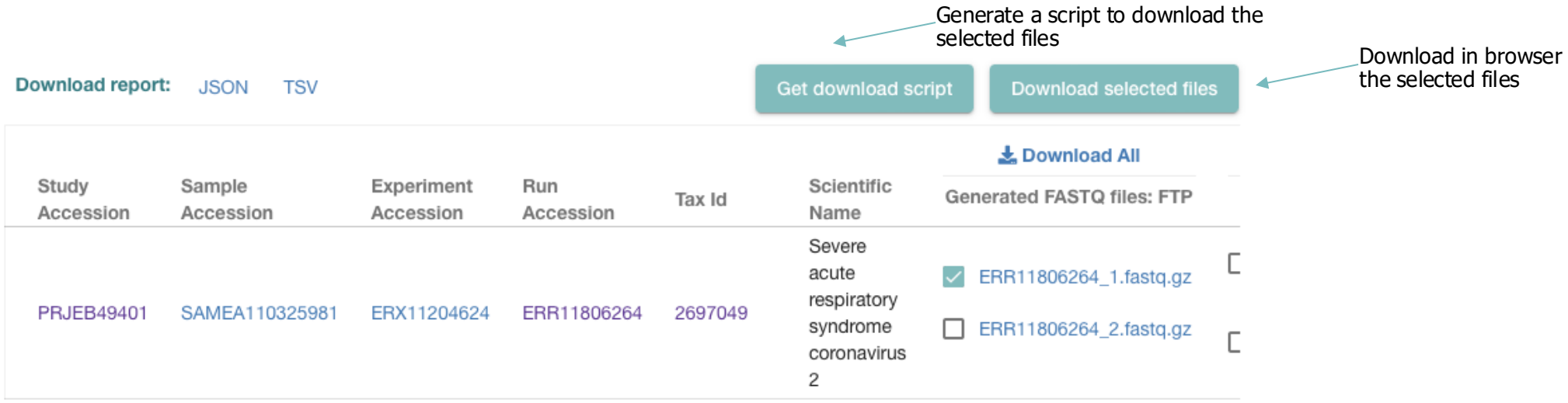
Data search services at ENA

https://www.ebi.ac.uk/ena/browser/home

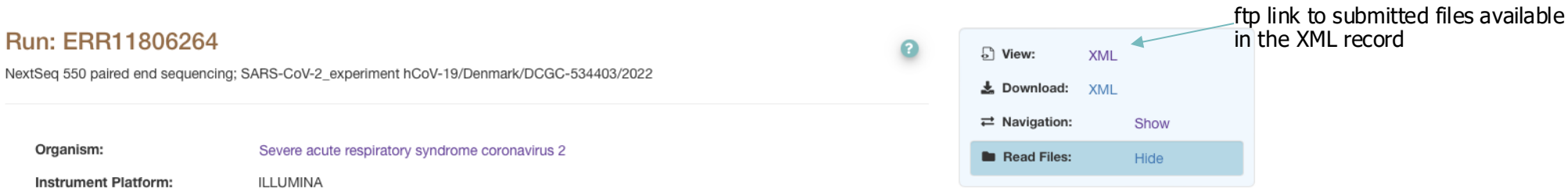
- Records searchable by accession numbers and free text search
- Links to available formats to download, crosslinks and related resources
- Summary reports downloadable in tabular or json formats.
- Advanced search available via the browser or using the API. https://www.ebi.ac.uk/ena/portal/api/

Data retrieval services (1/2)

- Single-file downloads can be done via the browser view page.



- Or by getting the ftp link by viewing the xml record



Data retrieval services (2/2)

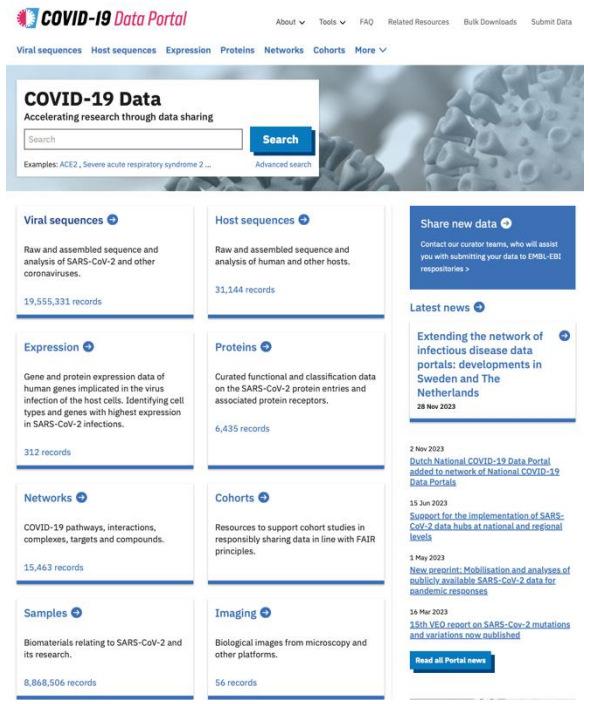
The ENA file uploader is the recommended option to retrieve multiple accessions

https://github.com/enasequence/ena-ftp-downloader/

- Java-based tool that can run in interactive or command line mode.
- Accepts single, list of accessions, advanced search queries, parallel downloads, retries
- Supports FTP and Aspera transfer protocols

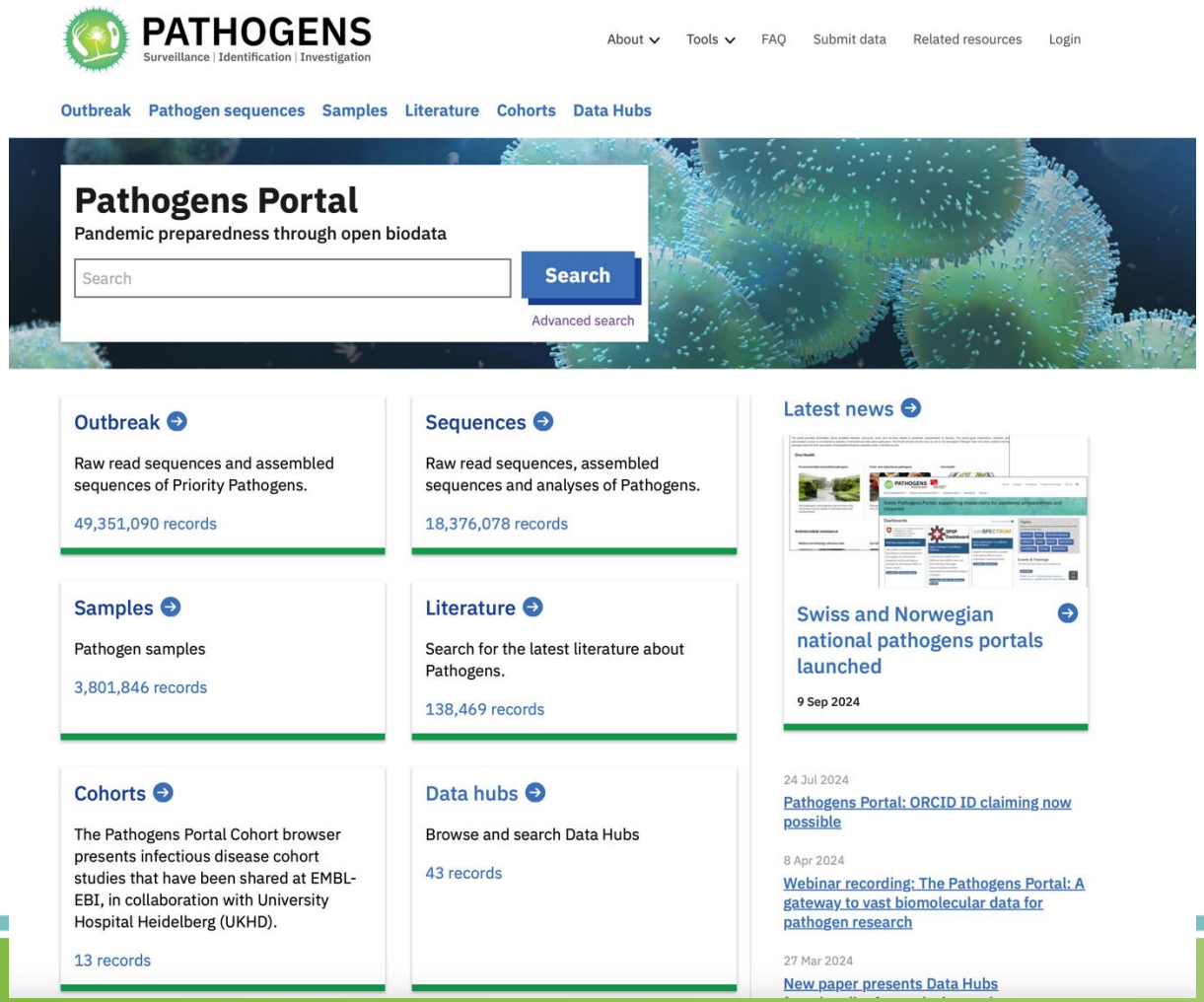
Data Hubs and entry points to ENA resources

Many data hubs and portals have been developed by ENA and their partners to facilitate rapid and large volume data sharing.



https://www.covid19dataportal.org

>9 million assembled sequences
11 national COVID-19 portals



ENA Pathogens Platform

All pathogen, all disease approach
Hosts, vectors and pathogens

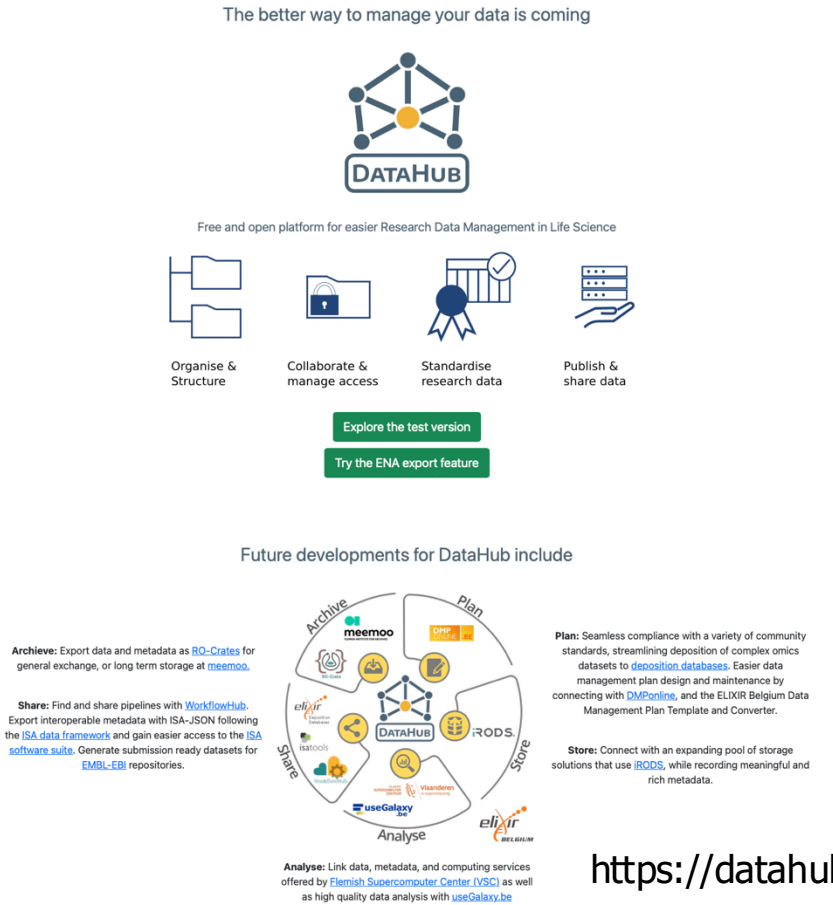
Antimicrobial resistance

Outbreaks

https://www.pathogensportal.org

Third party access point to ENA submission systems

Data brokering services have been developed by ENA partners to facilitate sequence data uploads, in particular during the COVID-19 pandemic.
For example Elixir Belgium provides data hub services and tools to submit data to ENA.



https://rdm.elixir-belgium.org/covid-19/sarscov2_submission.html
https://github.com/usegalaxy-eu/ena-upload-cli

https://datahub.elixir-belgium.org

GISAID (Global Initiative for Sharing Avian Influenza Data)

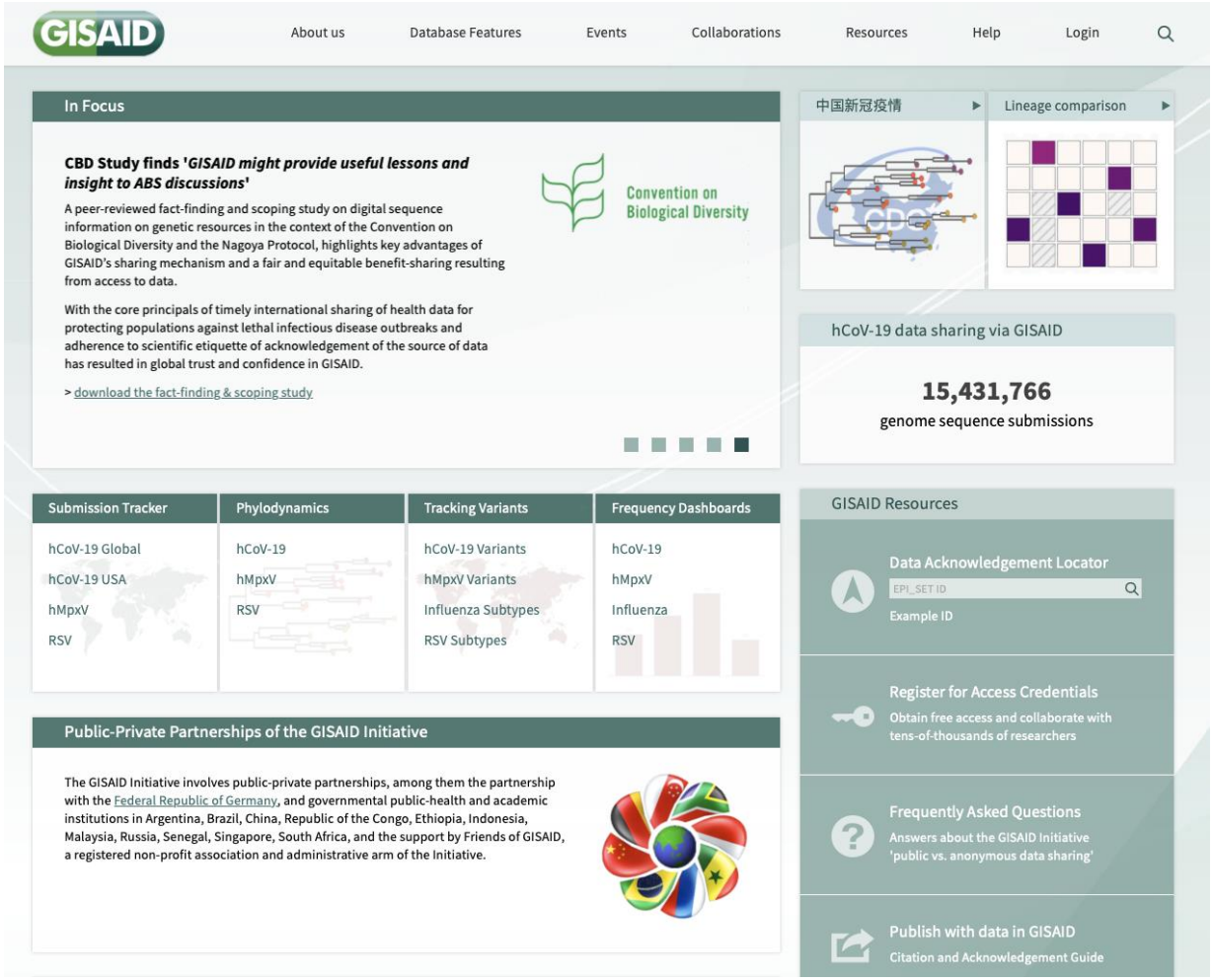
Global Initiative promoting rapid sharing of data from influenza, SARS-CoV-2 and more recently monkeypox viruses.

Restricted access to registered GISAID users after log-in

Open and free of charge access to all that agreed to identify themselves and agreed the GISAID sharing mechanism

Acknowledgement to the GISAID initiative and to the originating and submitting laborarories in case of use of the data in published work.

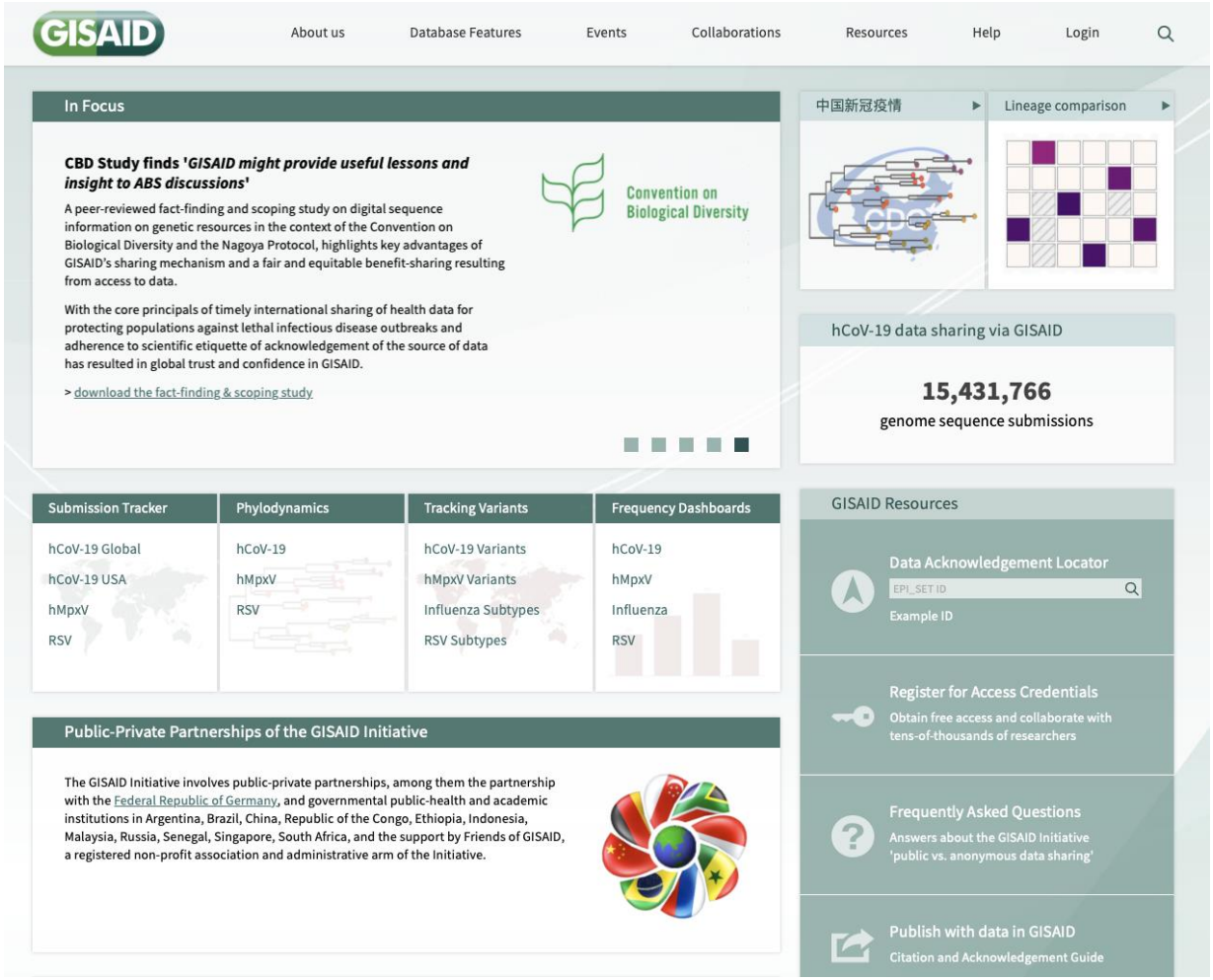
https://gisaid.org/about-us/mission/



GISAID (Global Initiative for Sharing Avian Influenza Data)

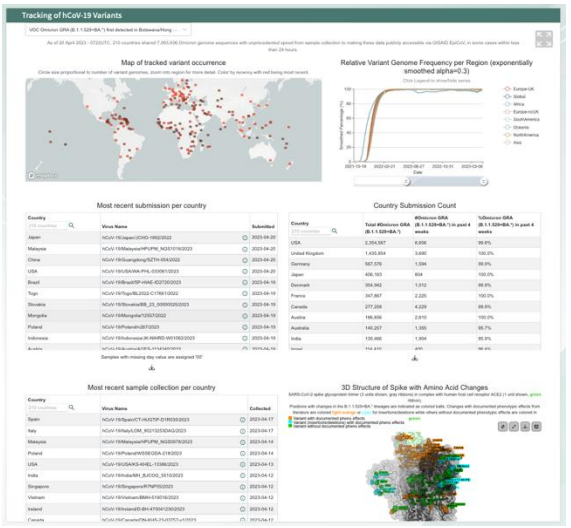
GISAID has several databases
- EpiFlu
- EpiCoV
- EpiRSV
- EpiPox
- Different levels of access

Different tools/features for each databases
  - Dependent on virus biology

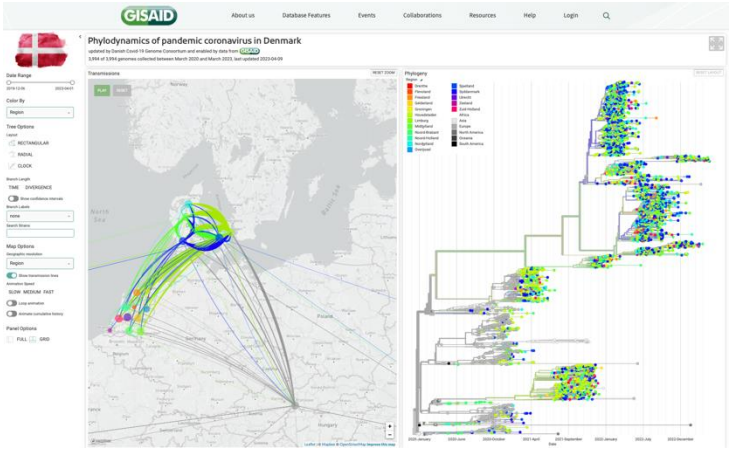


Free-access features and tools available on GISAID

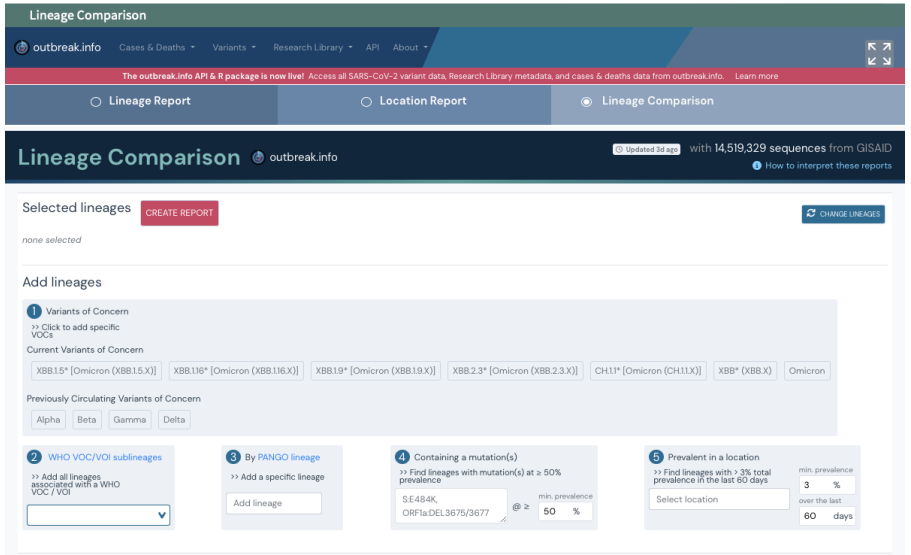
Submission tracker



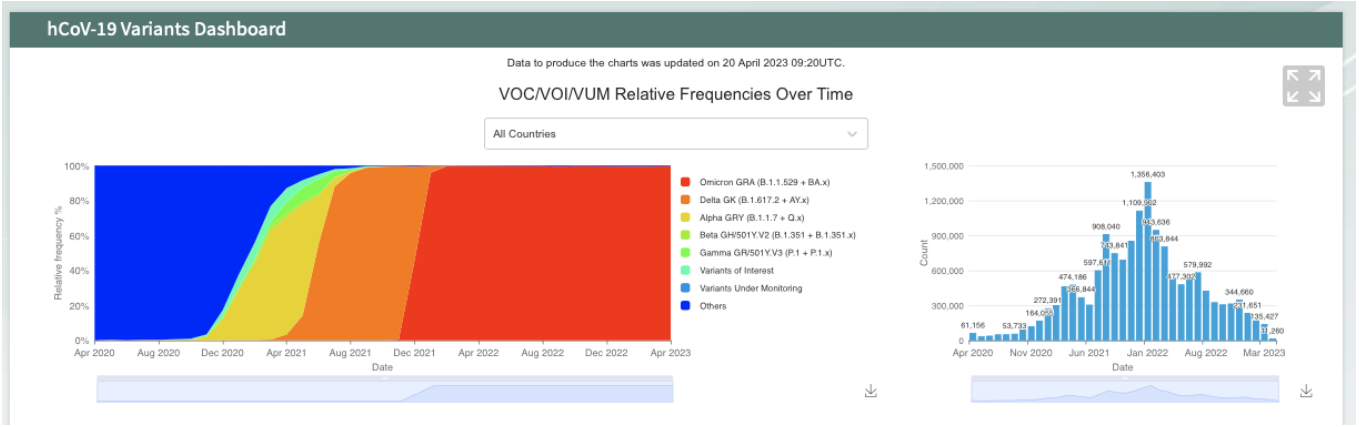
Phylodynamics

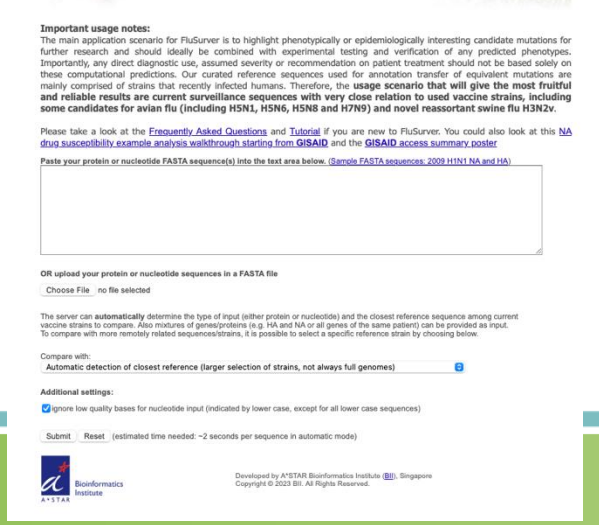


Lineage comparison



Dashboard

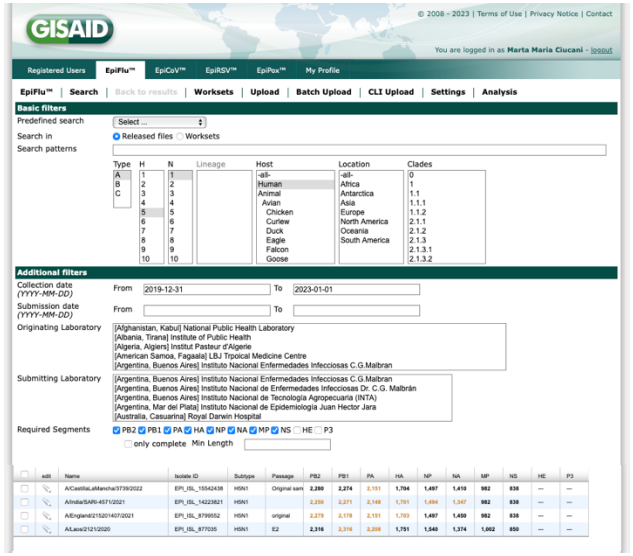




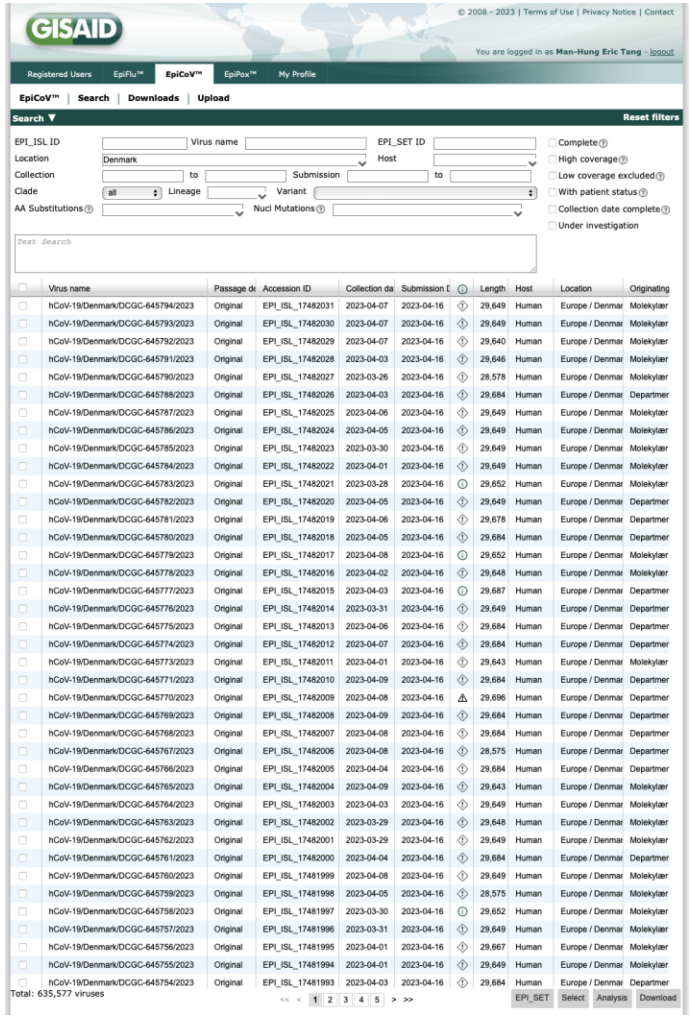
FluSurver

EpiFlu and EpiCoV

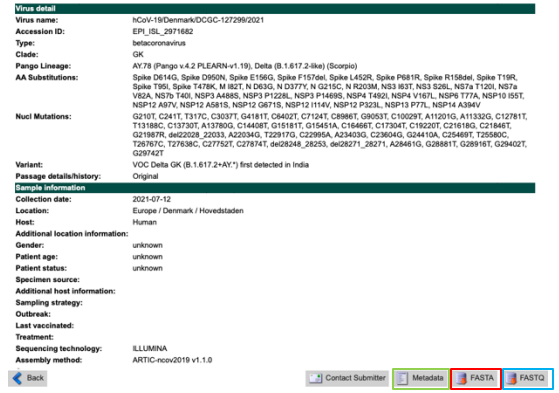
EpiFlu



EpiCoV



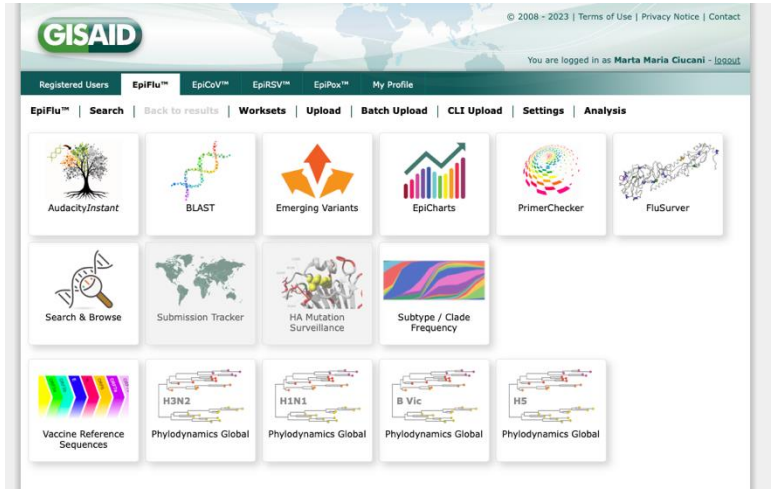
Online resource to search and download assemblies and raw reads and metadata

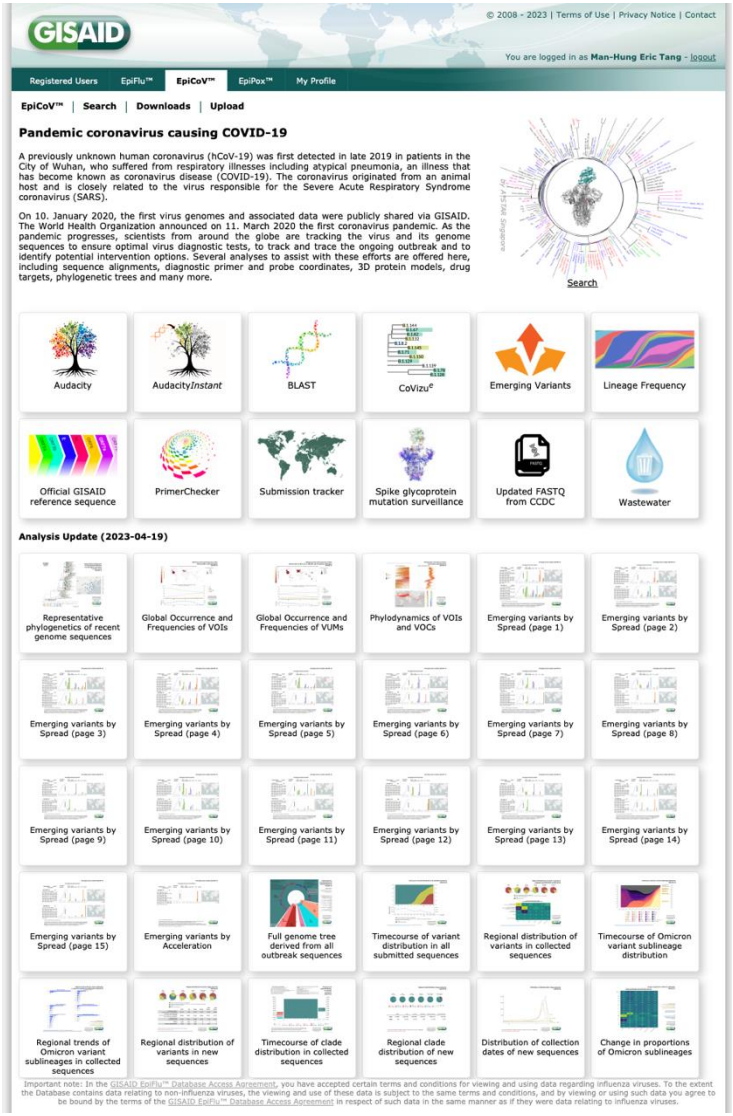


Tools restricted to GISAID members

Analysis tools developed by GISAID members/partners

EpiFlu

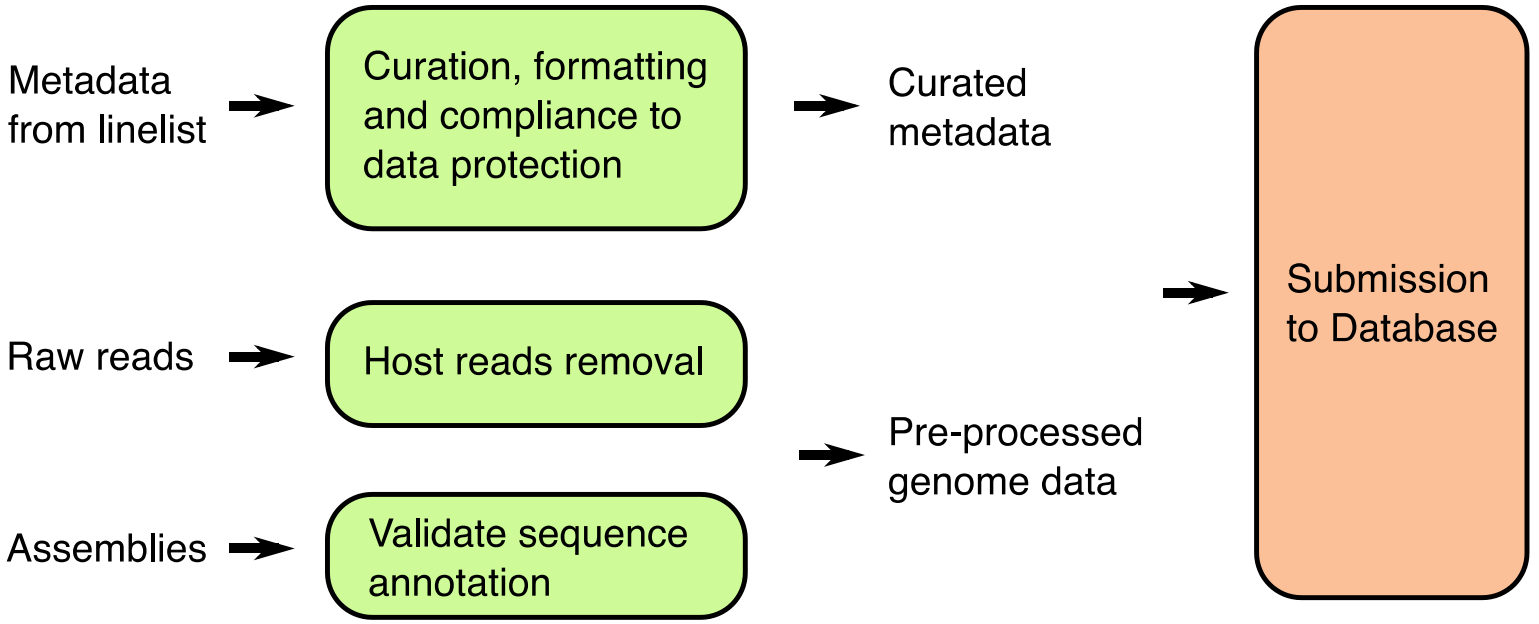




EpiCoV

Submitting data to ENA

General workflow for submitting sequence data



Checklist-based metadata collection

GISAID submission metadata template

EpiCov Columns	EpiCov Example
submitter	pmat-ssi
fn	sequences.fasta
covv_virus_name	hCoV-19/Denmark/DCGC-579315/2022
covv_type	betacoronavirus
covv_passage	Original
covv_collection_date	2022-09-07
covv_location	Europe/Denmark/Nordjylland
covv_add_location	
covv_host	Human
covv_add_host_info	n_infections=1
covv_gender	unknown
covv_patient_age	unknown
covv_patient_status	unknown
covv_specimen	unknown
covv_outbreak	unknown
covv_last_vaccinated	unknown
covv_treatment	unknown
covv_seq_technology	ILLUMINA
covv_assembly_method	ARTIC-ncov2019 v1.1.0
covv_coverage	5602.89170450142
covv_orig_lab	Department of Bacteria, Parasites and Fungi, Statens Serum Institut, Copenhagen, Denmark
covv_orig_lab_addr	Artillerivej 5, 2300 Copenhagen S, Denmark
covv_provider_sample_id	
covv_subm_lab	Department of Bacteria, Parasites and Fungi, Statens Serum Institut, Copenhagen, Denmark
covv_subm_lab_addr	Artillerivej 5, 2300 Copenhagen S, Denmark
covv_subm_sample_id	hCoV-19/Denmark/DCGC-579315/2022
covv_authors	Danish Covid-19 Genome Consortium

ENA submission checklist for virus pathogens

Checklist	tax_id	#units	2697049
ERC000033	scientific_name		Severe acute respiratory syndrome coronavirus 2
ENA virus pathogen reporting standard checklist	sample_alias		hCoV-19/Denmark/DCGC-587685/2022
	sample_title		hCoV-19/Denmark/DCGC-587685/2022
	collection date		27/09/2022
	geographic location (country and/or sea)		Denmark
	geographic location (region and locality)		Sjaelland
	sample capture status		Active surveillance in response to outbreak
	host common name		Human
	host subject id		not provided
	host health state		not provided
	host sex		not provided
	host scientific name		Homo sapiens
	collector name		Danish Covid-19 Genome Consortium
	collecting institution		Department of Bacteria, Parasites and Fungi, Statens Serum Institut, Copenhagen, Denmark
	isolate		not provided

Submitted metadata must comply to the data protection requirements.
Host-associated information is often either not provided, restricted or anonymized.

Metadata are usually provided by linelists, databases, spreadsheets, etc.

Host (Human) reads removal

It is often required to remove any reads mapping to Human genome before submitting sequence read data.

Two main approaches:
- whitelist approach (keep what resembles the target genome)
  For example:

ReadItAndKeep: rapid decontamination of SARS-CoV-2 sequencing reads

Martin Hunt, Jeremy Swann, Bede Constantinides, Philip W Fowler, Zamin Iqbal

Bioinformatics, Volume 38, Issue 12, June 2022, Pages 3291–3293, https://doi.org/10.1093/bioinformatics/btac311

- blacklist approach (remove anything that resembles Human DNA)
  For example:
  SRA human scrubber https://github.com/ncbi/sra-human-scrubber

STAT: a fast, scalable, MinHash-based *k*-mer tool to assess Sequence Read Archive next-generation sequence submissions

Kenneth S. Katz, Oleg Shutov, Richard Lapoint, Michael Kimelman, J. Rodney Brister & Christopher O'Sullivan

Genome Biology **22**, Article number: 270 (2021) | Cite this article

https://doi.org/10.1186/s13059-021-02490-0

These solutions are command-line based and many submission systems use an integrated human host removal tool

Virus assemblies validation tool

Schäffer *et al. BMC Bioinformatics* (2020) 21:211
https://doi.org/10.1186/s12859-020-3537-3

BMC Bioinformatics

SOFTWARE Open Access

VADR: validation and annotation of virus sequence submissions to GenBank

Alejandro A. Schäffer[1,2], Eneida L. Hatcher[2], Linda Yankie[2], Lara Shonkwiler[2,3], J. Rodney Brister[2], Ilene Karsch-Mizrachi[2] and Eric P. Nawrocki[2*]

*Correspondence:
nawrocke@ncbi.nlm.nih.gov
[2]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD, 20894 USA
Full list of author information is available at the end of the article

Abstract

Background: GenBank contains over 3 million viral sequences. The National Center for Biotechnology Information (NCBI) previously made available a tool for validating and annotating influenza virus sequences that is used to check submissions to GenBank. Before this project, there was no analogous tool in use for non-influenza viral sequence submissions.

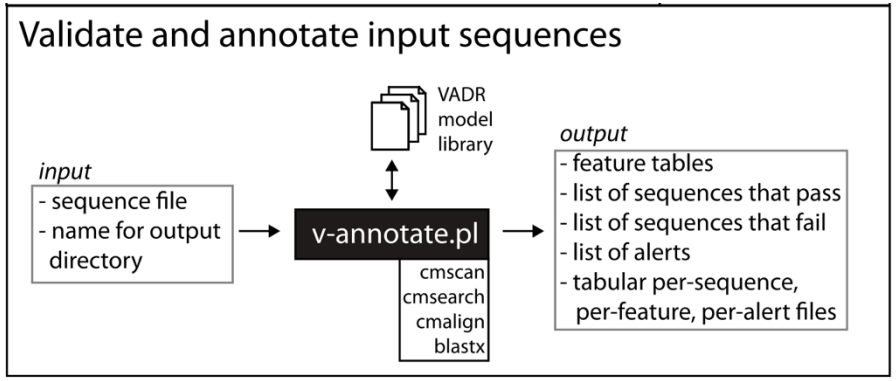


Fig. 1 VADR workflow schematic illustrating uses of the two main VADR scripts. *v-build.pl* can be used once to build a single model or repeatedly to build a library of models. *v-annotate.pl* can be used with a model or model library to validate and annotate input sequences

Supported genomes

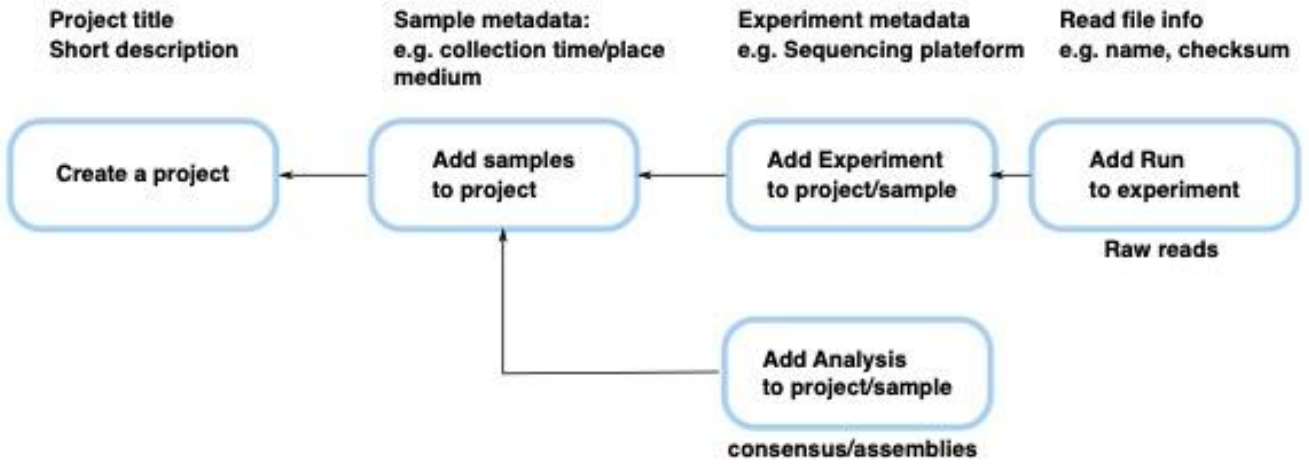
- SARS-CoV-2 and other Coronaviridae RefSeq models
- Influenza models
- Mpox virus (MPXV) RefSeq model
- RSV models
- Norovirus and other Caliciviridae RefSeq models
- Dengue virus and other Flaviviridae RefSeq models
- Metazoan Cytochrome c oxidase I (COX1) models models

Automated submission screening tool used in GenBank for viral genome sequence submissions

https://github.com/ncbi/vadr/

https://doi.org/10.1186/s12859-020-3537-3

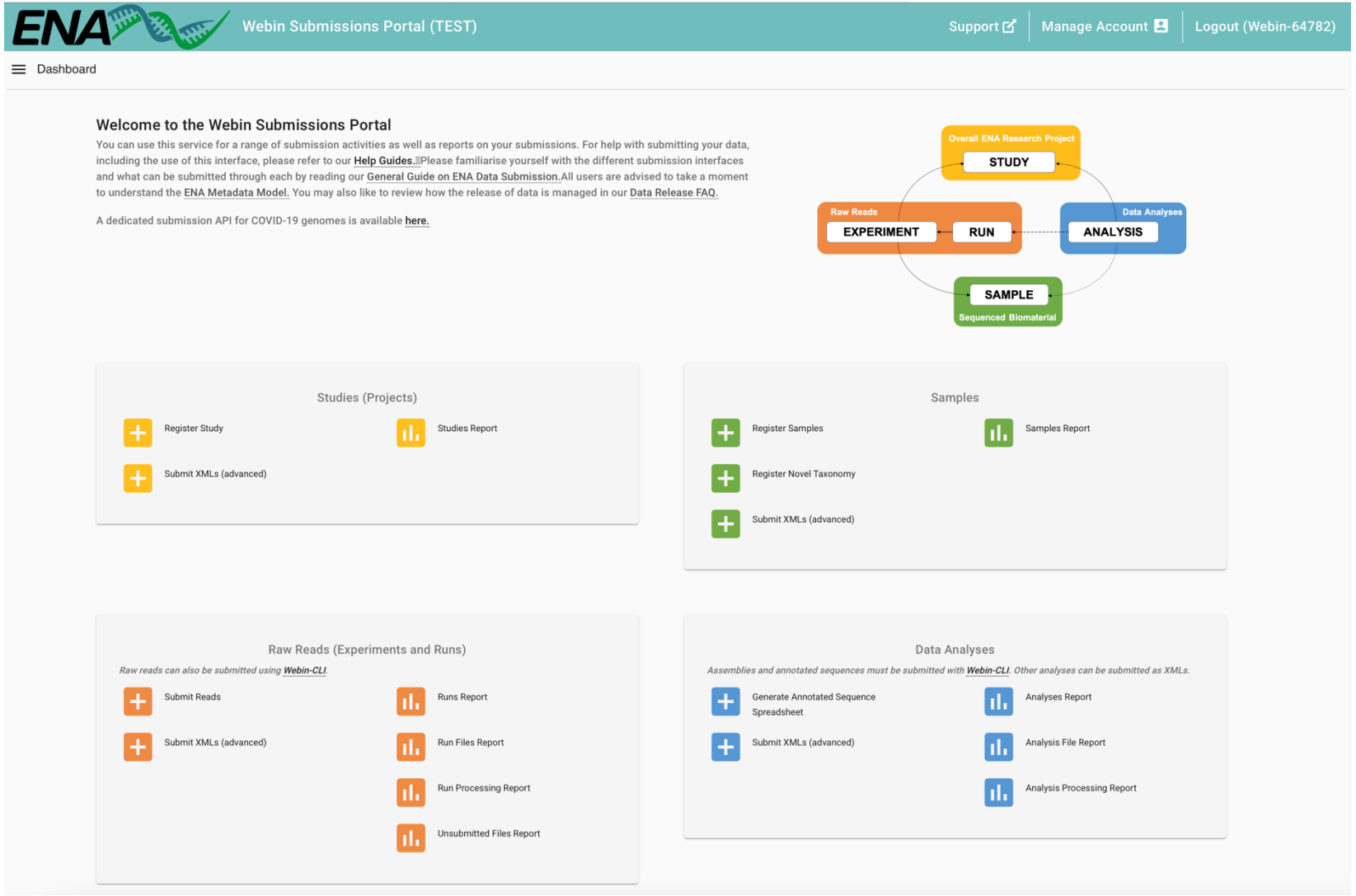
ENA submission workflow



https://ena-docs.readthedocs.io/en/latest/submit/general-guide.html

https://ena-browser-docs.readthedocs.io/en/latest/help_and_guides/sars-cov-2-submissions.html

The ENA web-interactive submission platform (Webin)



https://wwwdev.ebi.ac.uk/ena/submit/webin/

User-interface designed to handle small submissions (< 1000 samples)

TEST (submission deleted after 24 hours) and production servers

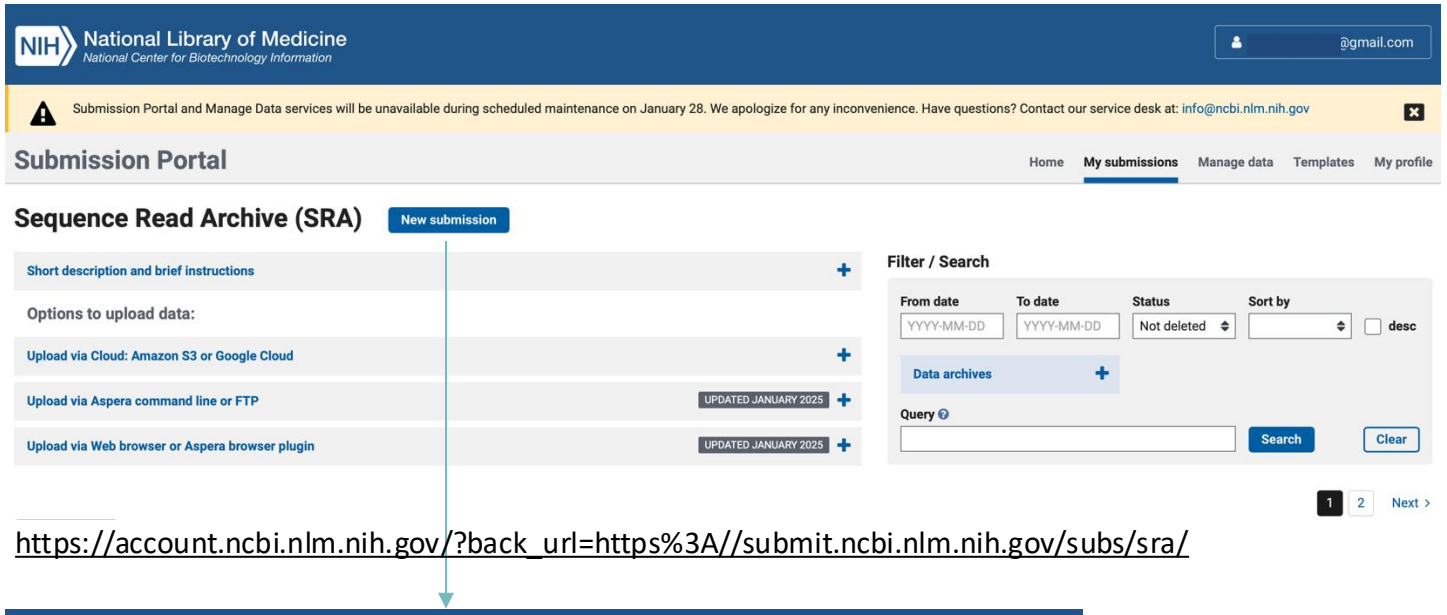
Submission account needed

Spreadsheet-based or XML-based submissions are supported

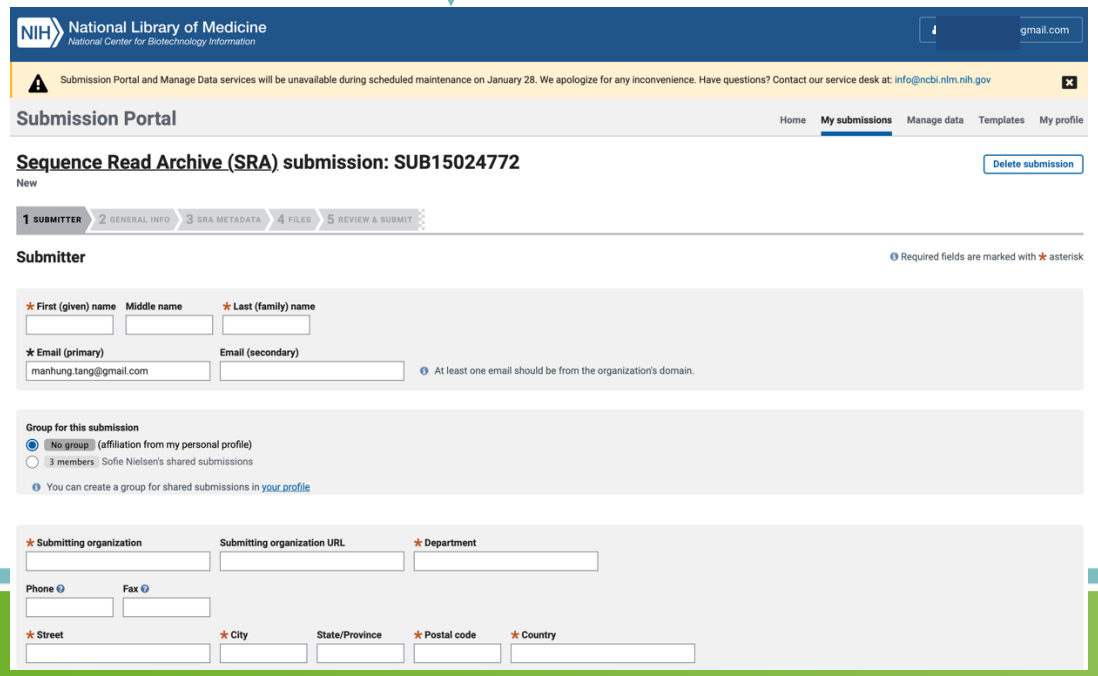
Reports allow to check the status of the submission of each object

Support tab: access point to contact the ENA team to report issues

The NCBI SRA interactive submission portal



https://account.ncbi.nlm.nih.gov/?back_url=https%3A//submit.ncbi.nlm.nih.gov/subs/sra/



Spreadsheet-based submissions

Requires a google or an institutional account

Production server only (each submission is permanent)

Similar structure as ENA interactive submissions

Requires two spreadsheets, one containing the sample metadata, the other containing read file information

References

https://www.cbd.int/

https://www.cbd.int/doc/recommendations/wg2020-04/wg2020-04-rec-02-en.pdf

https://www.ncbi.nlm.nih.gov/genbank/

https://www.ddbj.nig.ac.jp/index-e.html

https://www.ebi.ac.uk/ena/browser/about

https://www.ebi.ac.uk/ena/browser/checklists

https://gisaid.org/about-us/mission/

https://www.nlm.nih.gov/about/2021CJ_NLM.pdf

https://www.ncbi.nlm.nih.gov/sra/docs/sequence-data-processing/

https://ena-docs.readthedocs.io/en/latest/submit/general-guide.html

10.1126/science.abi4496

10.1126/science.298.5597.1333b

10.1038/s41588-022-01033-y

10.1093/bioinformatics/btac311

10.1186/s13059-021-02490-0

10.1186/s12859-020-3537-3

Practical exercises

Part 1: Make your own interactive submission to ENA

Part 2: Get familiar with the ENA data search and retrieval services

Acknowledgements

The creation of this training material was commissioned by ECDC to Statens Serum Institut with the direct involvement of Man-Hung Eric Tang