



Training in genomic Epidemiology and Public Health Bioinformatics "Bridging the Gap" Day 9: Introduction to sequence databases and data sharing



# **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



#### INFORMATION - ABOUT THE SECRETARIAT -Sign up for an account | Sign In English - Search Convention on BIODIVERSITY CONVENTION CARTAGENA PROTOCOL NAGOYA PROTOCOL COUNTRIES PROGRAMMES **Biological Diversity** THE NAGOYA PROTOCOL ON ACCESS AND BENEFIT-SHARING Entered into force on 12 October 2014 ACCESS AND BENEFIT-SHARING ACCESS AND BENEFIT-SHARING THURSDAY // 1.16.2025 NAGOYA PROTOCOL 142 ratifiations The Nagoya Protocol on Access and Benefit-sharing **Ratifications** > ABS under the GBF > About the Nadova Protoco > Nagova Protocol Text > History PARTIES Nagoya Protocol on Access and Benefit-sharing > Becoming a Party > List of Parties > National information - country profiles > Key Steps towards implementation KEY PROTOCOL ISSUES With the ratification of Costa Rica, the Nagoya > ABS Clearing-House Protocol has 142 > Assessment and review ratifications/accessions. Booklets available in: > Awareness-raising Ar | En | Es | Fr | Ru | Zh | Courtesy Translations > Capacity-building and development > Compliance with the Protocol The Nagoya Protocol on Access to Genetic Resources and the Fair and Equitable Sharing of Benefits Arising from their > Cooperation COP-MOP-5 Utilization to the Convention on Biological Diversity is an international agreement which aims at sharing the benefits arising > Digital sequence information on genetic resources from the utilization of genetic resources in a fair and equitable way. It entered into force on 12 October 2014, 90 days after > Financial mechanism the date of deposit of the fiftieth instrument of ratification. Learn more about the Nagoya Protocol. > Global multilateral benefit-sharing mechanism > Model contractual clauses, codes of conduct, guidelines and best practices and/or standards > Monitoring and reporting > Resource mobilization Visit the ABS Clearing-House: > Specialized international ABS instruments COP-MOP > COP-MOP Bureau ABSCH > COP-MOP decisions **Quick Links to** ACTIVITIES AND COMMUNICATIONS The Access and Benefit-sharing Clearing-House (ABS Clearing-House) is a platform for exchanging information on access **ABS Resources** > Meetings and Documents and benefit-sharing established by Article 14 of the Protocol, as part of the Clearing-House of the Convention established > Notifications under Article 18, paragraph 3 of the Convention. The ABS Clearing-House is a key tool for facilitating the implementation of > Statements and Press releases the Nagoya Protocol, by enhancing legal certainty and transparency on procedures for access and benefit-sharing, and for RESOLIRCES monitoring the utilization of genetic resources along the value chain, including through the internationally recognized > 10-year anniversary of the Protocol certificate of compliance. By hosting relevant information regarding ABS, the ABS Clearing-House will offer opportunities > ABS Clearing-House E-learning modules for connecting users and providers of genetic resources and associated traditional knowledge. > Awareness-raising Learn more about the ABS Clearing-House. **CEPA** Toolkit > Bonn Guidelines

#### https://www.cbd.int/

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

(0)

#### [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,<sup>2</sup> 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]];]

[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;]

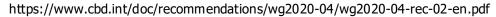
(s) [Tracking and tracing could be [used][useful] for limited specialized subsets of digital sequence information on genetic resources [but is currently not feasible technically or financially at a large scale [and could also lead to a significant environmental footprint]];]

(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;]

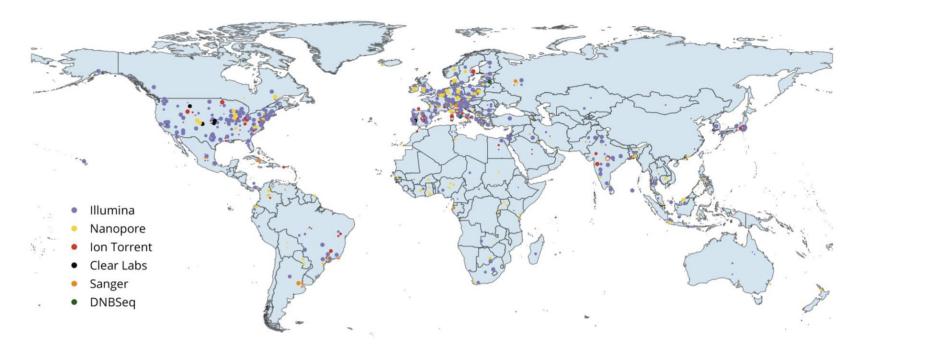
Need for ethical and equitable data sharing to achieve timeliness and accuracy in epidemic surveillance and research

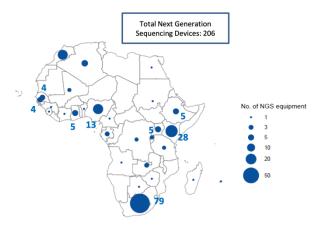




### **Geographic disparities of sequencing technologies availability during the COVID-19 pandemic**







Next Generation Sequencing Devices in Africa

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

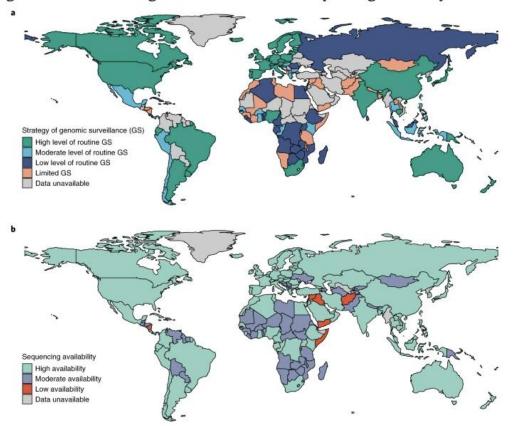
Map of sequencing tools available across global regions

Source: worldbank.org publication <a href="https://documents1.worldbank.org/curated/en/09971621222240993/pdf/P17618002a6bfd00c09e39087a18be85684.pdf">https://documents1.worldbank.org/curated/en/09971621222240993/pdf/P17618002a6bfd00c09e39087a18be85684.pdf</a>

based on data submitted to GISAID

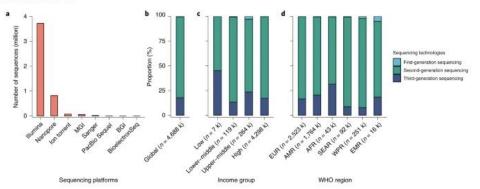
### **Geographic disparities and gaps in data sharing during the COVID-19 pandemic**

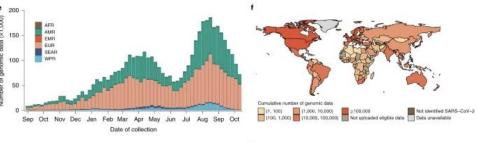
Fig. 1: Global SARS-CoV-2 genomic surveillance and sequencing availability.

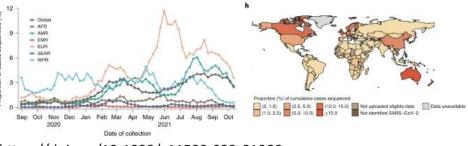


a, The global distribution of four strategies for SARS-CoV-2 genomic surveillance.
b, The global availability of SARS-CoV-2 sequencing. 'Data unavailable' include locations that do not belong to the 194 Member States or do not have applicable data. Data shown here are as of 31 October 2021.
Administrative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

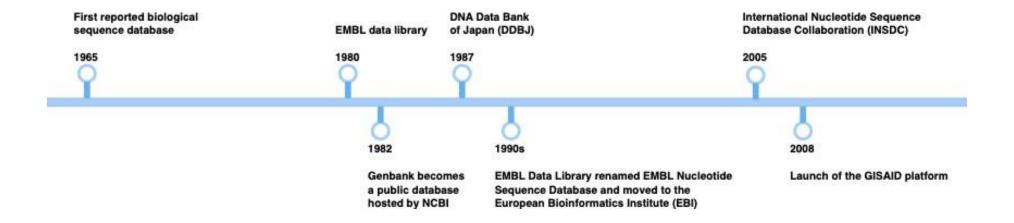




# **Biological sequences databases, a brief overview**

### **Historical timeline**





# The INSDC (International Nucleotide Sequence Database Collaboration)



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

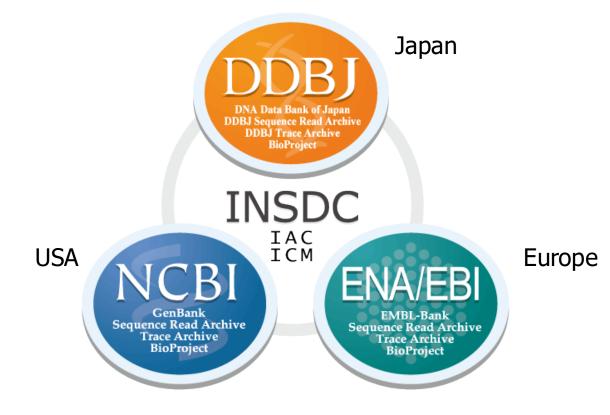
-No restriction or licencing fees attached to records

-All submitted records remain permanently accessible

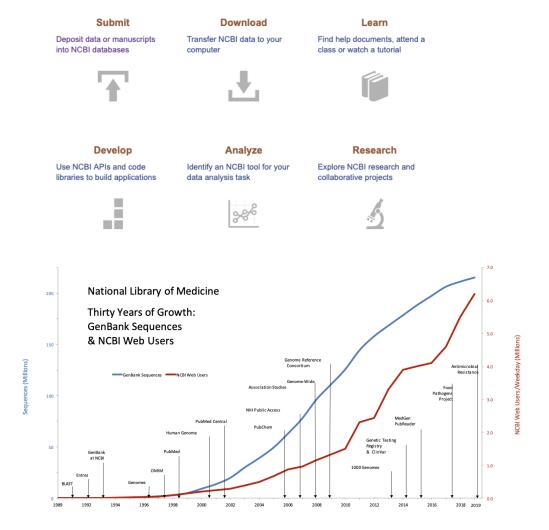
-Information disclosed to the web is fully public

-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



eccoc

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)

https://www.nlm.nih.gov/about/2021CJ\_NLM.pdf

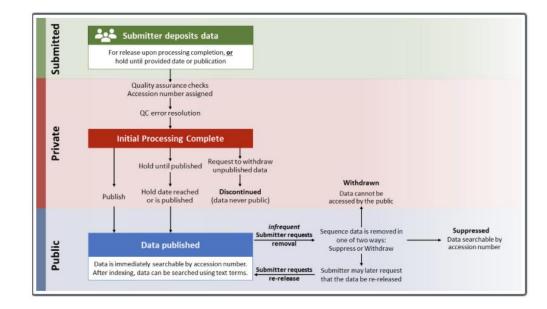
### **NCBI Genbank**



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.

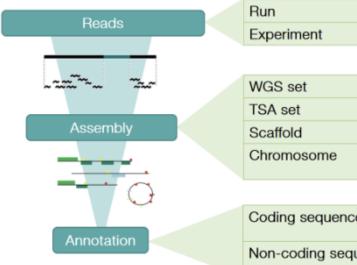
Provides all the sequence data information available in other INSDC databases

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.

| DBJ  |  |   |  |   |  |  |  |   |  |  |
|--|--|---|--|---|--|--|--|---|--|--|
| Bank of Japan  | out DDBJ   |   |  |   | How to Use   |  |  | Report/Sta  | itistics   |  |
| HOME > Search and Ar   | nalysis > ARSA > ARSA (  | Search Result)  |  |   |  |  |  |   |  |  |
| ARSA (Searc  | h Result)  |   |  |   |  |  |  |   |  |  |
| Search Condition   | on   |   |  |   |  |  |  |   |  |  |
| Search Result  |  |   |  |   |  |  |  |   |  |  |
| Facet  |  |   |  |   |  |  |  |   |  |  |
| List of Entries  |  |   |  |   |  |  |  |   |  |  |
| PrimaryAccession   | of founds: 26241 OFI<br>Number Definition                                  | SequenceLength  | MolecularType  | Organism                                  |  |  |  |   |  |  |
| PA370103 Defi<br>PA370105 Defi                                   | nition: WO 2022014620-7<br>nition: WO 2022014620-7                         | A/4: SARS-CoV-2 vaco<br>A/6: SARS-CoV-2 vaco                        | ine. SequenceLen<br>ine. SequenceLen                         | gth:3819 M<br>gth:3819 M                  | lolecularType:DNA Organism<br>lolecularType:DNA Organism                               | Severe acute respiratory syndrome c<br>Severe acute respiratory syndrome c<br>Severe acute respiratory syndrome c  | oronavirus 2<br>oronavirus 2   |   |  |  |
| PA370107 Defi<br>PA370109 Defi<br>PA370111 Defi                  | nition WO 2022014620-/   | A/10: SARS-CoV-2 vat<br>A/12: SARS-CoV-2 vat                        | cine. Sequencel.e  | ngth:3819                                 | MolecularTypeDNA Organis<br>lolecularTypeDNA Organis                                   | Severe acute respiratory syndrome c<br>Severe acute respiratory syndrome<br>Severe acute respiratory syndrome c  | coronavirus 2<br>oronavirus 2  |   |  |  |
| PA370100 Defi<br>PA370102 Defi<br>PA370104 Defi<br>PA370106 Defi | nition WO 2022014620-/   | A/3: SARS-CoV-2 vaco<br>A/5: SARS-CoV-2 vaco                        | ine. SequenceLen<br>ine. SequenceLen                         | gth:3819 M                                | olecularType DNA Organism<br>olecularType DNA Organism                                 | Severe acute respiratory syndrome c<br>Severe acute respiratory syndrome c<br>Severe acute respiratory syndrome c<br>Severe acute respiratory syndrome c | oronavirus 2<br>oronavirus 2   |   |  |  |
| PA370108 Defe  | nition WO 2022014620-/   | A/9: SARS-CoV-2 vaco  | ine. SequenceLen   | gth:3819 M                                | olecularType:DNA Organian  | Severe acute respiratory syndrome c  | oronavirus 2   |   |  |  |
| PD715743 Defi<br>OF935880 Defi<br>OF935882 Defi                  | nition:KR 102022001282   | 26-A/2: SARS-CoV-2 N  | EUTRALIZING BIND   | NG MOLECUL                                | E BINDING TO EPITOPES OF :   | SARS-CoV-2 SPIKE PROTEIN. Secure   | ncel.ength:7 Molecu  | arType PRT Or<br>arType PRT Or                      | ganism synthetic construct   |  |
| OF935888 Defi  | nition KR 102022001282<br>nition KR 102022001282<br>nition KR 102022001282 | 26-A/6: SARS-CoV-2 №<br>26-A/8: SARS-CoV-2 №<br>26-A/10: SARS-CoV-2 | EUTRALIZING BINDI<br>EUTRALIZING BINDI<br>NEUTRALIZING BINDI | NG MOLECUL<br>NG MOLECUL<br>DING MOLECU   | E BINDING TO EPITOPES OF<br>E BINDING TO EPITOPES OF<br>JLE BINDING TO EPITOPES OF     | SARS-CoV-2 SPIKE PROTEIN. Seque<br>SARS-CoV-2 SPIKE PROTEIN. Seque<br>SARS-CoV-2 SPIKE PROTEIN. Seque  | nceLength:19 Molec<br>nceLength:7 Molecu<br>enceLength:5 Molecu                      | ularType PRT C<br>larType PRT Or<br>ularType PRT C  | ganism synthetic construct<br>ganism synthetic construct<br>reganism synthetic construct                                     |  |
| OF935890 Defi<br>OF935892 Defi<br>OF935894 Defi                  | nition KR 102022001282   | 26-A/14: SARS-CoV-2<br>26-A/16: SARS-CoV-2                          | NEUTRALIZING BINI<br>NEUTRALIZING BINI                       | DING MOLECU                               | JLE BINDING TO EPITOPES OF<br>JLE BINDING TO EPITOPES OF                               | SARS-CoV-2 SPIKE PROTEIN. Sequ<br>SARS-CoV-2 SPIKE PROTEIN. Sequ   | enceLength:19 Mole<br>enceLength:7 Molec<br>enceLength:5 Molec                       | ularType:PRT C                                      | Organism synthetic construct<br>organism synthetic construct<br>organism synthetic construct                                 |  |
| OF935896 Defi<br>OF935898 Defi<br>OF935900 Defi<br>OF935902 Defi | nition KR 102022001282   | 26-A/20: SARS-CoV-2<br>26-A/22: SARS-CoV-2                          | NEUTRALIZING BING  | DING MOLECU                               | JLE BINDING TO EPITOPES OF<br>JLE BINDING TO EPITOPES OF                               | SARS-CoV-2 SPIKE PROTEIN. Sequ<br>SARS-CoV-2 SPIKE PROTEIN. Sequ   | enceLength:19 Mole<br>enceLength:7 Molec<br>enceLength:5 Molec<br>enceLength:19 Mole | ularType.PRT C                                      | Organism synthetic construct<br>Organism synthetic construct<br>Organism synthetic construct                                 |  |
| OF935902 Defi<br>OF935904 Defi<br>OF935906 Defi<br>OF935908 Defi | nition KR 102022001282   | 26-A/26: SARS-CoV-2<br>26-A/28: SARS-CoV-2                          | NEUTRALIZING BING<br>NEUTRALIZING BING                       | DING MOLECU                               | JLE BINDING TO EPITOPES OF<br>JLE BINDING TO EPITOPES OF                               | SARS-CoV-2 SPIKE PROTEIN. Sequ<br>SARS-CoV-2 SPIKE PROTEIN. Sequ   | encel ength? Molec   | ularType PRT 0                                      | Providence of the synthetic construct  |  |
| OF935910 Defi<br>OF935912 Defi<br>OF935914 Defi                  | nition KR 102022001282<br>nition KR 102022001282<br>nition KR 102022001282 | 26-A/32: SARS-CoV-2<br>26-A/34: SARS-CoV-2<br>26-A/36: SARS-CoV-2   | NEUTRALIZING BIN<br>NEUTRALIZING BIN<br>NEUTRALIZING BIN     | DING MOLECU<br>DING MOLECU<br>DING MOLECU | JLE BINDING TO EPITOPES OF<br>JLE BINDING TO EPITOPES OF<br>JLE BINDING TO EPITOPES OF | SARS-CoV-2 SPIKE PROTEIN. Segu<br>SARS-CoV-2 SPIKE PROTEIN. Segu<br>SARS-CoV-2 SPIKE PROTEIN. Segu   | enceLength:7 Molec<br>enceLength:5 Molec<br>enceLength:22 Mole                       | ularType PRT C<br>ularType PRT C<br>cularType PRT C | Organism synthetic construct<br>Organism synthetic construct<br>Organism synthetic construct                                 |  |
| OF935918 Defi<br>OF935920 Defi                                   | nition KR 102022001282<br>nition KR 102022001282<br>nition KR 102022001282 | 26-A/38: SARS-CoV-2<br>26-A/40: SARS-CoV-2<br>26-A/42: SARS-CoV-2   | NEUTRALIZING BIN<br>NEUTRALIZING BIN<br>NEUTRALIZING BIN     | DING MOLECU<br>DING MOLECU<br>DING MOLECU | JLE BINDING TO EPITOPES OF<br>JLE BINDING TO EPITOPES OF<br>JLE BINDING TO EPITOPES OF | SARS-CoV-2 SPIKE PROTEIN. Sequ<br>SARS-CoV-2 SPIKE PROTEIN. Sequ<br>SARS-CoV-2 SPIKE PROTEIN. Sequ   | encel ength 22 Mole  | ularType PRT C<br>cularType PRT                     | Arganism synthetic construct<br>Organism synthetic construct   |  |
| OF935924 Defi<br>OF935926 Defi                                   | nition KR 102022001282   | 26-A/46: SARS-CoV-2<br>26-A/48: SARS-CoV-2                          | NEUTRALIZING BINI<br>NEUTRALIZING BINI                       | DING MOLECU                               | JLE BINDING TO EPITOPES OF<br>JLE BINDING TO EPITOPES OF                               | SARS-CoV-2 SPIKE PROTEIN. Seque<br>SARS-CoV-2 SPIKE PROTEIN. Seque   | enceLength 5 Molec<br>enceLength 22 Mole   | ularType:PRT C<br>cularType:PRT C                   | Reginism:synthetic construct<br>Organism:synthetic construct<br>Organism:synthetic construct                                 |  |
| OF935930 Defi<br>OF935932 Defi                                   | nition KR 102022001282   | 26-A/52: SARS-CoV-2<br>26-A/54: SARS-CoV-2                          | NEUTRALIZING BIN   | DING MOLECU                               | JLE BINDING TO EPITOPES OF<br>JLE BINDING TO EPITOPES OF                               | SARS-CoV-2 SPIKE PROTEIN. Sequ<br>SARS-CoV-2 SPIKE PROTEIN. Sequ   | encel.ength:5 Molec  | ularType:PRT C                                      | Inganism synthetic construct<br>Inganism synthetic construct<br>Organism synthetic construct                                 |  |
| OF935936 Defi<br>OF935938 Defi                                   | nition KR 102022001282   | 26-A/58 SARS-CoV-2<br>26-A/60 SARS-CoV-2                            | NEUTRALIZING BIN   | DING MOLECU                               | JLE BINDING TO EPITOPES OF<br>JLE BINDING TO EPITOPES OF                               | SARS-CoV-2 SPIKE PROTEIN. Sequ<br>SARS-CoV-2 SPIKE PROTEIN. Sequ   | enceLength:7 Molec<br>enceLength:5 Molec<br>enceLength:22 Mole<br>enceLength:7 Molec | ularType:PRT C<br>cularType:PRT                     | Arganism synthetic construct<br>Arganism synthetic construct<br>Organism synthetic construct                                 |  |
| OF935942 Defi<br>OF935944 Defi<br>OF935946 Defi                  | nition KR 102022001282   | 26-A/64: SARS-CoV-2<br>26-A/66: SARS-CoV-2                          | NEUTRALIZING BINI<br>NEUTRALIZING BINI                       | DING MOLECU                               | JLE BINDING TO EPITOPES OF<br>JLE BINDING TO EPITOPES OF                               | SARS-CoV-2 SPIKE PROTEIN. Sequ<br>SARS-CoV-2 SPIKE PROTEIN. Sequ   | enceLength5 Molec  | ular Type PRT C                                     | Arganism:synthetic construct<br>Arganism:synthetic construct<br>Organism:synthetic construct<br>Arganism:synthetic construct |  |
| OF935948 Defi<br>OF935950 Defi<br>OF935952 Defi                  | nition KR 102022001282<br>nition KR 102022001282<br>nition KR 102022001282 | 26-A/70: SARS-CoV-2<br>26-A/72: SARS-CoV-2<br>26-A/74: SARS-CoV-2   | NEUTRALIZING BINI<br>NEUTRALIZING BINI<br>NEUTRALIZING BINI  | DING MOLECU<br>DING MOLECU<br>DING MOLECU | JLE BINDING TO EPITOPES OF<br>JLE BINDING TO EPITOPES OF<br>JLE BINDING TO EPITOPES OF | SARS-CoV-2 SPIKE PROTEIN. Segu<br>SARS-CoV-2 SPIKE PROTEIN. Segu<br>SARS-CoV-2 SPIKE PROTEIN. Segu   | enceLength 5 Molec<br>enceLength 22 Mole<br>enceLength 7 Molec                       | ularType PRT C<br>cularType PRT<br>ularType PRT C   | Organism synthetic construct<br>Organism synthetic construct   |  |
| OF935954 Defi<br>OF935956 Defi<br>OF935958 Defi                  | nition KR 102022001282<br>nition KR 102022001282<br>nition KR 102022001282 | 26-A/76: SARS-CoV-2<br>26-A/78: SARS-CoV-2<br>26-A/80: SARS-CoV-2   | NEUTRALIZING BIN<br>NEUTRALIZING BIN<br>NEUTRALIZING BIN     | DING MOLECU<br>DING MOLECU<br>DING MOLECU | JLE BINDING TO EPITOPES OF<br>JLE BINDING TO EPITOPES OF<br>JLE BINDING TO EPITOPES OF | SARS-CoV-2 SPIKE PROTEIN. Sequ<br>SARS-CoV-2 SPIKE PROTEIN. Sequ<br>SARS-CoV-2 SPIKE PROTEIN. Sequ   | encel.ength 5 Molec<br>encel.ength 22 Mole<br>encel.ength 7 Molec                    | ularType PRT C<br>cularType PRT<br>ularType PRT C   | Organism synthetic construct<br>Organism synthetic construct   |  |
| OF935960 Defi<br>OF935962 Defi                                   | nition KR 102022001282   | 26-A/82: SARS-CoV-2   | NEUTRALIZING BINE  | DING MOLECU                               | ILE BINDING TO EPITOPES OF   | SARS-CoV-2 SPIKE PROTEIN. Sequ   | enceLength 5 Molec<br>enceLength 22 Mole   | ularType PRT C                                      | Organism synthetic construct<br>Organism synthetic construct   |  |

## 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<br>organelles, plasmids and viral segments) |
|                      |  |
| Coding sequences     | Sequence regions reported as being protein-<br>coding regions                      |
| Non-coding sequences | Sequence regions reported as representing non-<br>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

### **Data types shared by the INSDC**



| Data type                | DDBJ Center           | EMBL-EBI   | NCBI                  |  |
|--------------------------|-----------------------|--|-----------------------|--|
| Next generation reads    | Sequence Read Archive |  | Sequence Read Archive |  |
| Sequence Read<br>Archive | Trace Archive         | European Nucleotide  | Trace Archive         |  |
| Annotated sequence       | DDBJ                  | <u>Archive (ENA)</u>   | <u>GenBank</u>        |  |
| Samples                  | <u>BioSample</u>      |  | <u>BioSample</u>      |  |
| Studies                  | <u>BioProject</u>     |  | <u>BioProject</u>     |  |
|                          |                       | Overall ENA Research Project       STUDY       Bata Analysis       EXPERIMENT       Raw Reads       Bata Analysis       SAMPLE       Sequenced Biomaterial | 25                    |  |

## 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 Colum         | n Selection                     |                              |                                  |              |                    |                 |              | ~               |
|--------------------|---------------------------------|------------------------------|----------------------------------|--------------|--------------------|-----------------|--------------|-----------------|
| Select Galax       | y Server : Curre                | ent selection is <b>http</b> | s://usegalaxy                    | .org/tool_ri | unner?tool_id=eb   | i_sra_main      |              | ~               |
| Download repo      | ort: JSON                       | TSV                          |                                  |              | Downloa            | ad Files as ZIP | Download se  | lected files    |
| Study<br>Accession | Secondary<br>Study<br>Accession | Sample<br>Accession          | Secondary<br>Sample<br>Accession | Tax Id       | Scientific<br>Name | Center<br>Name  | First Public | Last<br>Updated |
| PRJEB1234          | ERP002070                       | SAMEA1905789                 | ERS200899                        | 4555         | Setaria<br>italica | NCGR            | 2013-06-06   | 2018-11-1       |
| PRJEB1234          | ERP002070                       | SAMEA1905397                 | ERS200900                        | 4555         | Setaria<br>italica | NCGR            | 2013-06-06   | 2018-11-1       |

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

### The Sample object





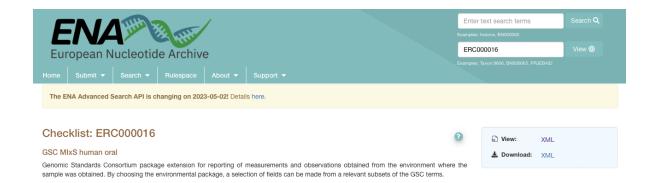
A sample object represents a unit of biomaterial associated to a single collection event

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 <u>https://www.ebi.ac.uk/ena/browser/checklists</u>

### **Example of a ENA sample checklist**



Checklist Fields Filter fields... Q Field Field Name (Field Restriction) Requirement (Units) Format Filter by type: collection date regular expression @ mandatory geographic location (country and/or (7) text choice options 🔹 🔻 mandatory seal restricted regular expression (2) geographic location (latitude) mandatory DD restricted regular expression @ geographic location (longitude) mandatory DD text geographic location (region and 0 free text optional locality) broad-scale environmental context mandatory (7) free text local environmental context free text mandatory environmental medium ⑦ free text mandatory source material identifiers (7) free text optional sample material processing ⑦ free text optional isolation and growth condition ⑦ free text optional propagation (?) free text optional restricted regular expression (2) amount or size of sample collected (2) optional options 🔹



Structured data collection at ENA is implemented using sample checklists.

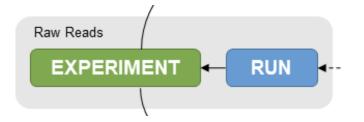
Sample data must be collected in a csv or XML format.

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





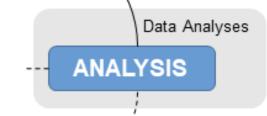
Raw reads are described on ENA by the 'Experiment' and 'Run' objects

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.

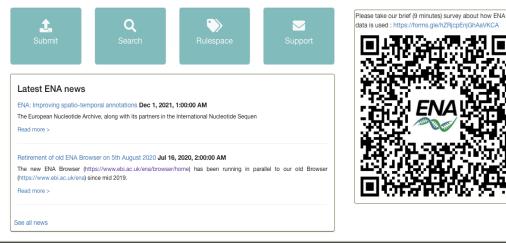
#### **Data search services at ENA**



#### European Nucleotide Archive

The European Nucleotide Archive (ENA) provides a comprehensive record of the world's nucleotide sequencing information, covering raw sequencing data, sequence assembly information and functional annotation. More about ENA.

Access to ENA data is provided through the browser, through search tools, through large scale file download and through the API.





The European Nucleotide Archive (ENA) is part of the ELIXIR infrastructure The ENA is an ELIXIR Core Data Resource. Learn more .

The European Nucleotide Archive (ENA) is a Global Core Biodata Resource The ENA is a GBC Global Core Biodata Resource. <u>Learn more .</u>

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

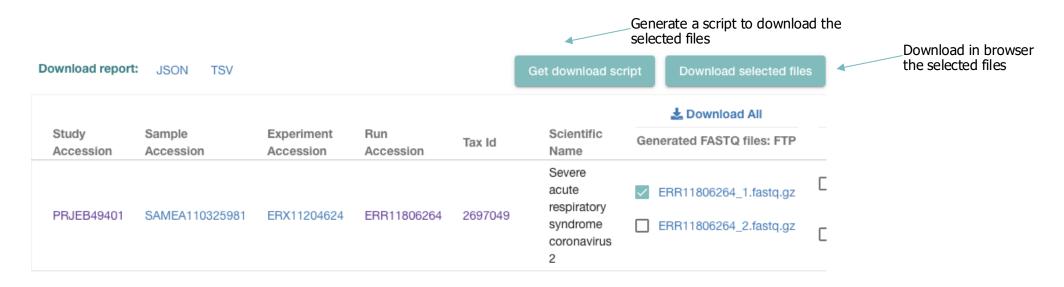


- Records searchable by accession numbers and free text search
- Links to available formats to download, crosslinks and related resources
- Summary reports downloadable in tabular or json formats.
- Advanced search available via the browser or using the API. <u>https://www.ebi.ac.uk/ena/portal/api/</u>

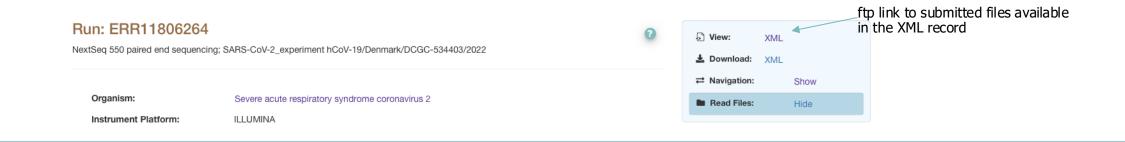
## Data retrieval services (1/2)



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



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

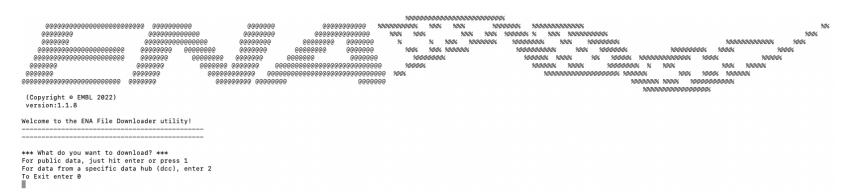


# Data retrieval services (2/2)



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

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

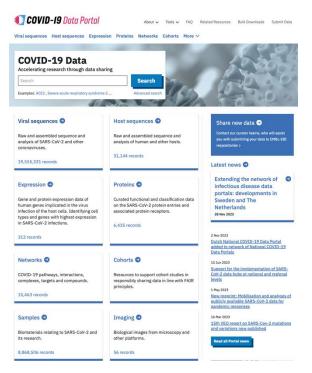


- Java-based tool that can run in interactive or command line mode.
- Accepts single, list of accessions, advanced search queries, parallel downloads, retries
- Supports FTP and Aspera transfer protocols

## **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.

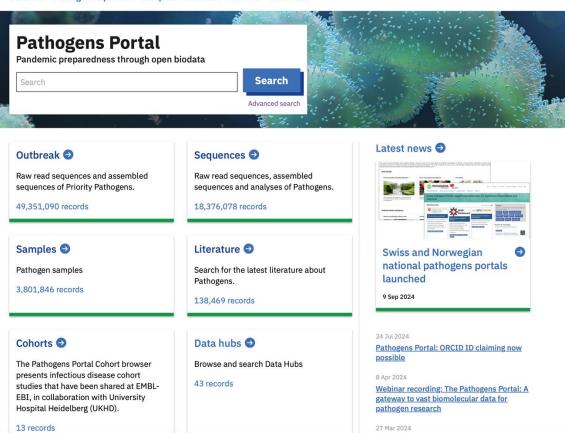


#### https://www.covid19dataportal.org

>9 million assembled sequences 11 national COVID-19 portals



Outbreak Pathogen sequences Samples Literature Cohorts Data Hubs



About V

Tools ✓ FAQ Submit data Related resources

New paper presents Data Hubs

Login

#### ENA Pathogens Platform

All pathogen, all disease approach Hosts, vectors and pathogens

Antimicrobial resistance

Outbreaks

https://www.pathogensportal.org

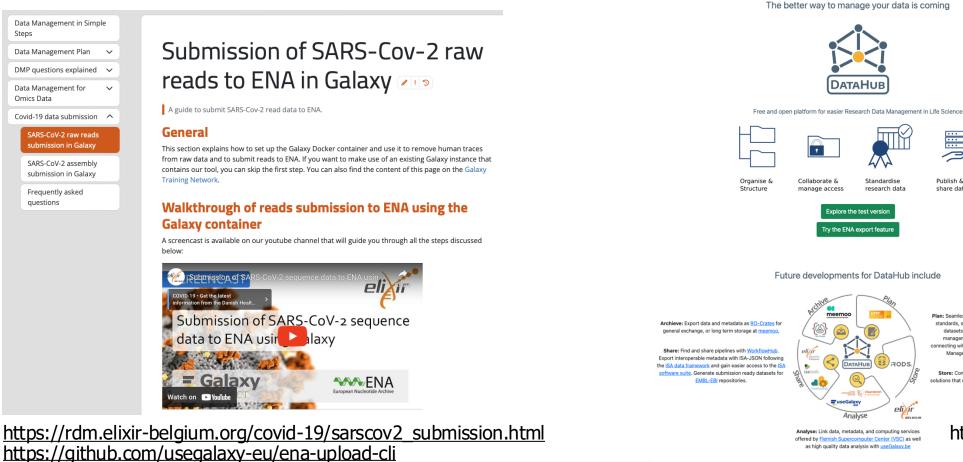
# Third party access point to ENA submission systems



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.

Steps



Publish & share data Plan: Seamless compliance with a variety of community standards, streamlining deposition of complex omics datasets to deposition databases. Easier data management plan design and maintenance by

Store: Connect with an expanding pool of storage solutions that use iRODS, while recording meaningful and rich metadata

connecting with DMPonline, and the ELIXIR Belgium Data

Management Plan Template and Converter.

https://datahub.elixir-belgium.org

## **GISAID (Global Initiative for Sharing Avian Influenza** Data)

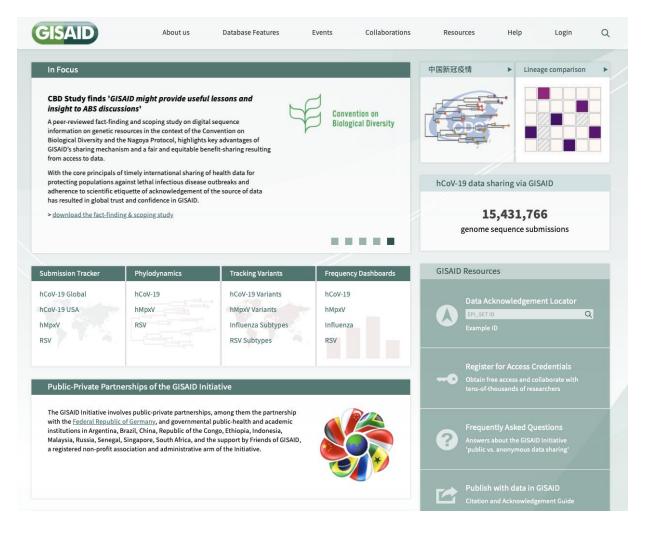
Global Initiative promoting rapid sharing of data from influenza, SARS-CoV-2 and more recently monkeypox viruses.

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/





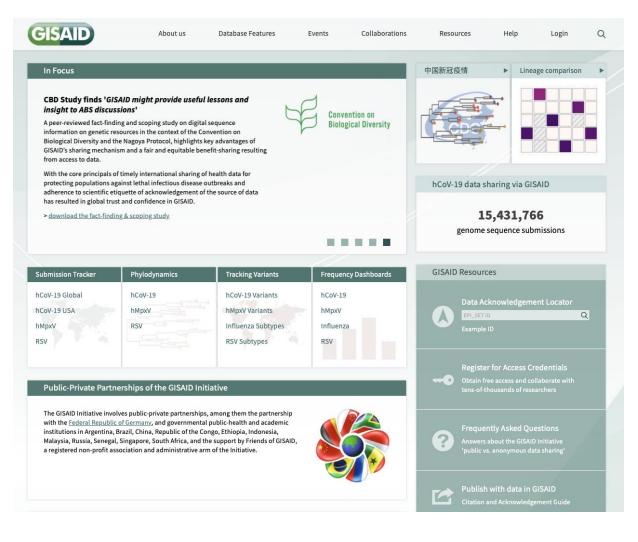
# **GISAID (Global Initiative for Sharing Avian Influenza** Data)

#### GISAID has several databases

- EpiFlu
- EpiCoV
- EpiRSV
- EpiPox
- Different levels of access

Different tools/features for each databases

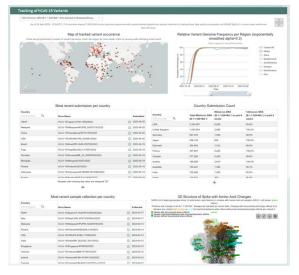
- Dependent on virus biology



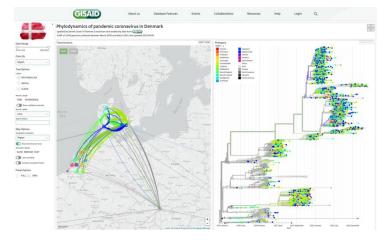
### Free-access features and tools available on GISAID



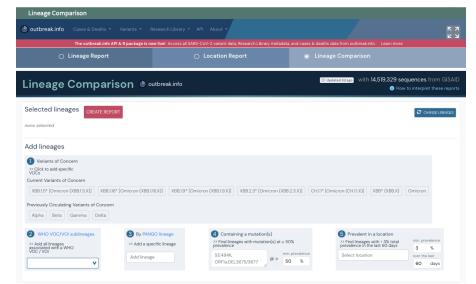
#### Submission tracker



#### Phylodynamics



#### Lineage comparison



FluSurver



#### Important usage notes:

The main application scenario for FluGurve is to highlight phenotyscially or epidemiologizally interesting candidate mutations for Influther research and should ideally be combined with beginners and service and any predicted phenotypes. Importantly, any direct diagnostic use, assumed averity or recommendation on patient treatment should not be based solely on these computational predictions. Dur candler inference supports and for annotation treatment should not be based solely on these computational predictions. Dur candler inference support and monitorial treatment should not be based solely on and reliable results are current surveillance sequences with very close relation to used vaccine strains, including some candidates for avian full (including ISNI), ISNIS and (TSNI) and novel reasonation strains with IN3NX.

Please take a look at the <u>Frequently Asked Questions</u> and <u>Tutorial</u> if you are new to FluSurver. You could also look at this <u>NA</u> drug susceptibility example analysis walkthrough starting from **GISAID** and the **GISAID** access summary poster

Paste your protein or nucleotide FASTA sequence(s) into the text area below. (Sample FASTA sequences: 2009 H1N1 NA and HA)

#### OR upload your protein or nucleotide sequences in a FASTA file

Choose File no file selected

The server can automatically determine the type of input (either protein or nucleotide) and the closest reference sequence among currer vaccine strains to compare. Also motures of genes/proteins (e.g., HA and HA or all genes of the same patient) can be provided as input. To compare with more remotily related sequences/strains, it is possible to select a specific reference strain by choosing below.

Compare with: Automatic detection of closest reference (larger selection of strains, not always full genomes)

#### Additional settings:

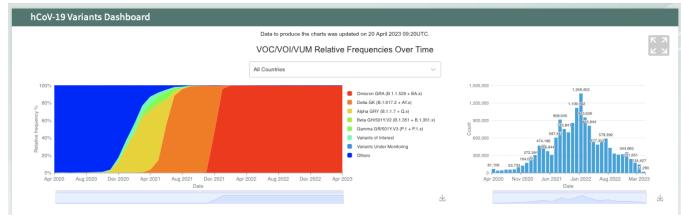
ignore low quality bases for nucleotide input (indicated by lower case, except for all lower case sequences)

Submit Reset (estimated time needed: ~2 seconds per sequence in automatic mode)



Developed by A\*STAR Bioinformatics Institute (BII), Singapore Copyright © 2023 BII. All Rights Reserved.

#### Dashboard



### **EpiFlu and EpiCoV**



#### EpiFlu

| EpiFlu<br>Basic ( |                       |                 |             |  | W™ E         | piPox'**        | My Pro    |          |              |         |              |       |      |      |    |    |
|-------------------|-----------------------|-----------------|-------------|--|--------------|-----------------|-----------|----------|--------------|---------|--------------|-------|------|------|----|----|
| Basic 1           |                       |                 | k to res    | ults   Workse                                  | te i lle     | load   E        | latch U   | bead     | CIT          | Jpload  | L Eat        | tings | Anal | uele |    |    |
|                   | filters               |                 | ik to resi  |  |              |                 | aten o    | pioau    |              | opioau  | 1 36         | ungs  | Alla | y515 |    |    |
|                   | ined search           | Se              | lect        | 0  |              |                 |           |          |              |         |              |       |      |      |    |    |
| earch             | n in                  |                 |             | iles Worksets                                  |              |                 |           |          |              |         |              |       |      |      |    |    |
| Search            | n patterns            |                 |             |  |              |                 |           |          |              |         |              |       |      |      |    |    |
|                   |                       | Typ             | н           | N Lineage                                      | н            | lost            |           | Locat    | ion          | CI      | Clades       |       |      |      |    |    |
|                   |                       | A               | 1           | 1  |              | all-            |           | -all-    |              | 0       |              |       |      |      |    |    |
|                   |                       | B               | 2           | 2  |              | Human           |           | Africa   |              | 1       | 1            |       |      |      |    |    |
|                   |                       | -               | 4           | 4  | 1            | Avian           |           | Asia     |              | 1.      | 1.1          |       |      |      |    |    |
|                   |                       |                 | 5           | 5  |              | Chicken         |           | Europ    | e<br>America |         | 1.2          |       |      |      |    |    |
|                   |                       |                 | 7           | 7  |              | Duck            |           | Ocea     | nia          | 2       | 1.2          |       |      |      |    |    |
|                   |                       |                 | 8           | 8  |              | Eagle<br>Falcon |           | South    | n America    |         | 1.3<br>1.3.1 |       |      |      |    |    |
|                   |                       |                 | 10          | 10   |              | Goose           |           |          |              |         | 1.3.2        |       |      |      |    |    |
| dditi             | ional filter          | 5               |             |  |              |                 |           |          |              |         |              |       |      |      |    |    |
|                   | tion date<br>-MM-DD)  | From            | n 2019-     | -12-31   |              | To 2            | 023-01-0  | 1        |              |         |              |       |      |      |    |    |
|                   | ssion date<br>-MM-DD) | From            | n 🗌         |  |              | To              |           |          |              |         | 1            |       |      |      |    |    |
|                   | ating Labor           | atory [fate     | hanistan k  | Kabul] National Put                            | lic Health I | aboratory       |           |          |              |         |              |       |      | -    |    |    |
| ,                 |                       | [Alb            | ania, Tiran | a] Institute of Publi                          | c Health     | ,               |           |          |              |         |              |       |      |      |    |    |
|                   |                       |                 |             | ns] Institut Pasteur (<br>moa. Fagaala] LBJ    |              | dicine Cent     | re        |          |              |         |              |       |      |      |    |    |
|                   |                       |                 |             | enos Aires] Institut                           |              |                 |           | iosas C. | G.Malbra     | n       |              |       |      |      |    |    |
| ubmit             | tting Labor           |                 |             | enos Aires] Institut                           |              |                 |           |          |              |         |              |       |      |      |    |    |
|                   |                       |                 |             | enos Aires] Instituti<br>enos Aires] Instituti |              |                 |           |          |              | Malbrar | 1            |       |      |      |    |    |
|                   |                       |                 |             | ar del Plata] Instituti                        |              |                 | logia Jui | in Hecto | r Jara       |         |              |       |      |      |    |    |
| -                 | ed Seamen             | 0               |             | uarina) Royal Darv                             |              |                 |           |          |              |         |              |       |      |      |    |    |
| equin             | eu seymen             |                 | _           | mplete Min Leng                                | _            | nin 🖸 Ma        | HE UP     | 3        |              |         |              |       |      |      |    |    |
|                   |                       |                 | only cor    | inprece min ceny                               |              |                 |           |          |              |         |              |       |      |      |    |    |
|                   |                       |                 |             |  |              |                 |           |          |              |         |              |       |      |      |    |    |
|                   | edit Name             |                 |             | Isolate ID                                     | Subtype      | Passage         | PB2       | PB1      | PA           | HA      | NP           | NA    | MP   | NS   | HE | P3 |
|                   | ·                     | laLaMancha/3731 | 2022        | EPI_ISL_15542438                               | HSN1         | Original sar    |           | 2,274    | 2,151        | 1,704   | 1,497        | 1,410 | 982  | 838  | -  | -  |
|                   | No. Atrida            | GARI-4571/2021  |             | EPI_15L_14223821                               | H5N1         |                 | 2,250     | 2,271    | 2,148        | 1,701   | 1,494        | 1,347 | 982  | 838  |    | -  |
|                   |                       | nd/215201407/20 |             | EPI 15L 8799552                                | H5N1         | original        | 2,279     | 2,178    | 2,151        | 1,703   | 1,497        | 1,450 | 982  | 838  |    | -  |

#### EpiCoV

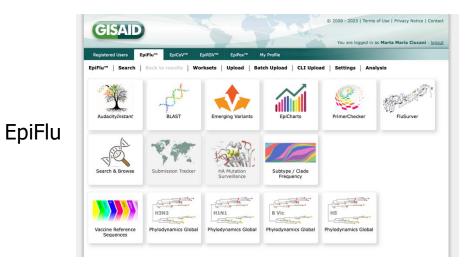
|        | gistered Users | EpiFlu™ EpiCoV™        | EpiPox'   | " My Profile                         |               |              |  |        |       |                   |                        |
|--------|----------------|------------------------|-----------|--------------------------------------|---------------|--------------|--|--------|-------|-------------------|------------------------|
| iearc  |                |                        |           |                                      |               |              |  |        |       | F                 | leset filter           |
| EPI_IS | SI ID          | Vin                    | us name   |                                      | EPI           | SET ID       |  |        |       | Complete          |                        |
| ocati  |                | Denmark                |           |                                      | Hos           |              |  |        | _     | High coverage     |                        |
| Collec | tion           | to                     |           | Submission                           |               | to           |  |        | - ĭ   | Low coverage ex   | cluded®                |
| Clade  |                | all 🛟 Lineage          |           | Variant                              |               |              |  |        | •     | With patient stat | us 🔊                   |
| IA Su  | ubstitutions 💿 |                        | ~         | Nucl Mutations                       |               |              |  |        | ¥ 0   | Collection date c | omplete 🔊              |
|        |                |                        |           |                                      |               |              |  |        |       | Under investigat  | ion                    |
| Text   | Search         |                        |           |                                      |               |              |  |        |       |                   |                        |
|        |                |                        |           |                                      |               |              |  |        | 6     |                   |                        |
|        | Virus name     |                        | Passage d | Accession ID                         | Collection da | Submission [ | ()   | Length | Host  | Location          | Originating            |
| 0      | hCoV-19/Der    | nmark/DCGC-645794/2023 | Original  | EPI_ISL_17482031                     | 2023-04-07    | 2023-04-16   | ٩  | 29,649 | Human | Europe / Denmar   | Molekylær              |
|        | hCoV-19/Der    | nmark/DCGC-645793/2023 | Original  | EPI_ISL_17482030                     | 2023-04-07    | 2023-04-16   | ٩  | 29,649 | Human | Europe / Denmar   | Molekylær              |
|        | hCoV-19/Der    | nmark/DCGC-645792/2023 | Original  | EPI_ISL_17482029                     | 2023-04-07    | 2023-04-16   | ٩  | 29,640 | Human | Europe / Denmar   | Molekylær              |
|        | hCoV-19/Der    | nmark/DCGC-645791/2023 | Original  | EPI_ISL_17482028                     | 2023-04-03    | 2023-04-16   | ٩  | 29,646 | Human | Europe / Denmar   | Molekylær              |
|        | hCoV-19/Der    | nmark/DCGC-645790/2023 | Original  | EPI_ISL_17482027                     | 2023-03-26    | 2023-04-16   | ٩  | 28,578 | Human | Europe / Denmar   | Molekylær              |
|        | hCoV-19/Der    | nmark/DCGC-645788/2023 | Original  | EPI_ISL_17482026                     | 2023-04-03    | 2023-04-16   | ٩  | 29,684 | Human | Europe / Denmar   | Departmer              |
|        | hCoV-19/Der    | nmark/DCGC-645787/2023 | Original  | EPI_ISL_17482025                     | 2023-04-06    | 2023-04-16   | ٩  | 29,649 | Human | Europe / Denmar   | Molekylær              |
|        | hCoV-19/Der    | nmark/DCGC-645786/2023 | Original  | EPI_ISL_17482024                     | 2023-04-05    | 2023-04-16   | ٢  | 29,649 | Human | Europe / Denmar   | Molekylær              |
|        | hCoV-19/Der    | nmark/DCGC-645785/2023 | Original  | EPI_ISL_17482023                     | 2023-03-30    | 2023-04-16   | ٩  | 29,649 | Human | Europe / Denmar   | Molekylær              |
|        |                | nmark/DCGC-645784/2023 | Original  | EPI_ISL_17482022                     | 2023-04-01    | 2023-04-16   | ٩  | 29,649 | Human | Europe / Denmar   | Molekylær              |
|        |                | nmark/DCGC-645783/2023 | Original  | EPI_ISL_17482021                     | 2023-03-28    | 2023-04-16   | 0  | 29,652 | Human |                   | Molekylær              |
|        |                | nmark/DCGC-645782/2023 | Original  | EPI_ISL_17482020                     | 2023-04-05    | 2023-04-16   | ٩  | 29,649 | Human |                   | Departmer              |
|        |                | nmark/DCGC-645781/2023 | Original  | EPI_ISL_17482019                     | 2023-04-06    | 2023-04-16   | ٩  | 29,678 | Human |                   | Departmer              |
|        |                | nmark/DCGC-645780/2023 | Original  | EPI_ISL_17482018                     | 2023-04-05    | 2023-04-16   | ٩  | 29,684 | Human | coreport o entria | Departmer              |
|        |                | nmark/DCGC-645779/2023 | Original  | EPI_ISL_17482017                     | 2023-04-08    | 2023-04-16   | 0  | 29,652 | Human |                   | Molekylær              |
|        |                | nmark/DCGC-645778/2023 | Original  | EPI_ISL_17482016                     | 2023-04-02    | 2023-04-16   | ٩  | 29,648 | Human |                   | Molekylær              |
|        |                | nmark/DCGC-645777/2023 | Original  | EPI_ISL_17482015                     | 2023-04-03    | 2023-04-16   | 0  | 29,687 | Human |                   | Departmer              |
|        |                | nmark/DCGC-645776/2023 | Original  | EPI_ISL_17482014                     | 2023-03-31    | 2023-04-16   | ٢  | 29,649 | Human |                   | Departmer              |
|        |                | nmark/DCGC-645775/2023 | Original  | EPI_ISL_17482013                     | 2023-04-06    | 2023-04-16   | ٢  | 29,684 | Human |                   | Departmer              |
|        |                | nmark/DCGC-645774/2023 | Original  |                                      | 2023-04-07    | 2023-04-16   | <ul><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><li>(1)</li><l< td=""><td>29,684</td><td>Human</td><td></td><td>Departmer</td></l<></ul> | 29,684 | Human |                   | Departmer              |
|        |                | nmark/DCGC-645773/2023 | Original  | EPI_ISL_17482011<br>EPI_ISL_17482010 | 2023-04-01    | 2023-04-16   | 1  | 29,643 | Human |                   | Molekylær<br>Departmer |
|        |                | nmark/DCGC-645770/2023 | Original  | EPI_ISL_17482010<br>EPI_ISL_17482009 | 2023-04-09    | 2023-04-16   | <  | 29,696 | Human |                   | Departmer              |
|        |                | nmark/DCGC-645769/2023 | Original  | EPI_ISL_17482009<br>EPI_ISL_17482008 | 2023-04-08    | 2023-04-16   | <u>م</u>   | 29,696 | Human |                   | Departmer              |
|        |                | nmark/DCGC-645768/2023 | Original  | EPI_ISL_17482005                     | 2023-04-09    | 2023-04-16   | 1  | 29,684 | Human |                   | Departmer              |
|        |                | mark/DCGC-645767/2023  | Original  | EPI ISL 17482006                     | 2023-04-08    | 2023-04-16   | 1  | 28,575 | Human |                   | Departmer              |
|        |                | nmark/DCGC-645766/2023 | Original  | EPI_ISL_17482005                     | 2023-04-04    | 2023-04-16   | 1  | 29,684 | Human |                   | Departmer              |
|        | hCoV-19/Der    | nmark/DCGC-645765/2023 | Original  | EPI_ISL_17482004                     | 2023-04-09    | 2023-04-16   | ٢  | 29,643 | Human | Europe / Denmar   | Molekylær              |
|        | hCoV-19/Der    | nmark/DCGC-645764/2023 | Original  | EPI_ISL_17482003                     | 2023-04-03    | 2023-04-16   | ٢  | 29,649 | Human |                   | Molekylær              |
|        | hCoV-19/Der    | nmark/DCGC-645763/2023 | Original  | EPI_ISL_17482002                     | 2023-03-29    | 2023-04-16   | ٩  | 29,648 | Human | Europe / Denmar   | Molekylær              |
|        | hCoV-19/Der    | nmark/DCGC-645762/2023 | Original  | EPI_ISL_17482001                     | 2023-03-29    | 2023-04-16   | ٩  | 29,649 | Human | Europe / Denmar   | Molekylær              |
|        | hCoV-19/Der    | nmark/DCGC-645761/2023 | Original  | EPI_ISL_17482000                     | 2023-04-04    | 2023-04-16   | ٢  | 29,684 | Human | Europe / Denmar   | Departmer              |
|        | hCoV-19/Der    | nmark/DCGC-645760/2023 | Original  | EPI_ISL_17481999                     | 2023-04-08    | 2023-04-16   | ٢  | 29,649 | Human | Europe / Denmar   | Molekylær              |
|        | hCoV-19/Der    | nmark/DCGC-645759/2023 | Original  | EPI_ISL_17481998                     | 2023-04-05    | 2023-04-16   | ٩  | 28,575 | Human | Europe / Denmar   | Molekylær              |
|        | hCoV-19/Der    | nmark/DCGC-645758/2023 | Original  | EPI_ISL_17481997                     | 2023-03-30    | 2023-04-16   | 0  | 29,652 | Human | Europe / Denmar   | Molekylær              |
|        | hCoV-19/Der    | nmark/DCGC-645757/2023 | Original  | EPI_ISL_17481996                     | 2023-03-31    | 2023-04-16   | ٩  | 29,649 | Human | Europe / Denmar   | Molekylær              |
|        | hCoV-19/Der    | nmark/DCGC-645756/2023 | Original  | EPI_ISL_17481995                     | 2023-04-01    | 2023-04-16   | ٩  | 29,667 | Human | Europe / Denmar   | Molekylær              |
|        | hCoV-19/Der    | nmark/DCGC-645755/2023 | Original  | EPI ISL 17481994                     | 2023-04-01    | 2023-04-16   | ٢  | 29.649 | Human | Europe / Denmar   | Molekvlær              |

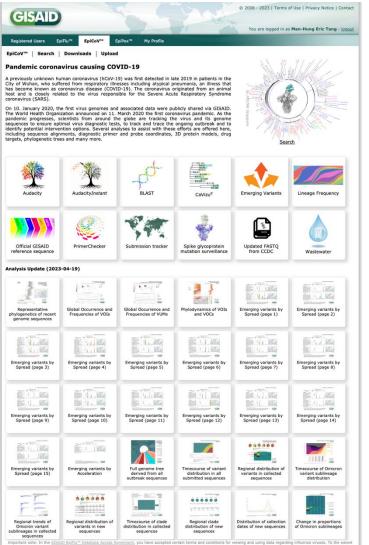
#### Online resource to search and download assemblies and raw reads and metadata

| Virus detail                     |   |
|----------------------------------|---|
| Virus name:                      | hCoV-19/Denmark/DCGC-127299/2021  |
| Accession ID:                    | EPI_ISL_2971682   |
| Type:                            | betacoronavirus   |
| Clade:                           | GK  |
| Pango Lineage:                   | AY.78 (Pango v.4.2 PLEARN-v1.19), Delta (B.1.617.2-like) (Scorpio)  |
| AA Substitutions:                | Spike 06140, Spike D65001, Spike E159G, Spike F157del, Spike L4529, Spike P6819, Spike R158del, Spike T19R,<br>Spike T965, Spike T476K, M1827, N0530, N 03777, N 0215C, N R2030, N338 1367, N3532B, N573 1210, N57a<br>V82A, N57b T401, N5F3 A4685, N5F3 P1228L, N5F3 P14666, NSF4 T4821, N5F4 V167L, N5F6 T77A, N5F10 1557,<br>N5F12 A677V, N5F12 A5815, N5F12 G6715, N5F12 1144V, N5F12 27321, N5F13 A577V, N5F14 A584V, N5F14 A587V, N5F144, N5F144V, N |
| Nucl Mutations:                  | G210T_C241T_T317C_C302T_G4181T_C6402T_C7124T_C8886T_69653T_C1002874_T12016_A1133G6_C12781T_<br>T31988C_C13730T_A1378D6_C14406T_G15181T_G15415A_C146465_C17340T_C16220T_C216186C_C21846T_<br>C21807F_6422028_22033_A220340_T228170_C22985A_A22403G_C23804G, G24410A_C25468T_T25580C_<br>T28787C_T27838G_C27752T_C27874T_64228248_28253_6428271_28271_A28461G_G28881T_G28815B_G28402T_<br>C28742T   |
| Variant:                         | VOC Delta GK (B.1.617.2+AY.*) first detected in India   |
| Passage details/history:         | Original  |
| Sample information               |   |
| Collection date:                 | 2021-07-12  |
| Location:                        | Europe / Denmark / Hovedstaden  |
| Host:                            | Human   |
| Additional location information: | 4   |
| Gender:                          | unknown   |
| Patient age:                     | unknown   |
| Patient status:                  | unknown   |
| Specimen source:                 |   |
| Additional host information:     |   |
| Sampling strategy:               |   |
| Outbreak:                        |   |
| Last vaccinated:                 |   |
| Treatment:                       |   |
| Sequencing technology:           | ILLUMINA  |
| Assembly method:                 | ARTIC-ncov2019 v1.1.0   |
| Back                             | Contact Submitter   |

#### **Tools restricted to GISAID members**

Analysis tools developed by GISAID members/partners





Important note: In the GISAID EVITY<sup>®</sup> 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 (SIDAID EVITY)<sup>®</sup> more aviruses. All the viewing and using data is explored by the terms of the (SIDAID EVITY)<sup>®</sup> more aviruses. All the viewing are aviruses aviruses.

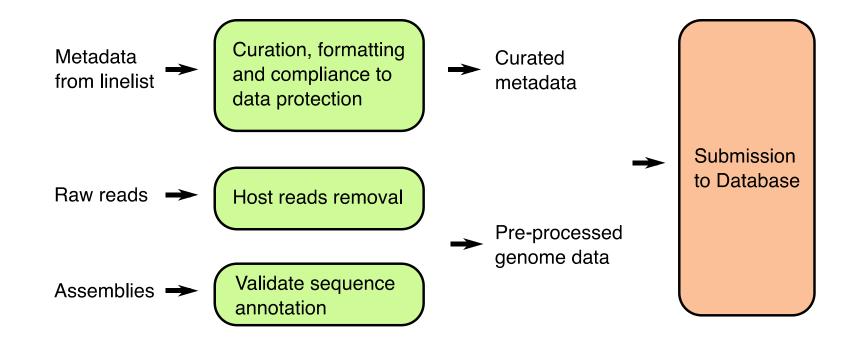
EpiCoV



# 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  |
| an a |   |
| covv_assembly_method                     | ARTIC-ncov2019 v1.1.0<br>5602.89170450142   |
| covv_coverage                            |   |
| covv_orig_lab                            | Department of Bacteria, Parasites and Fungi,<br>Statens Serum Institut, Copenhagen, Denmark |
| covv_orig_lab_addr                       | Artillerivej 5, 2300 Copenhagen S, Denmark  |
| covv_provider_sample_id                  |   |
| cover subm lab                           | Department of Bacteria, Parasites and Fungi,  |
| covv_subm_lab                            | 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                                | _<br>scientific_name                         |        | Severe acute respiratory syndrome coronavirus 2  |
| ENA virus pathogen<br>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<br>(country and/or sea)  |        | Denmark  |
|  | geographic location<br>(region and locality) |        | Sjaelland  |
|  | sample capture status                        |        | Active surveillance in response to outbreak  |
|  | L L  |        |  |
|  | 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<br>Fungi, Statens Serum Institut,<br>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**

Check for

updates

# VADR: validation and annotation of virus sequence submissions to GenBank

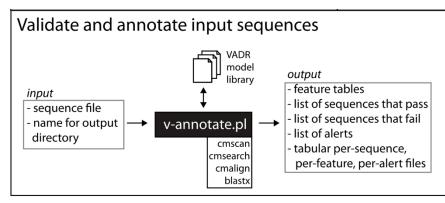
Alejandro A. Schäffer<sup>1,2</sup>, Eneida L. Hatcher<sup>2</sup>, Linda Yankie<sup>2</sup>, Lara Shonkwiler<sup>2,3</sup>, J. Rodney Brister<sup>2</sup>, Ilene Karsch-Mizrachi<sup>2</sup> and Eric P. Nawrocki<sup>2\*</sup>

#### \*Correspondence: nawrocke@ncbi.nlm.nih.gov 2National Center for Biotechnolog

<sup>2</sup>National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD, 20894 USA Tull list of author information is available at the end of the article section 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

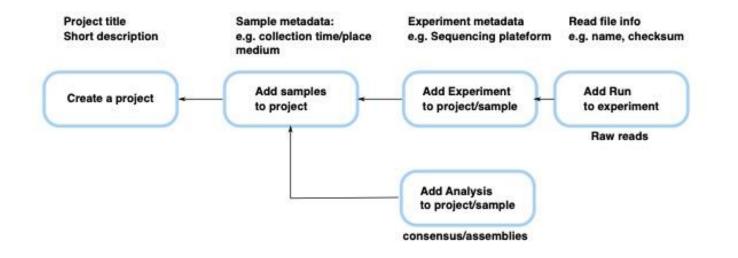
Automated submission screening tool used in GenBank for viral genome sequence submissions

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

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

#### **ENA submission workflow**



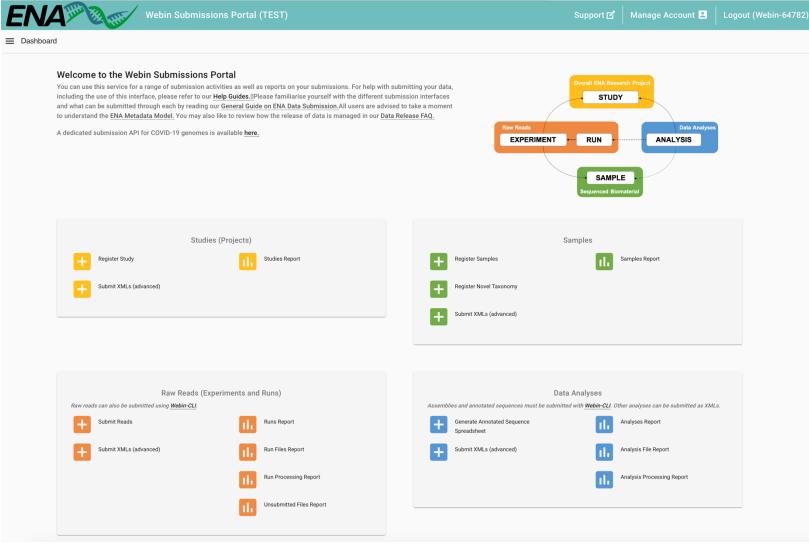


https://ena-docs.readthedocs.io/en/latest/submit/general-guide.html

https://ena-browser-docs.readthedocs.io/en/latest/help\_and\_guides/sars-cov-2-submissions.html

### The ENA web-interactive submission platform (Webin)





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

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

### The NCBI SRA interactive submission portal

| NIH National Library of Medicine<br>National Center for Biotechnology Information              |  |                               |                 |                      |                 | ĝg        | mail.com   |
|--|--|-------------------------------|-----------------|----------------------|-----------------|-----------|------------|
| Submission Portal and Manage Data services will be unavailable during                          | scheduled maintenance on January 28. We apologize for any incon                          | venience. Have question       | ns? Contact our | r service desk at: i | nfo@ncbi.nlm.ni | h.gov     | ×          |
| Submission Portal  |  |                               | Home            | My submissions       | Manage data     | Templates | My profile |
| Sequence Read Archive (SRA) New submiss  | ion  |                               |                 |                      |                 |           |            |
| Short description and brief instructions   | +  | Filter / Search               |                 |                      |                 |           |            |
| Options to upload data:  |  | From date                     | To date         | DD Not dele          | Sort            | by        | desc       |
| Upload via Cloud: Amazon S3 or Google Cloud  | +  | Data archives                 | ] [             | +                    |                 | •         |            |
| Upload via Aspera command line or FTP  | UPDATED JANUARY 2025   | Query @                       |                 |                      |                 |           |            |
| Upload via Web browser or Aspera browser plugin  | UPDATED JANUARY 2025   |                               |                 |                      | Sea             | arch      | Clear      |
|  |  |                               |                 |                      |                 | 1         | 2 Next >   |
| https://account.ncbi.nlm.nih.gov/?t  | <u>ack_url=https%3A//submit.nc</u>   | <u>cbi.nlm.nih</u>            | .gov/si         | ubs/sra/             | /               |           |            |
|  |  |                               |                 |                      |                 |           |            |
| NIH National Library of Medicine   |  | 4                             | gmail.com       |                      |                 |           |            |
| Submission Portal and Manage Data services will be unavailable during scheduled maintenance or | 1 January 28. We apologize for any inconvenience. Have questions? Contact our service de | esk at: info@ncbi.nlm.nih.gov | ×               |                      |                 |           |            |



#### 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

| Submission Portal and Manage Data  | services will be unavailable during schedule | ed maintenance on Jar                | nuary 28. We apologize f  | or any inconvenience. | Have questions? Contai | t our service desk at: | info@ncbi.nlm.nil   | h.gov          | ×           |
|--|--|--------------------------------------|---------------------------|-----------------------|------------------------|------------------------|---|----------------|-------------|
| Submission Portal  |  |                                      |                           |                       | Hom                    | e My submissions       | Manage data   | Templates      | My profil   |
| Sequence Read Archive  | <u>e (SRA)</u> submission: S                 | UB150247                             | 72                        |                       |                        |                        |   | Delete su      | ubmission   |
| 1 SUBMITTER 2 GENERAL INFO 3 SRA M   | METADATA 4 FILES 5 REVIEW & SUBM             | т                                    |                           |                       |                        |                        |   |                |             |
| Submitter  |  |                                      |                           |                       |                        | •                      | Required fields and the second sec | are marked wit | h ★ asteris |
|  | * Lest (family) name<br>Email (secondary)    | <ul> <li>At least one err</li> </ul> | nail should be from the o | rganization's domain. |                        |                        |   |                |             |
| Group for this submission<br>No group (affiliation from my personal<br>3 members Sofie Nielsen's shared submi<br>You can create a group for shared submi | nissions                                     |                                      |                           |                       |                        |                        |   |                |             |
|  | Submitting organization URL                  | * Department                         |                           |                       |                        |                        |   |                |             |
| Phone  Fax   tag  tag  tag  tag  tag  tag  tag  ta   | * City State/Province                        | * Postal code                        | * Country                 |                       |                        |                        |   |                |             |

# 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