DMP and Open Data training
Session 3: Metadata

Retour d'expérience de soumission en banque de données internationnales

helene.chiapello@inrae.fr

&

thomas.denecker@france-bioinformatique.fr

Hélène Chiapello : https://orcid.org/0000-0001-5102-0632
Thomas Denecker : https://orcid.org/0000-0003-1421-7641

Pourquoi soumettre mes données ?

- Open science
- La reproductibilité des expériences
- Donner accès à mes données
- Archiver mes données
- Publication d’articles
- Analyser mes données

3 bases de données

ENA
European Nucleotide Archive

GISAID

GEO
Gene Expression Omnibus


ENA
European Nucleotide Archive

Qui a déjà soumis à l’ENA ?

C’était facile ?

La base de données

Plateforme ouverte pour la gestion, le partage, l'intégration, l'archivage et la diffusion des données de séquençage.

Connecté avec UniProt, RNAcentral, EBI Metagenomics, Ensembl, Ensembl Genomes, ArrayExpress, ...

Des données variées: génomique animale, la biotechnologie marine, la biodiversité, la surveillance des agents pathogènes et la biologie des cellules souches

La documentation

ENA Training Modules

latest

Search docs

ENA DATA SUBMISSION

- General Guide On ENA Data Submission
- How to Register a Study
- How to Register Samples
- Preparing Files for Submission
- How to Submit Raw Reads
- How to Submit Assemblies
- How to Submit Targeted Sequences
- How to Submit Other Analyses

ENA DATA DISCOVERY & RETRIEVAL

- General Guide on ENA Data Retrieval
- How to Explore an ENA Project
- How to Download Data Files
- How To Perform An Advanced Search
- How to Access ENA Programmatically

ENA DATA UPDATES

- Updating Metadata Objects
- Updating Assemblies
- Updating Annotated Sequences

TIPS AND FAQS

- Data Release Policies
- Common Run Submission Errors
- Tips for Sample Taxonomy
- Requesting New Taxon IDs
- Metagenome Submission Queries
- Locus Tag Prefixes
- Archive Generated FASTQ Files
- Third Party Tools

Docs » ENA: Guidelines and Tutorials

Edit on GitHub

ENA: Guidelines and Tutorials

Welcome to the guidelines for submission and retrieval for the European Nucleotide Archive. Please use the links to find instructions specific to your needs. If you're completely new to ENA, you can see an introductory webinar at the bottom of the page.

ENA Data Submission

- General Guide On ENA Data Submission
- How to Register a Study
- How to Register Samples
- Preparing Files for Submission
- How to Submit Raw Reads
- How to Submit Assemblies
- How to Submit Targeted Sequences
- How to Submit Other Analyses

ENA Data Discovery & Retrieval

- General Guide on ENA Data Retrieval
- How to Explore an ENA Project
- How to Download Data Files
- How To Perform An Advanced Search
- How to Access ENA Programmatically

ENA Data Updates

- Updating Metadata Objects
- Updating Assemblies
- Updating Annotated Sequences

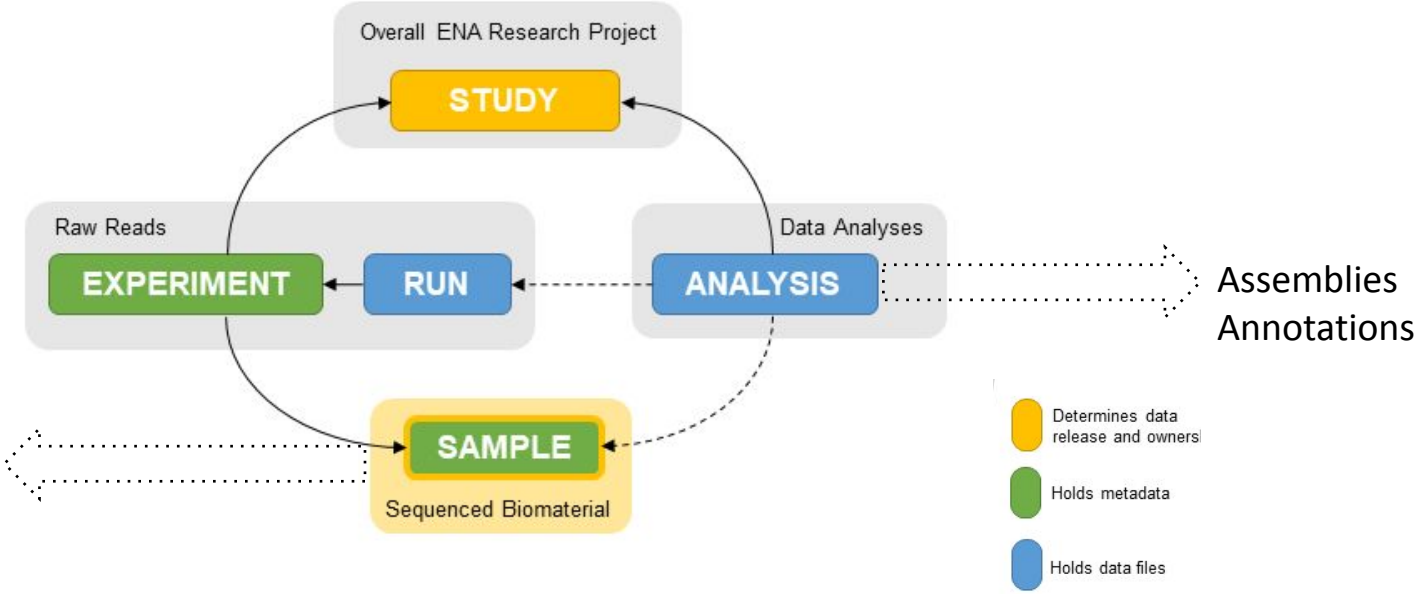
Tips and FAQs

- Data Release Policies
- Common Run Submission Errors
- Tips for Sample Taxonomy
- Requesting New Taxon IDs
- Metagenome Submission Queries
- Locus Tag Prefixes
- Archive Generated FASTQ Files
- Third Party Tools

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

Modèle des métadonnées

ISA compliant !



All **samples** submitted to ENA must conform to a **Checklist**

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

Description des expériences et validation

Metadata validation

Permitted values for platform

- LS454: 454 technology use 1-color sequential flows
- ILLUMINA: Illumina is 4-channel flowgram with 1-to-1 mapping between basecalls and flows
- PACBIO_SMRT: PacificBiosciences platform type for the single molecule real time (SMRT) technology.
- ION_TORRENT: Ion Torrent Personal Genome Machine (PGM) from Life Technologies.
- CAPILLARY: Sequencers based on capillary electrophoresis technology manufactured by LifeTech (formerly Applied BioSciences).
- OXFORD_NANOPORE: Oxford Nanopore platform type. nanopore-based electronic single molecule analysis.
- BGISEQ
- DNBSEQ

https://ena-docs.readthedocs.io/en/latest/submit/reads/webin-cli.html?permitted-values-for-instrument

Les checklists de l'ENA pour les "samples"

- A **checklist** defines the **minimum and optional metadata** expected to describe biological samples
- ENA are based on the **Genomic Standards Consortium (GSC)** recommandations
- The **most suitable checklist** depends on the type of the sample: https://www.ebi.ac.uk/ena/browser/checklists
- All ENA checklist are defined by an **access number** like ERCxxx (Ena R Checklist xxx)
  - example: GSC MIxS plant associated https://www.ebi.ac.uk/ena/browser/view/ERC000020

Listes des checklists pour les “Sample”

EMBL-EBI | Services | Research | Training | About us | EMBL-EBI

ENA European Nucleotide Archive

Enter text search terms — Search

Examples: histone, BN000065

Enter accession — View

Examples: Taxon:9606, BN000065, PRJEB402

Home | Submit | Search | Rulespace | About | Support

Sample Checklists

There is a minimum amount of information required during ENA sample registration and all samples must conform to a defined checklist of expected metadata values. The most suitable checklist for sample registration depends on the type of the sample.

These sample checklists have been developed to meet the needs of different research communities. Different communities have different requirements on the minimum metadata expected to describe biological samples.

Filter checklists...

Accession	Name	Description
ERC000012	GSC MIxS air	Genomic Standards Consortium package extension for reporting of measurements and observations obtain...
ERC000013	GSC MIxS host associated	Genomic Standards Consortium package extension for reporting of measurements and observations obtain...
ERC000014	GSC MIxS human associated	Genomic Standards Consortium package extension for reporting of measurements and observations obtain...
ERC000015	GSC MIxS human gut	Genomic Standards Consortium package extension for reporting of measurements and observations obtain...
ERC000016	GSC MIxS human oral	Genomic Standards Consortium package extension for reporting of measurements and observations obtain...
ERC000017	GSC MIxS human skin	Genomic Standards Consortium package extension for reporting of measurements and observations obtain...
ERC000018	GSC MIxS human vaginal	Genomic Standards Consortium package extension for reporting of measurements and observations obtain...

Exemple COVID-19

ENA European Nucleotide Archive

Enter text search terms | Search

Examples: histone, BN000065

ERC000033 | View

Examples: Taxon:9606, BN000065, PRJEB402

Home | Submit | Search | Rulespace | About | Support

Checklist: ERC000033

View: XML

Download: XML

ENA virus pathogen reporting standard checklist

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

Checklist Fields

Filter fields...

Filter by type:

- Human surveillance data
- Collection event information
- sample collection
- host disorder
- host description
- Virus isolate information
- General collection event information
- Serology detection
- Infraspecies information
- Associated host information
- host details
- Environmental information

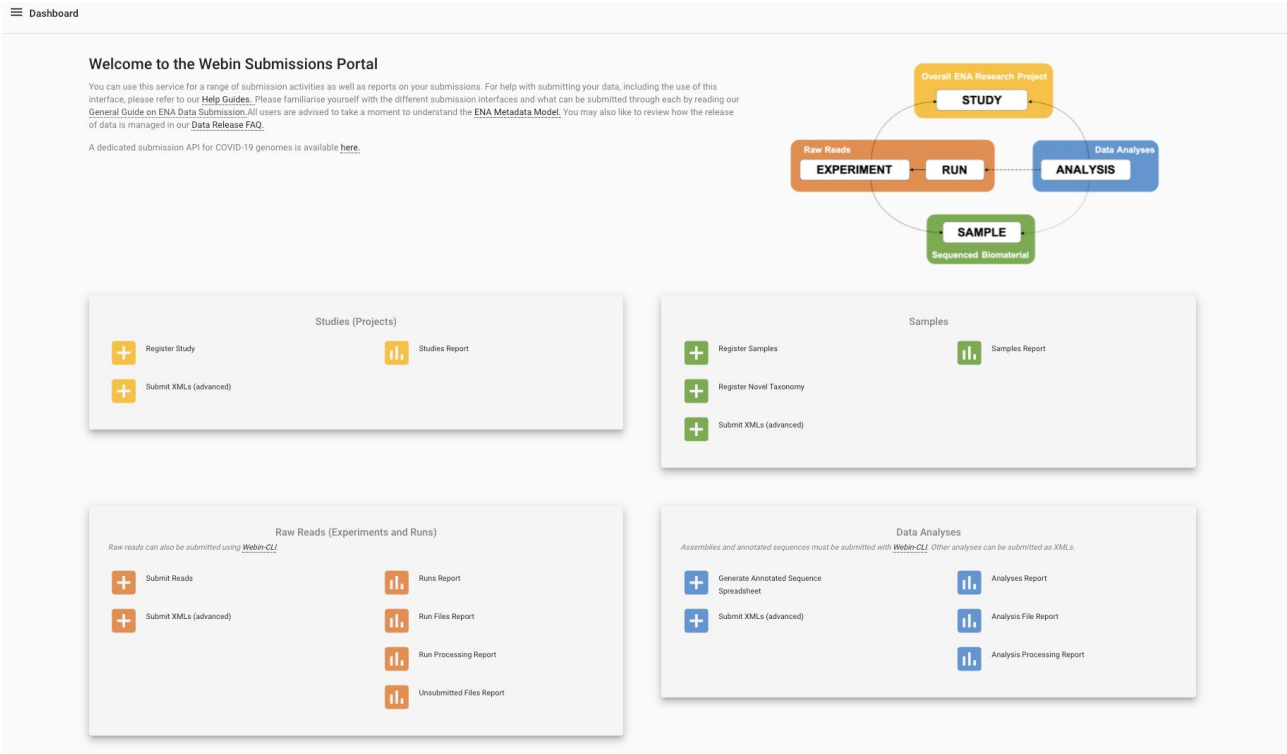
Field Name	Field Format	(Field Restriction)	Requirement	(Units)
subject exposure	free text		optional	
subject exposure duration	free text		optional	
type exposure	free text		optional	
personal protective equipment	free text		optional	
hospitalisation	text choice	options	optional	
illness duration	free text		optional	
illness symptoms	free text		optional	
collection date	restricted text	regular expression	recommended	
geographic location (country and/or sea)	text choice	options	mandatory	
geographic location (latitude)	restricted text	regular expression	recommended	DD
geographic location (longitude)	restricted text	regular expression	recommended	DD
geographic location (region and locality)	free text		recommended	

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

Méthodes de soumission

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

Interactive



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

Web-Cli

v4.2.1 Latest

Compare

Rajkumar-D released this 26 days ago v4.2.1 0d34c7a

- sequence context: Added support for BioSample accessions, SRA Sample accessions and SRA Sample aliases in the ORGANISM field in addition to the already supported NCBI taxonomy names and IDs.

Assets 4

webin-cli-4.2.1-sources.jar	109 KB
webin-cli-4.2.1.jar	61.5 MB
Source code (zip)	
Source code (tar.gz)	

Programmatic

- SUBMISSION (XML Schema)
- STUDY (XML Schema)
- SAMPLE (XML Schema)
- EXPERIMENT (XML Schema)
- RUN (XML Schema)
- ANALYSIS (XML Schema)
- DAC (XML Schema)
- POLICY (XML Schema)
- DATASET (XML Schema)
- PROJECT (XML Schema)

Exemple : submission.xml

```
<SUBMISSION>
     <ACTIONS>
         <ACTION>
             <ADD/>
         </ACTION>
    </ACTIONS>
</SUBMISSION>
```

Cas particulier COVID-19

COVID-19 Data Portal

About ▾ News Partners Related resources FAQ Bulk downloads Submit data

Viral Sequences Host Sequences Expression Proteins Networks Samples Imaging Literature

Submit new data

Information on how to submit COVID-19 data

We have a new drag-and-drop data submission tool, suitable for viral sequence submissions. We are inviting volunteers to try it out - please register your interest below.

Data types

- Viral, non-human and cell line sequence data
- Human molecular biology data
- Linked viral and human molecular biology data
- Viral and non-human proteomics data
- Structural biology data
- Viral and non-human molecular interaction data
- Viral and non-human metabolomics data
- Viral and other non-human molecular biology data
- Compound and target data
- Clinical and epidemiological data
- Non-biological data

ENA

Viral, non-human and cell line sequence data

This class includes sequence data from studies targeting virus alone or with co-occurring species. It also includes sequencing from non-human host species (such as from species acting as models for infection) and human cell lines (where data are consented for full open publication). All sequencing library types, all platforms, all library methods and all levels of processing (from raw data to assembled sequences) are included in this class.

Deposition actions:

Users should submit data to ENA
Specific deposition instructions are available for viral data submission
Users are encouraged to contact ENA at virus-dataflow@ebi.ac.uk

General depositions and those from users who are managing their data in SARS-CoV-2 Data Hubs are also included in this class.

Drag and Drop viral sequence submission tool

We have a new drag-and-drop data submission tool, which is suitable for many viral sequence submissions. Please register your interest and we will be in contact to assess the suitability of the tool for your data set.

Register

ENA European Nucleotide Archive

Enter text search terms — Search

Examples: histone, BN000065

ERC000033 — View

Examples: Taxon:9606, BN000065, PRJEB402

Home | Submit | Search | Rulespace | About | Support

Checklist: ERC000033

View: XML
Download: XML

ENA virus pathogen reporting standard checklist

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

Checklist Fields

Filter fields...

Filter by type:

- Human surveillance data
- Collection event information
- sample collection
- host disorder
- host description
- Virus isolate information
- General collection event information
- Serology detection
- Infraspecies information
- Associated host information
- host details
- Environmental information

Field Name	Field Format	(Field Restriction)	Requirement	(Units)
subject exposure	free text		optional	
subject exposure duration	free text		optional	
type exposure	free text		optional	
personal protective equipment	free text		optional	
hospitalisation	text choice	options	optional	
illness duration	free text		optional	
illness symptoms	free text		optional	
collection date	restricted text	regular expression	recommended	
geographic location (country and/or sea)	text choice	options	mandatory	
geographic location (latitude)	restricted text	regular expression	recommended	DD
geographic location (longitude)	restricted text	regular expression	recommended	DD
geographic location (region and locality)	free text		recommended	

https://www.ebi.ac.uk/ena/browser/view/ERC000033

Les outils complémentaires

Tools & Data Resources

Search all tools & data resources

Tools

Clustal Omega
Multiple sequence alignment of DNA or protein sequences. Clustal Omega replaces the older ClustalW alignment tools.
Web API | Multiple sequence alignment

InterProScan
InterProScan searches sequences against InterPro's predictive protein signatures.
Web API | Protein feature detection | Sequence motif recognition

BLAST [protein]
Fast local similarity search tool for protein sequence databases.
Web API | Sequence similarity search

BLAST [nucleotide]
Fast local similarity search tool for nucleotide sequence databases.
Web API | Sequence similarity search

HMMER
Fast sensitive protein homology searches using profile hidden Markov models (HMMs) for querying against both sequence and HMM target databases.
Web API | Sequence similarity search | Protein function prediction

See all tools

Data resources

Ensembl
Genome browser, API and database, providing access to reference genome annotation
Web API

UniProt
A comprehensive resource for protein sequence and functional annotation.
Web API

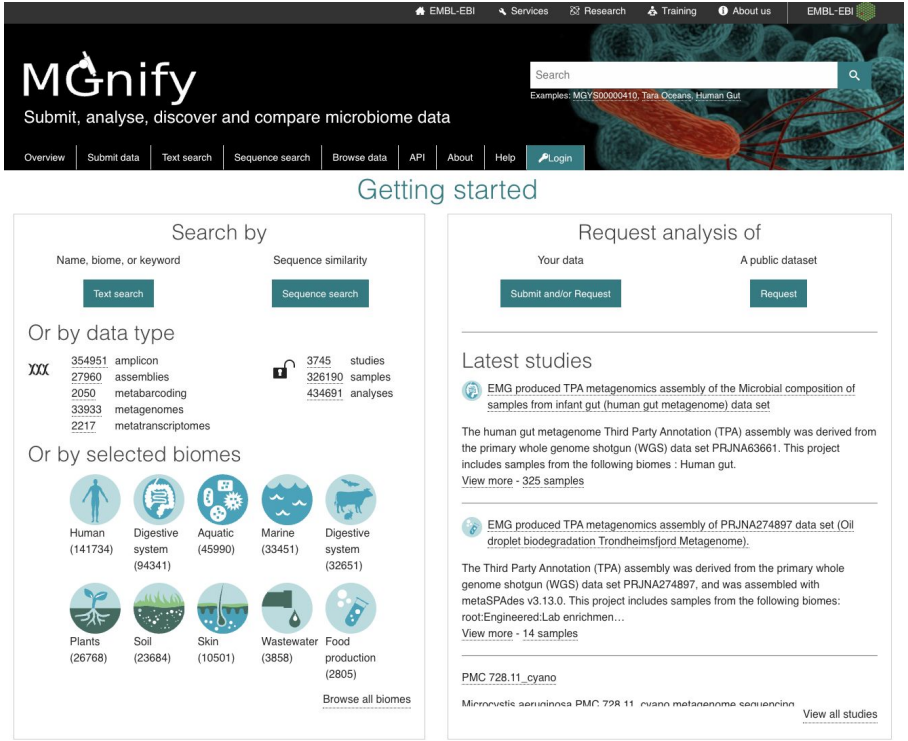
PDBe
The European resource for the collection, organisation and dissemination of 3D structural data (from PDB and EMDB) on biological macromolecules and their complexes.
Web API

Europe PMC
A database to search the worldwide life sciences literature
Web API

Expression Atlas
An added-value database that shows which genes/proteins are expressed under which conditions, and how expression differs between conditions.

ChEMBL
An open data resource of binding, functional and ADMET bioactivity data.
Web API

See all data resources



https://www.ebi.ac.uk/services/all

GEO

Gene Expression Omnibus

Qui a déjà soumis à GEO ?

C’était facile ?

La base de données

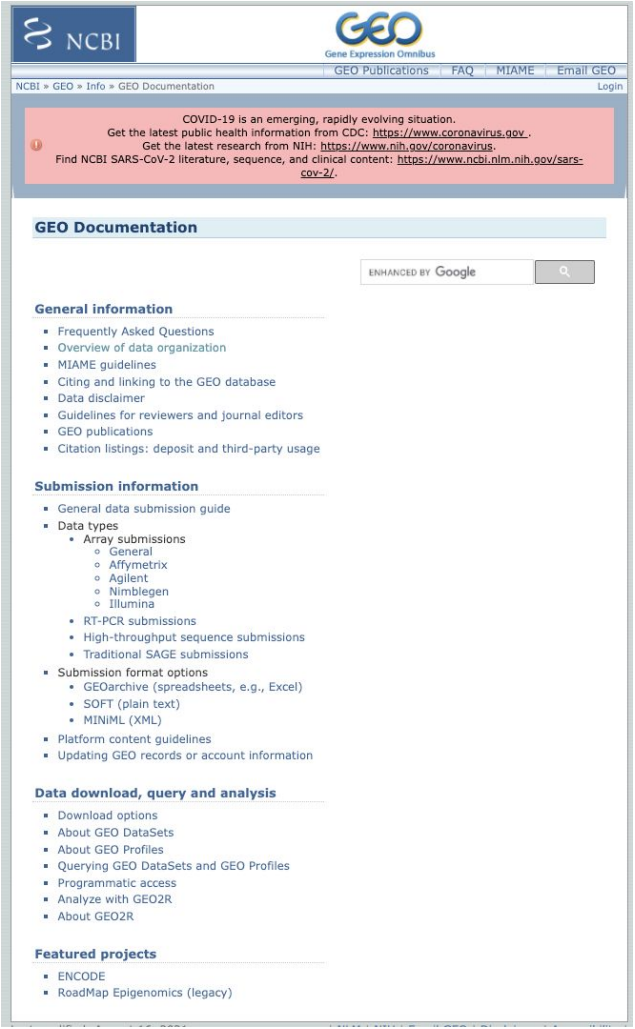
GEO est un dépôt public international qui archive et distribue librement des données de:

- microarray ;
- de NGS ;
- et d'autres formes de données de génomique fonctionnelle à haut débit .

soumises par la communauté des chercheurs.

Documentation

https://www.ncbi.nlm.nih.gov/geo/info/



NCBI

GEO Gene Expression Omnibus

GEO Publications | FAQ | MIAME | Email GEO

NCBI » GEO » Info » GEO Documentation

Login

COVID-19 is an emerging, rapidly evolving situation.
Get the latest public health information from CDC: https://www.coronavirus.gov .
Get the latest research from NIH: https://www.nih.gov/coronavirus.
Find NCBI SARS-CoV-2 literature, sequence, and clinical content: https://www.ncbi.nlm.nih.gov/sars-cov-2/.

GEO Documentation

ENHANCED BY Google

General information

- Frequently Asked Questions
- Overview of data organization
- MIAME guidelines
- Citing and linking to the GEO database
- Data disclaimer
- Guidelines for reviewers and journal editors
- GEO publications
- Citation listings: deposit and third-party usage

Submission information

- General data submission guide
- Data types
  - Array submissions
    - General
    - Affymetrix
    - Agilent
    - Nimblegen
    - Illumina
  - RT-PCR submissions
  - High-throughput sequence submissions
  - Traditional SAGE submissions
- Submission format options
  - GEOarchive (spreadsheets, e.g., Excel)
  - SOFT (plain text)
  - MINiML (XML)
- Platform content guidelines
- Updating GEO records or account information

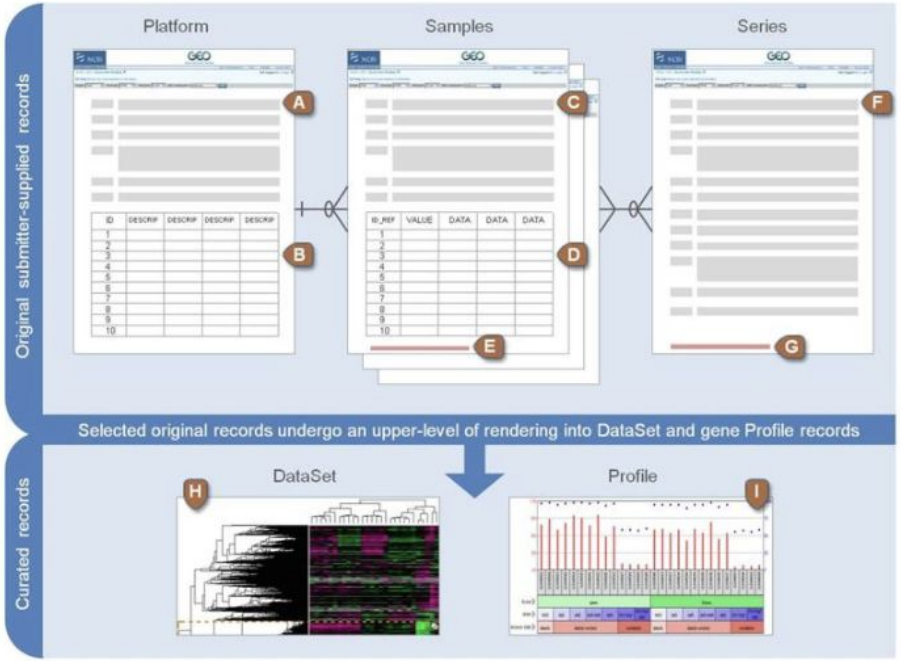
Data download, query and analysis

- Download options
- About GEO DataSets
- About GEO Profiles
- Querying GEO DataSets and GEO Profiles
- Programmatic access
- Analyze with GEO2R
- About GEO2R

Featured projects

- ENCODE
- RoadMap Epigenomics (legacy)

Organisation des données



Platform	Platform records are supplied by submitters A Platform record is composed of a summary description of the array or sequencer and, for array-based Platforms, a data table defining the array template.Each Platform record is assigned a unique and stable GEO accession number (GPLxxx). A Platform may reference many Samples that have been submitted by multiple submitters. Example Platform record »	A	**Text description of the array or sequencer**
		B	**Text tab-delimited table of the array template**
Sample	Sample records are supplied by submitters A Sample record describes the conditions under which an individual Sample was handled, the manipulations it underwent, and the abundance measurement of each element derived from it. Each Sample record is assigned a unique and stable GEO accession number (GSMxxx). A Sample entity must reference only one Platform and may be included in multiple Series. Example Sample record »	C	**Text description of the biological sample and protocols to which it was subjected**
		D	**Text tab-delimited table of processed hybridization result** (may optionally include raw data columns)
		E	**Original raw data file, or processed sequence data file**
Series	Series records are supplied by submitters A Series record links together a group of related Samples and provides a focal point and description of the whole study. Series records may also contain tables describing extracted data, summary conclusions, or analyses. Each Series record is assigned a unique and stable GEO accession number (GSExxx). Example Series record »	F	**Text description of the overall experiment**
		G	**Tar archive of original raw data files, or processed sequence data files**

https://www.ncbi.nlm.nih.gov/geo/info/overview.html

Fichiers

GEOarchive format

GEOarchive is a flexible spreadsheet-based submission format useful for batch deposit of experiments. GEOarchive submissions can be created in any spreadsheet software, usually Microsoft Excel.

A GEOarchive submission consists of several parts as follows:

Metadata spreadsheet	'Metadata' refers to descriptive information and protocols for the overall experiment and individual Samples. This information is supplied by completing all fields of the appropriate metadata spreadsheet template which can be downloaded from the GEOarchive templates and examples section below.
Matrix table	The matrix table is a spreadsheet containing the final, normalized values that are comparable across rows and Samples, and preferably processed as described in any accompanying manuscript. A complete data matrix should be supplied, not a summary subset. It is possible to include additional data columns in the table, for example, Affymetrix Detection calls and P-values, or background or flag columns. See the Affymetrix template for an example.
Raw data files	In addition to the normalized data provided in the Matrix table, submitters are required to provide raw data, usually in the form of supplementary raw data files. This facilitates the unambiguous interpretation of the data and potential verification of the conclusions as described in the MIAME and MINSEQE standards. Affymetrix submissions must include CEL files. Non-Affymetrix GEOarchive submissions should include the original software-generated scan quantification files, for example, GenePix GPR files. Next-generation sequence submissions must include files containing reads and quality scores.
Platform	If your experiments are performed using a commercial array (e.g., Affymetrix GeneChip) or other array already deposited in GEO, please use the FIND PLATFORM tool to find the GEO accession number (GPLxxxx) for inclusion in the '*platform*' column in the *SAMPLES* section of the metadata spreadsheet. If your array does not already exist in GEO, please include a *PLATFORM* section in your metadata spreadsheet and include Platform annotation columns in your matrix table. The Platform data must include meaningful, trackable, sequence identifiers (e.g. GenBank/RefSeq accessions, locus tags, clone IDs, oligo sequences, chromosome locations, etc - see the Platform content guidelines for full list). References to in-house databases or top BLAST hits are not sufficient. Platform submission is not necessary for SAGE or next-generation sequence submissions.

Bundle all parts (Excel file containing the metadata spreadsheet and matrix spreadsheet, raw data files) together into a .zip, .rar, or .tar archive using a program like WinZip, and transfer to GEO using the 'Transfer files to GEO with web form' option on the Submit to GEO page. Incomplete submissions will result in processing delays.

Submit

GEOarchive templates and examples

The first step in creating your GEOarchive submission is to download the appropriate template (Excel spreadsheet) from the list below. Each Excel file consists of several worksheets, including a metadata template, and examples of metadata and matrix tables. Click the tabs at the bottom of the worksheet window to switch between worksheets. Mouse over field names in the templates to view content guidelines.

Microarray

For the following microarray vendors, please download templates from the vendor-specific instructions pages:

- Affymetrix submissions
- Agilent submissions
- Nimblegen submissions
- Illumina submissions

For microarrays not from the vendors above, please use a 'Generic' template. For generic microarray submissions where the Platform is already deposited in GEO, please download the most appropriate template:

- Generic single channel submission template
- Generic dual channel submission template
- Generic merged dye-swap submission template
- Generic tiling ChIP-chip submission template

For generic microarray submissions where the Platform is not deposited in GEO, please download the most appropriate template:

- Generic single channel submission template, including Platform
- Generic dual channel submission template, including Platform
- Generic merged dye-swap submission template, including Platform
- Generic tiling ChIP-chip submission template, including Platform

To submit only a Platform, please download the following template (this option is appropriate only if you have no hybridization or sequence data to deposit):

- Platform-only template

High-throughput sequencing

For high-throughput sequence submissions, please refer to full instructions at:

- High-throughput sequence submissions

Other data types

For NanoString submissions, please use one of the 'Generic single channel' templates as appropriate:

- Generic single channel submission template
- Generic single channel submission template, including Platform

For high-throughput RT-PCR submissions, please refer to full instructions at:

- RT-PCR submissions

For traditional SAGE submissions, please refer to full instructions at:

- Traditional SAGE submissions

https://www.ncbi.nlm.nih.gov/geo/info/spreadsheet.html#GAtemplates

Exemple Excel Illumina

GA_illumina_expression.xls [Mode de compatibilité]

SERIES											
title	Genome-wide analysis of mechano-responsive gene expression by tenocytes in fascicles subjected to cyclic tensile strain										
summary	Analysis of mechano-regulation of tenocyte metabolism at gene expression level. The hypothesis tested in the present study was that cyclic tensile strain influence the balance of anabolism/catabolism of tenocytes. Results provide important information of the response of tenocyte										
overall design	Total RNA obtained from isolated tendon fascicles subjected to 1 or 24 hours in vitro cyclic tensile strain compared to unstrained control fascicles.										
contributor	Jane,Doe										
contributor	John,A,Smith										
SAMPLES											
# The corresponding example matrix table is included in the next worksheet.											
Sample name	title	source name	organism	idat file	characteristics: Strain	characteristics: age	characteristics: tiss	molecule	label	description	platform
Sample 1	Fascicle Strained 24h rep1	Rat tail tendon	Rattus norvegicus	4307579061_B_Grn.idat	Wistar	5 months	tail tendon	total RNA	biotin	replicate 1	GPL6101
Sample 2	Fascicle Unstrained 24h rep1	Rat tail tendon	Rattus norvegicus	4307579072_A_Grn.idat	Wistar	5 months	tail tendon	total RNA	biotin	replicate 1	GPL6101
Sample 3	Fascicle Strained 1h rep2	Rat tail tendon	Rattus norvegicus	4307579062_B_Grn.idat	Wistar	5 months	tail tendon	total RNA	biotin	replicate 2	GPL6101
PROTOCOLS											
extract protocol	RNA was extracted with Trizol reagent, followed by clean-up and DNase I treatment with QIAGEN RNeasy mini kit in accordance with the prescribed protocol provided with the kit. Quality control was performed with Agilent Bioanalyser.										
label protocol	Biotinylated cRNA were prepared with the Ambion MessageAmp kit for Illumina arrays										
hyb protocol	Standard Illumina hybridization protocol										
scan protocol	Standard Illumina scanning protocol										
data processing	The data were normalised using quantile normalisation with IlluminaGUI in R										
value definition	quantile normalized										

Metadata Template | Matrix normalized | Matrix non-normalized | Metadata Example | Matrix normalized Example | Matrix non-normalized Example

Les outils complémentaires : GeoToR

exemple : GSE25724

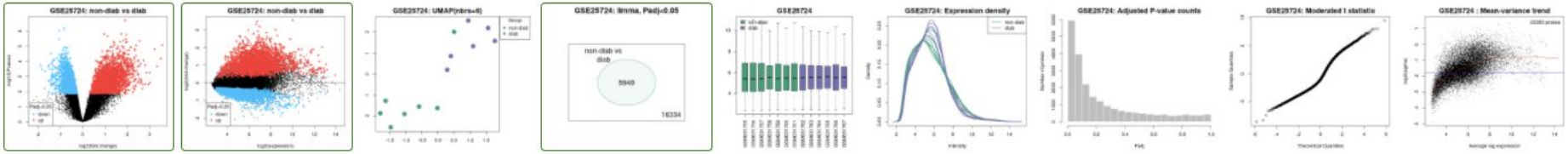
GEO accession GSE25724 Set Expression data from type 2 diabetic and non-diabetic isolated human islets

Samples Define groups Selected **13** out of **13** samples

Columns Set

Group	Accession	Title	Source name	Tissue	Disease state	Age	Gender	Characteristics
non-diab	GSM631755	Non-diabetic islets, rep1	human islets, non-diabetic	pancreatic islets	non-diabetic	47 yrs	male	bmi (kg/m2): 27.7
non-diab	GSM631756	Non-diabetic islets, rep2	human islets, non-diabetic	pancreatic islets	non-diabetic	33 yrs	male	bmi (kg/m2): 22.9
non-diab	GSM631757	Non-diabetic islets, rep3	human islets, non-diabetic	pancreatic islets	non-diabetic	47 yrs	male	bmi (kg/m2): 28.4
non-diab	GSM631758	Non-diabetic islets, rep4	human islets, non-diabetic	pancreatic islets	non-diabetic	54 yrs	male	bmi (kg/m2): 23.1
non-diab	GSM631759	Non-diabetic islets, rep5	human islets, non-diabetic	pancreatic islets	non-diabetic	76 yrs	female	bmi (kg/m2): 25.9
non-diab	GSM631760	Non-diabetic islets, rep6	human islets, non-diabetic	pancreatic islets	non-diabetic	77 yrs	female	bmi (kg/m2): 23.8
non-diab	GSM631761	Non-diabetic islets, rep7	human islets, non-diabetic	pancreatic islets	non-diabetic	73 yrs	female	bmi (kg/m2): 22
diab	GSM631762	Type 2 diabetic islets, rep1	human islets, diabetic	pancreatic islets	type 2 diabetes	79 yrs	male	bmi (kg/m2): 27.5
diab	GSM631763	Type 2 diabetic islets, rep2	human islets, diabetic	pancreatic islets	type 2 diabetes	76 yrs	male	bmi (kg/m2): 26
diab	GSM631764	Type 2 diabetic islets, rep3	human islets, diabetic	pancreatic islets	type 2 diabetes	73 yrs	female	bmi (kg/m2): 29
diab	GSM631765	Type 2 diabetic islets, rep4	human islets, diabetic	pancreatic islets	type 2 diabetes	75 yrs	female	bmi (kg/m2): 26.5
diab	GSM631766	Type 2 diabetic islets, rep5	human islets, diabetic	pancreatic islets	type 2 diabetes	54 yrs	female	bmi (kg/m2): 23.9
diab	GSM631767	Type 2 diabetic islets, rep6	human islets, diabetic	pancreatic islets	type 2 diabetes	66 yrs	male	bmi (kg/m2): 23.1

Visualization



https://www.ncbi.nlm.nih.gov/geo/info/geo2r.html

GISAID

Qui a déjà soumis à GISAID ?

C’était facile ?

Présentation de la base

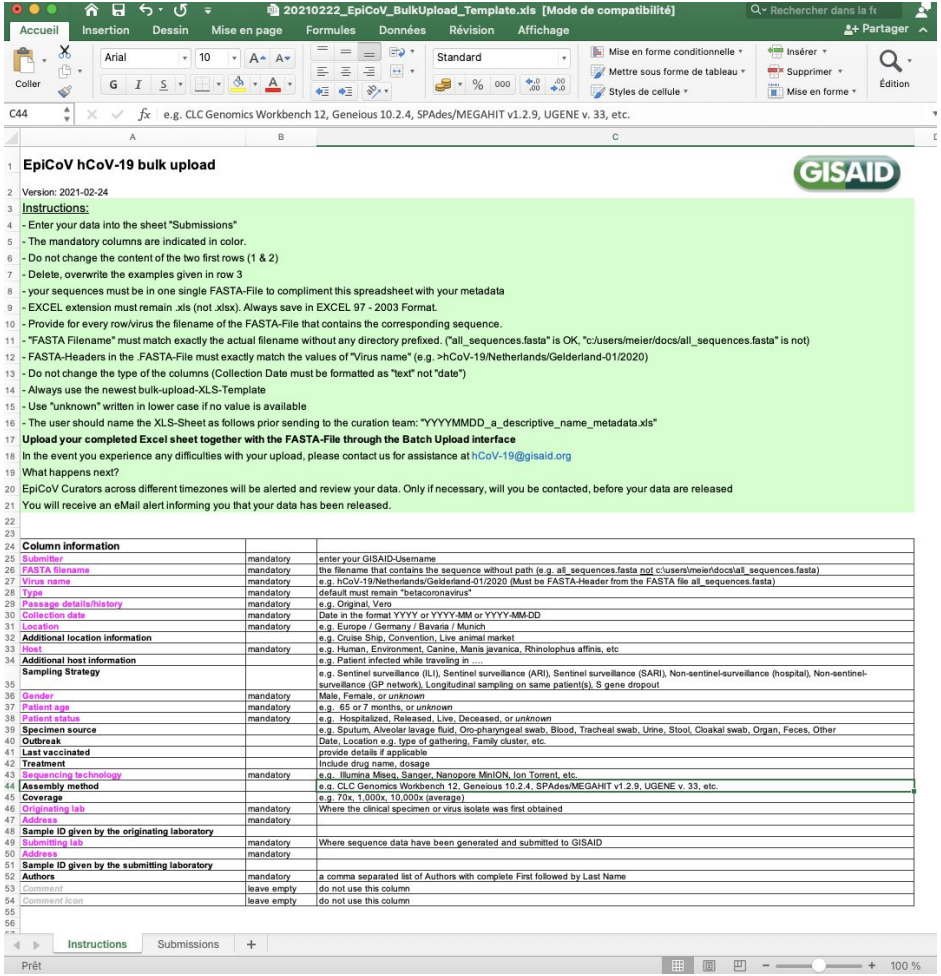
Données de tous les virus de la grippe et du **coronavirus à l'origine du COVID-19** : séquence génétique et les données cliniques et épidémiologiques associées aux virus humains, ainsi que les données géographiques et spécifiques aux espèces associées aux virus aviaires et autres virus animaux, pour aider les chercheurs à comprendre comment les virus évoluent et se propagent pendant les épidémies et les pandémies.

GISAID le fait en surmontant les obstacles et les restrictions dissuasifs, qui découragent ou empêchent le partage des données virologiques avant la publication officielle.

L'Initiative garantit que le libre accès aux données de GISAID est fourni gratuitement à toutes les personnes qui ont accepté de **s'identifier et de respecter le mécanisme de partage de GISAID régi par son accord d'accès à la base de données.**

Le fichier de métadonnées

Fichier excel



20210222_EpiCoV_BulkUpload_Template.xls [Mode de compatibilité]

C44: e.g. CLC Genomics Workbench 12, Geneious 10.2.4, SPAdes/MEGAHIT v1.2.9, UGENE v. 33, etc.

EpiCoV hCoV-19 bulk upload

Version: 2021-02-24

Instructions:

- Enter your data into the sheet "Submissions"
- The mandatory columns are indicated in color.
- Do not change the content of the two first rows (1 & 2)
- Delete, overwrite the examples given in row 3
- your sequences must be in one single FASTA-File to compliment this spreadsheet with your metadata
- EXCEL extension must remain .xls (not .xlsx). Always save in EXCEL 97 - 2003 Format.
- Provide for every row/virus the filename of the FASTA-File that contains the corresponding sequence.
- "FASTA Filename" must match exactly the actual filename without any directory prefixed. ("all_sequences.fasta" is OK, "c:/users/meier/docs/all_sequences.fasta" is not)
- FASTA-Headers in the .FASTA-File must exactly match the values of "Virus name" (e.g. >hCoV-19/Netherlands/Gelderland-01/2020)
- Do not change the type of the columns (Collection Date must be formatted as "text" not "date")
- Always use the newest bulk-upload-XLS-Template
- Use "unknown" written in lower case if no value is available
- The user should name the XLS-Sheet as follows prior sending to the curation team: "YYYYMMDD_a_descriptive_name_metadata.xls"

Upload your completed Excel sheet together with the FASTA-File through the Batch Upload interface

In the event you experience any difficulties with your upload, please contact us for assistance at hCoV-19@gisaid.org

What happens next?

EpiCoV Curators across different timezones will be alerted and review your data. Only if necessary, will you be contacted, before your data are released

You will receive an eMail alert informing you that your data has been released.

Column information		
Submitter	mandatory	enter your GISAID-Username
FASTA filename	mandatory	the filename that contains the sequence without path (e.g. all_sequences.fasta not c:\users\meier\docs\all_sequences.fasta)
Virus name	mandatory	e.g. hCoV-19/Netherlands/Gelderland-01/2020 (Must be FASTA-Header from the FASTA file all_sequences.fasta)
Type	mandatory	default must remain "betacoronavirus"
Passage details/history	mandatory	e.g. Original, Vero
Collection date	mandatory	Date in the format YYYY or YYYY-MM or YYYY-MM-DD
Location	mandatory	e.g. Europe / Germany / Bavaria / Munich
Additional location information		e.g. Cruise Ship, Convention, Live animal market
Host	mandatory	e.g. Human, Environment, Canine, Manis javanica, Rhinolophus affinis, etc
Additional host information		e.g. Patient infected while traveling in ....
Sampling Strategy		e.g. Sentinel surveillance (ILI), Sentinel surveillance (ARI), Sentinel surveillance (SARI), Non-sentinel-surveillance (hospital), Non-sentinel-surveillance (GP network), Longitudinal sampling on same patient(s), S gene dropout
Gender	mandatory	Male, Female, or *unknown*
Patient age	mandatory	e.g.. 65 or 7 months, or *unknown*
Patient status	mandatory	e.g.. Hospitalized, Released, Live, Deceased, or *unknown*
Specimen source		e.g. Sputum, Alveolar lavage fluid, Oro-pharyngeal swab, Blood, Tracheal swab, Urine, Stool, Cloakal swab, Organ, Feces, Other
Outbreak		Date, Location e.g. type of gathering, Family cluster, etc.
Last vaccinated		provide details if applicable
Treatment		Include drug name, dosage
Sequencing technology	mandatory	e.g.. Illumina Miseq, Sanger, Nanopore MinION, Ion Torrent, etc.
Assembly method		e.g. CLC Genomics Workbench 12, Geneious 10.2.4, SPAdes/MEGAHIT v1.2.9, UGENE v. 33, etc.
Coverage		e.g. 70x, 1,000x, 10,000x (average)
Originating lab	mandatory	Where the clinical specimen or virus isolate was first obtained
Address	mandatory	
Sample ID given by the originating laboratory		
Submitting lab	mandatory	Where sequence data have been generated and submitted to GISAID
Address	mandatory	
Sample ID given by the submitting laboratory		
Authors	mandatory	a comma separated list of Authors with complete First followed by Last Name
Comment	leave empty	do not use this column
Comment Icon	leave empty	do not use this column

Instructions | Submissions

Prêt 100 %

WEB - Single

GISAID

© 2008 - 2021 | Terms of Use | Privacy Notice | Contact

You are logged in as **Thomas Denecker** - logout

Registered Users | EpiFlu™ | EpiCoV™ | My profile

EpiCoV™ | Search | Downloads | Upload

Single Upload

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

Virus detail

Virus name*
hCoV-19/Country/Identifier/2021

Accession ID

Type: betacoronavirus

Passage details/history*
Example: Original, Vero

Sample information

Collection date*
Example: 2021-03-27, 2021-03 (collection in March, specific day unknown), 2021 (collection in 2021, month and day unknown)

Location*
Continent / Country or Territory / Region

Additional location information: *Travel history; Residence; Cruise ship; ...*

Host*: *Human, Environment, Canis lupus*

Additional host information
Example: Underlying health conditions; other host relevant characteristics

Outbreak Detail
Example: Date, Place, Family cluster

Sampling strategy: *Baseline surveillance; Active surveillance; Clinical trial; ...*

Gender*: *Male, Female, or unknown*

Patient age*
Example: 65, 7 months, or unknown

Patient status*: *Hospitalized, Released, Live, Deco*

Specimen source: *Sewage, Sputum, Alveolar lavage fluid, Oropharyngeal swab, Mid-Turbinate swab, Nasopharyngeal swab, Blood, Tracheal swab, Urine, Stool, Clo*

Last vaccinated
provide details if applicable

Treatment
Example: Include drug name, dosage

Sequencing technology*
Example: Illumina Miseq, Sanger, Nanopore MinION, Ion Torrent, etc.

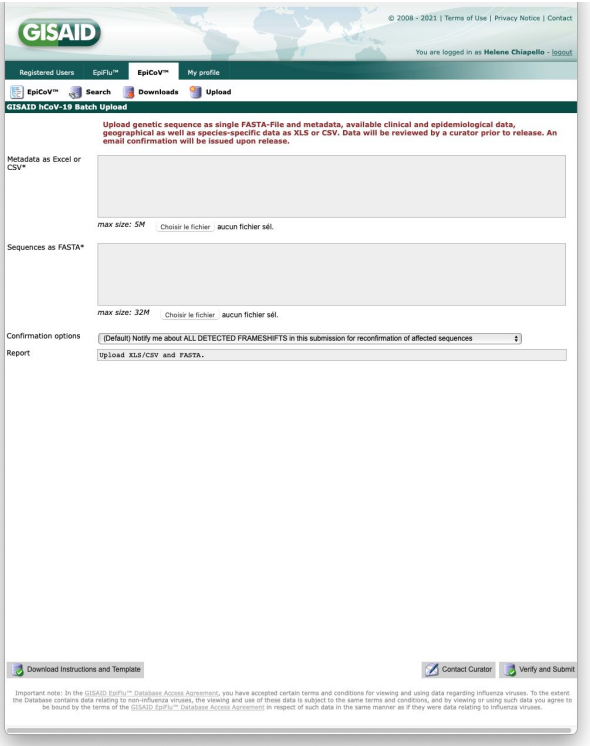
Assembly method
Example: CLC Genomics Workbench 12, Geneious 10.2.4, SPAdes/MEGAHIT v1.2.9, UGENE v. 33, etc.

Coverage
Example: 70x, 1,000x, 10,000x (average)

Institute information

Originating lab*
Where the clinical specimen or virus isolate was first obtained

Web - Batch upload



GISAID CLI2

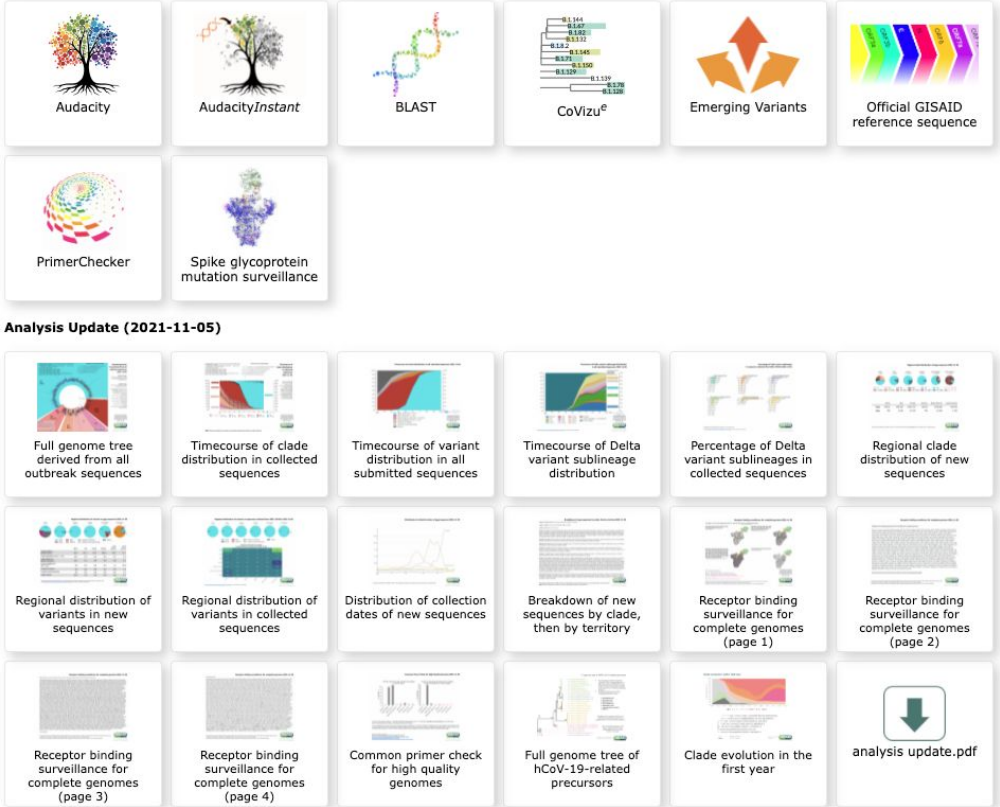
Version 2 Command Line Interface (CLI) for batch uploading

```
usage: cli2 upload [-h] [--database {EpiCoV,EpiFlu,EpiRSV}] [--token TOKEN] --metadata METADATA --fasta FASTA
                   [--frameshift {catch_all,catch_novel,catch_none}] [--failed FAILED] [--proxy PROXY] [--debug] [--log LOG]

Perform upload of sequences and metadata to GISAID's curation zone.

optional arguments:
-h, --help            show this help message and exit
--database {EpiCoV,EpiFlu,EpiRSV}
                      Target GISAID database. (default: EpiCoV)
--token TOKEN         Authentication token. (default: ./gisaid.authtoken)
--metadata METADATA   The csv-formatted metadata file. (default: None)
--fasta FASTA         The fasta-formatted nucleotide sequences file. (default: None)
--frameshift {catch_all,catch_novel,catch_none}
                        'catch_none': catch none of the frameshifts and release immediately; 'catch_all': catch all frameshifts and require email
                        confirmation; 'catch_novel': catch novel frameshifts and require email confirmation. (default: catch_all)
--failed FAILED       Name of CSV output to contain failed records. (default: ./failed.out)
--proxy PROXY         Proxy-configuration for HTTPS-Request in the form: http(s)://username:password@proxy:port. (default: None)
--debug               Switch off debugging information (dev purposes only). (default: True)
--log LOG             All output logged here. (default: ./upload.log)
```

Les outils complémentaires



Data brokering à l’IFB

Pourquoi le développer à l’IFB

Constat

- Les soumissions sont souvent complexes et difficiles à réaliser par les équipes expérimentales.
- Les métadonnées sont souvent mal comprises, ce qui entraîne des soumissions incomplètes, redondantes et incohérentes.

L'ENA a demandé à l’IFB de devenir le data broker français

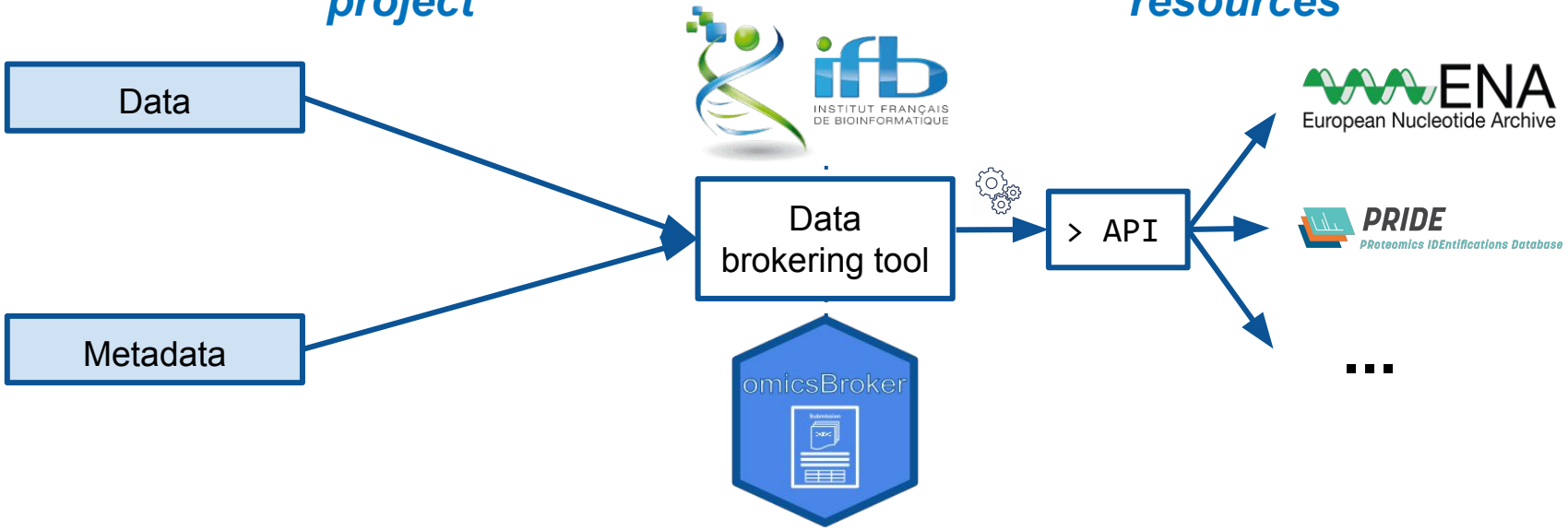
Idée principale : offrir un service national de data brokering à IFB pour **simplifier** et **rationaliser** les échanges de données entre les ressources internationales et le nœud Elixir français IFB.

3 types d'activités : le développement d'outils, la formation et le support aux utilisateurs.

Data Brokering service developed by IFB

IFB services to manage and centralize data and metadata of a project

IFB services to submit data and metadata of a project to international resources



The omicsBroker tool

omicsBroker is a tool to easily annotate and submit **omics data** to **international repositories**

Prototype disponible (soumission dans la zone de test de l'ENA)

- Développé en Django
- Disponible en Docker

Futurs développements

- Gestionnaire de soumission,
- API,
- ...

Exemple du prototype

Metadata table

Excel

	Experience name	Organism	Platform	Instrument	Library layout	Insert size
1						
2						
3						
4						
5						
6						
7						
8						
9						
10						

Descriptions

Search

Platform

Definition

Platform name. Permitted values : https://ena-docs.readthedocs.io/en/latest/submit/reads/webin-cli.html#permitted-values-for-platform

Value

LS454 ; ILLUMINA ; PACBIO_SMRT ; ION_TORRENT ; CAPILLARY ; OXFORD_NANOPORE ; DNBSEQ

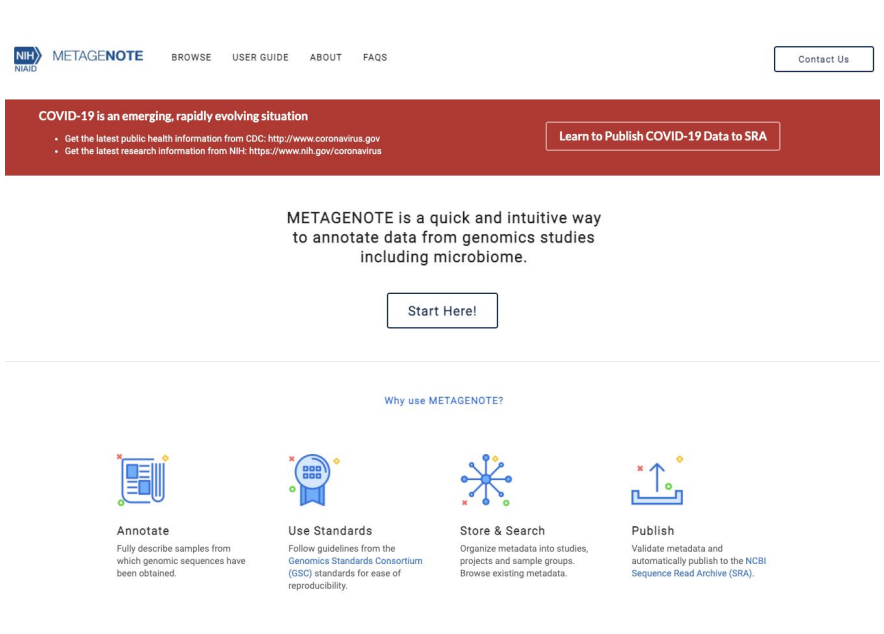
Harmonized Name

PLATFORM

* Mandatory

Des outils de data brokering déjà disponibles

https://www.gfbio.org/



https://metagenote.niaid.nih.gov/