ena-browser-docs Documentation

Release latest

Jan 12, 2023

About ENA

Welcome to the documentation for the European Nucleotide Archive (ENA) Web Browser. Please use the links below and to the left to navigate through information on the ENA browser, our collaboration projects and data policies.

For submission guidelines and retrieval tutorials, please see the ENA Training Documentation.

CHAPTER 1

About the European Nucleotide Archive

The European Nucleotide Archive (ENA) is an open, supported platform for the management, sharing, integration, archiving and dissemination of sequence data.

Database of record and platform for data management: ENA comprises both the globally comprehensive data resource that preserves the world's public-domain output of sequence data (Read more about ENA Content here) and a rich portfolio of tools and services to support the management of sequence data.

A data foundation: as nucleotide sequencing becomes increasingly central to applied areas such as healthcare and environmental sciences, ENA has become a foundation upon which scientific understanding of biological systems may be assembled. Our users comprise data submitters, data coordinators for sequence-based studies, direct data consumers and secondary service providers (e.g., UniProt, RNAcentral, EBI Metagenomics, Ensembl, Ensembl Genomes, ArrayExpress) that build on ENA services and content.

Data coordination: our data coordination partnerships span the life sciences, covering such areas as livestock genomics, marine biotechnology, biodiversity, pathogen surveillance and stem cell biology. Within these partnerships, we support data operations variously through the provision of technology, standards, data analysis, training, support and web/API data portals. Read more here.

Ambition: we are committed to the utility of the ENA platform and to achieving the broadest reach and utility of sequencing technology and data.

CHAPTER 2

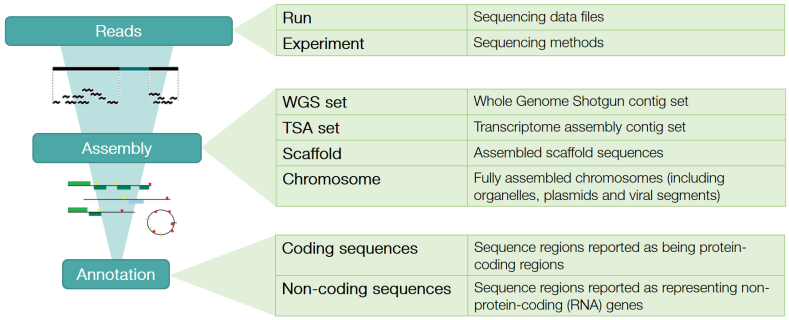
ENA Content

The European Nucleotide Archive (ENA) captures and presents information relating to experimental workflows that are based around nucleotide sequencing. A typical workflow includes the isolation and preparation of material for sequencing, a run of a sequencing machine in which sequencing data are produced and a subsequent bioinformatic analysis pipeline. ENA records this information in a data model that covers input information (sample, experimental setup, machine configuration), output machine data (sequence traces, reads and quality scores) and interpreted information (assembly, mapping, functional annotation).

Data arrive at ENA from a variety of sources. These include submissions of raw data, assembled sequences and annotation from small-scale sequencing efforts, data provision from the major European sequencing centres and routine and comprehensive exchange with our partners in the International Nucleotide Sequence Database Collaboration (INSDC).

Provision of nucleotide sequence data to ENA or its INSDC partners has become a central and mandatory step in the dissemination of research findings to the scientific community. ENA works with publishers of scientific literature and funding bodies to ensure compliance with these principles and to provide optimal submission systems and data access tools that work seamlessly with the published literature.

ENA is organised into three tiers: reads, assemblies and annotations.



Data from the ENA tiers are organised into domains. See here for the full list of ENA data domains with descriptions and example records.

Although the ENA has almost 30 years of history, the data and services are constantly changing to reflect growing volumes of data, ever improving sequencing technology and the broadening of applications to which sequencing is now put. Latest developments and changes to services are announced here and users are encourage to join the ENA mailing list ena-announce.

As part of the global effort to improve access to and usability of nucleotide sequencing data, we collaborate extensively in the development of our services and technologies and in standards activities.

The ENA is developed and maintained at the EMBL-EBI under the guidance of the INSDC International Advisory Committee and a Scientific Advisory Board.

CHAPTER 3

How to Cite Data in ENA

In all cases, the top-level Project accession should be cited as well as a link to where the data can be found in the browser, for example:

"the data for this study have been deposited in the European Nucleotide Archive (ENA) at EMBL-EBI under accession number PRJEBxxxx (https://www.ebi.ac.uk/ena/browser/view/PRJEBxxxx)."

If there is a particular scenario where using the top level accession would not be suitable, for example, if you have multiple publications that reference individual components within a single ENA project (and therefore the project accession provides too much ambiguity), then the following accessions are also considered accessions that could be used for publication:

- Assemblies (e.g. format GCA_123456789.1)
- BioSamples in the context of associated data (e.g. format SAMEA123456)
- Assembled/Annotated Sequences including contig, scaffold and chromosome sequences generated from an assembly submission (e.g. format A12345.1)

All accessions issued by ENA, along with all accessions issued by its international partner institutions under INSDC, can be viewed in the ENA browser by pasting the accession into the Text search box or through the URL: http://www.ebi.ac.uk/ena/browser/view/<accession>

e.g. http://www.ebi.ac.uk/ena/browser/view/BN000065

Numeric ranges for accessions with the same prefix are also supported via http://www.ebi.ac.uk/ena/browser/view/<accession1>-<accession2>

e.g. http://www.ebi.ac.uk/ena/browser/view/BN000066-BN000070

3.1 ORCID Data Claiming

The ORCID system (https://orcid.org/) will be familiar to many as a way to claim publications, grants and other types of contributions against an ID that is unique to you.

ENA studies/projects can now be claimed against your ORCID ID, in the same way that publications can. This is implemented through the EBI search interface : https://www.ebi.ac.uk/ebisearch/orcidclaimdocumentation.ebi

To claim your data, simply search in the 'ENA Study' search box using accession numbers or keywords to find your project(s). This will return a list of results for your review. To claim projects, select the relevant checkboxes and click 'Claim to ORCID' at the top of the search results. This will prompt you to log in to your ORCID account, if you haven't already done so.

CHAPTER 4

Data Coordination

We provide focused support for scientific projects and initiatives through our Data Coordination programme, in which we collaborate with partners across and beyond the life sciences to support the application of molecular methods. Addressing the challenges of data-intensive methods, interdisciplinary science, multi-omics and complex and deep data, we typically provide web and programmatic platforms for data sharing and publication, comprising tools, analysis support, cloud compute accessibility, data navigation and visualisation systems and user training and support.

4.1 Our Current Data Coordination Projects Include:

- European COVID-19 Data Platform supported by various partners and projects
- Pathogen epidemiology supported by BY-COVID, ReCoDID and VEO
- Marine science such as Tara Oceans and Ocean Sampling Day, AtlantECO and BlueCloud
- European sequencing facility access
- Investigating the lifelong effects of early life stress on health
- Improving taxonomy and service linking through a universal taxonomic framework for Eukaryotic organisms as well as efforts to link sequences to taxonomies and natural history collections
- Offering data coordination support for projects generating biodiversity reference genomes including Darwin Tree of Life, Aquatic Symbiosis Genomics and Biodiversity Genomics Europe

4.2 Previous Data Coordination Projects Include:

- Livestock functional genomics
- Stem cell research through EBiSC and HiPSCi

Contact us to discuss Data Coordination collaboration.

CHAPTER 5

ENA and INSDC Policies

The International Nucleotide Sequence Database Collaboration (INSDC) has been an international collaboration between DDBJ, EMBL, and GenBank for over 20 years. Its advisory committee, the International Advisory Committee (IAC), is made up of European, Japanese and US chapters; membership of the European chapter overlaps that of the ENA Scientific Advisory Board (SAB). In 2002, the IAC endorsed and reaffirmed the existing data-sharing policy of the three databases that make up the INSDC, which is stated below.

Individuals submitting data to the international sequence databases managed collaboratively by DDBJ, EMBL, and GenBank should be aware of the following:

1. The INSDC has a uniform policy of free and unrestricted access to all of the data records their databases contain. Scientists worldwide can access these records to plan experiments or publish any analysis or critique. Appropriate credit is given by citing the original submission, following the practices of scientists utilising published scientific literature.
2. The INSDC will not attach statements to records that restrict access to the data, limit the use of the information in these records, or prohibit certain types of publications based on these records. Specifically, no use restrictions or licensing requirements will be included in any sequence data records, and no restrictions or licensing fees will be placed on the redistribution or use of the database by any party.
3. All database records submitted to the INSDC will remain permanently accessible as part of the scientific record. Corrections of errors and update of the records by authors are welcome and erroneous records may be removed from the next database release, but all will remain permanently accessible by accession number.
4. Submitters are advised that the information displayed on the Web sites maintained by the INSDC is fully disclosed to the public. It is the responsibility of the submitters to ascertain that they have the right to submit the data.
5. Beyond limited editorial control and some internal integrity checks (for example, proper use of INSDC formats and translation of coding regions specified in CDS entries are verified), the quality and accuracy of the record are the responsibility of the submitting author, not of the database. The databases will work with submitters and users of the database to achieve the best quality resource possible.

The INSDC is an outstanding example of success in building an immensely valuable, widely used public resource through voluntary cooperation across the international scientific community. This success has been achieved by following the guidelines and principles outlined above.

5.1 Data availability policy

While the INSDC databases hold public data, there are several levels of data availability which control access to these data. See the INSDC Data Availability Policy for full details of INSDC data access and control.

The two main levels to data availability are when data are confidential pre-publication and then after public release.

Confidential Data	Public Data
A data owner can indicate during study/project registration that confidentiality is required until an owner-managed release date or publication in the literature, whichever comes earlier. During the confidential phase, data are not available publicly through any means.	A project is subsequently and automatically released as Public on reaching the specified release date or when the relevant INSDC accession cited online or in a publication prior to this date. In the event that a release date must be extended, data owners can extend the release of their data before it becomes public.

5.1.1 Removing data from the public browser

ENA general policy is that data which has been released into the public domain should remain public. As the submitter you need to make sure you specify the correct release date when submitting and send release date extension requests to ENA at least two weeks before the release date. Once the data has been fully released, the availability of the data is then managed at ENA and you must contact us in the event of there being an issue with the public availability of your data.

In particular, please contact us **in the event that:**

1. You realise that your data is incorrect or contaminated with no immediate opportunity to be updated.
2. You failed to manage your project release date and your project is released earlier than intended. If this is the case, please provide a reason that your data requires suppression from the browser and provide a new date for the project release.
3. You requested a **Confidential** status or an extension to an existing release date, but the ENA, or their submissions brokering collaborator, has failed to apply the appropriate release date correctly.
4. Data are found to have been submitted to the databases without the permission of the rightful owner. *This is expected to be extremely rare and requires formal institutional contact with the submitting institution.*

In any case where the data has been distributed as public, the INSDC partners cannot exercise any control on the resultant use of the data by third parties, even if it is subsequently removed from the service.

5.2 ENA policy relating to compression of submitted data

The European Nucleotide Archive (ENA) is committed to the safeguarding into the future of the world's public domain nucleic acid sequencing data.

In order to provide economically sustainable archiving, ENA team is actively developing CRAM, a technology for raw sequence read data compression. This technology offers both lossless compression, in which read sequence and per-base quality information is faithfully preserved, and lossy models, in which data are selectively reduced to reach an optimal balance between data preservation and compression.

It is our aim with CRAM to provide a flexible technological framework in which data producers, the broad scientific community that consumes ENA data, and funding agencies are empowered to make decisions about the level of compression that can appropriately be applied to different data sets.

ENA does not currently apply CRAM compression on incoming data and will not in the future apply lossy compression on submitted data without prior announcement and prior consultation with principal stakeholders. In addition, for legacy data already submitted and loaded into ENA, we will not seek to apply lossy compression without discussion with data owners.

Users may be aware that we currently preserve original submitted data files. Once data are loaded, these files contain redundant information with that integrated into ENA. As such, we have never committed to preserving these submitted files and will, in due course, cease to sustain their storage.

5.3 Third party data

Third PArty data (TPA) are submitted to the International Nucleotide Sequence Databases as part of the process of publishing biological studies that include the assembly and/or annotation of *existing INSDC reads and primary sequences.* Publicly accessible TPA data are therefore linked to a publication or publications that document the derivation of the data supported by peer-reviewed scientific evidence.

The ENA Content team review and assist with TPA submissions on a case-by-case basis. Please contact us if you would like to submit a record which fits the above description.

Based on the nature of TPA data, i.e. a type of record that is generated from public INSDC Read or Sequence/Trace data, which is not owned by the submitting group, these records undergo a strict release policy. TPA sequences should be planned for publication in a peer-reviewed journal, which discusses the TPA records unambiguously and encompasses the concepts of (re-)annotation, (re-)assembly or a combination of these. Once TPA records have been accepted by the database, they must be cited by accession number in the peer-reviewed journal article.

5.4 Publication

Soren Brunak, Antoine Danchin, Masahira Hattori, Haruki Nakamura, Kazuo Shinozaki, Tara Matise, Daphne Preuss (2002) Nucleotide Sequence Database Policies

Science 298 (5597): 1333 15 Nov 2002

5.5 INSDC membership

Please refer here for details of current membership of the European chapter of the INSDC IAC and the ENA SAB and here for the full membership of the IAC.

CHAPTER 6

Data Standards

Harmonization of data and metadata collection becomes an essential effort in the age when data generation is often easier and more affordable then their organization and storage.

Compliance of submitted data to the relevant reporting standards promotes:

- consistent and adequate data description
- thorough data validation
- data discoverability
- data reproducibility
- data interoperability and usability

6.1 ENA/INSDC reporting standards

The European Nucleotide Archive requires, where appropriate, use of the following reporting standards:

- Feature Table – Description of nucleotide sequence provenance and functional annotation of nucleotide sequence domains.
- Third Party Data – ENA use the INSDC agreed standards for capturing and presenting TPA data. Contact us if you intend to submit data that comprises of assembled or annotated data of existing INSDC records.
- Missing values - Guidelines for registering metadata which is missing or restricted access.

6.2 Community-developed reporting standards

The European Nucleotide Archive supports use of many community-developed reporting standards in the form of sample checklists. Sample checklists are a defined set of minimum information required and validated during ENA sample registration. Sample checklists have been developed with different research communities and allow data submission to abide by different community-developed standards.

The full list can be viewed and explored here.

As part of our community engagement and standards development, the European Nucleotide Archive has a long-standing collaboration with the Genomic Standards Consortium (GSC). The GSC is an initiative of experts building or using genome collections and developing standards for harmonised metadata collection and analysis efforts across the wider genomics community.

The GSC supports a range of projects spanning sequencing projects, development of ontologies, metadata standards, software tools or data formats. Minimum information about any (x) nucleotide sequence (MIxS, Yilmaz et al, 2011) is the core GSC standard consisting of checklists for describing genomes (MIGS), metagenomes (MIMS) and marker sequences (MIMARKS).

CHAPTER 7

Funding

ENA is developed and operated under the support of the European Molecular Biology Laboratory (EMBL) and through grants from external bodies that include the European Commission, the British Biotechnology and Biological Sciences Research Council (BBSRC) and the Wellcome Trust (WT).

CHAPTER 8

Upcoming Events At ENA

ENA sometimes runs events or participates in those hosted by others. Details of upcoming meetings where you might find us are available below.

Date	Event	Description
Winter 2022	ENA Facilities Day	Facilities Day is an annual meeting of ENA staff with frequent users. We invite representatives of sequencing facilities, brokers, and other major service users to discuss developments at ENA, and how our services are used.

8.1 Upcoming Training Workshops

ENA contributes to various EBI training courses throughout the year. These are a fantastic opportunity to learn about a range of topics in Bioinformatics, and get some hands-on experience of using ENA submission and retrieval services. Additionally, if you are planning a course and would like to invite ENA to contribute to it, please make a ticket with our support form and the appropriate member of our team will get back to you.

Date	Event	Description
31st Oct - 4th Nov 2022	Genome-resolved metagenomics bioinformatics	This course will provide biologists who are embarking upon metagenomics research projects training to use publicly available resources to manage, share, analyse and interpret metagenomics data, including both marker gene and whole gene shotgun (WGS) approaches

CHAPTER 9

Home Page

9.1 Introduction

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.

9.2 Exploring the ENA Browser

From the homepage you can access:

- Submission Guidelines and Links
- The ENA Search Services
- Rulespace
- Our User Support Service

You can also see the latest news and tweets regarding the archive.

CHAPTER 10

Data Submission and Updates

10.1 Introduction

We offer a number of services through which data (including updates) can be submitted to the European Nucleotide Archive (ENA). These technologies provide options appropriate for the scale and frequency of submission, the expertise and capacity of the submitter and the nature of the data to be transferred. The choices below lead users most directly to the appropriate submission route.

10.2 Submitting data to ENA

From the data submission page you can access all of our submission services directly and access the submission guidelines available here.

CHAPTER 11

Sample Checklists

11.1 Introduction

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.

11.2 Standards

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. For any questions regarding checklists, contact us here.

11.3 List of checklists

Explore all checklists here.

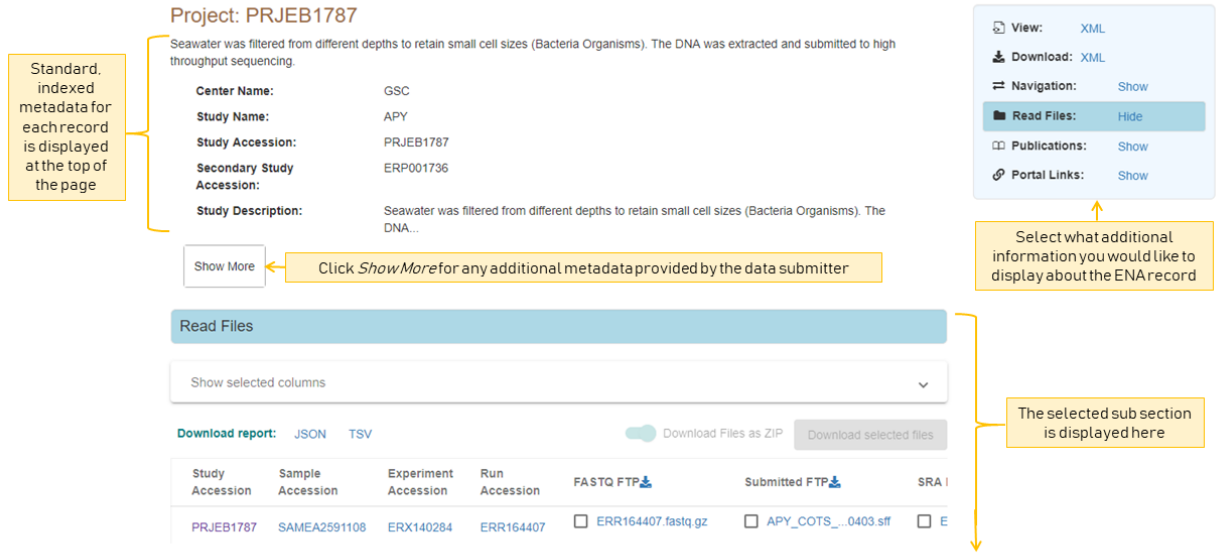
CHAPTER 12

Viewing ENA Records

12.1 ENA Records

You can view any public ENA records using their Accession (unique identifier) or as the result of a search.

12.1.1 Exploring an ENA record



All records have a selection of standard, indexed metadata which tells you about the database entry. For example, where it is submitted, basic information on the record's title and description etc. To show the full set of indexed

metadata for this record, you can click *Show More*. For additional custom metadata provided by the submitter, you can view this in the ENA record XML.

To explore a record further, you can use the navigation box in the top right of the view to show/hide different additional subsections of information.

This gives you access to any other associated data, such as if a project has any data files or publications associated with it.

Any links that looks like *accessions* (a series of letters then numbers) will take you to an associated record.

An ENA record can be one of the following types:

12.1.2 Record types

Record Type	Description	Example accessions
Projects/Studies	Contains information on a biological research project. This holds all the data generated as part of this research.	PRJEB1787/ERP001736
Samples	Represents biological samples collected and sequenced in real life	SAMEA2620084/ERS488919
Runs/Experiments	Hold raw read files and sequencing methods	ERR1701760/ERX1772048
Analyses	Hold results files of analyses performed on sequencing data and analysis methods	ERZ841272
WGS contig set	Hold Whole Genome Sequencing contig sets generated as part of a genome assembly.	CABHOY010000000.1 CABHOY010000000 CABHOY01
Assemblies	Represents an entire genome assembly and holds any contig sets or sequence records generated as part of the assembly	GCA_000001405.28
Assembled/Annotated Sequences	Any sequence records from coding or non-coding regions to full assembled chromosomes	CM000667.2
Taxon	The sequenced organism or metagenome of a sample	Taxon:9606
Sample Checklist	The checklist of metadata that the	ERC000013

12.2 View and Download Links

This panel contains the different data formats available for the current record.

12.2.1 Contigset (WGS/TLS/TSA)

- View/Download flatfile for the master record describing the set
  - EMBL https://www.ebi.ac.uk/ena/browser/api/embl/CABHOY010000000
- Download the full set of sequences in one file (Using ftp:// is recommended for programmatic downloads)
  - EMBL (from FTP) ftp://ftp.ebi.ac.uk/pub/databases/ena/wgs/public/cab/CABHOY01.dat.gz
  - FASTA (from FTP) ftp://ftp.ebi.ac.uk/pub/databases/ena/wgs/public/cab/CABHOY01.fasta.gz
  - FASTA (via API) https://www.ebi.ac.uk/ena/browser/api/fasta/CABHOY010000000

12.2.2 Project/Sample/Run/Analysis/Submission (XML types)

- View/Download metadata in XML format
  - https://www.ebi.ac.uk/ena/browser/api/xml/PRJEB402

12.2.3 Assembly (Genome Collection)

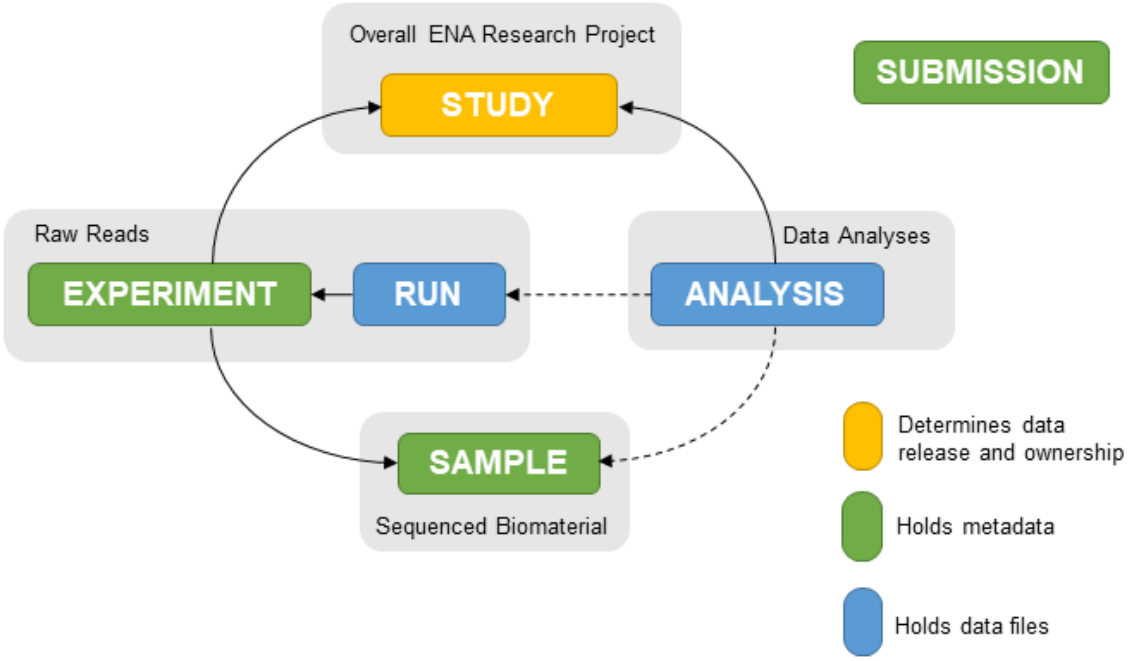
- View/Download metadata in XML format
  - https://www.ebi.ac.uk/ena/browser/api/xml/PRJEB402
- View/Download Sequence report as a text file
  - ftp://ftp.ebi.ac.uk/pub/databases/ena/assembly/GCA_900/GCA_900257/GCA_900257145.2_sequence_report.txt
- Download all sequences in the assembly, including chromosomes, WGS set and scaffolds as one file
  - EMBL https://www.ebi.ac.uk/ena/browser/api/embl/GCA_900257145.2?download=true&gzip=true
  - FASTA https://www.ebi.ac.uk/ena/browser/api/fasta/GCA_900257145.2?download=true&gzip=true

12.3 Navigation and Cross References

From here you can navigate through all associated records that were submitted within the same submission Project.

12.3.1 Organisation of a Project in ENA

When data is submitted to the archive, the data submitter establishes relationships between each record in a research project following the ENA metadata model:



This is so someone viewing the data can easily navigate between the records. This makes it easy to explore the data generated from biological samples which were sequenced/analysed within the same research project.

12.3.2 Cross-references

From this tab you can also see any links from the record out to external data resources that have used or generated these records as part of their services. These mappings are compiled as part of ENA's cross-reference service, and so only show data from resources that are registered with us. You can see more details on such registered resources here.

Note: This view only gives a view of the associated records submitted as part of the originally submitted research project and any registered cross-references. For a view showing all ENA records which are associated with this record (including any other links to this record within other ENA submission projects), you can see this in the Related ENA Records tab (available for Project, Sample and Taxon records).

12.4 Read Files

Display and download any associated raw read files. Please refer to Archive Generated Files for more information about file formats.

There are several ways to download read files:

12.4.1 1. Using ENA File Downloader Command Line Tool

The ENA File Downloader is a new Java based command line application. You have to submit one or more comma separated accessions, or a file with accessions that you want to download data for. This tool allows downloading of

read and analysis files, using FTP or Aspera. It has an easy to use interactive interface and can also create a script which can be run programmatically or integrated with pipelines.

Download the latest version from ENA Tools.

12.4.2 2. File Reports

You can download a Report of all the data displayed in the table or download files selected from the table. To download all files in the column, click the download icon in the table header.

To choose additional metadata to add to the table display and report, use the 'Show selected columns' expandable menu.

12.4.3 3. enaBrowserTools

You can also download files from ENA using the Python based scripts enaBrowserTools.

12.5 Analysis Files

Display and download any associated analysis files. There are three ways to download analysis files:

12.5.1 1. Using ENA File Downloader Command Line Tool

The ENA File Downloader is a new Java based command line application. You have to submit one or more comma separated accessions, or a file with accessions that you want to download data for. This tool allows downloading of read and analysis files, using FTP or Aspera. It has an easy to use interactive interface and can also create a script which can be run programatically or integrated with pipelines.

Download the latest version from ENA Tools.

12.5.2 2. File Reports

You can download a Report of all the data displayed in the table or download files selected from the table. To download all files in the column, click the download icon in the table header.

To choose additional metadata to add to the table display and report, use the 'Show selected columns' expandable menu.

12.5.3 3. enaBrowserTools

You can also download files from ENA using enaBrowserTools.

12.6 Publications

Explore publications that either cite the record or document the research where the record was generated.

This view provides links to the DOI or in some cases, a direct link to the PDF or article in Europe PMC.

12.7 Component Projects

In the case of an **Umbrella Project** (a project which is used to group many related sub-projects) there is the option to explore its Component Projects.

Component projects are the same as other project records in ENA but are grouped under one 'umbrella' meaning they will often have the same research motivation and will often represent a collaborative research effort.

12.8 Parent Projects

If a project has a parent project it is part of an **Umbrella Project** (a project which is used to group many related sub-projects).

Projects grouped under one 'umbrella' often have the same research motivation and will often represent a collaborative research effort. You can navigate to the parent project through this tab and view other related component projects through the 'Component Projects' tab.

12.9 Related ENA Records

This view provides a summary of all data associated with this record. Any submission in ENA that is associated with this record is available here.

This view is only available for three ENA record types:

Study: Here you can find all components of the project including any sequence or assembly records associated with the project.

Sample: Here you can find all sequencing records or analyses associated with the sample including assembly or sequence records. This view shows any third party uses of the sequencing data registered with ENA.

Taxon: Here you can see a summary of all ENA records registered with that particular taxon. This view also shows a summary of any records registered with descendant taxa.

12.10 Tax Tree

Here you can view the full tax tree of this taxon record.

From this view you can access all taxon records within this tax tree and explore ENA records that are registered with related taxa.

Click the arrows to expand the tree and explore the full lineage of the taxon.

12.11 Assembly Versions

If this assembly has been updated, you can view the past assembly versions here.

12.12 Assembly Statistics

Assembly statistics are generated for all assemblies submitted to INSDC.

Total Length (total sequence length) - total length of all top-level sequences.

Ungapped Length (total ungapped length) - total length of all top-level sequences ignoring gaps. Any stretch of 10 or more Ns in a sequence is treated like a gap.

Chromosomes & Plasmids (total number of chromosomes and plasmids) - total number of chromosomes, organelle genomes, and plasmids in the assembly.

Spanned Gaps - total number of gaps between contigs/scaffolds.

Unspanned Gaps - total number of unspanned gaps between scaffolds.

Regions/Patches/Alternative Loci - (number of regions with alternate loci or patches) - number of genomic regions that contain one or more alternate loci or patch scaffolds.

Scaffolds (number of scaffolds) - number of scaffolds including placed, unlocalized, unplaced, alternate loci and patch scaffolds.

Scaffold N50 - length such that scaffolds of this length or longer include half the bases of the assembly.

Contigs (number of contigs) - total number of sequence contigs in the assembly. Any stretch of 10 or more Ns in a sequence is treated as a gap between two contigs in a scaffold when counting contigs and calculating contig N50 & L50 values.

Contig N50 - length such that sequence contigs of this length or longer include half the bases of the assembly.

12.13 Chromosomes

When an assembly is is declared as assembled to full chromosome level on submission, chromosome sequences are generated for each chromosome submitted in the assembly.

These chromosomes are available as individual sequence records and can be explored in full here.

12.14 BlobToolKit

BlobToolKit is a set of computational tools developed to identify cross-species contamination within genome assemblies. A summary of results and graphics generated by BlobToolKit is displayed on the ENA browser to give data providers and consumers access to visualisation tools needed to identify contamination in public genome assembly data. BlobToolKit was developed by Richard Challis & Mark Blaxter at the University of Edinburgh.

For further information regarding BlobToolKit, please visit https://blobtoolkit.genomehubs.org.

Please send any questions or queries regarding BlobToolKit to blobtoolkit@genomehubs.org.

12.15 Checklist Fields

Sample Checklists are lists of fields that are required/recommended to be used during registration to describe samples (depending on the type of sample).

Explore the mandatory, recommended and optional fields defined as part of this checklist.

You can filter these fields further by requirement or by keywords in the field name or description.

In some cases, fields can be restricted by regular expression, a list of text choices, by valid taxonomy or by valid ontology terms.

12.16 3rd Party Curations

This tab presents the flow of 3rd party curations from the ELIXIR Contextual Data ClearingHouse (CDCH) data store.

The CDCH data store aims to provide a seamless method of exchange for curated contextual data available in external resources and community curation efforts, with ELIXIR data resources.

CHAPTER 13

ENA Search Services

The European Nucleotide Archive can be searched through interactively using a simple text search or explored in more detail using one of the ENA search services.

Customise your own search query and tailor your results to fit a specific set of search criteria (which can be saved with Rulespace):

13.1 Advanced Search

13.1.1 Introduction

Customise your own search query and retrieve a set of ENA records tailored to your search criteria.

All searches are performed against a subset of the archive specified by the *Data type* you choose to search against. You can then build your search query to specify what data you are looking for and select what fields you want to retrieve from your search. There are additional options to include/exclude specific datasets as well as filter the number of results you wish to return.

If you intend to repeat the same search at a later date, you can save this as a Rule using Rulespace. If you want to access the same data programmatically, you can copy the produced curl request and run this yourself against the ENA Portal API.

13.1.2 Data Types

Each data type is a subset of the data and metadata held within the ENA.

You must specify which datatype you wish to search across before you can narrow down your search results with any additional criteria.

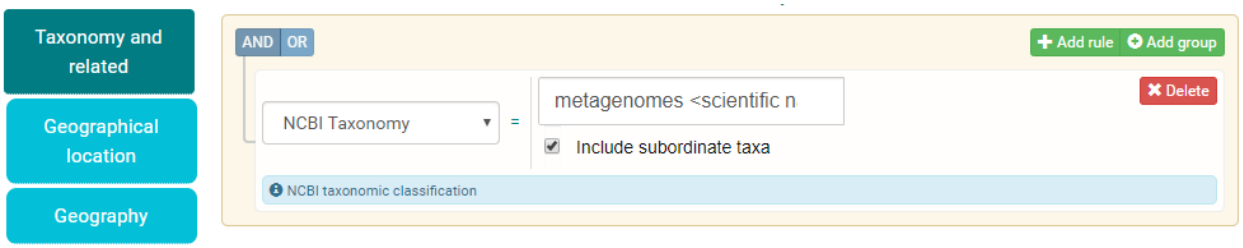
What is in each data type?

Data Type	Description	API Result type
Studies	All studies in ENA. Studies can be searched by the study metadata, e.g. title or by related taxonomy of data.	*study*
Studies used for raw reads	All studies that hold raw read datasets. Searching across this data type can give you information on these studies and can be filtered down by sample metadata (information about the collected samples in the study).	*read_study*
Studies used for nucleotide sequence analyses from reads	All studies that hold any nucleotide analyses. Searching across this data type can give you information on these studies and can be filtered down by sample metadata (information about the collected samples in the study).	*analysis_study*
Samples	All samples in ENA. Samples can be searched by sample metadata.	*sample*
Environmental samples	All samples in ENA that the submitter has flagged as 'evironmental' during sample registration. Please note, this does not include all environmental samples.	*environmental*
Experiments used for raw reads	All metadata associated with sequencing event used to generate raw reads submitted to ENA.	*read_experiment*
Raw reads		*read_run*

13.1.3 Search Query

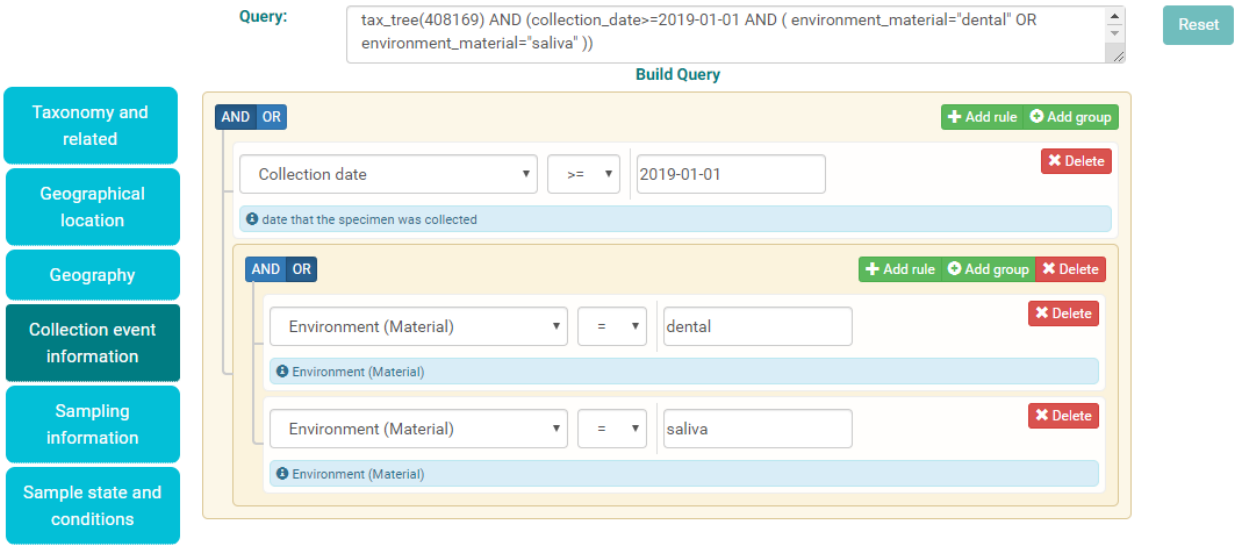
To build a query you need to create a list of rules that the resulting data sets should be restricted to.

This is done by clicking any relevant metadata types you would like to restrict (listed as buttons on the left) then selecting the relevant filters and specifying the desired restrictions for those:



When specifying taxonomy in your search, you can include all subordinate taxa within that tax tree and when searching by geographical range you can interactively drag the box or circle over the desired region to automatically fill out the location range.

These rules can be grouped and nested within AND or OR logical statements. For example, a query for *all metagenomic analyses where the sample was collected after 01 Jan 2019 AND the environmental material is either dental OR saliva* would look as follows:



13.1.4 Inclusion/Exclusion of datasets

If there are any known public datasets that do or do not fit the criteria you have specified that you wish to include or exclude from the results, you can list the accessions in a comma separated list here (with no spaces).

13.1.5 Return Fields

By default, you will receive the accession and description/title of the main datatype you are searching against. If you wish to customise the metadata which your search will return, you can manually select your search return fields from a list of all indexed fields for the specified datatype.

Select and order fields

To select fields you would like returned from your search, drag across any desired fields from the **Available Fields** list to the **Selected Fields** list. Alternatively, use the arrow buttons in the middle to move fields across from one list to the other.

The order of the **Selected Fields** list will define the order that you receive those metadata from your search. To specify the return order of these fields, you can drag and drop these into the desired order.

Field sets

Field sets are a pre-defined set of fields that can be returned together and are available for some data types. For example, for the analysis datatype, you can toggle the 'Submitted Files' field set which can be used to return all relavent fields relating to the original set of submitted files (e.g. this set includes the aspera, ftp and galaxy links for the submitted files, the size of the files (in bytes) and the files' md5 checksums).

13.1.6 Data Filters

Offset

You can specify an offset for the number of records you would like to skip from the beginning of your search. This can be used to view results beyond the maximum number of records that can be viewed in the results table (100 000) or to break up queries that result in a large number of records into smaller batches.

If you do not wish to skip any records, you can leave this field blank or enter an offset of '0'.

Limit

You can specify a data limit for the maximum number of records you would like to retrieve from your search.

If you wish to fetch the full result set enter '0'. Leaving the limit field blank applies the deafult limit of 100 000. For large result sets, to get all records please download the report (JSON/TSV) or copy and run the curl command outside of the browser.

13.1.7 Download ENA records

Here you can download the ENA records resulting from your search.

This will download the whole ENA record stored for each of the results. If you wish to only download the fields returned that were specified in your search, use one of the **Download report** options (JSON or TSV).

XML records

XML records are available for all standard metadata objects held within ENA (all results with the exception of sequence records).

XML records hold **all** the metadata for each object concatonated into a single bulk XML file. These XML metadata records are formatted in the standard ENA XML format (the same XML format that is used for data submission and for data to be displayed in the browser).

FASTA records

FASTA records hold all sequences resulting from your search concatonated into one FASTA file. FASTA records are only available when searching against sequence datatypes.

TEXT records

TEXT records hold all sequences resulting from your search and their annotation (if available) concatonated into a single EMBL flat file. TEXT records are only available when searching against sequence datatypes.

13.1.8 Download results report

This feature allows you to download all the results from your search in the format of a JSON or TSV file. Any data filters set by you will apply here.

13.1.9 Download associated data files

Pre-Conditions

To see file download columns in your results, you have to search against either the analysis or read_run data types and select the relevant fields that end with '_ftp'.

For example:

Data Type = analysis and **fields** = submitted_ftp

Data Type = read_run and **fields** = fastq_ftp / sra_ftp / submitted_ftp

Download data files

You can download the data files resulting from your search in one of four ways:

1. The ENA File Downloader is a new Java based command line application. You have to submit one or more comma separated accessions, or a file with accessions that you want to download data for. This tool allows downloading of read and analysis files, using FTP or Aspera. It has an easy to use interactive interface and can also create a script which can be run programatically or integrated with pipelines.

   Download the latest version from ENA Tools.

2. You can download a single file by clicking on its link in the FASTQ FTP, SRA FTP, or SUBMITTED FTP column.

3. You can select one or more files using the check boxes, and either download these as a bundled ZIP file or as individual files using the "Bundled ZIP" or "Individually" links above the table.

4. You can download ALL files resulting from your search as a bundled ZIP file by clicking the download icon in the column header.

Tips:

- If you wish to exclude any records from your search results before you download all the resulting files, you can go back and list these in the "Exclude Accessions" field and then repeat the search.
- If you selected multiple files and clicked the "Individually" link but only the first file is downloading, this could be because your browser is restricting multiple download pop-ups. Look for a browser warning or confirmation dialog to allow this.
- If selesting many files and using the download "Individually" option, you may wish to change the default download location of your browser. Look in your browser settings for this.
- You can also download files using a terminal from ENA using enaBrowserTools.

Search the archive for a specific sequence (Backed by NCBI BLAST+):

13.2 Sequence Similarity Search

13.2.1 Introduction

Submit a nucleotide sequence and receive a summary of all public INSDC assembled and annotated sequence records with regions of sequence similarity. Search using default parameters or optionally tailor your search further using the 'Search against' and 'Set parameters' options.

This service is backed by NCBI BLAST+ and you will be redirected to the NCBI BLAST+ service once the job is submitted. Please refer to the help & documentation for usage details and parameter options.

13.2.2 Search against a specific sequence set

Select a specific sequence set to refine your search to a specific sequence data type. If you wish to limit your search by taxonomic group or sequence data class this will restrict the search to the latest ENA release only.

13.2.3 Set parameters for your search

Searches can be performed using blastn, tblastx or tblastn. Please refer to the NCBI BLAST+ help & documentation for usage details and parameter options.

Explore external data resources which are linked to ENA records:

13.3 Cross Reference Search

13.3.1 Introduction

The ENA Xref service holds cross-references to a number of external data resources linked to ENA records. These cross-reference sources include both services operated by colleagues at EMBL-EBI (such as UniProt and Ensembl) as well as those operated outside EMBL-EBI (including SILVA and RFAM).

The update and frequency of each source is dependent on their own release cycle and/or internal processes, with ENA supporting updates as frequently as once a week.

These cross-references can also be explored programmatically using the Xref API.

If you would be interested in registering your resource as part of our cross-reference service, you can request to be added as a new Xref source here *and a member of the team will get into contact with you to discuss your eligibility.*

13.3.2 Xref Sources

A cross-reference source is the external data source that has been registered as a cross-reference within ENA. For more information on the available cross-reference sources, see the Source Details tab.

13.3.3 Xref Targets

A cross-reference target is the ENA data type which the external cross-reference is associated with. Current ENA targets:

- analysis
- assembly
- coding
- experiment
- noncoding
- run
- sample
- sequence
- study
- taxon
- trace
- wgsmaster

13.3.4 Expanded View

In some cases, the external data source will have provided additional useful information, you can choose to display this additional information by selecting the 'Expanded' view.

13.3.5 Search for Xrefs by ENA record

Return all cross-references associated with an ENA record by searching using the INSDC accession.

Look into the history of a sequence record and explore previous versions of an INSDC sequence:

13.4 Sequence Versions Archive

13.4.1 Introduction

The Sequence versions archive holds all publicly available versions of INSDC submitted sequences including the original dates of their release.

Search the full history of a sequence using its accession and view the sequence versions in FASTA or EMBL flat file format.

CHAPTER 14

Rulespace

14.1 Introduction

Explore your saved Rules with Rulespace. From here you can create a new rule or share and re-run your saved rules generated from the ENA Advanced Search.

14.2 Log into Rulespace

14.2.1 What is Rulespace?

Rulespace can be used to save, share and re-run custom search queries generated from the ENA Advanced Search.

Log in to access and manage your set of saved rules.

14.2.2 Log in using an AAP Account

The Authentication, Authorisation and Profile service (AAP) provides a central log-in for multiple different services at EMBL-EBI (and can be used by other services and organisations as required).

You can register for an AAP account here.

14.2.3 Log in using a Life Science Account

Logging in with your Life Science account enables you to log in to Rulespace use your home or organisation credentials (including the options to log in with your Google, Apple or ORCID accounts).

To log in with LIFE SCIENCE, you first need to register for a LIFE SCIENCE ID.

You can register with LIFE SCIENCE here.

14.3 Create a Rule

To create a rule, you need to create a custom search query using the ENA Advanced Search service. Follow the step by step guide on creating a custom search query and defining the fields you want returned from your search.

When you are happy with your search click *Search* then, on the results page you have the option to save the search to Rulespace:

14.4 Share a Rule

You can share your rules with others by providing them with one of four following options:

1. **The Search URL** - you can copy and share a direct URL to the results from your rule search. This URL links directly to the resulting metadata in the specified format (TSV or JSON) (not to the results page) so can be used to programmatically access the results of the search if you wish.

2. **The Rule URL** - you can copy and share a direct URL to the JSON object of the rule itself. This can be used to programmatically access information about the search parameters of the rule as well as information on when it was most recently updated.

3. **The Rule ID** - you can copy and share your rule ID which can be used to repeat the search you have saved in the ENA Advanced Search service.

4. **The Rule Name** - all rules have to be saved under a *unique name* so you can also copy and share your Rule name which can then be used to repeat your saved search in the ENA Advanced Search service.

Please note: Whenever the results of a Rulespace search are accessed, the search is performed again, live each time, so if data within the archive has changed since the last search, the results can vary since the search was last run.

14.5 Accessing Rulespace Programmatically

The features of the Rulespace service can also be accessed directly via the API at:

https://www.ebi.ac.uk/ena/rulespace/api/

For more information on the Rulespace API service, you can access the API documentation here.

CHAPTER 15

User Support

15.1 Introduction

By completing this Support form you are contacting the helpdesk of the European Nucleotide Archive (ENA). Please ensure you provide as much detail as possible here. This will ensure your query will be directed to the right person and answered as soon as possible.

Subscribe to the ENA-announce mailing list to receive alerts about ENA services.

15.2 Contact us

Complete the Support form here.

CHAPTER 16

Moving away from the release

16.1 Introduction

The ENA retired its periodic assembled/annotated sequence release in March 2020. The last release was number 143.

The European Nucleotide Archive (ENA) captures, preserves and presents the world's nucleotide sequence data. Since 1982 the European Nucleotide Archive has made more than 140 individual releases, providing a quarterly snapshot of ENA assembled/annotated sequence data. During this time, changes to the ways in which users access ENA data, have led us to develop a portfolio of data access tools, such as our daily FTP products and the ENA Browser API, which are currently offered in parallel to the traditional release. In recent years we have faced growing pressure on the release process in response to increases in data volume and have also seen a shift towards our newer services from the majority of users. Our release process has remained largely unchanged for the last two decades, and following an internal review we have concluded that it is no longer viable for us to continue the current release process as part of our presentation portfolio.

New data is already included in the ENA on a continuous basis and distributed daily from our ENA Browser, FTP and RESTful API services. The key change is that we will no longer make an additional separate quarterly release of the assembled/annotated subset of sequences. We will focus our resources on further developing and supporting our continuous distribution presentation products.

Additionally, as part of the release retirement we will no longer be creating cumulative FTP files in the FTP update folders (e.g. http://ftp.ebi.ac.uk/pub/databases/ena/sequence/update/). These cumulative files tracked daily changes in between release cycles and thus cannot continue to be produced sustainably. Release 143 will be the last available in the release folder (available here once released http://ftp.ebi.ac.uk/pub/databases/ena/sequence/release/), the update folder will be removed after this last release. Set based sequences have already been removed from the release and will continue to be added to the FTP in their corresponding folders (e.g. http://ftp.ebi.ac.uk/pub/databases/ena/wgs/public/; http://ftp.ebi.ac.uk/pub/databases/ena/tsa/public/; http://ftp.ebi.ac.uk/pub/databases/ena/tls/public/).

16.2 Deprecated: Release vs Update search results

We no longer maintain separately indexed datasets for RELEASE and UPDATE data for Sequence and Coding & NonCoding RNA records. Where RELEASE referred to the last ENA release, and UPDATE referred to any records

that had been added or modified since the last release.

After the final release (143) in March 2020, in our Advanced Search service, we've now merged the '_release' and '_update' data types for sequence, coding and non-coding. So the data types 'sequence_release' and 'sequence_update' were replaced with the data type 'sequence'. This affects users of our API and Browser advanced search services, who will need to use the updated data type end points.

The following guide has been created to assist users in moving away from the release. This guide outlines accessing assembled/annotated sequences, guidance on how to identify data based on a last updated timestamp and advice for establishing your own mirroring procedures using our portfolio of other access services.

16.3 Accessing assembled/annotated sequences

Assembled/annotated sequences can be obtained from our continuous daily distribution resources, with API, FTP and web browser-based options. For most use cases we would recommend the ENA Browser API as it provides the greatest specificity and flexibility for obtaining a tailored dataset of assembled/annotated sequences for your requirements.

16.3.1 ENA API

Assembled/annotated sequences can be identified and downloaded with our ENA Browser API. The http API Swagger interface lists the endpoints, documenting expected parameter and errors.

Examples: (we provide curl examples, but you could use wget or a web browser or a rest client)

Obtaining the latest version of a sequence record by accession:

In EMBL format:

```
curl -X GET "https://www.ebi.ac.uk/ena/browser/api/embl/BN000065"
```

In FASTA format

```
curl -X GET "https://www.ebi.ac.uk/ena/browser/api/fasta/BN000065"
```

Obtaining a specific version, including suppressed versions, of a sequence record by accession:

In EMBL format:

```
curl -X GET "https://www.ebi.ac.uk/ena/browser/api/embl/KF961410.1"
```

The ENA Browser API also allows the user to conduct a search for multiple Assembled/annotated sequences records and download them. In this example searching the sequence data type for human data distributed or updated since 19th August 2019: In EMBL format

```
curl 'https://www.ebi.ac.uk/ena/browser/api/embl/search?result=sequence&query=tax_
↪eq(9606)%20AND%20last_updated%3E%3D2019-08-18&limit=5' -o embl.txt
```

or FASTA

```
curl 'https://www.ebi.ac.uk/ena/browser/api/fasta/search?result=sequence&query=tax_
↪eq(9606)%20AND%20last_updated%3C%3D2019-08-18&limit=5' -o fasta.txt
```

We have added limits to the above examples to only return 5 records.

If not provided, limit defaults to 100000. To retrieve ALL records matching a query, user limit=0.

You can search using the sequence, coding or noncoding data type endpoints. In general when using the API search it is important to be as specific as possible with your query to save on downloading sequences that you do not require.

16.3.2 ENA FTP

The release folders, for example the sequence release folder (http://ftp.ebi.ac.uk/pub/databases/ena/sequence/release/) will contain the final release 143 made in March 2020. No further FTP releases will be made after release 143.

16.3.3 ENA Browser

For the majority of use cases we would recommend utilizing the ENA Browser API for obtaining assembled/annotated sequences. However, these are also available to search and download from the ENA Browser.

The ENA Browser provides direct access to sequences by accession, with subsequent options for downloading in EMBL or FASTA format; e.g. see https://www.ebi.ac.uk/ena/browser/view/BN000065

The ENA Browser also provides an Advanced Search for finding appropriate assembled/annotated sequences for download. This feature is also useful for assistance with constructing complex API queries. In particular one could use the graphical interface to construct the query and then export it for command line using the "Copy Curl Request" button.

Detailed guidance on the usage of Advanced Search is available in our Advanced Search documentation, but we make a brief mention here:

1. Start an advanced search at https://www.ebi.ac.uk/ena/browser/advanced-search
2. Select an assembled/annotated sequence data type such as 'sequence', 'coding' or 'noncoding'
3. (Recommended) Use the Query builder to be as specific as possible with the available filters to construct a query that will limit the resulting dataset to match your needs. e.g. Key filters include:
    - limiting by date. Database record -> last updated
    - taxon. Taxonomy and related -> NCBI taxonomy.
4. (optional) Select the fields you want in the resulting data. By default, the INSDC accession and description is provided.
5. (Optional) Use inclusion and exclusion lists of accessions to finely alter the returned records.
6. Once you have run your query you can click the hyperlinks to download the full data files in in either EMBL or FASTA format.
7. (Optional) If desired you can copy your query for command line use with the ENA APIs using the "Copy Curl Request" button.
8. (Optional) You can save this query for future use, by saving it to your Rulespace account using the 'Save To Rulespace' button, please refer to this guide for more information.

16.4 Periodic Snapshots & Support API

For sequence, coding and noncoding RNA data, we produce a periodic snapshot which includes all public records at that time point. These are available from FTP. These snapshots are different from the old release approach in these aspects:

1. Are more frequent. We aim to produce these twice a month.
2. Release numbers will not be updated in the flatfile DT lines

16.5 Assembled/Annotated Sequences

Latest snapshot is available at ftp.ebi.ac.uk/pub/databases/ena/sequence/snapshot_latest/.

snapshot_latest is a symlink that points to the most recent snapshot. This is also listed in the text file snapshot_latest.txt in the parent folder. In this folder, the records are divided into con, expanded_con and std subfolders. std subfolder contains all dataclasses that are not CON (STD, EST, GSS, PAT etc.) Records are in gzip files, further divided by taxonomic division, with upto 1,000,000 records per file.

16.6 Coding & Noncoding RNA Sequences

CDS and NCRNA subproducts from CON & STD (incl. EST, GSS etc) are treated the same was as Assembled/Annotated Sequences. The latest snapshots are available at

ftp.ebi.ac.uk/pub/databases/ena/coding/snapshot_latest and

ftp.ebi.ac.uk/pub/databases/ena/non-coding/snapshot_latest respectively.

But for subproducts from WGS/TSA/TLS sequences, the records are made available in a different manner. We group the coding records from a given WGS set in to one file. Then files are grouped set-name based on 3 character prefix into a tar file. e.g. ftp.ebi.ac.uk/pub/databases/ena/non-coding/snapshot_latest/wgs/aaa.tar contains Coding features from AAAA02, AAAB01 and so on.

Individual set files are also made available separately on FTP. e.g. Consider the WGS set WYAA01, that includes the individual WGS records WYAA01000001-WYAA01000116. The WGS sequence set for this is available on FTP at ftp.ebi.ac.uk/pub/databases/ena/wgs/public/wya.

Correspondingly, the coding subproducts from sequences WYAA01000001-WYAA01000116 are available together in ftp.ebi.ac.uk/pub/databases/ena/coding/wgs/public/wya with the name WYAA01.cds.gz

Similarly, the noncoding RNA file is available in ftp.ebi.ac.uk/pub/databases/ena/non-coding/wgs/public/wya with the name WYAA01.ncr.gz

So, if you wanted all coding from WGS, you would need to start at the ftp.ebi.ac.uk/pub/databases/ena/coding/wgs/public level, delve into each subfolder and download the *.cds.gz files.

16.6.1 Find Deleted (suppressed/killed) Records

For Sequence, Coding & Non-coding, to find deleted record IDs since a given date, call the API as follows:

https://www.ebi.ac.uk/ena/browser/api/deleted/sequence/2020-07-01

https://www.ebi.ac.uk/ena/browser/api/deleted/coding/2020-07-01

https://www.ebi.ac.uk/ena/browser/api/deleted/noncoding/2020-07-01

16.6.2 Find Changed Sets

To get a list of Coding or ncRNA set files that have been added/updated since a given date, without having to check through all the subfolders, we provide an API. Call it as follows.

https://www.ebi.ac.uk/ena/browser/api/changed_sets/coding/2020-07-01

and

https://www.ebi.ac.uk/ena/browser/api/changed_sets/noncoding/2020-07-01

16.7 How to identify data based on a last updated timestamp

One common usage of the ENA release was to obtain all assembled/annotated sequence data changed since the last release, either from the entire new release or from the incremental update folders. This can be fully replicated in the ENA Browser API or ENA Browser Advanced Search by using the "last_updated" query filter with a date value.

For the ENA Browser API search endpoint, you can include the 'last_updated' filter and provide a timestamp. This is essentially performing a 'less than or equal to' search, so will provide all records that are new or have been updated from the provided date to the present day). It is recommended that you further customize the query with further filters (for example taxon or geographic) to avoid unnecessarily downloading data you do not require.

Example in FASTA format

```
curl 'https://www.ebi.ac.uk/ena/browser/api/fasta/search?result=sequence&query=last_
↪updated%3E%3D2019-08-18&limit=5' -o fasta.txt
```

or in EMBL format

```
curl 'https://www.ebi.ac.uk/ena/browser/api/embl/search?result=sequence&query=last_
↪updated%3E%3D2019-08-18&limit=5' -o embl.txt
```

You can also provide multiple timestamp filters to give a specific from and to date range, rather than all data to this date, for example data for the first 5 days of August 2019:

```
curl 'https://www.ebi.ac.uk/ena/browser/api/fasta/search?result=sequence&query=last_
↪updated%3E%3D2019-08-01%20AND%20last_updated%3C%3D2019-08-05&limit=5' -o fasta.txt
```

We have added limits to the above examples to only return 5 records. Use limit=0 to retrieve ALL matching records. You can search using the sequence, coding or non-coding data type endpoints. In general when using the API search it is important to be as specific as possible with your query to save on downloading sequences that you do not require.

For the ENA Browser advanced search the 'last_updated' filter can be included in your query. It is located in the Database record filter section.

16.8 Establishing your own release mirroring procedures - Conducting your own release

This section covers the establishment of a mirroring of ENA assembled/annotated sequences without the ENA release. Successful mirroring includes the following concepts:

- Data provenance: Track the accessions obtained in your mirroring, so that the data can be obtained again in future.
- Periodic release: Obtain ENA assembled/annotated sequence data from a defined last updated timestamp.
- Data specificity: By preference use a filtered query to only obtain the data you need, unless you really do need to mirror everything.
- Recapturing the same data in future: Instructions for you or your users to use a summary file that you create to obtain the same dataset in future.

This equates to utilizing two separate ENA API services: - The Data Discovery API to obtain a summary for data provenance - The Browser API to obtain the data most efficiently.

16.8.1 Data provenance

Save the accessions and sequence versions that match your search criteria as a report, which will act as the master document for creating the release. To create such a list, you can query the ENA Portal API with search parameters and save the results to a TSV or JSON file, which you can then use to retrieve the EMBL format or FASTA format records from the ENA Browser API. If you would like to get the current public versions of the records even at a later time, in the query to Portal API, include 'sequence_version' in the fields list. A reason for doing this is to have a fixed list with which you could re-download the same set of records in the future. As records are added,updated or suppressed, the public dataset is regularly changing, and as such you may not get a certain record, or get a different version of a record were you to run the same query in a future date.

e.g.

```
curl 'https://www.ebi.ac.uk/ena/portal/api/search?result=sequence&query=last_updated
↪%3E%3D2019-08-01%20AND%20last_updated%3C%3D2019-08-05&fields=sequence_version,last_
↪updated' -o sequence_report.tsv
```

16.8.2 Periodic release and data specificity

Do the above based on your preferred time period for releases and use the last_updated search parameter.

16.8.3 Instructions for verifying changes since you conducted your release

At a future date, you could rerun the same query and save a new version of the report, which then can be compared with the original master report to look for any differences. We are working on an endpoint that you could upload the original report to and get the list of differences as a response. This is important step as you need to be aware of any sequences that have been killed, as these will not appear in the new data acquisition.

16.8.4 Instructions for obtaining same specific versions of sequences obtained in your release

If the sequence version has been captured in the report, you could retrieve the same specific versions at any time from the Browser API, except for any that may have been killed.

Using the accession and sequence_version fields from this report, you can then retrieve the specific version of the record from Browser API in EMBL or FASTA format. If your list is large, this is obviously not very efficient. So you could run the exact same query against the Browser API's search endpoint to retrieve all the matching records in EMBL or FASTA format at once.

e.g.

```
curl 'https://www.ebi.ac.uk/ena/browser/api/embl/search?result=sequence&query=last_
↪updated%3E%3D2019-08-01%20AND%20last_updated%3C%3D2019-08-05' -o sequences.txt
```

Either of the above, you could parallelize by using the offset and limit parameters to get different chunks of the data simultaneously.

```
curl 'https://www.ebi.ac.uk/ena/browser/api/fasta/search?result=sequence&query=last_
↪updated%3E%3D2019-08-01%20AND%20last_updated%3C%3D2019-08-05&offset=0&limit=100000'␣
↪-o sequences_1.txt

curl 'https://www.ebi.ac.uk/ena/browser/api/fasta/search?result=sequence&query=last_
↪updated%3E%3D2019-08-01%20AND%20last_updated%3C%3D2019-08-05&offset=100000&
↪limit=100000' -o sequences_2.txt
```

(continues on next page)

(continued from previous page)

etc.

Hint: If in the future you want to only retrieve records that have been added or changed since your last pull, it is important that you record the timestamp from when you run the current query and store this so that you can use it for repeating the process for your next update. Obviously you can now pick an update frequency that most suits your use case.

retrieved so far (e.g. using grep), and then use the offset parameter to get the rest from there onwards. If there is a significant delay between the first and the second call, please be aware that the indexed data may have been updated.

16.9 More information resources

Further documentation on the above services is available in their respective documentation: - ENA Discovery Portal API documentation - ENA Browser documentation

16.10 Further assistance

If you currently rely on any aspect of the separate assembled/annotated sequence release process for your work or resource, and cannot switch to one of our continuous distribution processes outlined above, please feel free to contact us to discuss your requirements.

In your query please list what features you utilised from the release process. We can discuss your requirements and determine how we might support your use case through

> one of our existing services or collaborate on an adapted or novel solution. Contacting us promptly with your requirements will allow us to ensure adequate time and resources to collaborate on a solution.

Please contact us with your questions or concerns at https://www.ebi.ac.uk/ena/browser/support with subject 'ENA release retirement'.

Spot an edit or improvement to this page? Please report it using our ENA Support Service quoting the URL of this page in your query.

CHAPTER 17

SARS-CoV-2 (Severe acute respiratory syndrome coronavirus 2) Submissions

Please see below for instructions on how to submit SARS-CoV-2 (Severe acute respiratory syndrome coronavirus 2) or COVID-19 related data. If you have any queries or require assistance with your submission please contact us at: virus-dataflow@ebi.ac.uk.

17.1 Registering Studies

Data submissions to the ENA require that you register a study to contextualise and group your data. Details of how to do this can be found in our Study Registration Guide. Please ensure you describe your study adequately, as well as provide an informative title.

Your ENA SARS-CoV-2 studies can now be claimed using your ORCID ID and/or assigned a DOI. Please see here and here for more information on these options.

17.2 Registering Samples

Having registered a study, please proceed to register your samples. These are metadata objects that describe the source biological material of your experiments. Following this, the sequence data can be registered (as described in later sections).

Instructions for sample registration can be found in our Sample Registration Guide. As part of this process, you must select a sample checklist to describe metadata. If you require any support regarding sample metadata, please contact virus-dataflow@ebi.ac.uk.

17.2.1 Viral Samples

The most appropriate checklist for SARS-CoV-2 viral submissions is the "ENA virus pathogen reporting standard checklist" - ERC000033. This presents 9 mandatory, 15 recommended and 11 optional fields (along with any additional user-defined fields).

Please use the organism name "Severe acute respiratory syndrome coronavirus 2" and taxonomic ID 2697049. It is recommended, as a minimum, that collection date and geographic location (e.g. country) are specified and sample capture status field is provided a value of 'active surveillance in response to outbreak'.

Please see below for a template SARS-CoV-2 viral sample xml for programmatic submission to the ENA:

```
<SAMPLE_SET>
  <SAMPLE alias="Test SARS-CoV-2 sample 1" center_name="EBI">
   <TITLE>Test SARS-CoV-2 Sample 1 Title</TITLE>
   <SAMPLE_NAME>
     <TAXON_ID>2697049</TAXON_ID>
     <SCIENTIFIC_NAME>Severe acute respiratory syndrome coronavirus 2</SCIENTIFIC_
↪NAME>
     <COMMON_NAME>SARS-CoV-2</COMMON_NAME>
   </SAMPLE_NAME>
   <SAMPLE_ATTRIBUTES>
     <SAMPLE_ATTRIBUTE>
       <TAG>geographic location (country and/or sea)</TAG>
       <VALUE>United Kingdom</VALUE>
     </SAMPLE_ATTRIBUTE>
     <SAMPLE_ATTRIBUTE>
       <TAG>collection date</TAG>
       <VALUE>2020-04-26</VALUE>
     </SAMPLE_ATTRIBUTE>
     <SAMPLE_ATTRIBUTE>
       <TAG>host common name</TAG>
       <VALUE>human</VALUE>
     </SAMPLE_ATTRIBUTE>
     <SAMPLE_ATTRIBUTE>
       <TAG>host subject id</TAG>
       <VALUE>1</VALUE>
     </SAMPLE_ATTRIBUTE>
     <SAMPLE_ATTRIBUTE>
       <TAG>host health state</TAG>
       <VALUE>diseased</VALUE>
     </SAMPLE_ATTRIBUTE>
     <SAMPLE_ATTRIBUTE>
       <TAG>host sex</TAG>
       <VALUE>female</VALUE>
     </SAMPLE_ATTRIBUTE>
     <SAMPLE_ATTRIBUTE>
       <TAG>host scientific name</TAG>
       <VALUE>homo sapien</VALUE>
     </SAMPLE_ATTRIBUTE>
     <SAMPLE_ATTRIBUTE>
       <TAG>collector name</TAG>
       <VALUE>Jane Smith</VALUE>
     </SAMPLE_ATTRIBUTE>
     <SAMPLE_ATTRIBUTE>
       <TAG>collecting institution</TAG>
       <VALUE>EMBL-EBI, Wellcome Genome Campus Cambridge CB10 1SD</VALUE>
     </SAMPLE_ATTRIBUTE>
     <SAMPLE_ATTRIBUTE>
       <TAG>isolate</TAG>
       <VALUE>hCoV-19/UK/Bristol/2020</VALUE>
     </SAMPLE_ATTRIBUTE>
     <SAMPLE_ATTRIBUTE>
       <TAG>sample capture status</TAG>
```

(continues on next page)

(continued from previous page)

```
      <VALUE>active surveillance in response to outbreak</VALUE>
    </SAMPLE_ATTRIBUTE>
    <SAMPLE_ATTRIBUTE>
      <TAG>ENA-CHECKLIST</TAG>
      <VALUE>ERC000033</VALUE>
    </SAMPLE_ATTRIBUTE>
  </SAMPLE_ATTRIBUTES>
 </SAMPLE>
</SAMPLE_SET>
```

17.2.2 Metagenomic Samples

Data submissions which include metagenomic samples, should be registered with a relevant metagenome taxonomy - visit our FAQ on Tips for Taxonomy for more information. A few examples include human lung metagenome - 433733, human saliva metagenome - 1679718, human tracheal metagenome - 1712573 or human metagenome - 646099, among many others. Please contact us if you require help with this.

The most appropriate sample checklists depending on the source of your biological samples, are likely to include:

- GSC MIxS host associated (ERC000013)
- GSC MIxS human associated (ERC000014)

Visit our ENA Sample Checklists page for a full listing of sample checklists.

When using the GSC MIxS checklists for your submission, please include the optional field 'host disease status' with a value of 'COVID-19'.

17.2.3 Other Sample Fields

If you have already submitted data to the GISAID database, a corresponding GISAID ID can be specified when using a sample checklist by creating a user-defined field named 'GISAID Accession ID'.

17.3 Submitting Reads

After registering your study and samples, you can submit your read files along with experimental (library-related) metadata. See our Read Submission Guide for detailed instructions on submitting reads.

We encourage submissions to include information on specific protocols used for the experiment. This should be provided in the library description. This can be, for example, the name and/or URL to a specific protocol. View our listing of the available full experimental metadata dictionaries.

Note: Submitted reads to ENA should not contain human identifiable reads. Please filter out human reads prior to submission - if required, here is a tool which can be used.

17.4 Submitting Assemblies

If submitting assemblies, you must have registered a study and a sample beforehand. It is also advised that the associated read data is also submitted. For instructions on assembly submission view our Assembly Submission Guide.

Assemblies can only be submitted using Webin-CLI program or Webin SARS-CoV-2 Genome Submission Web API

17.4.1 Submitting SARS-CoV-2 assembled sequences with Webin-CLI

In case of the Webin-CLI program *-context genome* should be used. During the process, you must define metadata in the manifest file(s). Please specify 'COVID-19 outbreak' as the 'ASSEMBLY_TYPE'.

Each assembly/consensus sequence should also be submitted with a **chromosome list file** (see here), which should be gzipped and referenced in the assembly manifest file, with 'CHROMOSOME_LIST'.

For SARS-CoV-2 submissions, the chromosome list file should contain the following tab-separated columns (with no column header line):

- fasta header
- chromosome number (arbitrary value, set to 1)
- chromosome type (Monopartite for coronaviruses)

e.g:

```
LR991698 1 Monopartite
```

Any assembly annotations, where provided, are captured according to INSDC Feature Table Definitions.

In alignment with INSDC partners, COVID-19 assemblies will **not** be assigned a GCA accession. However, sequence accessions will continue to be assigned, alongside ERZ analysis accessions which are the point of access for the submitted file(s). For more details on accessioning, view our Accessions Guide. To cite data, top-level project accessions (PRJ...) should be used as these are the most stable long-term accessions. View our guide to cite data for further details.

17.5 Submitting Targeted Sequences

If submitting targeted or annotated sequences, you must register a study as described above. See our Targeted Sequence Submission Guide for submission instructions. When submitting annotated sequences, you must select an appropriate Annotation Checklist. There are several virus-specific annotation checklists, with "Single Viral CDS" the most appropriate for complete or partial coding sequences from a viral gene. If your sequences do not fit the annotation checklists above please contact us at virus-dataflow@ebi.ac.uk.

Any annotations, where provided, are captured according to INSDC Feature Table Definitions.

If submitting single contig assemblies, or for any other support or issues around SARS-CoV-2 submissions please contact virus-dataflow@ebi.ac.uk.

17.6 Submitting Linked Human and Viral Datasets

For linked human and viral data submissions please contact virus-dataflow@ebi.ac.uk. Viral data should be submitted to ENA, with the corresponding human data being submitted to the European Genome Phenome Archive (EGA). The viral and human samples registered during each submission will reference each other to support data discovery and interoperability. This will involve the three types of samples below, but only two require user registration:

1. Human sample, registered at EGA.
2. Minimal metadata sample representing the human donor, automatically created at EGA upon registration of 1.

3. Viral sample, registered at ENA.

Samples 1 and 3 can be registered in any order, and Sample 2 will be used to link across human and viral datasets. To assist in this, the relevant biosample accessions should be provided to each archive.

If you have any questions regarding linked datasets please contact virus-dataflow@ebi.ac.uk.

17.7 Release of Data

We recommend that submitted data is set to public as soon as possible to enable early presentation in ENA and also on the COVID-19 data platform. Users are responsible for releasing data they submit to ENA. This is done by setting an appropriate release date, as detailed in our Data Release Policies.

17.8 DOI Issuing

We can now offer DOI issuing for SARS-CoV-2 projects. Digital Object Identifiers (DOIs) are persistent identifiers that can be assigned to any type of entity. From the DOI handbook:

> A DOI name is an identifier (not a location) of an entity on digital networks. It provides a system for persistent and actionable identification and interoperable exchange of managed information on digital networks. A DOI name can be assigned to any entity — physical, digital or abstract — primarily for sharing with an interested user community or managing as intellectual property. The DOI system is designed for interoperability; that is to use, or work with, existing identifier and metadata schemes. DOI names may also be expressed as URLs (URIs).

DOI issuing for ENA records is performed by creating a BioStudies record containing all relevant ENA projects (https://www.ebi.ac.uk/biostudies/about). We will generate this BioStudies record on your behalf and it will hold pointers to the ENA project(s) of your choosing.

To request a DOI for your data, please email virus-dataflow@ebi.ac.uk.

CHAPTER 18

Webin SARS-CoV-2 Genome Submission Web API

18.1 Introduction

Webin SARS-CoV-2 Genome Submission Web API is a JSON based service used to submit SARS-CoV-2 genome assemblies to the European Nucleotide Archive (ENA). For further information on submitting SARS-CoV-2 genomes, see our SARS-CoV-2 Submission Instructions.

There are two submission services:

- Test service
- Production service

The test service is recreated from the full content of the production service every day at 03.00 GMT/BST. Therefore, any submissions made to the test service will be removed by the following day.

When you are using the test service the receipt will contain the following message:

```
"info":["This submission is a TEST submission and will be discarded within 24 hours"]
```

18.2 Service endpoints

This service has two endpoints.

1. Submit SARS-CoV-2 genomes:

- Test service : https://wwwdev.ebi.ac.uk/ena/submit/webin-cli/api/v1/genome/covid-19
- Production service: https://www.ebi.ac.uk/ena/submit/webin-cli/api/v1/genome/covid-19

2. Validate but do NOT submit SARS-CoV-2 genomes:

- Test service: https://wwwdev.ebi.ac.uk/ena/submit/webin-cli/api/v1/genome/covid-19/validate
- Production service: https://www.ebi.ac.uk/ena/submit/webin-cli/api/v1/genome/covid-19/validate

The second endpoint can be used to test if a SARS-CoV-2 genome is valid without submitting it into ENA.

18.3 Submission process

18.3.1 Pre-register Study and Sample

Each submission must be associated with a pre-registered study and a sample.

Please find instruction on how to register studies and samples below: - Register a Study - Register a Sample

18.3.2 Authentication

This service supports basic HTTP authentication only. Please use your Webin submission account name (Webin-N) as the username with your Webin submission account password. New Webin submission accounts can be registered in Webin Portal.

When using curl the username and password are provided using the `-u` option:

```
curl -X 'POST'  -u Webin-N:password  'https://wwwdev.ebi.ac.uk/ena/submit/webin-cli...
```

18.3.3 JSON payload

Both the submission and validation endpoints require a JSON payload in the HTTP body. Below, mandatory fields are marked by * and field descriptions start with #:

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

Example

```
{
  "name": "Test",
  "study": "PRJEB46468",
  "sample": "ERS6670887",
  "coverage": 100,
  "program": " Minimap2",
  "platform": " OXFORD_NANOPORE ",
  "sequence": "CTCTCGATCGATCAAATTTGGGTTTAAGGCCCTTGGAATT",
  "description": "This is a test submission",
  "minGapLength": 1,
  "moleculeType": "genomic DNA",
  "authors": "EMBL-EBI",
  "address": "United Kingdom"
}
```

18.3.4 Submission

Example using curl

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

Example using python

```
import sys
import requests
import json

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

(continues on next page)

(continued from previous page)

```
      "name": "test_2", "study": "PRJEB46811", "sample": "ERS7306049",
      "coverage": 100, "program": "Ilumina", "platform": "Ilumina",
      "sequence": "CTCTCGATCGATCAAATTTGGGTTTAAGGCCCTTGGAATT",
      "description": "test", "minGapLength": 1, "moleculeType": "genomic DNA",
       "authors": "test", "address": "test"
    }
]

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

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

18.3.5 JSON response and HTTP status code

HTTP status code 200 indicates that the submission was successful. More information is available from the JSON response returned in the response body including the assigned accession number and any validation errors.

Please note that an accession will not be assigned when using the `/validate` endpoint.

HTTP status codes

Code	Description
200	OK
400	Bad Request
401	Forbidden
500	Internal Server error

JSON response body example: Successful test service submission

```
{
  "accession": "ERZ2881825",
  "alias": "webin-genome-test_1",
  "info": [
    "This submission is a TEST submission and will be discarded within 24 hours"
  ],
```

(continues on next page)

(continued from previous page)

```
  "error": []
}
```

JSON response body example: Successful production service submission

```
{
  "accession": "ERZ2881825",
  "alias": "webin-genome-test_1",
  "info": [],
  "error": []
}
```

JSON response body example: Successful validation

```
{
  "accession": null,
  "alias": null,
  "info": [],
  "error": []
}
```

JSON response body example: Failed validation

Invalid molecule type:

```
{
  "accession": null,
  "alias": null,
  "info": [],
  "error": [
    "ERROR: Invalid MOLECULETYPE field value: \"reads\". Valid values are: [genomic␣
↪DNA, genomic RNA, viral cRNA]. [manifest file: /tmp/288f4f48-132e-4e90-bb56-
↪5d8afe8af4c45476417061259693052/manifest.json, file name: /tmp/288f4f48-132e-4e90-
↪bb56-5d8afe8af4c45476417061259693052/manifest.json, field: MOLECULETYPE, value:␣
↪reads]"
  ]
}
```

No study and sample found:

```
{
  "accession": null,
  "alias": null,
  "info": [],
  "error": [
    "ERROR: Could not find study \"PRJEB46782\". The study must be owned by the␣
↪submission account used for this submission or it must be private or temporarily␣
↪suppressed and referenced by accession. Note that only a single study can be␣
↪referenced. Unknown study PRJEB46782 or the study cannot be referenced by your␣
↪submission account. Studies must be submitted before they can be referenced in the␣
↪submission. [manifest file: /tmp/612ca908-39af-4b72-b9a4-
↪f759bac7f144213552988004474 2773/manifest.json, file name: /tmp/612ca908-39af-4b72-
↪b9a4-f759bac7f144213552988004474 2773/manifest.json, field: STUDY, value: PRJEB46782
↪"
```

(continues on next page)

(continued from previous page)

```
    ]
}
```