EMBL-EBI

# ENA Browser API

Programmatic access to records held within the European Nucleotide Archive

EMBL-EBI
5-3-2020

## Change History

| Document Version | Date | Author | Notes |
|---|---|---|---|
| 1.0 | 2020-03-05 | Josephine Burgin | Initial version |
| 1.1 | 2020-11-09 | Vishnukumar Balavenkataraman Kadhirvelu | TLS set entries and result mapping to the ebisearch domain |

# Contents

# Introduction to ENA Browser API

The main function of the ENA browser API is to fetch and/or download public records from the European Nucleotide Archive. There is no need for authentication headers and all endpoints within the API use the HTTP GET method to request data.

The browser API is available from https://www.ebi.ac.uk/ena/browser/api. If you use this URL within a web browser, you will see some documentation regarding the different functions available, as well as forms for the different endpoints which allow you to send requests.

Access to the browser API will likely be through either inclusion within scripts or using a tool such as wget and curl.

For example, to retrieve an XML record using curl:

```
curl -X GET --header 'Accept: application/json'
'https://www.ebi.ac.uk/ena/browser/api/xml/SAMEA2591108'
```

## Endpoints

The Browser API endpoints are constructed from two parts:

- the record format to retrieve
- the function for retrieval

Most endpoints provide a direct to download function but others retrieve the data as a stream. In some cases, endpoints for data retrieval can also be downloaded with the parameter "download=true". More details can be found in the guidelines for each of the functions.

### ENA Browser API endpoints

| Record Format | Function | Purpose |
|---|---|---|
| **/embl** | /{accession} | Retrieve **embl** records by a single accession, range or a comma separated list (without spaces) |
| | /search | Direct download all **embl** records resulting from an advanced search |
| **/fasta** | /{accession} | Retrieve **fasta** format records by a single accession, range or a comma separated list (without spaces) |
| | /links/{linksResult} | Direct download all **fasta** records relating to a particular study/sample/taxon record |
| | /search | Direct download all **fasta** records resulting from an advanced search |
| | /textsearch | Retrieve all **fasta** records resulting from a key-word search (backed by EBI search) |
| | /textsearch/count | Count of all **fasta** records resulting from a key-word search (backed by EBI search) |
| **/text** | /{accession} | Retrieve **embl** records by a single accession, range or a comma separated list (without spaces) |
| | /links/{linksResult} | Direct download all **embl** records relating to a particular study/sample/taxon record |

| | | |
|---|---|---|
| | /search | Direct download all **embl** records resulting from an advanced search |
| | /textsearch | Retrieve all **embl** records resulting from a key-word search (backed by EBI search) |
| | /textsearch/count | Count of all **embl** records resulting from a key-word search (backed by EBI search) |
| /xml | /{accession} | Retrieve **xml** records by a single accession, range or a comma separated list (without spaces) |
| | /links/{linksResult} | Direct download all **xml** records relating to a particular study/sample/taxon record |
| | /search | Direct download all **xml** records resulting from an advanced search |
| | /textsearch | Retrieve all **xml** records resulting from a key-word search (backed by EBI search) |
| | /textsearch/count | Count of all **xml** records resulting from a key-word search (backed by EBI search) |
| /versions | /{accession} | Retrieve the **version** history of a record by its accession |

EMBL and TEXT record formats are equivalent – they both return embl flat file records. The text endpoints have been retained for users who are familiar with the old ENA browser where embl records were referred to as in TEXT format.

The doc endpoint will allow you to download the latest version of this API documentation in PDF format:

https://www.ebi.ac.uk/ena/browser/api/doc

All other endpoints are described in more detail the following chapters.

# Retrieving by accession

Records can be retrieved by accession using the `/{accession}` endpoint. This section does not cover the `/versions/{accession}` function which is described separately in a later section.

Not all ENA records can be downloaded in all formats. For example, raw read records and metadata objects (e.g. Study, Sample, Run, Taxon) can only be retrieved in XML format but other sequence-based objects can be retrieved in EMBL (or TEXT) or FASTA formats.

This table summarises which types of records can be retrieved for each record format and the valid accession format to retrieve these.

| Record Format | ENA record types that can be retrieved | Valid accession format for retrieval |
|---|---|---|
| /embl | Contig set | [A-Z]{4}[0-9]{2}<br>[A-Z]{6}[0-9]{2} |
| | Sequence | [A-Z]{1}[0-9]{5}.[0-9]+<br>[A-Z]{2}[0-9]{6}.[0-9]+<br>[A-Z]{2}[0-9]{8} |
| /fasta | Sequence | [A-Z]{1}[0-9]{5}.[0-9]+<br>[A-Z]{2}[0-9]{6}.[0-9]+<br>[A-Z]{2}[0-9]{8} |
| /text | Contig set | [A-Z]{4}[0-9]{2}<br>[A-Z]{6}[0-9]{2} |
| | Sequence | [A-Z]{1}[0-9]{5}.[0-9]+<br>[A-Z]{2}[0-9]{6}.[0-9]+<br>[A-Z]{2}[0-9]{8} |
| /xml | Study | (E\|D\|S)RP[0-9]{6,}<br>PRJ(E\|D\|N)[A-Z][0-9]+ |
| | Sample | (E\|D\|S)RS[0-9]{6,}<br>SAM(E\|D\|N)[A-Z]?[0-9]+ |
| | Run | (E\|D\|S)RR[0-9]{6,} |
| | Experiment | (E\|D\|S)RX[0-9]{6,} |
| | Analysis | (E\|D\|S)RZ[0-9]{6,} |
| | Assembly | GCA_[0-9]{9}.[0-9]+ |
| | Submission | (E\|D\|S)RA[0-9]{6,} |
| | Taxon | [0-9]+ |

Accessions for retrieval can be provided as a single accession (ERS123456), a range (ERS123456-ERS123459) or a comma separated list (ERS123456,ERS123458,ERS123460). Accessions in a range or list must be of the same record type for each API call.

### *Study records*

Study records represent a research project. The Study holds information about the motivation of the project as well as links to publications.

### *Sample records*

Sample records represent biological samples that were involved in a research project. They hold metadata regarding sample collection and processing (and sometimes analysis in special cases).

*Run records*

Run records hold the location of the sequencing data files.

*Experiment records*

Experiment records hold metadata of the sequencing event.

*Analysis records*

Analysis records hold metadata and the location of data files resulting from any subsequent analysis of sequencing data.

*Assembly records*

Assembly records are a metadata object that holds an overview of a genome assembly. They hold all the components of the genome assembly (e.g. the contig set, scaffold sequences or full chromosomes). They also hold generated assembly statistics.

*Sequence records*

Sequence records hold targeted assembled nucleotide sequences and sometimes annotation. This includes coding, non-coding and marker regions. These records do not include whole genome shotgun (WGS), transcriptome assembly (TSA) sequences or targeted locus study (TLS). These are instead separated into the contig set results described below.

*Contig set records*

Whole genome shotgun (WGS), transcriptome assembly (TSA) and targeted locus study (TLS) sequences are represented as sets grouped by a common set prefix, optionally with a set of annotations that describes all sequences in the set. Contig sets are not retrieved by individual contigs but as the whole set. Only the annotation of contig sets can be retrieved using the browser API due to the scale of the sequence sets.

*Submission records*

Submission records are records of a particular submission event. They hold all the accessions resulting from a single submission event as well as the date and submitting centre.

## Parameters of {accession} function

| Parameter | Valid Format | Description |
|---|---|---|
| **download** | boolean | Download the result as a file (false by default) |
| **lineLimit** | number | Limit the number of text lines returned (returns full text by default lineLimit=0 also returns all text) |
| **annotationOnly** | boolean | Only retrieve annotation, no sequence (false by default) |
| **expanded** | boolean | Get expanded records for CON sequences (false by default) |
| **gzip** | boolean | Download the result as a gzip file (false by default) |

# Performing an advanced search

An advanced search is performed via the `/search` endpoint. This returns a direct download of the records resulting from the search in the chosen record format.

## Parameters of search function

| Parameter | Valid Format | Description |
|---|---|---|
| **result** | string | The result type (data set) to search against |
| **query** | string | A set of search conditions joined by logical operators (AND, OR, NOT) and bound by double quotes. If none supplied, the full result set will be returned. |
| **includeAccessions** | string | A list of accessions that you would like to be included with the results of your query regardless of whether they match the provided query. |
| **excludeAccessions** | string | A list of accessions that you would like to be excluded from the results of your query regardless of whether they match the provided query. |
| **offset** | integer | How many records to skip in the search. Default value is 0. |
| **limit** | integer | The maximum number of records to retrieve. Default value is 100,000. If the full result set is to be fetched, the limit should be set to 0, use responsibly and only when required. It is better practice to make multiple size-limited calls using the offset feature than a single large call. |

## Result

To perform a search across ENA, you must first choose the correct datatype *result* to search across. Not all record formats are compatible with all *result*s.

This table summarises which *result* data types can be used for each record format and the record type that will be returned.

| Record Format | Data type 'result' that can be searched against | Resulting ENA record type |
|---|---|---|
| **/embl** | sequence<br>coding<br>noncoding | Sequence |
| | wgs_set<br>tsa_set<br>tls_set | Contig set |
| **/fasta** | sequence<br>coding<br>noncoding | Sequence |
| **/text** | sequence | Sequence |

| | | |
|---|---|---|
| | coding<br>noncoding | |
| | wgs_set<br>tsa_set<br>tls_set | Contig set |
| /xml | study<br>read_study<br>analysis_study | Study |
| | sample | Sample |
| | read_run | Run |
| | read_experiment | Experiment |
| | analysis | Analysis |
| | assembly | Assembly |
| | taxon | Taxon |

### *Study records*

Study records represent a research project. The Study holds information about the motivation of the project as well as links to publications. All studies can be searched across within the study result. To only search within studies containing raw reads, the read_study result can be used. read_study searches can be filtered by additional read-specific metadata. For studies containing analyses, the analysis_study result can be used. analysis_study searches can be filtered by additional analysis-specific metadata. These are summarised in the tables in the 'Searchable fields' section later in this section.

### *Sample records*

Sample records represent biological samples that were involved in a research project. They hold metadata regarding sample collection and processing (and sometimes analysis in special cases). These can be searched across within the sample result and filtered using sample-related metadata summarised in the tables in the 'Searchable fields' section later in this section.

### *Run records*

Run records hold the location of the sequencing data files. These can be searched across within the read_run result. The read_run result allows for searching using associated sample and read metadata as well. These are summarised in the tables in the 'Searchable fields' section later in this section.

### *Experiment records*

Experiment records hold metadata of the sequencing event. These can be searched across within the read_run result. The read_run result allows for searching using associated sample and read metadata. These are summarised in the tables in the 'Searchable fields' section later in this section.

### *Analysis records*

Analysis records hold metadata and the location of data files resulting from any subsequent analysis of sequencing data. These can be searched across within the analysis result. The analysis results allows for searching using associated sample and analysis metadata. These are summarised in the tables in the 'Searchable fields' section later in this section.

### *Assembly records*

Assembly records are a metadata object that holds an overview of a genome assembly. They hold all the components of the genome assembly (e.g. the contig set, scaffold sequences or full

chromosomes). They also hold generated assembly statistics. These can be searched across within the assembly result.

### *Sequence records*

Sequence records hold targeted assembled nucleotide sequences and sometimes annotation. This includes coding, non-coding and marker regions. These records do not include whole genome shotgun (WGS), transcriptome assembly (TSA) sequences or targeted locus study (TLS). These are instead separated into the contig set results described below. Sequence records can be searched across within the sequence result or more specifically the noncoding result for non-coding sequences.

### *Contig set records*

Whole genome shotgun (WGS), transcriptome assembly (TSA) and targeted locus study (TLS) sequences are represented as sets grouped by a common set prefix, optionally with a set of annotations that describes all sequences in the set. Contig sets are not retrieved by individual contigs but as the whole set. Only the annotation of contig sets can be retrieved using the browser API due to the scale of the sequence sets. Contig sets can be searched across within the wgs_set, tsa_set or tls_set results.

## Building a Query

When no query is defined, all records from the selected result will be displayed. In most cases, however, a subset of those records are required. To define this subset, there are a number of filter fields available. The query can be built from any number of fields, with logical operators and parentheses used to order the execution of each. Any text or controlled vocabulary values used within the query must be bound by double quotes. For example:

```
query=booleanField=true AND (stringField="value" OR cvField="CV1")
```

## Standard filter types

The majority of searchable fields use a standard data type. The table below lists all available operator for each of these filter types.

| Filter type | Operators |
|---|---|
| boolean | = |
| controlled vocabulary | =, != |
| date | =, !=, <, <=, >, >= |
| number | =, !=, <, <=, >, >= |
| text | =, != |

Text searches are case insensitive and a wildcard character (*) can be used at the beginning or end of a string value for partial matching.

### *Function filter types*

In addition to the standard filter types listed above, there are two query filters that are based on functions: geospatial and taxonomic.

### Geospatial

All geospatial coordinates are represented in decimal degrees.

| Function | Description | Parameters | Example |
|---|---|---|---|
| **geo_box1** | All locations within a box defined by the lower left (SW) and upper right (NE) points | SW latitude, SW longitude, NE latitude, NE longitude | geo_box1(-20, 10, 20, 50) |
| **geo_box2** | All locations within a box defined a centre point and a radius in km | latitude, longitude, radius (km) | geo_box2(35, 100, 300) |
| **geo_circ** | All locations within a circle defined by a centre point and a radius in km | latitude, longitude, radius (km) | geo_circ(35, 100, 300) |
| **geo_lat** | All locations within a latitude range given by a latitude and a radius in km | latitude, radius (km) | geo_lat(0, 100) |
| **geo_north** | All locations north of a given latitude (inclusive) | latitude | geo_north(80) |
| **geo_south** | All locations south of a given latitude (inclusive) | latitude | geo_south(-80) |
| **geo_point** | An exact latitude/longitude position | latitude, longitude | geo_point(9.12, -79.7) |

### Taxonomy

Three functions are available for performing taxonomic searches. These make it possible to filter on a single taxon (via NCBI taxon ID or scientific name) or a branch of the NCBI taxonomic tree.

| Function | Description | Parameters | Example |
|---|---|---|---|
| **tax_eq** | All records that match the given NCBI taxonomy identifier | NCBI taxon ID | tax_eq(9606) |
| **tax_tree** | All records that match the given NCBI taxonomy identifier or are descendants of it | NCBI taxon ID | tax_tree(2759) |
| **tax_name** | All records that match the given NCBI scientific name | NCBI scientific name | tax_name("Homo%20sapiens") |

## Searchable fields

Searchable fields are dependent on type:

### *Sample fields*

In addition to searching across sample results, Sample fields can also be used for refining data within the read and analysis results.

| Searchable field | Description | Filter type |
|---|---|---|
| accession | accession number | text |
| altitude | Altitude (m) | number |
| assembly_quality | Quality of assembly | text |
| assembly_software | Assembly software | text |
| binning_software | Binning software | text |
| bio_material | identifier for biological material including institute and collection code | text |
| broker_name | broker name | controlled vocab. |
| cell_line | cell line from which the sample was obtained | text |
| cell_type | cell type from which the sample was obtained | text |
| center_name | Submitting center | text |
| checklist | checklist name (or ID) | controlled vocab. |
| collected_by | name of the person who collected the specimen | text |
| collection_date | date that the specimen was collected | date |
| completeness_score | Completeness score (%) | number |
| contamination_score | Contamination score (%) | number |
| country | locality of sample isolation: country names, oceans or seas, followed by regions and localities | text |
| cultivar | cultivar (cultivated variety) of plant from which sample was obtained | text |
| culture_collection | identifier for the sample culture including institute and collection code | text |
| depth | Depth (m) | number |
| description | brief sequence description | text |
| dev_stage | sample obtained from an organism in a specific developmental stage | text |
| ecotype | a population within a given species displaying traits that reflect adaptation to a local habitat | text |
| elevation | Elevation (m) | number |
| environment_biome | Environment (Biome) | text |
| environment_feature | Environment (Feature) | text |
| environment_material | Environment (Material) | text |

| | | |
|---|---|---|
| environmental_package | MIGS/MIMS/MIMARKS extension for reporting (from environment where the sample was obtained) | controlled vocab. |
| environmental_sample | identifies sequences derived by direct molecular isolation from an environmental DNA sample | boolean |
| experimental_factor | variable aspects of the experimental design | text |
| first_public | date when made public | date |
| germline | the sample is an unrearranged molecule that was inherited from the parental germline | boolean |
| host | natural (as opposed to laboratory) host to the organism from which sample was obtained | text |
| host_body_site | name of body site from where the sample was obtained | text |
| host_genotype | genotype of host | text |
| host_growth_conditions | literature reference giving growth conditions of the host | text |
| host_phenotype | phenotype of host | text |
| host_sex | physical sex of the host | controlled vocab. |
| host_status | condition of host (e.g. diseased or healthy) | text |
| host_tax_id | NCBI taxon id of the host | number |
| identified_by | name of the taxonomist who identified the specimen | text |
| investigation_type | the study type targeted by the sequencing | controlled vocab. |
| isolate | individual isolate from which sample was obtained | text |
| isolation_source | describes the physical, environmental and/or local geographical source of the sample | text |
| last_updated | date when last updated | date |
| location | geographic location of isolation of the sample | geospatial |
| mating_type | mating type of the organism from which the sequence was obtained | text |
| ph | pH | number |
| project_name | name of the project within which the sequencing was organized | text |
| protocol_label | the protocol used to produce the sample | text |
| salinity | Salinity (PSU) | number |
| sample_accession | sample accession number | text |
| sample_alias | submitter's name for the sample | text |
| sample_collection | the method or device employed for collecting the sample | text |

| | | |
|---|---|---|
| sample_material | sample material label | text |
| sample_title | brief sample title | text |
| sampling_campaign | the activity within which this sample was collected | text |
| sampling_platform | the large infrastructure from which this sample was collected | text |
| sampling_site | the site/station where this sample was collection | text |
| secondary_sample_accession | secondary sample accession number | text |
| sequencing_method | sequencing method used | text |
| serotype | serological variety of a species characterized by its antigenic properties | text |
| serovar | serological variety of a species (usually a prokaryote) characterized by its antigenic properties | text |
| sex | sex of the organism from which the sample was obtained | controlled vocab. |
| specimen_voucher | identifier for the sample culture including institute and collection code | text |
| strain | strain from which sample was obtained | text |
| sub_species | name of sub-species of organism from which sample was obtained | text |
| sub_strain | name or identifier of a genetically or otherwise modified strain from which sample was obtained | text |
| target_gene | targeted gene or locus name for marker gene studies | text |
| taxonomic_classification | Taxonomic classification | text |
| taxonomic_identity_marker | Taxonomic identity marker | text |
| taxonomy | NCBI taxonomic classification | taxonomy |
| temperature | Temperature (C) | number |
| tissue_lib | tissue library from which sample was obtained | text |
| tissue_type | tissue type from which the sample was obtained | text |
| variety | variety (varietas, a formal Linnaean rank) of organism from which sample was derived | text |

*Read fields*

The following fields can be used to search against any of the read results: read_run, read_experiment and read_study. In addition to the fields in the table below, any sample fields can also be used in the search query.

| Searchable field | Description | Filter type |
|---|---|---|
| study_accession | study accession number | text |
| secondary_study_accession | secondary study accession number | text |
| experiment_accession | experiment accession number | text |
| run_accession | run accession number | text |
| sample_accession | sample accession number | text |
| secondary_sample_accession | secondary sample accession number | text |
| study_alias | submitter's name for the study | text |
| experiment_alias | submitter's name for the experiment | text |
| run_alias | submitter's name for the run | text |
| sample_alias | submitter's name for the sample | text |
| study_title | brief sequencing study description | text |
| experiment_title | brief experiment title | text |
| first_public | date when made public | date |
| last_updated | date when last updated | date |
| broker_name | broker name | controlled vocab. |
| center_name | Submitting center | text |
| instrument_model | instrument model used in sequencing experiment | controlled vocab. |
| instrument_platform | instrument platform used in sequencing experiment | controlled vocab. |
| library_layout | sequencing library layout | controlled vocab. |
| library_name | sequencing library name | text |
| library_selection | method used to select or enrich the material being sequenced | controlled vocab. |
| library_source | source material being sequenced | controlled vocab. |
| library_strategy | sequencing technique intended for the library | controlled vocab. |
| nominal_length | average fragmentation size of paired reads | number |
| nominal _sdev | standard deviation of fragmentation size of paired reads | number |
| base_count | number of base pairs | number |
| read_count | number of reads | number |
| submitted_format | format of submitted reads | text |
| parent_study | parent study accession number | text |
| taxonomy | NCBI taxonomic classification | taxonomy |

*Analysis fields*

The following fields can be used to search against any of the analysis results: analysis and analysis_study. In addition to the fields in the table below, any sample fields can also be used in the search query

| Searchable field | Description | Filter type |
|---|---|---|
| analysis_accession | analysis accession number | text |
| study_accession | study accession number | text |
| secondary_study_accession | secondary study accession number | text |
| sample_accession | sample accession number | text |
| secondary_sample_accession | secondary sample accession number | text |
| analysis_alias | submitter's name for the analysis | text |
| study_alias | submitter's name for the study | text |
| sample_alias | submitter's name for the sample | text |
| analysis_title | brief sequence analysis description | text |
| study_title | brief sequencing study description | text |
| broker_name | broker name | controlled vocab. |
| center_name | Submitting center | text |
| analysis_type | type of sequence analysis | controlled vocab. |
| assembly_type | analysis Assembly type | controlled vocab. |
| first_public | date when made public | date |
| last_updated | date when last updated | date |
| parent_study | parent study accession number | text |
| taxonomy | NCBI taxonomic classification | taxonomy |

*Assembly fields*

As the assembly result represents the latest public version of the assembly, the majority of fields that can be searched are specific to the latest version. The assembly_name field however contains all assembly version names, therefore a search for an assembly name may not return the version that matches the name.

| Searchable Field | Description | Filter type |
|---|---|---|
| accession | accession number | text |
| study_accession | study accession number | text |
| sample_accession | sample accession number | text |
| secondary_sample_accession | secondary sample accession number | text |
| assembly_name | genome assembly name for all live versions | text |
| assembly_title | brief genome assembly description | text |
| study_name | sequencing study name | text |
| study_title | brief sequencing study description | text |
| study_description | detailed sequencing study description | text |
| assembly_level | assembly level | controlled vocab. |

| | | |
|---|---|---|
| genome_representation | whether this is a full or partial genome | controlled vocab. |
| strain | strain from which sample was obtained | text |
| assembly_type | analysis Assembly type | controlled vocab. |
| taxonomy | NCBI taxonomic classification | taxonomy |
| last_updated | date when last updated | date |

*Sequence fields*

The fields in the table below are searchable for sequence results. Most of the source feature qualifiers are available to search against.

| Searchable Field | Description | Filter Type |
|---|---|---|
| accession | accession number | text |
| altitude | Altitude (m) | number |
| base_count | number of base pairs | number |
| bio_material | identifier for biological material including institute and collection code | text |
| cell_line | cell line from which the sample was obtained | text |
| cell_type | cell type from which the sample was obtained | text |
| collected_by | name of the person who collected the specimen | text |
| collection_date | date that the specimen was collected | date |
| country | locality of sample isolation: country names, oceans or seas, followed by regions and localities | text |
| cultivar | cultivar (cultivated variety) of plant from which sample was obtained | text |
| culture_collection | identifier for the sample culture including institute and collection code | text |
| dataclass | sequence data class | controlled vocab. |
| description | brief sequence description | text |
| dev_stage | sample obtained from an organism in a specific developmental stage | text |
| ecotype | a population within a given species displaying traits that reflect adaptation to a local habitat | text |
| environmental_sample | identifies sequences derived by direct molecular isolation from an environmental DNA sample | boolean |
| first_public | date when made public | date |
| germline | the sample is an unrearranged molecule that was inherited from the parental germline | text |
| haplotype | combination of alleles that are linked together on the same physical chromosome | text |

| | | |
|---|---|---|
| host | natural (as opposed to laboratory) host to the organism from which sample was obtained | text |
| identified_by | name of the taxonomist who identified the specimen | text |
| isolate | individual isolate from which sample was obtained | text |
| isolation_source | describes the physical, environmental and/or local geographical source of the sample | text |
| keywords | keywords associated with sequence | text |
| lab_host | scientific name of the laboratory host used to propagate the source organism for the sample | text |
| last_updated | date when last updated | date |
| location | geographic location of isolation of the sample | geospatial |
| mating_type | mating type of the organism from which the sequence was obtained | text |
| mol_type | in vivo molecule type of the sequence | controlled vocab. |
| organelle | membrane-bound intracellular structure from which the sequence was obtained | controlled vocab. |
| plasmid | name of naturally occurring plasmid from which the sequence was obtained | text |
| sample_accession | sample accession number | text |
| secondary_sample_accession | secondary sample accession number | text |
| sequence_md5 | MD5 checksum of sequence | text |
| serotype | serological variety of a species characterized by its antigenic properties | text |
| serovar | serological variety of a species (usually a prokaryote) characterized by its antigenic properties | text |
| sex | sex of the organism from which the sample was obtained | controlled vocab. |
| specimen_voucher | identifier for the sample culture including institute and collection code | text |
| strain | strain from which sample was obtained | text |
| study_accession | study accession number | text |
| sub_species | name of sub-species of organism from which sample was obtained | text |
| sub_strain | name or identifier of a genetically or otherwise modified strain from which sample was obtained | text |
| tax_division | taxonomic division | controlled vocab. |
| taxonomy | NCBI taxonomic classification | taxonomy |

| | | |
|---|---|---|
| tissue_lib | tissue library from which sample was obtained | text |
| tissue_type | tissue type from which the sample was obtained | text |
| topology | sequence topology: circular or linear | controlled vocab. |
| variety | variety (varietas, a formal Linnaean rank) of organism from which sample was derived | text |

*Contig set fields*

The fields in the table below are searchable for both the wgs_set, tsa_set and tls_set results. The information common to the whole WGS set (therefore contained within the set's master record) is available for each record. Any individual sequence-level source feature information that differs from the master record cannot be searched for.

| Searchable Field | Description | Filter Type |
|---|---|---|
| accession | accession number | text |
| altitude | Altitude (m) | number |
| base_count | number of base pairs | number |
| bio_material | identifier for biological material including institute and collection code | text |
| cell_line | cell line from which the sample was obtained | text |
| cell_type | cell type from which the sample was obtained | text |
| collected_by | name of the person who collected the specimen | text |
| collection_date | date that the specimen was collected | date |
| country | locality of sample isolation: country names, oceans or seas, followed by regions and localities | text |
| cultivar | cultivar (cultivated variety) of plant from which sample was obtained | text |
| culture_collection | identifier for the sample culture including institute and collection code | text |
| description | brief sequence description | text |
| dev_stage | sample obtained from an organism in a specific developmental stage | text |
| ecotype | a population within a given species displaying traits that reflect adaptation to a local habitat | text |
| environmental_sample | identifies sequences derived by direct molecular isolation from an environmental DNA sample | boolean |
| first_public | date when made public | date |
| germline | the sample is an unrearranged molecule that was inherited from the parental germline | text |

| | | |
|---|---|---|
| haplotype | combination of alleles that are linked together on the same physical chromosome | text |
| host | natural (as opposed to laboratory) host to the organism from which sample was obtained | text |
| identified_by | name of the taxonomist who identified the specimen | text |
| isolate | individual isolate from which sample was obtained | text |
| isolation_source | describes the physical, environmental and/or local geographical source of the sample | text |
| keywords | keywords associated with sequence | text |
| lab_host | scientific name of the laboratory host used to propagate the source organism for the sample | text |
| last_updated | date when last updated | date |
| location | geographic location of isolation of the sample | geospatial |
| mating_type | mating type of the organism from which the sequence was obtained | text |
| mol_type | in vivo molecule type of the sequence | controlled vocab. |
| sample_accession | sample accession number | text |
| secondary_sample_accession | secondary sample accession number | text |
| serotype | serological variety of a species characterized by its antigenic properties | text |
| serovar | serological variety of a species (usually a prokaryote) characterized by its antigenic properties | text |
| sex | sex of the organism from which the sample was obtained | controlled vocab. |
| specimen_voucher | identifier for the sample culture including institute and collection code | text |
| strain | strain from which sample was obtained | text |
| study_accession | study accession number | text |
| sub_species | name of sub-species of organism from which sample was obtained | text |
| sub_strain | name or identifier of a genetically or otherwise modified strain from which sample was obtained | text |
| tax_division | taxonomic division | controlled vocab. |
| taxonomy | NCBI taxonomic classification | taxonomy |
| tissue_lib | tissue library from which sample was obtained | text |
| tissue_type | tissue type from which the sample was obtained | text |

| | | |
|---|---|---|
| variety | variety (varietas, a formal Linnaean rank) of organism from which sample was derived | text |

*Noncoding fields*

The fields in the table below are searchable for noncoding results. As for searches against sequences, most of the source feature qualifiers are available to search against. Selected RNA-feature specific information has also been included.

| Searchable Field | Description | Filter Type |
|---|---|---|
| accession | accession number | text |
| anticodon | location of the anticodon of tRNA and the amino acid for which it codes | text |
| base_count | number of base pairs | number |
| collected_by | name of the person who collected the specimen | text |
| collection_date | date that the specimen was collected | date |
| country | locality of sample isolation: country names, oceans or seas, followed by regions and localities | text |
| dataclass | sequence data class | controlled vocab. |
| description | brief sequence description | text |
| dev_stage | sample obtained from an organism in a specific developmental stage | text |
| environmental_sample | identifies sequences derived by direct molecular isolation from an environmental DNA sample | boolean |
| experiment | a brief description of the nature of the experimental evidence | text |
| first_public | date when made public | date |
| function | function attributed to a sequence | text |
| gene | symbol of the gene corresponding to a sequence region | text |
| gene_synonym | synonymous, replaced, obsolete or former gene symbol | text |
| inference | a structured description of non-experimental evidence | text |
| keyword | keywords associated with sequence | text |
| lab_host | scientific name of the laboratory host used to propagate the source organism for the sample | text |
| last_updated | date when last updated | date |
| location | geographic location of isolation of the sample | geospatial |
| locus_tag | a submitter-supplied, systematic, stable identifier for a gene and its associated features | text |
| marker | marker classification | text |

| | | |
|---|---|---|
| mol_type | in vivo molecule type of the sequence | controlled vocab. |
| organelle | membrane-bound intracellular structure from which the sequence was obtained | controlled vocab. |
| product | name of the product associated with the feature | text |
| rna_class | classification of RNA | controlled vocab. |
| sample_accession | sample accession number | text |
| secondary_sample_accession | secondary sample accession number | text |
| sequence_md5 | MD5 checksum of sequence | text |
| strain | strain from which sample was obtained | text |
| study_accession | study accession number | text |
| tax_division | taxonomic division | controlled vocab. |
| taxonomy | NCBI taxonomic classification | taxonomy |

### Including and excluding records

To include or exclude specific records as part of the return set, you can include/exclude these within the search by specifying the type of record to include/exclude then provide accessions. Multiple accessions can be specified using commas without spaces.

Format:

`includeAccessionType=study&includeAccessions=PRJEB25206`

`excludeAccessionType=sample&excludeAccessions=SAMEA4051446,SAMEA4051447`

### Size limit and pagination of result

A search could return a few or millions of results. As a safety feature to protect against unintended download of large files (or time-outs in scripts that process directly from the URL stream), the default page size is set to 100,000 records. This can be changed using the limit parameter. The limit can be raised for larger page sizes, alternatively if you wish to retrieve all results, the limit should be set to 0, however, it is better practice to make multiple size-limited calls.

The offset parameter is available to use with the limit for pagination and should be set to how many records should be skipped. For example, to fetch the first two pages of a search result with a page size of 200,000 records:

Page one: `limit=200000`

Page two: `offset=200000&limit=200000`

# Performing a text search

A text based search is performed via the `/textsearch` endpoints of each record type. This returns a stream of records resulting from the search in the chosen record format.

This is a simpler search interface than the advanced search functionality based on free text searching across different domains of data within the ENA. It is backed by EBI Search which is a text search engine used across multiple services hosted at the European Bioinformatics Institute (EMBL-EBI).

## Parameters of textsearch function

| Parameter | Valid Format | Description |
|---|---|---|
| **domain** | string | The domain (data set) to search against. |
| **query** | string | Keyword(s) to search against. |
| **offset** | integer | How many records to skip in the search. Default value is 0. |
| **limit** | integer | The maximum number of records to retrieve. Default value is 100,000. If the full result set is to be fetched, the limit should be set to 0, use responsibly and only when required. It is better practice to make multiple size limited calls using the offset feature than a single large call. |
| **gzip** | boolean | Download the result as a gzip file (false by default) |

### Domain

To perform a text based search across ENA, you must first choose the correct datatype *domain* to search across. Not all record formats are compatible with all *domains*.

This table summarises which *domain* data types can be used for each record format and the record type that will be returned.

| Record Format | Data type 'domain' that can be searched against | Resulting ENA record | Resulting ENA record type |
|---|---|---|---|
| **/fasta** | embl<br>coding<br>non-coding | Sequence | sequence<br>coding<br>noncoding |
| **/text** | embl<br>coding<br>non-coding | Sequence | sequence<br>coding<br>noncoding |
| | wgs_masters<br>tsa_masters<br>tls_masters | Contig set | wgs_set<br>tsa_set<br>tls_set |
| **/xml** | project<br>sra-study | Study | study<br>read_study<br>analysis_study |
| | sra-sample | Sample | sample |
| | sra-run | Run | read_run |

| | | | |
|---|---|---|---|
| | sra-experiment | Experiment | read_experiment |
| | sra-analysis | Analysis | analysis |
| | genome_assembly | Assembly | assembly |
| | sra-submission | Submission | |

*Study records*

Study records represent a research project. The Study holds information about the motivation of the project as well as links to publications. The project domain covers all projects held within the ENA. The sra-study domain includes only those containing raw read data or analyses.

*Sample records*

Sample records represent biological samples that were involved in a research project. They hold metadata regarding sample collection and processing (and sometimes analysis in special cases). These can be searched across the sample domain.

*Run records*

Run records hold the location of the sequencing data files. These can be searched across the sra-run domain.

*Experiment records*

Experiment records hold metadata of the sequencing event. These can be searched across the sra-experiment domain.

*Analysis records*

Analysis records hold metadata and the location of data files resulting from any subsequent analysis of sequencing data. These can be searched across the sra-analysis domain.

*Assembly records*

Assembly records are a metadata object that holds an overview of a genome assembly. They hold all the components of the genome assembly (e.g. the contig set, scaffold sequences or full chromosomes). They also hold generated assembly statistics. These can be searched across the genome_assembly domain.

*Sequence records*

Sequence records hold targeted assembled nucleotide sequences and sometimes annotation. This includes coding, non-coding and marker regions. These records do not include whole genome shotgun (WGS), transcriptome assembly (TSA) or targeted locus study (TLS) sequences. These are instead separated into the contig set results described below. Sequence records can be retrieved using the general embl domain or specifically for coding or non-coding regions using the coding and non-coding domains.

*Contig set records*

Whole genome shotgun (WGS), transcriptome assembly (TSA) and targeted locus study (TLS) sequences are represented as sets grouped by a common set prefix, optionally with a set of annotations that describes all sequences in the set. Contig sets are not retrieved by individual contigs but as the whole set. Only the annotation of contig sets can be retrieved using the browser API due to the scale of the sequence sets. WGS sets are available to search across within the wgs_masters domain, TSA sets within the tsa_masters domain and TLS sets withing the tls_masters domain.

## Building a Query

The query parameter in a text based search is required. It should be in the format of one or multiple search terms separated by while spaces and combined by logic operators. For example:

```
query=(reductase OR transferase) AND glutathione
```

A wildcard character (*) can be used at the beginning or end of a string value for partial matching and enclosing a term in double quotes (" ") can be used for exact matches.

If you would like to target a specific field within the metadata in your search, you can specify this in the format *field:term.* For example:

```
query=description:dopamine
```

The default order of results is based on their relevance, i.e. the proximity of the terms in the entries.

### *Escaping special characters*

The following characters require escaping (using a ' \ ' before the character to escape) in order to be correctly interpreted as part of a search term:

+ - & | ! ( ) { } [ ] ^ " ~ * ? : \ /

## Size limit and pagination of result

A search could return a few or millions of results. As a safety feature to protect against unintended download of large files (or time-outs in scripts that process directly from the URL stream), the default page size is set to 100,000 records. This can be changed using the limit parameter. The limit can be raised for larger page sizes, alternatively if you wish to retrieve all results, the limit should be set to 0, however, it is better practice to make multiple size-limited calls.

The offset parameter is available to use with the limit for pagination and should be set to how many records should be skipped. For example, to fetch the first two pages of a search result with a page size of 200,000 records:

Page one: `limit=200000`

Page two: `offset=200000&limit=200000`

# Retrieving related ENA records

To directly download records related to a particular accession, use `/links/{linksResult}`.

This endpoint is for retrieving all records related to a specific 'Study', 'Sample' or 'Taxon' record. As a result, the following three path variables are allowed for `{linksResult}` :

`/links/study?{parameters}`

`/links/sample?{parameters}`

`/links/taxon?{parameters}`

## Parameters of links/{linksResult} function

| Parameter | Valid Format | Description |
|---|---|---|
| **result** | string | The result (data-type) of the related records to retrieve. |
| **accession** | string | The accession of the record to get links for (study, sample or taxon). |
| **subtree** | boolean | Whether to include the subtree of the taxon. |
| **gzip** | boolean | Download the result as a gzip file (false by default) |

### Result

To retrieve the related records of a study, sample or taxon, you must choose the correct datatype *result* to retrieve. Not all record formats are compatible with all *result*s.

This table summarises which *result* data types can be used for each record format and the record type that will be returned.

| Record Format | Data type 'result' that can be used for retrieval | Resulting ENA record type |
|---|---|---|
| **/fasta** | sequence<br>coding<br>noncoding | Sequence |
| **/text** | sequence<br>coding<br>noncoding | Sequence |
| | wgs_set<br>tsa_set<br>tls_set | Contig set |
| **/xml** | study<br>read_study<br>analysis_study | Study |
| | sample | Sample |
| | read_run | Run |
| | read_experiment | Experiment |
| | analysis | Analysis |

| | | |
|---|---|---|
| | assembly | Assembly |
| | taxon | Taxon |

*Study records*

Study records represent a research project. The Study holds information about the motivation of the project as well as links to publications.

*Sample records*

Sample records represent biological samples that were involved in a research project. They hold metadata regarding sample collection and processing (and sometimes analysis in special cases).

*Run records*

Run records hold the location of the sequencing data files.

*Experiment records*

Experiment records hold metadata of the sequencing event.

*Analysis records*

Analysis records hold metadata and the location of data files resulting from any subsequent analysis of sequencing data.

*Assembly records*

Assembly records are a metadata object that holds an overview of a genome assembly. They hold all the components of the genome assembly (e.g. the contig set, scaffold sequences or full chromosomes). They also hold generated assembly statistics.

*Sequence records*

Sequence records hold targeted assembled nucleotide sequences and sometimes annotation. This includes coding, non-coding and marker regions. These records do not include whole genome shotgun (WGS), transcriptome assembly (TSA) or targeted locus study (TLS) sequences. These are instead separated into the contig set results described below.

*Contig set records*

Whole genome shotgun (WGS), transcriptome assembly (TSA) and targeted locus study (TLS) sequences are represented as sets grouped by a common set prefix, optionally with a set of annotations that describes all sequences in the set. Contig sets are not retrieved by individual contigs but as the whole set. Only the annotation of contig sets can be retrieved using the browser API due to the scale of the sequence sets.

## Accession

The accession to retrieve links for has to be provided. This can only be a study or sample accession or an NCBI tax ID and must be a valid accession of the linksResult specified.

This table shows valid accession formats for each linksResult option:

| linksResult | Valid accession format |
|---|---|
| /study | (E\|D\|S)RP[0-9]{6,}<br>PRJ(E\|D\|N)[A-Z][0-9]+ |
| | [A-Z]{1}[0-9]{5}.[0-9]+<br>[A-Z]{2}[0-9]{6}.[0-9]+<br>[A-Z]{2}[0-9]{8} |

| | |
|---|---|
| /sample | (E\|D\|S)RS[0-9]{6,}<br>SAM(E\|D\|N)[A-Z]?[0-9]+ |
| /taxon | [0-9]+ |

# Retrieving record version history

Version history of versioned records can be retrieved via the `/versions/{accession}` endpoint. This will return a summary of all versions of the specified record in JSON format. The version number of an assembly or sequence record is indicated at the end of the accession after a '.' e.g. GCA_000001405.**28**. For contigs, the version is captured in the master accession e.g. OMHF**01**.

Not all records are versioned. The following table summarises versioned record types and valid accessions to retrieve the version summary:

| ENA record types that are versioned | Valid accession format for retrieval (unversioned accession) |
|---|---|
| Assembly | GCA_[0-9]{9} |
| Sequence | [A-Z]{1}[0-9]{5}<br>[A-Z]{2}[0-9]{6}<br>[A-Z]{2}[0-9]{8} |
| Contig set | [A-Z]{4}[0-9]{2}<br>[A-Z]{6}[0-9]{2} |

## Example:

https://www.ebi.ac.uk/ena/browser/api/versions/GCA_000001405


```
{
  "accession": "GCA_000001405",
  "versions": [
    {
      "accession": "GCA_000001405",
      "sequenceVersion": 28,
      "xml": "https://www.ebi.ac.uk/ena/browser/api/xml/GCA_000001405.28",
      "status": "public"
    },
    {
      "accession": "GCA_000001405",
      "sequenceVersion": 27,
      "xml": "https://www.ebi.ac.uk/ena/browser/api/xml/GCA_000001405.27",
      "status": "public"
    },
… continued
```

# Examples

All examples given below show the ENA browser API URL only. Examples of using these URLs with *curl* are given in the Introduction section.

## Download a range of analysis records

download a range of analysis accessions in XML format.

https://www.ebi.ac.uk/ena/browser/api/xml/ERZ1030230-ERZ1030297?download=true

## Fetch a list of sequences in fasta format

Fetch a list of sequences in a concatenated fasta format.

https://www.ebi.ac.uk/ena/browser/api/fasta/AA046425%2CAA046385%2CAA046323

## Download the first 500 analyses from a search for high quality binned metagenomes

Download the first 500 binned metagenomes when searching for bins with a completeness score of >90% and contamination score <10%.

https://www.ebi.ac.uk/ena/browser/api/xml/search?result=analysis&query=assembly_type%3D%22binned%20metagenome%22%20AND%20contamination_score%3C10%20AND%20completeness_score%3E90&limit=500

## Fetch 5 non-coding sequences resulting from a text search

Fetch the first 5 (most relevant) non-coding sequences resulting from a search for tRNA and Listeria.

https://www.ebi.ac.uk/ena/browser/api/text/textsearch?domain=non-coding&query=tRNA%20AND%20Listeria&limit=5

## Download all studies containing analyses of Homo sapiens

Fetch all studies containing analyses of Homo sapiens – all analysis studies linked to the taxon 9606.

https://www.ebi.ac.uk/ena/browser/api/xml/links/taxon?result=analysis_study&accession=9606&subtree=false

## Retrieve a summary of all versions of the human genome

Retrieve a summary of all versions of the reference human assembly (GCA_000001405).

https://www.ebi.ac.uk/ena/browser/api/versions/GCA_000001405