ENA Portal API

Programmatic access to data held within ENA

Change History

Document Version	Date	Author	Notes
1.0		Nicole Silvester	Initial version.
1.1		Suran Jayathilaka	Authentication changes. Sending dummy credentials no longer required for anonymous search.
1.2	2018-02-02	Suran Jayathilaka	Asynchronous search: Submit a search and download results later.
1.2.1	2018-02-07	Suran Jayathilaka	Changed notifyEmail param name to asyncSearchNotifyEmail
1.2.2	2018-03-15	Suran Jayathilaka	Changed asyncSearchNotifyEmail param name to email. asyncSearchNotifyEmail is still supported.
1.2.3	2018-04-30	Suran Jayathilaka	Email parameter is mandatory when fields=all is set.
1.3	2018-10-24	Suran Jayathilaka	Email parameter is not required for fields=all searches
1.4	2019-12-10	Suran Jayathilaka	Updated support info. Removed async search details as long search timeout doesn't occur anymore.

Introduction to ENA portal API

The main function of the ENA portal API is to provide access for users to search against all available data in ENA. In this API we introduce the term "data portal". This a way to allow a search to be performed against different sets of data, which can include a mixture of public and pre-publication (private) data. This concept of data portal, as well as the data portals currently available, is described in more detail in the following chapter "Performing a search".

As the API provides access to pre-publication data, user authentication is required when performing searches for such data. Any Webin (submission) or DCC (data hub) account is supported. If you wish to search against only public data, you can search without passing credentials. By default, any username provided that **does not** begin with "Webin-" or "dcc_" is assumed to be an anonymous access.

The portal API is available from https://www.ebi.ac.uk/ena/portal/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. However, beyond providing ease of use for testing, we don't recommend you use the API via a web browser as there are other tools available more appropriate for that purpose (such as Advanced Search within the ENA Browser).

Access to the portal API will likely be through either inclusion within scripts or using a tool such as *wget* and *curl*. When using any of these options, it is important to remember that when performing authenticated searches, a request header is needed for supplying the username and password. For anonymous searches on public data, username and password are not required.

For example, to download a search result using *curl* (anonymously):

```
curl -X GET --header 'Accept: application/json'
'https://www.ebi.ac.uk/ena/portal/api/search?<search_definition>'
```

The same search with authentication:

```
curl -X GET --header 'Accept: application/json' -u username:password
'https://www.ebi.ac.uk/ena/portal/api/search?<search_parameters>'
```

There are three important points to note about the above command:

1. The search URL is bound by single quotes (or double quotes if you're using curl in a Windows command line) . Due to the search URL consisting of several parameters, joined by the "&" character and often containing double quotes around text field

values, this is required for the entire URL to be read by *curl*. This is also true when using *wget*.

2. HTTPS is used in the URL instead of HTTP.

3. The above command will write the data to the stream (it will be printed to your screen and not saved). To save into a named output file, you should use the "-o" option in *curl* (or the "-O" option in *wget*). For example:

```
curl -o output.txt -X GET --header 'Accept: application/json' -u
username:password
'https://www.ebi.ac.uk/ena/portal/api/search?<search_parameters>'
```

Endpoints

While there is one main function of the API, several additional endpoints are available to support searches. These are listed in the table below.

ENA portal API endpoints

Endpoint	Purpose	Authentication Header required
doc	Download the latest documentation for the API.	No
search	Perform a search against a single data group (result) of the data available in ENA. This is optionally against all public data in ENA or against a subset of data (which can be both public and private) defined by a "data portal". Both GET and POST are supported.	Yes for a data portal specific search. No to search across all public data.
results	Fetch a list of all results available to search against. The list will be different based on the data portal identified in the request.	No
searchFields	Fetch a list of all searchable fields for a given result.	No
returnFields	Fetch a list of all returnable fields for a given result.	No
controlledVocab	Fetch a list of all available controlled vocabulary values for a given field.	No

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

```
https://www.ebi.ac.uk/ena/portal/api/doc?format=pdf
```

All other endpoints are described in the following chapters.

Performing a search

A search is performed via the `/search` endpoint. Several parameters are available, but only one is required: result. All available parameters are listed in the table below.

Parameters for the search endpoint

Parameter	Data type	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.
fields	string	A list of fields (comma separated) to be returned in the result. If none supplied, the accession and description/title of the main result object will be returned. To get ALL available fields for the result, you can use the convenience value of "fields=all".
dataPortal	string	The data portal ID.
sortFields	string	A list of fields (comma separated) in the order of which the results should be sorted.
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.
dccDataOnly	string	Whether to limit the search to only DCC records. By default, all public data is also included in the search.
includeMetagenomes	string	Whether to include public metagenome data in the search. By default, these are not included. Note that any metagenome data associated with a DCC hub will always be included in a search against that DCC.
format	string	What format the results should be returned as: TSV or JSON. By default, a TSV report is provided.
download	boolean	Whether to download the result as a file, rather than read it from the stream. By default, this is false.

Data portals

A "data portal" defines the set of data that is available for searching. There are currently two data portals available: *ena* and *pathogen*. If no data portal is defined in the request, the ENA portal will be used by default.

ENA portal

The ENA portal covers all public data held within ENA.

Pathogen portal

As there is not yet a public repository defining all pathogenic taxa and strains, the pathogen portal contains all data from the Bacteria, Virus, Kinetoplastida and Amoebozoa lineages from the NCBI taxonomic tree. The pathogen portal can be accessed anonymously, in which case only public data can be searched. Alternatively, a user can authenticate their search using either a Webin account or a DCC (data hub) account. When using a Webin account, all data hubs that the user is registered with will be included in the search.

For authentication, usernames that being with "dcc_" or "Webin-" are accepted. Any other values are ignored and the search is treated as an anonymous search.

Results

A result is a set of data that can be searched against and returned. These are largely based on the different data types available within ENA, however there is also some sub-divisions created by the data release/distribution practice.

Results may be added at any time to a data portal and the documentation may not yet be updated to those available. You can find the latest list of results available for a data portal using the `/result` endpoint. This requires just one parameter: dataPortal.

Results available for each data portal

Result	ENA	Pathogen
read_run	Y	Y
read_experiment	Y	Y
read_study	Y	Y
analysis	Y	Y
analysis_study	Y	Y
sample	Y	Y
study	Y	
assembly	Y	
sequence_release	Y	
sequence_update	Y	
wgs_set	Y	
tsa_set	Y	
coding_release	Y	
coding_update	Y	
noncoding_release	Y	

Result	ENA	Pathogen
noncoding_update	Y	
taxonomy	Y	

Sample result

Samples are of particular importance within the read and analysis domains, where all of the metadata describing the experiment and analysis are only available within the sample record. Assembled and annotated sequences (and their feature level products, coding and noncoding) traditionally hold sample annotation within their source feature, however samples are also gaining a more important role within these data types. This is due to the wider scope of information sample records can hold, as well as standardisation of fields (and formats) for communities based on sample checklists. Assemblies are one area that are gaining larger links to sample records.

Study result

While the *read_study* and *analysis_study* results cover all studies that are referenced by read and analysis data, the *study* result covers all available studies, including those that are linked to genome assemblies and other assembled and annotated sequences. The other main difference is that for the study result only study-specific fields are available to be searched and returned while the other two also have data-specific (read/analysis) fields.

Read results

There are three different results that can searched within the read data based on the three data types: runs (*read_run* result), experiments (*read_experiment* result) and studies (*read_study* result). These data types are all linked and therefore the same fields are available for searching against for all three results. However the *read_experiment* and *read_study* results are subsets of the *read_run* result and therefore only a subset of the read_run fields are returnable for each.

Analysis results

There are two different results that can be searched within the analysis data based on the two data types: analyses (*analysis* result) and studies (*analysis_study* result). As for the read results, these have the same fields available for searching, but the the *analysis_study* result is a subset of the *analysis* result.

Assembly result

While all other data types have only a single public version, there can be multiple public assembly versions. The *assembly* result returns information on the latest assembly version. This is something to bear in mind when searching against the assembly name as the record returned may not be the version that matches the name.

Sequence results

There are two results available when searching against assembled and annotated sequences, based on the release and distribution cycle of this data type. There is a quarterly release containing all sequences that are public at that time and daily updates that include all new and updated sequences. The latter is represented as a cumulative update of changes since the last release. The two results covering this data are *sequence_release* and *sequence_update*. Both of these results have the same fields available to both search against and return in the resultant report.

This data set does not include whole genome shotgun (WGS) or transcriptome analysis (TSA) sequences. These are instead separated into the contig set results described below.

Note that due to this segmentation of data, a sequence that has been suppressed since the last release can still be returned within the *sequence_release* result.

Contig set results

Whole genome shotgun (WGS) and transcriptome analysis (TSA) sequences are represented as sets grouped by a common set prefix, with a set of annotations that describes all sequences in the set. These sets can contain a few to millions of sequences and are usually required as a whole as opposed to as individual sequences. For this reason, these sequences are treated differently from all other assembled and annotated sequences. There are two results covering this data: *wgs_set* and *tsa_set*. When performing a search against one of these results, the search is being made against the master record, holding all common information on the set. As such, the accession returned in the resultant report is the master accession. However the files that are linked within these results contain the full set (with the exception of the master_file field), allowing users to obtain all sequences of interest in one go.

Coding results

Coding records are generated from the CDS features within assembled and annotated sequences. For this reason, they follow the same release and distribution cycle as described above with all the same caveats. The two results available for this data type are *coding_release* and *coding_update*.

Noncoding results

Noncoding records are generated from several features within assembled and annotated sequences; mostly the different RNA features. For this reason, they follow the same release and distribution cycle as described above with all the same caveats. The two results available for this data type are *noncoding_release* and *noncoding_update*.

Specifying the required output

Choosing which fields to return

By default, the output from a search will be tab separated format (TSV) report consisting of two columns, the accession and title/description. If wishing to retrieve a different set of columns in the report, the *fields* parameter should be used. A comma-separated list of

returnable fields for the result should be supplied. The order of the columns will match the order of the request. For example, to drop the description field and fetch the scientific name, strain and collection date for a sample along with the accession, one would use:

```
fields=sample_accession,scientific_name,strain,collection_date
```

If the accession column for the result is not listed, it will be prepended to the list of requested fields. For example, the following list of fields will produce exactly the same report as the example given above:

```
 fields=scientific_name,strain,collection_date
```

To get ALL available fields for the result, you can use the convenience value of "fields=all".

Selecting the sort order of the result

In most cases, the search results should be ordered based on the accession and this is therefore the default behaviour. In some cases however, it may be more useful to order by a different field in order to group similar data together. In this case, the sortFields parameter should be used, which again takes a comma-separated list of fields. For example, building on the above example for samples, it might be desirable to group by scientific name, strain and then by accession. To do this, one would use:

```
sortFields=scientific_name,strain,sample_accession
```

Setting the format of the output

The output from the search is presented as a tab separated report with the first row providing the field IDs as column headers. This is the most useful format for human users (as opposed to fully programmatic/scripted use) as it can be imported into spreadsheet programs like Excel. Many script writers may also find it a preferred option due to simple splitting of the columns based on a tab. However, some programmatic users might prefer JSON formatted output. This can be requested using the *format* parameter:

```
format=json
```

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.

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`

Pathogen portal specific search options

By default a search against the pathogen portal will look at all public data plus any pre-publication data that is shared within any of the data hubs (DCC accounts) that are authorised for the user account. In some cases, only the data within the data hub(s) should be searched. In this case, the *dccDataOnly* parameter should be used and set to true. In the case of a user accessing the API from a DCC account, `dccDataOnly=true` will result in the search being performed against all public and pre-publication data associated with that DCC data hub. If a Webin account was being used which was associated with two DCC accounts, the search would cover all data associated with <u>both</u> of the accounts.

We may add an option at a later date to refine searches to single DCC data hub (or subset) to which a Webin account is registered.

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.

To fetch a list of all available values for controlled vocabulary fields, the following endpoint is available: `/controlledVocab?field=<fieldname>`

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.

Geospatial functions

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)

Function	Description	Parameters	Example
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.

Taxonomic functions

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

The set of fields available for searching is dependent on the data portal and the result type. These fields are listed here. Note that fields may be added to the API faster than they are added to this documentation.

A full list of the latest result fields can be fetched from the following endpoint:

`/searchFields?result=<resultId>`

By default the searchable fields for that result for the ENA data portal will be returned. To request the fields for a specific portal, the *dataPortal* parameter needs to be used. For example to fetch the *sample* fields for the *pathogen* portal:

`/searchFields?result=sample&dataPortal=pathogen`

Sample fields

The set of fields searchable against samples is specific to a data portal. This is due to the different priorities of information to be searched, and therefore indexed, for each use case. Sample fields can also be used for refining data within the read and analysis results.

ENA sample fields

Field	Filter type	Field	Filter type
sample_accession	text	secondary_sample_accession	text
sample_alias	text	sample_title	text
broker_name	controlled vocab.	center_name	text
altitude	number	bio_material	text
cell_line	text	cell_type	text
checklist	controlled vocab.	collected_by	text
collection_date	date	country	text
cultivar	text	culture_collection	text
depth	number	dev_stage	text
ecotype	text	elevation	number
environment_biome	text	environment_feature	text
environment_material	text	environmental_package	controlled vocab.
environmental_sample	boolean	experimental_factor	text
first_public	date	germline	boolean
host	text	host_body_site	text

Field	Filter type	Field	Filter type
host_genotype	text	host_growth_conditions	text
host_phenotype	text	host_sex	controlled vocab.
host_status	text	host_tax_id	number
identified_by	text	investigation_type	controlled vocab.
isolate	text	isolation_source	text
last_updated	date	location	geospatial
mating_type	text	ph	number
project_name	text	protocol_label	text
salinity	number	sample_collection	text
sampling_campaign	text	sampling_platform	text
sampling_site	text	sequencing_method	text
serovar	text	sex	controlled vocab.
specimen_voucher	text	strain	text
sub_species	text	sub_strain	text
target_gene	text	taxonomy	taxonomy
temperature	number	tissue_lib	text
tissue_type	text	variety	text

Pathogen sample fields

Field	Filter type	Field	Filter type
sample_accession	text	secondary_sample_accession	text
sample_alias	text	sample_title	text
broker_name	controlled vocab.	center_name	text
checklist	controlled vocab.	collected_by	text
collecting_institute	text	collection_date	date
country	text	dev_stage	text
environmental_sample	boolean	first_public	date
host_common_name	text	host_scientific_name	text
host_sex	controlled vocab.	host_status	text
host_subject_id	text	host_tax_id	text

Field	Filter type	Field	Filter type
influenza_test_method	text	influenza_test_result	text
investigation_type	controlled vocab.	isolate	text
isolation_source	text	last_updated	date
location	geospatial	other_pathogens_result	text
other_pathogens_tested	text	receipt_date	date
region	text	sampling_site	text
serovar	text	sewage_type	text
sex	controlled vocab.	strain	text
taxonomy	taxonomy	tissue_type	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.

Read search fields for all data portals

Field	Filter type	Field	Filter type
study_accession	text	secondary_study_accession	text
experiment_accession	text	run_accession	text
sample_accession	text	secondary_sample_accession	text
study_alias	text	experiment_alias	text
run_alias	text	sample_alias	text
study_title	text	experiment_title	text
first_public	date	last_updated	date
broker_name	controlled vocab.	center_name	text
instrument_model	controlled vocab.	instrument_platform	controlled vocab.
library_layout	controlled vocab.	library_name	text
library_selection	controlled vocab.	library_source	controlled vocab.
library_strategy	controlled vocab.	nominal_length	number
base_count	number	read_count	number
submitted_format	text	parent_study	text
taxonomy	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.

Analysis search fields for all data portals

Field	Filter type	Field	Filter type
analysis_accession	text	study_accession	text
secondary_study_accession	text	sample_accession	text
secondary_sample_accession	text	analysis_alias	text
study_alias	text	sample_alias	text
analysis_title	text	study_title	text
broker_name	controlled vocab.	center_name	text
analysis_type	controlled vocab.	first_public	date
last_updated	date	parent_study	text
taxonomy	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.

Assembly fields

Field	Filter type	Field	Filter type
accession	text	study_accession	text
sample_accession	text	secondary_sample_accession	text
assembly_name	text	assembly_title	text
study_name	text	study_title	text
study_description	text	assembly_level	controlled vocab.
genome_representation	controlled vocab.	strain	text
taxonomy	taxonomy		

Sequence fields

The fields in the table below are searchable for both the *sequence_release* and *sequence_update* results. Most of the source feature qualifiers are available to search against.

Sequence search fields

Field	Filter type	Field	Filter type
accession	text	study_accession	text
sample_accession	text	secondary_sample_accession	text
description	text	sequence_md5	text
altitude	number	base_count	number
bio_material	text	cell_line	text
cell_type	text	collected_by	text
collection_date	date	country	text
cultivar	text	culture_collection	text
dataclass	controlled vocab.	dev_stage	text
ecotype	text	environmental_sample	boolean
first_public	date	germline	boolean
haplotype	text	host	text
identified_by	text	isolate	text
isolation_source	text	keywords	text
lab_host	text	last_updated	date
location	geospatial	mating_type	text
mol_type	controlled vocab.	organelle	controlled vocab.
plasmid	text	serotype	text
serovar	text	sex	controlled vocab.
specimen_voucher	text	strain	text
sub_species	text	sub_strain	text
tax_division	controlled vocab.	taxonomy	taxonomy
tissue_lib	text	tissue_type	text
topology	controlled vocab.	variety	text

Contig set fields

The fields in the table below are searchable for both the *wgs_set* and *tsa_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.

Contig set search fields

Field	Filter type	Field	Filter type
accession	text	study_accession	text
sample_accession	text	secondary_sample_accession	text
description	text	altitude	number
bio_material	text	cell_line	text
cell_type	text	collected_by	text
collection_date	date	country	text
cultivar	text	culture_collection	text
dev_stage	text	ecotype	text
environmental_sample	boolean	first_public	date
germline	boolean	haplotype	text
host	text	identified_by	text
isolate	text	isolation_source	text
keywords	text	lab_host	text
last_updated	date	location	geospatial
mating_type	text	mol_type	controlled vocab.
serotype	text	serovar	text
sex	controlled vocab.	specimen_voucher	text
strain	text	sub_species	text
sub_strain	text	tax_division	controlled vocab.
taxonomy	taxonomy	tissue_lib	text
tissue_type	text	variety	text

Coding fields

The fields in the table below are searchable for both the *coding_release* and *coding_update* results. As for searches against sequences, most of the source feature

qualifiers are available to search against. Selected CDS-feature specific information has also been included.

Coding search fields

Field	Filter type	Field	Filter type
accession	text	parent_accession	text
study_accession	text	sample_accession	text
secondary_sample_accession	text	description	text
sequence_md5	text	allele	text
altitude	number	artificial_location	controlled vocab.
base_count	number	bio_material	text
cell_line	text	cell_type	text
codon_start	controlled vocab.	collected_by	text
collection_date	date	country	text
cultivar	text	culture_collection	text
dataclass	controlled vocab.	dev_stage	text
ec_number	text	ecotype	text
environmental_sample	boolean	exception	text
experiment	text	first_public	date
function	text	gene	text
gene_synonym	text	germline	boolean
haplotype	text	host	text
identified_by	text	inference	text
isolate	text	isolation_source	text
keywords	text	lab_host	text
last_updated	date	location	geospatial
locus_tag	text	map	text
mating_type	text	mol_type	controlled vocab.
old_locus_tag	text	operon	text
organelle	controlled vocab.	product	text
protein_id	text	pseudo	text
pseudo_gene	text	ribosomal_slippage	boolean

Field	Filter type	Field	Filter type
serotype	text	serovar	text
sex	controlled vocab.	specimen_voucher	text
standard_name	text	strain	text
sub_species	text	sub_strain	text
tax_division	controlled vocab.	taxonomy	taxonomy
tissue_lib	text	tissue_type	text

Noncoding fields

The fields in the table below are searchable for both the *noncoding_release* and *noncoding_update* 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.

Noncoding search fields

Field	Filter type	Field	Filter type
accession	text	study_accession	text
sample_accession	text	secondary_sample_accession	text
description	text	sequence_md5	text
anticodon	text	base_count	number
collected_by	text	collection_date	date
country	text	dataclass	controlled vocab.
dev_stage	text	environmental_sample	boolean
experiment	text	first_public	date
function	text	gene	text
gene_synonym	text	inference	text
keywords	text	lab_host	text
last_updated	date	location	geospatial
locus_tag	text	mol_type	controlled vocab.
organelle	controlled vocab.	product	text
rna_class	controlled vocab.	strain	text
tax_division	controlled vocab.	taxonomy	taxonomy

Returnable fields

While there is a large overlap between the searchable and returnable fields for each result, there are some differences. The returnable fields for each result are listed here. Note that fields may be added to the API faster than they are added to this documentation.

A full list of the latest result fields can be fetched from the following endpoint:

```
/returnFields?result=<resultId>
```

By default the returnable fields for that result for the ENA data portal will be retrieved. To request the fields for a specific portal, the *dataPortal* parameter needs to be used. For example to fetch the *sample* fields for the *pathogen* portal:

```
/returnFields?result=sample&dataPortal=pathogen
```

Sample fields

The *sample* result is unique amongst all results in that it is custom built for each data portal. Each data portal has different priorities with respect to the plethora of sample attributes, giving rise to a different, targeted, set of fields for each. If a field is not available for the data portal you are using, please contact our helpdesk using https://www.ebi.ac.uk/ena/browser/support to request its addition. If there is sufficient use of and/or demand for the field, we will add it.

These fields listed for each data portal below are available when searching against the *sample* result, but also when searching against the read and analysis results.

ENA *sample* result fields

Field	Field	Field
altitude	bio_material	broker_name
cell_line	cell_type	center_name
checklist	collected_by	collection_date
country	cultivar	culture_collection
depth	dev_stage	ecotype
elevation	environment_biome	environment_feature
environment_material	environmental_package	environmental_sample
experimental_factor	first_public	germline
host	host_body_site	host_genotype
host_gravidity	host_growth_conditions	host_phenotype
host_sex	host_status	host_tax_id

Field	Field	Field
identified_by	investigation_type	isolate
isolation_source	location	mating_type
ph	project_name	protocol_label
salinity	sample_accession	sample_alias
sample_collection	sample_title	sampling_campaign
sampling_platform	sampling_site	scientific_name
secondary_sample_accession	sequencing_method	serovar
sex	specimen_voucher	strain
sub_species	sub_strain	submitted_host_sex
submitted_sex	target_gene	tax_id
temperature	tissue_lib	tissue_type
variety		

Pathogen *sample* result fields

Field	Field	Field
broker_name	center_name	checklist
collected_by	collecting_institute	collection_date
country	dev_stage	environmental_sample
first_public	host_common_name	host_scientific_name
host_sex	host_status	host_subject_id
host_tax_id	influenza_test_method	influenza_test_result
investigation_type	isolate	isolation_source
location	other_pathogens_result	other_pathogens_tested
receipt_date	region	sample_accession
sample_alias	sample_title	sampling_site
scientific_name	secondary_sample_accession	serovar
sewage_type	sex	strain
submitted_host_sex	submitted_sex	tax_id
tissue_type		

Read fields

In addition to the fields listed below for *read_run* searches, any *sample* field can also be returned. Unlike for samples, the available columns for each of these results are consistent across all data portals.

read_run fields for all data portals

Field	Field	Field
base_count	broker_name	center_name
cram_index_aspera	cram_index_ftp	cram_index_galaxy
experiment_accession	experiment_alias	experiment_title
fastq_aspera	fastq_bytes	fastq_ftp
fastq_galaxy	fastq_md5	first_public
instrument_model	instrument_platform	last_updated
library_layout	library_name	library_selection
library_source	library_strategy	nominal_length
read_count	run_accession	run_alias
sample_accession	sample_alias	scientific_name
secondary_sample_accession	secondary_study_accession	sra_aspera
sra_bytes	sra_ftp	sra_galaxy
sra_md5	study_accession	study_alias
study_title	submission_accession	submitted_aspera
submitted_bytes	submitted_format	submitted_ftp
submitted_galaxy	submitted_md5	tax_id

read_experiment fields for all data portals

Field	Field	Field
experiment_accession	experiment_alias	experiment_title
instrument_model	instrument_platform	library_layout
library_name	library_selection	library_source
library_strategy	secondary_study_accession	study_accession
study_alias	study_title	

Field	Field	Field
secondary_study_accession	study_accession	study_alias
study_title		

Analysis fields

In addition to the fields listed below for *analysis* searches, any *sample* field can also be returned. As with read results, the available columns for each of these results are consistent across all data portals.

analysis fields for all data portals

Field	Field	Field
analysis_accession	analysis_alias	analysis_title
analysis_type	broker_name	center_name
first_public	last_updated	sample_accession
sample_alias	scientific_name	secondary_sample_accession
secondary_study_accession	study_accession	study_alias
study_title	submitted_aspera	submitted_bytes
submitted_ftp	submitted_galaxy	submitted_md5
tax_id		

analysis_study fields for all data portals

Field	Field	Field
secondary_study_accession	study_accession	study_alias
study_title		

Assembly fields

The following fields can be returned for the *assembly* result, based on the latest public version of the assembly.

assembly fields

Field	Field	Field
accession	assembly_level	assembly_name

Field	Field	Field
assembly_title	base_count	genome_representation
sample_accession	scientific_name	secondary_sample_accession
strain	study_accession	study_description
study_name	study_title	tax_id

Sequence fields

The following fields are available to fetch for both the *sequence_release* and *sequence_update* results.

sequence_release and *sequence_update* fields

Field	Field	Field
accession	altitude	base_count
bio_material	cell_line	cell_type
collected_by	collection_date	country
cultivar	culture_collection	dataclass
description	dev_stage	ecotype
environmental_sample	first_public	germline
haplotype	host	identified_by
isolate	isolation_source	keywords
lab_host	last_updated	location
mating_type	mol_type	organelle
plasmid	sample_accession	scientific_name
sequence_md5	serotype	serovar
sex	specimen_voucher	strain
study_accession	sub_species	sub_strain
submitted_sex	tax_division	tax_id
tissue_lib	tissue_type	topology
variety		

Contig set fields

The fields available to fetch for the *wgs_set* and *tsa_set* results differ slightly, mostly in terms of the file fields. We expect to be updating the TSA data set in the coming months and these two field sets will likely converge into a single list.

wgs_set fields

Field	Field	Field
accession	altitude	bio_material
cell_line	cell_type	collected_by
collection_date	country	cultivar
culture_collection	description	dev_stage
ecotype	embl_file	environmental_sample
fasta_file	first_public	germline
haplotype	host	identified_by
isolate	isolation_source	keywords
lab_host	last_updated	location
master_file	mating_type	mol_type
sample_accession	scientific_name	serotype
serovar	set_files	sex
specimen_voucher	strain	study_accession
sub_species	sub_strain	submitted_sex
tax_division	tax_id	tissue_lib
tissue_type	variety	

tsa_set fields

Field	Field	Field
accession	altitude	bio_material
cell_line	cell_type	collected_by
collection_date	country	cultivar
culture_collection	description	dev_stage
ecotype	environmental_sample	first_public
germline	haplotype	host
identified_by	isolate	isolation_source
keywords	lab_host	last_updated
location	mating_type	mol_type
sample_accession	scientific_name	serotype

Field	Field	Field
serovar	set_files	sex
specimen_voucher	strain	study_accession
sub_species	sub_strain	submitted_sex
tax_division	tax_id	tissue_lib
tissue_type	variety	

Coding fields

The following fields are available to fetch for both the *coding_release* and *coding_update* results.

coding_release and *coding_update* fields

Field	Field	Field
accession	allele	altitude
artificial_location	base_count	bio_material
cell_line	cell_type	codon_start
collected_by	collection_date	country
cultivar	culture_collection	dataclass
description	dev_stage	ec_number
ecotype	environmental_sample	exception
experiment	first_public	function
gene	gene_synonym	germline
haplotype	host	identified_by
inference	isolate	isolation_source
keywords	lab_host	last_updated
location	locus_tag	map
mating_type	mol_type	old_locus_tag
operon	organelle	parent_accession
product	protein_id	pseudo
pseudo_gene	ribosomal_slippage	sample_accession
scientific_name	sequence_md5	serotype
serovar	sex	specimen_voucher
standard_name	strain	study_accession

Field	Field	Field
sub_species	sub_strain	submitted_sex
tax_division	tax_id	tissue_lib
tissue_type	topology	trans_splicing
transl_except	transl_table	variety

Noncoding fields

The following fields are available to fetch for both the *noncoding_release* and *noncoding_update* results.

noncoding_release and *noncoding_update* fields

Field	Field	Field
accession	anticodon	base_count
bio_material	cell_line	cell_type
collected_by	collection_date	country
cultivar	culture_collection	dataclass
description	dev_stage	ecotype
environmental_sample	experiment	first_public
function	gene	gene_synonym
germline	host	identified_by
inference	isolate	isolation_source
keywords	lab_host	last_updated
location	locus_tag	mating_type
mol_type	organelle	product
rna_class	scientific_name	sequence_md5
serotype	serovar	sex
specimen_voucher	strain	study_accession
sub_species	sub_strain	submitted_sex
tax_division	tax_id	tissue_lib
tissue_type	topology	variety

Field definitions

Attribute names can vary greatly, especially those used to describe samples. We map all similar fields into a single representation, and normalise the format of the data wherever possible. To assist determining which fields are of interest in building search queries and result reports, a full listing of all fields and their definitions are detailed below.

Field	Description
accession	Accession number
allele	Name of the allele for the given gene
altitude	Altitude (in metres)
analysis_accession	Analysis accession number
analysis_alias	Submitter's name for the analysis
analysis_title	Brief analysis description
analysis_type	Type of sequence analysis
anticodon	Location of the anticodon of tRNA and the amino acid for which it codes
artificial_location	Indicates location is modified to adjust for the presence of a frameshift or internal stop codon
assembly_level	Assembly level (contig, scaffold, chromosome, complete genome)
assembly_name	Genome assembly name. When searching, considers all all live versions of the assembly. When returning, the name of the latest version.
assembly_title	Brief assembly description
base_count	Number of base pairs
bio_material	Identifier for biological material including institute and collection code
breed	Breed
broker_name	Name of broker for the submission
cell_line	Cell line from which the sample was obtained
cell_type	Cell type from which the sample was obtained
center_name	Submitting centre
checklist	Accession of ENA sample checklist
codon_start	Indicates the offset of the first complete codon relative to the first base of the coding feature
collected_by	Name of the person who collected the specimen

Field	Description
collecting_institute	Name of the institution to which the person collecting the specimen belongs. Format: Institute Name, Institute Address
collection_date	Date that the specimen was collected
country	Locality of sample isolation: country names, oceans or seas, followed by regions and localities
cram_index_aspera	Aspera links for CRAM index files
cram_index_ftp	FTP links for CRAM index files
cram_index_galaxy	Galaxy links for CRAM index files
cultivar	Cultivar (cultivated variety) of plant from which sample was obtained
culture_collection	Identifier for the sample culture including institute and collection code
dataclass	Sequence data class
depth	The distance below the surface of the water at which a measurement was made or a sample was collected (in metres)
description	Brief sequence description
dev_stage	Sample obtained from an organism in a specific developmental stage
ec_number	Enzyme commission number for enzyme product of sequence
ecotype	A population within a given species displaying traits that reflect adaptation to a local habitat
elevation	The elevation of the sampling site as measured by the vertical distance from mean sea level (in metres)
embl_file	Flat file for the set
environment_biome	Environment (biome). Biomes are defined based on factors such as plant structures, leaf types, plant spacing, and other factors like climate. Examples include: desert, taiga, deciduous woodland, or coral reef
environment_feature	Environmental feature level includes geographic environmental features. Examples include: harbor, cliff, or lake
environment_material	The environmental material level refers to the material that was displaced by the sample, or material in which a sample was embedded, prior to the sampling event. Examples include: air, soil, or water
environmental_package	MIGS/MIMS/MIMARKS extension for reporting of measurements and observations obtained from one or more of the environments where the sample was obtained
environmental_sample	Identifies sequences derived by direct molecular isolation from an environmental DNA sample
event_label	Label given to sampling event
exception	Indicates that the coding region cannot be translated using standard biological rules

Field	Description
experiment	A brief description of the nature of the experimental evidence
experiment_accession	Experiment accession number
experiment_alias	Submitter's name for the experiment
experiment_title	Brief experiment title
experimental_factor	Experimental factors are essentially the variable aspects of an experiment design which can be used to describe an experiment, or set of experiments, in an increasingly detailed manner
fasta_file	FASTA file for the set
fastq_aspera	Aspera links for FASTQ files
fastq_bytes	Size (in bytes) of FASTQ files
fastq_ftp	FTP links for FASTQ files
fastq_galaxy	Galaxy links for FASTQ files
fastq_md5	MD5 checksum of FASTQ files
first_public	Date when made public
function	Function attributed to a sequence
gene	Symbol of the gene corresponding to a sequence region
gene_synonym	Synonymous, replaced, obsolete or former gene symbol
genome_representation	Whether the genome assembly is a full or partial genome
geo_accession	GEO accession
germline	Indicates whether the sample is an unrearranged molecule that was inherited from the parental germline
haplotype	Combination of alleles that are linked together on the same physical chromosome
host	Natural (as opposed to laboratory) host to the organism from which sample was obtained
host_body_site	Name of body site where the sample was obtained from, such as a specific organ or tissue
host_common_name	Common name of the natural host organism from which the sample was obtained
host_genotype	Genotype of the host
host_gravidity	Whether or not subject is gravid, including date due or date post-conception where applicable
host_growth_conditions	Literature reference giving growth conditions of the host
host_phenotype	Phenotype of the host
host_scientific_name	Scientific name of the natural (as opposed to laboratory) host to the organism from which sample was obtained

Field	Description
host_sex	Physical sex of the host
host_status	Condition of host (eg. diseased or healthy)
host_subject_id	A unique identifier by which each subject can be referred to, de-identified
host_tax_id	NCBI taxon ID of the host
identified_by	Name of the taxonomist who identified the specimen
inference	A structured description of non-experimental evidence
influenza_test_method	Method by which the current assessment of a sample as flu positive/negative is made
influenza_test_result	Classification of a sample as flu positive or negative based on the test performed and reported
instrument_model	Instrument model used in sequencing experiment
instrument_platform	Instrument platform used in sequencing experiment
investigation_type	The study type targeted by the sequencing
isolate	Individual isolate from which sample was obtained
isolation_source	Describes the physical, environmental and/or local geographical source of the sample
keywords	Keywords associated with sequence
lab_host	Scientific name of the laboratory host used to propagate the source organism for the sample
last_updated	Date when record was last updated
library_layout	Sequencing library layout
library_name	Sequencing library name
library_selection	Method used to select or enrich the material being sequenced
library_source	Source material being sequenced
library_strategy	Sequencing technique intended for the library
location	Geographic location of isolation of the sample. Latitude and longitude are given in decimal degrees. When using latitude and longitude in the geospatial search functions, positive and negative values should be given to represent direction. When returned in the results report, N/S and E/W are displayed with latitude and longitude.
locus_tag	A submitter-supplied, systematic, stable identifier for a gene and its associated features
map	Map position of feature
marine_region	Geographical origin of the sample as defined by the marine region
marker	Marker classification

Field	Description
master_file	Flat file for the set master
mating_type	Mating type of the organism from which the sequence was obtained
mol_type	in vivo molecule type of the sequence
nominal_length	Average fragmentation size of paired reads
old_locus_tag	Deprecated submitter-supplied, systematic, stable identifier for a gene and its associated features
operon	Name of the group of contiguous genes transcribed into a single transcript
organelle	Membrane-bound intracellular structure from which the sequence was obtained
other_pathogens_result	Classification of a sample as positive or negative based on the test performed and reported
other_pathogens_tested	Classification of pathogenic organisms other than influenza virus tested in the current assessment of a sample
parent_accession	Parent sequence accession number
parent_study	Parent study accession number
ph	pH measurement
plasmid	Name of naturally occurring plasmid from which the sequence was obtained
product	Name of the product associated with the feature
project_name	Name of the project within which the sequencing was organised
protein_id	A stable protein identifier issued by INSDC
protocol_label	The protocol used to produce the sample
pseudo	Indicates whether the feature is non-functional
pseudo_gene	Indicates that this feature is a pseudogene
read_count	Number of reads
receipt_date	Date on which the sample was received
region	Geographical origin of the sample as defined by the specific region name followed by the locality name
ribosomal_slippage	Indicates ribosomal slippage (change to an alternative reading frame) during protein translation
rna_class	Classification of RNA
run_accession	Run accession number
run_alias	Submitter's name for the run
salinity	Salinity of water at the time of taking the sample

Field	Description
sample_accession	Sample accession number
sample_alias	Submitter's name for the sample
sample_collection	The method or device employed for collecting the sample
sample_title	Brief sample title
sampling_campaign	The activity within which this sample was collected
sampling_platform	The large infrastructure from which this sample was collected
sampling_site	The site/station where this sample was collection
scientific_name	Scientific name of the organism from which the sample was derived
secondary_sample_accession	Secondary sample accession number
secondary_study_accession	Secondary study accession number
sequence_md5	MD5 checksum of the sequence
sequencing_method	Sequencing method used
serovar	Serological variety of a species characterized by its antigenic properties
set_files	Flat file for the set
sewage_type	Type of sewage based on origin
sex	Sex of the organism from which the sample was obtained
specimen_voucher	Identifier for the sample culture including institute and collection code
sra_aspera	Aspera links for NCBI SRA format files
sra_bytes	Size (in bytes) of NCBI SRA format files
sra_ftp	FTP links for NCBI SRA format files
sra_galaxy	Galaxy links for NCBI SRA format files
sra_md5	MD5 checksum of NCBI SRA format files
standard_name	Accepted standard name for a feature
strain	Strain from which sample was obtained
study_accession	Study accession name
study_alias	Submitter's name for the study
study_description	Detailed sequencing study description
study_title	Brief sequencing study description
sub_species	Name of sub-species of organism from which sample was obtained
sub_strain	Name or identifier of a genetically or otherwise modified strain from which sample was obtained

Field	Description
submission_accession	Submission accession number
submitted_aspera	Aspera links for submitted files
submitted_bytes	Size (in bytes) of submitted files
submitted_format	Format of submitted reads
submitted_ftp	FTP links for submitted files
submitted_galaxy	Galaxy links for submitted files
submitted_host_sex	Physical sex of the host as provided by the submitter. This is in contrast to the "host_sex" field which contains standardised values.
submitted_md5	MD5 checksum of submitted files
submitted_sex	Sex of the organism from which the sample was obtained. This is in contrast to the "sex" field which contains standardised values.
target_gene	Targeted gene or locus name for marker gene studies
tax_division	Taxonomic division of the organism from which the sample was obtained
tax_id	NCBI taxon ID of the organism from which the sample was obtained
taxonomy	A virtual field representing NCBI taxonomic classification. Searchable using one of the taxonomy functions
temperature	Temperature of the sample at time of sampling
tissue_lib	Tissue library from which sample was obtained
tissue_type	Tissue type from which the sample was obtained
topology	Sequence topology: circular or linear
trans_splicing	Indicates exons from two RNA molecules are ligated in intermolecular reaction to form mature RNA
transl_except	A single codon translation that does not conform to genetic code
transl_table	Indicates the genetic code table used if other than universal genetic code table
variety	Variety (varietas, a formal Linnaean rank) of organism from which sample was derived

Examples

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

Fetch available results

Fetch the list of results that can be searched against in the pathogen data portal.

`https://www.ebi.ac.uk/ena/portal/api/results?dataPortal=pathogen`

Fetch searchable fields

Fetch the list of fields that can be searched for the assembly result.

`https://www.ebi.ac.uk/ena/portal/api/searchFields?result=assembly`

Fetch returnable fields

Fetch the list of fields that can be returned in the report for the analysis result

`https://www.ebi.ac.uk/ena/portal/api/returnFields?result=analysis`

Fetch controlled vocabulary

Fetch the list of controlled vocabulary for the checklist field. Note that the value and label contain different content. The value column is what should be used in the search query but the label column gives the title of each checklist.

`https://www.ebi.ac.uk/ena/portal/api/controlledVocab?field=checklist`

Fetch the list of controlled vocabulary for the instrument_model field. Note that as the value column contains understandable values, the label column holds the same information.

`https://www.ebi.ac.uk/ena/portal/api/controlledVocab?field=instrument_model`

Search against public samples

- Find all public samples in the ENA data portal. Return a list of accession and title, ordered by accession.

`https://www.ebi.ac.uk/ena/portal/api/search?result=sample&limit=0`

- Find all public samples in the pathogen data portal. Return a list of accession, tax ID, scientific name, collection date, country and location. Sort results by scientific name, followed by sample accession.

https://www.ebi.ac.uk/ena/portal/api/search?result=sample&dataPortal=pathogen&fields=sample_accession,tax_id,scientific_name,collection_date,country,location&sortFields=scientific_name,sample_accession&limit=0

Search against public samples with paging

- Submit the same search as above, without sorting, but requesting ALL result fields, and using an offset and limit, in json format. Useful for paging.

https://www.ebi.ac.uk/ena/portal/api/search?result=sample&dataPortal=ena
&fields=all&offset=100&limit=100&format=json

Authenticated search for a dataPortal and results download

- Using CURL, submit a search with the username and password.

curl -X GET --header 'Accept: text/plain' -u 'username:password'
'https://www.ebi.ac.uk/ena/portal/api/search?result=read_run&limit=20000
0&dataPortal=pathogen'

Search against data hub (DCC) data only

- Find all read data that is registered with data hub dcc_chopin. Return a list of the run accessions with FTP FASTQ file links. Include the MD5 checksums for the files. Note that this example URL doesn't include the data hub account information, this must be included in the request header. Note also that the URL has been changed to HTTPS so that the account information is secure during transit.

https://www.ebi.ac.uk/ena/portal/api/search?result=read_run&fields=fastq
_ftp,fastq_md5&dccDataOnly&limit=0

Search for read data using sample fields

- Find all public Salmonella read data collected in 2016. Return a list of the run accessions with FTP FASTQ file links. Include the MD5 checksums for the files.

https://www.ebi.ac.uk/ena/portal/api/search?result=read_run&query=collec
tion_date>=2016-01-01%20AND%20collection_date<=2016-12-
31%20AND%20tax_tree(590)&fields=fastq_ftp,fastq_md5&limit=0

- Perform the same search, but return the sample accession and collection date in addition to the read file data. Note that while the above search will return one row for each run accession, the following search will return one row for each run-sample combination.

https://www.ebi.ac.uk/ena/portal/api/search?result=read_run&query=collec
tion_date>=2016-01-01%20AND%20collection_date<=2016-12-
31%20AND%20tax_tree(590)&fields=fastq_ftp,fastq_md5,sample_accession,col
lection_date&limit=0