European Nucleotide Archive: Using the primary nucleotide sequence resource

Clara Amid [1]

- DNA & RNA
- Intermediate
- 3 hours

This course will show you how to search and retrieve nucleotide sequence data from the European Nucleotide Archive (ENA). The ENA provides sequence assembly details and biological annotation, and uses versioning to track sequence changes.

Learning objectives:

- Know what ENA is and where the data comes from
- Know how to access and navigate the ENA Browser and the taxonomy portal
- Be able to search ENA with text (such as a gene name) and a simple sequence
- Know how to extract the information you require from an annotated entry
- Know how to drill down to a specific region of a sequence and its associated annotation
- Know how to access old versions of a sequence

Why do we need ENA?

It is important to maintain a record of all the original (primary) sequence data produced, not only because it forms an invaluable resource for biological research, but because this data acts as important reference material from which a wealth of secondary data are derived, such as consensus genomic sequences and translated protein sequences.

The volume of sequencing data being produced is increasing with the advancements made in sequencing technology. Not only do we require complete genomes, but we want access to all the variation seen between individuals, for example disease alleles [2]. Such volumes of data present a challenge in terms of organisation and accessibility (1 [3]).

Understanding how sequence databases archive and annotate their data can greatly improve your ability to find what you want quickly, as well as to correctly interpret the information you find. ENA provides access to all nucleotide sequence data, including assembled and annotation [4]-enriched data, as well as raw data [5] as soon as it becomes available, regardless of the sequencing technology used (Figure 1).
What is ENA?

The European Nucleotide Archive [6] (ENA) is a comprehensive resource of primary nucleotide sequence information. ENA provides access to both assembled sequence and unassembled (raw) sequence reads, but places them in separate databases in order to optimise accessibility and analysis (2 [3]). Figure 2 provides a schematic representation of how the data is stored in ENA.

Figure 2. ENA is composed of 3 databases: EMBL-Bank [7] for assembled data, and the Sequence Read Archive [8] and Trace Archive for raw data [5].

ENA consists of three databases:

1) EMBL-Bank consists of:

- **Assembled** sequence data, where the submitter has assembled the sequence into one long contiguous length.
- **Annotation** [4] information that describes the biological function of specific regions of the sequence.
(such as protein-coding regions, exons and introns), which is provided by the submitter.

(2) **Sequence Read Archive** (SRA) consists of (3 [3]):

- **Reads** of raw data consisting of typically short, unassembled fragments of sequence generated using **Next Generation Sequencing** [9] (NGS) technology.

(3) **Trace Archive** consists of:

  - **Reads** of raw data consisting of unassembled fragments of sequence generated using **capillary sequencing technology** [10].

Note that the unassembled sequence data contained in the **SRA** and **Trace Archive** can be difficult to work with because these raw data forms inherently show considerable duplication of sequence data and sequencing error. Sequence reads can be downloaded (see section on ‘**How to export sequence and download data**’ [11]).

**Where does the data come from**

**Sharing data - the INSDC agreement**

All nucleotide sequences, including both assembled and **raw data** [5], come from direct submissions. However, ENA is not the only resource to accept nucleotide sequence data. In total, there are three major nucleotide sequence resources:

- **ENA** (provided by EBI)
- **GenBank** + the US **Trace and Sequence Read Archives** (provided by NCBI)
- **DDBJ** + the Japanese **Trace and Sequence Read Archives** (provided by the National Institute of Genetics)

It is important to have all nucleotide sequence data available within each of these three resources, regardless of where it has been submitted. Therefore, the three partners formed the International Nucleotide Sequence Database Collaboration (**INSDC** [12]) and agreed to exchange all sequence data on a daily basis and to provide free unrestricted access to the data (Figure 3) (4 [3]). As a result, it does not matter to which database a sequence is submitted, all three INSDC databases will obtain the same sequence data.
Figure 3. Daily exchange of data between INSDC partners.

Even though the INSDC resources contain the same sequence data, they do differ in how they organise the data, the tools they provide to analyse the data, and their links to external databases that provide supplementary information.

Sources of submitted sequence

The ENA resource accepts sequence submissions generated using any type of sequencing technology, whether it is raw sequence reads or assembled data, and with or without annotation [4]. The data is submitted by independent researchers, large sequencing consortia and patent offices.
Figure 4. Sequencing is like making a DNA puzzle: the chromosome is fragmented into short segments (library) that can be sequenced (reads), then the data is re-assembled and annotated.

ENA contains sequence in the form of raw sequence reads, assembled data and data annotated with biological information (Figure 4). Therefore, sequence data can differ widely in both length and quality. Ideally, there should be deep coverage where each base is read several times, but this is not always possible.

It is important to note that there is no filtering [13] of the data, therefore all submitted sequence is represented, even if it is identical to that in an existing entry. This means that a certain level of sequence redundancy exists in ENA.

Data quality

Some sequence and annotation [4] validation is performed by ENA, including checking taxonomy, describing features and providing tools for identifying any vector contamination. The ENA curation [14] team contact authors to amend data where necessary. It is important to note that, because ENA contains original sequence data, the sequence records can only be updated by the submitter (author). If an author does not correct the data, then errors can persist in the database.

UniProt [15], the protein sequence archive, contains useful information about the accuracy of ENA coding sequences (CDS [16]). Most of their protein sequence data is derived from translations of CDS in ENA. When creating a curated UniProt/SwissProt [17] protein sequence entry, they must review all the CDS information available for a gene product, and record this information in the entry (Figure 5).
Figure 5. Sections from a UniProtKB/SwissProt entry containing information on CDS in ENA.

[A] Cross-references section contains a list of entries in ENA that code for a gene product.

[B] ENA sequence entries are listed with notes on the accuracy of each sequence; these notes are compiled by UniProt curators.

[C] General annotation section contains comments about a gene product, including any cautions regarding the translated sequences.

[D] Sequence caution details any errors found by the UniProt curator [18] in each translated sequence.

How is the sequence assembled?

Sequence assembly is performed by the submitter, not by ENA curators (Figure 6). ENA sometimes receives the initial data as unassembled sequence reads, which are gradually complemented with additional layers of interpretation, such as assembly or annotation [4], by the author.

As a result, data is always changing. Sources of this change include:
- Assembly of sequence into larger fragments;
- Reducing the visibility of obsolete entries (once assembled);
- Sequence modifications;
- Daily updates;
- Corrections submitted by author.

**Figure 6.** An example where the resubmission of assembled sequence results in the creation of new entries and the removal of those they replace to the Sequence Version Archive (see below). Sequence might first enter ENA as SRA (Sequence Read Archive) fragmented sequence reads; it might be re-submitted as assembled WGS (Whole Genome Shotgun) sequence overlap contigs; it might be re-submitted again with further assembly as CON (Constructed) sequence entries, with the older WGS entries being consigned to the Sequence Version Archive.

SRA is the Sequence Read Archive, WGS is the whole genome shotgun data class representing sequence overlap contigs, and CON are the constructed sequences consisting of scaffolds, super-scaffolds and chromosomes.

Obsolete entries are accessible through the Sequence Version Archive (SVA). See section on 'Finding old archived entries'.

**How is the sequence annotated**

Annotation is optionally submitted by the sequencing group when they submit the sequence or at a later stage. However, ENA will accept the submission of Third Party Annotation (TPA) from people other than the sequencing group as long as there is a published reference for the annotation. TPA is placed in a separate entry from the original submission (Figure 7).

There are two types of TPA datasets:

- **TPA:experimental** data are supported by peer-reviewed experimental evidence.
- **TPA:inferential** are supported by peer-reviewed indirect experimental evidence. For instance, annotation inferred through homologous genes of experimentally-determined function.
Figure 7. ENA entry containing TPA: Experimental data.

[A] Description identifies this entry as containing TPA data.

[B] Keywords, such as TPA, can be used in a text search.

[C] Assembly provides a link to the original sequence entry/entries.

[D] References to peer-reviewed journals are required for TPA annotation.

Ensembl [26] is another source of genome annotation. Try the online course: Ensembl: Browsing chordate genomes [27].
How is the data structured?

Data classes and taxonomic divisions

We have learnt that ENA consists of three very large databases: EMBL-Bank [7], SRA and TA. Because EMBL-Bank is the only one with assembled data, it is the most complex of the three databases. EMBL-Bank contains multiple types of data, everything from whole genome shotgun assemblies to cDNA libraries and patent data. To make the data easier to access, EMBL-Bank has structured the data in two ways:

- By **Data Class** [21], which divides entries according to the type of data or method used to obtain it. For example, the WGS (whole genome shotgun) data class.

- By **Taxonomic Division**. For example, the HUM (human) taxonomic division.

A sequence can only belong to one class and to one division. For data class, a sequence will belong to the class that best describes that sequence. For taxonomic division, the most specific taxonomy is used to categorise a sequence.

EMBL-Bank is unique in using data classes and **taxonomic divisions** [28] to create intersecting slices of data. In other words, the database is first divided into data classes, then the data is subdivided by taxonomic division (Figure 8). This allows you search or download smaller volumes of data. By contrast, most other databases will allow you to access data divided by data class or by taxonomic division, but not by both.

**Figure 8.** Data is first split into classes, then it is split into intersecting slices by taxonomy.
**[A] Intersecting slice** of data consisting of ‘Mouse’ + ‘EST’ gives a reduced search set (other INSDC databases only provide parallel data slices, e.g. ‘EST’ for all taxonomy or ‘Mouse’ for all data classes).

**Data classes**

For assembled data, which is found in the EMBL-Bank database, each sequence is assigned to a single **data class** [21].

A full list of data classes can be found in the help pages [29].

The majority data classes include **STD** (standard), **CON** (constructed) and **WGS** (Whole Genome shotgun).

**Transcript data**

In ENA, **transcript** information is found in both the EMBL-Bank and the SRA databases. Within EMBL-Bank, transcript information can be found listed under several data classes, depending upon how the sequence was obtained:

- **EST** class contains raw **expressed sequence tag** [30] sequence that is of variable quality (single-pass reads).
• HTC class are high-throughput assembled transcript sequences.

• TSA class are transcriptome [31] shotgun assembly sequences consisting of derived from the SRA or TA databases.

• STD (standard) class can contain transcript information, and can be search using the ‘mol_type’ field, filtering [13] for ‘mRNA’.

A good way to search for coding transcript data is to query the EMBL-CDS [32] (coding sequence) dataset, which is derived from the different data classes in EMBL-Bank. Because this contains coding regions derived from both genomic and transcript records, you would need to filter results using ‘mol_type = mRNA’.

### Taxonomic divisions

Sequences from related species are grouped together by taxonomic divisions [28]. In all, there are 15 taxonomic divisions that are used in ENA:

<table>
<thead>
<tr>
<th>HUM human</th>
<th>FUN fungi</th>
<th>VRL viral</th>
</tr>
</thead>
<tbody>
<tr>
<td>MUS mouse</td>
<td>INV invertebrate</td>
<td>ENV environmental</td>
</tr>
<tr>
<td>ROD rodent</td>
<td>PLN plant</td>
<td>SYN synthetic</td>
</tr>
<tr>
<td>MAM mammal</td>
<td>PRO prokaryote</td>
<td>TGN transgenic</td>
</tr>
<tr>
<td>VRT vertebrate</td>
<td>PHG phage</td>
<td>UNC unclassified</td>
</tr>
</tbody>
</table>

Each sequence is only assigned to one taxonomic division (otherwise the sequence would be duplicated in different parts of the database). However, as you can see from the list above, some taxonomic divisions overlap. Therefore, sequences are classified according to the most specific division. For example, a mouse sequence could belong to MUS, ROD, MAM or VRT divisions, but it is classified as MUS as this is the most specific category (lowest taxonomic node).

### Taxonomic exclusions

Once a sequence is placed in the most specific taxonomic division, it is then excluded from all remaining taxonomic divisions so as not to duplicate data. For example, the mouse sequence is found in the MUS divisions, therefore it is excluded from the ROD, MAM and VRT divisions, even though a rat is a mammal and a vertebrate (Figure 9).
Figure 9. Sequences are assigned to the most specific taxonomic division.

[A] MUS (mouse) division contains only mouse sequences.

[B] ROD (rodent), MAM (mammal) and VRT (vertebrate) all exclude mouse sequences because a sequence can only occur in one taxonomic division.

These exclusions will become important when we look at "How to search and browse ENA". For example, if you want mouse sequences, you must be careful not to select ROD (rodent), as mouse sequences are excluded from this division. Later we will learn that there are exceptions to this rule when searching with the ENA browser, which merges taxonomic divisions together to make searching simpler.

What if no taxonomy is associated with a sequence?

All EMBL-Bank entries are assigned to a taxonomic division. However, some sequences are not associated with any formal taxonomy (e.g. synthetic sequences), while other sequences are derived from organisms whose taxonomy could not be determined (e.g. metagenomic sequences). ENA has special taxonomic divisions to overcome this problem:

Environmental sequences

ENV: contains metagenomic and environmental sequences where the taxonomy is unknown or is from a high taxonomic node such as a kingdom (Figure 10).
The taxonomy is displayed as: Organism = uncultured bacterium, uncultured eukaryote, metagenome [34].

Figure 10. ENA entry containing metagenomic data where no species was identified.

[A] **Organism** is only described in general terms ('termite gut metagenome') because little is known about the species the sequence came from.

[B] **Taxonomic Division** classified as ENV because the precise organism could not be identified.

**Synthetic sequences**
SYN contains synthetic or experimentally altered sequences.

- The taxonomy is displayed as: Organism = synthetic construct.

Transgenic sequences

TGN contains transgenic sequences. Transgenics occur when genetic material from one species is transferred to another species, either naturally or through genetic engineering techniques.

- Taxonomy is provided for both donor and recipient organisms.

Unclassified sequences

UNC contains unclassified sequences typically obtained from patents, for which taxonomic data is not always available.

- The taxonomy is displayed as: Organism = unidentified

When to use ENA

You can use ENA if you:

- have a gene or transcript of interest and you would like to know its sequence;
- would like to know about non-coding sequence found between genes (intergenic regions);
- want to explore a region around a gene of interest to find neighbouring genes;
- want to know how a gene, transcript or region was assembled into one contiguous sequence;
- want to identify orthologues of a gene in other species, or paralogues within a species;
- want to know about any sequence variations associated with a disease or mutational study;
- want to fetch a batch of sequences for analysis and study, including via programmatic access;
- want to see what sequence data is available for a specific organism;
• have an interest in a project, such as a **metagenomics project**, and you want to see all the sequence data they have generated;

• would like to know the sequence associated with a **chromosomal rearrangement**, such as the Philadelphia 22q11 chromosome associated with leukaemia;

• want to explore **Next Generation Sequencing** [9] data as soon as it is released, prior to its assembly.

ENA would not be your first choice if you:

• are interested in comprehensive **genomic information**, it is better to use the genome browser **Ensembl** [35] or **Ensembl Genomes** [36], which puts sequence data into context with additional **annotation** [4] not in ENA;

• are interested in **patent sequences**, it is better to use the **Non-redundant Patent Sequence Databases**, because they contain all the patent sequences found in ENA plus additional unique sequences and patent-specific information.

**How to access ENA**

**Accessing ENA through the EBI homepage**

You can access the ENA browser through the EBI homepage using the link shown by the red arrow in Figure 11.
Figure 11. ENA can be accessed directly from the EBI homepage.

The ENA browser

From the ENA Browser you can (Figure 12):

- search and browse ENA data;
- download bulk data from ENA;
- programmatically access ENA data;
- submit sequences to ENA.
Figure 12. The ENA homepage.

[A] **Search & Browse** link provides information on searching ENA, downloading ENA data in bulk, and gaining programmatic access to ENA.

[B] **Submit & Update** link provides information on how to submit sequences to ENA.

**How to search and browse ENA**

You can search ENA using the search boxes on the ENA Browser (Figure 13), or via programmatic access (see section: [How to Access ENA](#)).
Figure 13. You can search ENA using text or sequence.

[A] **Text search** queries ENA.

[B] **Sequence search** queries ENA using DNA or RNA sequences in a variety of formats.

[C] **Advanced search** allows you to refine a sequence search using a range of options.

[D] **Search & Browse** links provides information on programmatic access to ENA.

[E] **EBI search** is an EBI-wide search engine that queries multiple databases in addition to ENA.

If you don't find what you want and need to search again, instead of going back to the ENA homepage, try using the search box at the top of every ENA page (Figure 14).

Figure 14. You can continue searching using the query boxes at the top of the tabbed pages.
[A] **Text search** queries ENA.

[B] **Sequence search** queries ENA using DNA or RNA sequences in a variety of formats.

## How to search ENA with text

**Text search**

You can search ENA using free text, such as:

- a **gene name**, for example, P53;
- a **disease name**, such as diabetes mellitus;
- an **ENA accession number**, for example, BN000065;
- a **UniProt** accession number, for example, Q00987;
- a **keyword**, for example, mRNA cap structure;
- a **data class**, for example, CON;
- a **taxonomic division**, for example, HUM.

If you search using an ENA accession number, your search results will take you directly to the appropriate entry. If you search using any other text, your search results will consist of a list of entries (Figure 15). The browser also supports accession ranges and lists.
Figure 15. Searching ENA using an accession number or simply any free text.

[A] Text search box will query any free text.

[B] Searching using an ENA accession number (e.g. BN000065) takes you directly to that entry.

[C] Searching using any other free text (e.g. kinase) provides a list of relevant entries.

1. Open the ENA Browser [40] in a new window.
2. Type the search term human into the text search box [A].
3. Click ‘Search’ [D] to obtain search results.

Search terms consisting of two or more words are searched together, therefore all terms must be found to report a match. For example, searching for lung cancer is equivalent to searching for lung AND cancer (the AND is automatically inserted). If you want to search for terms independently insert an OR between them, for example searching for lung OR cancer will report matches that have either lung or cancer or both terms (caution: this will return an exceedingly long list).

Text search results page

When you search ENA using any term other than an ENA accession [38] number, then you will obtain a list of search results grouped by data type (Figure 16).

Assembled nucleotide entries from EMBL-Bank [7] include:

- Annotated Nucleotide Sequences: entries from the STD, EST, GSS, HGT, HTC, PAT, STS and TSA data classes;
- Whole Genome Shotgun Sequences: entries from the WGS data class [21];
- Genomic Constructed Sequences: entries from the CON data class;
- Protein-coding Sequences: entries from the CDS [16] data class.

Raw nucleotide data from SRA can be viewed grouped by a study, sample, experiment, run or submission.

Other data includes:

- Projects group together both assembled and raw data [5] for a specific sequencing project;
- Taxonomy group together both assembled and raw data for a taxonomic group.
European Nucleotide Archive: Using the primary nucleotide sequence resource
Published on EMBL-EBI Train online (https://www.ebi.ac.uk/training/online)

Figure 16. ENA browser results page for a text search on ‘human’.

[A] **Assembled Nucleotide Sequences** provides a list of relevant entries grouped by data class from EMBL-Bank Release and EMBL-Bank Update (data deposited after the current release date).

[B] **Raw Nucleotide Sequences** provides a list of SRA (Next Gen sequencing data) entries grouped by study, sample, experiment, run or submission

[C] **Other** provides a list of assembled and raw data grouped by the sequencing project (‘Projects’) or taxonomy (‘Taxa’ [41]).

[D] **Expand** to view entries in each category.

Trace Archive data can be obtained by searching on either the TI accession or TI name.

**How to search ENA with sequence**

**Sequence search**

Only the [EMBL-Bank](https://www.ebi.ac.uk/ena) [7] database of ENA can be searched with a nucleotide sequence; the SRA and TA databases contain [raw data](https://www.ebi.ac.uk/ena) [5] composed of very short redundant sequences that make them unsuitable for sequence searching.
The ENA browser allows you to search the entire EMBL-Bank database using either a DNA or RNA query sequence. Searching with a sequence is useful if you:

- have a sequence but are not sure of the gene name;
- have an unknown sequence you want to identify;
- want to find orthologues of a gene in other species, or paralogues within a species;
- want to identify sequence variants for a gene, including disease or mutant alleles [2];
- want to check whether you have identified a novel sequence.

The sequence search box on the ENA Browser will accept either an EMBL-Bank accession [38] number (where it will automatically insert the sequence from that accession):

...or a nucleotide sequence:

![Sequence Search](image)

The nucleotide sequence can be in plain text (i.e. straight sequence with no header) or FASTA [42] format (as above).

This simple sequence search will query both EMBL-Bank, Ensembl [35] and Ensembl Genomes [36] for similar sequences.

**Accessing the advanced sequence search**

There is a link from the ENA browser to an advanced sequence search page (Figure 17), which allows you to refine your search to a specific section of EMBL-Bank [7], Ensembl [35] or Ensembl Genomes [36], as well as to change search method.
Figure 17. Link from the ENA browser to the advanced sequence search page.

[A] Advanced Search link on the ENA browser will direct you to the advanced sequence search page.

1. Open the ENA browser [40] in a new window.

2. Click on the ‘Advanced Search’ option [A].

Advanced sequence search

There are several search options available on the advanced search page that allow you to refine your search (Figure 18).
European Nucleotide Archive: Using the primary nucleotide sequence resource
Published on EMBL-EBI Train online (https://www.ebi.ac.uk/training/online)

Figure 18. Advanced sequence search page; at the foot of the page you can change the search method.

[A] Search box will take a sequence or an EMBL-Bank [7] accession [38] number (same as for the simple sequence search).

[B] File Upload allows you to search on sequences held in a file (see help box below for allowable sequence formats).

[C] Search modes allow you to select different types of searches.

[D] Collection allows you to restrict your search to a section of EMBL-Bank, Ensembl [35], or Ensembl Genomes [36] (for example, an EMBL-Bank taxonomic division or a species in Ensembl). If nothing is selected, then the search will be carried out on all EMBL-Bank + all Ensembl + all Ensembl Genomes, which could be a lengthy search.

- When an EMBL-Bank taxonomic division is selected, it will automatically include more specific taxonomic divisions [28] in addition to the selected division.
- For example, searching the Vertebrate division will include sequences in the Human, Mouse,
Rodent and Mammal divisions, so that *all vertebrate species* will be included in the search.

[E] Masking gives you the option of:

- **no masking** (default);
- **soft masking**, where the user puts repetitive sequences in lower case to exclude them from the search, but which still remain visible in the resulting alignments.

[F] Further options allows you to search new EMBL-Bank sequences only.

1. Enter the EMBL-Bank accession number 'AAA62278' into search box [A].
2. In the drop-down menu for Collection [C], select 'All Sequences'.
3. Click 'Submit Query'.

Sequences can be in any of the EMBOSS sequence formats, which includes FASTA [42] format. For a list of these formats, please see the EMBOSS User Manual [43].

CAUTION: Microsoft Word format is NOT a sequence format.

**Sequence search results page**

The results of a sequence search are listed in tables, where the matches for EMBL-Bank [7] and Ensembl [35] are separated so they can be dealt with independently (Figure 19).
Figure 19. Example of a results page for a sequence search using EMBL-Bank CDS accession ‘AAB07223.1’ against both EMBL-Bank and Ensembl.

[A] Search Completion Bar goes green as the search proceeds.

[B] Query Sequence Details provides the name, description and length of the search sequence.

[C] Ensembl Results are separated from [D] EMBL-Bank Results so they can be dealt with independently.

[E] Filter allows you to reduce search results using a text-based filter.
   • For example, filtering [13] using the term ‘Homo sapiens’ will return only human entries.
• Note that there are separate filters for Ensembl and EMBL-Bank results.

[F] Show alignments displays the alignments for the matches in the table (Note: click on 'Next' to view further results).

A closer look at the sequence search results page

Taking a closer look at the search results, you will see that they are ordered by e-value [44], starting with the lowest e-value which is the most significant match (Figure 20).

![Figure 20. A close-up of the Ensembl table in the advanced search results.](image)

[A] Select columns allows you to hide/show columns in the table.

[B] Alignment Length column shows the length of the matching region between the query and target sequences.

[C] Target Length column shows the length of the target sequence.

[D] Identity (%) column shows the % of nucleotides that are identical in the query and target sequences.

[E] e-value column displays a calculated estimate of the significance of the match.

It is better to use the e-value as a measure of how significant a match is rather than % identity, because the e-value takes account of the database size and the length of the query sequence, in addition to the number of matching nucleotides.

Partial and full-length matches: to determine whether a match covers the full length of the query and/or target sequence, or only part of the sequence (partial match), take a look at the Query Sequence Length (Figure 19 [B]), the Target Sequence Length (Figure 20 [E]) and the length of the match between them (Figure 20[D]). In addition, viewing the alignments (by expanding the description line, Figure 20 [C], or by expanding ‘Show all alignments’, Figure 22 [C]) will show you where the query and target sequences match.

Additional sequence search tools
Additional search engines are available on the [Sequence Similarity Searching](https://www.ebi.ac.uk/training/online) page. These include [BLAST](https://www.ebi.ac.uk/training/online), [PSI-BLAST](https://www.ebi.ac.uk/training/online), [FASTA](https://www.ebi.ac.uk/training/online), SSEARCH and specialised search programs. If you wish to search the entire [EMBL-Bank](https://www.ebi.ac.uk/training/online) and/or [Ensembl](https://www.ebi.ac.uk/training/online) database, then the fast Exonerate search on the ENA browser is your best option.

However, you might want to consider trying one of the search programs on the Sequence Similarity Searching page if you want to:

- search a **subsection of EMBL-Bank** not available through the ENA browser search (Figure 21);
- search **other databases** with nucleotide information, for example patent, structure or immunoglobulin databases;
- carry out a **specialised type of search**, for example searching a protein database with a DNA sequence (FASTX or BLASTX), or searching with a set of short [oligonucleotide](https://www.ebi.ac.uk/training/online) sequences (FASTM);
- search with a **very short nucleotide sequence**, where a true Smith-Waterman program such as SSEARCH would perform better;
- be able to adjust **search parameters** to fine-tune your search query.
European Nucleotide Archive: Using the primary nucleotide sequence resource
Published on EMBL-EBI Train online (https://www.ebi.ac.uk/training/online)

[Diagram of database selection]

- **Database Information**
- **EMBL-Bank**
- **FASTA validated database**
- **Search for FASTA related database in EMBL**

**FASTA - Nucleotide Fragment Search**
This tool provides nucleotide similarity searching against protein databases using the FASTA programs.

**Use this tool**

**STEP 1 - Select your databases**

- **NUCLEOTIDE DATABASES**
  - EMBL-Bank
    - EMBL Release
    - EMBL Updates
    - EMBL Coding Sequence
    - Others
    - IMGT

**STEP 2 - Enter your input sequence**

- Enter or paste a DNA sequence in any supported format.
- Or upload a file.

**STEP 3 - Set your parameters**

- **PROGRAM**: FASTA
  - The default settings will fulfill the needs of most users and, for that reason, are not visible.
  - More options...

**STEP 4 - Submit your job**

- Be notified by email (Tick this box if you want to be notified by email when the results are available)

[Browsable database](https://www.ebi.ac.uk/training/online)
Figure 21. Sequence Similarity & Analysis search page detailing the selection of databases available, including EMBL-Bank subdivisions.

[A] Databank selection includes all the taxonomic divisions and data classes from EMBL-Bank, plus additional databases such as:

- Immunoglobulin databases IMGT/LIGM-DB, IMGT/HLA, IPD-KIR & IPD-MHC;
- Human genome variation database HGVS/DB;
- Patent sequence databases NR Patent DNAs Level-1 & Level-2;
- Nucleotide Structure Sequences.

[B] Databank Selection (close-up) highlighting the different subdivisions of EMBL-Bank that can be queried.

[C] EMBL-Bank Taxonomic Divisions allows you to restrict your search to a specific taxonomic division, for example 'EMBL VRT' (EMBL-Bank vertebrate sequences).

[D] EMBL-Bank Divisions/Classes Cross-section allows you to restrict your search to a specific taxonomic division AND a specific data class, for example 'EMBL GSS VRT' (EMBL-Bank GSS class of vertebrate sequences).

- The taxonomic division/data class cross-sections are only available through Sequence Similarity & Analysis tools.

[E] EMBL-Bank Data Classes allows you to restrict your search to a specific data class, for example 'EMBL GSS' (EMBL-Bank GSS class).

A full course on Sequence Search Strategies will be here soon.

How to search ENA with taxonomy

How to use the taxonomy portal

To look for information on what sequence is available for a species, the taxonomy portal allows easy navigation via a taxonomic tree and a summary of the sequence available at each taxonomic node (see Navigating the taxonomic tree). The taxonomy portal allows you to look at the total coverage for any organism, or for any node in the taxonomic tree.

When doing a text search using taxonomy, it is best to use the scientific taxonomic name as it is more precise. However, you can also use a common name, but you are more likely to get a range of different taxa. For example, Figure 22 shows a query on the common name 'honey bee', which returns results for four different taxa.
Figure 22. Results of an ENA browser text search on ‘Honey bee’; taxonomy results are found in the ‘Other’ section.

[A] Taxa results provides the taxonomic portal summaries.

[B] By expanding the Taxa results, you get a list of closest matching taxonomies.

A closer look at the taxonomy portal

By expanding the 'Taxa' results' section, you can see a summary of the nucleotide information available for a taxon (Figure 23):
Figure 23. Taxonomy portal detailing the nucleotide information available for *Apis mellifera* (honey bee).

[A] Taxonomy for which information is displayed.

[B] **Taxonomy Portal** tab (current view) provides a summary of the nucleotide information available for [A].

[C] **Navigation** tab provides a taxonomy tree so you can navigate between taxonomy nodes.

[D] **Genetic code** tab provides the translation [51] tables used to translate the coding sequences for this species.

[E] Summary of the nucleotide information available.

**Navigating the taxonomy tree**

The taxonomy tree also provides an easy way to explore what nucleotide information is available for related taxonomic groups (Figure 24):
Figure 24. Navigation tab showing the taxonomic tree for *Apis mellifera* (honey bee).

[A] The taxonomic tree displays the complete lineage of a taxon. A summary of nucleotide data is available for each node in the tree.

[B] Each node can be expanded in order to navigate to related taxa [41]. In this example, the node for *Apis* [genus] has been expanded.
The current position in the taxonomic tree is highlighted in black.

Troubleshooting

Restricting your search by taxonomy is a good way of cutting out unwanted data, especially if all you need are sequences from one or a few related species. However, you need to be careful that you don't exclude relevant data from your search. There are several points to consider (Figures 25-28):

How specific is the taxonomy you require?

ENA contains information on strains, varieties and breeds for many taxonomic groups, whether or not the sequences varies between them.

![Figure 25](source: https://www.ebi.ac.uk/training/online)

Figure 25. There are several dog sequences in EMBL-Bank; this one is for the Alsatian breed.

What if the sequence you require has no taxonomy associated with it?

ENA contains sequences for which no species has been identified, such as those from environmental studies, synthetic constructs, transgenics and patents. These are found under special taxonomic divisions [28] (see How is the data structured [52]).

![Figure 26](source: https://www.ebi.ac.uk/training/online)

Figure 26. EMBL-Bank entry displaying the source information for an unknown bacterium.

Could your sequence be classified in a different way?

Some sequences are difficult to classify and require caution when searching so as not to miss valuable data. For example, endogenous viruses are usually classified by the host organism in which they were sequenced, or as being viral if isolated and sequenced. Therefore, endogenous viruses should be searched under both the VRL (virus) division and the taxonomic division of the host organism.
Figure 27. EMBL-Bank entry of the endogenous virus gamma-3, which is classified as being from the organism *Canis lupus familiaris* (dog) in the MAM (mammal) division, because it was sequenced as part of the dog genome.

Are you looking to compare sequences from a group of related organisms?

Taxonomic divisions help divide the data into manageable chunks, but be careful which search engine you use:

- The **ENA browser** merges divisions together to provide results that are complete for a taxonomic group; for example, if you search ‘Rodents’ you will include ROD + MUS divisions.

- The **Sequence Search & Analysis tools** keep the divisions separate to allow more flexibility when searching; for example, if you search ‘EMBL Rodent’ you will only include the ROD division (to search ROD + MUS you must select both).

Does the organism you are interested in have any alternative names?

Organisms can sometimes have different accepted names, or synonyms, because they were referenced in the literature differently or because different resources use different taxonomic classifications; this is important when you link out to external resources.


Are you certain the taxonomic name for your organism is unique?

Sometimes different organisms can share the same taxonomic name (homonyms).
Figure 28. ENA taxonomy search reveals two organisms with the same genus species name: *Agathis montana de Laub* is a conifer tree, while *Agathis montana Shest* is a wasp.

Where is the taxonomic information in an entry?

When viewing an EMBL-Bank entry, it is good practice to check the Source Feature(s) section to ensure the sequence you are looking at is what you expect it to be (Figure 29). In this section you can also pick up valuable additional information, such as sample information, the origin of each region of a transgenic sequence, or notes on how a sequence was isolated, or where it occurs in the genome.

Figure 29. Source Feature(s) section of EMBL-Bank entry FR695060, which provides detailed information on where the sequence was isolated.

Finding old archived entries

Sequence Version Archive

In the section 'How sequence is assembled' [54], we saw that sequence entries are subject to change, for instance when a sequence changes from being partially to fully assembled, or when annotation [4] is added/changed for a sequence. The EMBL-Bank [7] database, which contains the assembled sequences in ENA, only displays the most recent version of a sequence and its annotation. However, older versions of EMBL-Bank entries can still be accessed through the ENA Sequence Version Archive [22] (SVA) (Figure 30) [5] [3]).
Figure 30. ENA Sequence Version Archive search for finding old versions of EMBL-Bank entries.

[A] Accession [38] Number or Sequence Version lets you search using an EMBL-Bank accession number (for example BN000065) or a sequence version (for example BN000065.2, which is version 2 of entry BN000065).

[B] Snapshot enables you to see a specific EMBL-Bank entry (which you must enter in [A]) as it looked on a particular date.

SVA results - querying an accession number

If you search on an EMBL-Bank [7] accession [38] number in the SVA, then you will get a list of all the version changes for that EMBL-Bank entry and the date they became public (Figure 31). You can select to view and/or save any of these versions. You can also cross-compare different versions to see exactly what has changed.
Figure 31. SVA search results displaying a list of all the versions of EMBL-Bank entry BN000065.

[A] View allows you to see a snapshot of the EMBL-Bank entry at the date listed.

[B] Check two or more versions and click [C] Compare Selected to see a comparison of the selected versions of the entry, with all the differences highlighted.

[D] You can save old versions of an entry in EMBL or Fasta [42] formats.
SVA results - comparing different versions of an entry

If you select to compare two or more versions of an EMBL-Bank entry, the SVA will display the full entry with everything that has been removed or inserted highlighted, so you can quickly see what changes have occurred (Figure 32).

**Figure 32.** SVA search results showing a comparison of two versions of EMBL-Bank entry BN000065 (14-NOV-2006 and 05-OCT-2004), where the lines inserted or removed are highlighted.
[A] SVA search, where two different versions of entry BN000065 have been selected (red arrows).

[B] SVA results showing which lines have been added to the entry highlighted in green [C], and lines that have been removed from the entry highlighted in orange [D].

Exploring an EMBL-Bank entry

Overview of an EMBL-Bank entry

EMBL-Bank [7] provides an easy-to-read view of the data, where information such as taxonomy and annotation [4] features are grouped into separate sections. In addition there is a graphical view of assembly and annotation features. EMBL-Bank also has a plain text view that is useful for programmatic access (Figure 33).
European Nucleotide Archive: Using the primary nucleotide sequence resource
Published on EMBL-EBI Train online (https://www.ebi.ac.uk/training/online)
Figure 33. EMBL-Bank entry for BN000065; on the left is the default view and on the right the plain text view.

[A] In the default view, the **entry summary** provides information on the organism, [data class](#) [21], taxonomic division and sequence; you can also download sequence and change the view of the entry.

[B] **Navigation** provides links to other resources, including the taxonomy portal, [Ensembl](#) [35] and the [Sequence Version](#) [19] Archive (of old versions of the entry).

[C] **Overview** provides a graphical display of assembly and annotation data.

[D] **Source Feature(s)** provides information on the source of the sequence, such as the organism, organelle or country it was isolated in.

[E] **Other Features** provides detailed information about the function of different regions of the sequence.

[F] **Assembly** provides detailed information on how the sequence has been constructed from lower level sequences.

[G] **References** enable you to view the paper(s) citing the sequence and its annotation.

[H] **Sequence** can be used to search for similar sequences in the database.

[I] The text view of the same entry; this can be accessed by clicking on ‘TEXT’ in section [A]. This view is useful if you are writing programs as it provides all the line codes that identifies the line type; for example ‘DE’ identifies the ‘Description’ line.

### EMBL-Bank entry - General summary section

The top of an [EMBL-Bank](#) [7] entry provides a general summary of the data, the ability to change the view of the entry and to download information (Figure 34).  

Figure 34. EMBL-Bank entry BX548174 displaying general entry information.

[A] Accession [38] and Description of the entry (line codes AC and DE); 'BN000065.1' shows it is version 1 of sequence BN000065.

[B] View enables you to change the view to plain TEXT or XML [55], or just the sequence in FASTA [42] format.

[C] Download enables you to download the entry as TEXT or XML, or the sequence in FASTA format.

[D] Navigation bar allows you to jump to a specific section of the entry.

[E] Summary of entry data, including the date the entry became public and the date of the last revision (line code DT).

[F] Keywords can be used in a text search (line code KW).

[G] Secondary Accession(s) allow tracking of split/merged entries as well as entries used to construct a sequence (line code AC).

[H] Lineage provides the full lineage of the organism; clicking on any node of the lineage will take you the taxonomy portal (line code OC).

EMBL-Bank entry - Navigation section

EMBL-Bank [7] provides cross-references to almost forty other databases, including Ensembl [26], UniProtKB [56], InterPro [57], RFAM [58], WormBase [59], GrainGenes [60], dictyBase [61], FlyBase [62], VectorBase [63], GOA [64], PDB [65] and IMGT/HLA [66]. An entry will contain links to the external database(s) that have information on
the sequence, providing a valuable source of additional annotation [4] (Figure 35).

**Figure 35.** EMBL-Bank entry BX548174 showing Navigation section (DR line); each cross-reference [67] has a link to the relevant database entry, such as the RNA database Rfam.

[A] **Up arrow** allows you to navigate up to a higher level record

[B] **Tree** symbol allows you to navigate the taxonomy tree (line code FT)

[C] **Across arrow** provides a cross-reference database link (line code DR)

**EMBL-Bank entry - Overview section**

The overview section provides an at-a-glance graphical display of the assembly and annotation [4] features of the sequence (Figure 36). Annotation features describe where genes, mRNA, exons, introns, CDS [16] (coding sequence) and other features are located on the sequence. This information is supplied by the author, or occasionally as third party annotation (see section 'How is the sequence annotated [68]?).
Figure 36. EMBL-Bank entry BN000065 showing the Overview section.

[A] **Base range** enables you to zoom in to a specific region of the sequence.

[B] **Overview** shows the full length of the sequence as a grey bar, with a red box around the region being described below.

[C] **Assembly** shows the clones used in the assembly of the sequence (line code AS).

[D] **Source** describes the source of the sequence (line code FT).

[E] **Features** such as genes, mRNA, exon [69], CDS and intron [70] are shown relative to their position on the sequence (line code FT).
The source features section details where the sequence came from (Figure 37). For more information please see the section on 'How to search ENA with taxonomy' [71].

![Figure 37. EMBL-Bank entry Z71230 showing the Source Feature(s) section.](image)

[A] Taxon provides a link to the taxonomy portal, which provides a summary of all the sequence available for an organism (line code FT).

**EMBL-Bank entry - Other features section**

In addition to the graphical display of the annotation [4] features we saw in the Overview section, EMBL-Bank [7] also provides a detailed description of each feature in the 'Other Feature(s)' section (Figure 38). There are over fifty different features that can provide annotation for a sequence, and over seventy different qualifiers that help refine these features. Which features are described in a particular entry depends on the data the author submitted (ENA curators do not add features; they are provided by either the author or by third party annotation).
[A] **Base range** allows you to restrict the annotation features to those within a specific sequence range.

[B] **Show main features only** restricts the display to main features such as mRNA and **CDS** [16].

[C] **Features** describe the annotation for the sequence; there are >50 features, including CDS, mRNA, exon [69] and **intron** [70] (line code FT).

[D] **Qualifiers** refine each feature (line code FT).

In this example, the feature ‘CDS’ is further refined by the qualifiers ‘gene’, ‘product’ and **translation** [51].

[E] **Navigation** provides cross-references to other databases, including **UniProtKB** [72], **InterPro** [73] and **GOA** [74] (line code FT).

For a full list of the features and qualifiers available, please see [here](#) [75].
EMBL-Bank entry - References section

Literature references relating to the submitted sequence, including third party annotation [4], are provided in the reference list (Figure 39). Cited literature should be considered as a pointer to scientific information and not a credit for the elucidation of the sequence.

Figure 39. EMBL-Bank entry BN000065 showing the reference section.

[A] Abstract can be expanded to view the abstract.

[B] Links are provided to the full paper as a pdf, doi, html [76] or as a cross reference to CiteXplore [77].

EMBL-Bank entry - Sequence section

Either the full or part of the sequence can be viewed in FASTA format (Figure 40).
Figure 40. EMBL-Bank entry showing the sequence section.

[A] **Base range** allows you to restrict the sequence to a specified range.

[B] **Find similar sequences** will launch a sequence search on the displayed sequence.

[C] **Sequence** is shown in FASTA format (line code SQ).

**Exploring an SRA entry**

**Overview of an SRA entry**

The SRA (Sequence Read Archive [8]) provides access to unassembled data (Figure 41). **Accession** [38] numbers starting with 'ER' are assigned by EBI, while those starting 'SR' are assigned by NCBI and 'DR' by DDBJ. In other words, the original database the sequence was submitted to is identifiable by these two letter codes. This is important, because with SRA sequences submitted to EBI (ER entries) you can access both the original sequence as submitted by the author as well as normalized sequence data as 'Fastq' [79] Files'. By contrast, when using ENA to view SRA sequences submitted to NCBI (SR entries) or DDBJ (DR entries), you can only view the 'Fastq Files'.
Figure 41. SRA entry ERP000001.
[A] Summary provides general information on the submitted sequence and its source.

[B] Tabs section the SRA data for ease of handling:

- **Navigation** provides links to related SRA data (see next section);
- **Fastq Files** provides links to download normalized SRA sequence data (see 'Exporting SRA sequences and data [11]');
- **Submitted Files** provides links to download SRA sequence data as submitted by the author (see 'Exporting SRA sequences and data [11]');
- **References** provide links to literature cited in the entry.

[C] The tab shown is the 'Fastq Files' tab with cross-links to individual studies and sequence downloads.

More information on SRA entries can be found here [80].

**SRA entry - Navigation tab**

The Navigation tab allows you to explore all the Sequence Read Archive [8] data associated with an entry. In this way you can browse SRA data by sample, experiment or run, and even see the original submission data (Figure 42).

![Navigation tab](image)

**Figure 42.** SRA entry ERP000001 showing the Navigation tab.

[A] Down arrow allows you to navigate down through a tree in order to explore Sample, experiment or run SRA data.

[B] Back arrow allows you to navigate back to the original SRA submission data.

[C] For example, by following the link to ERS000001-ERS000012, you will have a list of the twelve SRA samples associated with entry ERP000001.
SRA entry - References

Literature references relating to the SRA sequence are provided under the References tab (Figure 43). Cited literature should be considered as a pointer to scientific information and not a credit for the elucidation of the sequence.

![Figure 43. SRA entry ERP000001 showing the References tab.]

[A] Abstract can be expanded to view the abstract.

[B] Links are provided to the full paper as a html [76], pdf, doi or as a cross reference to CiteXplore [77].

How to export sequence and download data

Exporting sequences and annotation

You can download single or multiple sequences, with or without their annotation [4], from any of the ENA databases, including:

- Downloading a single EMBL-Bank [7] sequence or full entry;
- Downloading multiple EMBL-Bank sequences or full entries;
- Downloading sequences or full entries from the taxonomy portal;
- Downloading SRA sequences and data using SRA-DataDownloader;
Bulk downloads using **ftp** [81].

More information on exporting data from ENA can be obtained from the [help pages](https://www.ebi.ac.uk/training/online). [82]

### Exporting single EMBL-Bank sequences and annotation

Once you have found the EMBL-Bank [7] entry you want, you can use the download links (Figure 44) at the top right of every entry page to easily download either:

- the sequence in **FASTA** [42] format;
- the full entry in either **TEXT** or **XML** [55] format.

![Figure 44. EMBL-Bank entry for BN000065 showing the download links at the top of every ENA entry page.](image)

[A] **TEXT** enables you to download the full entry in flat file format.

[B] **FASTA** enables you to download the sequence in FASTA format.

[C] **XML** enables you to download the full entry in XML format.

### Exporting multiple EMBL-Bank sequences and annotation

Alternatively, you can download multiple EMBL-Bank [7] sequences or full entries either by:
Following the links from the search page results;
Uploading your file of accessions.

You have the option of selecting the range of entries to download, which is particularly useful if your search query returns a large number of results (Figure 45).

Figure 45. Results page from a text search on 'human' displaying the download options.

[A] TEXT, XML [55], FASTA [42] download options available for all the sequence or full entries displayed in the results.

- Note: that there separate download capabilities for Assembled Nucleotide Sequences, Raw Nucleotide Sequences, Projects and Taxa [41].

[B] From-to range enables you to select which range of sequences or entries you wish to download.

- In this figure, there are 140885 results, therefore you may want to select fewer results to download.

[C] Upload file of accessions enables you to upload a file containing a list of accessions, which is then displayed as a list of resulting entries; you can then follow the download links [A] and data range capabilities [B] explained above.

Exporting sequences using the taxonomy portal

The ENA browser enables taxonomy from any node of the taxonomic tree to be downloaded in XML [55] format (Figure 46). In addition, bulk taxonomy downloads are possible through the ftp [81] site.
Figure 46. Taxonomy Portal displaying the download options.

[A] **XML** enables you to download all the taxonomy data in XML format.

[B] **Navigation** enables you to navigate the taxonomy tree, where data from each node of the tree can be downloaded using the XML link in [A].

Taxonomy data can also be downloaded in bulk from the **ftp** site. More information about the taxonomy data available for ftp download can be viewed on the [ENA help pages](https://www.ebi.ac.uk/training/online).

**Exporting SRA sequences and data**

The ENA browser enables a range of options when downloading raw sequence data from the [Sequence Read Archive](https://www.ebi.ac.uk/training/online) (SRA) (Figure 47). SRA data can be downloaded in normalised **fastq format** [85], or in the original format as submitted by the author. The SRA data can be grouped by study, sample, experiment, run or submission, where each group of sequences can be downloaded separately. Because these sequence files are often very...
large, in addition to ftp [81] download, the ENA browser enables downloading using the high-speed file transfer software Aspera [86]. Alternatively, you can upload SRA data into the Galaxy [87] platform.

**Figure 47.** ENA browser displaying SRA search results and the download capabilities.

(A) **Bulk download Fastq/Submitted files** provides the ability to select and download multiple files at once.

NOTE: this is the BEST route to download SRA sequences as you can choose which files to download (see Figure 48).

(B) **Fastq Files** provide SRA sequences in normalised fastq format.

(C) **Submitted Files** are the unaltered SRA sequence files submitted by the author.

(D) ftp, Aspera, Galaxy download links provide the ability to:

- use FTP or Aspera to download individual submitted files and Fastq files;
- upload individual Fastq files to Galaxy.

ADVICE: please consider using ‘Bulk download Fastq/Submitted files’ to download multiple files at once.
Figure 48. SRA-Filedownloader allows you to select the files you want to download.

1. Select SRA entry: type the accession [38] number into box [A] and click [B] Search.

2. Select data format: you can select either normalised fastq files [C] or unaltered files as submitted by the author [D].

3. Select files: you can either select specific files by checking the appropriate boxes [E], or choose to select all files [F].

   4. Download files [G].

Note: you will need to download Aspera first for the Aspera link on this page to be active. If you do not have Aspera, then a link to the Aspera download site will be provided at the bottom of the ENA browser page.

For information on SRA formats, see the ENA help pages [88].

For more information on downloading SRA data, see the ENA help pages [89].

There is a tutorial [90] on retrieving Sequence Read Archive data.
Bulk downloads using ftp

ENA data can be downloaded in bulk by ftp [81], including data from EMBL-Bank [7], SRA, Trace Archive and taxonomy data (Figure 49).

For example, EMBL-Bank sequences that can be directly downloaded from the ftp [91] site include:

- the entire EMBL-Bank release;
- new and updated entries made available after the latest release;
- specific data classes, such as Coding Sequence (CDS [16]) Whole Genome Shotgun (WGS), Mass Genome Annotation [4] (MGA) or the Construct (CON).

Figure 49. ftp site for the download of data from the EMBL-Bank full release.

More information on the EMBL-Bank data available through ftp can be obtained from the help pages [92].

The main directories (in flat-file format) for EMBL-Bank are:

EMBL-Bank full release:

ftp://ftp.ebi.ac.uk/pub/databases/embl/release/ [93]
EMBL-Bank entries released/updated after release:

ftp://ftp.ebi.ac.uk/pub/databases/embl/new/ [94]
ftp://ftp.ebi.ac.uk/pub/databases/embl/wgs/ [95]
ftp://ftp.ebi.ac.uk/pub/databases/embl/con/ [96]

EMBL-CDS (coding sequence) entries:

ftp://ftp.ebi.ac.uk/pub/databases/embl/cds/ [97]

Accessing ENA data programmatically

Programmatic access using ENA browser REST URLs

Using ENA browser REST [98] URLs, a wide variety of data is accessible in a variety of different formats. Single or multiple identifiers (including data ranges) can be used to retrieve up to 100,000 records at a time, which can be gzip-compressed or uncompressed. It is also possible to request specific taxonomy, or to retrieve archived versions of the data.

Examples

Here are some examples of what can be achieved through ENA browser REST URLs:

- Retrieve EMBL-Bank records in XML [55] or flat file formats;
- Retrieve EMBL-Bank records using sequence versions;
- Retrieve EMBL-Bank graphical images;
- Retrieve Taxon records in XML or Darwin Core XML formats;
- Retrieve a list of SRA submitted files or FASTQ [79] files;
- Retrieve SRA metadata [99] in XML format;
- Retrieve Trace sequences in FASTA or FASTQ formats;
- Retrieve Trace metadata in XML format.
More information on programmatic access to ENA data can be obtained from the ENA help pages [82].

More information on ENA Browser REST URLs can be obtained from the ENA help pages [100].

**Programmatic access using Dbfetch**

Dbfetch provides an easy way to retrieve data from multiple databases, including ENA, in a consistent manner (Figure 50). Dbfetch can be used from any web browser, as well as within a web-aware scripting tool that uses wget, lynx or similar.

![Figure 50. EBI Dbfetch tool showing the range of databases and data formats available.](image-url)
[A] Databases within Dbfetch include EMBL-Bank [7], the EMBL-Bank data classes CDS [16] (coding sequence) and CON (constructed), EMBL-SVA (sequence version [19] archive) and Ensembl [35], as well as a variety of protein and immunoglobulin databases.

[B] Search items takes EMBL-Bank accessions, including accession [38] ranges.

[C] Formats available for EMBL-Bank include EMBL, FASTA [42] and EMBL-XML [55] formats; other formats are available for different databases.

[D] Click Retrieve to obtain ENA sequence or entry data.

More information on Dbfetch is available here [101].

More information on Dbfetch URL syntax can be found here [102].

How to submit data to ENA

Types of submissions

ENA accepts a range of nucleotide data, including raw sequence reads, assembly information and functional annotation [4]. All submitted public data is exchanged on a daily basis between ENA, NCBI and DDBJ as part of the International Nucleotide Sequence Database Collaboration (INSDC). Therefore, no matter which databank you submit to, your data will appear in all three.

Submissions of assembled data to EMBL-Bank [7] and raw unassembled data or large-scale metadata [99] to SRA are handled through the Webin system [6 [3]]. Webin provides an easy-to-use format that guides you through the
submission process. All sequence submissions are processed through secure lines:

- Submissions are password protected and can be done through a secure encrypted https network;
- Data can be held confidentially prior to publication.

For more information on submitting data to ENA, see the help pages [103].

How to access the submission pages

You can access the submission page from the ENA browser homepage using the link shown by the red arrow in Figure 51, which will take you to the EMBL-Bank [7] Webin and SRA Webin submission page (Figure 52).

![Figure 51. The submission page can be accessed from the ENA browser homepage.](image-url)
Figure 52. The EMBL-Bank Webin and SRA Webin submission page.

[A] Login or register provides access to EMBL-Bank Webin; first-time submitters must use this link to register online.

[B] Login provides access to SRA Webin once you have registered; first time submitters must contact datasubs [at] ebi.ac.uk [C] to register.

Webin - EMBL-Bank submissions

EMBL-Bank [7] accepts sequence submissions through the Webin system.

To use EMBL-Bank Webin, you must first register on the log-in page [104].

Webin has a number of advantages:

- Webin uses pre-formatted templates that allow you to select the submission forms that best suit the type of data you have;

- Quality checks ensure data is as complete and accurate as possible.

More information on submitting data to EMBL-Bank can be found [here](#) [106].

**Webin - SRA submissions**

The SRA Webin system accepts both sequence reads and analysis data generated by next-generation sequencing methods.

To use SRA Webin, you must first register by contacting [datasubs@ebi.ac.uk](mailto:datasubs@ebi.ac.uk); registered users can then log-in using the [log-in page](#) [107].

Note: SRA only accepts data which is intended for unrestricted release. Controlled access data should be submitted to the European Genome-phenome Archive ([EGA](#) [108]).

More information on submitting data and [metadata](#) [99] to SRA can be found [here](#) [109].

There are tutorials available on: [Webin submission](#) [110]

**Summary**

- **Why do we need ENA?** ENA is a comprehensive archive of original nucleotide sequence data and its [annotation](#) [4].

- **What is ENA?** ENA provides a browser and underlying databank. ENA consists of 3 databases: EMBL-Bank [7] for assembled sequence and its annotation; Sequence Read Archive [8] (SRA) for unassembled (raw) data generated by Next Generation sequencing [9] methods; Trace Archive (TA) for unassembled data generated by capillary sequencing methods.

- **Where does the data come from?** Data is submitted as unassembled or assembled sequence directly to one of the three INSDC databases: ENA, NCBI or DDBJ. All data is exchanged daily between these databases.

- **How is the sequence assembled?** Assembly is performed by the submitter, not by ENA curators.
- **How is the sequence annotated?** Annotation is usually submitted along with the sequence, but ENA does take third party annotation as long as there is a published reference.

- **How is the data structured?** ENA is divided into data classes and taxonomic divisions [28]. A sequence is assigned to the most specific taxonomic division. There are special divisions for sequences not associated with any conventional taxonomy.

- **How do you access ENA?** You can access the ENA browser through the EBI homepage [111] or by going directly to ENA [40].

- **How can you search ENA?** You can search ENA through the ENA browser [40] or the Sequence Search & Analysis [112] tools. You can search with gene names, disease names, keywords, accession [38] IDs or sequence.

- **How do I download data from ENA?** EMBL-Bank sequences, taxonomy and annotation can be exported by the ENA browser REST [98] URLs or by ftp [81]. SRA and Trace Archive data can be downloaded by ftp or Aspera [113].

- **How do I submit data to ENA?** Assembled and annotated sequences can be submitted to EMBL-Bank using the Webin template submission forms or through pre-prepared EMBL-Bank flat files [114] for complete genomes and long sequences. SRA sequences can also be submitted to ENA.

### Guided examples of using ENA

The examples allow you to revise your knowledge gained on this course by providing guided examples of how the ENA can be used. The following scenarios demonstrate some uses of the ENA browser:

- Finding a protein-coding sequence;
- Exploring taxonomy;
- Finding genes using a sequence.

### Finding a protein-coding sequence

#### Search for a gene

**Background**
Human \textit{BRCA1} (breast cancer 1, early onset) is a tumour suppressor gene coding for a protein involved in DNA repair. Mutations in the \textit{BRCA1} gene in the germ line can result in the individual developing \textit{Hereditary Breast and Ovarian Cancer Syndrome}.

\textbf{Scenario}

Imagine you are working on the \textit{BRCA1} gene, studying the effects of different genetic mutations on the function of the encoded protein. You need to know the coding sequence of human \textit{BRCA1}, and of any known variants, so you can make targeted mutations to the gene sequence that will code for alternate amino acids in the protein. You can use ENA to search for the \textit{BRCA1} coding sequence.

![Search ENA](image)

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{search_ena.png}
\caption{To search ENA, type 'BRCA1 AND Homo sapiens' into the text search box.}
\end{figure}

1. Open the ENA Browser in a new window.

2. Type the search term \textbf{BRCA1 AND Homo sapiens} into the text search box (this is a more stringent search than typing ‘BRCA1 AND human’).

3. Click ‘Search’ to obtain search results.

\textbf{Results - ENA Browser summary}

The ENA Bowser provides a summary of the nucleotide data available for the human \textit{BRCA1} gene in both the \textbf{EMBL-Bank} (assembled/annotated data) and SRA (raw data) databases (Figure 53). You can now explore the data further by expanding and viewing the EMBL-Bank protein-coding sequences available for this gene.
Figure 53. Results page showing the ENA Browser summary of the assembled (EMBL-Bank database) and raw (SRA database) sequence data available for human BRCA1; the EMBL-Bank database 'Protein-coding Sequence' section has been expanded.

1. Expand the **Protein-coding Sequences** [A] results by clicking on '+’ [B]

Note: because the 'Protein-coding Sequences' section has already been expanded in Figure 52, it displays a ‘-’.

2. You should now have an expanded view [C] of all the coding sequence available for BRCA1.

3. Expand the entry for **AAC37594** (circled) by clicking on '+’ [D].
Note that many reads in the SRA database are likely to contain sequence from the BRCA1 locus, but will not have been annotated as such and will, therefore, not appear in the results of this search.

Results - obtaining BRCA1 coding sequence

Within one of the EMBL-Bank [7] database entries for the BRCA1 gene, you can download the sequence or browse the biological annotation [4] available (Figure 54). Note that there are several protein-coding sequences available for human BRCA1 in the EMBL-Bank database, which a researcher would want to view.
1. To download the BRCA1 coding sequence, click on 'FASTA' in the top right-hand corner of the entry [A].

2. You should now have a pop-up text file containing the sequence in FASTA format.

3. Delete the text file. Although this file could be saved for later use by a researcher, we were only viewing it as a demonstration.

3. To explore the BRCA1 protein-coding sequence, take a look at the biological annotation available in the graphical display [B].
Exploring taxonomy

Search with taxonomy

Background

*Homo sp. Altai* (Denisova hominin) is an ancient hominin that was present in Eurasia at the time modern humans evolved 200,000 years ago. Denisovans share a common origin with Neanderthals, but are believed to have distinct morphological features and a separate evolutionary history. Mitochondrial DNA has been successfully sequenced from a finger bone and a tooth found in the Denisova Cave in Russia.

Scenario

Imagine you are digging in the Altai Mountains of southern Siberia, and in dark cave you find a fragment of skull bone. From the artefacts found near the bone, you believe you have found evidence of the Denisova hominin, but you can’t be sure as past inhabitants of the region include Neanderthals and modern humans. You manage to isolate mitochondrial DNA and you want to compare it with other hominin species. You use EMBL-Bank [7] to see what ancient DNA sequences from the genus *Homo* are available.

![Image of ENA search interface](image.png)

**Figure 55.** To search ENA for *Homo sp. Altai* sequence, type 'Homo Altai' into the text search box.

1. Open the ENA Browser [40] in a new window.
2. Type the search term *Homo Altai* into the text search box.
3. Click ‘Search’ to obtain search results.

Results - the Taxonomy Portal tab

The Taxonomy Portal provides a summary of the nucleotide data available for *Homo sp. Altai*. You can now explore the data further by viewing the annotated nucleotide sequences or protein-coding sequences available, or...
you can use the Navigation tab to see what other ancient genomes are present in ENA.

Figure 56. Results page showing the Taxonomy Portal summary of the sequence data available for Homo sp. Altai.

[A] Assembled Nucleotide Sequences ([EMBL-Bank](#)) shows all the annotated nucleotide and protein-coding sequences available for Homo sp. Altai.

[B] Portal tab provides a summary of all nucleotide data available for Homo sp. Altai, including unassembled data.

1. Expand the Taxa [41] results to view the Portal tab displayed above.

2. Click on the Navigation tab.
Results - the Navigation tab

With the Navigation tab, you can easily browse the taxonomic tree. The results show that there is nucleotide data available for three species in the genus Homo: *Homo sp. Altai*, *Homo sapiens* and *Homo sapiens neanderthalensis* (Figure 57).

![Figure 57. The Navigation tab displaying the taxonomic tree for Homo sp. Altai.](image)

[A] Navigation tab displays a taxonomic tree for Homo sp. Altai, which can be navigated by expanding/collapsing the nodes in the tree.
The Homo [genus] node has been expanded to show all the species in this genus with nucleotide data: Homo sp. Altai, Homo sapiens and Homo sapiens neanderthalensis.

Finding genes using a sequence

Finding a gene

Background

If you do not have a gene name or an accession number for your sequence of interest, you can use the sequence search facility of the ENA Browser to identify the closest matches in ENA. Sequence searching is also useful for finding potential homologues that are related to your gene of interest.

Scenario

Imagine that you have isolated a gene associated with an autoimmune disease in humans. You know part of the DNA sequence and that mutations in the sequence are associated with a T cell-mediated autoimmune disease, but you do not know the gene it belongs to. You can use the ENA Browser's sequence search to see the closest matches in the ENA database.

The sequence you identified is:

TGGAAAGATAATTAATAAGACATGGGAAATGGGA
AGCTGATAACGTGGGGAGAGGTTTTGCTTGTGTTTC
ACCAAGAGAAAATCAGCTTCCTGTTTGGATACCCACT
AAACATTTGAAGTTTACAAATGAACCCATCACAGATG
CAAAGAAAGCGCCTCCACCGAG
Figure 58. To search ENA, copy/paste your sequence into the 'Sequence Search' box (red arrow) and click search.

1. Open the ENA Browser in a new window.

2. Copy/paste your sequence into the Sequence Search box.

3. Click ‘Search’ to obtain search results.

Results - closest sequence matches

Our search will provide matches in both ENA and Ensembl, so you can look at the sequences and their annotation in ENA, then look at the Ensembl results to see the alignment of your sequence to the genome(s).

Note: ENA Browser’s sequence search does not search raw sequences in either SRA or the Trace Archive, because the very short, redundant nature of raw sequences make sequence searching very difficult.
Figure 59. Close-up of the sequence search results page displaying the closest matching sequences in EMBL-Bank. The top hit is AF333072 (red arrow), which shows 100% identity (i.e. all the nucleotides in the query sequences align with the target sequence).

1. Scroll down to the EMBL-Bank results.
2. Click on the entry AF333072.

Results - analysing the results

Entry AF333072 is described as *Homo sapiens* HERV-K18, but HERV-K18 is a virus so what's going on?
**Figure 60.** Close-up of results page of EMBL-Bank entry ABF333072 showing a graphical overview of the annotation available for this sequence and the source of the sequence.

[A] The graphical **Overview** section summarises the features associated with this sequence: gag, pol and env genes.

[B] **Source** of the sequence is *Homo sapiens*.

Although it is a human sequence, the **gag**, **pol** and **env** genes are usually associated with viruses, so why is this sequence not classified as being viral?

Genomes often contain 'foreign' DNA, such as endogenous viruses (ERVs) and transposable elements. ERVs are thought to have arisen from ancient viral infections, but through the course of evolution have remained permanently integrated within their host genome to be passed down to subsequent generations. When you sequence a genome and find these 'foreign elements', how do you classify them? Are they part of the host genome or are they separate entities?

How a sequence is classified depends on the origin of the sequence.
If the virus was isolated and sequenced, then it would be classified as a viral sequence (VRL taxonomic division).

However, if the viral sequence was obtained from sequencing another organism, then it would be classified by the host organism.

In entry AF333072, the HERV-K18 endogenous virus was inserted into intron [70] 1 of the human CD48 gene (see Note encircled in red in Figure 59), and the sequence was obtained from sequencing the human genome, therefore it is classified as human (HUM taxonomic division). On average the human genome contains 25-50 copies of endogenous HERV type K retroviruses.

When searching for endogenous viral sequences, be careful not to restrict your search to just the VRL taxonomic class. Some may be under the VRL division, but others might be under their host taxonomy.

Exercises

The exercises allow you to apply your knowledge gained on this course. You will be asked how ENA can be used to solve given tasks. You can start by clicking on one of the exercise titles provided.

For hints, click on 'Need some help?'

Check your answer by viewing 'How we did it'.

Finding orangutan sequence data
Scenario

Orangutans, literally 'men of the forest', are found in the rainforests of Asia. You are studying the differences between two different species of orang-utan, those from Borneo (*Pongo pygmaeus* species) and those from Sumatra (*Pongo abelii* species). You require nucleotide sequence data from both species.

Exercise

Use the ENA browser to find out what sequence data is available in ENA for both *Pongo pygmaeus* and *Pongo abelii*.

- How many protein-coding sequences are available for each species?
- Is there any data on *Pongo pygmaeus* x *Pongo abelii* crosses available?

Need some help?

1. Start by searching for one of the species.

2. Use the Taxonomy Portal 'Navigation' tab to find the organisms you want information on.

How we did it

1. Start with the [ENA browser](https://www.ebi.ac.uk/ena) Text Search on the homepage.

2. Type the search term *Pongo pygmaeus* into the text search box.

3. Under Other, expand the [Taxa](https://www.ebi.ac.uk/ena) results.
4. Now expand the results for taxa 9600: *Pongo pygmaeus* (Bornean organutan).

5. You should have the **Portal** tab open, which lists all the assembled and raw nucleotide sequences available for *P. pygmaeus*. Scan down to the line showing the details for the number of **Protein-coding Sequences** available.

![Figure 1A. Portal tab showing the number of protein-coding sequences available for *P. pygmaeus*.](image)

6. Now click on the **Navigation** tab. You should have the full lineage for *P. pygmaeus*.

7. Expand the node for *Pongo [genus]*. You can see that in addition to *P. pygmaeus* there is also sequence available for *P. abelli* and for the cross *P. abelli x pygmaeus*. 
Figure 1B. Navigation tab with 'Pongo [genus]' expanded. You can obtain sequence information from each node.

8. Click on **Pongo abelii: Sumatran orangutan**.

9. You should have the **Portal** tab open for *P. abelii*. Scan down to the line showing the details for the number of **Protein-coding Sequences** available.

Note: if you click on the number of protein-coding sequence entries, you will be provided with a browsable list of these entries.

**Finding information on transgenic sequences**
Background

Malaria accounts for over 2% of the world's deaths. Female mosquitoes of the *Anopheles* genus are the primary hosts as they require a blood meal to develop eggs, the males feeding on nectar rather than blood. When a female ingests infected blood, malarial sporozoites develop in her gut before migrating to the salivary glands. When the mosquito next feeds, malarial parasites in the saliva are transmitted into the skin. As resistance to anti-malarial drugs becomes an increasing problem, researchers have tried other approaches, such as trying to prevent malarial parasites from invading mosquito salivary glands in order to break the infection cycle.

Scenario

You want to investigate the interactions between mosquitoes and malaria-causing parasites. You found an interesting paper detailing work done with *Anopheles stephensi*, an Asian species of malarial mosquito, on salivary gland-parasite interactions. You would like to look at the genetic constructs they used in order to design your own experiments involving other species from the *Anopheles* genus. For this, you need to know the details of the genetic construct they used. The paper mentions that the nucleotide data for their transgenic construct was deposited to GenBank under the accession number 'AB212907'.

- From which organism does each region of the transgenic sequence originate?
- What feature(s) are described for each sequence region?
- Give the nucleotide ranges for each sequence region.
- If the accession number 'AB212907' was not known, could you think of another way of finding this entry?

Need some help?

1. Use the accession number to begin your search.
2. Look at the graphical 'Overview' to see a summary of each sequence region.

3. Look at the 'Source Features' and 'Other Features' sections too.

How we did it

1. Start with the ENA browser [40] Text Search on the homepage.

2. Type the search term AB212907 into the text search box.

3. Scroll down to the graphical Overview section.

![Figure 2A. Graphical overview section of EMBL-Bank entry AA212907, where the blue box/arrow tells you the app promoter is from Anopheles stephensi, the red box/arrow tells you the DsRed CDS [16] (coding sequence) is a synthetic construct, and the green box/arrow tells you the Antryp1 terminator is from Anopheles gambiae.]

1. Scroll down to the Source Feature(s) section. This section describes the source of the sequence in more detail, including the nucleotide ranges for each region of the sequence.
Identifying a sequence

Background

Every year fungal diseases destroy valuable crops. In the 1970s, the USA almost lost its entire corn crop to fungal parasites, bringing economic disaster to many farmers. Much time and money is spent researching the development of new fungicides and disease-resistant transgenic plants.
Scenario

A farmer's corn field is destroyed by a particularly virulent but unknown fungal species. You manage to get a partial sequence of the gene you think is responsible for spread of the fungus. What does the culprit gene encode [115], and which species is it from?

The sequence you identified is:

```
GAAAAACTTTCTCAAAACAGCTGGCCGCCCCATTTAG
CTGAATTGACTGACTACCGCCCGCAGGACCTTCTTCT
TTCACTCCTCATTTGCTCTGGGCGAAATCACC
ATGCAGATTATAATGTGGTATACTCCTCCTTTTGC
GGCAGCAATGCTTCCAGTGGTTCACTCCTTAGGGATTA
ATTGTAGGGGCACTCCTGCTGTTATCCGGCGGG
AACCTTATGGTCCG
```

Need some help?

1. Search the given sequence using the ENA browser.
2. Look at the top hits in the results.

How we did it

1. Start with the ENA browser [40] Sequence Search on the homepage.
2. Copy/paste the entire sequence into the search box.
Figure 3A. Search results showing the closest matches to the sequence and the organism they come from, namely the Ustilago maydis virus P4.

1. Click on the top result L12226.

2. Scroll down to the Other Features section to get the name of the gene product.

Figure 3B. Other Features section of EMBL-Bank entry L12226 showing the gene product to be KP4 toxin.

Quiz 1: Understanding ENA

Questions: 13
Attempts allowed: Unlimited
Available: Always
Pass rate: 75%
Backwards navigation: Allowed

Your feedback

Please tell us what you thought about this course. Your feedback is invaluable and helps us to improve our courses and thus enhance your learning experience.

Get help and support on ENA
ENA helpdesk. You can contact the ENA helpdesk for technical support or enquiries at datasubs [at] ebi.ac.uk.

Submissions. Please see our submissions page [116]. You can contact ENA with submission queries at datasubs [at] ebi.ac.uk.

Mailing lists. For updates and publication notifications join our mailing group at ena-announce [at] ebi.ac.uk.

News. For news about recent updates, new documents and announcements, please see our News page [117].

FAQs. Get help from ENA by reading the website [40].

Publications. Read more about ENA data and analysis in publications [118].

Courses held at EBI. ENA takes part in several in-house courses [119] at the EBI.

Workshops. We go on the road with other EBI resources to run workshops worldwide in external training [120] events.

References

For a full list of ENA references, please see the ENA Publication [121] page.


Contributors
Jennifer McDowall

EMBL-EBI
Patent Informatics and Bioinformatics Trainer

Jennifer McDowall gained her PhD in Medical Genetics from the University of British Columbia, Canada. Jennifer has held several research positions in both academia and the pharmaceutical industry, as well as a faculty position at the Open University, Canada. She joined EMBL-EBI as a senior curator for the InterPro database after working on the Human Genome Project at the Sanger Institute, where she was part of the team that sequenced the human X chromosome.

Clara Amid

EMBL-EBI
Content Project Leader - Cochrane team: European Nucleotide Archive

Source URL: https://www.ebi.ac.uk/training/online/course/european-nucleotide-archive-using-primary-nucleotide-data-resources-ebi/how-export-sequence-and-download-data

Links
[1] https://www.ebi.ac.uk/training/online/trainers/amid
[2] https://www.ebi.ac.uk/training/online/glossary/allele
[3] https://www.ebi.ac.uk/training/online/course/nucleotide-sequence-data-resources-ebi/references
[4] https://www.ebi.ac.uk/training/online/glossary/annotation
[5] https://www.ebi.ac.uk/training/online/glossary/raw-data
[6] https://www.ebi.ac.uk/training/online/glossary/european-nucleotide-archive
[7] https://www.ebi.ac.uk/training/online/glossary/sequence-read-archive
[8] https://www.ebi.ac.uk/training/online/glossary/next-generation-sequencing
[9] https://www.ebi.ac.uk/training/online/glossary/capillary-sequencing-technology
[10] https://www.ebi.ac.uk/training/online/course/nucleotide-sequence-data-resources-ebi/how-export-sequence-and-download-data
[13] https://www.ebi.ac.uk/training/online/glossary/contigs
[14] https://www.ebi.ac.uk/training/online/glossary/curation
[15] https://www.ebi.ac.uk/training/online/glossary/cds
[17] https://www.ebi.ac.uk/training/online/glossary/curator
[18] https://www.ebi.ac.uk/training/online/glossary/sequence-version
[19] https://www.ebi.ac.uk/training/online/glossary/data-class
[21] https://www.ebi.ac.uk/training/online/course/nucleotide-sequence-data-resources-ebi/finding-old-archived-entries