European Nucleotide Archive: Quick tour

Clara Amid [1]

- DNA & RNA
- Beginner
- 0.5 hour

This quick tour provides a brief introduction to the European Nucleotide Archive, the EBI’s repository for nucleotide sequence data.

Learning objectives:

- Basic understanding of the European Nucleotide Archive and how it can help you to access and submit DNA sequence data
- Know where to find out more about the European Nucleotide Archive

What is the European Nucleotide Archive?

The European Nucleotide Archive [2] (ENA) provides a comprehensive, accessible and publicly available repository for nucleotide sequence data. The ENA attracts users from a multitude of research disciplines and serves as an underlying data infrastructure for other EBI services, including Ensembl [3], Ensembl Genomes [4], UniProt [5] and ArrayExpress [6]. Data submitted to the ENA are validated by automated quality checking and, where possible, manual inspection and curation [7].

The foundation for the ENA was the EMBL Data Library, which was established in EMBL Heidelberg [8] in the early 1980s (later renamed as the EMBL Nucleotide Sequence Database, EMBL-Bank [9]). Once started as a primary database [10] for assembled and annotated sequences, the ENA’s remit has expanded enormously in response to advances in sequencing technology and the broad applications of sequence data. The ENA now incorporates raw data [11] from electrophoresis-based sequencing machines as well as raw reads from next-generation sequencing platforms. By consolidating information from these three tiers, the ENA provides access to the whole scale of sequencing information: from raw data, through assembly and mapping information that relates very fragmented raw sequence reads into contigs [12] and higher order structures, such as scaffolds and chromosomes [12], through to high-level functional annotation [13] (see Figure 1).

Why do we need the ENA?

Nucleotide sequence information is crucial to our understanding of biology, from genetics and molecular interactions through to organism-wide processes. Free access to nucleotide sequence data is therefore essential for life science research, even for basic tasks such as primer design, comparing sequences to those in the public domain and gene expression [14] analysis. As large-scale sequencing becomes faster and cheaper, the need to deposit, search and analyse information in a central archive that is publicly available and easily accessible continues to grow.
Figure 1 The ENA’s three-tiered data architecture. Individual sequence reads are represented with submitted assemblies and read alignments. Assemblies are annotated with features such as genes and regulatory regions.


What can I do with the ENA?

- Permanently archive your sequence data and disseminate to the global research community.
- Share your pre-publication data with collaborators in multi-centred sequencing studies.
- Reduce your local hardware requirements for archiving next-generation sequence data.
- Locate, retrieve and aggregate existing sequence data for analysis and meta-analysis using the EBI Toolbox and third party tools.
- Report novel annotation relating to existing sequence as part of the Third Party Data policy.
- Browse existing sequence and annotation referred to in the literature.
- Find all sequence and annotation available for a gene of interest.
- Use sequence similarity to search data (including unassembled raw data) and find out what is known about your new sequence.
- Link through from nucleotide data to a host of integrated resources, such as genomes (Ensembl [3] and Ensembl Genomes [4]), the scientific literature (CiteXplore [16]), protein products (UniProt [17]) and protein families, motifs and domains (InterPro [18]).

Searching and getting data from the ENA

Search and download data

The ENA data are available through several routes (outlined below). For more information, see the ENA search and browse page [19].

Search from the ENA homepage

Carry out searches using accession IDs for nucleotide sequences, text searches (against annotation fields, such as gene symbols and culture collection identifiers), and rapid comprehensive sequence similarity searches.

Web browser

All types of data held in the ENA are available from the browser in integrated views, which include annotated sequences, projects, coding sequences, taxa and experiments in the Sequence Read Archive [22]. See more information about the ENA browser [23].

Parallel web services [24]

Programmatic data access [23] to these services is also provided.

FTP [25] server

For large-scale data downloads, there are FTP and Aspera [26] channels.

Search results
Submitting data to the ENA

Many journals and funders require authors to submit their sequence information to a database that is a member of the International Nucleotide Sequence Database Collaboration [27] (INSDC; see information on collaborations in get help and support on the ENA [28]) prior to publication. The advantage of submitting your sequence data to the ENA is that your data will be permanently available and readily accessible to scientists worldwide. After submission, accession [20] numbers are assigned to identify your sequence and any related information. These accession numbers should be included in your manuscript. You can choose whether to make your data publicly available immediately or wait until your paper is published.

Manual submissions

You can register new sequencing projects and submit assembled sequence and annotation [13] to the ENA using Webin, the EBI’s preferred web-based submission system. Webin provides interactive web forms that are tailored to the type of data to be submitted and that capture and validate the information required. Sequence and annotation can also be uploaded to Webin in several formats. Webin is available from the ENA’s submissions login page [29].

To submit small-scale raw sequence data, send an e-mail request to datasubs [at] ebi.ac.uk (submissions) and a secure Webin-box will be set up. Users can upload data and metadata [30] files (prepared in third party editors) into the Webin-box by FTP [25] or Aspera [26] and using a dedicated webpage.

Automated submissions

The EBI works closely with sequencing centres and other facilities to ensure the timely incorporation of data into
the ENA. A number of options are available for automated submissions, including submission accounts for annotated sequences, the RESTful web-based submission service for next-generation sequence metadata, as well as FTP and Aspera drop-boxes. All enquiries should be directed to datasubs [at] ebi.ac.uk (submissions).

Updating ENA content

Records can eventually become out of date; authors might need to make corrections to sequence and assemblies, or they might discover new features that need to be added through annotation. Because such findings are rarely published in journals, it is important that authors communicate their new findings to the ENA. Authors wishing to do so should use the update procedure available from the submissions page [31].

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 the ENA

EBI-EMBL, Hinxton, UK

Support

- For enquiries about submitting data to the ENA, please email the datasubs [at] ebi.ac.uk (submissions team)
- Further information can be found on the About the European Nucleotide Archive [32] page

References


Collaborators

The ENA achieves comprehensive coverage through partnership with other global bioinformatics service providers: the National Center for Biotechnology Information [38] (NCBI) in the US and the DNA Data Bank of Japan [39] (DDBJ). These three partners form the International Nucleotide Sequence Database Collaboration [40] (INSDC),
which has been underway for over a quarter of a century and now serves as a model for data sharing in the life sciences. Each of the three groups collects a proportion of the total sequence data reported worldwide, and all new and updated database entries are exchanged between the groups on a regular basis.

Principles of the INSxDC include the provision of free and unrestricted access to data for all users and the maintenance of permanently accessible records.

Funding

ENA is developed and operated under the support of the European Molecular Biology Laboratory (EMBL) and through grants from external bodies that include the Seventh Framework Programme of the European Commission (EC-FP7), the British Biotechnology and Biological Sciences Research Council (BBSRC) and the Wellcome Trust (WT).

Contributors

Guy Cochrane
EMBL-EBI
Team Leader: European Nucleotide Archive

Guy Cochrane received his PhD in cancer cell and molecular biology from the University of East Anglia in 1999 and carried out postdoctoral work in the molecular biology of photoreception at the University of Cambridge prior to joining EMBL-EBI. Guy is now the Team Leader for the European Nucleotide Archive (ENA). Under Guy's leadership, the team have launched important new services, such as the Sequence Read Archive for raw data from next generation sequencing platforms and the CRAMtools sequence data compression platform. His other work involves editing for several journals and meetings, and he has been involved in standardisation activities in many areas of bioinformatics.

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-quick-tour

Links
[1] https://www.ebi.ac.uk/training/online/trainers/amid