EMBL-EBI

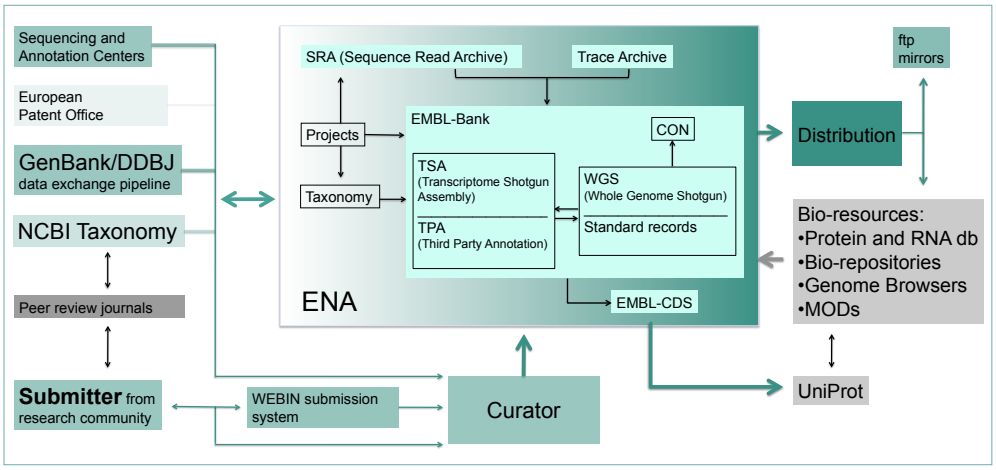
European Bioinformatics Institute
is an Outstation of the
European Molecular Biology Laboratory.

ENA
European Nucleotide Archive

ENA as an Information Hub

Petra ten Hoopen, Ruth Akhtar, Richard Gibson, Bob Vaughan, Nadeem Faruque and Guy Cochrane

ENA; http://www.ebi.ac.uk/ena/ ; datasubs@ebi.ac.uk



ENA data flow

The European Nucleotide Archive (ENA) accepts directly sequenced nucleic acid molecules from Sequencing and Annotation Centers, individual researchers or the European Patent Office. While for large data sets submission pipelines are created, submitters of small scale data sets use the WEBIN submission tool. Curators guide and help through the submission process. Deposited sequences are owned solely by the submitter and co-authors who only can give editorial rights to the ENA team. We do not have direct contact with editors of peer reviewed Journals. However, sequence records contain a reference to the citation where the submitted sequence data is published. This info is provided and updated by submitters or via GenBank in the USA. A pipeline for a data exchange and daily distribution between collaborating INSDC partners – ENA, GenBank and DDBJ - ensures availability of all sequences deposited at any of the three primary databases. Classification of sequenced organisms is approved by our advisers from the NCBI Taxonomy. Accessioned sequence records are distributed to our mirror sites and screened by other bio-resources, such as RNA and protein databases, genome collections and model organism services that use ENA entries as both source and supporting evidence for their records. Integration of the growing wealth of molecular information is a great challenge that brings opportunities for ENA to serve as a bioinformatics data information hub, allowing, through its provision of permanent identifiers for sequence and project records, community-recognized identifiers for navigation across databases.

ENA data content

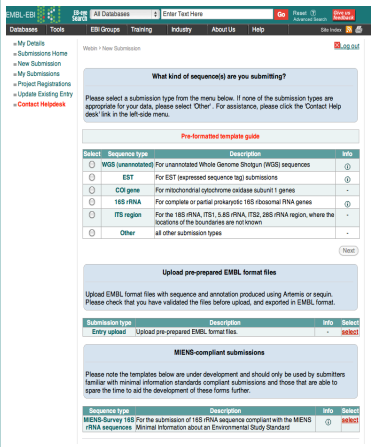
The major contributor to the sequence volume in ENA is the **Sequence Read Archive (SRA**) storing submissions of next-generation sequencing reads and corresponding metadata described in the study, sample and experiment. Traditional **EMBL-Bank** stores i/ partial or complete assembled nucleic acid molecules with functional annotation derived from a direct or third party experimental evidence (**Standard** and **TPA** data classes), ii/ assembled contigs (Whole Genome Shotgun assemblies – **WGS**), iii/ sequenced replicons with a partially known order of assemblies (**CON**) and iv/ assemblies of EST transcripts (Transcriptome Shotgun Assembly-**TSA**). **EMBL-CDS database** filters coding sequences from the EMBL-Bank. ENA also hosts the **European Trace Archive** maintained previously at the Wellcome Trust Sanger Institute, UK. Classification of each sequenced organism from every data class is verified from the **Taxonomy** database which is synchronized daily with the NCBI Taxonomy. Permanent Accession numbers are assigned to each EMBL-Bank sequence and each SRA object. Taxon-based large-scale sequencing projects are registered and assigned IDs that link data from EMBL-Bank and SRA. Graphical browsing and a new sequence similarity search facilitate a free and publicly available access to the ENA content.

Curator role

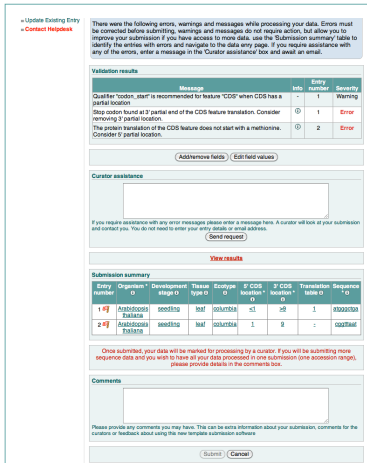
The curation team guides submitters through the submission process. They take the unique opportunity to obtain directly from submitting researchers exact provenance information on the sequenced sample and on the methodology surrounding its preparation for sequencing. Curators sort submitted data, fix errors and resolve taxonomy issues. They provide a helpdesk and generally mediate communication between the scientific community and ENA software engineers. Submitters contact the curators for updates of their sequence records. Curators also maintain the annotation guidelines and are involved in the data integration.

WEBIN submission system

A small scale submitter can create an account and login to the WEBIN submission system.
For frequent types of sequences submitters can choose from several pre-formatted templates with pre-selected features and qualifiers. A list of available templates is currently being extended. If none of the templates is suitable for the submitting sequences submitters are guided to the interactive web application with the full spectrum of INSDC agreed features and qualifiers used for more specific annotation. Alternatively, submitters can upload sequences as EMBL-formatted files. Pre-formatted templates or uploaded files are validated. Based on a feedback from the rule-based validator submitters are encouraged to correct their data prior to submission to the ENA.
Submitters can now ask for help also during the submission process. Curators can view data and advise about corrections before sequences are actually submitted.
For updates of existing entries submitters or co-authors of the cited publication can contact the ENA team for changes of the sequence, the annotation, the reference or the confidentiality of their sequence records.



Initial Webin submission page



Validator feedback page

Acknowledgements

Funding is provided by the European Molecular Biology Laboratory, the European Commission and the Wellcome Trust. Funding for open access charge: EMBL.

Petra ten Hoopen, PhD
Curator
ENA
petra@ebi.ac.uk

EMBL- EBI
Wellcome Trust Genome
Campus
Hinxton
Cambridge
CB10 1SD
UK

T +44 (0) 1223 494 444
F +44 (0) 1223 494 468
http://www.ebi.ac.uk