

Improvements to services at the European Nucleotide Archive

Rasko Leinonen^{1,*}, Ruth Akhtar¹, Ewan Birney¹, James Bonfield², Lawrence Bower¹, Matt Corbett¹, Ying Cheng¹, Fehmi Demiralp¹, Nadeem Faruque¹, Neil Goodgame¹, Richard Gibson¹, Gemma Hoad¹, Christopher Hunter¹, Mikyung Jang¹, Steven Leonard², Quan Lin¹, Rodrigo Lopez¹, Michael Maguire¹, Hamish McWilliam¹, Sheila Plaister¹, Rajesh Radhakrishnan¹, Siamak Sobhany¹, Guy Slater², Petra Ten Hoopen¹, Franck Valentin¹, Robert Vaughan¹, Vadim Zalunin¹, Daniel Zerbino¹ and Guy Cochrane¹

¹European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD and

²Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK

Received October 15, 2009; Accepted October 16, 2009

ABSTRACT

The European Nucleotide Archive (ENA; <http://www.ebi.ac.uk/ena>) is Europe's primary nucleotide sequence archival resource, safeguarding open nucleotide data access, engaging in worldwide collaborative data exchange and integrating with the scientific publication process. ENA has made significant contributions to the collaborative nucleotide archival arena as an active proponent of extending the traditional collaboration to cover capillary and next-generation sequencing information. We have continued to co-develop data and metadata representation formats with our collaborators for both data exchange and public data dissemination. In addition to the DDBJ/EMBL/GenBank feature table format, we share metadata formats for capillary and next-generation sequencing traces and are using and contributing to the NCBI SRA Toolkit for the long-term storage of the next-generation sequence traces. During the course of 2009, ENA has significantly improved sequence submission, search and access functionalities provided at EMBL–EBI. In this article, we briefly describe the content and scope of our archive and introduce major improvements to our services.

BRIEF HISTORY

ENA was established in the early 1980s as the EMBL Data Library (later renamed as the EMBL Nucleotide Sequence Database, EMBL–Bank) and focused initially

on richly annotated nucleotide sequences. After breakthrough improvements in sequencing technologies culminating in the wide-scale adoption of the chain-termination method developed by Sanger (1,2), a further function of the archive, initially operated by the Wellcome Trust Sanger Institute as the Trace Archive, was the storage of high-throughput sequence reads with associated quality and instrumentation information. The growth of the Trace Archive accelerated notably with the emergence of the shotgun approach as the method of choice for genome sequencing and increased further with the commercialization of highly parallel next-generation sequencing technologies first by Roche's 454 (<http://www.454.com/>) followed by Illumina's Genome Analyzer (<http://www.illumina.com/pages.ilmn?ID=204>) and Applied Biosystems' SOLiD System (http://www3.appliedbiosystems.com/AB_Home/applicationstechnologies/SOLiD-System-Sequencing-B/index.htm) (3). After inclusion of the Trace Archive and the establishment of the Sequence Read Archive (SRA) in 2008, an archival resource for next-generation sequences, ENA had completed its transformation into a comprehensive nucleotide sequence archive.

FREE AND UNRESTRICTED ACCESS

ENA, along with NCBI (4) and DDBJ (5), is an active member of the International Nucleotide Sequence Database Collaboration (INSDC), established to promote worldwide collaborative data exchange. The principal policy of INSDC is to provide free and unrestricted permanent access to all archived nucleotide data. All primary data in the INSDC belong to the submitters and can only be updated with submitter

*To whom correspondence should be addressed. Tel: +44 1223 494608; Fax: +44 1223 494468; Email: rasko@ebi.ac.uk

Table 1. ENA-Annotation and ENA-Assembly data classes

Data class	Description
Expressed sequence tag (EST)	High-throughput short transcribed cDNA (mRNA) sequences
Genome survey sequence (GSS)	High-throughput short genomic sequences
High throughput cDNA sequencing (HTC)	Unfinished cDNA (mRNA) sequences
High throughput genome sequencing (HTG)	Unfinished genomic sequences
Patent sequence (PAT)	Patent sequences
Sequenced tagged site (STS)	Short unique genomic sequences
Standard sequence (STD)	High-quality annotated sequences
Third party annotation sequence (TPA)	Re-annotated and re-assembled sequences
Transcriptome shotgun assembly (TSA)	Computationally assembled sequences
Whole genome shotgun (WGS)	Shotgun sequences
Constructed sequences (CON)	Sequence assemblies primarily from WGS sequences

Table 2. ENA-Reads data classes

Data class	Description
Trace Archive	Sequence traces with base, quality and intensity information from capillary sequencing instruments
Sequence read archive (SRA)	Sequence traces with base, quality and intensity information from next-generation sequencing instruments

consent. For full policy details, please refer to <http://www.insdc.org/page.php?page=policy>.

STRUCTURE

The ENA consists of ENA-Annotation, ENA-Assembly, and ENA-Reads tiers. The oldest records lie within ENA-Annotation and ENA-Assembly sections (Table 1). Capillary and next-generation sequence traces are included in ENA-Reads (Table 2). Capillary traces are stored in the Trace Archive and next-generation sequences in the SRA. Different data classes are designed to capture the full spectrum of nucleotide-sequence-related information starting from the sequencing experiments through complete assemblies and annotations up to high-level sample and project information. ENA-Annotation contains rich high-level functional annotation captured in the INSDC feature table format. ENA-Assembly is designed for efficient storage of assembly information and ENA-Reads for the efficient storage of sequence trace information. Entries from different data classes are connected together through high-level sample and project records to create rich linkage between different types of data.

CONTENT

In October 2009, ENA-Annotation and ENA-Assembly contained 163 million records covering 283 billion bases. Whole-genome shotgun sequences continue to be the dominant source of new sequences (30% sequences and 53% of bases) followed by expressed sequence tags (EST) (38% sequences and 12% of bases). The growth of the Trace Archive, part of ENA-Reads, is markedly reduced, increasing only 6.2% in the last year to 1.96 billion sequences and 1.77 trillion bases. The SRA, containing next-generation sequences, has rapidly grown to 83 billion spots covering 7.4 trillion bases, making the SRA the fastest growing section of ENA.

In ENA, the number of sequenced taxa has grown to 460 000 organisms and the number of scientific literature citations has exceeded 270 000.

IMPROVED INTERACTIVE SUBMISSION TOOL

We have made significant improvements to our interactive submission tool (Webin) with the addition of a new template-based system. Webin templates are text documents containing information common to large numbers of similar records and variable fields expected to be of use for a given data type. At the end of the submission process, submitted information is expanded using the template to create full database records. The Webin launcher, the entry point to all interactive submissions, has been extended to offer an appropriate set of common use case templates for submitters and to guide them through the submissions process.

Presently, we have configured templates for most commonly occurring types of ENA-Annotation submissions, including a MIENS (Minimum Information about an ENvironmental Sequence) standard compliant template, and we may add additional templates complying with other standards as they become available. We also plan to expand this system to cover SRA and project submissions. Upon submission and template expansion, the resulting entries are analysed with a rule-based validator and users are informed of any warnings and errors generated as part of the data validation process. All users wishing to submit large number of sequences with a fixed number of variable fields are encouraged to contact datasubs@ebi.ac.uk for creation of new templates which can be rapidly integrated into Webin. The Webin submission tool is available at <http://www.ebi.ac.uk/embl/Submission/webin.html>.

On the first page, users are asked to choose one of the available sequence submission types (Figure 1). This will determine which template will be used for submission.

What kind of sequence(s) are you submitting?

Webin will guide you through the submission process, from general questions about your data, to information about publications, and then the specific metadata and sequence information which can be entered or uploaded. You can return to previous pages during submission, and will also be given the chance to review your data at the end. For assistance, please click the **'Contact Helpdesk'** link in the left-side menu.

Select	Sequence type	Description
<input type="radio"/>	EST submission	This submission type is suitable for EST submissions
<input type="radio"/>	16S ribosomal RNA submission	For the submission of complete or partial 16S rRNA genes. The entire sequence presented is presumed to be 16S rRNA, with direction and partiality being determined automatically after submission
<input type="radio"/>	COI submission	This submission type is suitable for COI gene submissions
<input type="radio"/>	MIENS-compliant 16S rRNA	For the submission of 16S rRNA sequence compliant with the MIENS Minimal Information about an Environmental Study Standard

My sequence is not listed above...

If your sequence(s) do not match any of the above options [click here](#).

Figure 1. Selection of the sequence submission type.

Our template-based submission tool supports both constant and variable parameters for templates. Parameters are selected on the second page from a list of mandatory and optional fields (Figure 2). Constant common parameters are selected and filled in on the third page and the variable parameters are uploaded on the fourth page using a comma separated text file. This file is generated by Webin for the user based on the variable field selection and contains one column for each variable field. It is expected to be filled up by the submitter, e.g. by using Excel, and to contain the information for each sequence on its own row. Finally, the summary page provides an overview of the progress of the submission (Figure 3). Data is validated using the 'validate' button after which it can be submitted to the archive. Curator assistance can be requested from most pages.

SRA AUTOMATED SUBMISSION TOOL

The SRA accepts sequence submissions generated by the next-generation sequencing platforms. New submitters are advised to contact datasubs@ebi.ac.uk for the creation of a submission account and a secure data upload area. An automated submission service is provided to all registered submitters and is recommended for all users providing regular submissions. Immediate feedback is given of metadata validation errors and a service is provided for querying the data file processing status.

The first step in the submission process is to upload data files in platform specific, SRF or fastq formats using FTP or Aspera protocols into the secure data upload area. Aspera (<http://www.asperasoft.com/>) is a commercial

UDP-based data transfer protocol capable of better utilization of available network bandwidth than the TCP-based FTP protocol.

The second step is the preparation of submission, study, sample, experiment and run SRA metadata XML files. Studies and samples contain high-level project and sample information. Each experiment is associated with a single study and one or more samples. Experiments contain one or more runs which are associated with the submitted data files. The final step is to use our RESTful web-based service (<https://www.ebi.ac.uk/ena/submit/drop-box/>) to submit the data files and the SRA XML objects. Interactive submissions use the submission form and fully automated submissions take advantage of the RESTful service.

ENA BROWSER

We have developed a new web-based data retrieval and visualization tool which has been first deployed for the SRA, Project and Taxonomy data, and which will soon be expanded to cover the remaining ENA-Reads data (from the Trace Archive) and ENA-Assembly and ENA-Annotation. Data can be visualized and downloaded in XML, HTML and flat file formats. Retrievals can be made by single accession numbers, e.g. <http://www.ebi.ac.uk/ena/data/view/SRP000031&display=html>, ranges of accession numbers, e.g. <http://www.ebi.ac.uk/ena/data/view/ERX000025-ERX000034&display=html>, or by lists of accession numbers, e.g. <http://www.ebi.ac.uk/ena/data/view/ERR001087,ERR001088&display=html>. Numeric project and taxonomy identifiers must be

Tell us which information you are submitting

Please specify which fields are relevant to your entry. Click 'include' to add fields. You will provide the data for the fields in later sections.

Source

Field	Info	Include
Organism	(i)	✓
Strain name	(i)	✓
Clone identifier	(i)	<input type="checkbox"/>
Clone library name	(i)	✓
Isolate name	(i)	<input type="checkbox"/>
Country	(i)	✓
Isolation source	(i)	✓
Cell type	(i)	✓
Tissue type	(i)	✓

[Include all](#) [Exclude all](#)

Figure 2. Selection of the fields to include in the submission.

Summary

When all sections are complete, click the **Finish** button to process the data and submit

EST submission -This submission type is suitable for EST submissions

Entry number and release date ✓ - Complete

Number of entries

Public release date

Date format: 15/07/2009. Limited to 2 years ahead.

Add citations (i) ✓ - Complete

Select All | Select None

Select	Author	Title	Type	Edit
<input type="checkbox"/>	Bower L. [...]	Novel sequences in Fruit Fly	Unpublished journal article	Edit

Field values (i)

✓ - Complete

✓ - Complete

✓ - Complete

Validate data and submit

Figure 3. Submission summary page.

Project : 17929 : Soil isolate			
Taxon : 406425 : Burkholderia cenocepacia MC0-3 More			
Strain	MC0-3		
Locus Tag Prefix	BCENMC03		
Type	Genome Sequencing		
Submitter	DOE Joint Genome Institute		
Consortium	US DOE Joint Genome Institute		
Replicons			
Type	Name	Accession	Length
Chromosome	2	CP000959	3.20Mbp
Chromosome	1	CP000958	3.53Mbp
Chromosome	3	CP000960	1.22Mbp
Cellular features			
Gram stain	-		
Shape	Rod		
Endospores	No		
Motility	Yes		
Environment			
Oxygen Req.	Facultative		
Habitat	Multiple		
Temperature			
Opt. temp.	66.6C		
Range	Mesophilic		
Pathogen			
Pathogenic in	Human		
Disease	Necrotizing pneumonia, chronic infections		
See Also			
GOLD			

Figure 4. ENA Browser project page.

prefixed with 'Project:' and 'Taxon:', e.g. <http://www.ebi.ac.uk/ena/data/view/Project:10724&display=html> (Figure 4) and <http://www.ebi.ac.uk/ena/data/view/Taxon:9606&display=html> (Figure 5). Display in XML and HTML format is requested by using 'display=xml' and 'display=html' attributes, respectively. Download in gzip compressed format is possibly by using 'download=gzip' in place of 'display' attribute. SRA data can be downloaded either in submitted or fastq format by clicking links displayed in the SRA submission and run pages.

The ENA browser has been fully integrated with the EB-Eye indexer accessible from the header section of all

EBI web pages. Users search on accession numbers, description text or other free text to find appropriate data in the ENA Browser.

ENA SEQUENCE SIMILARITY SEARCH

Early in 2010, we expect to launch a new sequence similarity search service based on Exonerate (6) and Velvet (7). Exonerate servers will be used for searching all assembled sequences. We have extended Velvet, a de Bruijn graph-based sequence assembler, to support sequence similarity searches against assemblies induced from trace and short

Taxon : 9606 : Homo sapiens : human

Rank : species

Lineage

- > cellular organisms
- > **Eukaryota**
- > Fungi/Metazoa group
- > **Metazoa**
- > Eumetazoa
- > Bilateria
- > Coelomata
- > Deuterostomia
- > **Chordata**
- > **Craniata**
- > **Vertebrata**
- > Gnathostomata
- > Teleostomi
- > **Euteleostomi**
- > Sarcopterygii
- > Tetrapoda
- > Amniota
- > **Mammalia**
- > Theria
- > **Eutheria**
- > **Euarchontoglires**
- > **Primates**
- > **Haplorrhini**
- > Simiiformes
- > **Catarrhini**
- > Hominoidea
- > **Hominidae**
- > Homininae
- > **Homo**

Children

- > **Homo sapiens neanderthalensis**

Genetic Code : Translation table 1
Mitochondrial Genetic Code : Translation table 2

Figure 5. ENA Browser taxonomy page.

read sequences. We have implemented Velvet as a server that uses the Exonerate client server protocol so that we can run the Exonerate client against both Exonerate and Velvet servers. We have extended the exonerate client to support multiple and redundant servers to maximize the availability of our sequence search service. The result for the user will be a simple search page from which searches across comprehensive data can be launched, using Exonerate or Velvet methods as appropriate according to the nature of the data to be searched.

Presently, sequence similarity searches for ENA data are available using web, as well as EBI SOAP and REST Web Services interfaces (8). Search against ENA-Annotation sequences is available using NCBI-Blast (9) at <http://www.ebi.ac.uk/Tools/sss/ncbiblast/nucleotide.html> and Fasta (10) at <http://www.ebi.ac.uk/Tools/sss/fasta/nucleotide.html>. WGS sequences and full genomes are available for Fasta search at <http://www.ebi.ac.uk/Tools/sss/fasta/wgs.html> and <http://www.ebi.ac.uk/Tools/sss/fasta/genomes.html>, respectively.

FUNDING

Funding for open access charge: European Molecular Biology Laboratory and the Wellcome Trust.

Conflict of interest statement. None declared.

REFERENCES

1. Sanger, F. and Coulson, A.R. (1975) A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J. Mol. Biol.*, **94**, 441–448.
2. Sanger, F., Nicklen, S. and Coulson, A.R. (1977) DNA sequencing with chain-terminating inhibitors. *Proc. Natl Acad. Sci. USA*, **74**, 5463–5467.
3. Ansorge, W.J. (2009) Next-generation DNA sequencing techniques. *N. Biotechnol.*, **25**, 195–203.
4. Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J. and Sayers, E.W. (2009) GenBank. *Nucleic Acids Res.*, **37**, D26–D31.
5. Sugawara, H., Ikeo, K., Fukuchi, S., Gojobori, T. and Tateno, Y. (2009) DDBJ dealing with mass data produced by the second generation sequencer. *Nucleic Acids Res.*, **37**, D16–D18.
6. Slater, G. and Birney, E. (2005) Automated generation of heuristics for biological sequence comparison. *BMC Bioinform.*, **6**, 31.

7. Zerbino, D.R. and Birney, E. (2008) Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome Res.*, **18**, 821–829.
8. McWilliam, H., Valentin, F., Goujon, M., Li, W., Narayanasamy, M., Martin, J., Miyar, T. and Lopez, R. (2009) Web services at the European Bioinformatics Institute. *Nucleic Acids Res.*, **37**, W6–W10.
9. Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
10. Pearson, W.R. and Lipman, D.J. (1988) Improved tools for biological sequence comparison. *PNAS*, **85**, 2444–2448.