D28–D31 *Nucleic Acids Research, 2011, Vol. 39, Database issue* *Published online 23 October 2010*
doi:10.1093/nar/gkq967

The European Nucleotide Archive

Rasko Leinonen*, Ruth Akhtar, Ewan Birney, Lawrence Bower, Ana Cerdeno-Tárraga, Ying Cheng, Iain Cleland, Nadeem Faruque, Neil Goodgame, Richard Gibson, Gemma Hoad, Mikyung Jang, Nima Pakseresht, Sheila Plaister, Rajesh Radhakrishnan, Kethi Reddy, Siamak Sobhany, Petra Ten Hoopen, Robert Vaughan, Vadim Zalunin and Guy Cochrane

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

Received September 15, 2010; Accepted October 3, 2010

ABSTRACT

The European Nucleotide Archive (ENA; http://www.ebi.ac.uk/ena) is Europe's primary nucleotide-sequence repository. The ENA consists of three main databases: the Sequence Read Archive (SRA), the Trace Archive and EMBL-Bank. The objective of ENA is to support and promote the use of nucleotide sequencing as an experimental research platform by providing data submission, archive, search and download services. In this article, we outline these services and describe major changes and improvements introduced during 2010. These include extended EMBL-Bank and SRA-data submission services, extended ENA Browser functionality, support for submitting data to the European Genome-phenome Archive (EGA) through SRA, and the launch of a new sequence similarity search service.

THE EUROPEAN NUCLEOTIDE ARCHIVE

The European Nucleotide Archive (ENA) operates as a public archive for nucleotide sequence data. By bringing together databases for raw sequence data, assembly information and functional annotation, the ENA provides a comprehensive and integrated resource for this fundamental source of biological information. Central to the ENA is the provision of submission services, including interactive and programmatic submission tools, search services, including text and sequence similarity search tools and data presentation and retrieval services. The ENA works closely together with NCBI (1) and DDBJ (2) as partners in the International Nucleotide Sequence Database Collaboration (3). The principal policy of INSDC is to provide free and unrestricted permanent access to all archived nucleotide data. All primary data in the INSDC belongs to the submitters and can only be updated with submitter consent. For full policy details please refer to: http://www.insdc.org/policy.html.

CONTENT

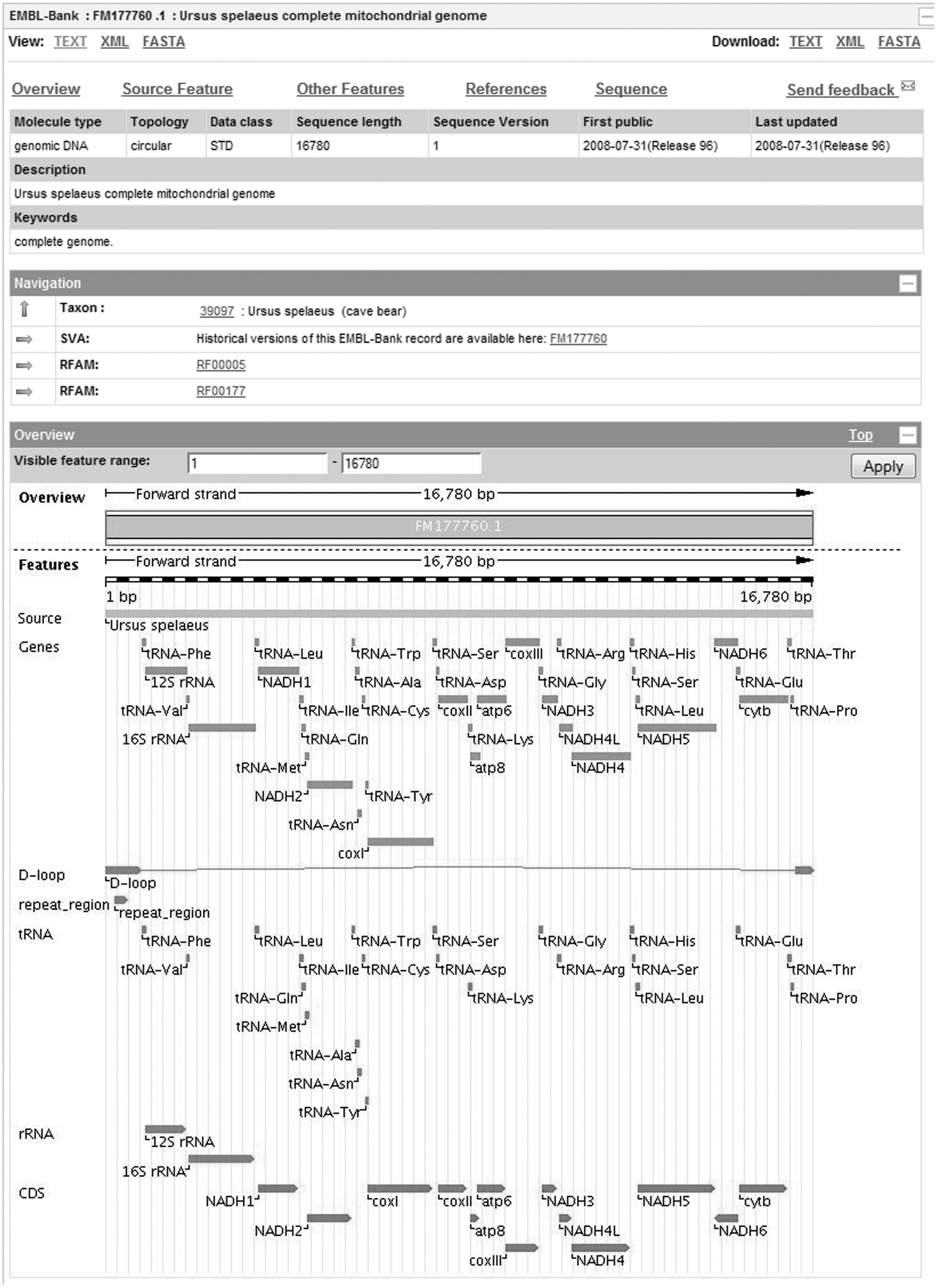
In October 2010, the ENA contained ~500 billion raw and assembled sequences consisting of ~50 trillion base pairs. In the last 3 years, the next-generation sequence reads stored in the Sequence Read Archive (SRA) have become the largest and fastest growing source of new data accounting now for ~95% of all base pairs made available by ENA. At the same time, the number of completed genome sequences has risen to over 1400 for cellular organisms and 3000 for viruses and phages (http://www.ebi.ac.uk/genomes/).

SUBMISSIONS OF RAW DATA FROM NEXT GENERATION PLATFORMS

The SRA accepts sequence submissions from next-generation sequencing platforms. New submitters should contact datasubs@ebi.ac.uk for the creation of a submission account and a secure data upload area. Submitters first upload data files into the secure data-upload area in one of the supported data formats, then prepare and submit study, sample, experiment, run and submission XML files to SRA. Detailed submission instructions are available here: http://www.ebi.ac.uk/ena/about/page.php?page = sra_submissions.

We have extended the SRA submission service to support submissions of authorized access data, typically clinical samples that have been sequenced under a confidentiality and consent agreement. Authorized access data can now be submitted through the SRA submission service into the European Genome-phenome Archive

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



EMBL-Bank : FM177760 .1 : Ursus spelaeus complete mitochondrial genome

View: TEXT XML FASTA

Download: TEXT XML FASTA

Overview Source Feature Other Features References Sequence Send feedback

Molecule type	Topology	Data class	Sequence length	Sequence Version	First public	Last updated
genomic DNA	circular	STD	16780	1	2008-07-31(Release 96)	2008-07-31(Release 96)

Description

Ursus spelaeus complete mitochondrial genome

Keywords

complete genome.

Navigation

⇧	Taxon :	39097 : Ursus spelaeus (cave bear)
⇨	SVA:	Historical versions of this EMBL-Bank record are available here: FM177760
⇨	RFAM:	RF00005
⇨	RFAM:	RF00177

Overview Top

Visible feature range: 1 - 16780 Apply

Figure 1. The complete mitochondrial genome for Ursus spelaeus (cave bear) from the Max Planck Institute for Evolutionary Anthropology submitted to EMBL-Bank in 2010.

(EGA; http://www.ebi.ac.uk/ega). Data submitted to EGA are not part of the public SRA database and are excluded from the INSDC data exchange. Permission to view and retrieve authorised access data can only be granted by the external data access committee (DAC) responsible for the data concerned. Please contact ega-helpdesk@ebi.ac.uk for more information about EGA policies. A secure data upload area is required to submit authorised access data through the SRA submission service. It is also possible to submit EGA's policy, dataset and DAC objects through SRA.

SRA will shortly accept sequence read submissions in Binary Alignment/Map (BAM) format (4). A BAM file is a binary compressed representation of the Sequence Alignment/Map (SAM) format. With sequence read alignments becoming an increasingly common intermediate in primary analysis, BAM format is emerging as a popular choice for storing sequence reads with alignments. The SRA is currently finalizing an archive BAM specification which will standardize the use of BAM files for primary data archival purposes. Once completed, BAM submissions to SRA archives will be required to follow this specification.

SUBMISSIONS OF ASSEMBLED AND ANNOTATED SEQUENCES

EMBL-Bank is a comprehensive public database of nucleotide sequences, associated biological annotation and bibliographic information. It contains a large diversity of data from patent, expressed sequence tag, whole genome shotgun and other high-throughput sequences, through genomic assemblies and richly annotated sequence fragments to whole replicons (5). Submitters should navigate to http://www.ebi.ac.uk/ena/about/page.php?page=submissions for access to all submission services. Advice regarding EMBL-Bank submissions is available from datasubs@ebi.ac.uk.

We have extended the web-based EMBL-Bank submission service in a number of ways. For providers of genome-scale data, we have added functionality that allows data submissions in EMBL-Bank flat file format. For smaller scale submissions, we have added new templates to the EMBL-Bank submission service. Each template focuses on a particular commonly occurring type of sequence and annotation data and collects required information from the submitters using a web form or spreadsheet upload. New templates are available for unannotated WGS submissions with only source organism annotation, and for protein coding and phylogenetic-marker regions. The template mechanism, introduced in 2009, has been well received and attracts now up to half of all web-based EMBL-Bank submissions.

DATA SEARCH, BROWSING AND RETRIEVAL

ENA data can be browsed and retrieved in XML, HTML, fasta, fastq and flat file formats using the ENA Browser which can be used both interactively and programmatically through REST URLs. In 2010, we extended the ENA Browser to cover EMBL-Bank and Trace archive records and introduced several improvements including a graphical EMBL-Bank annotation and assembly viewer and intuitive navigation between different ENA data classes. For full details of the ENA browser URL syntax please refer to: http://www.ebi.ac.uk/ena/about/page.php?page=browser. For example, the following URL returns the complete mitochondrial genome for 'Ursus spelaeus' (cave bear) (6): http://www.ebi.ac.uk/ena/data/view/FM177760 (Figure 1). Data can be queried using the EB-Eye free text search functionality available in the header section of all EBI web pages (7). ENA results are available under the 'Nucleotide Sequences' category and linked to the ENA Browser. Free text search is also available from the ENA home page: http://www.ebi.ac.uk/ena.

Rapid and comprehensive sequence similarity searches against ENA data are supported through a new service based on Exonerate (8) technology: http://www.ebi.ac.uk/ena/search/ (Goodgame, N., manuscript in preparation). All nucleotide sequences archived by the INSDC and made available as part of EMBL-Bank are covered by our service. This includes all ENA sequences except raw reads from the Trace Archive and SRA. Experimental search support for a limited number of raw reads is provided through De-Bruijn servers based on Velvet (9), using the Exonerate client-server protocol and being fully integrated with our search service. This search is available by selecting the 'Experimental De Bruijn search' option from the search page. The EMBL-Bank sequence search service is currently being expanded for more specific purposes according to community requests.

Bulk download of EMBL-Bank data is supported through FTP at ftp://ftp.ebi.ac.uk/pub/databases/embl/, and SRA and Trace Archive data through FTP at ftp://ftp.sra.ebi.ac.uk/ and Aspera through fasp.sra.ebi.ac.uk.

ENA COMMUNITY

The ENA team welcomes feedback and suggestions relating to all of our services at datasubs@ebi.ac.uk. We are always interested in hearing from potential collaborators who have an interest in working with and integrating our services.

FUNDING

The ENA is funded by the European Molecular Biology Laboratory, European Commission and the Wellcome Trust. Funding for open access charge: European Molecular Biology Laboratory.

Conflict of interest statement. None declared.

REFERENCES

1. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Sayers,E.W. (2010) GenBank. *Nucleic Acids Res.*, **38**, D46–D51.
2. Kaminuma,E., Mashima,J., Kodama,Y., Gojobori,T., Ogasawara,O., Okubo,K., Takagi,T. and Nakamura,Y. (2010) DDBJ launches a new archive database with analytical tools for next-generation sequence data. *Nucleic. Acids Res.*, **38**, D33–D38.

3. Cochrane,G. *et al.* (2011) The International Nucleotide Sequence Database Collaboration in 2010. *Nucleic Acids Res.*, http://www.insdc.org/.
4. Li,H., Handsaker,B., Wysoker,A., Fennell,T., Ruan,J., Homer,N., Marth,G., Abecasis,G., Durbin,R. and 1000 Genome Project Data Processing Subgroup (2009) The Sequence alignment/map (SAM) format and SAMtools. *Bioinformatics*, **25**, 2078–2079.
5. Leinonen,R., Akhtar,R., Birney,E., Bonfield,J., Bower,L., Corbett,M., Cheng,Y., Demiralp,F., Faruque,N., Goodgame,N. *et al.* (2010) Improvements to services at the European Nucleotide Archive. *Nucleic Acids Res.*, **38**, D39–D45.
6. Krause,J., Unger,T., Nocon,A., Malaspinas,A.S., Kolokotronis,S.O., Stiller,M., Soibelzon,L., Spriggs,H., Dear,P.H., Briggs,A.W. *et al.* (2008) Mitochondrial genomes reveal an explosive radiation of extinct and extant bears near the Miocene-Pliocene boundary. *BMC Evol. Biol.*, **8**, 220.
7. Valentin,F., Squizzato,S., Goujon,M., McWilliam,H., Paern,J. and Lopez,R. (2010) Fast and efficient searching of biological data resources—using EB-eye. *Brief. Bioinform.*, **11**, 375–384.
8. Slater,G. and Birney,E. (2005) Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics*, **6**, 31.
9. Zerbino,D.R. and Birney,E. (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.*, **18**, 821–829.