D32–D36 *Nucleic Acids Research, 2017, Vol. 45, Database issue* *Published online 28 November 2016*
doi: 10.1093/nar/gkw1106

European Nucleotide Archive in 2016

Ana Luisa Toribio*, Blaise Alako, Clara Amid, Ana Cerdeño-Tarrága, Laura Clarke, Iain Cleland, Susan Fairley, Richard Gibson, Neil Goodgame, Petra ten Hoopen, Suran Jayathilaka, Simon Kay, Rasko Leinonen, Xin Liu, Josué Martínez-Villacorta, Nima Pakseresht, Jeena Rajan, Kethi Reddy, Marc Rosello, Nicole Silvester, Dmitriy Smirnov, Daniel Vaughan, Vadim Zalunin and Guy Cochrane

European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, Cambridge CB10 1SD, UK

Received October 22, 2016; Revised October 25, 2016; Editorial Decision October 26, 2016; Accepted October 31, 2016

ABSTRACT

The European Nucleotide Archive (ENA; http://www.ebi.ac.uk/ena) offers a rich platform for data sharing, publishing and archiving and a globally comprehensive data set for onward use by the scientific community. With a broad scope spanning raw sequencing reads, genome assemblies and functional annotation, the resource provides extensive data submission, search and download facilities across web and programmatic interfaces. Here, we outline ENA content and major access modalities, highlight major developments in 2016 and outline a number of examples of data reuse from ENA.

INTRODUCTION

For the last third of a century, the European Nucleotide Archive (ENA; http://www.ebi.ac.uk/ena) has served as a cornerstone of the world's bioinformatics infrastructure. In 2016, the resource continues as a broad and heavily used platform for the sharing, publication, safeguarding and reuse of globally comprehensive public nucleotide sequence data and associated information.

ENA content spans a spectrum of data types from raw reads to asserted annotation and offers a broad range of services to the scientific community. Submissions services cater for our broad range of data providers, from the small-scale submitting research laboratory to the major sequencing centre. Data access services, such as our search and retrieval interfaces support users, from those browsing casually to those integrating ENA content through programmatic access into their own analyses and software applications. Our helpdesk provides responsive support to several thousand active data submitters and many times this number of data consumers.

As a member of the International Nucleotide Sequence Database Collaboration (INSDC; http://www.insdc.org/) (1), ENA partners with the National Institute of Genetics' DNA DataBank of Japan (2) and the National Center for Biotechnology's GenBank and Sequence Read archive (NCBI GenBank and SRA) (3) to provide globally comprehensive coverage through routine data exchange, to build the scientific standards for data exchange and to promote the timely sharing of well-structured sequence data.

In this paper, we outline ENA content, introduce the core services on offer from the resource and provide entry points for users into these services. We then outline a number of developments that have been made in the last year. Finally, we highlight examples of use of ENA data.

CONTENT

We continue to see significant growth in ENA content, with doubling times currently at 32 months for raw sequence data. At the time of writing, ENA comprised 3×10^{15} base pairs, over 1.5 million taxa, 192 thousand studies, 770 million sequence records and references to almost 200 thousand literature publications.

ENA content is organized into a number of data types (see Figure 1). Core data types, such as raw sequence data and derived data, including sequences, assemblies and functional annotation, are supplemented with accessory data, such as studies and samples, to provide experimental context. Primary data (such as reads) and several derived data types (sequence, assembly and analysis) are submitted, while remaining derived data types (coding, non-coding, marker and environmental) are derived from submitted content as part of processing and indexing within ENA. Further information is provided from http://www.ebi.ac.uk/ena/submit/data-formats.

Data are highly integrated to provide discoverability, reusability and cross-data set interoperability. This integration is achieved at three levels; between records of the same class (such as through consistent standards of annotation in Feature Tables associated with sequence records), be-

*To whom correspondence should be addressed. Tel: +44 1223 494680; Fax: +44 1223 494468; Email: anat@ebi.ac.uk

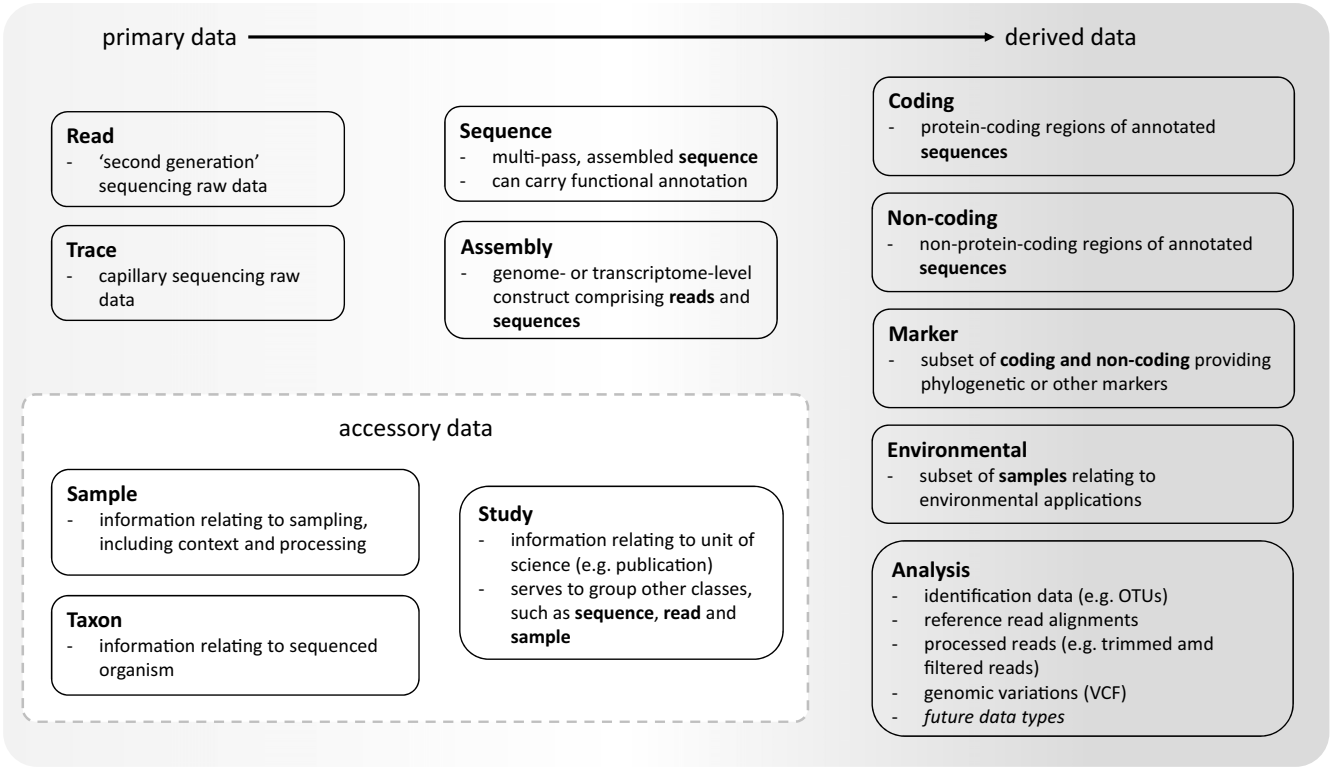

Figure 1. Organization of principal data types in ENA.

tween records of different classes (such as a sequence record making reference to a sample and a taxon) and between ENA records and records in external resources (such as a coding record carrying a cross-reference to a record in the UniProt KnowledgeBase) (4). Further discoverability and reusability are afforded through collaborations with external expert communities, in which we work on the development and implementation in ENA of reporting standards initiatives such as the Minimal Information about a Sequencing Experiment (5) and the Global Microbial Identifier (http://www.globalmicrobialidentifier.org/).

Given the aggressive technological advance that we continue to observe in sequencing technology, we expect ongoing change in the nature of incoming data (such as new read formats) and continued broadening of the application base served by sequencing (and hence new derived data structures). As such, we consider our role as a 'pioneer' database to be important and we strive to be agile and responsive in responding to new requirements from the scientific community. We support raw data from all sequencing platforms and have in place a technical system that allows rapid extension as new data emerge. For new applications, our 'analysis' data type has been put in place to allow us rapidly to respond to new derived data types; indeed, reference read alignments and genomic variation data are examples of extended data types that have been supported using this approach.

SUBMISSION

ENA offers a comprehensive range of submission options through the Webin system (http://www.ebi.ac.uk/ena/submit). This system provides both an interactive web application that offers, *inter alia*, spreadsheet upload support and a powerful RESTful programmatic submission interface. The former is recommended for infrequent submitters and first-time users (including those setting up programmatic submission systems). The latter is recommended for those with informatics skills who wish to establish ongoing regular data flow for a project or submitting centre. (We note that data volume *per se* is not a useful guide in choosing which tool to use; since both web and programmatic interfaces work in conjunction with file upload over FTP (supported within the web application and through external clients), with web and RESTful interfaces providing the user with transactional control over the submission.) For depositions of multi-omics data, submitters should first contact our helpdesk with a brief outline of the data set in question and we will work with colleagues within the European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI) BioStudies (https://www.ebi.ac.uk/biostudies/) and BioSamples (http://www.ebi.ac.uk/biosamples) (6) data resources to provide instructions on how to proceed.

ACCESS

We offer a range of web, programmatic and FTP services that support user access to ENA data, covering search, browse and download functions.

Three search services are supported: A simple text search box, available in the header of all pages on the ENA website (http://www.ebi.ac.uk/ena), provides rapid search across all data and web content and, when ENA accession numbers are searched, directly links to views of the requested records. The Advanced search service (http://www.ebi.ac.uk/ena/data/warehouse/search) allows users to select domains of interest (see Figure 1) and apply specific filters, such as taxonomic, geographical, date-related, to annotated fields within the selected domain. While the web forms within this search allow users to build up queries, the query language itself is exposed to the user and made available for direct editing for more complex Boolean searches. Results views from Advanced search are offered as lists of ENA records and as 'reports', which provide user-configured tabular views of annotations across records. Finally, sequence similarity searches (http://www.ebi.ac.uk/ena/data/sequence/search) are provided based on BLAST (7) and other tools supported centrally by EMBL-EBI.

The ENA Browser (e.g. http://www.ebi.ac.uk/ena/data/view/BN000065) allows users to navigate ENA data. Starting from results pages from any of the searches, or using direct links into records, a wealth of navigation options is on offer to link to similar records, records from the same taxon, study records that collate sequences and much more. All data types are supported with HTML views and, in most cases, XML, text and/or other downloadable options are also supported.

Programmatic access is richly supported at ENA. ENA search and download functions are all supported with RESTful services, allowing full embedding of ENA as a data source into high-throughput analysis and secondary software applications. Full details are provided at http://www.ebi.ac.uk/ena/browse/programmatic-access.

SUPPORT AND TRAINING

The ENA helpdesk provides the central point for users seeking support on any of the ENA services, from submissions to access. The main communication channel for the helpdesk is e-mail (contact datasubs@ebi.ac.uk), which operates through a central ticketing system that allows prioritization and cover during staff absence in order to maximize responsiveness. We encourage users to participate in our in-person or online 'surgery' events (http://www.ebi.ac.uk/ena/support/ena-surgeries) which allow more synchronous support in small user group settings. The helpdesk team builds documentation (available on the ENA website) and a number of downloadable training materials (http://www.ebi.ac.uk/ena/support/training-material). Finally, the team operates an in-person training programme that provides modules into generalist next generation sequencing courses and workshops run by specific user communities.

SELECTED DEVELOPMENTS IN 2016

Genome assembly data, especially from bacterial isolates, continue to grow relative to other data types. In order to serve these data better, we have supplemented the existing HTML view of assemblies with an XML view, available from the browser and through RESTful services. This XML view provides for the first time full access to all assembly information, including a new sequence report for all top-level sequences (above contigs) in an assembly.

Late in 2016, we have launched a major set of changes into the Webin data submission system. Previously, submissions of non-assembly-related sequence data with annotation were captured using the Webin system and fed a manually operated process for validation and preparation for loading into ENA by staff. With the new system, these data are validated, prepared and loaded without need for staff involvement. For those cases where submitted data pass validation, this allows far more rapid turnaround for users. Only in those cases where validation fails do the team's biocurators get involved in communicating with data submitters to help to curate the data. Although a major technical change, other than some visual changes to the user interface, for most users the submission process has remained largely unchanged, simply with faster turnaround.

Increasing interest in the application of sequencing technology to environmental, including host-associated, samples has led to demand for the capture of 'identification' data that allow users to report taxonomic, OTU and functional annotations derived from shotgun metagenomics and metabarcoding data. During the year, we have configured a new data type for the representation of these data (as a new class, 'ENVIRONMENTAL_IDENTIFICATION' of analysis record). Submission of such data is now supported through the programmatic interface and we plan to roll support out to the interactive Webin interface. We encourage those with identification data to contact the helpdesk for submissions support for this new data type.

We offer a variety of downloadable software to support users of ENA. During the year, we have improved the ENA validator tool with better release management allowing full software versioning. We have release the 'ENA sequence API' which is a set of utilities that we use internally for the processing of flatfile sequence data; we believe that this software will be useful for those developing tools upstream and downstream of ENA services.

We continue our work with the 'checklist' concept, for which we prepare, often with expert communities, lists of fields of information with formalized structure (such as dictionaries, ranges, regular expressions, etc.) that directly configure our submission and validation systems to ensure the capture of sufficiently described and well-structured data. One development here is that we now support external ontologies for fields in checklists. We expect to start to roll ontologies into a number of checklists in 2017; these will provide assistance to users in the submission process (e.g. look-up of terms to use in annotation) and improve data discoverability and reusability for those consuming ENA content. A second development is that the checklists are now available from the ENA browser to serve as a guide

to data submitters and data consumers. (see, for example, http://www.ebi.ac.uk/ena/data/view/ERC000038).

Finally, in 2016, we have added a new sequence similarity search method, MegaBLAST (8) (available from http://www.ebi.ac.uk/ena/data/sequence/search).

DATA REUSE

ENA serves as a foundation for onward sequence-based science at many scales. Deep technical integration exists for a number of data resources that consume ENA content amongst their sources. At EMBL-EBI, the Expression Atlas (9) and RNAcentral (10) resources, for example, provide transcriptomics and non-coding RNA specialized processing and integration to their respective users. Outside EMBL-EBI, resources such as SILVA (11) and Eukaryotic Promoter Database (12), for example, provide reference ribosomal RNA gene sets and promoter data to their respective users. Lighter integration, where scientific research groups discover and retrieve data from ENA for meta-analysis are also common. Here we summarize some examples.

In May 2015, ENA released the new complete genomes of *Toxoplasma gondii* VEG and *Neospora caninum* LIV, two eukoryote parasites (13). The genomes have been improved by resequencing (corrections and gap bridging) of the original sequence. The work was done by a different group than the original genomes; RNA-seq was also used to improve the annotation with novel transcript features. The new genome assembly of *Neospora caninum* contains 14 chromosomes and 1 unplaced scaffold (http://www.ebi.ac.uk/ena/data/view/LN714474-LN714488) while the new genome assembly of *Toxoplasma gondii* genome is made of 14 chromosomes and 12 unplaced scaffolds (http://www.ebi.ac.uk/ena/data/view/LN714489-LN714514). Note that the resequencing and RNA-seq data sets are available and linked from the cross-reference (DR lines) of the corresponding flat files with accessions starting 'ERR'.

Another example of meta-analysis is the recent case of a novel virus identified in a eukaryote genome. In this case, the scientists worked on the existing genome sequence of *Centruroides exilicauda* (scorpion) to identify two variants of a novel virus (14). Both variant genomes received new ENA accession numbers (http://www.ebi.ac.uk/ena/data/view/LN846618-LN846619) and the alignment files (in BAM format) for both assemblies are also available in the ENA Browser (http://www.ebi.ac.uk/ena/data/view/ERZ112709-ERZ112710).

The reuse of existing data sets can include multiple ENA samples. This is the case of the assembly of two new leptin-like genes from birds (15). The leptin genes are GC-rich and therefore difficult to detect by conventional sequencing. The sequences submitted to ENA are made up of reads from different samples of different sequencing projects. Each submitted leptin-like gene received an ENA accession number (duck leptin mRNA, http://www.ebi.ac.uk/ena/data/view/LN794245; chicken leptin mRNA, http://www.ebi.ac.uk/ena/data/view/LN794246). The alignments that show how these sequences were constructed from INSDC reads are also available in BAM format from the ENA Browser (http://www.ebi.ac.uk/ena/data/view/ERZ115817-ERZ115818).

In addition to the above examples, the reuse of existing annotated sequences from ENA allows researchers to perform comparative analysis and build phylogenetic trees. This has been carried out in the case of *Citrobacter rodentium*, a natural murine pathogen whose genome has been sequenced (16). Comparative genomic studies against existing human pathogen genomes, combined with phylogenetic analysis using a number of existing gastrointestinal pathogen sequences, brought further understanding of this mouse pathogen as a unique model of the human enteropathogenic and enterohaemorrhagic *Escherichia coli* (EPEC and EHEC) infection.

FUNDING

European Molecular Biology Laboratory (EMBL); The Horizon 2020 Programme of the European Union under COMPARE [643476], EXCELERATE [676559]; EMBRIC [654008]; UK Biotechnology and Biological Sciences Research Council under Metagenomics Portal [BB/M011755/1], MG-RAST-EBI [BB/N018354/1]; RNA Central [NN/J019321/1]. Funding for open access charge: European Molecular Biology Laboratory (EMBL).
Conflict of interest statement. None declared.

REFERENCES

1. Cochrane,G., Karsch-Mizrachi,I., Takagi,T. and International Nucleotide Sequence Database Collaboration (2016) The International nucleotide sequence database collaboration. *Nucleic Acids Res.*, **44**, D48–D50.
2. Mashima,J., Kodama,Y., Kosuge,T., Fujisawa,T., Katayama,T., Nagasaki,H., Okuda,Y., Kaminuma,E., Ogasawara,O., Okubo,K. *et al.* (2016) DNA data bank of Japan (DDBJ) progress report. *Nucleic Acids Res.*, **44**, D51–D57.
3. Clark,K., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Sayers,E.W. (2016) GenBank. *Nucleic Acids Res.*, **44**, D67–D72.
4. UniProt Consortium (2015) UniProt: a hub for protein information. *Nucleic Acids Res.*, **43**, D204–D212.
5. Yilmaz,P., Kottmann,R., Field,D., Knight,R., Cole,J.R., Amaral-Zettler,L., Gilbert,J.A., Karsch-Mizrachi,I., Johnston,A., Cochrane,G *et al.* (2011) Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. *Nat. Biotechnol.*, **29**, 415–420.
6. Faulconbridge,A., Burdett,T., Brandizi,M., Gostev,M., Pereira,R., Vasant,D., Sarkans,U., Brazma,A. and Parkinson,H. (2014) Updates to BioSamples database at European Bioinformatics Institute. *Nucleic Acids Res.*, **42**, D50–D52.
7. Boratyn,G.M., Camacho,C., Cooper,P.S., Coulouris,G., Fong,A., Ma,N., Madden,T.L., Matten,W.T., McGinnis,S.D., Merezhuk,Y. *et al.* (2013) BLAST: a more efficient report with usability improvements. *Nucleic Acids Res.*, **41**, W29–W33.
8. Zhang,Z., Schwartz,S., Wagner,L. and Miller,W. (2000) A greedy algorithm for aligning DNA sequences. *J. Comput. Biol.*, **7**, 203–214.
9. Petryszak,R., Keays,M., Tang,Y.A., Fonseca,N.A., Barrera,E., Burdett,T., Füllgrabe,A., Fuentes,A.M., Jupp,S., Koskinen,S *et al.* (2016) Expression Atlas update–an integrated database of gene and protein expression in humans, animals and plants. *Nucleic Acids Res.*, **44**, D746–D752.
10. RNAcentral Consortium (2015) RNAcentral: an international database of ncRNA sequences. *Nucleic Acids Res.*, **43**, D123–D129.
11. Quast,C., Pruesse,E., Yilmaz,P., Gerken,J., Schweer,T., Yarza,P., Peplies,J. and Glöckner,F.O. (2013) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.*, **41**, D590–D596.
12. Dreos,R., Ambrosini,G., Périer,R.C. and Bucher,P. (2015) The eukaryotic promoter database: expansion of EPDnew and new promoter analysis tools. *Nucleic Acids Res.*, **43**, D92–D96.

13. Ramaprasad,A., Mourier,T., Naeem,R., Malas,T.B., Moussa,E., Panigrahi,A., Vermont,S.J., Otto,T.D., Wastling,J. and Pain,A. (2015) Comprehensive evaluation of *Toxoplasma gondii* VEG and *Neospora caninum* LIV genomes with tachyzoite stage transcriptome and proteome defines novel transcript features. *PLoS One*, **10**, e0124473.
14. Buck,C.B., Van Doorslaer,K., Peretti,A., Geoghegan,E.M., Tisza,M.J., An,P., Katz,J.P., Pipas,J.M., McBride,A.A., Camus,A.C. *et al.* (2016) The Ancient Evolutionary History of Polyomaviruses. *PLoS Pathog.*, **12**, e1005574.
15. Seroussi,E., Cinnamon,Y., Yosefi,S., Genin,O., Smith,J.G., Rafati,N., Bornelöv,S., Andersson,L. and Friedman-Einat,M. (2016) Identification of the long-sought leptin in chicken and duck: expression RT pattern of the highly GC-rich avian leptin fits an autocrine/paracrine RT rather than endocrine function. *Endocrinology*, **157**, 737–51.
16. Petty,N.K., Bulgin,R., Crepin,V.F., Cerdeño-Tárraga,A.M., Schroeder,G.N., Quail,M.A., Lennard,N., Corton,C., Barron,A., Clark,L. *et al.* (2010) The *Citrobacter rodentium* genome sequence reveals convergent evolution with human pathogenic *Escherichia coli*. *J. Bacteriol.*, **192**, 525–538.