D38–D42 *Nucleic Acids Research, 2012, Vol. 40, Database issue* *Published online 22 November 2011*
doi:10.1093/nar/gkr994

The DNA Data Bank of Japan launches a new resource, the DDBJ Omics Archive of functional genomics experiments

Yuichi Kodama, Jun Mashima, Eli Kaminuma, Takashi Gojobori, Osamu Ogasawara, Toshihisa Takagi, Kousaku Okubo and Yasukazu Nakamura*

Center for Information Biology and DNA Data Bank of Japan, National Institute of Genetics, Research Organization for Information and Systems, Yata, Mishima 411-8510, Japan

Received September 7, 2011; Revised October 17, 2011; Accepted October 18, 2011

ABSTRACT

The DNA Data Bank of Japan (DDBJ; http://www.ddbj.nig.ac.jp) maintains and provides archival, retrieval and analytical resources for biological information. The central DDBJ resource consists of public, open-access nucleotide sequence databases including raw sequence reads, assembly information and functional annotation. Database content is exchanged with EBI and NCBI within the framework of the International Nucleotide Sequence Database Collaboration (INSDC). In 2011, DDBJ launched two new resources: the 'DDBJ Omics Archive' (DOR; http://trace.ddbj.nig.ac.jp/dor) and BioProject (http://trace.ddbj.nig.ac.jp/bioproject). DOR is an archival database of functional genomics data generated by microarray and highly parallel new generation sequencers. Data are exchanged between the ArrayExpress at EBI and DOR in the common MAGE-TAB format. BioProject provides an organizational framework to access metadata about research projects and the data from the projects that are deposited into different databases. In this article, we describe major changes and improvements introduced to the DDBJ services, and the launch of two new resources: DOR and BioProject.

INTRODUCTION

Since 1987, the DNA Data Bank of Japan (DDBJ; http://www.ddbj.nig.ac.jp) has been operating public resources for biological information at the National Institute of Genetics (NIG) by providing submission, archive, search, download and analysis services. The objective of DDBJ is to support and promote the sharing and use of biological data as a public resource. The traditional DDBJ archive has collected annotated nucleotide sequences by collaborating with the EMBL-Bank at the European Bioinformatics Institute (EBI) and GenBank at the National Center for Biotechnology Information (NCBI) as partners in the International Nucleotide Sequence Database Collaboration (INSDC) (1). In the late 1990s, automated capillary platforms were introduced to many sequencing centres, which began to produce raw data at the genomic scale. Researchers soon recognized the importance of this raw data for the qualitative assessment of sequences and high-level re-analysis such as re-base-calling and re-assembly. DDBJ therefore installed the Trace Archive (http://trace.ddbj.nig.ac.jp/dta) to store the DNA sequence chromatograms (traces), base-calls and quality estimates from single-pass reads of various large-scale sequencing projects. More recently, massively parallel new generation sequencing platforms are producing biological sequencing data in unprecedented amounts. To ensure the wider scientific community can access and use these enormous amounts of data for scientific discovery and application, DDBJ established the 'Sequence Read Archive' (DRA; http://trace.ddbj.nig.ac.jp/dra) in 2008 to store and disseminate raw next-generation sequencing data (2). The traditional DDBJ and the DDBJ Trace and Sequence Read Archives share all submitted public data with the other members of INSDC.

This year, DDBJ established two new resources: the 'DDBJ Omics Archive' (DOR; http://trace.ddbj.nig.ac.jp/dor) and BioProject (http://trace.ddbj.nig.ac.jp/bioproject). The new-generation sequencing platforms produce data from gene expression, epigenomics and variation studies. In these functional genomics studies, the new-generation platforms are replacing the microarray because of their much greater resolving power and accuracy. These functional genomics data have been archived by the Gene Expression Omnibus (GEO) at

*To whom correspondence should be addressed. Tel: +81 55 981 6859; Fax: +81 55 981 6889; Email: yanakamu@genes.nig.ac.jp

NCBI (3) and by the ArrayExpress at EBI (4). DOR is a resource of high-throughput experimental data for functional genomics generated by sequencing as well as using microarray platforms. DOR accepts functional genomics experiments in the MAGE-TAB format used by ArrayExpress. All public DOR data are exchanged with ArrayExpress.

As new-generation technologies significantly increase sequencing throughput, data from a single project can be dispersed across more than one archival database. The DDBJ BioProject resource provides an overview of diverse types of projects and organizes the data from the projects deposited into the archival databases maintained by members of INSDC. The DDBJ submission portal 'D-way' was developed to be a single submission account to enable users to submit their data to all relevant DDBJ archival databases (traditional DDBJ, Omics Archive, Trace Archive, Sequence Read Archive and BioProject).

The public analysis service of new generation sequences is provided by the 'DDBJ Read Annotation Pipeline' (http://p.ddbj.nig.ac.jp) (2). This pipeline processes the uploaded raw sequence reads on cloud computers and supports data submissions to the DDBJ archival databases. DDBJ is developing into a more comprehensive and integrated public domain resource for biological research by providing archival databases and analysis services.

DDBJ ARCHIVAL DATABASES

DDBJ traditional assembled sequence archive

Here we describe the development of the DDBJ database as a member of INSD during the course of 2011, which is appended to last year's report (2). Between July 2010 and June 2011, 18 296 211 entries/13 576 228 536 bp were released from INSD as core traditional nucleotide flat files, except for whole-genome shotgun (WGS), mass sequence for genome annotation (MGA) and third party annotation (TPA) files (5). DDBJ contributed 12.7% of the entries and 10% of the base pairs added to the core nucleotide data of INSD during this period. Most of the nucleotide data were submitted by Japanese researchers; the rest came from China, Korea, Taiwan and other countries. DDBJ has also continually distributed sequence data related to patent applications transferred from the Japan Patent Office (JPO, http://www.jpo.go.jp) and the Korean Intellectual Property Office (KIPO, http://www.kipo.go.kr/en). In addition to the core nucleotide data, DDBJ has released a total of 2 228 313 WGS entries, 35 271 312 MGA entries, 660 TPA entries, 6374 TPA-WGS entries and 1272 TPA-CON entries as of 15 July 2011.

Noteworthy large-scale data released from DDBJ are listed in Table 1. The genome data of the vase tunicate (*Ciona intestinalis*) has been re-assembled and re-annotated from the first draft sequences. The first draft genome of *C. intestinalis* was published in 2002 by the international genome sequencing project (6). Main contributors to the sequencing project were from the US Department of Energy Joint Genome Institute and Kyoto University. The international collaboration for the *C. intestinalis* genome was discontinued and the genome sequences reported in INSD remained as the first version. However, the genome data were updated by the Kyoto University team independently of the sequencing collaboration (7). Since the updated version of the *C. intestinalis* genome has been used as a standard in the ascidian research community, it would be useful for researchers to share the updated genome data through INSD. Since the Kyoto University team submitted the updated genome data, DDBJ accepted and released the

Table 1. List of large-scale data released by DDBJ from July 2010 to June 2011

Type	Organism	Accession number [number of entries, number of base pairs (bp)]
GSS	Fission yeast (*S. pombe*)	FT321169-FT434719 (113 551 entries, 55 932 252)
EST	Fission yeast (*S. pombe*)	FY072959-FY174037 (101 079 entries, 51 492 242)
EST	Kale (*Brassica oleracea* var. *acephala*)	DK455950-DK575153 (119 204 entries, 93 446 677)
Metagenome	Uncultivated thermophilic archaeon (*Candidatus Caldiarchaeum subterraneum*)	Fosmid clones: AB201309, AP011633, AP011650, AP011675, AP011689, AP011708, AP011723, AP011724, AP011727, AP011745, AP011751, AP011786, AP011796, AP011826-AP011902 (90 entries, 3 092 718) Scaffold CON: BA000048 (1 entry, 1 680 938)
GSS	Uncultivated thermophilic archaeon (*Candidatus Caldiarchaeum subterraneum*)	AG993735-AG999698 (5964 entries, 3 194 765)
Genome	*Jatropha curcas*	HTG: AP011961-AP011977 (17 entries, 1 356 476) WGS: BABX01000001-BABX01150417 (150 417 entries, 285 858 490)
GSS	Mouse (*Mus musculus domesticus*)	GA000001-GA131507 (131 507 entries, 130 569 356)
GSS	Rice (*Oryza sativa* Japonica Group)	FT872362-FT932077 (59 716 entries, 26 152 462)
Full length cDNA	Domesticated barley (*Hordeum vulgare* subsp. *vulgare*)	AK353559-AK377172 (23 614 entries, 40 190 236)
GSS	African rice (*Oryza glaberrima*)	FT434720-FT872361 (437 642 entries, 206 317 130)
Genome (Chromosome 3H)	Domesticated barley (*Hordeum vulgare* subsp. *vulgare* cv. Haruna Nijo)	BACC01000001-BACC01008583 (8583 entries, 28 015 997)
Genome (Re-assemble and re-annotation)	Vase tunicate (*Ciona intestinalis*)	TPA-WGS: EAAA01000001-EAAA01006374 (6374 entries, 112 163 512) TPA-scaffold CON: HT000001-HT001272 (1272 entries, 115 226 814)
EST	Tammar wallaby (*Macropus eugenii*)	FY469875-FY736474 (265 835 entries, 212 531 387)

updated version as its first case of TPA re-assemble. TPA re-assemble was agreed among INSD partners, GenBank, EMBL-Bank and DDBJ, as a strategy to accept some important genomes that were re-assembled by researchers other than the original submitters. Moreover, DDBJ has released the following: Expressed sequence tags (EST) and genome survey sequences (GSS) of fission yeast (*Schizosaccharomyces pombe*) submitted by Osaka City University, Japan; the kale (*Brassica oleracea* var. *acephala*) EST submitted by Kirin Holdings Company, Ltd, Japan; Metagenome and GSS of an uncultivated thermophilic archaeon (*Candidatus Caldiarchaeum subterraneum*) submitted by Japan Agency for Marine-Earth Science and Technology, Japan; the biofuel crop (*Jatropha curcas*) genome submitted by Kazusa DNA Research Institute, Japan; the mouse (*Mus musculus domesticus*) GSS submitted by the RIKEN BioResource Center, Japan; GSS of Japanese rice (*Oryza sativa* Japonica Group) and African rice (*O. glaberrima*) and full length cDNAs of domesticated barley (*Hordeum vulgare* subsp. *vulgare*) submitted by National Institute of Agrobiological Sciences, Japan; WGS derived from chromosome 3H of domesticated barley (*H. vulgare* subsp. *vulgare* cv. Haruna Nijo) submitted by Okayama University, Japan; the tammar wallaby (*Macropus eugenii*) EST submitted by National Institute of Genetics, Japan.

DDBJ Sequence Read Archive

DRA accepts raw sequencing data submissions from new-generation sequencing platforms (8). New submitters should contact DRA for the creation of a D-way submission account and a secure data upload area. Sequencing data files should then be uploaded into the secure data upload area. Submitters can prepare and submit all their information, including study, experiment, sample, run and analysis metadata associated with the sequencing data by using the web-based submission tool 'MetaDefine'. In the MetaDefine intuitive interface, users can create metadata by simply entering the necessary pieces of information into the appropriate fields and can revise the content according to the feedback given by the metadata validation. Data search and download services have also been included in DRA (http://trace.ddbj.nig.ac.jp/DRASearch). Users can search all the public SRA data of DDBJ/EBI/NCBI by accession numbers, several other parameters (e.g. organism and study type) and free-text keywords. The sequencing data can be downloaded by FTP in either the SRA toolkit format or fastq format via ftp://ftp.ddbj.nig.ac.jp/ddbj_database/dra. DRA works closely with several large-sequencing projects. For example, in the Genome Science Project (http://www.genome-sci.jp), DRA serves as a data hub to support data flow from sequencing centres to participant researchers and the DDBJ Read Annotation Pipeline. DRA continues to support large-scale and collaborative projects by coordinating data flow, submission and analysis.

DDBJ Omics Archive

DDBJ has supported the Center for Information Biology Gene Expression Database (CIBEX; http://cibex.nig.ac.jp), which has collected more than 3700 hybridizations from microarray experiments since 2004 (9). DDBJ has participated in the Functional Genomics Data (FGED; http://www.mged.org) society which aims to assure that the investment in functional genomics data generates maximum public benefit by defining minimum information specifications for reporting data and standards for biological research data. CIBEX accepts microarray data compliant with the Minimum Information About a Microarray Experiment (MIAME) guideline developed by FGED, which is required to interpret and verify the results of an experiment (10). Many scientific journals require the deposition of MIAME-compliant microarray data to GEO at NCBI and ArrayExpress at EBI or CIBEX before publication of the relevant paper. In functional genomics research areas, new-generation sequencers play a key role in producing data by measuring the abundance of DNA and RNA molecules. The shift from microarrays to new generation sequencing platforms for functional genomics investigations has resulted in much greater resolving power and accuracy for such experiments. To accommodate the increasing volume of high-throughput functional genomics data from sequencing- and microarray-based technologies, DDBJ established a new resource called 'DDBJ Omics Archive' (DOR; http://trace.ddbj.nig.ac.jp/dor). DOR accepts functional genomics data (e.g. gene expression, small RNA profiling and epigenetics, etc.) in the MAGE-TAB format specified by FGED and used by ArrayExpress. DOR encourages microarray and sequencing data submissions compliant with MIAME and Minimum Information about a high-throughput SeQuencing Experiment (MINSEQE) guidelines, respectively. Submitters need to prepare the MAGE-TAB submission files by using spreadsheet applications or third-party tools until our own submission tool is released. DOR issues internationally recognized accession numbers in the E-DORD-n format to the submitted experiments. Public data are exchanged between ArrayExpress and DOR. Once new-generation sequencing data sets are deposited in DOR, the raw sequencing data are submitted to DRA with the SRA XML metadata generated from the supplied MAGE-TAB files. CIBEX will stop accepting new submissions once DOR becomes fully operational. Access to the data that have been deposited into CIBEX will be provided. The CIBEX data will be exported to DOR on the request of the data owner.

DDBJ BioProject

As new high-throughput sequencing technologies significantly increase the volume of data that can be generated, distinct types of data (e.g. genome, transcriptome and targeted locus data) from a single project can be deposited into different archival databases. The BioProject resource is a redesigned, expanded, replacement of the NCBI Genome Project resource. The redesign adds tracking of several data elements including more precise information

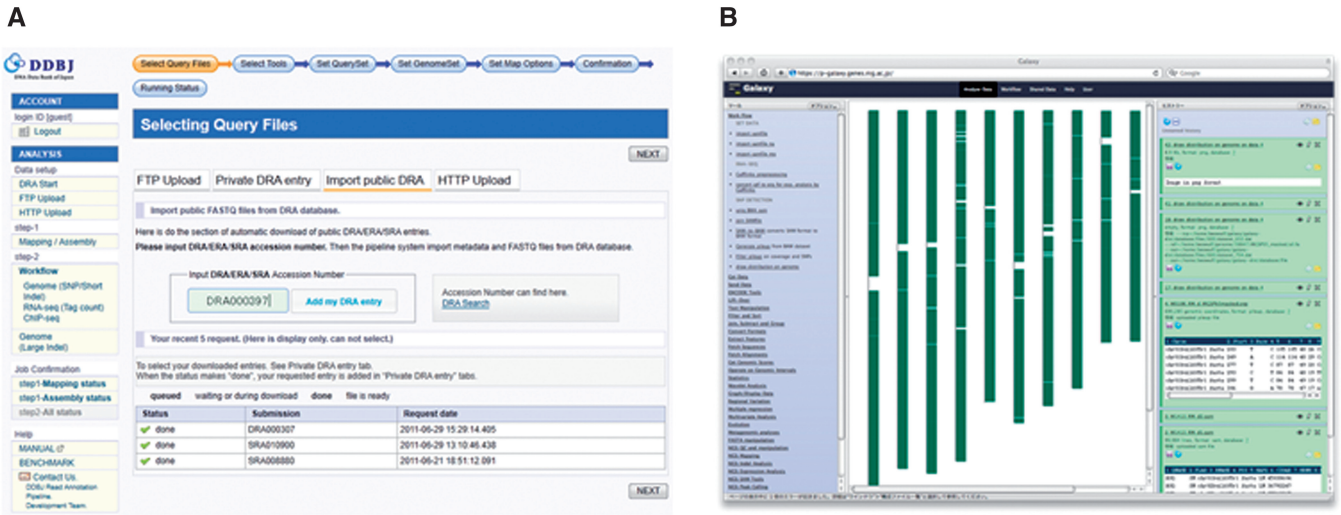

Figure 1. (**A**) DRA query import panel of DDBJ Pipeline. (**B**) Generated image for genomic locations of SNPs.

about a project's scope, material, objectives, funding source and general relevance categories. DDBJ BioProject issues accession numbers with the prefix 'PRJD' to the submitted projects and exchanges the released data with other members of INSDC. Data submitted to INSDC-associated databases cross-reference to the BioProject identifier to support navigation between the project and the project's datasets. BioProject enables users to retrieve a project's datasets across multiple archival databases. Related projects can be grouped under an umbrella project. An umbrella project may group projects that are part of a single collaborative effort but represent distinct studies that differ in methodology, sample material or resulting data type. The definition of a set of related data, a 'project', is very flexible and supports the need to define a complex project and various distinct sub-projects. Registration for a DDBJ BioProject accession is encouraged for the types of projects that result in submission of a very large volume of data, submissions from multiple members of a collaboration or submissions to multiple archival databases. A BioProject accession is required for submissions of microbial and eukaryotic genomes. At DDBJ, BioProject unites the records in the traditional DDBJ archive, Sequence Read Archive, Trace Archive and Omics Archive.

DDBJ ANALYSIS SERVICES

DDBJ Read Annotation Pipeline

The DDBJ Read Annotation Pipeline (DDBJ Pipeline, http://p.ddbj.nig.ac.jp) analyzes and annotates raw new-generation sequencing data and generates a template metadata file to support submission of analyzed data to the DDBJ archival databases (2). DDBJ Pipeline is a cloud computing-based analysis system in which users can access NIG supercomputers through a web application with a graphical user interface. DDBJ Pipeline consists of two processes: a basic process for reference sequence mapping and *de novo* assembly, and a successive analytical process for structural and functional annotations. In the basic process, popular mapping and assembly tools are available, and reference sequences can be retrieved by their INSDC accession numbers from the DDBJ database using SOAP access (11). Major enhanced functions are as follows: (i) DRA data import, (ii) New annotation workflows in the analytical process, (iii) Deletion policy of query and result data, (iv) MD5 checksum for download files and (v) Usage statistics information. DDBJ Pipeline implemented an automatic loader which loads DRA data from the public DDBJ FTP server in FASTQ file formats for query (Figure 1A). When loading is complete, the Pipeline system sends an e-mail notification to users. New workflows, genomic contig annotation, SNP detection (Figure 1B) and RNA-seq transcript annotation have been implemented in the analytical process. To save disk space for data storage, we set a policy to store query and result data up to 30 days. To verify the download process, MD5 checksums have been added to compressed download files. Usage statistics of job submissions and web access are displayed on the website.

ADDITIONAL INFORMATION

More information is available on the DDBJ website at http://www.ddbj.nig.ac.jp. News is delivered by really simple syndication (RSS), Twitter and mail magazines.

ACKNOWLEDGEMENTS

We gratefully acknowledge the support of all members of DDBJ for data collection, annotation and release and for software development. In particular, we thank Dr Hideki Nagasaki, Dr Satoshi Saruhashi, Takako Mochizuki, Naoko Sakamoto and Natsuko Sakakura for consultations on DRA and Pipeline, and Prof. Hideaki

Sugawara and Prof. Yoshio Tateno for support in the form of database maintenance and INSD collaboration.

FUNDING

Ministry of Education, Culture, Sports, Science and Technology of Japan (MEXT) (management expense grant for National University Cooperation, to DDBJ); MEXT Grant-in-Aid for Scientific Research on Innovative Areas 'Genome Science' (to DDBJ Sequence Read Archive and DDBJ Pipeline, partial). Funding for open access charge: The DDBJ management expenses grant (from MEXT).

Conflict of interest statement. None declared.

REFERENCES

1. Karsch-Mizrachi,I., Nakamura,Y. and Cochrane,G. (2012) The international nucleotide sequence database collaboration. *Nucleic Acids Res.*, **40**, D33–D37.
2. Kaminuma,E., Kosuge,T., Kodama,Y., Aono,H., Mashima,J., Gojobori,T., Ogasawara,O., Okubo,K., Takagi,T. and Nakamura,Y. (2011) DDBJ progress report. *Nucleic Acids Res.*, **39**, D22–D27.
3. Barrett,T., Troup,D.B., Wilhite,S.E., Ledoux,P., Evangelista,C., Kim,I.F., Tomashevsky,M., Marshall,K.A., Phillippy,K.H., Sherman,P.M. *et al.* (2011) NCBI GEO: archive for functional genomics data sets–10 years on. *Nucleic Acids Res.*, **39**, D1005–D1010.
4. Parkinson,H., Sarkans,U., Kolesnikov,N., Abeygunawardena,N., Burdett,T., Dylag,M., Emam,I., Farne,A., Hastings,E., Holloway,E. *et al.* (2011) ArrayExpress update–an archive of microarray and high-throughput sequencing-based functional genomics experiments. *Nucleic Acids Res.*, **39**, D1002–D1004.
5. Cochrane,G., Bates,K., Apweiler,R., Tateno,Y., Mashima,J., Kosuge,T., Mizrachi,I.K., Schafer,S. and Fetchko,M. (2006) Evidence standards in experimental and inferential INSDC Third Party Annotation data. *OMICS*, **10**, 105–113.
6. Dehal,P., Satou,Y., Campbell,R.K., Chapman,J., Degnan,B., De Tomaso,A., Davidson,B., Di Gregorio,A., Gelpke,M., Goodstein,D.M. *et al.* (2002) The draft genome of *Ciona intestinalis*: insights into chordate and vertebrate origins. *Science*, **298**, 2157–2167.
7. Satou,Y., Mineta,K., Ogasawara,M., Sasakura,Y., Shoguchi,E., Ueno,K., Yamada,L., Matsumoto,J., Wasserscheid,J., Dewar,K. *et al.* (2008) Improved genome assembly and evidence-based global gene model set for the chordate *Ciona intestinalis*: new insight into intron and operon populations. *Genome Biol.*, **9**, R152.
8. Kodama,Y., Shumway,M. and Leinonen,R. (2012) The sequence read archive: explosive growth of sequencing data. *Nucleic Acids Res.*, **40**, D54–D56.
9. Kodama,Y., Kaminuma,E., Saruhashi,S., Ikeo,K., Sugawara,H., Tateno,Y. and Nakamura,Y. (2010) Biological databases at DNA Data Bank of Japan in the era of next-generation sequencing technologies. *Adv. Exp. Med. Biol.*, **680**, 125–135.
10. Brazma,A., Hingamp,P., Quackenbush,J., Sherlock,G., Spellman,P., Stoeckert,C., Aach,J., Ansorge,W., Ball,C.A., Causton,H.C. *et al.* (2001) Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat. Genet.*, **29**, 365–371.
11. Kwon,Y., Shigemoto,Y., Kuwana,Y. and Sugawara,H. (2009) Web API for biology with a workflow navigation system. *Nucleic Acids Res.*, **37**, W11–W16.