*D22–D27* *Nucleic Acids Research, 2011, Vol. 39, Database issue* *Published online 9 November 2010*
*doi:10.1093/nar/gkq1041*

# DDBJ progress report

**Eli Kaminuma, Takehide Kosuge, Yuichi Kodama, Hideo Aono, Jun Mashima, Takashi Gojobori, Hideaki Sugawara, 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 27, 2010; Revised October 7, 2010; Accepted October 11, 2010

## ABSTRACT

**The DNA Data Bank of Japan (DDBJ, http://www.ddbj.nig.ac.jp) provides a nucleotide sequence archive database and accompanying database tools for sequence submission, entry retrieval and annotation analysis. The DDBJ collected and released 3 637 446 entries/2 272 231 889 bases between July 2009 and June 2010. A highlight of the released data was archive datasets from next-generation sequencing reads of Japanese rice cultivar, Koshihikari submitted by the National Institute of Agrobiological Sciences. In this period, we started a new archive for quantitative genomics data, the DDBJ Omics aRchive (DOR). The DOR stores quantitative data both from the microarray and high-throughput new sequencing platforms. Moreover, we improved the content of the DDBJ patent sequence, released a new submission tool of the DDBJ Sequence Read Archive (DRA) which archives massive raw sequencing reads, and enhanced a cloud computing-based analytical system from sequencing reads, the DDBJ Read Annotation Pipeline. In this article, we describe these new functions of the DDBJ databases and support tools.**

## INTRODUCTION

The DNA Data Bank of Japan (DDBJ, http://www.ddbj.nig.ac.jp) is one of the three databanks that comprise the DDBJ/EMBL-Bank/GenBank International Nucleotide Sequence Database (INSD), which was established through close collaboration with the European Bioinformatics Institute (EBI) in Europe and the National Center for Biotechnology Information (NCBI) in the USA. The DDBJ is administered by the Center for Information Biology and DDBJ (CIB-DDBJ) of the National Institute of Genetics (http://www.nig.ac.jp/index-e.html) with funding endorsement from the Japanese Ministry of Education, Culture, Sports, Science and Technology (MEXT). All researchers can submit their data to one of the three summit databanks to register it with the INSD. The enrolled data are exchanged every day, so the three collaborating databanks share virtually the same data at any given time. In this report, we introduce advances in the DDBJ databases and related services based on the updates from last year. Specifically, we describe new entries in the DDBJ databases, a new archive database for quantitative biological data and improvements in the database services of data registration, analysis and retrieval. All resources are available from http://www.ddbj.nig.ac.jp/index-e.html, and archive data can be downloaded at ftp://ftp.ddbj.nig.ac.jp/.

## RECENT DATA GROWTH IN DDBJ DATABASES

### Development of DDBJ primary databases

Here, we introduce the development of the DDBJ database as a member of the INSD during the course of 2010, which is appended to last year's report (1). Between July 2009 and June 2010, 14 296 738 entries/10 572 329 252 bp were released from the 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 (2). The DDBJ collected and released original data consisting of 3 637 446 entries/2 272 231 889 bp during this period. The DDBJ contributed 25.4% of the entries and 21.5% of the base pairs added to the INSD during this period. Most of the nucleotide data were submitted from Japanese researchers; the rest came from China, Korea, Taiwan and other countries. The DDBJ has also continually distributed patent data transferred from the Japan Patent Office (JPO, http://www.jpo.go.jp/index.htm) and the Korean Intellectual Property Office (KIPO, http://www.kipo.go.kr/en/). In addition to the core nucleotide data, the DDBJ has released a total of 2 049 801 WGS entries, 35 270 259 MGA entries and 628 TPA entries as of 21 September 2010.

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

**Table 1.** List of large-scale data released from the DDBJ from July 2009 to June 2010

| Type | Organism | Accession number (number of entries) |
|---|---|---|
| Genome (update) | Rice (*Oryza sativa* Japonica Group, cultivar Nipponbare) | AP008207–AP008218 (12) |
| EST | Eggplant (*Solanum melongena*) | FS000001–FS098086 (98 086) |
| EST | Turkey berry (*Solanum torvum*) | FS098087–FS126465 (28 379) |
| WGS | Tomato (*Solanum lycopersicum*) | BABP01000001–BABP01100783 (100 783) |
| WGS | *Vigna radiata* | BABL01000001–BABL01046645 (46 645) |
| MGA | *Drosophila melanogaster* | AMAAA0000001–AMAAA0023916 (23 916); AMAAB0000001–AMAAB0024086 (24 086) |
| GSS | False killer whale (*Pseudorca crassidens*) | DE647769–DE737775 (90 007) |
| EST | *Sus scrofa* | FS639971–FS722296 (82 326) |
| MGA | *Mus musculus* | ANAAA0000001–ANAAA0033164 (33 164); ANAAB0000001–ANAAB0051442 (51 442) |
| GSS | Tomato (*Solanum lycopersicum*) | FT227487–FT321168 (93 682) |
| GSS | *Mus musculus domesticus* | DH839446–DH961576 (122 131) |
| EST | Purple witchweed (*Striga hermonthica*) | FS438984–FS506797 (67 814) |
| WGS scaffold DRA | Rice (*Oryza sativa* Japonica Group, cultivar Koshihikari) | BABO01000001–BABO01654543 (654 543); DG000025–DG000036(12); DRA000010 |
| MGA | *Mus musculus* | ALAAA0000001–ALAAA0130942 (130 942); ALAAB0000001–ALAAB0116883 (116 883); ALAAC0000001–ALAAC0092019 (92 019); ALAAD0000001–ALAAD0057749 (57 749) |
| GSS EST | Fission yeast (*Schizosaccharomyces pombe*) | FT321169–FT434719 (113 551); FY072959–FY174037 (101 079) |

In addition to traditional nucleotide data, the DDBJ has released raw sequencing data output from the DDBJ Trace Archive (DTA, http://trace.ddbj.nig.ac.jp/dta/dta_index_e.shtml) and the DDBJ Sequence Read Archive (DRA, http://trace.ddbj.nig.ac.jp) (3). The DTA contains raw sequencing data obtained from gel/capillary platforms such as Applied Biosystems ABI 3730. The DRA collects raw data from next-generation sequencing platforms (NGSes). As of 21 September 2010, the DDBJ has released 2 DTA submissions and 32 DRA submissions.

Noteworthy large-scale data released from the DDBJ are listed in Table 1. The genome data for Japanese rice, *Oryza sativa* Japonica Group cultivar Nipponbare, underwent an important update (4). With this update, the rice chromosome version was changed from build 3 to build 4, and approximately 28 000 coding sequence features were annotated to the genome as a result of the Second Rice Annotation Project Meeting (RAP2), managed by the International Rice Genome Sequencing Project (5) (http://rapdb.dna.affrc.go.jp/). The DDBJ released WGS, scaffold and DRA datasets obtained from the Japanese rice cultivar Koshihikari, which were submitted by the National Institute of Agrobiological Sciences (6). Koshihikari is the most popular rice in Japan; it has occupied the most cultivated rice acreage for the last 30 years. Since the genome sequence of the japonica rice cultivar Nipponbare has been reported, the Koshihikari data will be useful for comparative genome analysis among rice cultivars.

Moreover, the DDBJ has released the expressed sequence tags (ESTs) of eggplant (*Solanum melongena*) and turkey berry (*Solanum torvum*) submitted by the National Institute of Vegetable and Tea Science, Japan; the tomato (*Solanum lycopersicum*) WGS submitted by the Kazusa DNA Research Institute; the *Vigna radiata* WGS submitted by the National Center for Genetic Engineering and Biotechnology, Thailand; the *Drosophila melanogaster* MGA submitted by the National Institute of Advanced Industrial Science and Technology, Japan; the false killer whale (*Pseudorca crassidens*) genome survey sequence (GSS) submitted by the Korea Research Institute of Bioscience and Biotechnology; the *Sus scrofa* EST submitted by the National Institute of Agrobiological Sciences, Japan; the *Mus musculus* MGA submitted by Kinki University, Japan; the tomato (*Solanum lycopersicum*) GSS submitted by the University of Tsukuba, Japan; the *Mus musculus domesticus* GSS submitted by the RIKEN BioResource Center; the purple witchweed (*Striga hermonthica*) EST submitted by RIKEN; the *Mus musculus* MGA submitted by the National Institute of Neuroscience, Japan and both GSS and EST data for fission yeast (*Schizosaccharomyces pombe*) from Osaka City University, Japan.

## Development of DDBJ secondary databases

A secondary database is constructed by re-analyzing or modifying the primary data consisting of nucleotide sequence flat files released from the INSD. The DDBJ provides users with various types of secondary databases. DDBJ Amino Acid Database (DAD) records amino acid sequences extracted from values of /translation qualifiers in the nucleotide flat files. The DAD consists of 17 348 613 entries (4 825 871 820 amino acids) as of June 2010. Gene Trek in Prokaryote Space (GTPS, http://gtps.ddbj.nig.ac.jp/) (7) is a prokaryotic genome database that has been re-annotated by a sophisticated common protocol. GTPS assigns reliability grades to entire re-annotated protein-coding genes according to the result of blast and motif scans. GTPS can predict genes that are not annotated originally. As of 21 September 2010, GTPS

contains 862 bacterial (Archaea and Eubacteria) genomes. The GTPS database is updated once a year. In addition to this periodic update, the entire GTPS dataset will soon be updated, and about 1000 bacterial genomes will be re-annotated. The Genome Information Broker (GIB, http://gib.genes.nig.ac.jp/) (8) is a comprehensive data repository of complete microbial genomes in the public domain. It collects complete bacterial genome sequences and annotations soon after the data are available in the INSD. As of 21 September 2010, 1238 genomes of bacterial strains are stored in the GIB. The Genome Information Broker for Viruses (GIB-V, http://gib-v.genes.nig.ac.jp/) (9) is a repository of complete virus genomes or segment data automatically collected from INSD release data. As of 21 September 2010, 69 294 viral genomes and segments can be obtained from the GIB-V. Genomes TO Protein structures and function (GTOP) (10) is a database consisting of data analyses of proteins identified by genome projects. The GTOP database mainly uses sequence homology analysis and information on 3D structures.

## NOVEL DATABASE SERVICES

### DDBJ Omics Archive: new archive for quantitative genomics data

Next-generation sequencing platforms are gradually replacing the DNA microarray for measuring molecular abundances at the genomic scale. To accommodate quantitative genomics data from traditional and new platforms, the DDBJ has decided to launch a new archival database, the DDBJ Omics aRchive (DOR). The DOR has agreed to collaborate with ArrayExpress at the EBI to exchange data. The DOR archives data in compliance with two international guidelines, Minimum Information about a High-Throughput Sequencing Experiment (MINSEQE) and Minimum Information about a Microarray Experiment (MIAME), as ArrayExpress does (11). As NGSes are used to quantify DNA/RNA molecules, researchers submit their raw data to the DRA and their processed data to the DOR. The DOR will establish a submission brokering system in which researchers deposit necessary data sets to the DOR, and the raw data are automatically registered to the DRA.

The DDBJ has supported a microarray database, the Center for Information Biology Gene Expression database (CIBEX, http://cibex.nig.ac.jp) (12) by providing maintenance service. The DOR integrates CIBEX data and exports the data to ArrayExpress (13).

## IMPROVEMENTS IN DATABASE SERVICES

### DDBJ patent sequence database: amino acid entries

We also enhanced the patent sequence databases and added new search and download services. The DDBJ has maintained two patent sequence databases, JPO and KIPO, which consist of patent sequences originally included in patent publications. These databases supply information in DDBJ patent flatfile format and include patent information since 1997. The first enhancement of the DDBJ patent sequence databases was to add taxonomic organism information to all DDBJ flatfiles. This was done as follows. First, the NCBI taxonomy ID was added as the taxon in the /db_xref qualifier on the basis of the original organism names in data submitted to JPO and KIPO. Next, individual organism qualifiers/organism, were added using taxon information. Then, the common name on the SOURCE line was similarly appended. Finally, the lineage on the ORGANISM line was constructed. As a result, we have released the revised nucleic acid and amino acid entries for JPO and KIPO beginning in May 2010. The organism information will be updated once per year. The second enhancement of these databases is to append amino acid patent sequences for anonymous FTP download and BLAST search. Beginning February 2010, DDBJ's anonymous FTP site, ftp://ftp.ddbj.nig.ac.jp/ddbj_database/patent/, has provided the cumulated amino acid sequences for JPO and KIPO with gzip compression. Moreover, beginning August 2010, DDBJ's BLAST search service has contained databases of amino acid sequences from the JPO, KIPO, EPO (European Patent Office) and USPTO (United States Patent and Trademark Office).

### MetaDefine: new web-based submission tool of DRA

The DRA (http://trace.ddbj.nig.ac.jp) is an archival database for data from the primary analysis phase of next-generation sequencing. The DRA has constructed a world-wide Sequence Read Archive (SRA) by mirroring data with partner databases at the NCBI and EBI (3). To support efficient data submission to the DRA, we implemented a novel web-based metadata creation tool, MetaDefine, for the submission system (Figure 1). Instead of generating complex XML files, submitters can create metadata by simply entering the necessary information in the MetaDefine interface. Submitters can also create metadata by editing an existing submission or a template (e.g. Transcriptome) offered by the tool. MetaDefine validates the metadata's format and content and displays detailed error messages to help users prepare valid metadata.

DRA provides all public SRA data by FTP in FASTQ files (ftp://ftp.ddbj.nig.ac.jp/ddbj_database/dra/). A data retrieval system will soon be available. The DRA continues to automate the submission process and support data from new sequencers as a platform database.

### DDBJ Read Annotation Pipeline: functional reinforcement

The DDBJ Read Annotation Pipeline (DDBJ Pipeline, http://p.ddbj.nig.ac.jp) annotates NGS raw sequencing reads with high-throughput (Figure 2). The DDBJ Pipeline is a cloud computing-based pipeline analysis system, so users can access National Institute of Genetics (NIG) supercomputers through a web application with a graphical user interface. The DDBJ Pipeline consists of two processes: a basic analytical process for reference sequence mapping and *de novo* assembly, and an analytical process for structural and functional

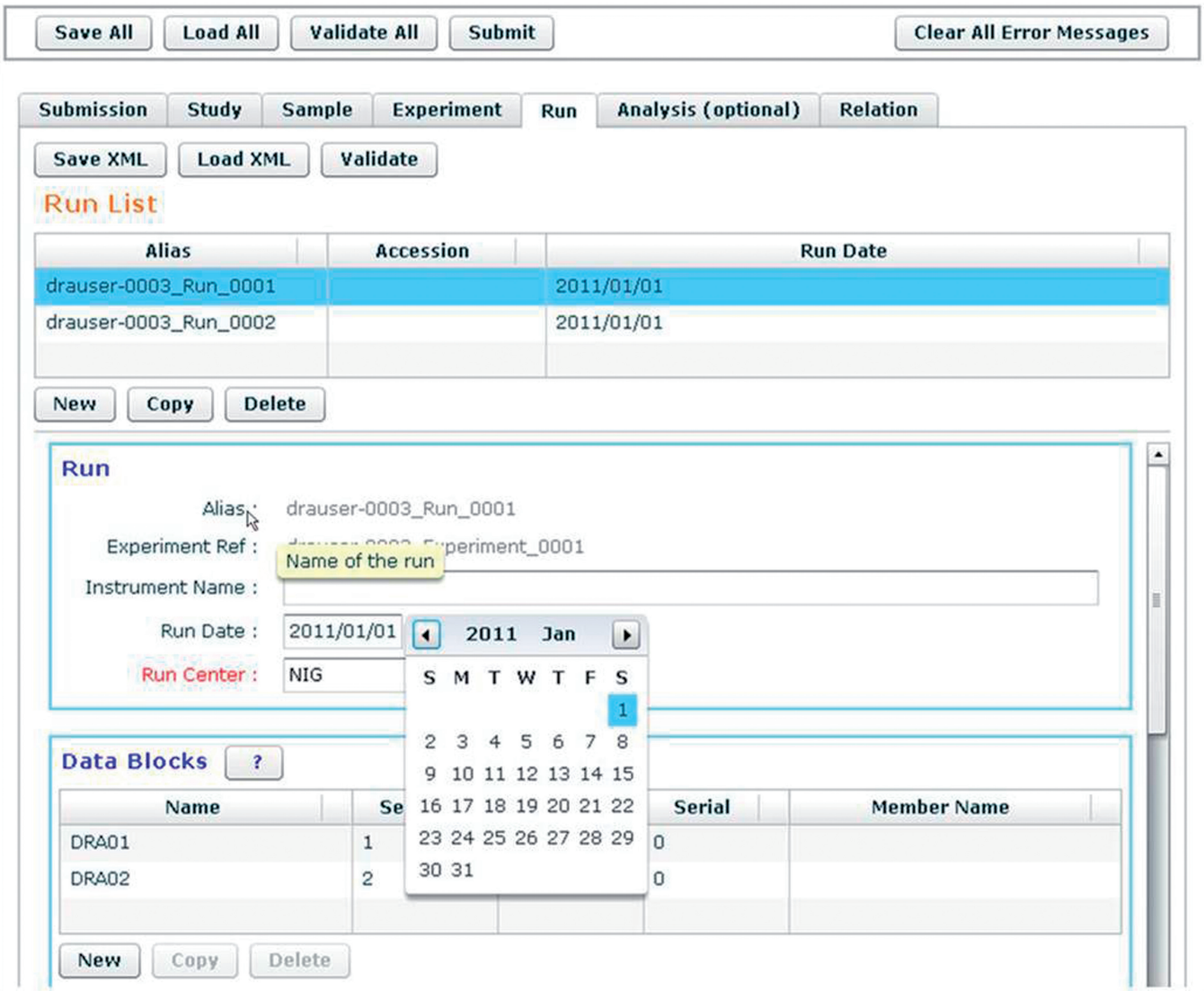


**Figure 1.** Web-based metadata creation tool MetaDefine implemented in DRA submission system.

annotations. In the basic process, popular mapping and assembly tools are available, and reference sequences can be retrieved by INSDC accession from the DDBJ database using SOAP access (14). Analytical statistics on the mapping error rate by read positions, ratio of mapped reads, coverage, sequence depth and maximum contig length can be generated. The analytical process is currently implemented, and several tools for detecting single nucleotide polymorphisms (SNPs) and deletion/insertion polymorphisms (DIPs) are available. The DDBJ Pipeline was enhanced by release of the following functions.

(i) Color-space support.
(ii) Ability to upload local query files (FASTA format) without DRA registration.
(iii) Ability to upload original reference sequences.
(iv) Email notification of job completion/errors.
(v) Improved computational performance.
(vi) Security support.

The color-space analytical tools of LifeTech SOLiD are supported. In addition to the ability to start from FASTQ-formatted files, which are generated from DRA raw files, the pipeline enables the use of users' original FASTA-formatted files to start without registration by uploading from a local computer. Data registration to the DRA is not required for trial data, and thus the DRA issues a temporary accession number. Moreover, original reference sequences can be uploaded. When jobs are finished or aborted, users receive email notification. The infrastructure has been improved from 128 GB of memory to 256 GB for *de novo* assembly. To enhance security, Hypertext Transfer Protocol Secure (HTTPS) has been made available.

In *de novo* assembly, a metadata file of the DDBJ WGS category is automatically generated for convenience in subsequent submissions to the DDBJ database. In the future, the DDBJ Pipeline will also support RNA-seq workflow and submission to the DOR database.

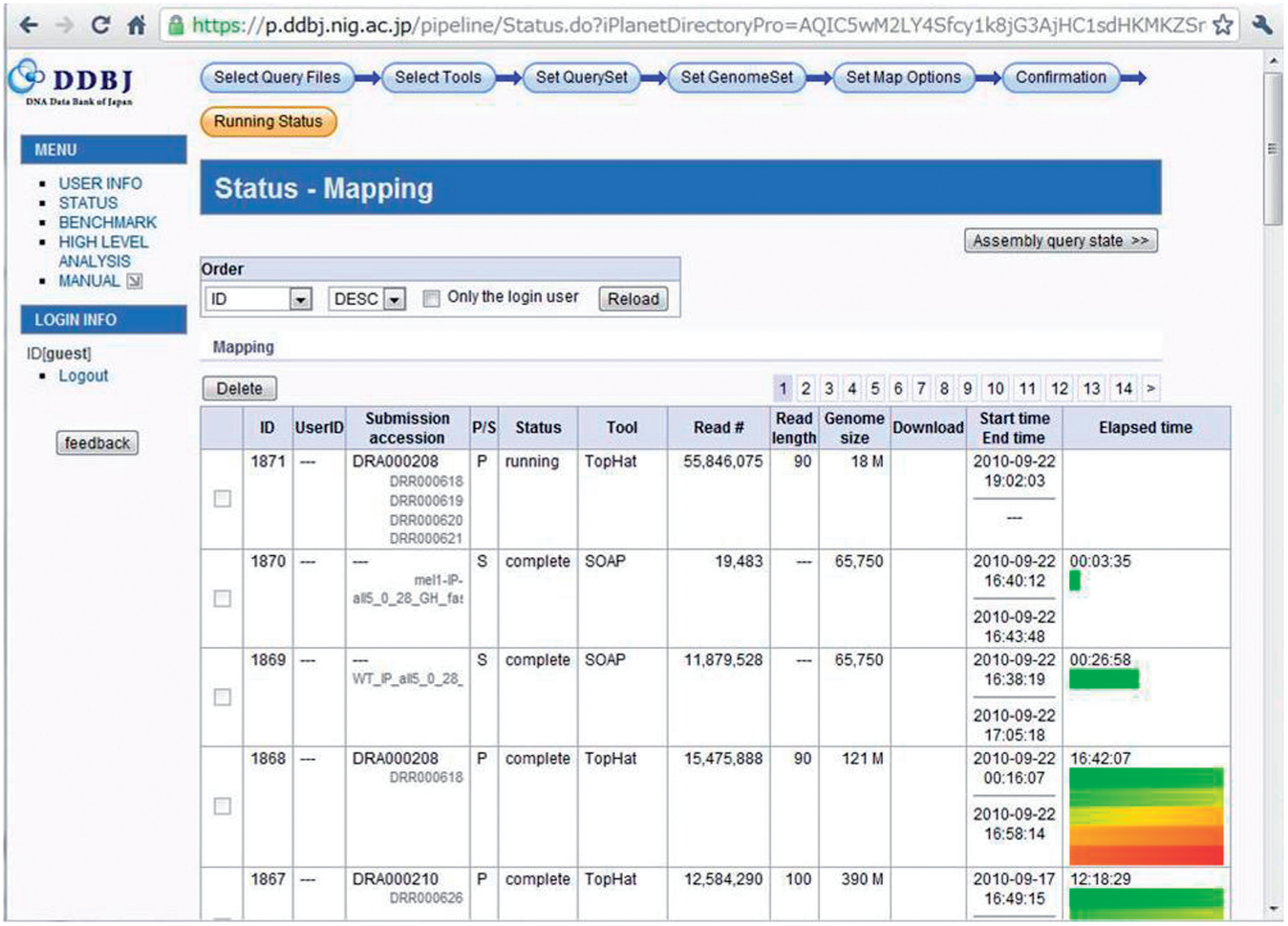


**Figure 2.** Cloud computing-based NGS analytical tool DDBJ Pipeline. Present running status of all jobs can be viewed in the Status panel.

## ACKNOWLEDGEMENTS

We gratefully acknowledge the support of all members of the DDBJ for data collection, annotation and release and for software development. In particular, we thank Takako Mochizuki, Dr Hideki Nagasaki, Dr Toshihisa Okido and Dr Satoshi Saruhashi for consultations on the DRA and the pipeline and Professor Yoshio Tateno for support in the form of database maintenance and INSD collaboration.

## FUNDING

Ministry of Education, Culture, Sports, Science and Technology of Japan with a management expense grant for national university cooperation (to DDBJ); Integrated Database Project (http://lifesciencedb.jp/en) of the Database Center for Life Science in Japan (to DDBJ Trace Archive, DDBJ Sequence Read Archive and DDBJ Pipeline, partially); Institute for Bioinformatics Research and Development, Japan Science and Technology Agency (to DDBJ Trace Archive, DDBJ Sequence Read Archive and DDBJ Pipeline, partially). Funding for open access charge: The DDBJ management expenses grant.

*Conflict of interest statement*. None declared.

## REFERENCES

1. 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.
2. 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.
3. Shumway,M., Cochrane,G. and Sugawara,H. (2010) Archiving next generation sequencing data. *Nucleic Acids Res.*, **38**, D870–D871.
4. Rice Annotation Project. (2008) The Rice Annotation Project Database (RAP-DB): 2008 update. *Nucleic Acids Res.*, **36**, D1028–D1033.
5. International Rice Genome Sequencing Project. (2005) The map-based sequence of the rice genome. *Nature*, **436**, 793–800.
6. Yamamoto,T., Nagasaki,H., Yonemaru,J., Ebana,K., Nakajima,M., Shibaya,T. and Yano,M. (2010) Fine definition of the pedigree haplotypes of closely related rice cultivars by means of genome-wide discovery of single-nucleotide polymorphisms. *BMC Genomics*, **11**, 267.
7. Kosuge,T., Abe,T., Okido,T., Tanaka,N., Hirahata,M., Maruyama,Y., Mashima,J., Tomiki,A., Kurokawa,M., Himeno,R. *et al.* (2006) Exploration and grading of possible genes from 183 bacterial strains by a common protocol to identification of new

genes: Gene Trek in Prokaryote Space (GTPS). *DNA Res.*, **13**, 245–254.
8. Fumoto,M., Miyazaki,S. and Sugawara,H. (2002) Genome Information Broker (GIB): data retrieval and comparative analysis system for completed microbial genomes and more. *Nucleic Acids Res.*, **30**, 66–68.
9. Hirahata,M., Abe,T., Tanaka,N., Kuwana,Y., Shigemoto,Y., Miyazaki,S., Suzuki,Y. and Sugawara,H. (2007) Genome Information Broker for Viruses (GIB-V): database for comparative analysis of virus genomes. *Nucleic Acids Res.*, **35**, D339–D342.
10. Fukuchi,S., Homma,K., Sakamoto,S., Sugawara,H., Tateno,Y., Gojobori,T. and Nishikawa,K. (2009) The GTOP database in 2009: updated content and novel features to expand and deepen insights into protein structures and functions. *Nucleic Acids Res.*, **37**, D333–D337.
11. Parkinson,H., Kapushesky,M., Kolesnikov,N., Rustici,G., Shojatalab,M., Abeygunawardena,N., Berube,H., Dylag,M., Emam,I., Farne,A. *et al.* (2009) ArrayExpress update–from an archive of functional genomics experiments to the atlas of gene expression. *Nucleic Acids Res.*, **37**, D868–D872.
12. Ikeo,K., Ishi-i,J., Tamura,T., Gojobori,T. and Tateno,Y. (2003) CIBEX: center for information biology gene expression database. *C. R. Biol.*, **326**, 1079–1082.
13. 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.
14. Yeondae,K., Yasumasa,S., Yoshikazu,K. and Hideaki,S. (2009) Web API for biology with a workflow navigation system. *Nucleic Acids Res.*, **37**, W11–W16.