Published online 24 October 2018 *Nucleic Acids Research, 2019, Vol. 47, Database issue* ***D69–D73***
doi: 10.1093/nar/gky1002

DDBJ update: the Genomic Expression Archive (GEA) for functional genomics data

Yuichi Kodama*, Jun Mashima, Takehide Kosuge and Osamu Ogasawara*

DDBJ Center, National Institute of Genetics, Shizuoka 411-8540, Japan

Received September 14, 2018; Revised October 03, 2018; Editorial Decision October 08, 2018; Accepted October 09, 2018

ABSTRACT

The Genomic Expression Archive (GEA) for functional genomics data from microarray and high-throughput sequencing experiments has been established at the DNA Data Bank of Japan (DDBJ) Center (https://www.ddbj.nig.ac.jp), which is a member of the International Nucleotide Sequence Database Collaboration (INSDC) with the US National Center for Biotechnology Information and the European Bioinformatics Institute. The DDBJ Center collects nucleotide sequence data and associated biological information from researchers and also services the Japanese Genotype–phenotype Archive (JGA) with the National Bioscience Database Center for collecting human data. To automate the submission process, we have implemented the DDBJ BioSample validator which checks submitted records, auto-corrects their format, and issues error messages and warnings if necessary. The DDBJ Center also operates the NIG supercomputer, prepared for analyzing large-scale genome sequences. We now offer a secure platform specifically to handle personal human genomes. This report describes database activities for INSDC and JGA over the past year, the newly launched GEA, submission, retrieval, and analysis services available in our supercomputer system and their recent developments.

INTRODUCTION

The DNA Data Bank of Japan (DDBJ, https://www.ddbj.nig.ac.jp) (1) is a public nucleotide sequence database established at the National Institute of Genetics (NIG, https://www.nig.ac.jp). Since 1987, the DDBJ Center has been collecting annotated nucleotide sequences as its traditional database service. This endeavor is conducted in collaboration with GenBank (2) at the National Center for Biotechnology Information (NCBI) and with the European Nucleotide Archive (ENA) (3) at the European Bioinformatics Institute (EBI). This collaborative framework is known as the International Nucleotide Sequence Database Collaboration (INSDC) (4), and its product database is called the International Nucleotide Sequence Database (INSD).

Within the INSDC framework, the DDBJ Center also services the DDBJ Sequence Read Archive (DRA) for raw sequencing data and alignment information from high-throughput sequencing platforms (5), BioProject for sequencing project metadata, and BioSample for sample information (1,6). This comprehensive resource of nucleotide sequences and associated biological information complies with INSDC policy guaranteeing free and unrestricted access to data archives (7).

In July 2018, the DDBJ Center launched a new public database, the Genomic Expression Archive (GEA, https://www.ddbj.nig.ac.jp/gea), which collects functional genomics data from microarray and high-throughput sequencing experiments. Besides the Gene Expression Omnibus (GEO) at the NCBI (8) and ArrayExpress at the EBI (9), the GEA issues accession numbers to functional genomics experiments, whose data are associated with metadata in a structured and standardized MAGE-TAB format (10), and public GEA data will be indexed by ArrayExpress. For publications under review, submitters can allow journal reviewers anonymous access to private GEA data cited in their manuscripts. With the GEA launch, the DDBJ Center now covers the archiving of sequences with functional annotation (traditional database) and molecular abundance (GEA).

In addition to these unrestricted-access databases, the DDBJ Center also services a controlled-access database, the Japanese Genotype–phenotype Archive (JGA, https://www.ddbj.nig.ac.jp/jga), in collaboration with the National Bioscience Database Center (NBDC, https://biosciencedbc.jp/en/) at the Japan Science and Technology Agency (1,11). The JGA stores genotype and phenotype data from human individuals who have signed consent agreements authorizing data usage for specific research only. The NBDC provides guidelines and policies for sharing human-derived data (https://humandbs.biosciencedbc.jp/en/guidelines) and reviews data submission and usage requests.

*To whom correspondence should be addressed. Tel: +81 55 981 6853; Fax: +81 55 981 6849; Email: ykodama@nig.ac.jp
Correspondence may also be addressed to Osamu Ogasawara. Tel: +81 55 981 6853; Fax: +81 55 981 6849; Email: oogasawa@nig.ac.jp

The DDBJ Center operates a commodity-based computer cluster, the NIG supercomputer. It provides an analytical environment for domestic researchers to examine large-scale biology data from its archival databases with preinstalled bioinformatics tools.

In the present article, we provide an update of the above services at the DDBJ Center, highlighting the new database resource: the GEA. All resources mentioned in this article are available at https://www.ddbj.nig.ac.jp, and most of the archival data can be downloaded from ftp://ftp.ddbj.nig.ac.jp. Major database and supercomputer services are summarized in Figure 1.

DDBJ ARCHIVAL DATABASES

Data contents: unrestricted- and controlled-access databases

From June 2017 to May 2018, the traditional DDBJ accepted 6011 nucleotide data submissions consisting of 5 926 950 entries, most of which were from Japanese research groups (4784 submissions; 79.6%). The DDBJ periodical release includes both conventional sequence and bulk sequence data such as whole-genome shotgun (WGS), transcriptome shotgun assembly (TSA), and targeted locus study (TLS) but does not include third-party data (TPA) records (12). Between June 2017 and May 2018, the DDBJ periodical release increased from 689 916 250 to 1 564 840 159 entries and from 1 333 798 893 388 to 3 795 161 222 944 base pairs. The DDBJ contributed 4.67% of the entries and 3.25% of the total base pairs in the INSD nucleotide sequence data. A detailed statistical breakdown of records is present on the DDBJ website (https://www.ddbj.nig.ac.jp/stats/release-e.html#total_data).

During June 2017 and May 2018, high-throughput sequencing data consisting of 43 174 runs have been registered to the DRA. As of 24 August 2018, the DRA has distributed 3.8 PB of sequencing data in SRA (2.7 PB) and FASTQ (1.1 PB) formats.

The JGA is a controlled-access database for human genotype and phenotype data (11) similar to the database of Genotypes and Phenotypes (dbGaP) at the NCBI (13) and the European Genome-phenome Archive at the EBI (14). As of 24 August 2018, the JGA has archived 146 studies, 248 043 samples, and 96 TB of individual-level human datasets submitted by Japanese researchers. The summaries of 88 studies are publicly available both on the JGA (https://ddbj.nig.ac.jp/jga/viewer/view/studies) and NBDC (https://humandbs.biosciencedbc.jp/en/data-use/all-researches) websites. A large-scale dataset of the BioBank Japan project (15) is available under JGA study accession number JGAS00000000114. This includes the microarray genotyping data of 182 505 individuals, the whole genome sequencing data of 1026 individuals, and 58 quantitative traits associated with 47 diseases of 200 849 individuals. To access the individual-level data of these public studies, users are required to send data usage requests to the NBDC (https://humandbs.biosciencedbc.jp/en/data-use).

The Genomic Expression Archive

The GEA is a public archive for functional genomics experiments including gene expression, epigenetics, and genome-protein interactions. Along with the GEO and ArrayExpress, the GEA supports two community-derived reporting standards—the Minimum Information About a Microarray Experiment (MIAME) (16) and the Minimum Information about a high-throughput nucleotide SEQuencing Experiment (MINSEQE, http://fged.org/projects/minseqe)—for unambiguous interpretation and data usage. These guidelines specify the provision of critical experimental elements in raw data, processed data, sample annotation, and protocols.

The GEA accepts functional genomics experiments in a structured and standardized MAGE-TAB format adopted by ArrayExpress (10). A submission to the GEA consists of four parts: raw and processed data, as well as associated metadata in Investigation Description Format (IDF) and in Sample and Data Relationship Format (SDRF) (Figure 2). IDF metadata provides an overview of the experiment, including experimental design, protocols, publication information, and submitter details. SDRF metadata provides sample characteristics and the relation between samples, microarray or sequencing platforms, and raw and processed data files.

The submission workflow of microarray and sequencing experiment involves three and four steps, respectively (Figure 2). The microarray submission consists of (i) preregistration of project information to BioProject, (ii) preregistration of sample information to BioSample and (iii) provision of raw and processed data to the GEA (Figure 2). Major commercial array designs at ArrayExpress can be referenced by citing their array design accession numbers. For a novel array design, the submitter needs to register its Array Design Format file to the GEA.

The sequencing submission consists of four steps. (i) preregistration of project information to BioProject, (ii) preregistration of sample information to BioSample, (iii) preregistration of raw data to the DRA and (iv) provision of processed data to the GEA (Figure 2). The submitter may complete the first three steps through a single DRA submission. Raw sequencing data are stored in the DRA and not in the GEA. This submission process is different from that of the GEO (or ArrayExpress), where the database submission pipeline brokers raw sequencing data of the GEO (or ArrayExpress) into NCBI (or EBI) SRA, respectively.

In the GEA submission portal, the submitter creates IDF and SDRF metadata files associated with raw and processed data. In the IDF tab, the submitter provides an overview of the experiment: title, description, type and design; and protocols, array design (microarray), and publications. In the SDRF tab, the submitter verifies a SDRF template generated from the selected BioSample records, DRA submission (sequencing) or array design (microarray), and the IDF. The submitter is also required to include information describing the material and data files obtained from the samples, as well as the experimental variables under investigation. Lastly, in the 'Overview' tab, the submitter confirms the created IDF and SDRF metadata files. Then the system validates the submitted IDF and SDRF files and sends error and warning messages according to validation rules (https://www.ddbj.nig.ac.jp/gea/validation-e.html).

After the metadata and data files are reviewed, the GEA issues unique accession numbers to the experiments and ar-

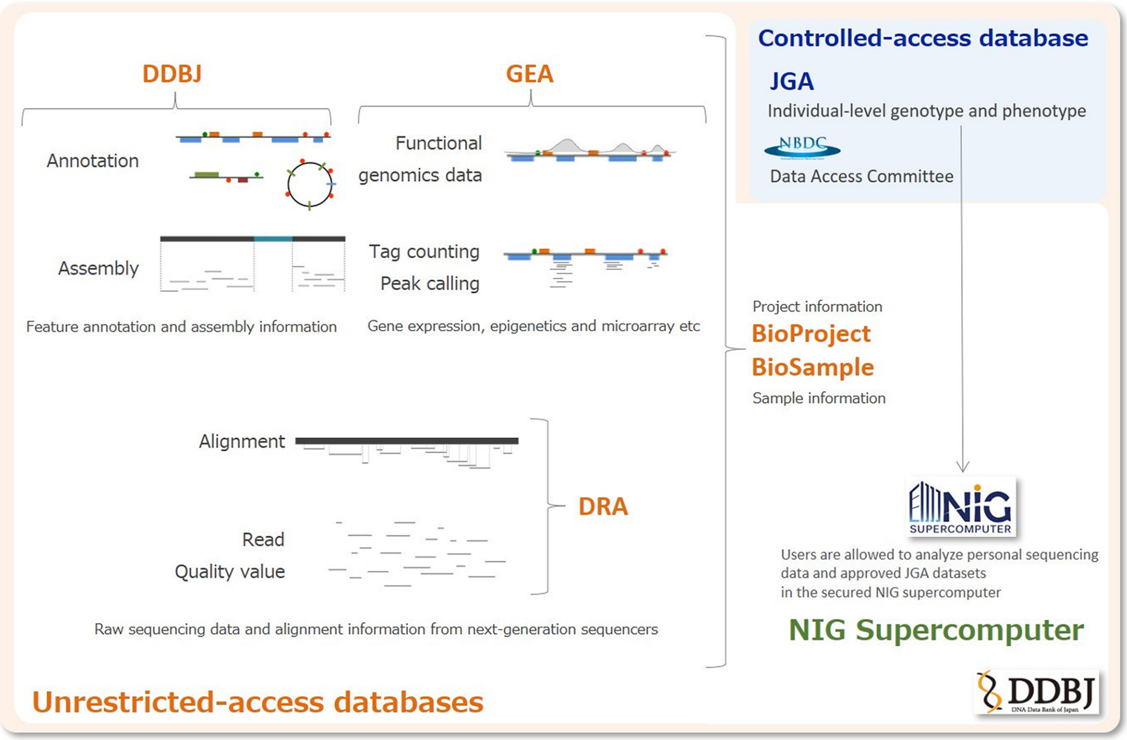
DDBJ
GEA
Controlled-access database
JGA
Individual-level genotype and phenotype
NBDC
Data Access Committee
Annotation
Functional genomics data
Assembly
Tag counting
Peak calling
Feature annotation and assembly information
Gene expression, epigenetics and microarray etc
Project information
BioProject
BioSample
Sample information
Alignment
DRA
Read
Quality value
NIG SUPERCOMPUTER
Users are allowed to analyze personal sequencing data and approved JGA datasets in the secured NIG supercomputer
Raw sequencing data and alignment information from next-generation sequencers
NIG Supercomputer
Unrestricted-access databases
DDBJ
DNA Data Bank of Japan

Figure 1. Database and supercomputer services of the DDBJ Center. The unrestricted-access databases (DRA, DDBJ, GEA, BioProject and BioSample) are illustrated with their data scopes. The controlled-access database (JGA) is operated in collaboration with NBDC, whose Data Access Committee reviews data submission and usage requests according to the NBDC guidelines for sharing human-derived data. The NIG supercomputer is a commodity-based cluster designed for analyzing large-scale sequencing data. In the secure platform designed for analyzing personal human genomes, users are allowed to analyze approved JGA datasets.

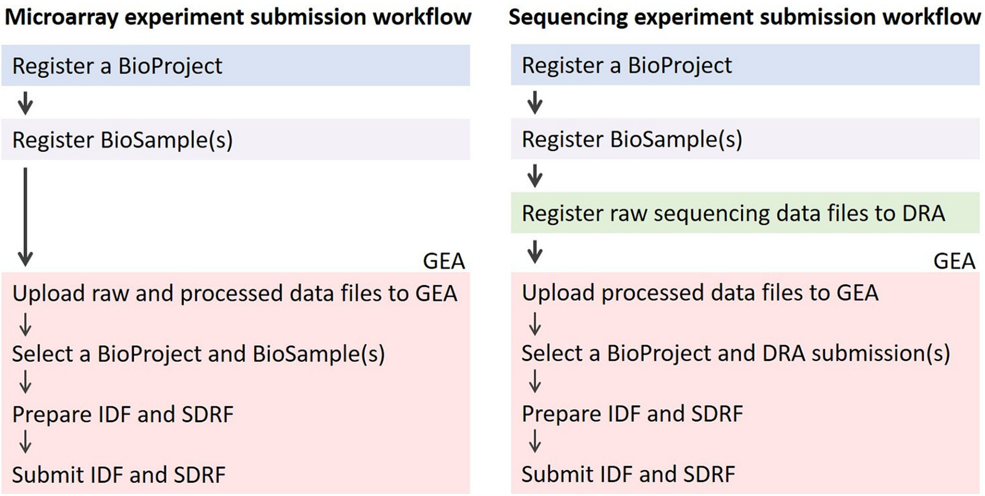
Microarray experiment submission workflow
Register a BioProject
Register BioSample(s)
GEA
Upload raw and processed data files to GEA
Select a BioProject and BioSample(s)
Prepare IDF and SDRF
Submit IDF and SDRF
Sequencing experiment submission workflow
Register a BioProject
Register BioSample(s)
Register raw sequencing data files to DRA
GEA
Upload processed data files to GEA
Select a BioProject and DRA submission(s)
Prepare IDF and SDRF
Submit IDF and SDRF

Figure 2. GEA submission workflow of microarray and sequencing experiments. Registration of BioProject and BioSample(s) are mandatory prior to the submission of both microarray and sequencing experiments. To submit microarray experiments to GEA, provision of raw and processed data files, and associated IDF and SDRF metadata are necessary. Raw data of sequencing experiments are stored in DRA and not in GEA.

ray designs in 'E-GEAD-*n*' and 'A-GEAD-*n*' formats, respectively (*n* being a number). Because the GEA and ArrayExpress share the accession namespace, the GEA numbers are searchable at ArrayExpress as well. The registered data may be kept private for a limited time, typically during the peer-review process of the relevant publication. The submitter can generate an access token for private GEA data in the submission portal for reviewers. The data would then become public either when the accession number associated with the data is published or at the release date designated by the submitter, whichever comes first. Public GEA experiments and array designs are accessible at its FTP (ftp://ftp.ddbj.nig.ac.jp/ddbj_database/gea) and will be indexed by ArrayExpress. In collaboration with the Database Center for Life Science (DBCLS), GEA metadata will become searchable at the DBCLS All Of gene Expression (AOE, http://aoe.dbcls.jp), an interface where public functional genomics data can be searched and viewed. The DDBJ Center also mirrors ArrayExpress data at its FTP site (ftp://ftp.ddbj.nig.ac.jp/mirror_database/arrayexpress).

The DDBJ Center archived 260 functional genomics experiments at the Center for Information Biology gene EXpression database (CIBEX) (17), which operated from 2003 to 2012. The CIBEX data, with their original CIBEX accession numbers, will be available at the GEA.

DDBJ SYSTEM UPDATE

Submission services of biological data

As sequencing technologies are changing and submissions are increasing, our submission processes are also transitioning from manual curation to automatic validation. As part of such efforts, we have implemented the DDBJ BioSample validator and it validates submitted BioSample records according to validation rules, autocorrect formats, and sends error messages and warnings (https://www.ddbj.nig.ac.jp/biosample/validation-e.html). The submitter can correct errors and improve content according to the validation messages displayed in the submission portal. The BioSample submission system automatically assigns accession numbers to submitted records without errors. The BioSample curators have shifted from manual, record-by-record curation to validation rule improvement.

Retrieval and analysis services of biological data

The DDBJ Center provides web interfaces of the getentry data retrieval system by accession numbers and the ARSA keyword search system for traditional DDBJ flat files (18), Web BLAST (19), ClustalW (20), VecScreen (http://ddbj.nig.ac.jp/vecscreen) and TXSearch (http://ddbj.nig.ac.jp/tx_search) services. For programmatic access to the data, the DDBJ Center also provides the Web API for Bioinformatics (WABI) (21), which includes BLAST, VecScreen, ClustalW, MAFFT (22), getentry and ARSA services. We also have the DDBJ Read Annotation Pipeline (https://p.ddbj.nig.ac.jp), which is the web service for analyzing high-throughput DNA sequencing reads (23).

We provide download services for public data in our archival databases at ftp://ftp.ddbj.nig.ac.jp/ddbj_database/ and those in other mirrored databases at ftp://ftp.ddbj.nig.ac.jp/mirror_database/.

The NIG supercomputer

The DDBJ Center also operates the NIG supercomputer, a commodity-based computer cluster designed for storing and analyzing large-scale sequencing data. The NIG supercomputer is composed of calculation nodes for general-purpose (554 thin nodes at 64 GB memory each) and memory-intensive tasks, including *de novo* sequence assembly (10 medium nodes, each with 2 TB of memory and one fat node with 10 TB of memory). The calculation nodes are interconnected with InfiniBand, and the total peak performance of its CPUs is 372 Tflops. To support massive I/O in the big-data analysis, the NIG supercomputer is equipped with 10.6 PB of the Lustre parallel distributed file system (http://lustre.org). Large-scale DRA and JGA data are stored in 30 PB of the hierarchical storage system (the IBM Elastic Storage Server). The NIG supercomputer is also installed with major biological datasets and popular bioinformatics tools. Users can transfer large data between their local computers and the NIG supercomputer by using the high-speed-transfer software IBM Aspera (https://asperasoft.com/).

The ever-increasing volume of personal sequencing data makes it difficult for researchers to prepare their own secure computer resources with sufficient storage and computing power and to transfer large data online from public databases including the JGA. To solve these issues, the DDBJ Center has provided a secured NIG supercomputer environment for analyzing personal sequencing data, which is composed of 32 thin nodes and 440 TB of storage and logically separates each user's computing environment (Figure 1). JGA users are allowed to analyze approved JGA datasets in the secured NIG supercomputer as an 'off-premise server' in addition to their own servers. Because the secured NIG supercomputer is connected with the JGA server through a high-speed network, users can smoothly download JGA datasets and analyze their own personal genomic data in the same environment. We charge users of the secured supercomputer to share operation and security costs with them (https://sc.ddbj.nig.ac.jp/ja/application/individual-genome-analysis-system, Japanese text only).

FUTURE DIRECTION

To process the growing number of data submissions and access requests to the NBDC and the JGA, the DDBJ Center and the NBDC are developing an integrated system which will seamlessly connect the NBDC request application processes and the JGA data submission/download processes by using the same user account system.

To manage the increasing volume of sequencing submissions despite limited resources and budget, we are shifting from manual, record-by-record curation to a more automatic validation-based processing system. We are developing validators for BioProject and the DRA and will implement them to the submission systems to automate submission processes. We are also working to automate bacterial genome submissions by using the DDBJ Fast Annotation and Submission Tool (DFAST, https://dfast.nig.ac.jp), which is a bacterial genome annotation pipeline producing submission-ready DDBJ annotation files (24).

ACKNOWLEDGEMENTS

We gratefully acknowledge Koji Watanabe, Naofumi Ishikawa, Yasukazu Nakamura, Kousaku Okubo, Toshihisa Takagi, Masanori Arita and all members of the DDBJ Center for their support in data collection, annotation, release and software development. We are thankful to Minae Kawashima, Licht Toyo-oka and Nobutaka Mitsuhashi of the NBDC as collaborators of the JGA project; Takatomo Fujisawa for validator development; Eli Kaminuma for the DDBJ Read Annotation Pipeline development; Yasuhiro

Tanizawa for DFAST collaboration; Yoshihiro Okuda for taxonomy search; and Hidemasa Bono of the DBCLS and Amy Tang, Ugis Sarkans and Robert Petryszak of the EBI for the GEA service launch. We would also like to thank Kento Aida, Shigetoshi Yokoyama, and Nobuyoshi Masatani of the National Institute of Information and Shinichi Miura and Satoshi Matsuoka of the Tokyo Institute of Technology for the computational infrastructure of the NIG supercomputer.

FUNDING

The DDBJ Center was supported by the Ministry of Education, Culture, Sports, Science and Technology of Japan (MEXT) and Research Organization of Information and Systems (ROIS); NBDC at the Japan Science and Technology Agency (JST) (to JGA); the NIG supercomputer infrastructure technology on inter-cloud was supported by JST CREST [JPMJCR1501], Japan. Funding for open access charge: MEXT grant to the DDBJ Center.

Conflict of interest statement. None declared.

REFERENCES

1. Kodama,Y., Mashima,J., Kosuge,T., Kaminuma,E., Ogasawara,O., Okubo,K., Nakamura,Y. and Takagi,T. (2018) DNA Data Bank of Japan: 30th anniversary. *Nucleic Acids Res.*, **46**, D30–D35.
2. Benson,D.A., Cavanaugh,M., Clark,K., Karsch-Mizrachi,I., Ostell,J., Pruitt,K.D. and Sayers,E.W. (2018) GenBank. *Nucleic Acids Res.*, **46**, D41–D47.
3. Silvester,N., Alako,B., Amid,C., Cerdeño-Tarrága,A., Clarke,L., Cleland,I., Harrison,P.W., Jayathilaka,S., Kay,S., Keane,T. *et al.* (2018) The European Nucleotide Archive in 2017. *Nucleic Acids Res.*, **46**, D36–D40.
4. Cochrane,G., Karsch-Mizrachi,I., Takagi,T. and Sequence Database Collaboration, I.N. (2016) The International Nucleotide Sequence Database Collaboration. *Nucleic Acids Res.*, **44**, D48–D50.
5. Kodama,Y., Shumway,M. and Leinonen,R. (2012) The Sequence Read Archive: explosive growth of sequencing data. *Nucleic Acids Res.*, **40**, D54–D56.
6. Federhen,S., Clark,K., Barrett,T., Parkinson,H., Ostell,J., Kodama,Y., Mashima,J., Nakamura,Y., Cochrane,G. and Karsch-Mizrachi,I. (2014) Toward richer metadata for microbial sequences: replacing strain-level NCBI taxonomy taxids with BioProject, BioSample and Assembly records. *Stand. Genomic Sci.*, **9**, 1275–1277.
7. Brunak,S., Danchin,A., Hattori,M., Nakamura,H., Shinozaki,K., Matise,T. and Preuss,D. (2002) Nucleotide sequence database policies. *Science*, **298**, 1333.
8. Clough,E. and Barrett,T. (2016) The Gene Expression Omnibus Database. In: *Methods in Molecular Biology*. Clifton, Vol. **1418**, pp. 93–110.
9. Kolesnikov,N., Hastings,E., Keays,M., Melnichuk,O., Tang,Y.A., Williams,E., Dylag,M., Kurbatova,N., Brandizi,M., Burdett,T. *et al.* (2015) ArrayExpress update–simplifying data submissions. *Nucleic Acids Res.*, **43**, D1113–D1116.
10. Rayner,T.F., Rocca-Serra,P., Spellman,P.T., Causton,H.C., Farne,A., Holloway,E., Irizarry,R.A., Liu,J., Maier,D.S., Miller,M. *et al.* (2006) A simple spreadsheet-based, MIAME-supportive format for microarray data: MAGE-TAB. *BMC Bioinformatics*, **7**, 489.
11. Kodama,Y., Mashima,J., Kosuge,T., Katayama,T., Fujisawa,T., Kaminuma,E., Ogasawara,O., Okubo,K., Takagi,T. and Nakamura,Y. (2015) The DDBJ Japanese Genotype-phenotype Archive for genetic and phenotypic human data. *Nucleic Acids Res.*, **43**, D18–D22.
12. 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.
13. Wong,K.M., Langlais,K., Tobias,G.S., Fletcher-Hoppe,C., Krasnewich,D., Leeds,H.S., Rodriguez,L.L., Godynskiy,G., Schneider,V.A., Ramos,E.M. *et al.* (2017) The dbGaP data browser: a new tool for browsing dbGaP controlled-access genomic data. *Nucleic Acids Res.*, **45**, D819–D826.
14. Lappalainen,I., Almeida-King,J., Kumanduri,V., Senf,A., Spalding,J.D., Ur-Rehman,S., Saunders,G., Kandasamy,J., Caccamo,M., Leinonen,R. *et al.* (2015) The European Genome-phenome Archive of human data consented for biomedical research. *Nat. Genet.*, **47**, 692–695.
15. Nagai,A., Hirata,M., Kamatani,Y., Muto,K., Matsuda,K., Kiyohara,Y., Ninomiya,T., Tamakoshi,A., Yamagata,Z., Mushiroda,T. *et al.* (2017) Overview of the BioBank Japan Project: Study design and profile. *J. Epidemiol.*, **27**, S2–S8.
16. 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.
17. Ikeo,K., Ishi-i,J., Tamura,T., Gojobori,T. and Tateno,Y. CIBEX: center for information biology gene expression database. *C. R. Biol.*, **326**, 1079–1082.
18. Ogasawara,O., Mashima,J., Kodama,Y., Kaminuma,E., Nakamura,Y., Okubo,K. and Takagi,T. (2013) DDBJ new system and service refactoring. *Nucleic Acids Res.*, **41**, D25–D29.
19. 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.
20. Larkin,M.A., Blackshields,G., Brown,N.P., Chenna,R., McGettigan,P.A., McWilliam,H., Valentin,F., Wallace,I.M., Wilm,A., Lopez,R. *et al.* (2007) Clustal W and Clustal X version 2.0. *Bioinformatics*, **23**, 2947–2948.
21. 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.
22. Katoh,K. and Standley,D.M. (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.*, **30**, 772–780.
23. Nagasaki,H., Mochizuki,T., Kodama,Y., Saruhashi,S., Morizaki,S., Sugawara,H., Ohyanagi,H., Kurata,N., Okubo,K., Takagi,T. *et al.* (2013) DDBJ read annotation pipeline: a cloud computing-based pipeline for high-throughput analysis of next-generation sequencing data. *DNA Res.*, **20**, 383–390.
24. Tanizawa,Y., Fujisawa,T. and Nakamura,Y. (2018) DFAST: a flexible prokaryotic genome annotation pipeline for faster genome publication. *Bioinformatics*, **34**, 1037–1039.