D870–D871 *Nucleic Acids Research, 2010, Vol. 38, Database issue* *Published online 3 December 2009*
doi:10.1093/nar/gkp1078

Archiving next generation sequencing data

Martin Shumway[1],*, Guy Cochrane[2] and Hideaki Sugawara[3]

[1]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA, [2]EMBL European Bioinformatics Institute, Hinxton, Cambridge, UK and [3]DNA Data Bank of Japan, National Institute of Genetics, Inter-University Research Institute of Corporation Research Organization of Information and Systems, Shizuoka, Japan

Received November 3, 2009; Accepted November 4, 2009

ABSTRACT

Next generation sequencing platforms are producing biological sequencing data in unprecedented amounts. The partners of the International Nucleotide Sequencing Database Collaboration, which includes the National Center for Biotechnology Information (NCBI), the European Bioinformatics Institute (EBI), and the DNA Data Bank of Japan (DDBJ), have established the Sequence Read Archive (SRA) to provide the scientific community with an archival destination for next generation data sets. The SRA is now accessible at http://www.ncbi.nlm.nih.gov/Traces/sra from NCBI, at http://www.ebi.ac.uk/ena from EBI and at http://www.ddbj.nig.ac.jp/sub/trace_sra-e.html from DDBJ. Users of these resources can obtain data sets deposited in any of the three SRA instances. Links and submission instructions are provided.

TEXT

Next generation sequencing platforms are revolutionizing genomics and genome science. These instruments are producing vastly more sequencing data than was ever possible with capillary technology, providing more power for resolution of genomic variation, reducing clonal bias in amplification and making practicable new assays such as full-length cDNA sequencing on a large scale. In addition, the shift from microarrays to next generation sequencing platforms for gene expression and epigenomics investigations has resulted in much greater resolving power and accuracy for those experiments. The new technologies offer tremendous promise for advancing fundamental knowledge about biology, particularly if the data are made widely available to the researchers. Based on the experience with the Trace Archive (established at NCBI and Wellcome Trust Sanger Institute in 2001 to archive and distribute capillary sequences to the scientific community) (1), NCBI set out in 2007 to design a successor archive to accommodate the next generation sequencing platforms (2). These platforms now include 454 (Roche Diagnostics Corporation, Branford, CT, USA), Illumina Genome Analyzer (Illumina, Inc., San Diego, CA, USA), SOLiD^{TM} (Life Technologies Corporation, Carlsbad, CA, USA), HeliScope (Helicos Biosciences Corporation, Cambridge, MA, USA), Complete Genomics (Complete Genomics Inc., Mountain View, CA, USA) and SMRT^{TM} (Pacific Biosciences Inc., Menlo Park, CA, USA).

The resulting Sequence Read Archive (SRA) is now accessible at www.ncbi.nlm.nih.gov/Traces/sra from NCBI, at http://www.ebi.ac.uk/ena from European Bioinformatics Institute (EBI) and at http://www.ddbj.nig.ac.jp/sub/trace_sra-e.html from DNA Data Bank of Japan (DDBJ). In order to adapt to the much greater output from next generation sequencing platforms, the SRA incorporates several improvements over the Trace Archive, including separation of metadata from the content, institution of a ‘run’ concept to cover the production unit (plate or flowcell) and the creation of a sequencing ‘experiment’ object to describe the sequencing library that the runs belong to.

The SRA data model was designed in collaboration with the EBI and the DDBJ under the auspices of the International Nucleotide Sequence Database Collaboration (INSDC) (http://www.insdc.org). The INSDC’s DDBJ/EMBL/GenBank database has been a critical resource in biomedicine. As new technologies have arisen, be they ESTs or whole genome shotgun records, DDBJ/EMBL/GenBank have adapted and expanded to maintain this valuable international shared resource. The expansion of Trace/SRA into the international collaboration continues the support for a uniform, international path to critical data sharing in biomedicine. The three SRAs will mirror data and share an accession space, essentially providing a world-wide archive. The EBI’s SRA implementation is described in (3) and DDBJ’s in (4).

In November 2009, the SRAs collectively hosted about 11 Terabases of biological sequence data. This included 170 full-length human genomes, over 900 bacterial

*To whom correspondence should be addressed. Tel: +1 301 402 4041; Fax: +1 301 402 9651; Email: shumwaym@ncbi.nlm.nih.gov

Published by Oxford University Press.

genomes, and ~100 expression and epigenomics studies. Over 90 published studies have been linked to SRA deposits. Most of the human genomes were produced by the 1000 Genomes Project, which is using sequencing data to perform a deep analysis of ordinary human variation in three healthy populations with the expectation of detecting common human genetic variants (defined as frequency 1% or higher) (www.1000genomes.org). The Project is submitting reads to the SRAs in real time as they are produced, allowing investigators, not associated with this project, direct access to its output.

The value of the SRAs to the scientific community will depend on the degree to which data from investigations are deposited. Accordingly, NCBI, EBI and DDBJ encourage researchers to consider depositing their data in one of the SRAs. We have tried to ease the burden of sequence submission in several ways: first time and occasional submitters can use an interactive interface and upload smaller data sets through a web browser; high-throughput users can submit data via an automated submission pipeline that uses XML to describe metadata and the community-developed Sequence Read Format (SRF) as a common container file format; and all three SRAs use a high-speed file transfer protocol called fasp (Aspera, Inc., Emeryville, CA, USA) that allows users to transfer files at speeds up to 400 Mbps, many times faster than ftp. For information about submitting to SRA, see http://www.ncbi.nlm.nih.gov/Traces/sra/static/SRA_Submission_Guidelines.pdf at NCBI, http://www.ebi.ac.uk/embl/Documentation/ENA-Reads.html at EBI and http://trace.ddbj.nig.ac.jp/dra/submission_e.shtml at DDBJ. Functional genomics studies utilizing short reads (e.g. ChIP-Seq and mRNA-Seq) can be submitted via the Gene Expression Omnibus and ArrayExpress resources; see instructions at http://www.ncbi.nlm.nih.gov/geo/info/seq.html and http://www.ebi.ac.uk/microarray/submissions_overview.html, respectively. Finally, NCBI and EBI are working on developing SRA instances specially designed for the archiving of human sequencing data sets under privacy control, usage restrictions or ethical constraints.

FUNDING

The EBI's next generation sequence archiving activities are supported by the Wellcome Trust, the European Commission and the European Molecular Biology Laboratory. DDBJ's work on SRA and Trace Archive is supported by the Ministry of Education, Culture, Sports, Science and Technology of Japan. Funding for open access charge: NCBI's SRA work was supported by the Intramural Research Program of the NIH, National Library of Medicine.

Conflict of interest statement. None declared.

REFERENCES

1. Wheeler,D.L., Barrett,T., Benson,D.A., Bryant,S.H., Canese,K., Chetvernin,V., Church,D.M., DiCuccio,M., Edgar,R., Federhen,S. *et al.* (2007) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **35**, D5–D12.
2. Wheeler,D.L., Barrett,T., Benson,D.A., Bryant,S.H., Canese,K., Chetvernin,V., Church,D.M., Dicuccio,M., Edgar,R., Federhen,S. *et al.* (2008) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **36**, D13–D21.
3. Cochrane,G., Akhtar,R., Bonfield,J., Bower,L., Demiralp,F., Faruque,N., Gibson,R., Hoad,G., Hubbard,T., Hunter,C. *et al.* (2009) Petabyte-scale innovations at the European Nucleotide Archive. *Nucleic Acids Res.*, **37**, D19–D25.
4. Sugawara,H., Ikeo,K., Fukuchi,S., Gojobori,T. and Tateno,Y. (2009) DDBJ dealing with mass data produced by the second generation sequencer. *Nucleic Acids Res.*, **37**, D16–D18.
5. Sayers,E.W., Barrett,T., Benson,D.A., Bryant,S.H., Canese,K., Chetvernin,V., Church,D.M., DiCuccio,M., Edgar,R., Federhen,S. *et al.* (2009) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **37**, D5–D15.