Published online 9 November 2010 *Nucleic Acids Research, 2011, Vol. 39, Database issue* ***D19–D21***
doi:10.1093/nar/gkq1019

The Sequence Read Archive

Rasko Leinonen[1,*], Hideaki Sugawara[2] and Martin Shumway[3] on behalf of the International Nucleotide Sequence Database Collaboration

[1]European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK, [2]Center for Information Biology and DNA Data Bank of Japan, National Institute of Genetics, Research Organization of Information and Systems, Yata, Mishima 411-8540, Japan and [3]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland 20894, USA

Received September 16, 2010; Accepted October 8, 2010

ABSTRACT

The combination of significantly lower cost and increased speed of sequencing has resulted in an explosive growth of data submitted into the primary next-generation sequence data archive, the Sequence Read Archive (SRA). The preservation of experimental data is an important part of the scientific record, and increasing numbers of journals and funding agencies require that next-generation sequence data are deposited into the SRA. The SRA was established as a public repository for the next-generation sequence data and is operated by the International Nucleotide Sequence Database Collaboration (INSDC). INSDC partners include the National Center for Biotechnology Information (NCBI), the European Bioinformatics Institute (EBI) and the DNA Data Bank of Japan (DDBJ). The SRA is accessible at http://www.ncbi.nlm.nih.gov/Traces/sra from NCBI, at http://www.ebi.ac.uk/ena from EBI and at http://trace.ddbj.nig.ac.jp from DDBJ. In this article, we present the content and structure of the SRA, detail our support for sequencing platforms and provide recommended data submission levels and formats. We also briefly outline our response to the challenge of data growth.

THE SEQUENCE READ ARCHIVE

The Sequence Read Archive (SRA) is an international public archival resource for next-generation sequence data established under the guidance of the International Nucleotide Sequence Database Collaboration (INSDC) (1). Instances of the SRA are operated by the National Center for Biotechnology Information (NCBI) (2), the European Bioinformatics Institute (EBI) (3) and the DNA Data Bank of Japan (DDBJ) (4). The mission of INSDC is to preserve public-domain sequencing data and to provide free, unrestricted and permanent access to the data. For INSDC policy details please refer to: http://www.insdc.org/policy.html. Authorized access data submissions, such as human samples sequenced under ethical consent agreements, should be submitted to dbGAP (http://www.ncbi.nlm.nih.gov/gap) at NCBI or to the European Genome-phenome archive (EGA) (http://www.ebi.ac.uk/ega) at EBI. Data submitted to dbGAP or EGA is not part of the public SRA. However, high-level metadata is made available through SRA. For a brief history of the SRA please refer to (5).

CONTENT

In mid-September 2010, the SRA contained >500 billion reads consisting of 60 trillion base pairs available for download including authorized access data submitted to dbGAP. Almost 80% of the sequencing data are derived from the Illumina GA platform. The SOLiD™ and Roche/454 platforms account for 15% and 5% of submitted base pairs, respectively. In terms of submitted base pairs, the most active SRA submitters include the Broad Institute, Washington University in St Louis, the Wellcome Trust Sanger Institute and Baylor College of Medicine with 34, 15, 13 and 12% share of sequenced bases, respectively. The largest individual global project generating next-generation sequence is the 1000 Genomes project (http://www.1000genomes.org) which has generated nearly half of all data submitted into the SRA. The most sequenced organisms are *Homo sapiens* with 65% and *Mus musculus* with 4% share of all bases. Human metagenome sequencing accounted for 16% of submitted bases.

PLATFORM SUPPORT

At present, support is offered for widely used sequencing platforms: Roche/454 (Roche Diagnostics Corp.), Illumina

*To whom correspondence should be addressed. Tel: +44 1223 494608; Fax: +44 1223 494408; Email: rasko@ebi.ac.uk

Genome Analyzer (Illumina Inc.) and SOLiD™ (Life Technologies Corp.). Support for HeliScope™ Single Molecule Sequencer (Helicos Biosciences Corp.), Complete Genomics Inc., SMRT™ (Pacific Biosciences Inc.) and Ion Torrent Systems Inc. will be available shortly.

RECOMMENDED DATA SUBMISSION LEVELS AND FORMATS

The SRA is intended as a repository of data from the primary analysis phase of sequencing. Experience of operation of the SRA over the last 3 years has allowed us to refine the levels to which we archive data. Storing early raw forms of data, such as images and signals, provides users greatest theoretical precision, but at the expense of significant costs. Data submitted to the SRA archives must always include base or SOLiD™ color calls and qualities. This is now also the recommended data submission level for Illumina Genome Analyzer (GA) and SOLiD™ platforms. Signal data for the Illumina GA and SOLiD™ platforms should no longer be submitted into the SRA archives, as the cost of signal data storage for these platforms is considered to be significantly higher than the value of making these data available for any further analysis. For the 454 platform, the submitted data should still include the signal information.

The recommended submission format for data from the Illumina GA and SOLiD™ platforms is Sequence Read Format (SRF); SRF files for the Illumina GA platform should be prepared using the DNA Sequence Read Toolkit (http://sourceforge.net/projects/sequenceread/files), and for the SOLiD™ using the SOLiD™ SRF conversion utility (http://solidsoftwaretools.com/gf/project/srf). For the 454 platform, the recommended submission format is Standard Flowgram Format (SFF). Both SRF and SFF files are highly compressed and should be submitted to SRA without applying any further compression.

METADATA STORAGE AND EXCHANGE

The SRA metadata model consists of six XML objects, each constrained by a schema. All metadata are exchanged on a daily basis between the SRA archives and can be accessed and retrieved from all three sites. The current versions (1.1) of the SRA XML Schemas are available at http://www.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?cmd=show&f=schema&m=software&s=schema from NCBI and at ftp://ftp.sra.ebi.ac.uk/meta/xsd/sra_1_1 from EBI. While the SRA stores and presents metadata using SRA XML documents, submissions may be prepared using a variety of tools and pipelines. The SRA XML objects are study, sample, experiment, run, analysis and submission. The SRA study object contains high-level project information including literature references, and may be linked to the INSDC projects database. Similarly, the SRA sample object contains detailed sample information. The SRA experiment and run objects contain instrument and library information and are directly associated with the sequence data. The SRA analysis object is used for the deposition of a variety of analysis results including reference alignments, multiple alignments and assemblies. The SRA submission object groups the other objects for submission into the SRA. Metadata XML objects are all accessioned with unique permanent identifiers that are used by all partners in the collaboration.

SEQUENCE DATA STORAGE AND EXCHANGE

The SRA follows the established INSDC data-exchange convention where public data are exchanged between the INSDC partners on a daily basis. This allows all public data to be accessed at each site regardless of the point of the original submission.

Before next-generation sequencing platforms existed, the most commonly used format for the representation of base calls and quality scores was the Sanger Fastq format (6). In 2001, a new community format was created which also supported the inclusion of the signal information: the ZTR format (7). SRF, a further development of ZTR, became the first widely used cross-platform format for storing next-generation sequence data. The SRF format gained a substantial user base from Illumina GA and SOLiD™ users, while the earlier SFF format (http://www.ncbi.nlm.nih.gov/Traces/trace.cgi?cmd=show&f=formats&m=doc&s=format#sff) became the standard for the 454 platform. In 2009, SAM and BAM (8) were introduced as generic formats for storing read alignments against reference sequences. Sequence alignments are increasingly generated as a primary analysis intermediate, and BAM is expected to replace SRF as the preferred submission format to the SRA; importantly, BAM supports not only aligned, but also unaligned reads which are also recommended to be submitted to SRA. The SRA archives are currently working together with community experts to define an archival BAM format with the goal of making submission and exchange of BAM files as easy as possible.

Efficient storage and compression of next-generation sequence data has always been one of the main objectives of the SRA. Internally, the SRA uses the NCBI SRA Toolkit (http://www.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?cmd=show&f=software&m=software&s=software) for storing and exchanging all next-generation sequence data. Critically, the toolkit contains a configurable storage and compression architecture allowing current best practices to co-exist with future ones. The NCBI SRA Toolkit has established itself as an important part of the SRA operations at NCBI, EBI and DDBJ, who now routinely validate and convert submitted data into the SRA Toolkit format. This format is used for data exchange by the SRA partners, converted to other formats such as Fastq, and made available to other applications through its standard API. For example, the NCBI BLAST has been extended to do sequence similarity searches using the files generated by the NCBI SRA Toolkit.

CHALLENGE OF DATA GROWTH

With the growth of the next-generation sequence data surpassing the growth of disk-storage capacity, the value of

storing different types of data is being evaluated. Experience from the last 3 years has allowed SRA to define the recommended platform specific data submission levels. The cost of archiving Illumina GA and SOLiD™ signal data are now considered to significantly exceed the value of making this data available for any subsequent analysis. At NCBI, this signal data is now stored on a less accessible secondary-storage system and is no longer guaranteed to be permanently available as part of the SRA archives. A complementary approach to limiting the cost of the archival storage is to implement more efficient compression strategies. Different types of data vary in their compressibility characteristics and some types of data can be compressed significantly more efficiently than others. Currently, one of the most promising compression strategies for next-generation sequences involves reference-based compression (9). The SRA is actively exploring better compression methods including approaches based on reference alignment of reads, and on the preservation of only the most valuable base quality information (Fritz,M.H. *et al.*, submitted for publication). The SRA strategy, then, is to balance data reduction and compression in light of infrastructure costs and usage patterns.

FUNDING

European Molecular Biology Laboratory, European Commission and the Wellcome Trust; Ministry of Education, Culture, Sports, Science and Technology of Japan (to D.D.B.J.'s work on SRA and Trace Archive); Intramural Research Program of the NIH, National Library of Medicine (to NCBI's SRA work). Funding for open access charge: European Molecular Biology Laboratory.

Conflict of interest statement. None declared.

REFERENCES

1. Cochrane,G. *et al.* (2011) The International Nucleotide Sequence Database Collaboration in 2010. *Nucleic Acids Res*, **39**, D15–D18.
2. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Sayers,E.W. (2010) GenBank. *Nucleic Acids Res.*, **38**, D46–D51.
3. Leinonen,R., Akhtar,R., Birney,E., Bonfield,J., Bower,L., Corbett,M., Cheng,Y., Demiralp,F., Faruque,N., Goodgame,N. *et al.* (2010) Improvements to services at the European Nucleotide Archive. *Nucleic Acids Res.*, **38**, D39–D45.
4. 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.
5. Shumway,M., Cochrane,C. and Sugawara,H. (2010) Archiving next generation sequencing data. *Nucleic Acids Res.*, **38**, D870–D871.
6. Cock,P.J.A., Fields,C.J., Goto,N., Heuer,M.L. and Rice,P.M. (2009) The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Res.*, **38**, 1767–1771.
7. Bonfield,J.K. and Staden,R. (2001) ZTR: a new format for DNA sequence trace data. *Bioinformatics*, **18**, 3–10.
8. Li,H., Handsaker,B., Wysoker,A., Fennell,T., Ruan,J., Homer,N., Marth,G., Abecasis,G., Durbin,R. and 1000 Genome Project Data Processing Subgroup. (2009) The Sequence alignment/map (SAM) format and SAMtools. *Bioinformatics*, **25**, 2078–2079.
9. Christley,S., Lu,Y., Li,C. and Xie,X. (2008) Human genomes as email attachments. *Bioinformatics*, **25**, 274–275.