The NCBI Short Read Archive (SRA), a new primary data archive resource

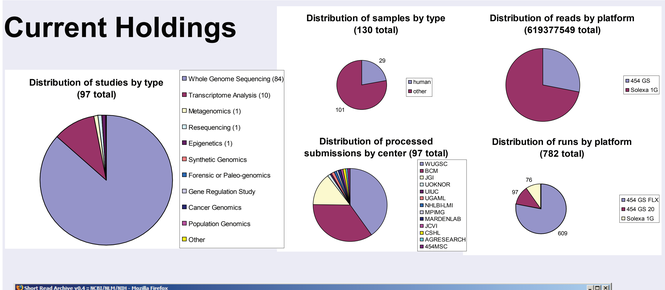
Martin Shumway, Eugene Yaschenko, Vladimir Alekseyev, Deanna Church, and James Ostell

National Center for Biotechnology Information (NCBI)
National Library of Medicine (NLM)
Bethesda MD USA

Abstract

The Short Read Archive (SRA) at the National Center for Biotechnology Information (NCBI) accepts deposits of sequencing data from the next generation of genome sequencing platforms. This new resource stores primary sequence, quality, and intensity data from experiments. A new data model separates notions of study, experiment, sample, and run data such that these elements can be flexibly reused. The SRA meets research needs by providing a public home for massive datasets, providing permanent accessions for sequencing projects, allowing investigators to query studies based on a rich set of index terms, and by providing users with links to downstream analyses and outside resources. A hold-until-publish feature allows a submitter to obtain an accession for their dataset but mask its publication until after the submitter releases it. Tag-value attribute pairs and linkout objects are used to capture much of the ancillary data that submitters might wish to provide along with their samples and experiments. These can be as rich or as lean as desired. A new method of encoding reads provides for flexible representation of creative sequencing chemistries.

Current Holdings



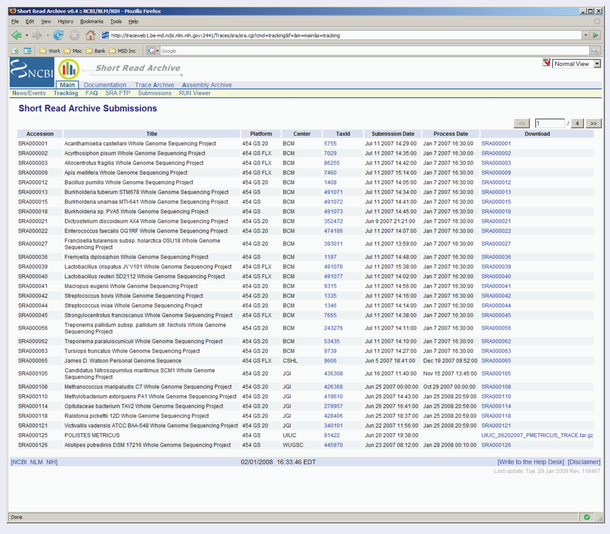


Figure 1: Tracking page shows deposits by study and makes data available in its original form.

Sizing

Table 1 – The SRA obtains an additional 2 – 4X compression on submission data that is already binary compressed.

Platform	Original	Submission Form	Pre-SRA	Post-SRA
Solexa	Text tab files	*tar, gzip*	100	23-25
454	sff	sff	100	43-46

Run Browser

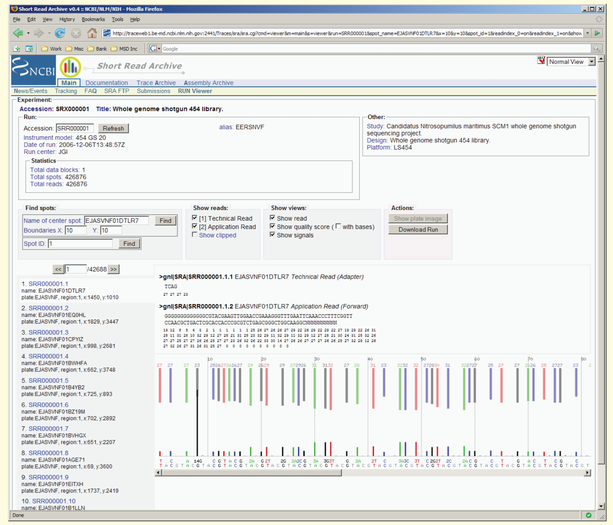


Figure 2: A spot is shown in the context of its flow sequence on the 454 platform. The sub-sequence used for applications (SRR000001.1.2) does not include the key sequence (TCAG) or the clipped portion at the 3' end.

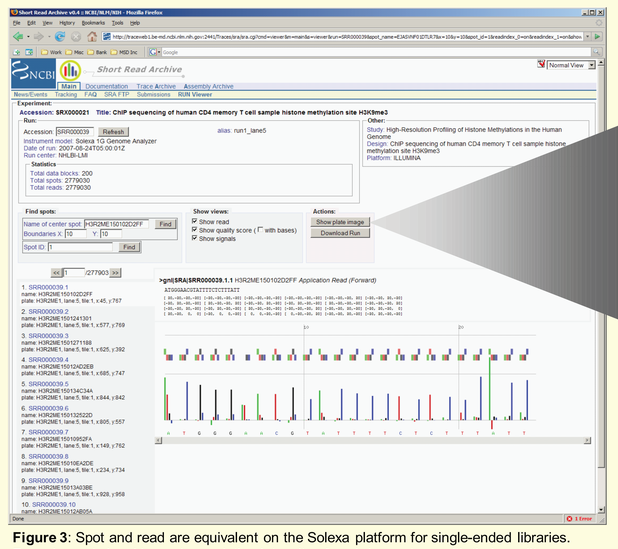


Figure 3: Spot and read are equivalent on the Solexa platform for single-ended libraries. Four channel signal bars show relative intensity of each base along with their quality scores displayed overhead.

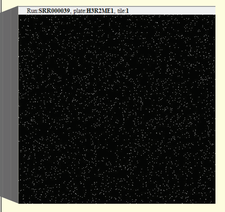


Figure 4: Portion of a plate image plot showing location of spots as computed by the primary analysis stage of the instrument's data processing.

Run Browser Basic Features

- Random access to reads by vendor assigned name
- Random access to reads by accession
- Random access to reads by address
- Partitions spot into technical and application reads
- Relates run to parent experiment, study, sample
- View ancillary information about run (run date, etc)
- Access and view neighbors
- Download run data
- Filter run data to reduced download set
- View aggregate run statistics and plots
- View complete intensity graph of a read

The Run Browser is intended to serve as a platform for accessing individual reads, analyzing their production data, and providing a facility for sub-selecting read sets for download. The design attempts to present short read data in a platform independent way.

Future developments include BLAST-like search capabilities. Many of these features will be driven by the user community.

Spot Abstraction

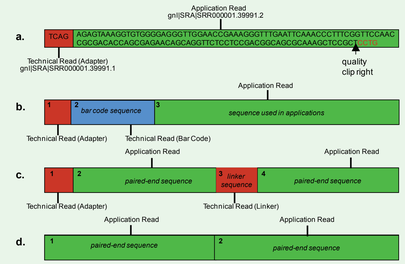


Figure 5: Short reads are extracted from an image cluster, or "spot", and partitioned during processing to present subsequences to downstream applications. The SRA presents the partition among "application" and "technical" reads of the spot's sequence. Annotations such as 3' quality clipping are stored as properties of the entire spot (**a**). Alternative spot layouts should be supported this way: barcoded reads (**b**), paired-end sequence with linker (**c**), and paired-end sequence without linker (**d**).

Data Model

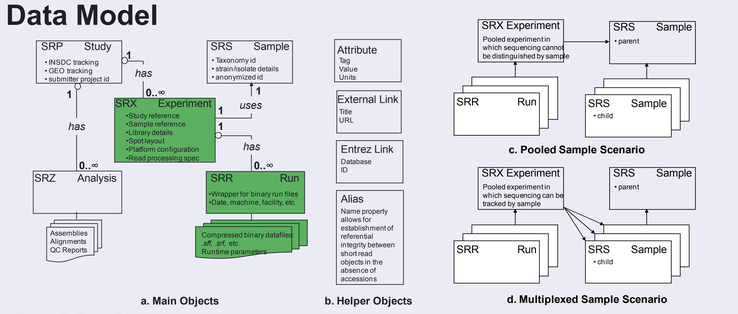


Figure 6: The SRA separates metadata from data (a), relates experiments to samples in flexible ways (c,d), and provides flexible data structures for decorating objects with properties, internal and external links, and names (b).

Entrez Model

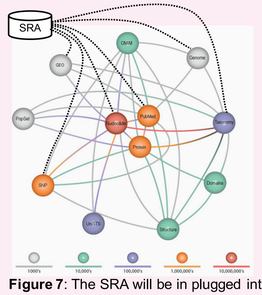


Figure 7: The SRA will be in plugged into the Entrez system of interlinked databases.

Submitting to SRA

Submissions Model

- High Throughput Submission with XML
- Interactive Submissions Tool
- Consultative Submissions
- Updates and modifications to metadata
- Sequencing Centers vs Individual Submitters

A distinct accession (SRA) is issued just for the submission session, separating submissions contact info, file manifests, exceptions from the data and metadata. Metadata can be submitted separately from data, and elements of a submission can arrive at the SRA at different points in time.

Hold Until Publish Feature

- Hold for [days]
- Hold until [date]
- Hold [until released]
- Hold for broker [broker] = poll broker authority

Restricted Access Feature

- Metadata are public
- Run data are private
- Appropriate for patient data, personal genomics
- Brokered by dbGaP, which manages access

Data Mirroring

Submissions will be mirrored to EBI and other INSDC partners on a frequent basis. Mirroring will take into account holds and restrictions. Short read accessions issued by EBI will be valid in the SRA.

Acknowledgements

SRA Technical Team

Misha Kimmelman
Andriy Klymenko
Sergey Ponomarev
Kurt Rodarmer
Dmitry Volodin
Aleksey Vyskokolov
Zinaida Belaia
Bob Sanders

Collaborators

European Bioinformatics Institute
Sanger Centre
James Bonfield
Asim Siddiqui

Vendors

454 Life Sciences/Roche Applied Science
Illumina, Inc.
Applied Biosystems
Helicos Bioscience Corp.

Contributors

Baylor College of Medicine
Joint Genome Institute
Washington University Genome Sequencing Center
National Heart, Lung, and Blood Institute
and others...

This research was supported by the Intramural Research Program of the NIH, National Library of Medicine.