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

Martin Shumway, Eugene Yaschenko, Vladimir Aleksyey, 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 identity 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 mature 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 hash-unique 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 textual objects are used to capture much of the ancillary data that submitters might wish to provide along with their sample and experiment data. These can be as rich or as lean as desired. A new method of encoding reads provides for flexible representation of creative sequencing chemistries.

Run Browser

Run Browser Basic Features
- Random access to reads by vendor assigned name
- Random access to reads by accession
- Random access to reads by address
- Reads sorted into application needs
- Flexible views to present experiment, study, sample
- View ancillary information about run (run date, etc)
- Access and view neighbors
- Download run data
- Filter run data to reduced downsampled set
- View aggregate run statistics and plots
- View complete intensity graph of a read

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 overlaid.

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.

Figure 5: A spot is shown in the context of its flow sequence on the 454 platform. The subsequence used for applications (SRR000001.12) does not include the key sequence (TCAG) or the clipped portion at the 3' end.

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).