



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

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

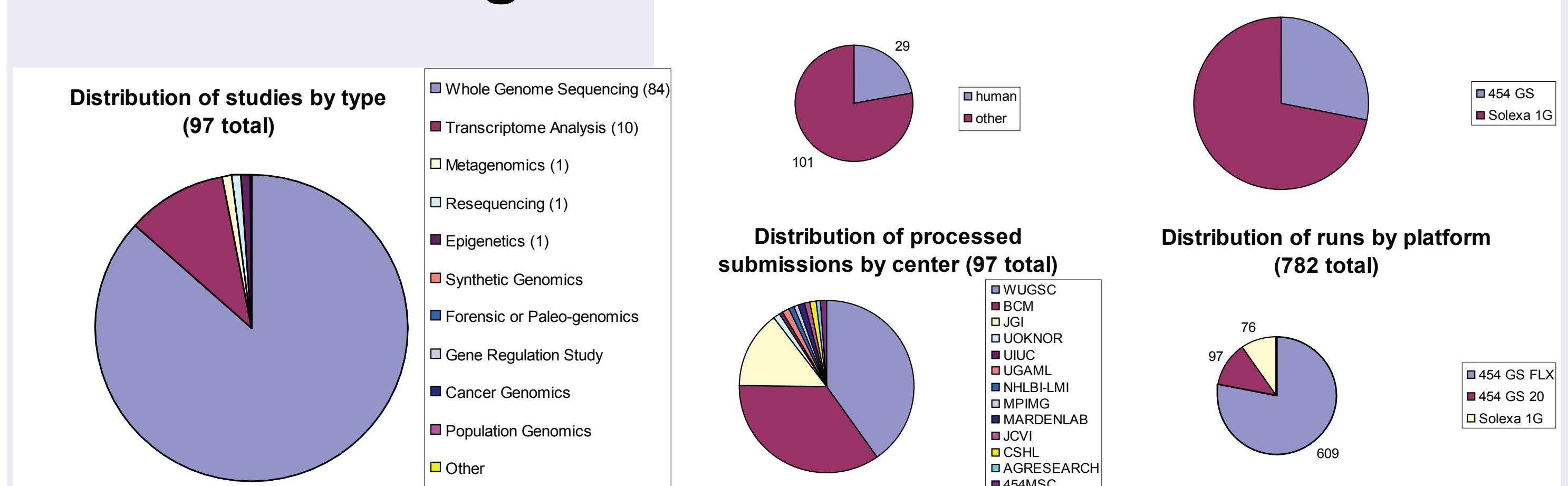
National Center for Biotechnology Information (NCBI)
National Library of Medicine (NLM)
Bethesda MD USA



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



Accession	Title	Center	Taxid	Submission Date	Process Date	Download
SRAD00001	Acetabularia castellanii Whole Genome Sequencing Project	BCM	5755	Jul 11 2007 14:29:00	Jan 7 2007 16:30:00	SRAD00001
SRAD00002	Acetabularia pinnatifida Whole Genome Sequencing Project	BCM	7029	Jul 11 2007 14:35:00	Jan 7 2007 16:30:00	SRAD00002
SRAD00003	Alcochlorus fragilis Whole Genome Sequencing Project	BCM	86255	Jul 11 2007 14:42:00	Jan 7 2007 16:30:00	SRAD00003
SRAD00009	Apis mellifera Whole Genome Sequencing Project	BCM	7460	Jul 11 2007 15:14:00	Jan 7 2007 16:30:00	SRAD00009
SRAD00012	Bacillus pumilus Whole Genome Sequencing Project	BCM	1409	Jul 11 2007 14:05:00	Jan 7 2007 16:30:00	SRAD00012
SRAD00013	Burkholderia tubum STM78 Whole Genome Sequencing Project	BCM	491071	Jul 11 2007 14:34:00	Jan 7 2007 16:30:00	SRAD00013
SRAD00015	Burkholderia unumae MTI-641 Whole Genome Sequencing Project	BCM	491072	Jul 11 2007 14:41:00	Jan 7 2007 16:30:00	SRAD00015
SRAD00018	Burkholderia sp. PVA5 Whole Genome Sequencing Project	BCM	491073	Jul 11 2007 14:45:00	Jan 7 2007 16:30:00	SRAD00018
SRAD00021	Drosophila obscura J44 Whole Genome Sequencing Project	BCM	352472	Jan 9 2007 21:21:00	Jan 7 2007 16:30:00	SRAD00021
SRAD00022	Enterococcus faecalis CIGRIF Whole Genome Sequencing Project	BCM	474186	Jul 11 2007 14:07:00	Jan 7 2007 16:30:00	SRAD00022
SRAD00027	Francisella tularensis subsp. holarctica OSU18 Whole Genome Sequencing Project	BCM	393011	Jul 11 2007 13:59:00	Jan 7 2007 16:30:00	SRAD00027
SRAD00036	Fraxinus diplophaga Whole Genome Sequencing Project	BCM	1197	Jul 11 2007 14:48:00	Jan 7 2007 16:30:00	SRAD00036
SRAD00039	Lactobacillus crispatus J1-V101 Whole Genome Sequencing Project	BCM	491076	Jul 11 2007 15:38:00	Jan 7 2007 16:30:00	SRAD00039
SRAD00040	Lactobacillus reuteri SD2112 Whole Genome Sequencing Project	BCM	491077	Jul 11 2007 14:02:00	Jan 7 2007 16:30:00	SRAD00040
SRAD00041	Macropus eugenii Whole Genome Sequencing Project	BCM	9315	Jul 11 2007 14:56:00	Jan 7 2007 16:30:00	SRAD00041
SRAD00042	Streptococcus bovis Whole Genome Sequencing Project	BCM	1335	Jul 11 2007 14:16:00	Jan 7 2007 16:30:00	SRAD00042
SRAD00044	Streptococcus iniae Whole Genome Sequencing Project	BCM	1346	Jul 11 2007 14:14:00	Jan 7 2007 16:30:00	SRAD00044
SRAD00045	Strongylocentrotus franciscanus Whole Genome Sequencing Project	BCM	7665	Jul 11 2007 14:38:00	Jan 7 2007 16:30:00	SRAD00045
SRAD00056	Trigonostylopsis subsp. pallidum str. Nichols Whole Genome Sequencing Project	BCM	243276	Jul 11 2007 14:11:00	Jan 7 2007 16:30:00	SRAD00056
SRAD00062	Tropaeolum parviflorum Whole Genome Sequencing Project	BCM	53435	Jul 11 2007 14:10:00	Jan 7 2007 16:30:00	SRAD00062
SRAD00063	Turrisia truncatula Whole Genome Sequencing Project	BCM	9739	Jul 11 2007 14:27:00	Jan 7 2007 16:30:00	SRAD00063
SRAD00065	James D. Watson Personal Genome Sequence	CHL	9606	Jan 5 2007 18:41:00	Dec 18 2007 05:52:00	SRAD00065
SRAD00105	Candidatus Nitrospumilus maritimus SCM1 Whole Genome Sequencing Project	JGI	436308	Jan 16 2007 11:40:00	Nov 15 2007 13:45:00	SRAD00105
SRAD00108	Methanococcus marisnigri C7 Whole Genome Sequencing Project	JGI	426368	Jan 25 2007 09:00:00	Oct 29 2007 09:00:00	SRAD00108
SRAD00110	Methylobacterium extorquens PA1 Whole Genome Sequencing Project	JGI	41910	Jan 26 2007 14:43:00	Jan 25 2008 20:59:00	SRAD00110
SRAD00114	Opilobacter bacterium TA02 Whole Genome Sequencing Project	JGI	278057	Jun 26 2007 18:41:00	Jan 25 2008 20:59:00	SRAD00114
SRAD00118	Ralstonia pectus 12D Whole Genome Sequencing Project	JGI	426406	Jun 25 2007 19:37:00	Jan 25 2008 20:59:00	SRAD00118
SRAD00121	Vibrio vulnificus ATCC 35064 Whole Genome Sequencing Project	JGI	340191	Jun 22 2007 11:56:00	Jan 25 2008 20:59:00	SRAD00121
SRAD00125	POLISTES METRICUS	ULUC	91422	Jun 20 2007 19:38:00	Nov 15 2007 13:45:00	SRAD00125
SRAD00126	Altipates pubertis DSM 17216 Whole Genome Sequencing Project	WUGSC	445970	Jun 23 2007 08:12:00	Jan 28 2008 08:10:00	SRAD00126

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	100	43-46

Run Browser

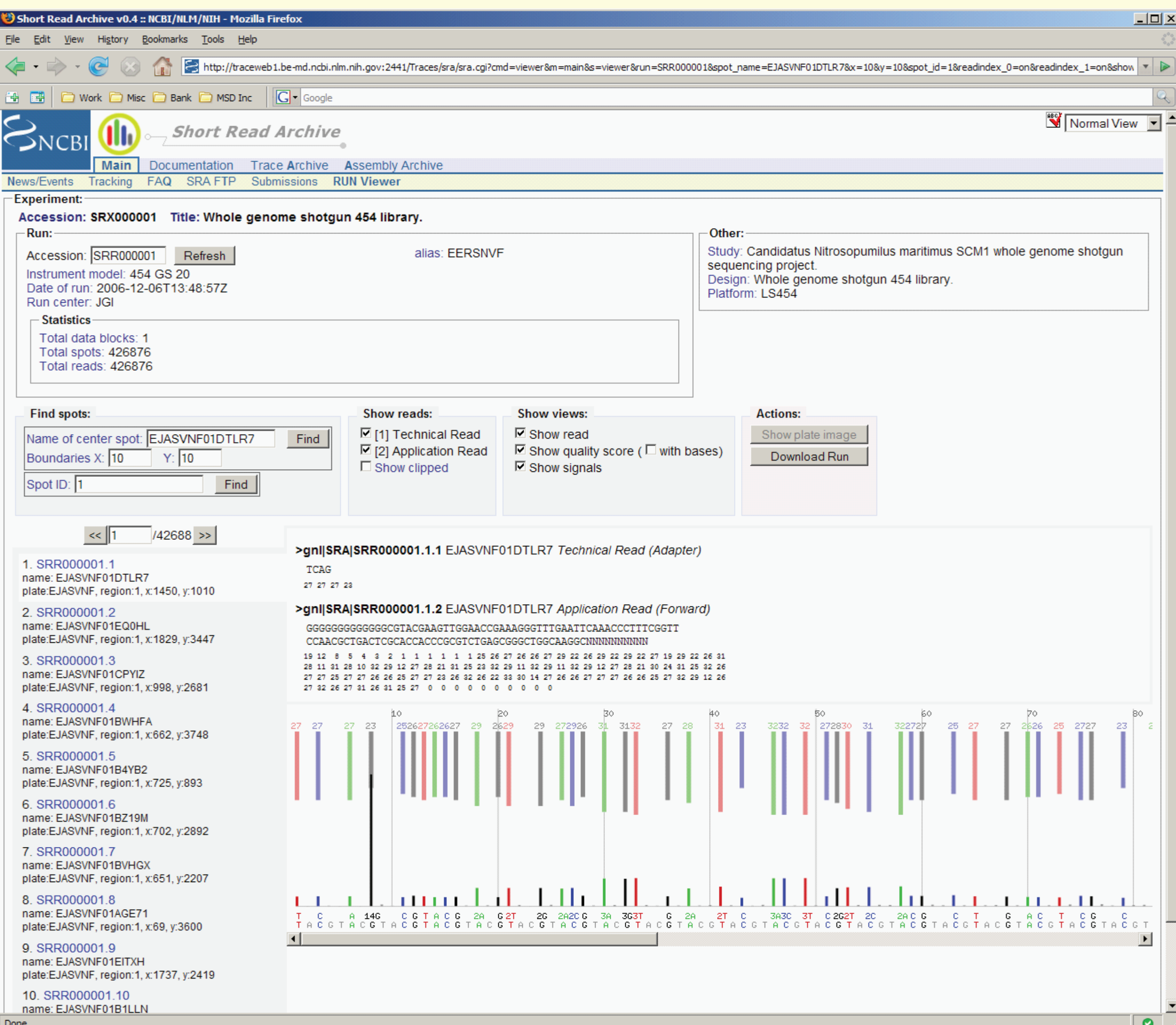


Figure 2: A spot is shown in the context of its flow sequence on the 454 platform. The sub-sequence used for applications (SRR000001.1.2) does not include the key sequence (TCAG) or the clipped portion at the 3' end.

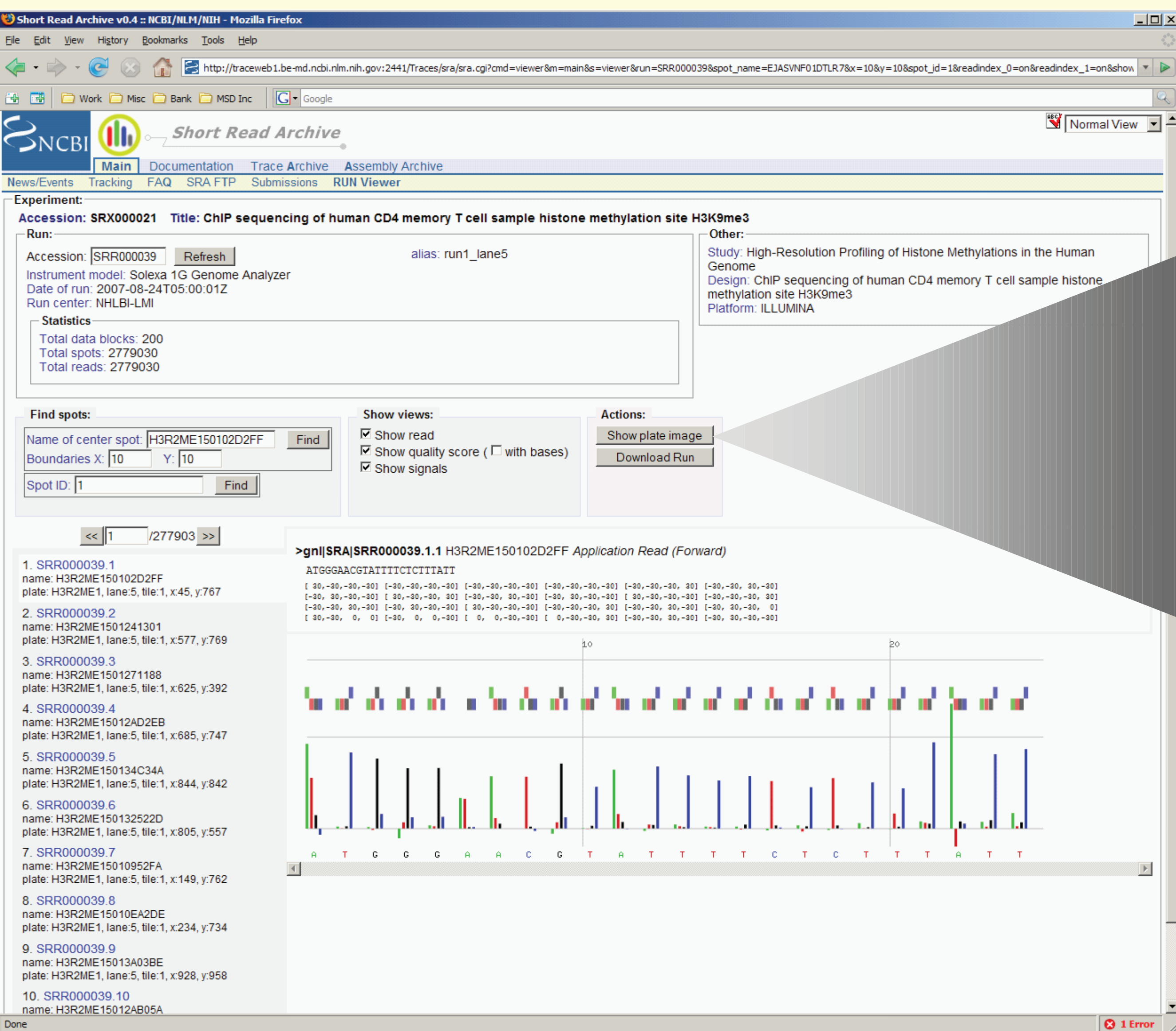


Figure 3: Spot and read are equivalent on the Solexa platform for single-ended libraries. Four channel signal bars show relative intensity of each base along with their quality scores displayed overhead.

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.

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.

Spot Abstraction

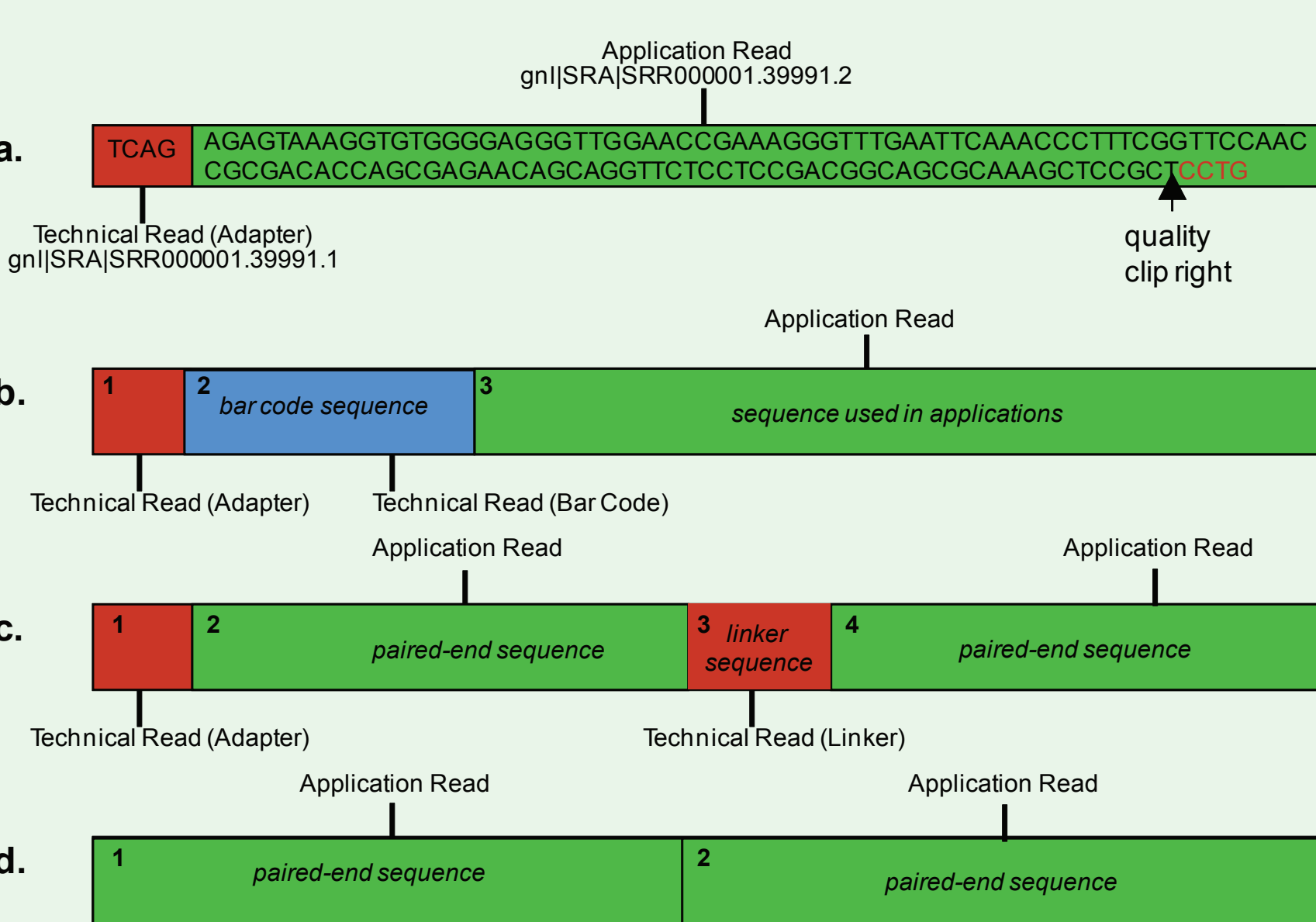


Figure 5: Short reads are extracted from an image cluster, or "spot", and partitioned during processing to present subsequences to downstream applications. The SRA presents the partition among "application" and "technical" reads of the spot's sequence. Annotations such as 3' quality clipping are stored 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

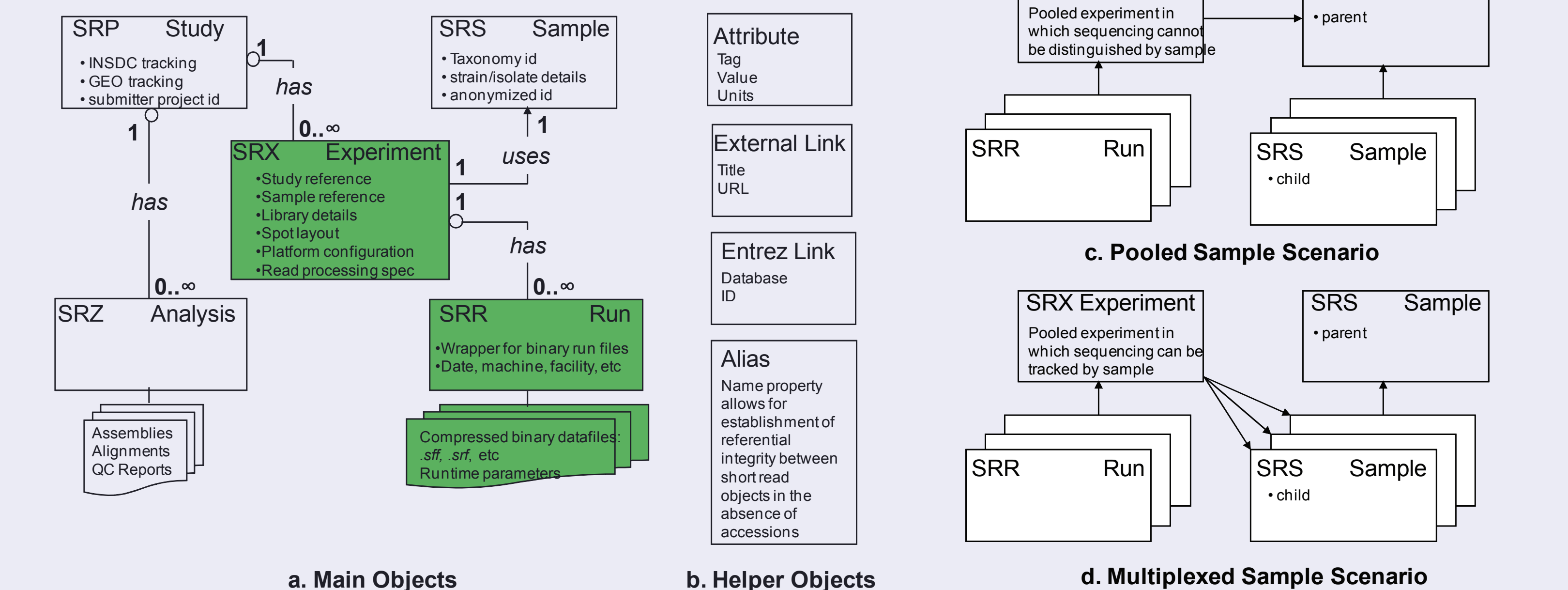


Figure 6: The SRA separates metadata from data (a), relates experiments to samples in flexible ways (c,d), and provides flexible data structures for decorating objects with properties, internal and external links, and names (b).

Entrez Model

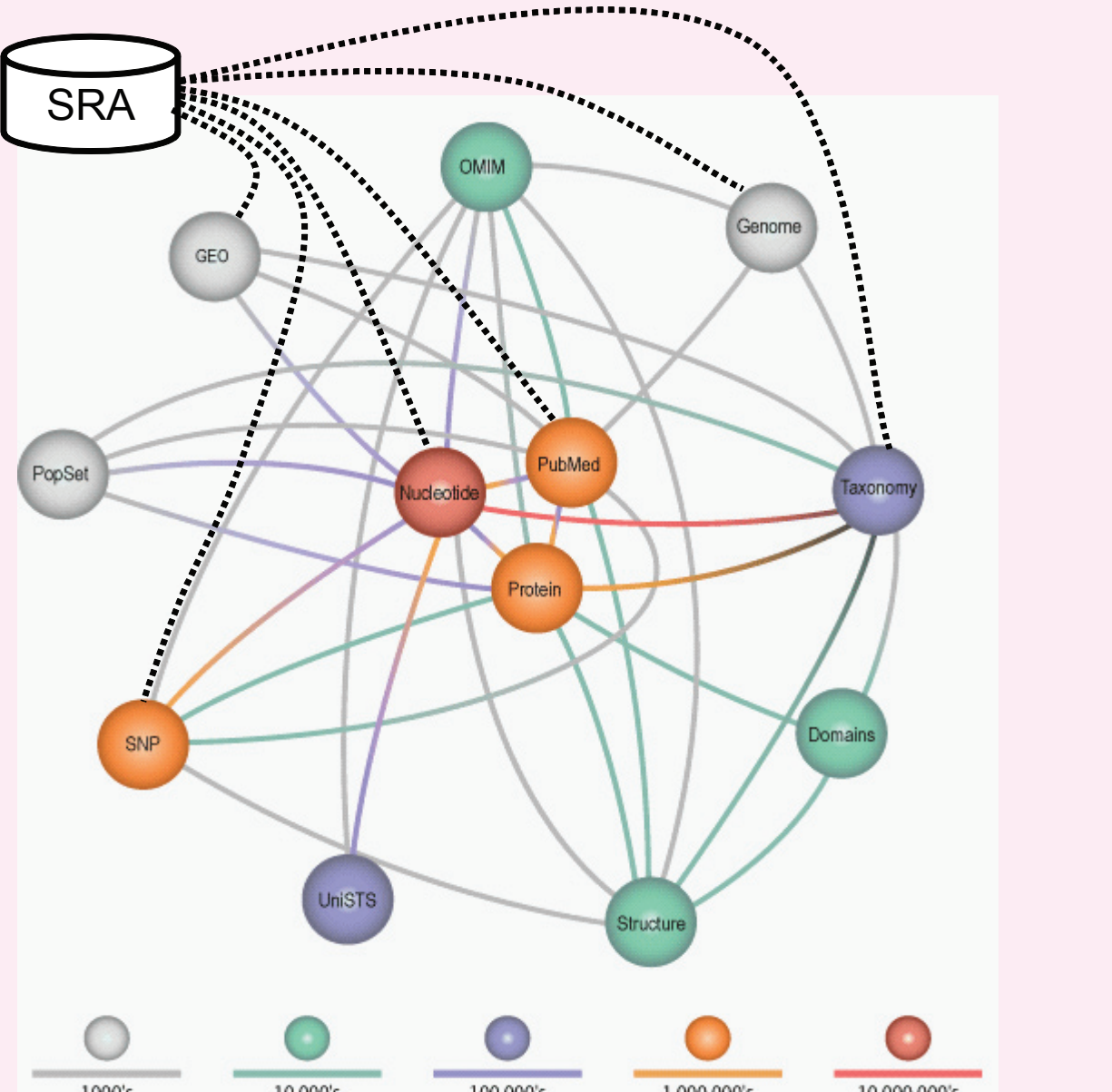


Figure 7: The SRA will be plugged into the Entrez system of interlinked databases.

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 Kimmelmann
Andry Klymenko
Sergey Ponomarev
Kurt Rodamer
Dmitry Volodin
Aleksey Vysokolov
Zinaida Belia
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.