Using the NCBI Short Read Archive (SRA) as a Scientific Resource

Martin Shumway, Eugene Yaschenko, Kurt Rodarmer, Michael Kimelman, and James Ostell

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

Current Holdings

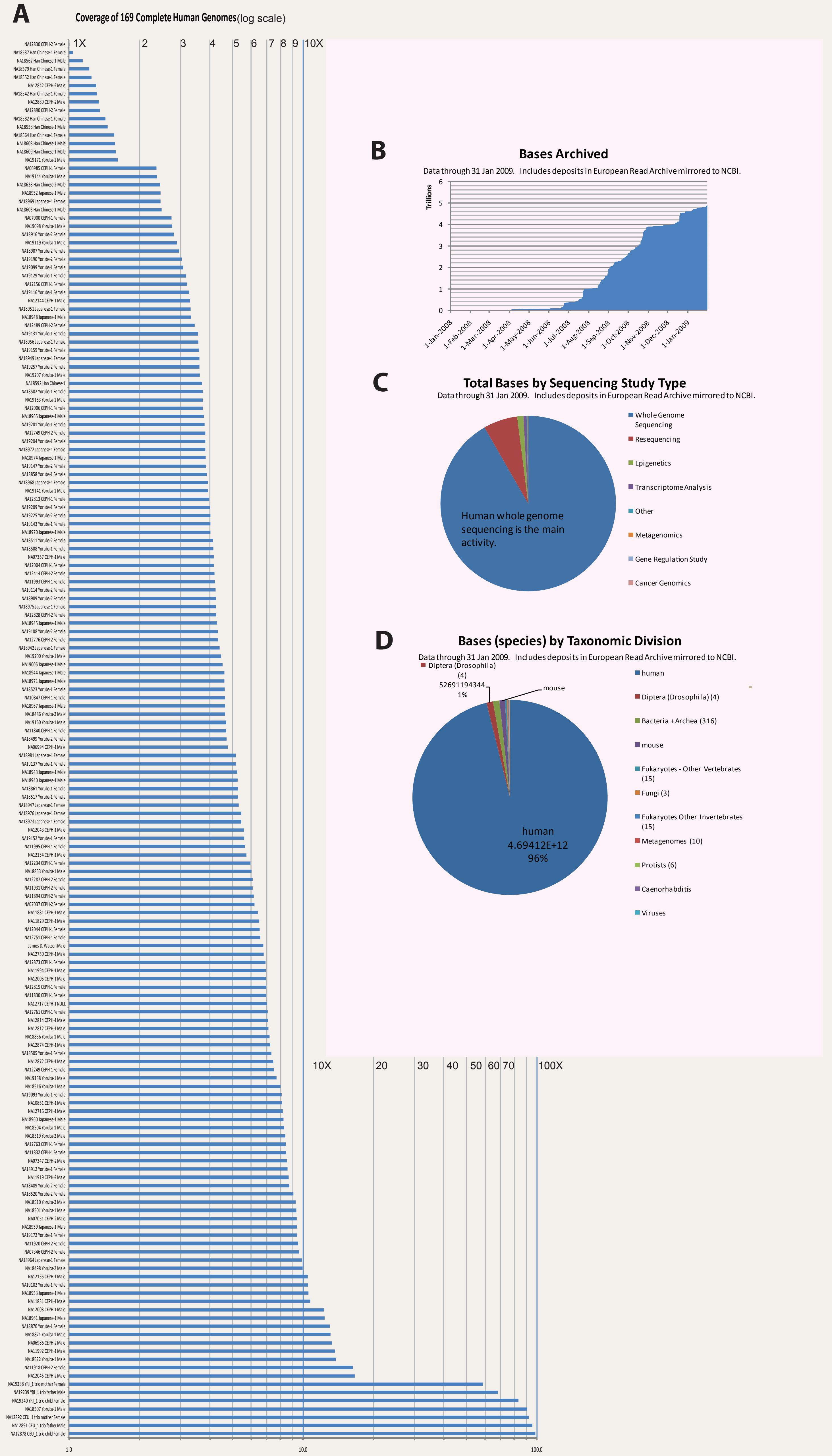


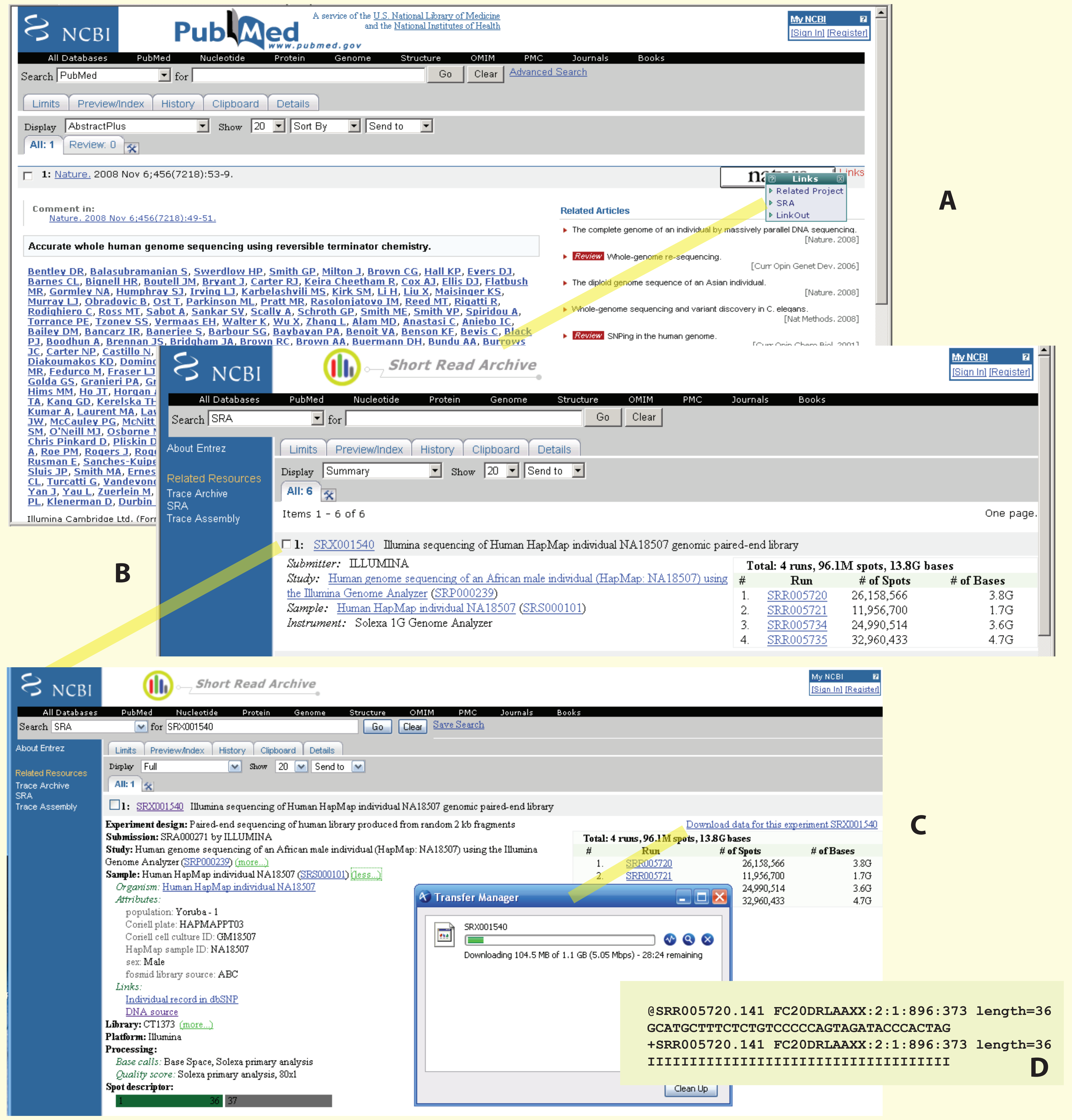
Figure 1: Deposits in the SRA as of 31 Jan 2009. (A) lists 169 human genomes sequenced to 1X or greater. Most of these were generated by the 1000 Genomes Project. (B) shows the growth in tera bases of deposited sequence. (C, D) show the breakdown of sequencing by study type and taxonomy.

Abstract

NCBI announces production release of the Short Read Archive (SRA), a repository of sequencing data from next generation sequencing platforms. Thus far the SRA has amassed over 4 TBp (Tera bases) of sequence from several thousand submissions executed via automated and interactive channels. The deposits are now indexed in the Entrez system, meaning that short read datasets can be retrieved via queries to PubMed, Genome Project, Taxonomy, and GEO. Short read data are provisioned in fastq format (bases and qualities), suitable for use with popular assemblers and aligners. The SRA regularly exchanges data with the European Read Archive (ERA) and the DNA Data Bank of Japan, providing users with worldwide scope for short read deposit and retrieval.

The SRA continues as a work-in-progress. Future developments include on-line sequence search capability, archival of short read alignment and assembly records, improved indexing and retrieval, a software library for interacting with archival data sets, and support for new platforms and sequencing technologies.

Retrieval Workflow



A service of the U.S. National Library of Medicine and the National Institutes of Health

Search PubMed for

Display AbstractPlus Show 20 Sort By Send to

All: 1 Review: 0

1: Nature. 2008 Nov 6;456(7218):53-9.

Comment in:
Nature. 2008 Nov 6;456(7218):49-51.

Accurate whole human genome sequencing using reversible terminator chemistry.

Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR, Boutell JM, Bryant J, Carter RJ, Keira Cheetham R, Cox AJ, Ellis DJ, Flatbush MR, Gormley NA, Humphray SJ, Irving LJ, Karbelashvili MS, Kirk SM, Li H, Liu X, Maisinger KS, Murray LJ, Obradovic B, Ost T, Parkinson ML, Pratt MR, Rasolonjatovo IM, Reed MT, Rigatti R, Rodighiero C, Ross MT, Sabot A, Sankar SV, Scally A, Schroth GP, Smith ME, Smith VP, Spiridou A, Torrance PE, Tzonev SS, Vermaas EH, Walter K, Wu X, Zhang L, Alam MD, Anastasi C, Aniebo IC, Bailey DM, Bancarz IR, Banerjee S, Barbour SG, Baybayan PA, Benoit VA, Benson KF, Bevis C, Black PJ, Boodhun A, Brennan JS, Bridgham JA, Brown RC, Brown AA, Buermann DH, Bundu AA, Burrows JC, Carter NP, Castillo N, Diakoumakos KD, Doming MR, Fedurco M, Fraser LJ, Golda GS, Granieri PA, Gr Hims MM, Ho JT, Horgan TA, Kang GD, Kerelska TH Kumar A, Laurent MA, La JW, McCauley PG, McNitt SM, O'Neill MJ, Osborne Chris Pinkard D, Pliskin D A, Roe PM, Rogers J, Roq Rusman E, Sanches-Kuipe Sluis JP, Smith MA, Ernes CL, Turcatti G, Vandevon Yan J, Yau L, Zuerlein M, PL, Klenerman D, Durbin

Illumina Cambridge Ltd. (For

Related Articles

- The complete genome of an individual by massively parallel DNA sequencing. [Nature. 2008]
- Review Whole-genome re-sequencing. [Curr Opin Genet Dev. 2006]
- The diploid genome sequence of an Asian individual. [Nature. 2008]
- Whole-genome sequencing and variant discovery in C. elegans. [Nat Methods. 2008]
- Review SNPing in the human genome.

Search SRA for

Display Summary Show 20 Send to

All: 6

Items 1 - 6 of 6 One page.

□ 1: SRX001540 Illumina sequencing of Human HapMap individual NA18507 genomic paired-end library

Submitter: ILLUMINA
Study: Human genome sequencing of an African male individual (HapMap: NA18507) using the Illumina Genome Analyzer (SRP000239)
Sample: Human HapMap individual NA18507 (SRS000101)
Instrument: Solexa 1G Genome Analyzer

Total: 4 runs, 96.1M spots, 13.8G bases

#	Run	# of Spots	# of Bases
1.	SRR005720	26,158,566	3.8G
2.	SRR005721	11,956,700	1.7G
3.	SRR005734	24,990,514	3.6G
4.	SRR005735	32,960,433	4.7G

Search SRA for SRX001540

Display Full Show 20 Send to

All: 1

□ 1: SRX001540 Illumina sequencing of Human HapMap individual NA18507 genomic paired-end library

Download data for this experiment SRX001540

Experiment design: Paired-end sequencing of human library produced from random 2 kb fragments
Submission: SRA000271 by ILLUMINA
Study: Human genome sequencing of an African male individual (HapMap: NA18507) using the Illumina Genome Analyzer (SRP000239) (more...)
Sample: Human HapMap individual NA18507 (SRS000101) (less...)
Organism: Human HapMap individual NA18507
Attributes:
population: Yoruba - 1
Coriell plate: HAPMAPPT03
Coriell cell culture ID: GM18507
HapMap sample ID: NA18507
sex: Male
fosmid library source: ABC
Links:
Individual record in dbSNP
DNA source
Library: CT1373 (more...)
Platform: Illumina
Processing:
Base calls: Base Space, Solexa primary analysis
Quality score: Solexa primary analysis, 80x1
Spot descriptor:
1 36 37

Total: 4 runs, 96.1M spots, 13.8G bases

#	Run	# of Spots	# of Bases
1.	SRR005720	26,158,566	3.8G
2.	SRR005721	11,956,700	1.7G
		24,990,514	3.6G
		32,960,433	4.7G

SRX001540
Downloading 104.5 MB of 1.1 GB (5.05 Mbps) - 28:24 remaining

```
@SRR005720.141 FC20DRLAAXX:2:1:896:373 length=36
GCATGCTTTCTCTGTCCCCCAGTAGATACCCACTAG
+SRR005720.141 FC20DRLAAXX:2:1:896:373 length=36
IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
```

Figure 2: A user finds an article of interest in PubMed. The user follows the SRA link (A), obtains a list of experiments run for the study (B), expands the detail of one such experiment (C), and downloads the data in fastq format (D).

Spot Abstraction

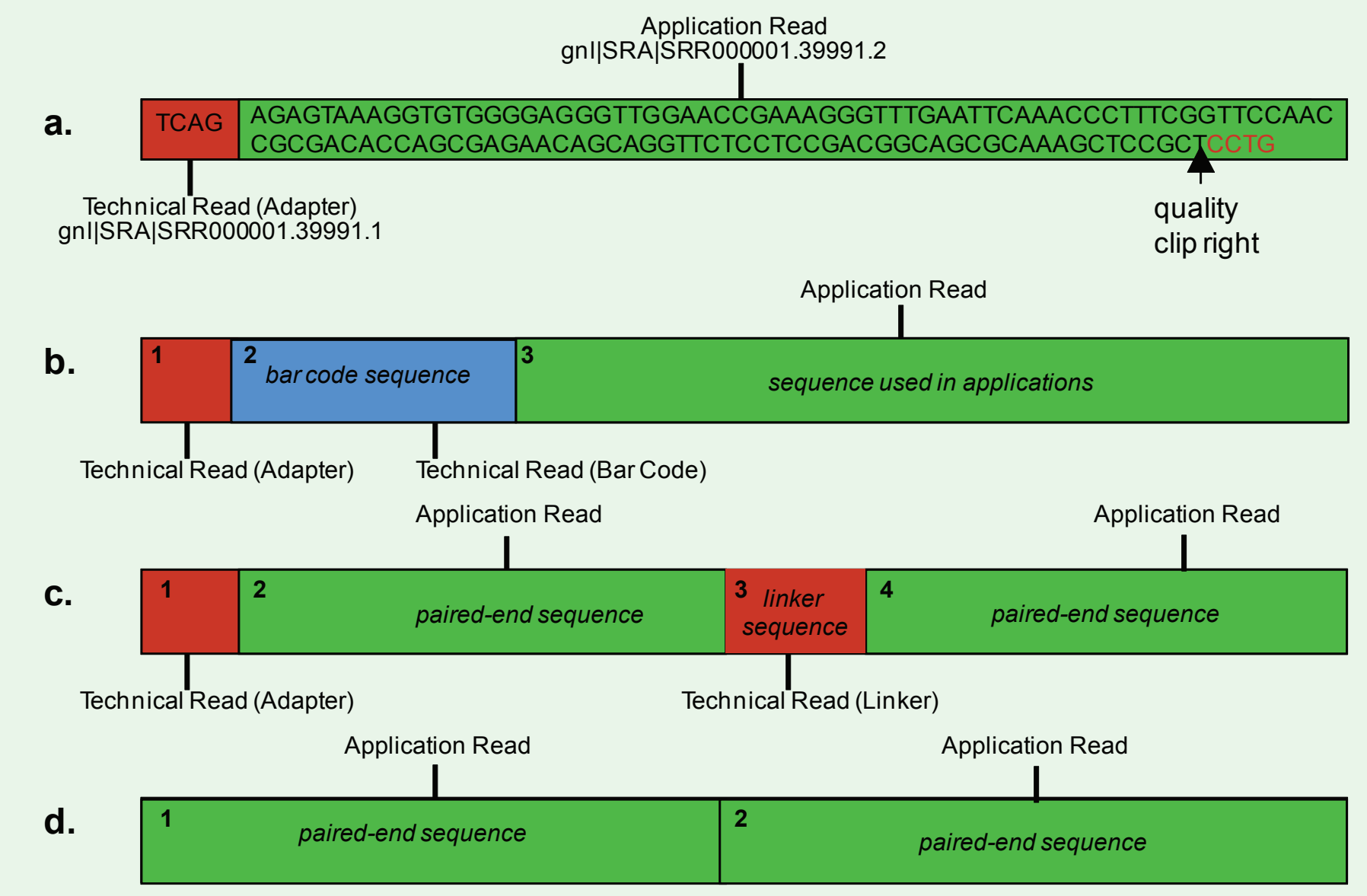


Figure 3: Short reads are extracted from an image cluster, or "spot", and partitioned during processing to present subsequences to downstream applications. The SRA presents the partition among "application" and "technical" reads of the spot's sequence. Annotations such as 3' quality clipping are stored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

Submitting to SRA

Submissions Model

- High Throughput Submission with XML
- Interactive Submissions Tool
- Consultative Submissions
- Submitter maintenance of 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 data and metadata for a given submission can arrive asynchronously.

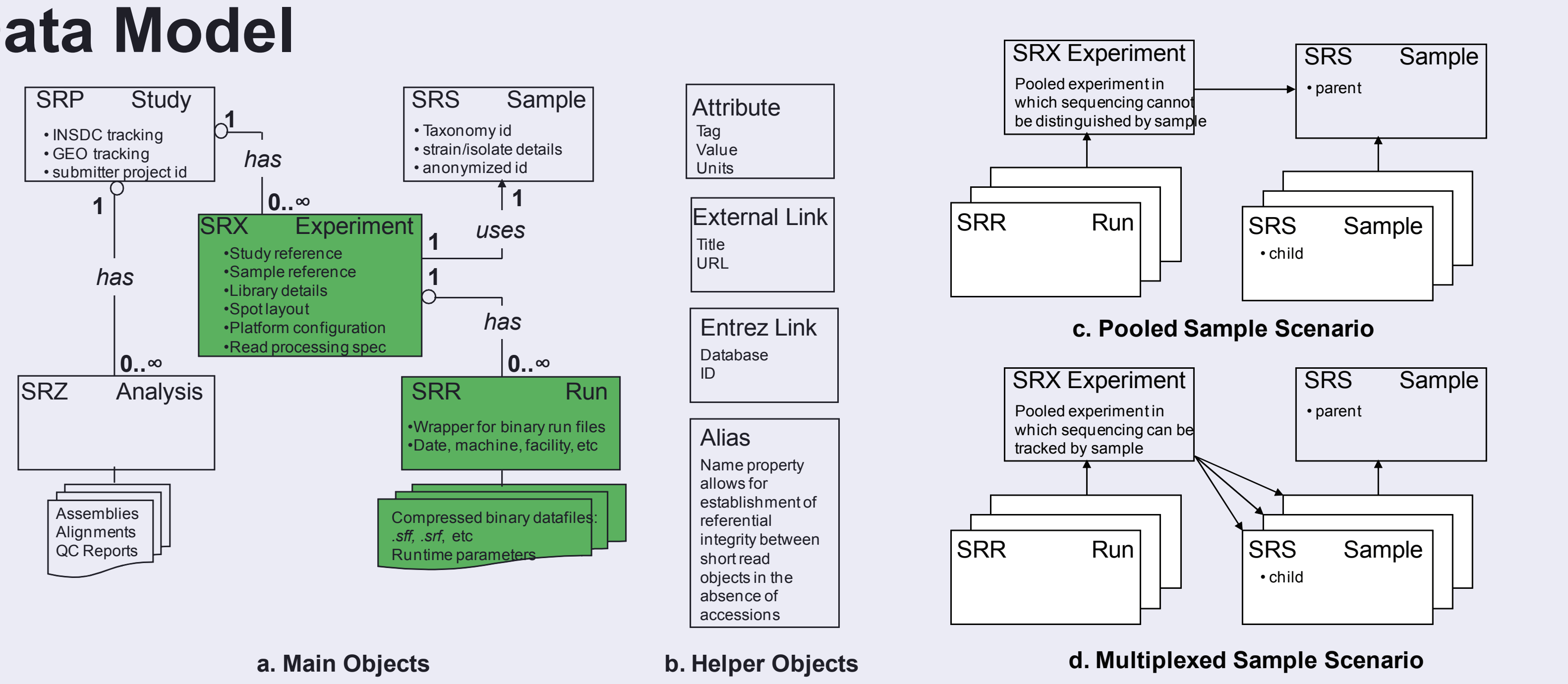
Hold Until Publish Feature

- 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 Model



a. Main Objects

b. Helper Objects

c. Pooled Sample Scenario

d. Multiplexed Sample Scenario

Figure 4: 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).

Deposit Network

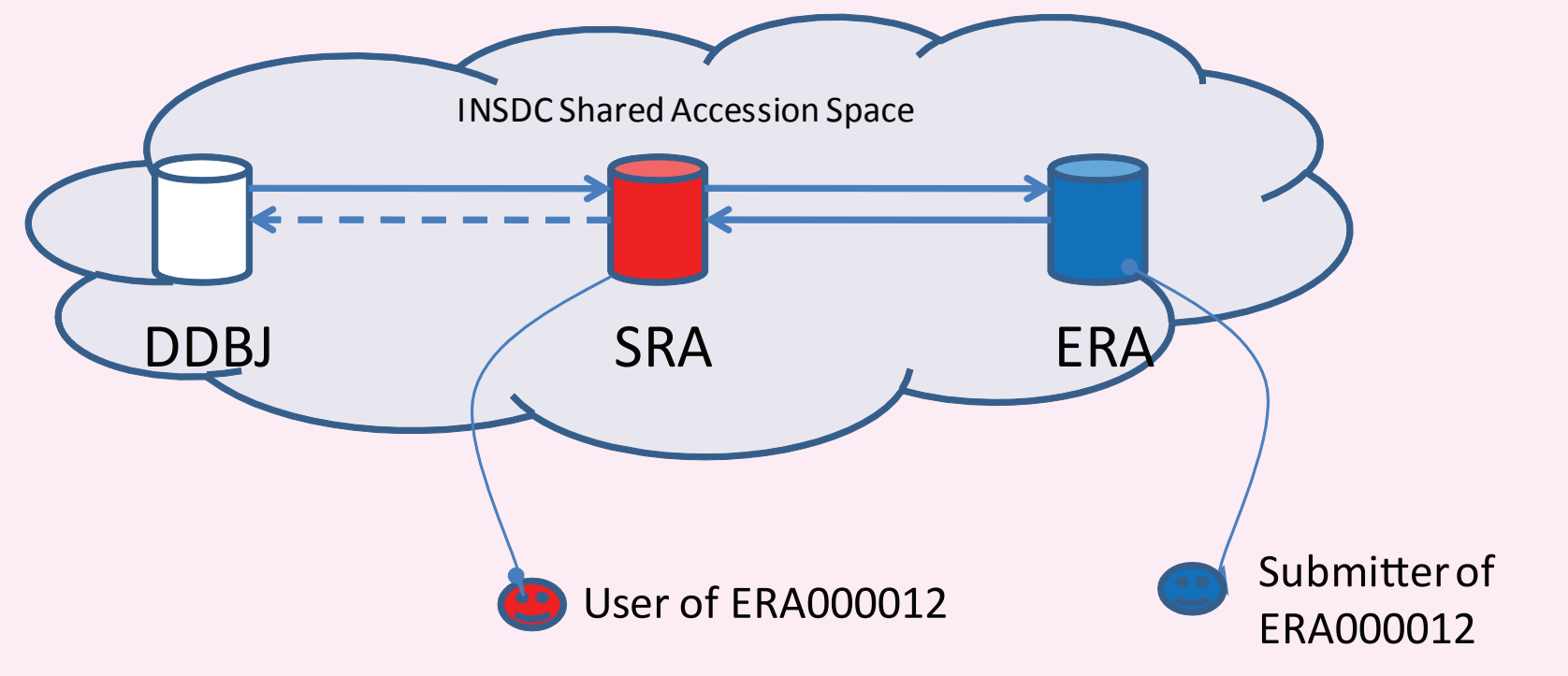


Figure 5: Deposits are currently brokered by DNA Data Bank of Japan (DDBJ) to NCBI, and are mirrored between the SRA at NCBI and the European Read Archive (ERA) at European Bioinformatics Institute (EBI). This gives users across the world immediate access to deposits at the respective archives.

A global name space ensures that accessions generated by the individual archives are recognized by the collaborating archives.

Acknowledgements

SRA Technical Team
Sergey Ponomarev
Robert Sanders
Aleksey Vysokolov
Andriy Klymenko
Andrei Petrow
Zinaida Belaia
Christopher O'Sullivan

Collaborators
European Bioinformatics Institute
DNA Data Bank of Japan
NHGRI

Vendors
454 Life Sciences/Roche Applied Science
Illumina, Inc.
Applied Biosystems, Inc.
Asperasoft, Inc.

Submitters
Baylor College of Medicine
Broad Institute
J. Craig Venter Institute
Joint Genome Institute
Sanger Centre
Washington University in St Louis
and many others...

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

Visit the NCBI Short Read Archive at http://www.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?