

Accelerating research with the NCBI Sequence Read Archive on the commercial cloud



S. Sherry, A. Boshkin, J. Brister, R. Connor, D. Deriy, A. Dillman, K. Durbrow, A. Efremov, L. Fleischmann, G. Godynskiy, D. Ignatovich, A. Iskhakov, A. Johnson, M. Kimelman, A. Klymenko, A. Kochergin, C. Leubsdorf, K. McDaniel, A. Mnev, C. O'Sullivan, S. Ponomarov, W. Raetz, K. Rodarmer Jr, K. Rodarmer Sr, O. Shutov, D. Soren, A. Stine, M. Vartanian, E. Yaschenko, and V. Zalunin.

Summary

The NCBI Sequence Read Archive (SRA) is NCBI's comprehensive collection of next generation sequence data from human and non-human organisms. At over 10 petabytes, SRA is NCBI's largest research archive and has exceeded the capability of most sites to replicate locally. Recognizing this, NCBI and NIH have recently replicated SRA content to both Amazon and Google commercial cloud platforms as part of the NIH STRIDES program. NCBI is providing community education, and training opportunities to build relationships with SRA users through webinars, online tutorials and a series of NCBI-hosted hackathons that bring together users to work with topical and problem-relevant subsets of SRA data.

Why SRA?

Taken as a broad collection of sequences sampled across the tree of life, SRA data are mined for new discoveries about genomic sequence, natural variation, antimicrobial resistant genes, gene expression, methylation states, and previously undescribed genes and species, strains, or viral isolates. By size, SRA is equal parts public and controlled access data — public data includes non-human sequences and human RNAseq and genomic data where individual consent has been provided for the open and unrestricted use of their data. Controlled access data is sequence information from human research study participants supported by NIH and access is restricted and controlled through dbGaP approval protocols.

How is NCBI making SRA access useful to me?

Moving SRA data to cloud platforms will benefit users by enabling meaningful and timely access to SRA's exponentially growing corpus of data.

This migration of data is accompanied by:

- improvements to the SRA Run Selector
 - a new data model that supports access to both submitted and normalized data formats
 - improvements to the SRA toolkit
 - conformance to NIH's new standards for Identity and Authorization Management.
- Researchers are invited to join the NCBI cloud user community, participate in Codeathon events, and explore our developing knowledge base of self-directed education resources to design powerful and affordable cloud-based analysis workflows.

SRA vision — Promote an ecosystem to advance genome research through cloud-centered compute environments. In practice this means supporting dual use through a hybrid storage model.

The Dual Uses of SRA

Reproducibility	Discovery
When user requires data from a specific study for reanalysis, replication or reprocessing tasks.	When user needs to search and compute on data that spans multiple submissions.
Not interested in standardized format.	Processing tasks regard SRA as a corpus with a standardized representation.
Requires access to study-specific markup in the submission.	Focus of NCBI to curate & optimize.

SRA serves data in both original and SRA formats

HOT STORAGE — ORIGINAL FORMAT	HOT STORAGE — SRA FORMAT
Reproducibility is prioritized for data submitted within the last year.	Discovery & Search are prioritized for SRA format where core elements are presented in a uniform representation.
COLD STORAGE for less-active data sets. Frozen data will be restored to user compute area within 24 hours.	SRA format will optimize data objects where possible, e.g. replacing reads with contigs+coverage.

A new SRA data flow — Submissions to SRA are now processed at NCBI into archive format. Both the original and archive versions of the data are then copied to regional storage systems in Amazon and Google.

Finding the data — With SRA run data now available in multiple locations, users have several options to identify the best source of the data for their analysis task. The SRA Run Browser shows all locations of a particular submission on the Data Access tab.

Data details are shown for both archive format and original submission formats.

There are three locations for data: NCBI (download only), Amazon (AWS), and Google (GCP).

Recent original files are available as cloud storage objects.

Older submissions are moved to cold storage and can be restored from there into user compute areas.

Data can be downloaded freely from NCBI or AWS Public Dataset Program. All other cloud locations must be accessed from within Google or Amazon.

worldwide can be downloaded from anywhere for free.

s3.us-east-1 is only free to access from machines running in Amazon's us-east-1 region.

gs.us is only free to access from machines running in Google's gs.us region.

All other access or data download will generate egress charges to the user.

Indicates whether a cloud service user account is necessary for data access. "anonymous" access means general public access.

Accessing SRA via Run Selector

SRA Run Selector now supports eRA login. Approved dbGaP users can quickly build lists of SRA runs by project.

1. Select facets

2. Select runs

3. Select cloud

4. Select execution

5. Select billing project

6. Select region, must be near the data

7. Select execution

SRA in the Cloud

Sequence Read Archive (SRA) data is now available on the Google and Amazon Web Services (AWS) clouds. NCBI tutorials will help you get acquainted with these tools in the cloud so you can search and retrieve data based on your needs.

Available Data

Publicly-available, unassembled read data is available for access and compute through the cloud providers listed above. Authorized-access human data will be available by the end of 2019. There are several ways you can access this data. Explore the links below to find out which approach is best for you.

Links

- Installing the SRA Toolkit
- Accessing SRA with Fusera
- SRA in BigQuery

BigQuery provides fast, programmatic access to SRA metadata and supports a large collection of client libraries.

Engage

SRA has deposited its metadata into BigQuery to provide programmatic access to these data. You can now search across the entire SRA using sequencing methodologies and sample attributes.

NCBI is piloting BigQuery access to help users leverage the benefits of elastic scaling and parallel execution of queries.

NCBI Codeathons

Bringing people with diverse backgrounds together to build tools for advanced analysis of biomedical data.

<https://ncbi-codeathons.github.io/>

NCBI Codeathon topic areas of potential interest to users of SRA

Virus Characterization and Discovery	Next Event: Virus 2	Nov 2019
Cloud-based search/retrieval to support data set identification based on organism/genetic content.		
RNAseq Analysis and Visualization	Next Event: Single cell RNAseq	Jan 2020
Standardized work flows to support data inter-operability and novel transcriptomic profile discovery.		
Pangenomics and Genome Graphs	Next Event: Pangenome 2	Aug 2020
Population wide comparisons of genetic data from humans and other multi-cellular organisms.		

Educational Resources for Investigators

Educational materials developed with NIH STRIDES program and other partners. Materials, courses, seminars, conference demonstrations. Hands on cloud data and tools engagement and education. User testing, feedback, and partnerships.

NCBI Cloud Tools & Workflows

Basic Local Alignment Search Tool (BLAST)

Compare nucleotide or protein sequences to sequence databases and calculate statistical significance

Dockerized BLAST+: Run BLAST on docker*

Coming soon! Cloud native version of BLAST

* Also available on GitHub

Sequence Read Archive Toolkit (SRA)

Explore the largest public repository of next generation sequence data

SRA toolkit: Access public SRA data in your desired format and cloud bucket.

Fusera: Interrogate SRA data outside a cloud bucket.

BigQuery: Filter SRA and BioSample metadata.

NCBI Webinar Using SRA Toolkit to access data from dbGaP and SRA

Prokaryotic Annotation Genome Pipeline (PGAP)

Annotate bacterial and archaeal genomes (chromosomes and plasmids)

Dockerized PGAP: PGAP on the cloud*

Assemble with dockerized SKESA before annotating with PGAP

* Also available on GitHub