National Institutes of Health
Office of Data Science Strategy

Sequence Read Archive Data Working Group Interim Report

Susan Gregurick, Ph.D.
Associate Director for Data Science and
Director, Office of Data Science Strategy

Kristin Ardlie, Ph.D.
Director of the GTEx Laboratory Data Analysis and Coordination Center
Broad Institute of Harvard and MIT

Council of Councils | May 20-21, 2021

Agenda

- Update on Current SRA Status
- Prior SRA Working Group’s Recommendations and Responses
- SRA Working Group’s New Charge
- New Recommendations

Agenda

- Update on Current SRA Status
- Prior SRA Working Group’s Recommendations and Responses
- SRA Working Group’s New Charge
- New Recommendations

Background – SRA in the Cloud

The NCBI Sequence Read Archive (SRA) is a crucial resource.

- One of NIH's largest and most diverse datasets, representing genome diversity throughout the
  tree of life.
- Essential for research in pathogen characterization, linking diseases with genetic and epigenetic variation, bioinformatics, and evolutionary biology.

SRA is now available in the cloud.

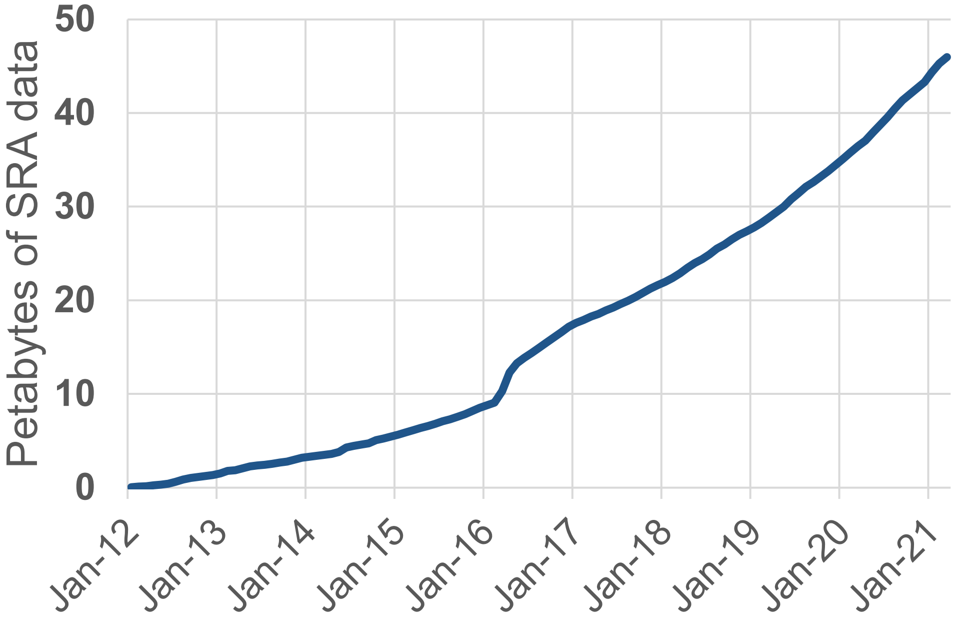
- Migration to Google Cloud Platform (GCP) and Amazon Web Services (AWS) began in 2019 through the STRIDES Initiative.
- First and largest biomedical dataset in the cloud.

SRA is large and frequently accessed.

- Currently over **13** million records, **16.7** PB of data, growing exponentially.
- During 2020, over **48** PB of SRA data was downloaded, and **>10%** of data was downloaded from cloud platforms.

Historic and Projected SRA Growth in Petabytes

SRA has experienced rapid historic growth*



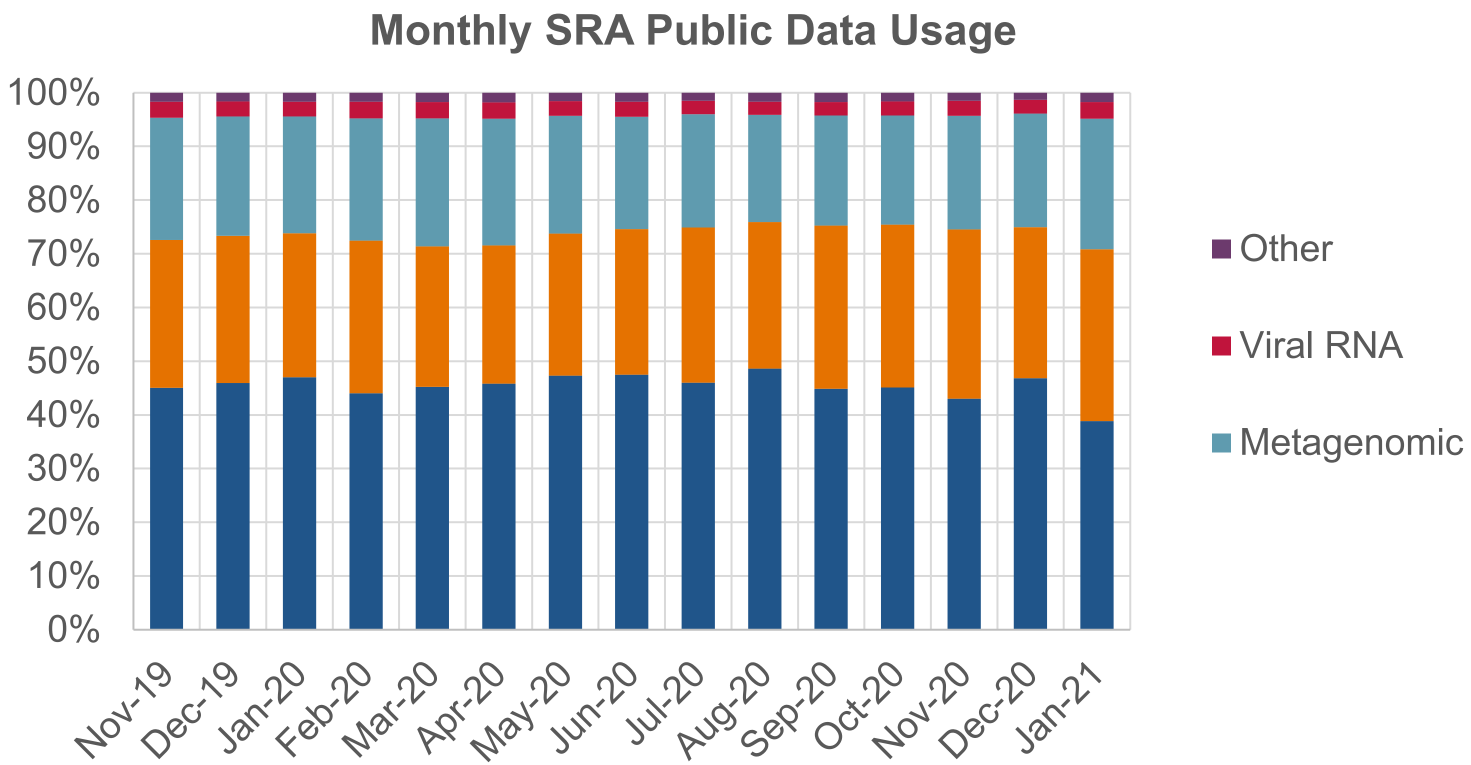
SRA is projected to grow rapidly over the next 5 years*

Date	ETL+BQS format (Petabytes)	Original, source format (Petabytes)
Sept 30, 2021	17.0	30.8
Sept 30, 2022	20.0	36.1
Sept 30, 2023	23.5	42.0
Sept 30, 2024	27.7	49.3
Sept 30, 2025	32.5	57.5

**Figures include originally submitted source data and normalized ETL data formats*

SRA data Usage by Types

- Analytics implemented for public data and can monitor usage based on a variety of data attributes
- Currently implementing analytics for controlled access data – controlled access is 31% of total SRA


Monthly SRA Public Data Usage

Agenda

- Update on Current SRA Status
- Prior SRA Working Group’s Recommendations and Responses
- SRA Working Group’s New Charge
- New Recommendations

Prior SRA Working Group Recommendations

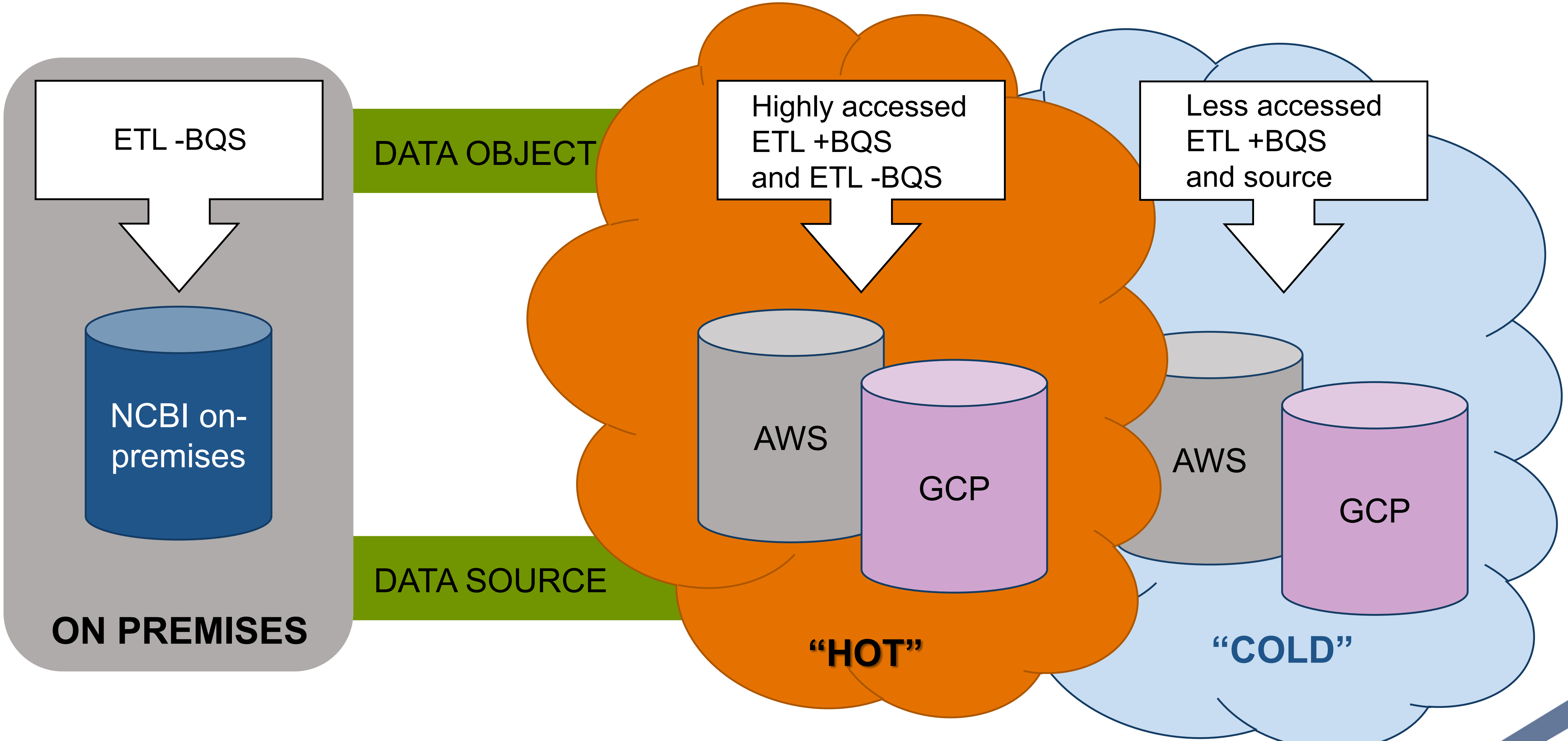
Recommendation: A new model for SRA data storage and retrieval in the cloud

- Two formats of data in the cloud: Original and normalized formats
- Data in hot or cold storages depending on format and usage

Addressed:

- Hybrid storage model: planned architecture, end state

Hybrid Storage Model: Planned Architecture, End State



Prior SRA Working Group Recommendations

Recommendation: Communication of the model

- Public channel to describe the cost models including cost and storage
- Education to researchers about data access, storage, computing in cloud usage

Addressed:

- NIH/ODSS conducted a Request For Information (RFI)

Findings from Request for Information

75 responders were received across 12 countries and affiliated with academic (77.3%), government (9.3%), industry (6.7%), and nonprofit (6.7%) institutions.

65% reported that they appreciated the NIH's efforts to maximize efficiency in retrieval through the new format.

Over 80% reported that BQS are critical for their analyses and requested that data be retained in the original FASTQ format.

Only 17% reported that they were currently using SRA in cloud.

52% responders expressed concerns about unpredictable cost of using cloud computing platforms and felt that cost was more easily managed on their local high performance computing system.

Prior SRA Working Group Recommendations

Recommendation: Continued research to inform changes to the model over time

- Monitor and adjust the model based on actual costs of user working in cloud
- NIH should consider funding efficiency optimization research for use in cloud computing

Addressed:

- Partnership with AWS Open Data Program (ODP) created new opportunity to update model and provide free access to ETL+BQS data
- Implemented cold storage and retrieval

Hybrid Storage Model Updated to Reflect Integration with AWS Open Data Project

AWS ODP supports free user access to SRA (ETL+BQS) data at no cost to NIH.

GCP	Source	ETL +BQS	ETL –BQS
Hot	Commercial bucket (898TB)	Commercial bucket (8.3PB)	Commercial bucket (2.8PB)
Cold	Commercial bucket (27.9PB)	Commercial bucket (7.3PB)	
AWS	**Source**	**ETL +BQS**	**ETL –BQS**
Hot		Commercial bucket (13.6PB)	Commercial bucket (N/A)
Hot	Open data bucket (8.38TB)	Open data bucket (>10.2PB)	
Cold	Commercial bucket (27.9PB)	Commercial bucket (2PB)	

Agenda

- Update on Current SRA Status
- Prior SRA Working Group’s Recommendations and Responses
- SRA Working Group’s New Charge
- New Recommendations

SRA Working Group FY21

Co-chairs

Susan Gregurick, Ph.D.
National Institutes
of Health

Kristin Ardlie, Ph.D.
Broad Institute

Members

Kevin B. Johnson, M.D.
Vanderbilt University
Medical Center

Jinghui Zhang, Ph.D.
St. Jude Children's
Research Hospital

Daniel Danciu, Ph.D.
ETH Zürich
Biomedical Informatics

Andy Bentley
KU Biodiversity Institute

Anthony Philippakis, M.D., Ph.D.
Broad Institute

Brandi Davis-Dusenbery, Ph.D.
Seven Bridges

Rob Edwards, Ph.D.
San Diego
State University

SRA Data Working Group Charge

Analysis and evaluation of strategies for, or changes to, SRA data storage, management, and access, **including impact for the biomedical research community**

Recommendations on data retention, data models and/or data usage that will keep costs to NIH within sustainable levels while maintaining community access to this large public data resource

Vision for future needs or opportunities, including sustaining SRA as a community resource.

Final report requested by the September 2020 Council of Councils meeting

Agenda

- Update on Current SRA Status
- Prior SRA Working Group’s Recommendations and Responses
- SRA Working Group’s New Charge
- New Recommendations

New Recommendations

- Reduce costs and ensure that data remain equitable and sustainable
- Explore tolerance and frequency for data retrieval in cost models
- Explore data usage, data type, search, and access in the cloud
- Consider more cloud vendors to host SRA data
- Consider the needs of users who do not use GCP or AWS platforms
- Promote cloud computing usage with representative examples
- Pursue training and user feedback (e.g., workshops, tutorials)
- Consider incentives for researchers using SRA to develop tools/algorithm
- Evaluate impact from SRA

Recommendations

1. **Reduce costs and ensure that data remain equitable and sustainable**

- Cost-effective cloud computing is challenging for many users.
- Communicate to stakeholders about storage and retrieval models and the open access program.
- Reach out to minority institutions for targeted cloud-based training

2. **Explore tolerance and frequency for data retrieval in cost models**

- Understand benefits of moving more data into cold storage? Understanding how cold storage retrieval time impacts users?
- Further assessment is needed to fully understand the cost model.

Recommendations

3. **Explore data usage, data type, search, and access in the cloud**
   - Understand relationship between data types and compute cost. Many computing costs are related to the data type.
   - Develop data-driven storage solutions, which involves defining the dynamics of SRA accession usage, identifying low-usage data, and moving low-usage data to cold storage

4. **Consider more cloud vendors to host SRA data**
   - More cloud vendors will bring a diversity of perspectives and insights to SRA and serve the broader research community.
   - Need to identify approaches that mitigate costs associated without replicating SRA across more cloud providers
   - Pilot approaches to cross-platform computing that could support more efficient storage of SRA data (fewer copies of each accession)

Recommendations

5. **Consider the needs of users who do not use GCP or AWS platforms**

- Education and training on to familiarize users with cloud technologies & data access
- Leverage cloud credits to promote more cloud users in the research community
- Provide data via user-friendly data analyses platforms such as Galaxy, Terra

6. **Promote cloud computing usage with representative examples**

- Provide normalized or pre-processed datasets
- Establish best practices of cloud computing with estimated schedules and transparent costs.
- Train-the-trainers (e.g. librarians, power users) who have a channel or mechanism for outreach within their institution or networks

Recommendations

7. **Pursue training and user feedback (e.g., workshops, tutorials)**

- Understand users and learn their comfort level with cloud and computing
- Efforts tailored on different user segments based on their familiarity with the cloud
- Accelerate awareness, education & training for novice and intermediate cloud users
- Easily accessed video (YouTube) tutorials to demonstrate common use cases

8. **Consider incentives for researchers using SRA to develop**

- Incentivize investigators to invest in community-driven efforts to develop tools and algorithms for using SRA in cloud.

Recommendations

9. Evaluate impact from SRA

- Conduct PubMed searches to track the success of research projects using SRA data/tools
- Partner with analysis platforms to obtain reports on SRA data access frequency and types
- Define the SRA user communities
  - users who contribute to the SRA
  - users who access data from the SRA
  - publications generated using SRA data
- Conduct a survey with researchers on the use of SRA data in their curricula for training
- Engage with training platforms (e.g., Galaxy) to obtain SRA usage information
- Obtain information on intended use from users during download through an optional description field
- Partnership with cloud vendors to provide cloud credits for training programs or workshop

Questions & Discussion