National Institutes of Health
Office of Data Science Strategy

Sequence Read Archive Data Working Group Final Report

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

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

September 17, 2021

Agenda

- Updates on Current SRA Status
- FY 21 SRA Working Group Charge
- FY 21 SRA Working Group Recommendations
- Key Performance Principles for Evaluating Success of SRA
- Future Work and Considerations

Agenda

- Updates on Current SRA Status
- FY 21 SRA Working Group Charge
- FY 21 SRA Working Group Recommendations
- Key Performance Principles for Evaluating Success of SRA
- Future Work and Considerations

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.

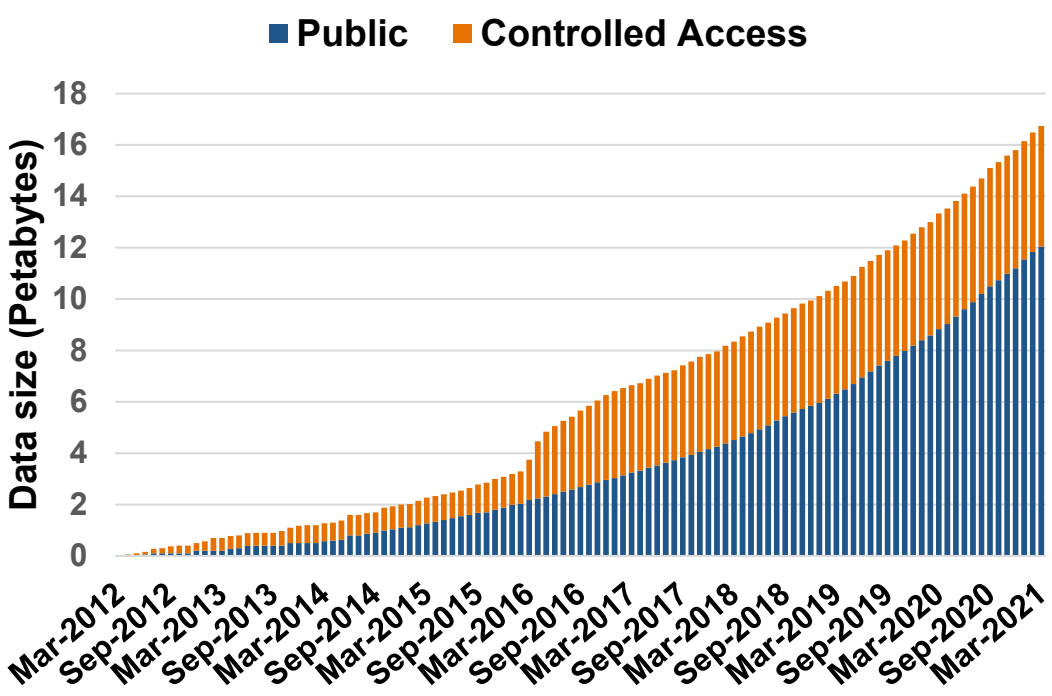
- Migration to Google Cloud Platform (GCP) and Amazon Web Services (AWS) began in 2019 through the STRIDES Initiative.
- First and largest biomedical dataset in the cloud.

SRA is large and frequently accessed.

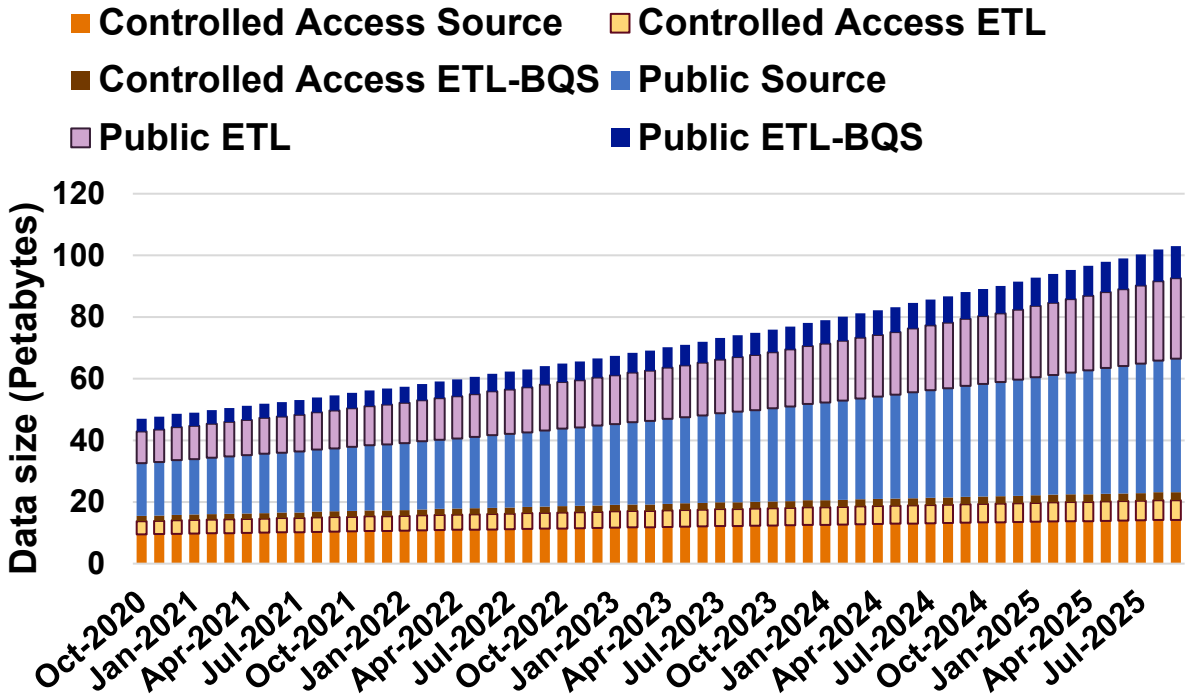
- Currently over **14.5** million records, **16.5** PB of data, growing exponentially.
- During 2020, over **48** PB of SRA data was downloaded, and **>10%** of data was downloaded from cloud platforms.

Historic and Projected SRA Growth in Petabytes

SRA has experienced rapid historic growth*



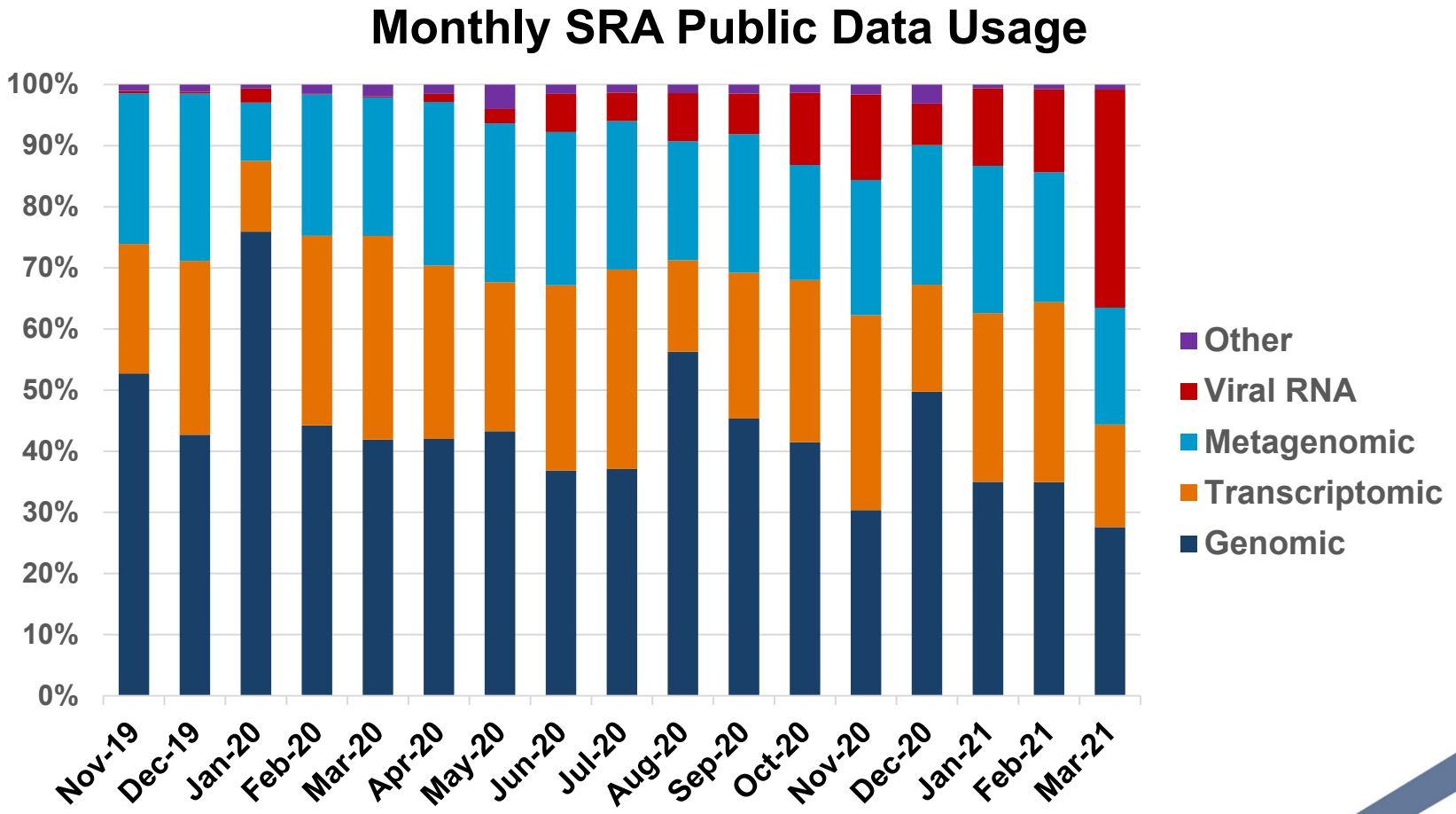
SRA is projected to grow rapidly over the next 5 years*



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

SRA Data Usage by Types

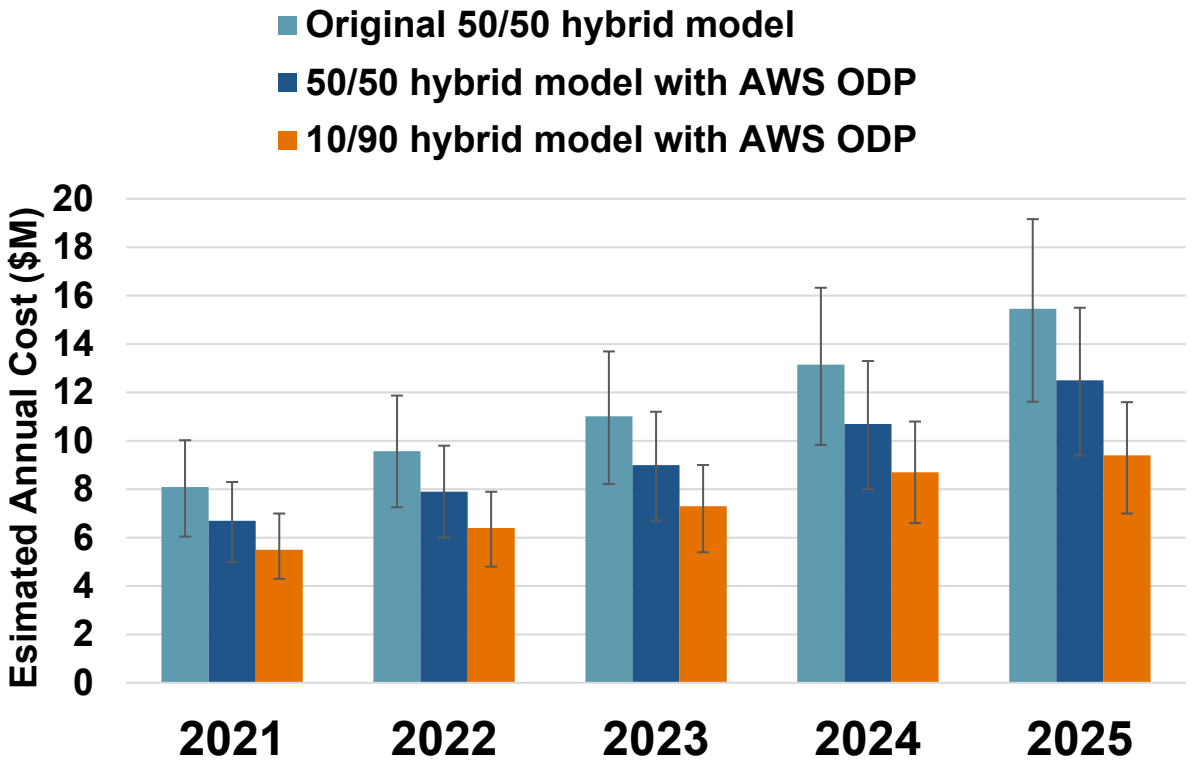
- Analytics implemented for public data and can monitor usage based on a variety of data attributes
- Currently implementing analytics for controlled access data – controlled access is 31% of total SRA



Projected SRA Cloud Storage and Costs to NIH

Current implementation of SRA hybrid storage model

	GCP Hot	GCP Cold	AWS ODP Hot	AWS Hot	AWS Cold
Source	1 PB	27 PB	1 PB		27 PB
ETL	8 PB	7.8PB	10.1 PB	3.7 PB	2 BP
ETL-BQS	2.8 PB	N/A	N/A	N/A	N/A



Agenda

- Updates on Current SRA Status
- FY 21 SRA Working Group Charge
- FY 21 SRA Working Group Recommendations
- Key Performance Principles for Evaluating Success of SRA
- Future Work and Considerations

FY21 SRA Working Group

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

Andy Bentley
KU Biodiversity Institute

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

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

Daniel Danciu, Ph.D.
ETH Zürich
Biomedical Informatics

**Brandi Davis-
Dusenbery, Ph.D.**
Seven Bridges

Rob Edwards, Ph.D.
San Diego
State University

FY21 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.

Agenda

- Updates on Current SRA Status
- FY 21 SRA Working Group Charge
- FY 21 SRA Working Group Recommendations
- Key Performance Principles for Evaluating Success of SRA
- Future Work and Considerations

FY21 SRA Working Group Recommendations

- Promote cloud usage and ensure SRA data usage with equity and sustainability
- Explore data usage, access frequency and tolerance for cloud data retrieval in cost model
- Consider incentives for researchers using SRA to develop tools/algorithms for cloud computing
- Evaluate impact from SRA

Recommendation 1:

Promote cloud usage and ensure SRA data usage with equity and sustainability

- Consider more cost-effective strategies for data deposition and use through communications and negotiations with cloud providers (open or commercial)
- Provide guidance and transparency for SRA enabled cloud computing
- Promote cloud computing usage with representative examples, training programs and user feedback (e.g., workshops, tutorials)
- Consider the needs of users who do not use GCP or AWS platforms, including those from under-resourced institutions

Recommendation 2:

Explore data usage, access frequency and tolerance for cloud data retrieval in cost model

- 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
- Understand cold storage retrieval time/cost impact to users
- Understand relationship between data types and compute cost

Recommendation 3:

Consider incentives for researchers using SRA to develop tools/algorithm for cloud computing

- NIH should incentivize investigators to promote cloud native analyses and collaborations through community-driven efforts to develop and enhance cloud-based tools and algorithms
- Advanced petabyte scale sequence searching tools are needed
- NIH should focus on exemplars like SARS-CoV-2 and metagenomic data, and associated metadata for improving platform agnostic cloud data usage

Recommendation 4:

Evaluate impact from SRA

- Consider citations in publications or other citable objects
- Partner with an analysis platform, develop metrics to capture user statistics and surveys
  - Conduct PubMed searches or reach out to both research community and journal editors to track the success of research projects using SRA data
  - Facilitate metrics development for assessing the SRA data usage, advocacy, integration, and SRA interoperability with the other NIH data repositories
  - Partner with an analysis platform to obtain reports on SRA data access frequency and types
  - Working group members suggested that NIH could consider conducting a survey with researchers on their 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 of data from users during download through a list of common user cases with an optional description field

Agenda

- Updates on Current SRA Status
- FY 21 SRA Working Group Charge
- FY 21 SRA Working Group Recommendations
- Key Performance Principles for Evaluating Success of SRA
- Future Work and Considerations

Key Performance Principles for Evaluating Success of SRA (A)

Category	SRA Key Performance	Key Performance Principle
A. Data quality	Creation and distribution of original, ETL, and ETL-BQS formats that meets criteria for fitness for purpose a. ETL format b. Specification of new ETL-BQS formats c. Original format	• *Data are readily available in common formats (e.g., original format, ETL, ETL-BQS) to support both cloud and non-cloud users' need.* • *As defined by a time period, the most frequently accessed datasets are always readily available from multiple sources.*
	Distribution of SRA, BioSample, and BioProject metadata with sequence data a. Support access to sequence metadata b. Support data communities with specific metadata needs	• *No intentional discrepancies in ability to discover, access, compute upon, or analyze data, data to adhere to FAIR principles.*
	Development of tools to support search of data a. Metadata-based search b. Sequence-based search	• *Sequence search and data analysis tools are available for supporting SRA and specific metadata sequence data, metadata to be digestible by third party tools.*
	Improve value of data a. Provide analysis of data (e.g., SRA Sequence Taxonomic Analysis Tool STAT) b. Support metadata packages	• *Training Datasets/Tools are updated, maintained and always available in ALL locations with no associated costs.*

Key Performance Principles for Evaluating Success of SRA (B)

Category	SRA Key Performance	Key Performance Principle
B. Equitable user access	Distribution of data in hot and cold storage	• *Optimize hot and cold storage distribution to save costs for both NIH and users.*
	Support data access for both cloud and non-cloud users a. Open Data and Public Dataset Programs b. Tools/interfaces to support data retrieval from hot and cold storage	• *Open Data and public datasets are immediately or up to less than 24hours (dependent on data size) available.* • *User costs are well defined and a mechanism to ameliorate those costs for under resourced investigators is developed.*
	Replicate SRA among STRIDES cloud service providers	• *Costs to NIH are under allocated amount in the budget and are increasing at a rate no greater than the yearly expected increase in that allocated amount.*
	User costs on data retrieval/egress to different cloud platforms	*listed above*
	Training and outreach for competency a. General; b. Minority-focused	• *Training Datasets/Tools are updated, maintained and always available in ALL locations with no associated costs.*
	Partnerships among US Government agencies	• *Collaboration with other US Government agencies to serve broader research communities by developing advanced tools and improving data repositories interoperability*

Agenda

- Updates on Current SRA Status
- FY 21 SRA Working Group Charge
- FY 21 SRA Working Group Recommendations
- Key Performance Principles for Evaluating Success of SRA
- Future Work and Considerations

Future Work and Considerations

- User-centered focus
  - Understand user cost and cost awareness education
  - Cloud benchmarks are needed
- Interoperability standards to extend impact and reduce cost
- Streamline guidance for cloud costs
- Provide intermediate or processed data on cloud
- Have a funding mechanism to support optimizing the existing cloud computing tools
- Promote multi-cloud optimization of highly used (and new) tools
- Promote submission of robust sample metadata

Questions & Discussion