Sequence Read Archive Data Working Group Final Report

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

Kevin B. Johnson, MD, MS
Cornelius Vanderbilt Professor and Chair, Biomedical Informatics
Professor of Pediatrics
Vanderbilt University Medical Center

September 11, 2020

NIH National Institutes of Health
Office of Data Science Strategy

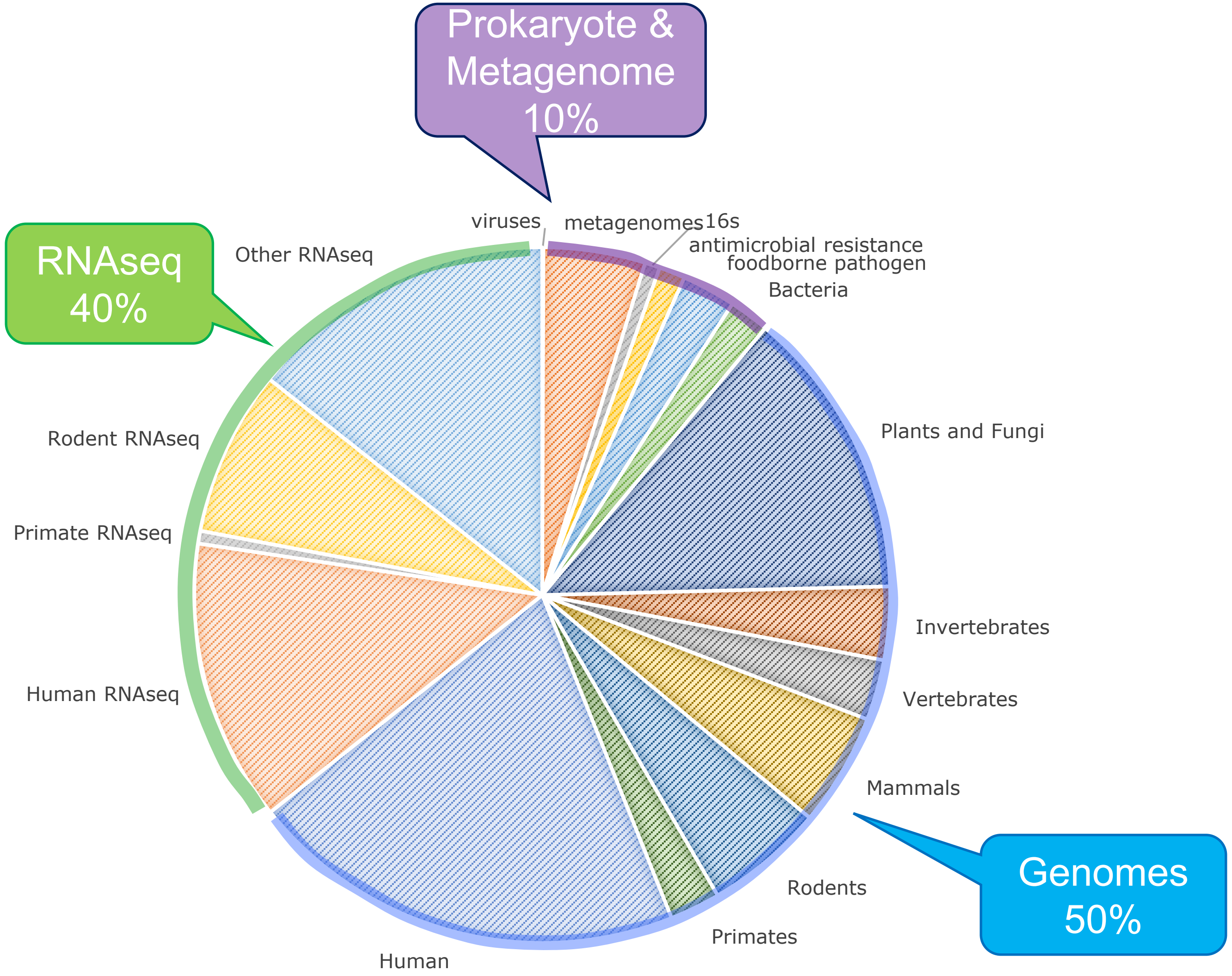
Agenda

- Background and Challenge
- Working Group Charge
- Recommendations

Agenda

- Background and Challenge
- Working Group Charge
- Recommendations

Public SRA Data – 8.8 PB



All SRA Data Under Discussion: 26.8 PB

	Controlled Access	Public Access
Google	4.6 PB	8.8 PB
AWS	4.6 PB	8.8 PB

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 10 million records, 13.4 PB of data, growing exponentially.
- During 2019, over 1.2 million visitors downloaded over 8.5 PB of SRA data, and 20% of the visits were from cloud IP addresses.

SRA Formats

- **Original format**
  - The format in which data are initially submitted to SRA; NCBI supports 20 possible file formats.
- **Normalized format**
  - A standardized format to which NCBI converts all SRA data, also called ETL: extract, transform, load.
  - Currently the only format available to researchers to download from NCBI site or access in the cloud.

Projected SRA Growth

Format	*July 2019*	*2020*	*2021*	*2022*	*2023*	**2024**	**2025**
Original	23	28	35	44	57	70	87
SRA Normalized (ETL)	10	16	20	26	33	39	49

Current size and projected* future growth of SRA by format type (in petabytes).

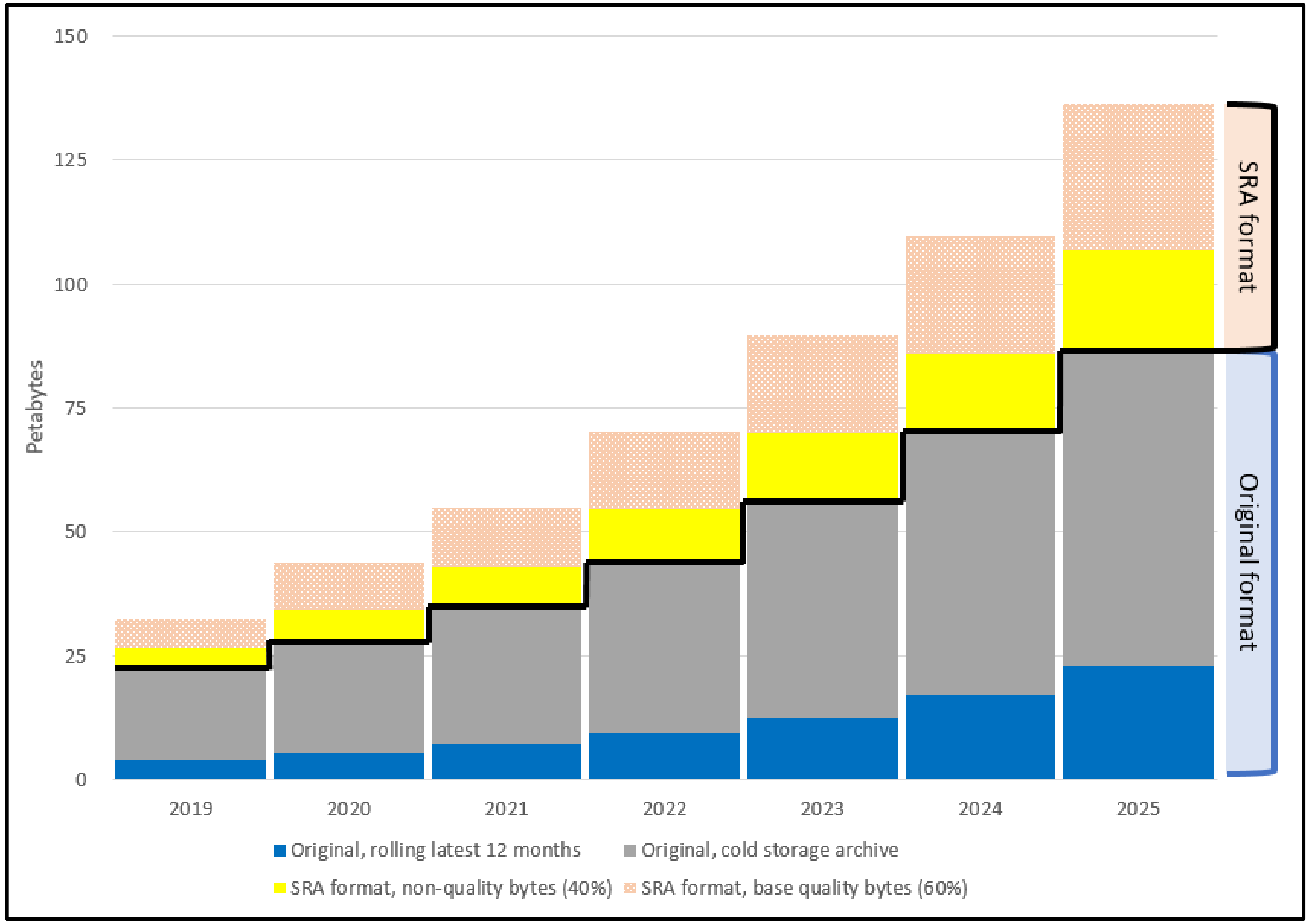
*Sizes extrapolated from a best-fitting exponential model of SRA growth using archive annual growth from 2015 to 2018.

SRA Projected Growth

- **SRA growth is exponential.** The two archive formats have different sizes and rates of growth as indicated.

- SRA normalized format can be subdivided into **two categories** of data types:
  - BQS (pink)
  - everything else (yellow)

- **BQS**: Base quality scores, or quantitative representations of the probability of an error at a base.
  - The largest single component of normalized data by type
  - Most file types have one BQS per letter of sequence
  - Difficult to compress because of the large number of possible values



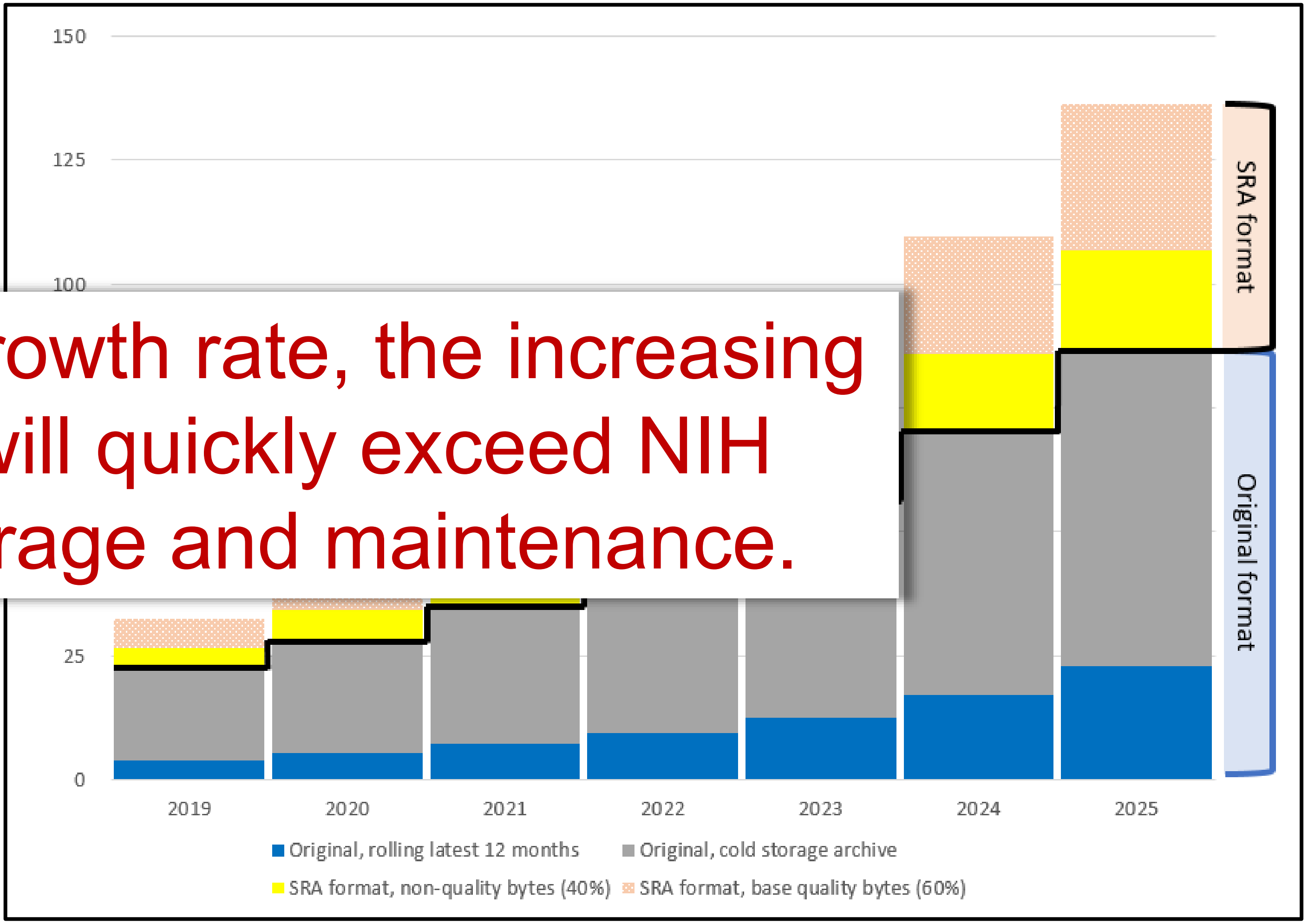
SRA Projected Growth

- **SRA growth is exponential.** The two archive formats have different sizes and rates of growth as indicated.

- SRA normalized format can be subdivided into **two categories** of data types:
  - BQS (pink)
  - everything els

- **BQS**: Base qua representations error at a base.
  - The largest single component of normalized data by type
  - Most file types have one BQS per letter of sequence
  - Difficult to compress because of the large number of possible values

At the current growth rate, the increasing size of SRA will quickly exceed NIH budget for storage and maintenance.



Agenda

- Background and Challenge
- Working Group Charge
- Recommendations

SRA Data Working Group

Co-chairs

Susan Gregurick, Ph.D.
National Institutes
of Health

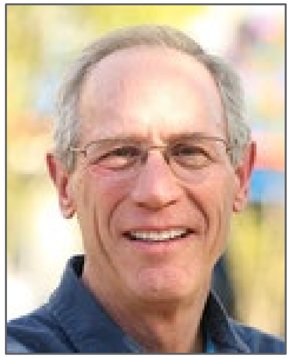
Kevin B. Johnson, M.D.
Vanderbilt University
Medical Center

Members

Kristin Ardlie, Ph.D.
Broad Institute of MIT,
Harvard University

Toby Bloom, Ph.D.
New York
Genome Center

Rob Edwards, Ph.D.
San Diego
State University



Rick Horwitz, Ph.D.
Allen Institute for
Cell Science

Hyun Min Kang, Ph.D.
University of Michigan
School of Public Health

**Debbie
Nickerson, Ph.D.**
University of Washington

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

SRA Data Working Group Charge

- Provide recommendations to the Council on key factors for storing and managing SRA data on cloud service provider environments
- Evaluate and identify solutions to maintain efficiencies in the storage footprint of SRA
  - Evaluate the use of BQS and format compression strategies

Final report requested by the September 2020 Council of Councils meeting

Agenda

- Background and Challenge
- Working Group Charge
- Recommendations
  - **A new model for SRA data storage and retrieval in the cloud**
  - **Communication of the model**
  - **Continued research to inform changes to the model over time**

Recommendations

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

	Hot storage	Cold storage
Normalized data with BQS	More actively accessed half	Less actively accessed half
Normalized data without BQS	All	
Original format data (with BQS)		All

- BQS would be retained in original format data, and two versions of SRA normalized format data would be maintained: one with quality scores and one without them.

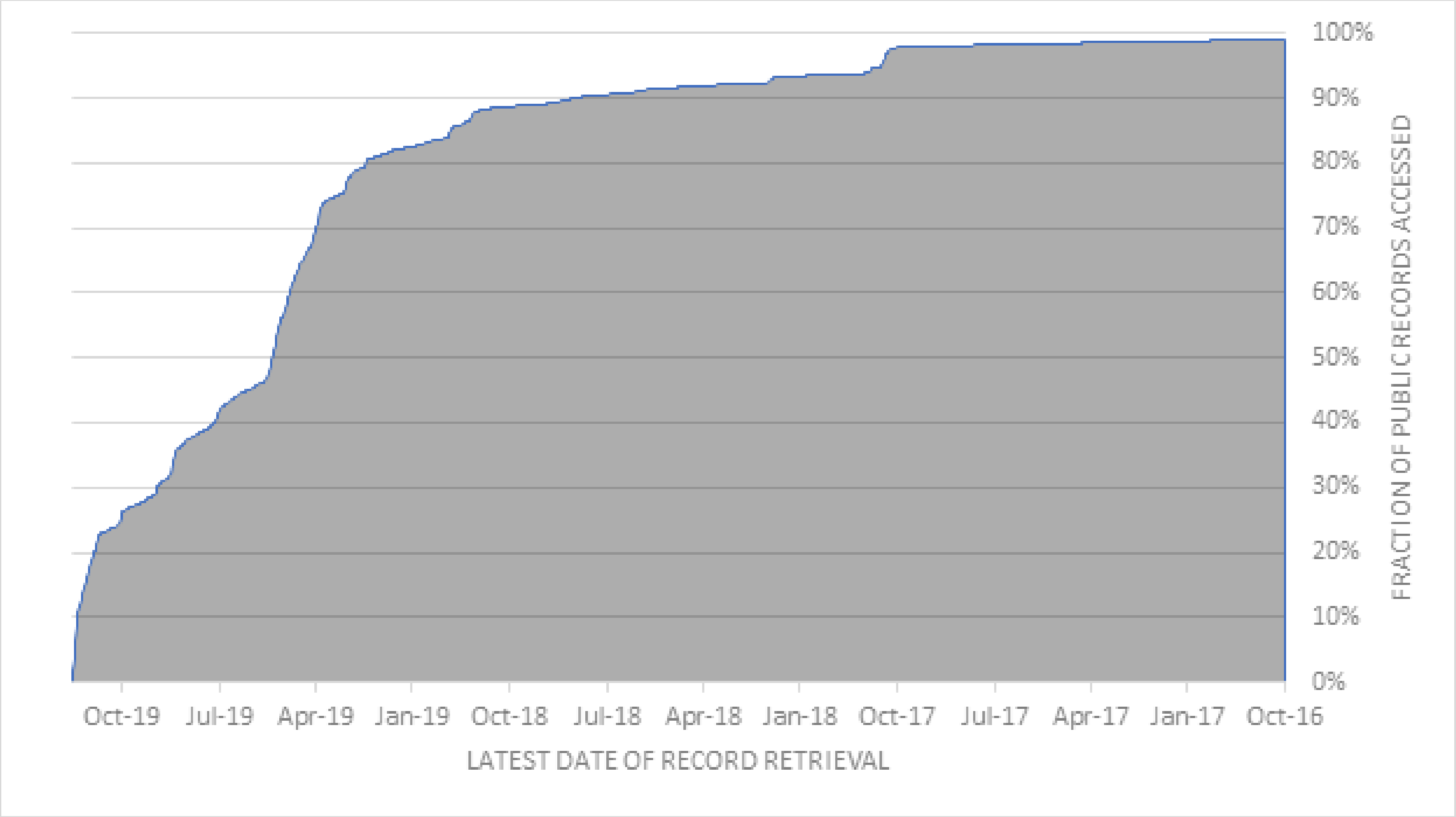
Recommendations

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

NCBI should monitor data usage and determine the appropriate cloud storage location for each dataset depending on usage data. Users should be informed as they are retrieving data where the data are stored and costs associated with retrieving it.

NCBI should provide limits on the amount of data users can request to be thawed without approval (i.e., provide a “circuit-breaker”) to prevent accidental overuse of NIH resources. The computing limits should be defined by a sliding time interval window to allow users to use up to their compute limits in varying timeframes fit to their research needs

SRA Data Access: 2016 – 2019



Cumulative distribution of waiting times until first request for SRA access indicates that 50% of the unique data records were accessed between May and October 2019.

Recommendations

Communication of the model

Communication materials should be developed around non cloud-based sources for data. If NIH decides to make BQS formats only available in the cloud, equity challenges should be addressed as part of the planning process.

Cost models, including how to estimate charges and determining who pays, should be clear and provided to the research community via ODSS and NCBI websites and other public-facing communication mechanisms (e.g., the NIH Guide).

Information provided should include specifics on costs for both storage and compute: What is the user paying? What is NIH paying?

Education via online tutorials or courses must be provided for users/potential users to understand when to use the cloud, how to access data in cold or hot storage, and how to monitor compute time.

Recommendations

Continued research to inform changes to the model over time

NIH should monitor costs of the current model over time to make adjustments based on the actual costs of researchers working in the cloud. Determine if different strategies are needed for different cloud service providers and continue to solicit feedback from the SRA user community

NIH should monitor cost and use and adjust policies accordingly to ensure that no subset of researchers are bearing an unfair burden (based on data format).

NIH should consider intramural and extramural support for efforts that explore the effects of various compression strategies, efforts to optimize code and efficiency in the cloud to reduce compute costs. NIH should engage academic, industry, and fee-for-service communities to address software optimization challenges

Questions & Discussion

Sequence Read Archive (SRA) Data Working Group

Recharge the working group of the Council of Councils

September 11, 2020

Council of Councils

NIH National Institutes of Health
Office of Data Science Strategy

SRA Data Working Group Charge

The charge of the SRA Data Working Group of the Council of Councils is to provide recommendations to the Council regarding evaluation of SRA data storage, management, and access in cloud service provider environments. The working group will focus on evaluation of SRA as a resource and other related issues, including but not limited to:

- 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.

2021 Priorities

The SRA Data Working Group will examine data related to SRA scientific impact, value to the community, access, cost, and usage, as well as other areas, to inform their considerations and evaluations.