# Sequence Read Archive Data Working Group Interim 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**

*January 24, 2020*

NIH National Institutes of Health
*Office of Data Science Strategy*

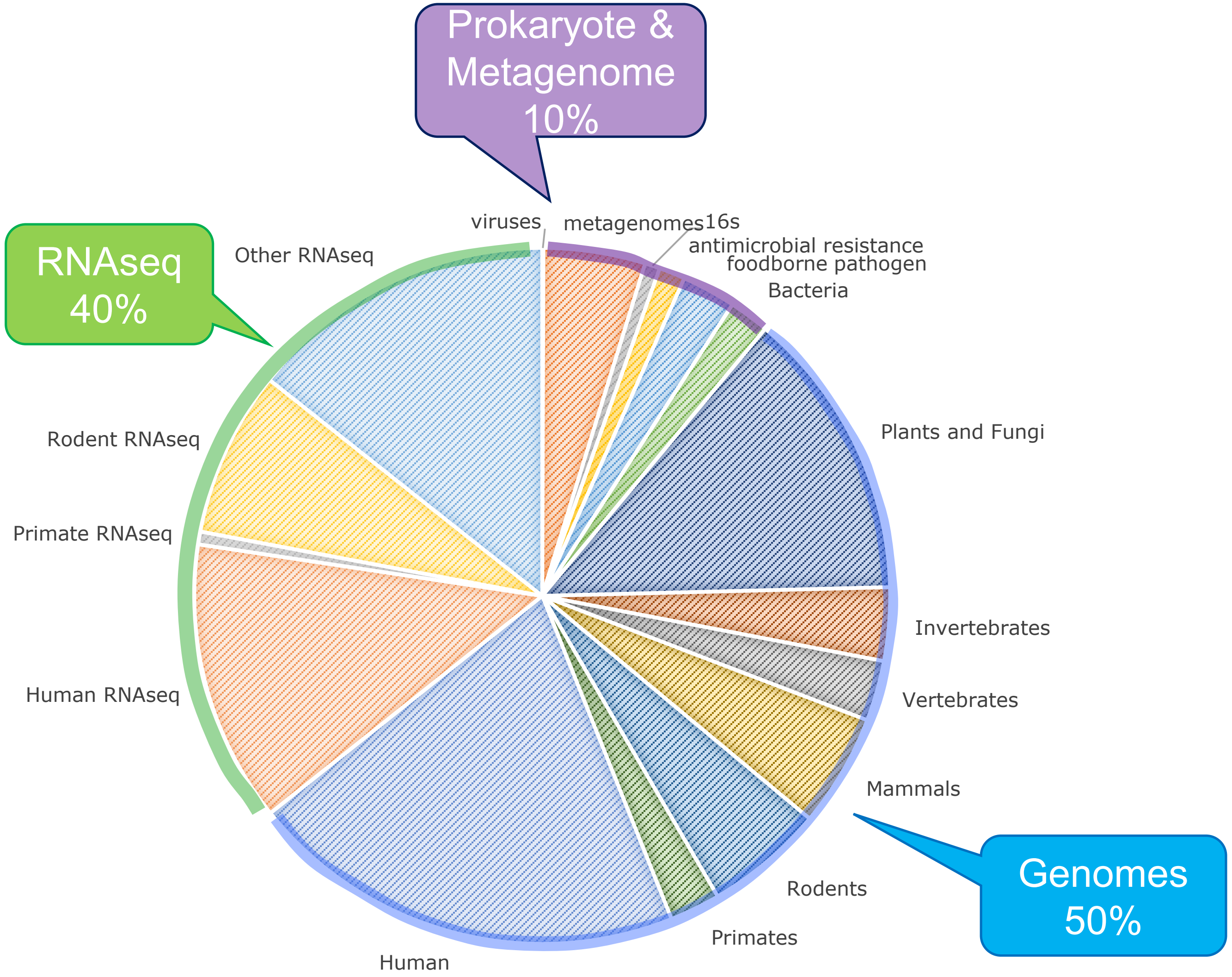
# Agenda

- Background and Challenge
- Working Group Charge
- Principles and Proposals
- Recommendations
- Future Considerations

# Agenda

- Background and Challenge
- Working Group Charge
- Principles and Proposals
- Recommendations
- Future Considerations

# Public SRA Data – 6 PB



# All SRA Data Under Discussion – 24 PB

| | Controlled Access | Public Access |
|---|---|---|
| Google | 6 PB | 6 PB |
| AWS | 6 PB | 6 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 9 million records, 12 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

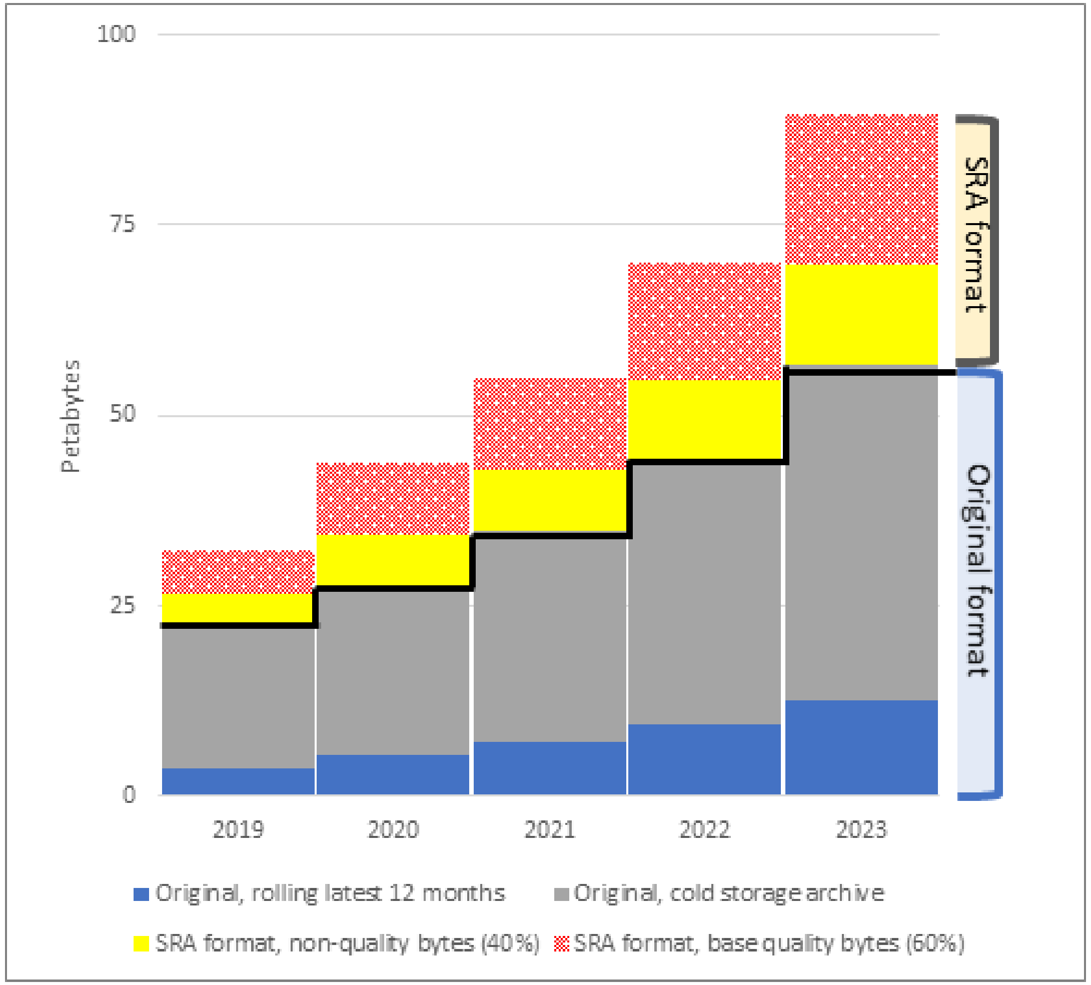
| *Format* | *July 2019* | *2020* | *2021* | *2022* | *2023* |
|---|---|---|---|---|---|
| **Original** | 23 | 28 | 35 | 44 | 57 |
| **SRA Normalized (ETL)** | 10 | 16 | 20 | 26 | 33 |

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.

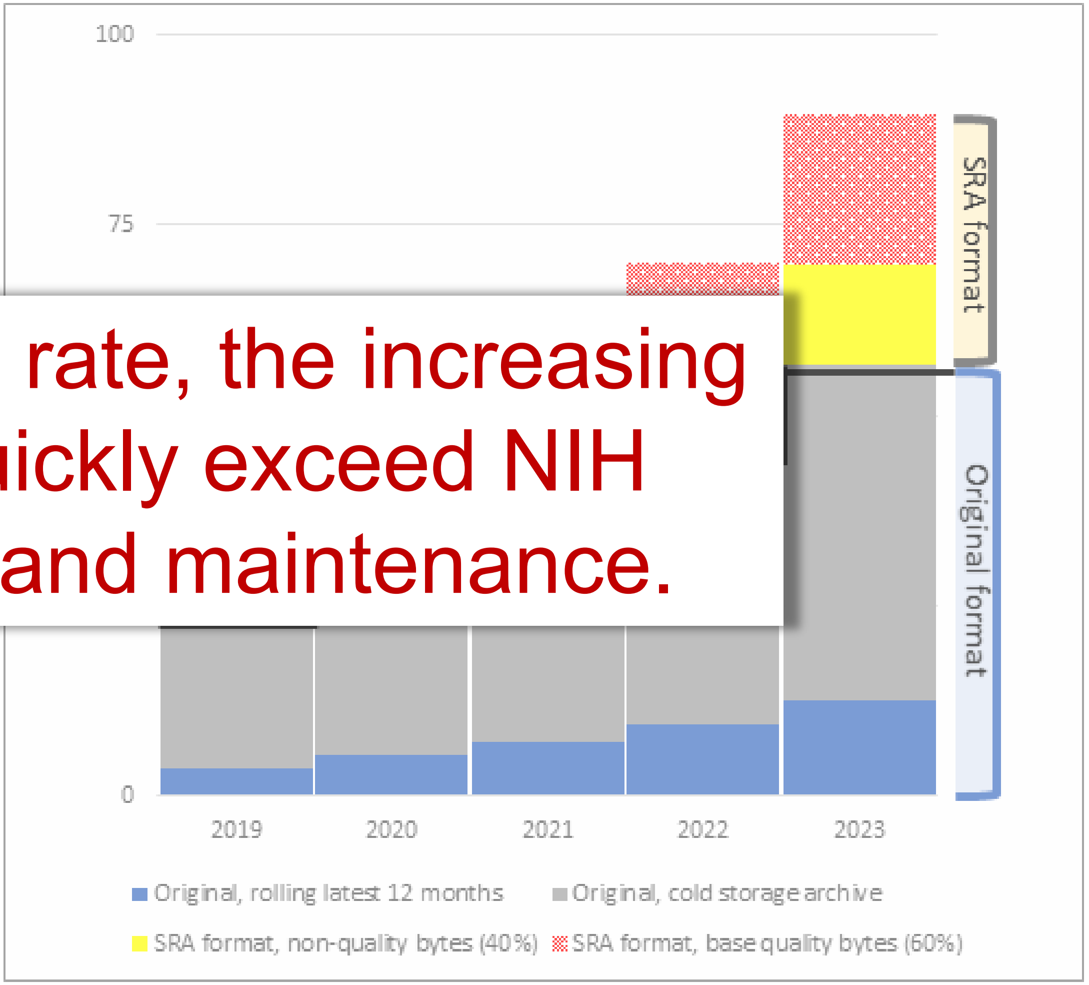
# SRA Projected Growth

- **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 (p
  - everyth
- **BQS**: Bas representa error at a
  - 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
- Principles and Proposals
- Recommendations
- Future Considerations

# SRA Data Working Group

## Co-chairs

**Susan Gregurick, Ph.D.**
National Institutes
of Health

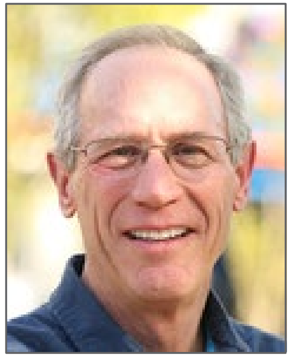
**Kevin B. Johnson, M.D.**
Vanderbilt University
Medical Center

## 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

***Initial draft report requested by the January 2020 Council of Councils meeting***

# Examples of How WG Labs Use SRA

### Kang Lab

- Aligned sequence reads from SRA to test software for specific data types with relatively new technologies (e.g. single cell RNA-seq, single-cell ATAC-seq, long read sequencing) to reproduce what was reported in the paper and improve upon it.

### Edwards Lab

- Developed tools to automatically discriminate between and annotate environmental microbial data from SRA (microbiome and metagenomics).
  Analysis updated monthly.
- Built a website where users can upload DNA or protein sequences for comparison to those datasets.

### Zhang Lab

- Used both restricted access and publicly available data to perform integrative genomic analyses and compare genomic, transcriptomic, and epigenetic profiles between cancer and matched non-cancerous cells or tissues.
- Developed and validated new bioinformatics tools and used SRA data to demonstrate the broad applicability of these methods across sample types.

# Agenda

- Background and Challenge
- Working Group Charge
- Principles and Proposals
- Recommendations
- Future Considerations

# Principles Considered

| | |
|---|---|
| **Continuous access to training datasets** | Some datasets are valuable public resources that are used as references for training, and these should be kept readily available in hot storage for immediate use. |
| **Quality of data available for analysis** | Changing the model for storage of the two formats of SRA data would not result in sacrifices to the data quality, as long as the full BQS data are preserved in a backup location. |
| **Prioritizing availability of frequently accessed datasets** | To minimize constraints on researchers, data should be transferred to the cloud and between hot and cold storage strategically (e.g., reverse chronologically). |
| **NIH costs** | NIH pays to provision and store SRA data in the cloud and would also pay to thaw data from cold to hot storage. Storage costs are increasing and are the focus of this Working Group. |
| **User costs** | Users pay for compute instances in the cloud and egress fees if they choose to download data from the cloud. These costs may create financial burdens for some users. |
| **Access to normalized and original formats** | Both forms have value and should remain available. |
| **Search and random access across entire SRA** | These are not tasks that are currently available to SRA users, so changing the data storage model would not result in sacrifices in this area. |
| **Barriers to access** | Want to avoid creating barriers to access to SRA for under-resourced institutions or researchers not currently in the cloud. |
| **Speed/wait time to access** | The projected 24- to 48-hour waiting period to access original or normalized data from cold storage would be acceptable, but should be tested. |
| **Flexibility and adaptability** | Recommendations should inform storage and access models that are flexible and adaptable and can change over time based on additional data. |

# Proposals Considered

| | **Proposal 1:** Eliminate BQS in normalized data in the cloud; keep some original data in hot storage | **Proposal 2:** Retain BQS in normalized data in the cloud; no original data available in the cloud | **Proposal 3:** Two versions of normalized data (with and without BQS) in the cloud; all original data in cold storage |
|---|---|---|---|
| Normalized data with BQS | Not available | Hot storage | Split between hot and cold storage |
| Normalized data without BQS | Hot storage | Not generated | Hot storage |
| Original format data (with BQS) | Split between hot and cold storage | Backup tape only | Cold storage |

# Proposals Considered

| Principles* | **Proposal 1:** Eliminate BQS in normalized data in the cloud; keep some original data in hot storage | **Proposal 2:** Retain BQS in normalized data in the cloud; no original data available in the cloud | **Proposal 3:** Two versions of normalized data (with and without BQS) in the cloud; all original data in cold storage |
|---|---|---|---|
| **Access to training data** | Improved | Unchanged | Unchanged |
| **NIH costs** | Now responsible for thaw charges; need models to determine hot/cold split | Storage costs may be unsustainable long-term | Now responsible for thaw charges; need models to determine hot/cold split |
| **Access to normalized and original formats** | Change to normalized format prevents some workflows | No access to original format data in the cloud | Both formats now available in the cloud |
| **Speed/wait time to access** | Must wait 24 – 48 hours for original data thaw | Unchanged | Must wait 24 – 48 hours for original and some normalized data thaw |

*All other principles are addressed by all three proposals.

# Agenda

- Background and Challenge
- Working Group Charge
- Principles and Proposals
- Recommendations
- Future Considerations

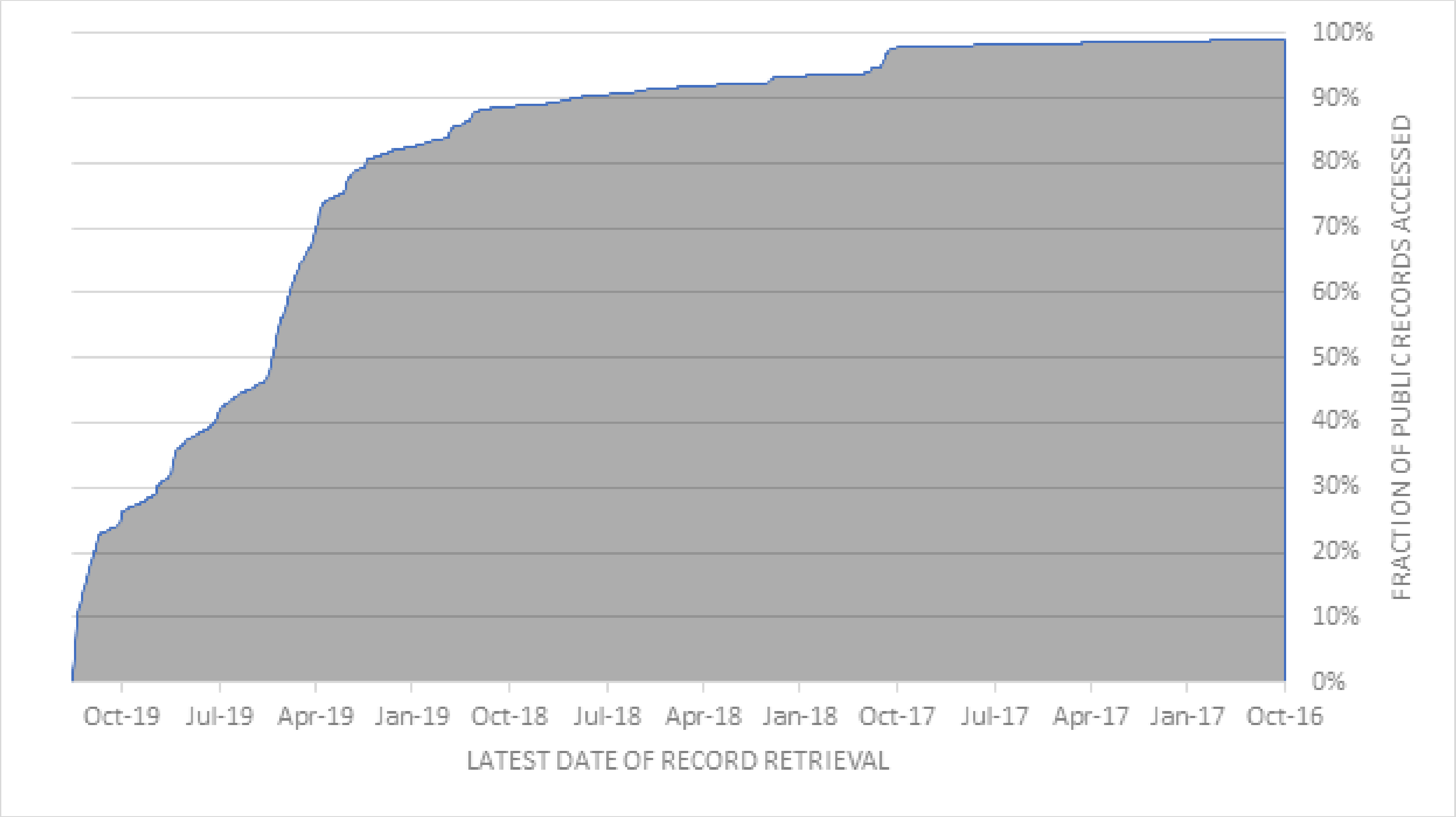
# Recommendations

## 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.
- Normalized data with BQS will be stored in hot or cold storage depending on usage; original formatted data (with BQS) will be retained in cold storage.
- NCBI would provide a “circuit-breaker” to limit the amount of data thawing users can request.

# SRA Data Access: 2016 – Present



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

Cost models 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 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

Since understanding of true costs is currently incomplete, the Working Group recommends that NIH monitor costs over time to adjust the model based on actual costs of people working in the cloud.

- Determine if different strategies are needed for different clouds, including what other strategies may be needed if additional clouds are added.

NIH should consider funding efficiency optimization research for use in the cloud to reduce cost for computing.

- The goal of these projects should be to design tools that ultimately can limit the need to egress data and incur costs.

# Agenda

- Background and Challenge
- Working Group Charge
- Principles and Proposals
- Recommendations
- Future Considerations

# Future Considerations

| Remaining question | Additional data/discussion sought |
|---|---|
| Which data should be immediately available, and which might be transferred to cold storage? | Studies of repeated access and age of accessed data sets. |
| Can the storage footprint of BQS in the cloud be reduced by compression or elimination in the long term? | A longitudinal study of the effects of binning (BQS consolidation) or eliminating BQS from normalized data on research workflows, algorithms, and analytical pipelines. |
| Can different data storage models be used for different cloud service providers, and are additional cloud service providers necessary? | RFI to understand how researchers are using, or anticipate using, SRA in the cloud. |
| How might NIH mitigate the potential risk in a cost model that relies heavily on cloud service providers? | Discussion of other models, including consideration of a *public research cloud*. |
| Can costs for cold storage and thawing of data be further optimized? | Information about how these costs might be negotiated with cloud service providers. |
| How can SRA data be integrated with data stored in other agencies' cloud platforms for analysis? | Exploration of use cases involving integration of SRA with other data platforms. |

# Timeline and Activities

| Task | Objective | Timeline | Notes |
|---|---|---|---|
| Finalize SRA working group interim report | Communicate findings and recommendations to community | By January 2020 | This report is an interim report for immediate efficiencies in SRA data storage needs. |
| Develop appropriate data collection methods for SRA in cloud | Develop a longer term SRA storage lifecycle recommendation | By March 2020 | Require information on repeated data access request from SRA over longer periods of time. Also need longitudinal study of changes to normalized SRA data formats (binning or eliminating BQS) on research workflows. |
| Compile analysis into final recommendations | Finalize SRA data lifecycle recommendations, including SRA format and storage | By Summer or Fall 2020 | These final recommendations could also provide guidelines for NIH cloud-based data storage and management principles. |

# Key Terms

| | |
|---|---|
| **Base quality scores (BQS)** | Quantitative representations of the probability of an error at a base; most file types have one BQS per letter of sequence. |
| **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. |
| **Cold storage** | A form of cloud storage in which data must be “thawed” before becoming available to users; this is generally less expensive than hot storage. |
| **Hot storage** | A form of cloud storage in which data are immediately available to users. |
| **Thaw** | The process of transferring data from cold to hot storage in the cloud. |
| **Binning** | An option for compression of BQS by combining similar scores into a smaller number of “bins.” |
| **Amazon Web Services (AWS)** | One of the two cloud service providers currently hosting SRA data through the STRIDES Initiative. |
| **Google Cloud Provider (GCP)** | One of the two cloud service providers currently hosting SRA data through the STRIDES Initiative. |