Council of Councils Working Group on Sequence Read Archive Data

Final Report -

September 2021

Table of Contents

Executive Summary *3*
Summary statement of SRA, size, and move to cloud .... **Error! Bookmark not defined.**
Summary statement of first WG findings .... **Error! Bookmark not defined.**
Summary statement on current WG and WG charge .... **Error! Bookmark not defined.**
Summary statement on WG Recommendations .... **Error! Bookmark not defined.**
Background and Challenge *5*
Status of SRA cloud storage and access *6*
Growth of SRA and Working Group Charge **7**
Projected growth of SRA in the context of the hybrid storage model **8**
Integration of SRA data into AWS Open Data Program **10**
Impact of hot/cold distribution on cost model projections and user access **11**
Communicating the hybrid SRA cloud storage model to the scientific community **13**
Key Performance Principles for maintaining SRA storage efficiency **14**
Relative cloud cost with AWS ODP **16**
WG Recommendations *18*
Promote cloud usage and ensure SRA data usage with equity and sustainability **18**
Explore data usage, access frequency and tolerance for cloud data retrieval in cost model **19**
Consider incentives for researchers using SRA to develop tools/algorithm for cloud computing **19**
Evaluate impact from SRA **20**
Future Work and Considerations *20*
Conclusions *23*
References *23*
Appendix 1: Organismal content and usage of public SRA *25*
Appendix 2: Assay type content of SRA *25*
Appendix 3: Glossary of Terms Used in this Report *26*

Executive Summary

The National Center for Biotechnology Information (NCBI), part of the National Library of Medicine, hosts one of NIH's largest and most diverse datasets, the Sequence Read Archive (SRA). The SRA is a broad collection of experimental DNA and RNA sequences that represent genome diversity across the tree of life. As of March 2021, the SRA contains 16.5 petabytes of normalized data and is continually growing. To advance the research communities use of SRA, the resource was replicated on Google Cloud Platform (GCP) and Amazon Web Service (AWS) cloud services in 2019 as part of the NIH Science and Technology Research Infrastructure for Discovery, Experimentation, and Sustainability (STRIDES) Initiative.

In March 2021, the SRA archive held over 14.5 million records in two formats. The original format (30 petybtes of user-accessible data) is received by NCBI from submitters and is instrument- and experiment-specific; these data were originally stored to tape. However, this made it difficult for researchers to access and use SRA, and thus these data are now being stored in the cloud. NCBI transforms these original format data into standard SRA normalized format (16.5 petabytes) for redistribution. The normalized format contains base quality scores (BQS) that provide information about the quality of each base in the sequence; however, because of the number of possible BQS for each base, these drastically increase file size, thus making BQS the largest cost driver for SRA storage in the cloud.

The NIH has charged the SRA Data Working Group of the NIH Council of Councils to provide input 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. The SRA continues to experience exponential growth in submission rates, and the normalized format data is projected to grow to 57.5 petabytes by 2025; at this rate, the increasing size will quickly exceed NIH budget for storage and maintenance of this valuable resource. The SRA Working Group's charge is to provide recommendations to the NIH on several key factors for managing this data resource in cloud service provider environments. The NIH is requesting that the working group analyze and evaluate strategies for, or changes to, SRA data storage, manangement, and access, including its impact on the research community. The working group was also charged with providing input on future needs and opportunities, including sustaining SRA as a community resource.

In deliberations leading up to their recommendations, the working group studied data on SRA growth, cost models for storage over time, frequency and breadth of data usage and access, tolerance for cloud data retrieval in cost models, and projections of future growth and use. In the data models and discussions, they considered how to promote cloud usage, ensure SRA data usage with equity and sustainability, and how to encourage and support developing tools/algorithms for cloud computing. The working group also summarized two categorical Key Performance Principles to offer a framework for evaluating current progress towards distributing SRA to cloud storage providers.

The recommendations in this report are intended to address immediate concerns about the SRA storage footprint, while the “Future Work and Considerations” section highlights the group's next steps in considering longer-term solutions for future efficiency of, and access to, this large data resource in the cloud. The final recommendations are:

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

- Consider a more holistic and global view of commercial vs. open access clouds that would provide cost effective strategies for data deposition and use.
- 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 Google Cloud Platform (GCP) or Amazon Web Services (AWS) platforms, including those from under-resourced institutions

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. Many computing costs are related to the data type.

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.
- Advance petabyte scale sequence searching tools and focus on exemplars like SARS-CoV-2 and metagenomic data, and associated metadata

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

The future work and considerations are:

- 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

Background and Challenge

The Sequence Read Archive (SRA) is a diverse collection of experimental DNA and RNA sequences derived from across the tree of life (see Appendices 1 and 2). The archive is hosted by the National Center for Biotechnology (NCBI) at the National Library of Medicine and provides a repository where submitters can share their data with attribution to a growing reservoir of deposited data that can be mined for historic trends and novel biomedical and other research discoveries. As such, SRA represents a dynamic dataset where the efforts of each data creator add over time to the increasing value of the corpus for subsequent analysis, supporting novel insights into genome architecture, genetic and epigenetic variation, gene expression patterns, discovery of novel organisms and viruses, and association of genetic signatures from complex microbiome and metagenomic samples with environmental attributes [1-7].

SRA holds both public data available to all researchers and controlled access data derived from human research studies that is available to qualified biomedical investigators who agree in advance to use the data appropriately. Sequence data in the archive is linked to sample metadata from the SRA records and associated BioSample and BioProject records that together provide information about the sequenced sample and can be used to associate genetic information in SRA with phenotypic, clinical, and environmental attributes [8]. Given its vast size and diversity, the SRA represents a crucial resource for the biomedical community and other scientific communities, and researchers have developed bioinformatic tools and methods that support deep analysis of SRA data – everything from assembly and annotation of new genomes, characterization of human pathogens, identification of processes that drive genome evolution, to prediction of the functional significance of rare human genetic variation [9-18].

As a result of the growth of the SRA sequence database, and its expanded use in answering research questions that require petabyte datasets, NIH partnered with Google Cloud Platform (GCP) and Amazon Web Services (AWS) in 2019, through the Science and Technology Research Infrastructure for Discovery, Experimentation, and Sustainability (STRIDES) Initiative, to move SRA to a cloud-based ecosystem. As of March 2021, all 14.5 million data records are available from each of the two cloud providers both in the original, instrument-specific format and the normalized SRA format totaling 16.5 petabytes of data. SRA is the largest dataset thus far to be included in the NIH effort to build a cloud-enabled biomedical dataverse, with new processes, tools, and architecture to drive development of an equitable ecosystem that makes NIH-funded data findable, accessible, interoperable, and reusable (FAIR) [19].

Moving SRA to commercial cloud service providers creates potential advantages for researchers, but also presents several considerations regarding the accessibility and sustainability of the resource. In 2019, the NIH engaged a Council of Councils SRA Data Working Group to provide guidance for managing the archive in cloud service provider environments. Specifically, NIH requested that the working group identify and evaluate compression strategies to maintain efficiency in the storage footprint of SRA. This first Working Group presented a set of recommendations including deployment of a hybrid model that included the use of “hot” and “cold” storage tiers and the incorporation of SRA data files without base quality scores. This working group also recommended that NIH continue to monitor costs and feedback from the biomedical and other scientific research communities and continue to research potential changes to the model over time. In 2020, SRA Data Working Group of the NIH Council of Councils was empaneled and recharged as part of this ongoing evaluation with a focus on SRA sustainability and how storage strategies impact the biomedical and other scientific research communities as well as costs to NIH.

Status of SRA cloud storage and access

The entire corpus of SRA data records is stored locally by NCBI and replicated on both the GCP and AWS cloud platforms. In accordance with the hybrid storage model, three different file types are maintained for each data record with two cloud vendors using same storage model:

(1) An **original** format as submitted to NCBI from data providers that reflects the results of a submitter's particular instrument output and analysis computational pipeline (e.g. BAM, FASTQ). These files had previously been stored on tape at NCBI, and this is the first time that this format is readily accessible, and it is now only available through the cloud providers.
(2) A **normalized** version (termed extract, transform, load or ETL) of the data is created for all submitted data. This format supports interoperability across data originally provided in multiple formats and more compact storage.
(3) A **normalized format without base quality scores** has also been introduced (termed ETL-BQS) wherein base quality scores (BQS) are entirely removed or replaced with binary read quality scores to further reduce the storage footprint.

Recommendations from the 2019 SRA Working Group included deliberations and studies of SRA growth, cost models for storage over time, frequency and breadth of data access and projections of future growth and use. This working group considered the size and value of BQS and the availability of both "hot" and "cold" storage in the cloud. "Hot" storage provides immediate access to data and is the default standard. Data in "cold" storage is not immediately accessible but can be stored at a reduced cost with a charge to "thaw" (move) data from cold to hot storage. The working group also used 10 principles to guide its recommendations: continuous access to training datasets, quality of data available for analysis, prioritizing availability of frequently accessed datasets, NIH costs, user costs, barriers to access, speed/wait time to access, access to normalized and original formats, search and random access, and flexibility and adaptability.

Based on these recommendations, the three SRA data formats are distributed among two storage tiers in each cloud provider, a hot tier that provides immediate access to data and a less expensive cold tier where data may not be immediately accessible. The hybrid data model posits that both original format data and the less used 50% of ETL data are stored in cold storage tiers on commercial cloud platforms. This is the current structure of SRA on the two cloud service providers, and the details of data size footprints across formats and tiers will be discussed in further detail below (Table 1).

To enable researchers to utilize SRA on the cloud providers, SRA data can be searched through the NCBI web resources as well as tabular query interfaces associated with AWS and GCP and retrieved from hot and cold storage through the web-based Cloud Data Delivery Service (CDDS) [20] and programmatically through the SRA toolkit for hot storage. Data can be copied from AWS and GCP hot storage locations using cloud provider CLI tools and standard command line tools like *wget* after retrieving the data location from NCBI's data location service or from NCBI web resources. Additionally, researchers may directly utilize data for cloud-based analyses, removing the need to make an additional physical copy either locally or in a cloud environment and reducing overall costs. During 2020, more than 48 petabytes of data were downloaded from SRA, and greater than 10% of this was accessed from cloud platforms.

Growth of SRA and Working Group Charge

SRA continues to grow exponentially and included nearly 17 petabytes of public and controlled access sequence data as of March 2020 (see Figure 1). While this growth correlates with increased value of the resource to the research community, it also has a profound impact on the financial sustainability of the resource going forward. The current SRA Working Group was charged with providing recommendations to the Council of Councils on both the cost and the impact of distributing this resource to cloud service provider environments. The SRA Working Group charge is:

(1) Analyze and evaluate strategies for, or changes to, SRA data storage, management, and access, including impact for the biomedical research community.
(2) Make recommendations on data retention, data models, and data usage that will keep costs to the NIH within sustainable levels while maintaining community access to this large public data resource.
(3) Provide a vision for future needs or opportunities, including sustaining the SRA as a community resource.

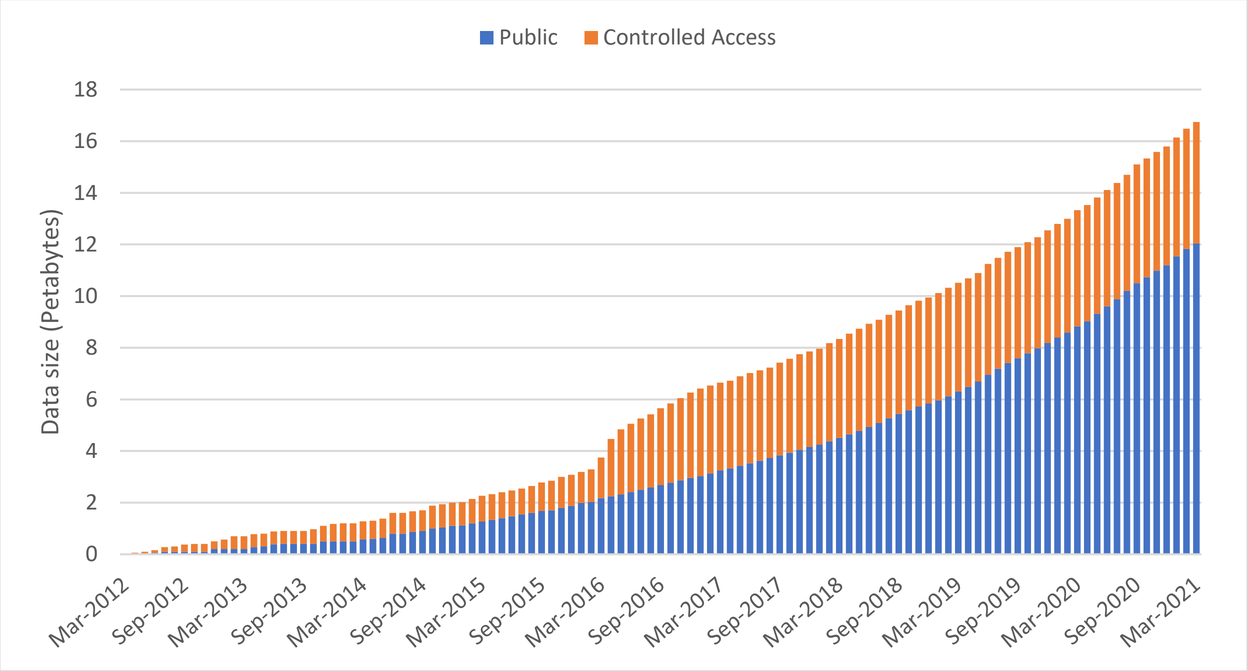

***Figure 1**. Historic growth of SRA by consent type, public and controlled access. Size of a single copy of SRA is plotted for both public and controlled access data in ETL format.*

Projected growth of SRA in the context of the hybrid storage model

The hybrid SRA storage model maintains three data formats for each submitted data record – original format, normalized ETL, and normalized ETL without base quality scores (ETL-BQS) – and projected growth in the number and size of submitted data records seen in Figure 1 is amplified across the three data formats, profoundly impacting the total footprint of the archive, a single copy of which is projected to grow by 50 PB over the next 4 years (Figure 2). The hybrid model defrays the increased cost of maintaining multiple formats for each data record by utilizing a lower cost cold storage tier for original format and less used, standardized ETL formatted data. The model posits that only more highly used ETL data and all ETL data without base quality scores (ETL-BQS) is maintained in the more expensive hot storage.

Replicating this model and associated storage footprint across multiple cloud providers increases accessibility, convenience, and cost efficiency to the research community while also potentially mitigating price monopoly concerns associated with a single cloud service provider. However, these benefits to researchers come with a significant increase in SRA operational costs, and requires careful assessment of the two tenants of the model, use of ETL-BQS format and cold storage tiers.

Base quality scores quantify the error probability for each base and represent 60-70% of the storage footprint in original format SRA data and on average 60% of the footprint of ETL formatted data. Because there are 63-94 possible BQS values for each nucleotide base, they are difficult to compress. Although many studies have investigated methods for BQS compression [21-25], no clear consensus approach has emerged within the research community, and different strategies may be preferred for different SRA data use cases. NCBI has pursued two approaches to reducing the footprint of BQS, first replacing them with single binary quality score for each read, and second, by dropping them entirely after aligning reads to a reference in a storage object that also retains unaligned reads. These approaches greatly reduce the overall storage footprint compared to ETL formatted data with base quality scores (see Figure 2), and if accepted by the research community, these provide advantages to both NIH storage cost and scientific use of data at scale.

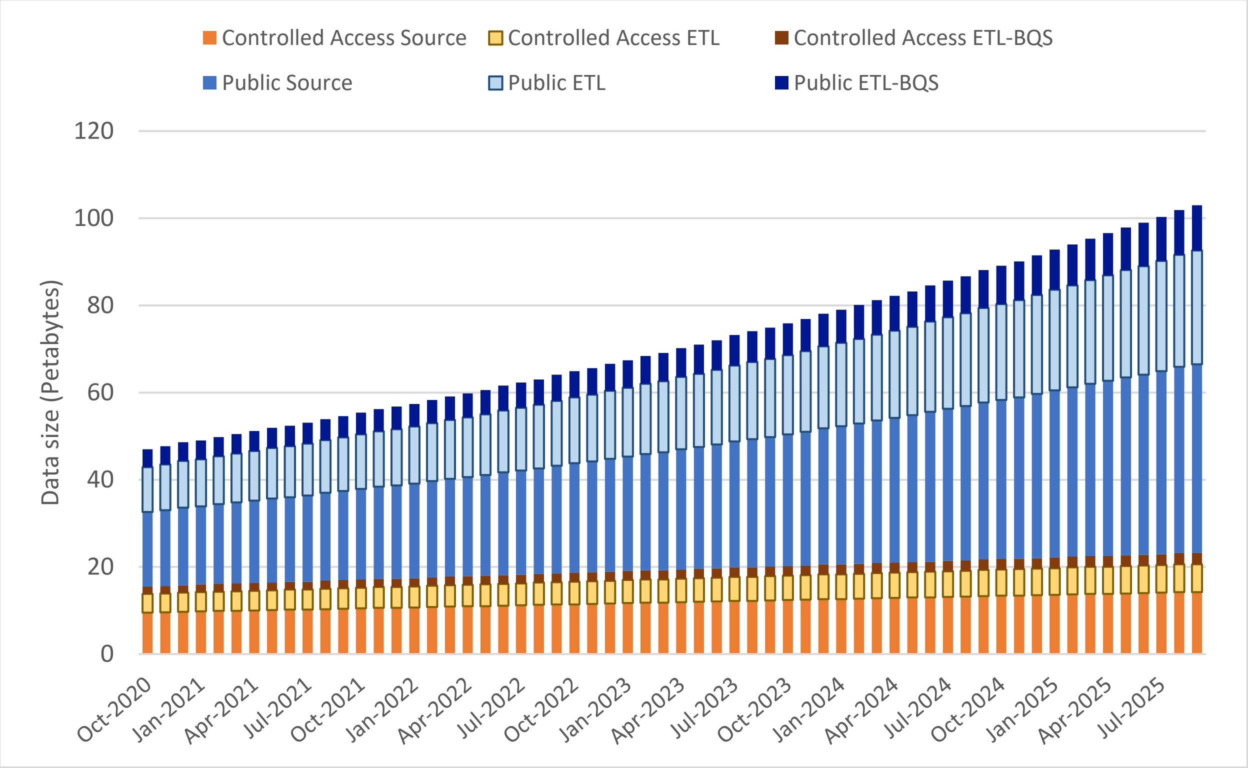

Figure 2*. Projected growth of a single copy of SRA by data format and consent type, public or controlled access. Projections were fitted to historical growth data using an exponential curve with smoothing. For simplicity, the 95% confidence intervals for the fit are omitted, but represent approximately a 25% increase or decrease in the sizes depicted in this figure.*

The largest portion of the overall SRA storage footprint is the original formatted data submitted to the archive (see Figure 2). While this format can contain additional data elements important to analysis or originate from newer sequencing technologies where the development of accurate base calling software is ongoing, for most use cases original format is functionally replaced with ETL. To make data available when needed while maintaining efficient storage costs, both original format and the less used portion of ETL data are maintained in lower cost cold storage tiers. This saves significantly on NIH costs but increases the “wait” time between data request and delivery. Delivery time is impacted by two factors, cold to hot restore time - which is dependent on cloud providers technology - and transfer speed from the NCBI hot bucket to the user bucket. Transfer speed is influenced by three factors, transfer speed within the cloud provider’s network, number of parallel machines that are used to deliver the data, and the size of files being transferred. Using logs from the NCBI CDDS, NIH investigated the time to deliver data from cold storage under real world conditions. Median wait times based on actual retrieval events through CDDS are 0.5 hours for data located in GCP cold storage and 14 hours for AWS cold data and indicates initial ‘thaw’ time. In addition, researchers may need to transform these data into other formats so that they can be used for analysis by other computational software, and this transformation will add additional time.

Integration of SRA data into AWS Open Data Program

One key development since 2020 is the expansion of the AWS Open Data Program (AWS ODP) within the STRIDES initiative. This program supports storage of data at no cost to NIH or users who can freely download data to their own cloud, local locations, or access the data for analysis in the cloud without copy. The initial STRIDES AWS ODP allocation of 2 PB to SRA was used to host datasets in original format requiring reprocessing of base calls due to frequent updates to base calling algorithms and pipelines. The new STRIDES agreement with AWS ODP supports an additional 12 PB of hot data storage which is being used to host public SRA records in the normalized ETL format.

Inclusion of additional storage space in the AWS ODP significantly impacts accessibility and essentially presents a best-case scenario wherein the scientific research community has no-cost, high-bandwidth access to normalized SRA data. The transfer of public SRA ETL data to the new AWS ODP cloud storage began in January 2021. During this time there has been a clear uptick in usage of AWS ODP hosted data (Figure 3), surpassing the usage observed from NCBI servers in March 2021. While this trend could reflect temporary activities, the increased usage came from both cloud and non-cloud users, underscoring the positive impact on all SRA users.

AWS ODP also provides the opportunity to create novel implementations for important datasets. For example, a specialized SARS-CoV-2 dataset is maintained in AWS ODP that includes original and ETL formatted SRA data, as well as ETL formatted data without base quality scores [26]. The ETL-BQS format implemented for SARS-CoV-2 ETL-BQS data removes base quality scores through alignment to reference genome sequences to create a multi-purpose data object that includes a consensus sequence as well as representation of read coverage, depth, and positional variation. Unaligned reads are retained, and both aligned and unaligned reads can be extracted from the object to use in analysis and processing pipelines that require them. To support rapid comparisons between SARS-CoV-2 data records, this dataset also includes gene and protein annotations as well as nucleotide and protein variant data stored in the Athena query service. NCBI is now providing the same SARS-CoV-2 SRA dataset through the GCP Public Datasets Program which also provides free user access to data at no cost to NIH.

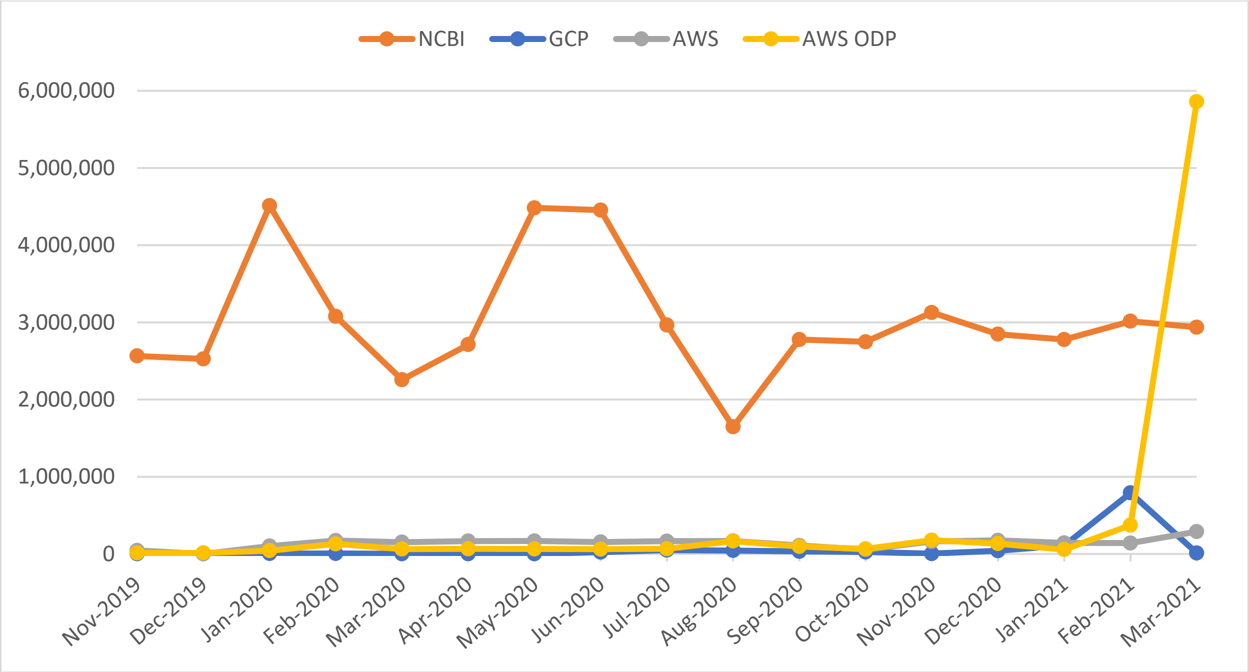

Figure 3*. SRA public data usage. Number of interactions with public SRA data records hosted by NCBI servers, GCP, commercial AWS, and AWS Open Data Program (AWS ODP) are plotted for each month since November 2019. Figures include only accessions accessed by external, non-NCBI users. During 2020 total SRA data usage from cloud sources represented 10% of total SRA retrievals. Transfer of public SRA in ETL format to AWS OPD was initiated in January 2021.*

Impact of hot/cold distribution on cost model projections and user access

The integration of AWS ODP augments the original hybrid SRA storage plan (see Table 1). While original formatted SRA data is distributed to cold storage tiers on both AWS and GCP, it is currently not necessary to store the less used 50% of ETL formatted data in the AWS cold tier. Instead, ETL formatted data is stored in AWS ODP hot tier, and only the amount of less used data necessary to sustain the 12 PB cap is moved to commercial AWS cold storage tier. Although the plan is to provide ETL-BQS data on both GCP and AWS commercial platforms, this format is currently only available on GCP.

	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

Table 1*. Current implementation of SRA hybrid storage model. *1 PB storage buffer remains in the original AWS ODP bucket to accommodate future source files requiring frequent reprocessing of base calls. **Distribution of data to AWS ODP is still in progress and when complete there will be no ETL data in the AWS hot commercial buckets and 12 PB of data in the AWS ODP bucket.*

Including AWS ODP in the SRA cloud storage allocation provides substantial cost savings to NIH compared to the original hybrid model (Figure 4). Nonetheless, projected storage costs to NIH will continue to rise even with this addition, increasing by 46% over current levels in FY 2025. Therefore, NIH has continued to investigate other cost mitigation strategies, including those that emphasize the use of cold storage tiers. NCBI analyzed overall usage for each data record as well as trends in time to next retrieval (i.e., the relationship between an initial request for a data record and subsequent requests for the same data record). Based on historic data usage, time to next retrieval was found to be essentially stochastic, so NCBI focused on optimizing storage tier distribution based on the frequency of data record use over the course of time.

Storage tier distribution models integrated both storage costs as well as retrieval costs associated with "thawing" data from cold storage. This latter cost is important to note as NIH is currently paying for cold storage retrieval costs with monthly caps placed on individual users. Based on this analysis, data records with more than 8 retrievals per year were more efficiently stored in hot storage tiers. Though more than 75% of data records are accessed annually, records with more than 8 retrievals per year comprise less than 10% of the total, implying that the data distribution between hot and cold storage tiers can be amended to improve NIH cost efficiency. The resultant hybrid model with 12 PB of ETL data allocated to the AWS ODP hot tier and only 10% of ETL allocated to commercial GCP hot tier reduces cost projection by 40% compared to the original hybrid model in FY2025 (Figure 4).

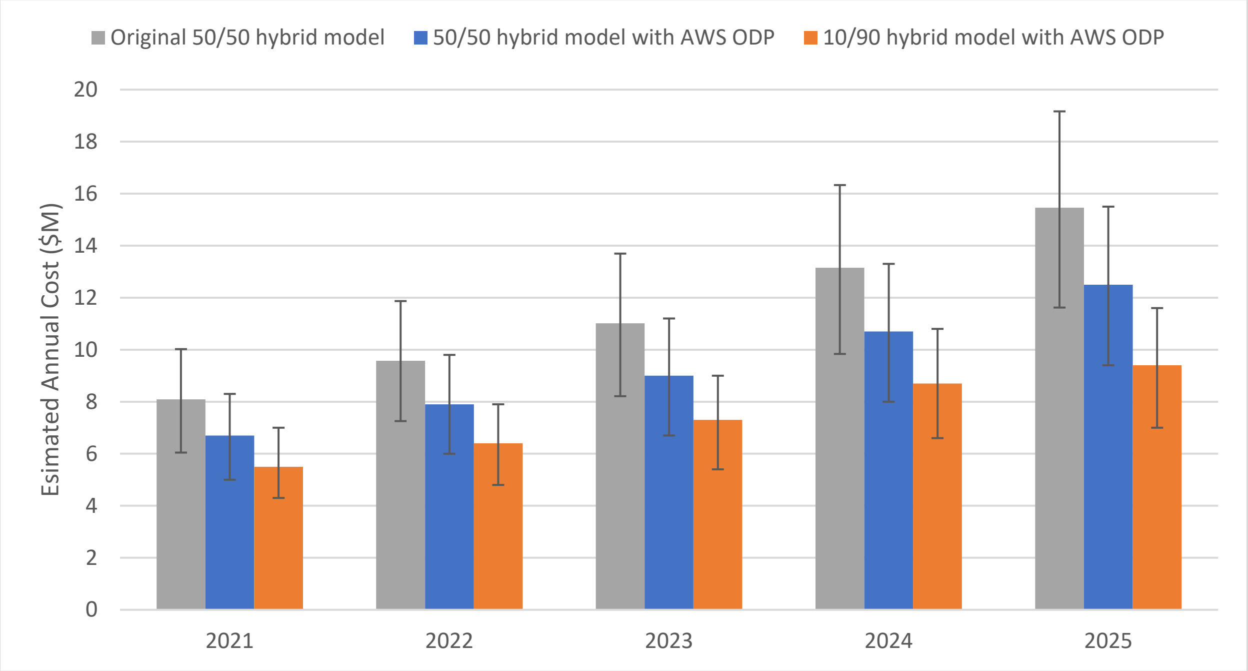

Figure 4*. Projected SRA cloud storage and retrieval costs to NIH. Projections include three different storage scenarios wherein original source, normalized ETL, and normalized ETL without base quality scores data is allocated to both GCP and AWS cloud platforms. The first scenario reflects the original 50/50 hybrid model wherein 50% of normalized ETL data is allocated to hot storage tiers, and 50% stored in cold storage tier. The second scenario (50/50 hybrid model with AWS ODP) includes integration of the AWS ODP into the original hybrid storage model. The third scenario (10/90 hybrid model with AWS ODP)*

distributes the less used 90% of ETL formatted data to the cold GCP storage tier. Storage cost projections are based on the growth of different data formats in Figure 2. In scenarios that include AWS ODP, the most used 12 PB of ETL formatted data is distributed to hot open data storage and the rest to commercial cold storage. Cold storage retrieval costs were calculated based on historic usage logs on NCBI servers and are included in the overall cost calculations presented in the graph. Error bars reflect the 95% confidence interval provided by the growth projection curves.

One potential issue with increasing the cold storage allocation is the impact on access caused by increased duration between data request and retrieval. These wait times vary between the two current STRIDES partner cloud providers, as do cold storage and thaw costs, but delays between data request and retrieval may be acceptable for typical use cases. This is particularly true when larger numbers of data records are requested, and any delays are offset by processing times required to deliver data from the commercial cloud provider bucket to the user cloud bucket or local machine. For example, the average retrieval time experienced for all CDDS data orders was 1.5 Gigabytes (GB) per minute, but the retrieval times for orders containing more than 100 GB of data averaged 3.8 GB per minute.

Communicating the hybrid SRA cloud storage model to the scientific community

The previous working group also recommended that NIH communicate the hybrid SRA cloud storage to scientific communities and evaluate updates to the model based on feedback. As part of these outreach efforts, NIH released a Request for Information (RFI), *Use of Cloud Resources and New File Formats for Sequence Read Archive Data* in July 2020 [27]. The key findings of this RFI are summarized below:

- 75 responses were received from 12 countries, and these were affiliated with academic (77.3%), government (9.3%), industry (6.7%), and nonprofit (6.7%) institutions.
- 65% of responders reported that they appreciated the NIH's efforts to maximize efficiency in retrieval through introduction of new data formats without BQS.
- Over 80% reported that BQS are critical for their analyses and requested that data be retained in the original FASTQ format.
- 17% reported that they were currently using SRA in cloud.
- 52% of responders expressed concerns about the unpredictable cost of using cloud computing platforms and felt that cost was more easily managed on their local high-performance computing system.

Responses from the RFI underscored two major themes; there is hesitancy among some SRA users about the move to the cloud storage, and there is continued need to communicate the data model – including information about SRA data formats, accessibility, and costs – to the user community.

Updates to the hybrid cloud storage model addresses some of the concerns expressed in the RFI. Moving ETL data to AWS ODP effectively deemphasizes the role of ETL-BQS data formats as part of the NIH cost mitigation strategy and extends the timeline for continued community outreach around these data formats. While the majority of RFI respondents appreciated the effort to deploy SRA formats without BQS to help reduce overall storage cost associated with the archive, clearly work remains in associating data formats with specific use cases and refining data formats to efficiently meet those needs.

Key Performance Principles for maintaining SRA storage efficiency

The previous NIH Council of Councils SRA Working Group emphasized the importance of aligning cloud storage models with community use cases and identified 10 principles to consider when proposing recommendations for maintaining efficiency in the SRA storage footprint. Based on those principles, the current NIH Council of Councils SRA Working Group summarized two categorical Key Performance Principles (Table 2) to offer a framework for evaluating current progress towards distributing SRA to cloud storage providers. The working group noted that metrics for the principles will need to be further developed and suggested a standing SRA Advisory Committee.

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	1, 2
	Distribution of SRA, BioSample, and BioProject metadata with sequence data d. Support access to sequence metadata e. Support data communities with specific metadata needs	5
	Development of tools to support search of data f. Metadata-based search g. Sequence-based search	6
	Improve value of data h. Provide analysis of data (e.g. SRA Sequence Taxonomic Analysis Tool STAT) i. Support metadata packages	4
B. Equitable user access	Distribution of data in hot and cold storage	8
	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 (e.g. CDDS)	3,7
	Replicate SRA among STRIDES cloud service providers	9
	User costs on data retrieval/egress to different cloud platforms	7,8

	Training and outreach for competency a. General b. Minority-focused	4
	Partnerships among US Government agencies	10

Table 2. *Key Performance Principles for evaluating success of SRA maintenance and use in the cloud.*

Key Performance Principles:

1. *Data are readily available in common formats (e.g., original format, ETL, ETL-BQS) to support both cloud and non-cloud users' need.*
2. *As defined by a time period, the most frequently accessed datasets are always readily available from multiple sources.*
3. *Open Data and public datasets are immediately or up to less than 24hours (dependent on data size) available.*
4. *"Training Datasets/Tools" are updated, maintained and always available in ALL locations with no associated costs.*
5. *No intentional discrepancies in ability to discover, access, compute upon, or analyze data, data to adhere to FAIR principles.*
6. *Sequence search and data analysis tools are available for supporting SRA and specific metadata sequence data, metadata to be digestible by third party tools.*
7. *User costs are well defined and a mechanism to ameliorate those costs for under resourced investigators is developed.*
8. *Optimize hot and cold storage distribution to save costs for both NIH and users.*
9. *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.*
10. *Collaboration with other US Government agencies to serve broader research communities by developing advanced tools and improving data repositories interoperability.*

One key positive development is the inclusion of original formatted data in cold storage. This is the first time this data has been readily available without special request and will allow access in cases of need and clear assessment of value against normalized ETL and ETL-BQS formats. The inclusion of ETL data formats in the AWS Open Data program has also greatly impacted progress within these principles. This development reduces NIH storage costs and obviates user retrieval costs. It also ensures easy and fast access to training datasets and other highly accessed SRA records. Integration of the AWS ODP storage platform and commercial cold cloud storage tiers with the web-based Cloud Data Delivery Service [20] provides a seamless approach to retrieval of public data regardless of cloud location and in doing so greatly lowers barriers to data access.

Integration of controlled access data into AWS ODP would improve accessibility to this data. That said, controlled access data presents a unique set of security requirements and access considerations including the need for special access keys or authorization schemes for cloud-hosted data and the distributed data storage model across multiple NIH Institutions. These will need to be addressed through direct collaboration with AWS ODP and the integration of NIH-wide authentication, authorization, and auditing systems like the Researcher Auth Service Initiative [28] and federated data access models like the NIH Cloud Platform Interoperability Effort [29].

Work remains in creating more cloud-native approaches to data search and analytics – both keyword and sequence string-based approaches – as well as approaches to identify higher and lesser "quality" datasets based on sequence and metadata metrics. Both activities are important to continued progress as they will directly impact the ability to locate data of interest and may also help to inform future iterations of the data model. While the hybrid model remains somewhat adaptable, it should be noted that it does not completely solve the problem of storing a growing resource within a constrained budget. This would be particularly true if new cloud providers were integrated into the STRIDES program and a complete copy of the SRA was replicated in the new platforms without commensurate increases in the budget. User compute costs will also continue to be a significant barrier to cloud data usage until the scientific community has sufficient knowledge, resources, and incentives to move current analysis pipelines from local environments to the cloud.

Relative cloud cost with AWS ODP

1. **NIH costs:** NIH currently bears many of the costs associated with SRA data storage and retrieval on multiple cloud platforms. Currently, NIH pays for the following:

  a. Costs of NCBI compute to provision data in AWS and GCP clouds.

  b. Costs for storage of SRA data on these platforms in two formats (original and normalized).

  Under models involving a split between hot and cold storage (with the goal of maintaining an affordable mix of hot and cold storage for active and infrequent SRA records), NIH would additionally be responsible for the cost of thawing any data from cold to hot storage. See Table 3 and documentation on user costs for estimated average costs for NIH and users for some typical workflows.

2. **User costs:** The working group discussed concerns about the financial burden on users who would have to pay egress costs to download their data from the cloud. Currently, NIH-funded or other federal agency funded investigators/institutions pay for the following:

  a. Costs of user compute instances in the GCP or AWS cloud.

  b. Egress fees if users wish to download the data from a cloud platform to their local computational environment or alternative cloud environment.

  c. Data storage costs (local or cloud-based) if the user decides to create a replica of the data.

  See Table 3 and documentation on user costs for estimated average costs for NIH and users for some typical workflows. It is noted that researchers directly accessing data in the cloud for downstream analysis without copying or downloading them achieve significant cost savings versus alternative models. However a relatively limited number of researchers currently have expertise to benefit from these savings.

For the purposes of Table 3, NCBI identified three common use cases for SRA data. The first two cases below use subsets of the data, while the third case (global search) may require access to the entire corpus of SRA data. These use cases are as follows:

1. Replication of a published study, which would require original format data with BQS;
2. Reanalysis of data for genetic variation detection, which requires normalized format data with BQS; and
3. Other discovery tasks, e.g. metagenomic profiling, expression analysis, and global sequence search, which can use normalized format data without BQS.

Role	Activity (relevant use case[s])	Data upload to cloud platforms	Storage Hot/Cold	Thaw	Egress	Compute
NCBI	SRA Submission Dataflow: estimates based on data from November 2019 only (all use cases)	500 TB/ month. Data ingress is free. NCBI pays SRA labor costs to process.				
NIH	SRA Archive (all use cases)		$150,000 / month (avg per cloud copy of SRA, ~10 PB)			
NIH **NCBI**	Develops and operates search algorithms (Use case 3)					Variable (cost)
NIH	Charge to thaw original format data (always cold) or normalized format data with BQS (mix of hot and cold) (Use cases 1 and 2)			$27/TB thawed (avg)		
User	(Optional) Egress data to different cloud platform (all use cases)				$100/TB (avg), No cost from AWS ODP	Variable (cost)
User	Compute on selected data in the cloud (all use cases)					Variable (cost)
Institution	Provide hardware platform to access the cloud (i.e., desktop and internet connection) (all use cases)					Substantially lower cost than hardware necessary for local compute
User	Download data for local compute (all use cases)				$100/TB (avg), No cost from AWS ODP	Costs per institutional policy
Institution	Provide hardware for local compute (all use cases)					Substantially higher cost than hardware necessary for cloud access

Table 3. Relative costs of activities for common use cases of SRA and their distribution across entities in pilot efforts.

WG Recommendations

Framed by the new charges, the current SRA Working Group presented a new set of recommendations after reviewing the Report from the previous Council of Councils Working Group on SRA and the Summary of Responses to the Request for Information (RFI) on use of cloud resources and new file formats for SRA Archive data, which was conducted by the NIH Office of Data Science Strategy (ODSS) in July 2020 [27]. The current SRA Working Group also studied updates to SRA status, as articulated in this report and in response to the prior working group recommendations. Considering all of this information, the SRA Working Group (WG) members agreed on the following new set of recommendations:

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

NIH's effort of implementing the hybrid storage model helps promote SRA's sustainability. However, this solution cannot keep pace with constrained budgets and the exponential growth in SRA data. NIH should provide cost effective strategies for the research community. For instance, NIH can communicate to stakeholders about storage and retrieval models and the open access program. Meanwhile, NIH needs to identify approaches that mitigate costs associated with cloud data storage without replicating SRA across more cloud providers. Different groups of users have experience and workflows on different cloud vendors. NIH should communicate and continue to engage different cloud vendors.

Working group members noted that cost-effective cloud computing is challenging for many researchers. NIH should consider providing normalized or pre-processed datasets to reduce researcher compute and usage cost; consider establishing best practices for cloud computing with estimated schedules and transparent costs. Researchers need to have more information to help select feasible workflows with the knowledge of the estimated analysis timeline and cloud computing cost. Additionally, continued education including guidance about inclusion of cloud computing costs in NIH or other federal agency funded awards is critical to support investigators at institutions without significant experience working in cloud environments.

NIH should increase outreach to minority serving institutions for targeted cloud-based training and consider lowering the cost of SRA data usage for example through the AWS Open Data Program. NIH should enhance its leverage of AWS ODP and cloud credits as a model to provide egress-free access and promote cloud usage of SRA in the research community. In addition, NIH should promote usage, training, and feedback by engaging researchers with different cloud usage experience, learn their comfort levels and tailor efforts to different user segments based on their familiarity with the cloud. For instance, providing easily accessible video tutorials to demonstrate common use cases and workflows, and providing user-friendly data access and analysis platforms to accelerate awareness, education and training for novice and intermediate cloud users. To improve effectiveness of training programs, NIH could consider focusing on training the-trainers (e.g., librarians, power users) who have a channel or mechanism for outreach within their institution or networks.

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

NIH needs to explore SRA data usage patterns, access frequency, and learn researchers' tolerance for cloud data retrieval. It needs to be evaluated whether amending storage models to include more data in cold storage will benefit NIH and save budget for funding more data storage. This should include impact of research delays in accessing cold-stored data and researchers' tolerance for data retrieval time. Further assessment of the cost model is also needed after implementing updates to the hybrid SRA storage model. For example, metrics need to be created to evaluate researcher data access frequencies, data access patterns and time cold data is held in hot storage. NIH should develop data-driven storage solutions, which involve defining the dynamics of SRA accession usage, identifying low-usage data, and moving low-usage data to cold storage. SRA WG members encourage the NIH NCBI team to develop robust APIs to SRA (e.g. RAS [28]) that can provide more information about user data access patterns. Partnering with data analysis ecosystems developed by various Institutes and Centers (for example the NHLBI BioData Catalyst, NHGRI Anvil, NCI Cloud Resources, and others) may not only expand the use of SRA data for less computationally savvy users, but also enable greater detail about the types of analyses and workflows that are being performed with SRA data.

Working group members noted that different research use cases require different SRA formats and NIH needs to understand the relationship between the SRA sequencing data formats and their compute cost. Many computing costs are highly related to the data formats. For instance, most mature analysis workflows require the original sequencing data in FASTQ format. The data conversation from compressed BAM format to FASTQ format is time consuming and expensive on the cloud. NIH should consider providing a normalized data format to be ready for downstream analysis, for example, offering gene reads counts data for RNA-Seq datasets. NIH should also consider proactive communication when data conversion is needed with research community.

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. Ideally this would involve leveraging and adapting the existing tools, such as updating sequencing assembly tools, genes, and mutation annotation tools. The working group stressed the need for, and importance of, advanced petabyte scale sequence searching tools. For improving platform agnostic cloud data usage, NIH should focus on exemplars like SARS-CoV-2 and metagenomic data, and associated metadata.

Evaluate impact from SRA

- Consider citations in publications or other citable objects
- Partner with analysis platforms, develop metrics to capture user statistics and surveys

The SRA working group members recommend that NIH develop and monitor more advanced metrics for assessing the impact of SRA for biomedical researchers and for the larger research and training community. The members suggested that NIH gather the following information to be used in the evaluation of the impact of SRA:

- 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.

Furture Work and Considerations

User-centered focus

Many of the design principles underlying our recommendations focused on the reality of the user community—a group largely unaware of how to use the cloud as a research platform. Moreover, this community is comprised of people from a variety of backgrounds, with a range of competencies, and from institutions with varying infrastructural agreements or procedures. This group cannot predict the impact of our decisions on this community. It will be incumbent upon the NIH (indeed the entire PHS) to understand and minimize the unintentional impact this change to cloud-based data and computation has on the research capabilities of the user community. We recommend periodic surveys focusing on a nationally representative sample of this community, collecting data from both researchers and research administrators to understand:

- Financial cost of use
- Educational gaps and exemplar strategies to overcome any local challenges
- Performance (time to complete projects related to STRIDES-accessed data and compute)
- Cost-shifting (new costs to the user or institution that were previously borne by the NIH or other facilities.)

These data should inform iterative improvements in the resources available to researchers/research institutions/the NCBI/NIH.

Interoperability standards to extend impact and reduce costs

Over the last several years significant progress and alignment has been made across multiple cloud-based environments to define standards for data access, exchange, and computation. Relevant examples include the NIH Researcher Authentication Service (RAS) and key standards from the Global Alliance for Genomics and Health (GA4GH) such as Data Repository Service (DRS). Implementation of these standard APIs can enable existing cloud-based analysis environments to interoperate with SRA data readily and cost-efficiently. This in turn expands the number of users that can directly benefit from SRA data *'in situ'*, ie without requiring an additional cloud- or local- copy of data. The NIH should continue to encourage development, implementation, and adoption of these standards to ultimately support the end user community and lower overall data storage and analysis costs.

Streamline guidance for cloud costs

This is a crucial moment in biomedicine, where we are coming to realize that compute and storage (as well as computational personnel) are an increasingly important component of modern biomedical research. This also means that the cost of supporting these capabilities is rising. At the same time, NIH funding models have not kept pace with this phase change, such that we are both expecting that our storage costs will remain constant (which would require defying a law of physics) and also creating perverse incentive structures where cloud costs come out of direct funding, but on prem computing comes out of indirect. For the NIH at large, cloud presents a clear cost savings mechanism, wherein, rather than paying for the costs of storing popular datasets at every medical center, we can instead pay for only one copy in the cloud. We should all recognize, however that this is a significant change in how biomedical research is done. The working group emphasized that the SRA team is doing impressive work in managing this transition and recommends that the NIH encourages use and availability of cloud credits to make use of SRA in the cloud.

Provide intermediate or processed data on cloud

Gene expression analysis is the one of the most common use cases for RNA-seq data. Currently, a user needs to download fastq files, performs mapping using tools such as STAR (Dobin et al, 2012) to an assembled reference genome, and then prepare feature counts using programs such as HTSeq (Quinlan and Hall, 2010). The feature count file has a small footprint (float data for 20,000 rows per human sample) compared with the raw RNA-seq data file. Availability of a feature count file will enable performing analysis such as differential gene expression across different cancer subtypes, construction of gene expression network, and identify potential drug targets uniquely expressed in patient samples without the need to download, store and analyzing the fastq files. Genetic variant data is another immediate data set that was accessed regularly, and pre-computed variant data files (in the VCF format) can eliminate the need for performing mapping and variant calling using the raw data.

Have a funding mechanism to support optimizing the existing cloud computing tools

Optimization of existing computing tool for Cloud usage is essential for bringing more users to the Cloud platform, which will ultimately reduce the cost for storage and computing for the broad research community. There are several considerations: 1) how to utilize the cloud resource more efficiently so that computing cost can be reduced using the low-priority nodes/queues; 2) how to take advantage the elasticity of cloud computing to ensure delivery of time-sensitive results (e.g. COVID 19 variant analysis in an outbreak); and 3) how to implement a time-out mechanism to protect users from getting charged with unexpected cost due to issues with low-quality or complexity of the data. Furthermore, tracking the usage of cloud-deployed tools may motivate software engineers to make effort in code optimization as broad adoption by the research community is a major motivation/reward for those who developed the software tools.

Promote multi-cloud optimization of highly used (and new) tools

Software and analysis tools need to be adapted and optimized specifically for each different cloud environment. Since most new tools are often developed by postdocs or computational students, this presents a huge barrier for the developer who has no vested interest in undertaking the efforts required to make their tools cloud agnostic. Tool optimization typically occurs in a local environment, or a single cloud environment, where tool use is describe (often in a publication), and then left for users to work out their own subsequent use and optimization. It's typically much easier for the user to test and optimize a tool in their own local environment (at little to no cost for 'errors') and then download data for analysis. For widely used tools some incentive should be provided to the tool creators to undertake this multi-cloud optimization of those tools, using the same set of data housed on different clouds. Additionally, small grant supplements could be made available to encourage new tool developers (postdocs etc) to also do the same?

Promote submission of robust sample metadata

To improve the possibility of reproducible science and foster the principles of FAIR data accessibility, integration, attribution as well as the subsequent linking of related data as outlined in the Biodiversity Collections Network (BCoN) Extended Specimen Network report (https://academic.oup.com/bioscience/article/70/1/23/5637849) and National Academies of Sciences, Engineering and Medicine report on Biological Collections (https://www.nationalacademies.org/our-work/biological-collections-their-past-present-and-future-contributions-and-options-for-sustaining-them), we recommend that NIH consider working collaboratively with the various stakeholder communities to develop and deploy standards and best practices for metadata associated with sequence information. These standards should include the ability to link voucher sample information to sequence depositions in order to validate taxonomic identity and provide attribution to collections housing these tissue and voucher materials. Standards should be widely dissemninated and best practices described in depositor guidelines to ensure complete and correct metadata submission to SRA and BioSample. Recommend that NIH consider methods to provide community input into curation and metadata associated with sequences to ensure data correctness, completeness, and quality. Recommend that NIH balance the needs to support the curation of metadata with other priorities.

Conclusions

This second working group is the continuation of an engagement process with the NIH to provide access to SRA data in the cloud in ways that are equitable, fair, and available to researchers across the biomedical landscape. NIH may consider an advisory committee as a long-term engagement process to provide recommendations to enrich SRA's impact to the biomedical research community. As NIH plans next steps based on the recommendations in this report, the implementation teams will continue to engage the broader research community to solicit input on data formats, data access, and cloud use. NIH will need to carefully design the strategies in ways to maximize the value of NIH cloud investments for the research community.

References

1. Wahba, L., et al., *An Extensive Meta-Metagenomic Search Identifies SARS-CoV-2-Homologous Sequences in Pangolin Lung Viromes.* mSphere, 2020. **5**(3).
2. Bono, H., *Meta-Analysis of Oxidative Transcriptomes in Insects.* Antioxidants (Basel), 2021. **10**(3).
3. Li, H., et al., *riboCIRC: a comprehensive database of translatable circRNAs.* Genome Biol, 2021. **22**(1): p. 79.
4. Hop, P.J., et al., *Genome-wide identification of genes regulating DNA methylation using genetic anchors for causal inference.* Genome Biol, 2020. **21**(1): p. 220.
5. Da, L., et al., *AppleMDO: A Multi-Dimensional Omics Database for Apple Co-Expression Networks and Chromatin States.* Front Plant Sci, 2019. **10**: p. 1333.
6. Vieira, G.A. and F. Prosdocimi, *Accessible molecular phylogenomics at no cost: obtaining 14 new mitogenomes for the ant subfamily Pseudomyrmecinae from public data.* PeerJ, 2019. **7**: p. e6271.
7. Souvorov, A., R. Agarwala, and D.J. Lipman, *SKESA: strategic k-mer extension for scrupulous assemblies.* Genome Biol, 2018. **19**(1): p. 153.
8. Sayers, E.W., et al., *Database resources of the National Center for Biotechnology Information.* Nucleic Acids Res, 2021. **49**(D1): p. D10-D17.
9. Gavrielatos, M., et al., *Benchmarking of next and third generation sequencing technologies and their associated algorithms for de novo genome assembly.* Mol Med Rep, 2021. **23**(4).
10. Marti-Carreras, J., et al., *NCBI's Virus Discovery Codeathon: Building "FIVE" -The Federated Index of Viral Experiments API Index.* Viruses, 2020. **12**(12).
11. Kolmykov, S., et al., *GTRD: an integrated view of transcription regulation.* Nucleic Acids Res, 2021. **49**(D1): p. D104-D111.
12. Li, C., et al., *Genome Variation Map: a worldwide collection of genome variations across multiple species.* Nucleic Acids Res, 2021. **49**(D1): p. D1186-D1191.

13. Kanakoglou, D.S., et al., *Effects of High-Dose Ionizing Radiation in Human Gene Expression: A Meta-Analysis.* Int J Mol Sci, 2020. **21**(6).

14. Zuo, Z., et al., *BBCancer: an expression atlas of blood-based biomarkers in the early diagnosis of cancers.* Nucleic Acids Res, 2020. **48**(D1): p. D789-D796.

15. Knierim, A.B., et al., *Genetic basis of functional variability in adhesion G protein-coupled receptors.* Sci Rep, 2019. **9**(1): p. 11036.

16. Wei, Y., J.R. Silke, and X. Xia, *An improved estimation of tRNA expression to better elucidate the coevolution between tRNA abundance and codon usage in bacteria.* Sci Rep, 2019. **9**(1): p. 3184.

17. Davies, M.R., et al., *Atlas of group A streptococcal vaccine candidates compiled using large-scale comparative genomics.* Nat Genet, 2019. **51**(6): p. 1035-1043.

18. Armstrong, G.L., et al., *Pathogen Genomics in Public Health.* N Engl J Med, 2019. **381**(26): p. 2569-2580.

19. Connor, R., et al., *NCBI's Virus Discovery Hackathon: Engaging Research Communities to Identify Cloud Infrastructure Requirements.* Genes (Basel), 2019. **10**(9).

20. NCBI. *Cloud Data Delivery Service* [cited 2021 May 20]; Available from: https://www.ncbi.nlm.nih.gov/Traces/cloud-delivery/.

21. Yu, Y.W., et al., *Quality score compression improves genotyping accuracy.* Nat Biotechnol, 2015. **33**(3): p. 240-3.

22. Wan, R., V.N. Anh, and K. Asai, *Transformations for the compression of FASTQ quality scores of next-generation sequencing data.* Bioinformatics, 2012. **28**(5): p. 628-35.

23. Ochoa, I., et al., *QualComp: a new lossy compressor for quality scores based on rate distortion theory.* BMC Bioinformatics, 2013. **14**: p. 187.

24. Janin, L., G. Rosone, and A.J. Cox, *Adaptive reference-free compression of sequence quality scores.* Bioinformatics, 2014. **30**(1): p. 24-30.

25. Shibuya, Y. and M. Comin, *Better quality score compression through sequence-based quality smoothing.* BMC Bioinformatics, 2019. **20**(Suppl 9): p. 302.

26. AWS. *NIH NCBI Sequence Read Archive (SRA) on AWS*. [cited 2021 May 20]; Available from: https://registry.opendata.aws/ncbi-sra/.

27. NIH. *Request for Information (RFI), Use of Cloud Resources and New File Formats for Sequence Read Archive Data*. [cited 2021 May 20]; Available from: https://grants.nih.gov/grants/guide/notice-files/NOT-OD-20-108.html.

28. NIH. *Researcher Auth Service Initiative*. Available from: https://datascience.nih.gov/researcher-auth-service-initiative.

29. NIH. *Cloud Platform Interoperability Effort* [cited 2021 May 20]; Available from: https://anvilproject.org/ncpi.

Appendix 1: Organismal content and usage of public SRA

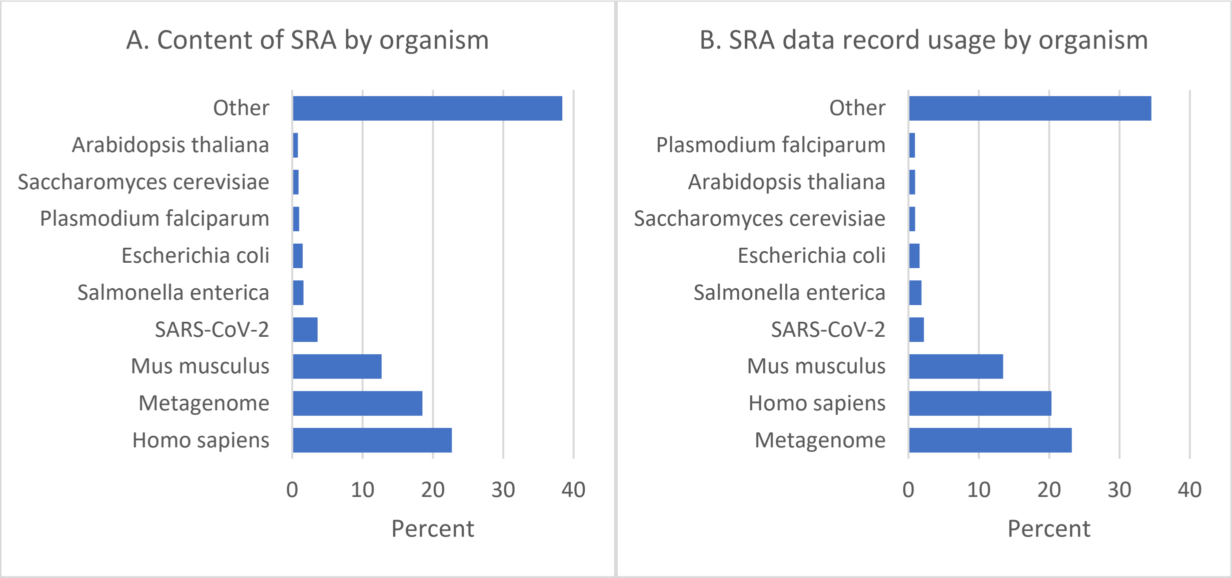

Figure A1*. Organismal content and usage of public SRA. Panel A shows the most prevalent organisms in the public portion of SRA. Panel B shows SRA data record usage in public SRA based on organism. In both graphs, the organism name associated with individual data records was used to collate data; all samples with the term "metagenomic" in the organism name were aggregated into a single value. Data are derived from SRA submissions and total usage logs – cloud and non-cloud – between 11/2019 to 3/2021. Data records associated with more than 10,000 individual organism names were retrieved 178 or more times during this period.*

Appendix 2: Assay type content of SRA

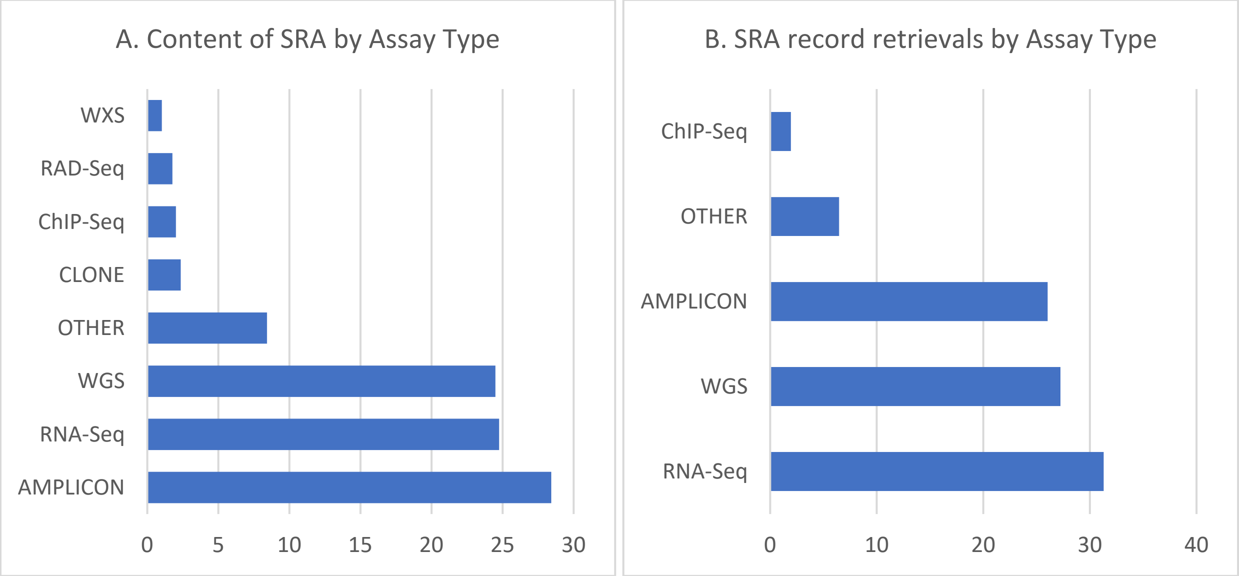

Figure A2*. SRA content and usage by Data Type. Panel A shows the most prevalent Data Types in SRA. Panel B shows SRA data record usage based on Data Type. Data are derived from SRA submissions and total usage logs – cloud and non-cloud – between 11/2019 to 3/2021.*

Appendix 3: Glossary of Terms Used in this Report

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.

Hot storage: A form of cloud storage in which data are immediately available to users.

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.

Thaw: The process of transferring data from cold to hot storage in the cloud.

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.