Council of Councils Working Group on Sequence Read Archive Data

Interim Report

January 24, 2020

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. At present, the SRA contains 12 petabytes of data and is continually growing. The SRA was moved to 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 2019 the SRA archive held 9 million records in two formats. The original format (23 petabytes) is received by NCBI from submitters and is instrument- and experiment-specific; these data are stored to tape. NCBI transforms these original format data into standard SRA normalized format (12.7 petabytes) for redistribution. The normalized format contains base quality scores (BQS) that provide information about the quality of the sequence; however, because of the number of possible BQS for each base, they drastically increase file size, thus making BQS the largest cost driver for SRA storage in the cloud.

The NIH has engaged the SRA Data Working Group of the NIH Council of Councils to provide input on how to address the long-standing challenge of ensuring SRA's sustainability as an archive of exponentially growing experimental data. The SRA continues to experience exponential growth in submission rates, and the normalized format data is projected to grow to 33 petabytes by 2023; at this rate, the increasing size will quickly exceed NIH budget for storage and maintenance. The SRA Working Group's charge is to provide recommendations to the Council of Councils on several key factors for managing this data resource in cloud service provider environments. The NIH is requesting that the working group identify and evaluate solutions to maintain efficiency in the storage footprint of SRA, specifically relevant to the use of BQS and format compression strategies.

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 access, and projections of future growth and use. In the data models and discussions, they 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.

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 this large data resource in the cloud. The immediate recommendations are:

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

- BQS should be retained in original format data, and two versions of SRA normalized format data should be maintained: one with BQS and one without them.
- In both AWS and GCP clouds, hot storage should contain the full set of normalized data without BQS, as well as the most actively accessed half of normalized data with BQS; cold storage should contain the less active half of normalized data with BQS, as well as all original format data.
- NCBI should monitor data usage and determine the appropriate storage location (hot or cold storage) for each dataset deposited into SRA to be provisioned in the cloud depending on these usage data.
- 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 massive overuse of NIH compute resources.

Communication of the model

- Cost models should be provided in clear language to the research community via the Office of Data Science Strategy STRIDES website, which can in turn be referenced by the NCBI SRA website and other public-facing communications mechanisms (e.g., the NIH Guide).
- The information provided should include specifics about costs for both storage and compute, as well as how these costs are expected to be distributed among users, institutions, and the NIH.
- Education must be provided for current and prospective cloud users so that they can understand which cloud service providers host SRA data, which data are available in hot and cold storage and how to access both types of data, and how to monitor compute time.

Continued research to inform changes to the model over time

- The working group recognizes that the current model has been selected based on limited data and recommends that NIH monitor costs over time to make adjustments based on the actual costs of researchers working in the cloud. This will include determining if different strategies are needed for different cloud service providers, as well as assessing what other strategies may be needed if additional cloud service providers become SRA hosts.
- NIH should consider funding efforts to optimize code and efficiency for use in the cloud to reduce the cost of computing in this environment. The goal of these projects should be to design tools that ultimately can limit users’ costs by reducing the need to egress data from the cloud.

Table of Contents

Background and Challenge

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. Researchers continue to mine these sequences for new discoveries about genome architecture, natural variation, gene expression, methylation states, and the identification of unknown species, viruses, and genes in microbiome and metagenome samples. Given its vast size and diversity, the SRA represents a crucial resource for the scientific research community. In the past five years, researchers have relied on data from SRA to assemble and annotate (i.e., identify genes within) new genomes, characterize pathogens, understand the processes that drive genome evolution, identify and predict the functional effects of rare human genetic variation, and develop new bioinformatics tools and methods [1–8]. At present, the SRA contains 9 million records totaling 12 petabytes of data, and its size is continually growing.

The NIH has engaged the SRA Data Working Group of the NIH Council of Councils to provide input on how to address the long-standing challenge of ensuring SRA's sustainability as an archive of exponentially growing experimental data. Specific use cases conducted by individual working group members' labs highlight examples of the value SRA provides the research community.

- The Edwards lab [9] focuses on understanding the environmental microbial datasets in the SRA. There are two main types of environmental microbial data: microbiome sequences (amplicon sequencing) and metagenome sequences (random community genome sequences) that require very different computational analyses and approaches. The lab developed tools to automatically discriminate between and annotate these two types of datasets, and these tools are run every month with the new data submitted to the SRA [10,11]. The lab built a website where users can upload DNA or protein sequences for comparison to those datasets [12].
- The Kang lab [13] downloads original or aligned sequence reads from the SRA when it wants 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 a paper and improve upon it. Sometimes, the lab may request a large amount of data when the data needed is population-scale (such as GTEx data, or PsychENCODE data).
- The Zhang lab [14] primarily focuses on pediatric cancer genomic research and uses the SRA in various ways to support its research program. It uses 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 [15,16]. The lab also uses SRA data to develop and validate new bioinformatics tools and to demonstrate the broad applicability of these methods across

sample types. For example, the lab has recently used public SRA data to validate performance of a new method for normalizing ChIP-seq data [17] and repeatedly used restricted access SRA data to demonstrate the applicability of methods developed with pediatric cancer data to adult cancer datasets [18].

The SRA was moved to Google Cloud Platform (GCP) and Amazon Web Services (AWS) cloud services in 2019 as part of the NIH Science and Technology Research Infrastructure for Discovery, Experimentation, and Sustainability (STRIDES) Initiative. Under STRIDES, NIH is developing a cloud-enabled biomedical data ecosystem, and the SRA is one of the first – and to date the largest – datasets that has been moved to the cloud. NIH is piloting new processes, tools, and equities to help drive development of a true data ecosystem that moves toward making NIH-funded data findable, accessible, interoperable, and reusable (FAIR). Migrating SRA data to the cloud represents an important use case for this process in three ways. First, SRA data are now available in both their original, instrument-specific format as well as the normalized SRA format, which promotes reusability for both replication and discovery. Second, the entire corpus of 9 million records is provisioned for immediate compute, which promotes availability and use. Finally, the metadata for all SRA submissions have been stored in Google's BigQuery service to make the data findable by user-defined custom searches.

Current status of SRA access in the cloud

In 2019 the SRA archive held 9 million records in two formats. Original format (23 petabytes) is received by NCBI from submitter and is instrument and experiment-specific; these data are stored to tape. NCBI transforms the normalized format (12.7 petabytes) into standard SRA format for redistribution. The SRA continues to experience exponential growth in submission rates, and the normalized format of data is projected to grow to 33 petabytes by 2023 (Table 1). The SRA holds both public data that is available without restriction 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. As of 2019 the public and controlled-access partitions of SRA were equal size. The SRA had over 1.2 million visitors who downloaded over 8.5 petabytes of sequence from NCBI servers in total in 2019. Twenty percent of those visits were from IP addresses in the cloud platforms (as opposed to addresses indicating the use of on-premise computing), suggesting that SRA has a sizable cloud-based user community ready to work on data stored locally in the cloud.

Cloud storage enables access to additional SRA data formats

As discussed above, each data record submitted to SRA exists in two forms:

(1) an **original** format as submitted to NCBI that reflects the results of a submitter's particular instrument output and analysis computational pipeline. These record formats vary across submissions in both the structure and meaning of data objects. While SRA includes original format data from 20 sequencing platform-specific schemas, 91 percent

of submissions were either sequences aligned to a reference genome, e.g. BAM (66%), or unplaced Illumina sequence, e.g. FASTQ for RNASeq (25%).

(2) To support the systematic reuse of these data, NCBI creates a **normalized** (extract, transform, load [ETL]) record of the data for compact storage, efficient redistribution, and standard representation.

Historically, the normalized format has been the only version of the SRA records distributed by NCBI. The larger set of submission records in their various original formats were archived to tape storage during processing as backups for system recovery events. In the summer and fall of 2019, NCBI copied all normalized format SRA data to AWS and GCP and adopted a policy that future submissions would be provisioned in both original and normalized format in the cloud for user access. Copying the legacy submissions of original format records from NCBI's tape archive to the cloud is ongoing, but given the performance constraints of the tape system hardware, NCBI estimates that it will take an additional 18-24 months to fully restore those files to the cloud. When the migration of legacy original format data into the cloud is complete in 2021, SRA expects to provision 20 petabytes of SRA data in normalized format and 35 petabytes in original format in both cloud platforms (Table 1).

Researchers have long requested access to both data formats, and NCBI has now enabled that access through cloud storage. This opens up new research opportunities. First, access to the original format corpus is available for users interested in **replicating** published results exactly as submitted. Second, the complete corpus of all normalized SRA data is available as a computational resource for new **discovery** tasks and algorithms that can now be run over all data.

A foundational search task would be to find all records in SRA with sequences that match a user's query. Traditionally, SRA data has been too large to comprehensively download to local data centers, and users searching for data of interest have been restricted to searching and downloading subsets of SRA using only metadata key terms attached to records. By replicating SRA content into the cloud, NIH has created an opportunity for data scientists to engineer powerful search and discovery algorithms. Researchers working on these platforms now enjoy full read access to public SRA submissions and could execute these sequence-based searches of SRA as part of a research ecosystem. Other discovery opportunities are available, such as powerful cloud-based search engines that can mine the full set of SRA metadata tags using community-developed queries and quickly assemble sets of records for downstream analysis.

Table 1. Current size and projected* future growth of SRA by format type (in petabytes).

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

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

Cloud storage options create a framework to affordably manage multiple formats

When designing a storage and retrieval architecture for petabyte-scale data in the cloud, it is important to understand both the retrieval patterns and the modes of storage available. Storage comes in two modes with costs based on the amount of time one is willing to wait for the object to be read, i.e. read latency. Hot storage provides immediate access to data and is the default standard. More infrequently accessed data can be stored at a discount using cold storage, which is not immediately accessible. However, cloud service providers apply a charge to "thaw" data from cold to hot storage, and users may have to wait up to 48 hours for that "cold" data to be available for analysis. Under the current STRIDES agreement, NIH is responsible for paying any thaw charges incurred by its users.

A review of the latest 36 months of SRA usage (Figure 1) indicated that 7.7 million records were retrieved. Both old and new records were retrieved at similar rates when rates are adjusted for database size over time, indicating continued value in at least some older SRA submissions.

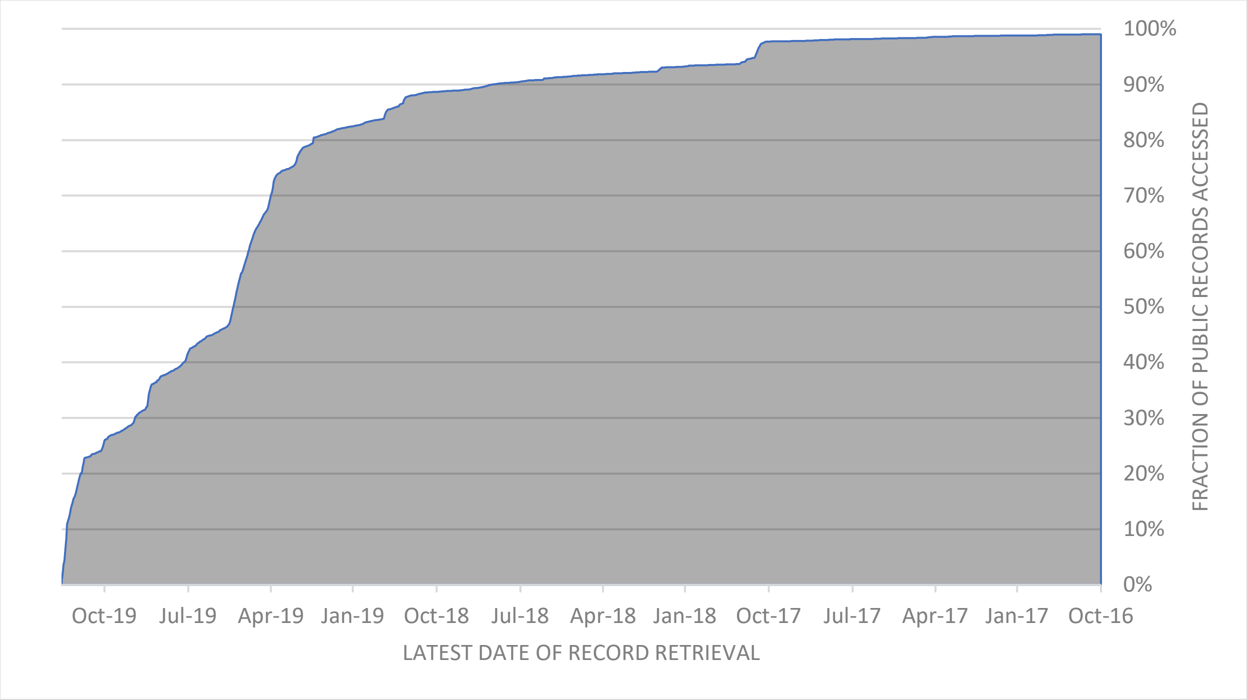

Figure 1 Cumulative distribution of waiting times until first request for SRA record indicates that 50% of the unique data records were accessed between May and October 2019. The next 49% of data records were requested at least once after October 2016. Only 1% were not requested in the three-year lookback.

Any given sequencing submission to SRA may contain some combination of the following data types: base calls, base quality scores (BQS), placement, mate pair, spot name, and platform-specific optional data. Of these the largest data type for datasets currently in the SRA is BQS, which was observed to be 60-70 percent of the submitted run-byte footprint for SRA normalized data as of 2019 (Figure 2). This makes BQS the largest cost driver for SRA storage in the cloud.

BQS are a string of characters the same length as the nucleotide sequence. They represent a quantification of the error probability for each base. Because there are 63-94 possible BQS values in contrast to the smaller number (usually four) of possible nucleotide values, BQS data are more difficult to compress than sequence data. Many studies have investigated methods for BQS compression, including combining similar scores into a smaller number of "bins" [19–24]. However, there is not currently a consensus among the research community about an optimal compression strategy, and different strategies may be preferred for different uses of sequence and BQS data.

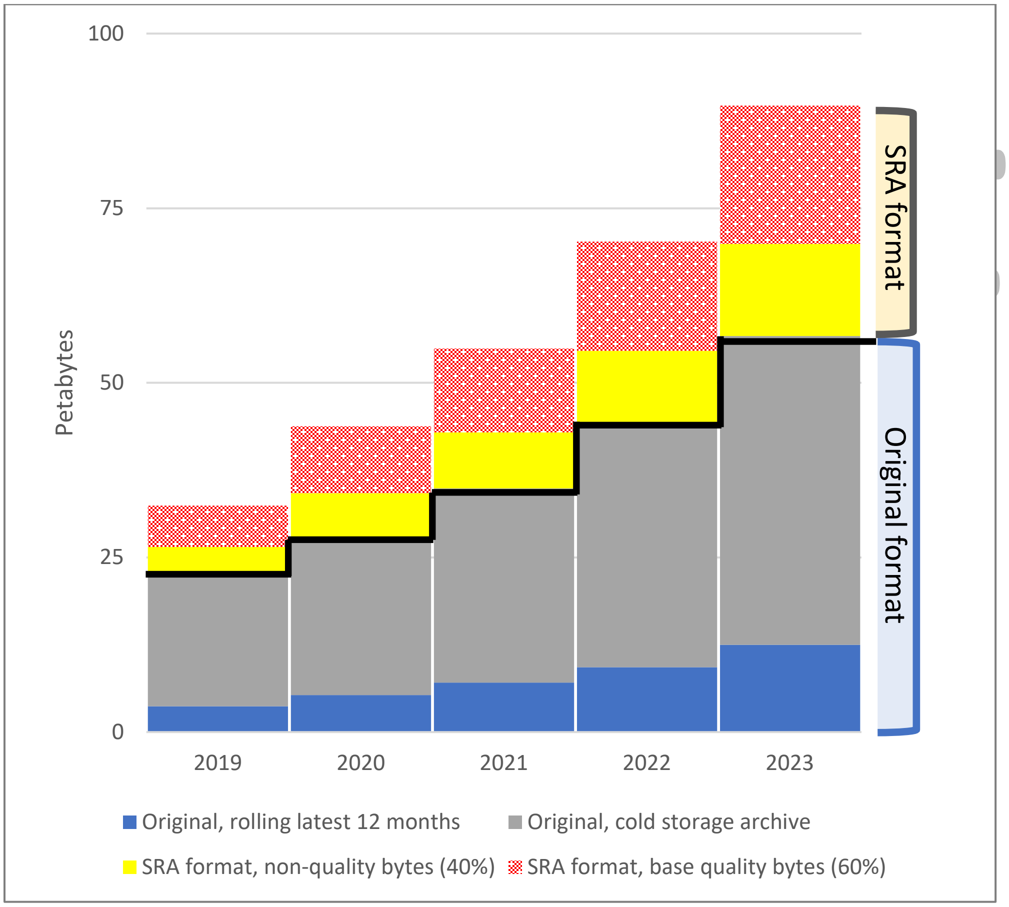

Figure 2: SRA growth is exponential. The two archive formats have different sizes and rates of growth as indicated. Storage budgets for these data requires that the older original format data be kept in cold storage. SRA normalized format can be subdivided into two categories of data types: BQS (pink) and everything else (yellow). BQS is the largest single component of normalized data by type.

Data growth necessitates new models to enable user access while controlling costs

The SRA Working Group's charge is to provide recommendations to the Council of Councils on several key factors for managing this data resource in cloud service provider environments. SRA data are growing at a pace that will quickly exceed NIH budget for storage and maintenance. The NIH is requesting that the working group identify and evaluate solutions to maintain efficiency in the storage footprint of SRA, specifically relevant to the use of BQS and format compression strategies. The data are in the clouds of two distinct service providers to facilitate maximum use and convenience to the research community, as well as to prevent price monopoly. However, each provider has different cost models for storage, compute, and egress. Furthermore, different members of the research community have tools and preferences for different cloud environments, not all of which are included under NIH's current STRIDES Initiative.

Use cases and key principles for consideration in developing models for SRA maintenance and use in the cloud

To assist in the working group's deliberations, 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.

The SRA Data Working Group members identified 10 principles to consider when proposing recommendations for maintaining efficiency in the SRA storage footprint:

1. **Continuous access to training datasets:** Some datasets are valuable public resources that are used as references for training. The working group suggested that a few such datasets be kept readily available in hot storage for immediate use in this capacity. Future discussions may include determining how to select relevant datasets for this purpose.
2. **Quality of data available for analysis:** The group assessed that changing the model for storage of the two formats of SRA data would not result in sacrifices to the data quality. However, eliminating BQS from normalized format data would reduce availability of information about data quality.
3. **Prioritizing availability of frequently accessed datasets:** The working group proposed that, to minimize effects on users, any new model be adopted strategically (e.g., reverse chronologically) regarding how data are transferred to the cloud and between hot and cold storage in the cloud. Many datasets are accessed episodically, but nearly all datasets are

accessed at some point. Statistical models of data access over time are needed to identify the best strategy for transferring datasets.

4. **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 2 for estimated average costs for NIH and users for some typical workflows.

5. **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 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.

   See Table 2 for estimated average costs for NIH and users for some typical workflows.
6. **Barriers to access:** The group wanted to avoid implementing changes that would make the SRA difficult to access, whether because of ability to access and compute in the cloud or because of the costs of computing in the cloud. This is especially in consideration of under-resourced institutions or researchers not already working in the cloud.
7. **Speed/wait time to access:** The working group agreed that while having to access data from cold storage would create an extra step for users, a 24- to 48-hour waiting period to access original or normalized data from cold storage would be acceptable. Such wait times are common for the least expensive cold storage options. The group further discussed testing the waiting period to determine the reliability of estimates for total time to thaw and compute on any data that would need to be retrieved from cold storage.
8. **Access to normalized and original formats:** The group agreed that both forms have value and should remain available to users via some mechanism.
9. **Search and random access:** The working group agreed that changes to the model of SRA data storage would not result in any sacrifices to search and random access across the entire SRA dataset, since these are not tasks that are currently available to SRA users. To enable full search and random access in the future, the working group recommended prioritizing retention of normalized format data in hot storage for these purposes.
10. **Flexibility and adaptability:** The group agreed that the recommendations should inform storage and access models that are flexible and adaptable and can change over time based on additional data.

Table 2. Relative costs of activities for common use cases of SRA and their distribution across entities.

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)	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)	Costs per institutional policy
Institution	Provide hardware for local compute (all use cases)					Substantially higher cost than hardware necessary for cloud access

Proposals for cloud storage models of SRA

The SRA Working Group outlined three proposals, all of which balanced cost to meet budget constraints:

Proposal 1: Eliminate BQS from the normalized format data but retain them in the original format data in the cloud. All normalized format data without BQS and up to 50% of the original format data would remain in hot storage in both clouds; the remainder would be maintained in cold storage in both clouds.

Proposal 2: Retain BQS within the normalized format data in the cloud. Only normalized format data with BQS would be available on hot storage in both clouds; the original format data would remain on tape backup only.

Proposal 3: Normalized format data would come in two versions: (a) normalized with BQS having actively used records (up to 50%) in hot storage and the rest in cold storage and (b) normalized without BQS having all records in hot storage to support search and discovery functions. All original format data would also be retained in the cloud as version (c) in cold storage.

Modeling storage costs under the three proposals

Cloud storage costs are a function of database size and storage modality (i.e., the necessary wait time to retrieval of data). NCBI analyzed AWS and GCP pricing models for two storage modalities: hot storage with immediate retrieval and cold storage with delayed, less expensive retrieval for infrequently accessed data.

The pattern of SRA usage over the past 36 months was used to model the predicted costs for each of the proposals under a range of policies regarding the distribution of data across hot and cold storage for active and inactive SRA records. It was straightforward to estimate cost components for full sets of data within a single storage modality, e.g. all original format data in cold storage and all normalized format without BQS in hot storage. To model the third proposal, in which normalized data with BQS are distributed over hot and cold storage, NCBI needed to estimate the likelihood of data reuse after a record was thawed. This was done as a simulation using the past 36 months of SRA usage data. In each simulation run, total costs to store and thaw SRA data were calculated using a hot storage data retention policy that ranged from one to 180 days. It was found that a retention policy of 36 days would provide a 50% chance that a record would be in hot storage at any point in time. This was judged to be affordable tradeoff between wait time for a researcher to obtain data and operational cost to NIH.

The working group also determined that Proposal 3 was best aligned with the 10 previously stated principles. Based on the cost models described above, they determined that a policy of retaining up to approximately 50% of the total size of all normalized data with BQS in hot storage would result in a budget aligned with the NIH strategic budget goals for this resource. Assuming that these data continue to be accessed at current rates, this is approximately equal to leaving any SRA data retrieved from cold storage in hot storage for 36 days after its initial thaw. Adopting this policy would result in the following implications for the three use cases:

- All data would be available in both AWS and GCP clouds in either hot or cold storage (necessary for all three use cases).
- Retrieval from cold storage should take less than 48 hours (reducing barriers to access for use cases 1 and 2).
- All normalized format data without BQS would be available in hot storage (enabling use case 3 without any need to thaw data).

- All original format data would be available in cold storage (enabling use case 1 once specific data are thawed).
- Normalized format data with BQS would be distributed between hot and cold storage (requiring thawing for some instances of use case 2 and allowing immediate access for others).

Recommendations

The working group members formulated the recommendations, integrating the 10 principles listed above into their deliberations.

They agree that the recommendations in this report are intended to address immediate concerns about the SRA storage footprint and that approaches incorporating these recommendations should be dynamic and responsive to future changes in use data. Furthermore, these recommendations propose building flexible models that can continue to incorporate changes in use data to inform future decision-making.

The recommendations are organized into three main concepts: (1) a new model for SRA data storage and retrieval in the cloud, (2) communication of the model, and (3) continued research to inform changes to the model over time.

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

- BQS should be retained in original format data, and two versions of SRA normalized format data should be maintained: one with BQS and one without them.
- In both AWS and GCP clouds, hot storage should contain the full set of normalized data without BQS as well as the most actively accessed half of normalized data with BQS; cold storage should contain the less active half of normalized data with BQS as well as all original format data.
- NCBI should monitor data usage and determine the appropriate storage location (hot or cold storage) for each dataset deposited into SRA to be provisioned in the cloud depending on these usage data.
- 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 massive overuse of NIH compute resources.

Communication of the model

- Cost models should be provided in clear language to the research community via the Office of Data Science Strategy STRIDES Initiative webpage, which can in turn be referenced by the NCBI SRA website and other public-facing communications mechanisms (e.g., the NIH Guide).

- The information provided should include specifics about costs for both storage and compute as well as how these costs are expected to be distributed among users, institutions, and the NIH.
- Education must be provided for current and prospective cloud users so that they can understand which cloud service providers host SRA data, which data are available in hot and cold storage and how to access both types of data, and how to monitor compute time.

Continued research to inform changes to the model over time

- The working group recognizes that the current model has been selected based on limited data and recommends that NIH monitor costs over time to make adjustments based on the actual costs of researchers working in the cloud. This will include determining if different strategies are needed for different cloud service providers and assessing what other strategies may be needed if additional cloud service providers become SRA hosts.
- NIH should consider funding efforts to optimize code and efficiency for use in the cloud to reduce the cost of computing in this environment. The goal of these projects should be to design tools that ultimately can limit users' costs by reducing the need to egress data from the cloud.

Future Work and Considerations

The SRA Working Group discussed several considerations that, while initially focused on SRA, could be applied to other NIH-supported high value dataset that are in, or will be moved to, the cloud. These considerations may require longitudinal analysis of data access and data usage or require additional community input and are therefore left for future work.

The working group highlighted that SRA is a very dynamic dataset that has both time-sensitive and research-sensitive considerations. SRA data is most frequently accessed immediately after it becomes available through NCBI. However, certain events, such as an outbreak of an infectious disease, and high-profile research results, such as new studies on human diseases, may cause SRA data to be repeatedly accessed in an unpredictable fashion. The working group felt that future studies of repeated access and age of accessed data sets would be important to conduct before a final recommendation could be made about the SRA data storage lifecycle.

The working group considered a proposal that would eliminate BQS from the normalized data available in the cloud; however, it felt that this would be too extreme a change at present given the workflows that currently rely on the presence of BQS. However, it acknowledged that reducing the large data footprint of BQS within the SRA is an important goal for the future and suggested additional study of appropriate compression strategies for BQS, such as binning the scores into a smaller number of possible values. The group also felt that a longitudinal study of the effects of binning or removing BQS on research workflows, algorithms, and analytical pipelines was warranted before a final recommendation could be made about strategies to reduce the data footprint of BQS.

The working group noted that researchers in the SRA community may also have workflows and pipelines in other cloud service providers, such as Microsoft Azure. The working group also felt that the nature of duplicating data in both GCP and AWS cloud service providers inherently creates complexities in data provenance. In the future, the working group would like to explore the possibility of decoupling models for data storage between GCP and AWS as well as better understand the community's need for additional cloud service providers.

The group felt that at a minimum an RFI to understand how researchers are using, or anticipate using, SRA in the cloud was warranted before a recommendation could be made about SRA data provenance in multiple cloud service providers.

The working group noted that there is potential risk in a cost model that relies heavily on cloud service providers, and the discussion of how NIH might mitigate that risk included the notion of a public research cloud, as is being considered by agencies such as the NSF.

The working group noted that the current STRIDES arrangement was not initially set up to optimize costs for cold storage and thawing of data, which are key components of the proposed model for SRA. Use of cold storage will have a large impact on the future of biomedical research. The group would like further information about how the NIH might consider cold storage in its future negotiations with cloud service provider partners.

Finally, the working group noted that the SRA is used by many different researchers well beyond the biomedical research communities supported by NIH. The working group felt it was important to explore the need to integrate analysis between SRA and data platforms supported by other agencies, including those using other cloud service providers. The group would like to explore the use cases involving integration of SRA with other data platforms before making a final recommendation on SRA data interoperability.

References

1. Davies MR, McIntyre L, Mutreja A, Lacey JA, Lees JA, Towers RJ, et al. Atlas of group A streptococcal vaccine candidates compiled using large-scale comparative genomics. Nat Genet. 2019;51: 1035–1043. doi:10.1038/s41588-019-0417-8

2. Vasconcelos EJR, Dasilva LF, Pires DS, Lavezzo GM, Pereira ASA, Amaral MS, et al. The Schistosoma mansoni genome encodes thousands of long non-coding RNAs predicted to be functional at different parasite life-cycle stages. Sci Rep. 2017;7. doi:10.1038/s41598-017-10853-6

3. Souvorov A, Agarwala R, Lipman DJ. SKESA: strategic k-mer extension for scrupulous assemblies. Genome Biol. 2018;19: 153. doi:10.1186/s13059-018-1540-z

4. Lek M, Karczewski KJ, Minikel E V., Samocha KE, Banks E, Fennell T, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature. 2016;536: 285–291. doi:10.1038/nature19057

5. Kuleshov M V., Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Res. 2016;44: W90–W97. doi:10.1093/nar/gkw377

6. Tang Z, Li C, Kang B, Gao G, Li C, Zhang Z. GEPIA: a web server for cancer and normal gene expression profiling and interactive analyses. Nucleic Acids Res. 2017;45: W98–W102. doi:10.1093/nar/gkx247

7. Armstrong GL, MacCannell DR, Taylor J, Carleton HA, Neuhaus EB, Bradbury RS, et al. Pathogen Genomics in Public Health. N Engl J Med. 2019;381: 2569–2580. doi:10.1056/NEJMsr1813907

8. EdwardsLab | Delivering the best in bioinformatics…. [cited 23 Dec 2019]. Available: https://edwards.sdsu.edu/research/

9. GitHub - linsalrob/partie: PARTIE is a program to partition sequence read archive (SRA) metagenomics data into amplicon and shotgun data sets. The user-supplied annotations of the data sets can not be trusted, and so PARTIE allows automatic separation of the data. [cited 23 Dec 2019]. Available: https://github.com/linsalrob/partie

10. GitHub - linsalrob/partie_hat: High throughput automated typing for the partition engine. [cited 23 Dec 2019]. Available: https://github.com/linsalrob/partie_hat

11. Home. [cited 23 Dec 2019]. Available: https://www.searchsra.org/

12. Hyun Min Kang, PhD | Faculty Profiles | U-M School of Public Health. [cited 23 Dec 2019]. Available: https://sph.umich.edu/faculty-profiles/kang-hyunmin.html

13. St. Jude Research | Zhang laboratory. [cited 23 Dec 2019]. Available: https://www.stjuderesearch.org/site/lab/zhang

14. Zhang J, McCastlain K, Yoshihara H, Xu B, Chang Y, Churchman ML, et al. Deregulation of DUX4 and ERG in ALL--Jinghui Zhang. Nat Genet. 2016;48: 1481–1489. doi:10.1038/ng.3691

15. Ma X, Liu Y, Liu Y, Alexandrov LB, Edmonson MN, Gawad C, et al. Pan-cancer genome and transcriptome analyses of 1,699 paediatric leukaemias and solid tumours. Nature. 2018;555: 371–376. doi:10.1038/nature25795

16. Jin H, Kasper LH, Larson JD, Wu G, Baker SJ, Zhang J, et al. ChIPseqSpikeInFree: a ChIP-seq normalization approach to reveal global changes in histone modifications without spike-in. Berger B, editor. Bioinformatics. 2019 [cited 23 Dec 2019]. doi:10.1093/bioinformatics/btz720

17. Chen X, Gupta P, Wang J, Nakitandwe J, Roberts K, Dalton JD, et al. CONSERTING: integrating copy-number analysis with structural-variation detection. Nat Methods. 2015;12: 527–30. doi:10.1038/nmeth.3394

18. Illumina. Reducing Whole-Genome Data Storage Footprint.

19. Yu YW, Yorukoglu D, Peng J, Berger B. Quality score compression improves genotyping accuracy. Nature Biotechnology. Nature Publishing Group; 2015. pp. 240–243. doi:10.1038/nbt.3170

20. Wan R, Anh VN, Asai K. Transformations for the compression of fastq quality scores of next-generation sequencing data. Bioinformatics. 2012;28: 628–635. doi:10.1093/bioinformatics/btr689

21. Ochoa I, Asnani H, Bharadia D, Chowdhury M, Weissman T, Yona G. QualComp: A new lossy compressor for quality scores based on rate distortion theory. BMC Bioinformatics. 2013;14. doi:10.1186/1471-2105-14-187

22. Janin L, Rosone G, Cox AJ. Adaptive reference-free compression of sequence quality scores. Bioinformatics. 2014;30: 24–30. doi:10.1093/bioinformatics/btt257

23. Shibuya Y, Comin M. Better quality score compression through sequence-based quality smoothing. BMC Bioinformatics. 2019;20: 302. doi:10.1186/s12859-019-2883-5

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.

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.