Sequence Read Archive (SRA) Data Working Group

Proposed Working Group of the Council of Councils

September 6, 2019

Council of Councils

National Institutes of Health
Office of Data Science Strategy

Topics to be Presented

- Background on SRA
- Proposed Charge for SRA Data Working Group

Implementing the Goals of NIH Strategic Plan for Data Science

Data Infrastructure

- **Leverage existing federal, academic, and commercial computer systems for data storage and analysis.**

Modernized Data Ecosystem

- Support storage and sharing of individual datasets

Data Management, Analytics, and Tools

- Improve discovery and cataloging resources

Workforce Development

Stewardship and Sustainability

- Establish sustainability models for data resources

GenBank	SRA
Released in 1982	Released in 2009
Assembled sequence	Fragmented sequence
2.6 billion records • 15 TB	9 million records • 12 PB
Rich annotations	No annotations
Rich metadata	Limited metadata
Typically represents single molecule or genome	Sequence captures ‘everything in the tube’
Records are small	Records are huge

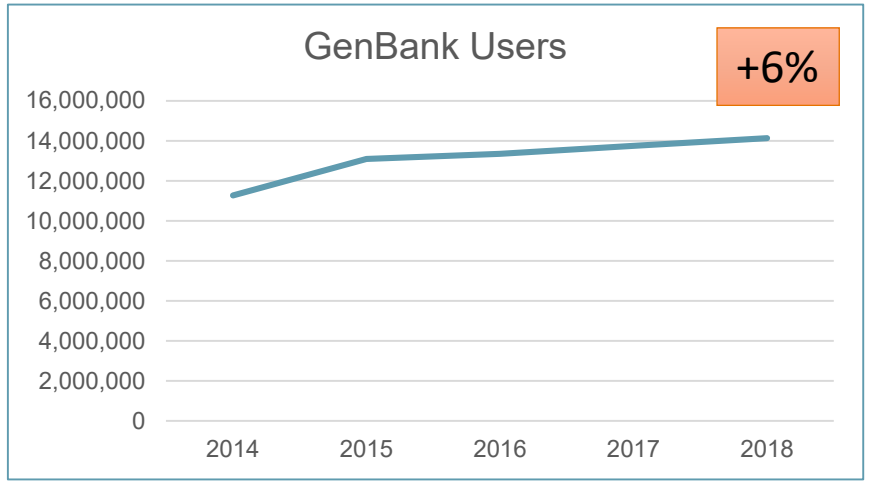
Sequence Read Archive

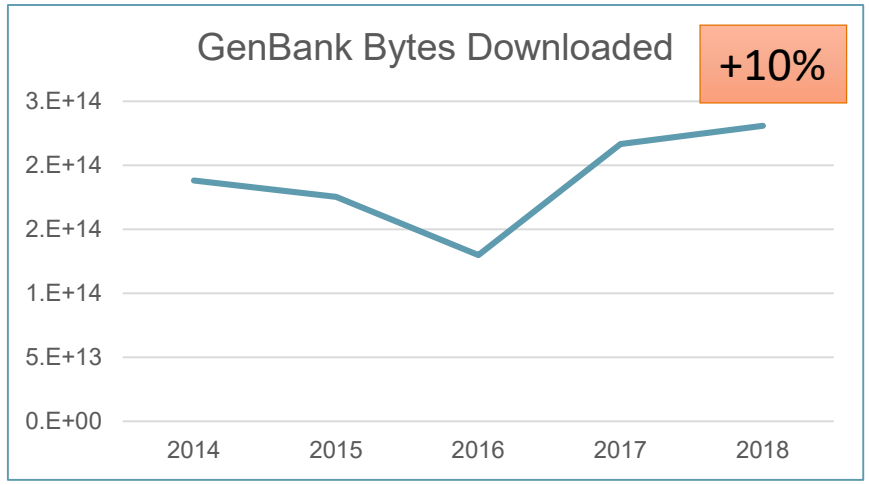
Objectives/Activities

- Archives raw oversampling NGS data for various organisms from several platforms
- Shares submitted NGS data with EMBL and DDBJ
- Serves as a starting point for “secondary analyses”
- Provides access to data from human clinical samples to authorized users who agree to the dataset’s privacy and usage mandates

GenBank

Daily access (web): 4,300 organizations

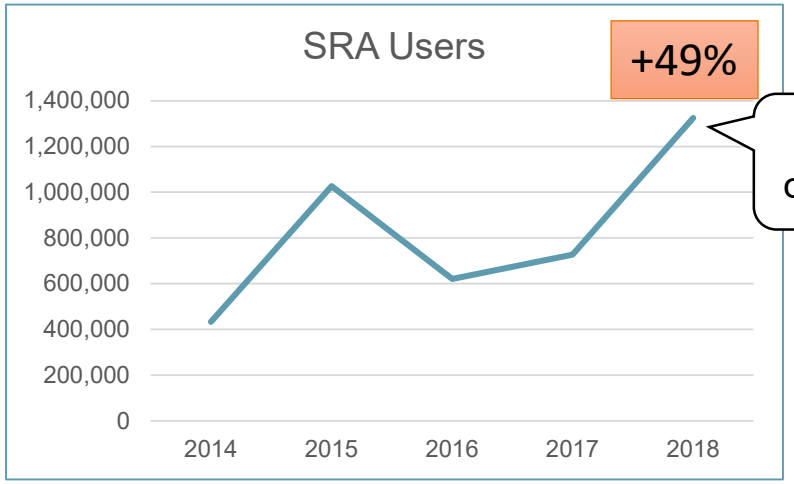




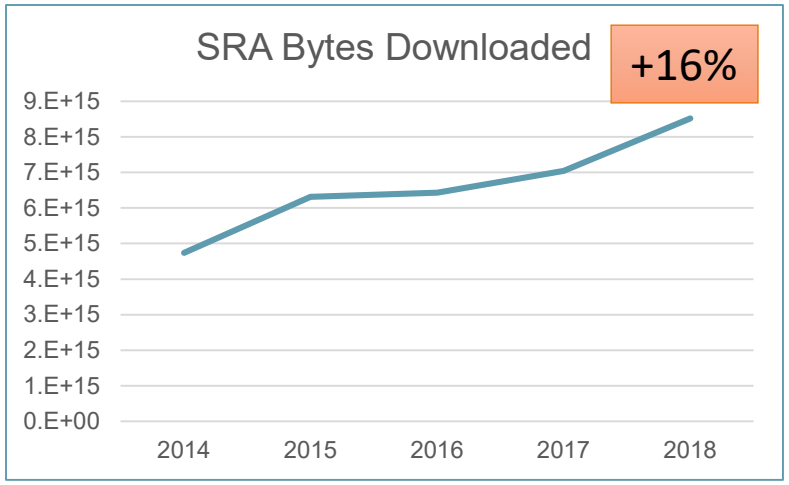
SRA

Daily access (web): 1,500 organization

Average annual growth rates



~20% of 2019 usage is on Amazon and Google.



SRA Data Working Group

Rationale:

Through the STRIDES initiative, in 2019 NIH has relocated the SRA database to Google and AWS cloud service providers. By moving SRA in the cloud, NIH provides unprecedent access and computational capabilities to the entire SRA compendium, up to 12 PB of data. However, with the costs of data storage and the increasing ability to sequence genomes and metagenomes, NIH is requesting recommendations for SRA and its current activities, future plans, and opportunities.

SRA Data Working Group Charge

The charge of the SRA Data Working Group Council of Councils Working Group is to provide recommendations to the Council on the following key factors for storing and managing SRA data on cloud service provider environments.

As the initial priority, the NIH is requesting the WG to evaluate and identify solutions to maintain efficiencies in the storage footprint of SRA, specifically evaluate the use of Base Quality Scores and format compression strategies. A Draft Report is requested by the January 2020 Council of Council meeting.

Over a longer timeframe the WG may be asked to evaluate and provide recommendations on other issues including but not limited to:

- Analysis of SRA and SRA services
- Technical recommendations on SRA improvements and efficiencies
- Recommendations on data retention, data models and/or data usage
- Vision for future needs or opportunities, as these related to SRA

Discussion and Vote