PRACTICAL STRATEGIES FOR USING SRA LITE AND CLOUD-OPTIMIZED DATA FORMATS

September 10, 2025, 1-3 PM

NIH Building 10, FAES Rm. 1

Derek Caetano-Anollés

National Library of Medicine
National Center for Biotechnology Information

TODAY'S SESSION

- Sequencing Data
- Sequence Read Archive
- SRA File Formats
- Using the SRA Toolkit

WORKSHOP MATERIALS

github.com/**ncbi/workshop-sra-hands-on**

CLICK HERE

Type it.

Or scan it.

Or click it.

WHAT'S THE DEAL WITH

DATA?

NIH National Library of Medicine
National Center for Biotechnology Information

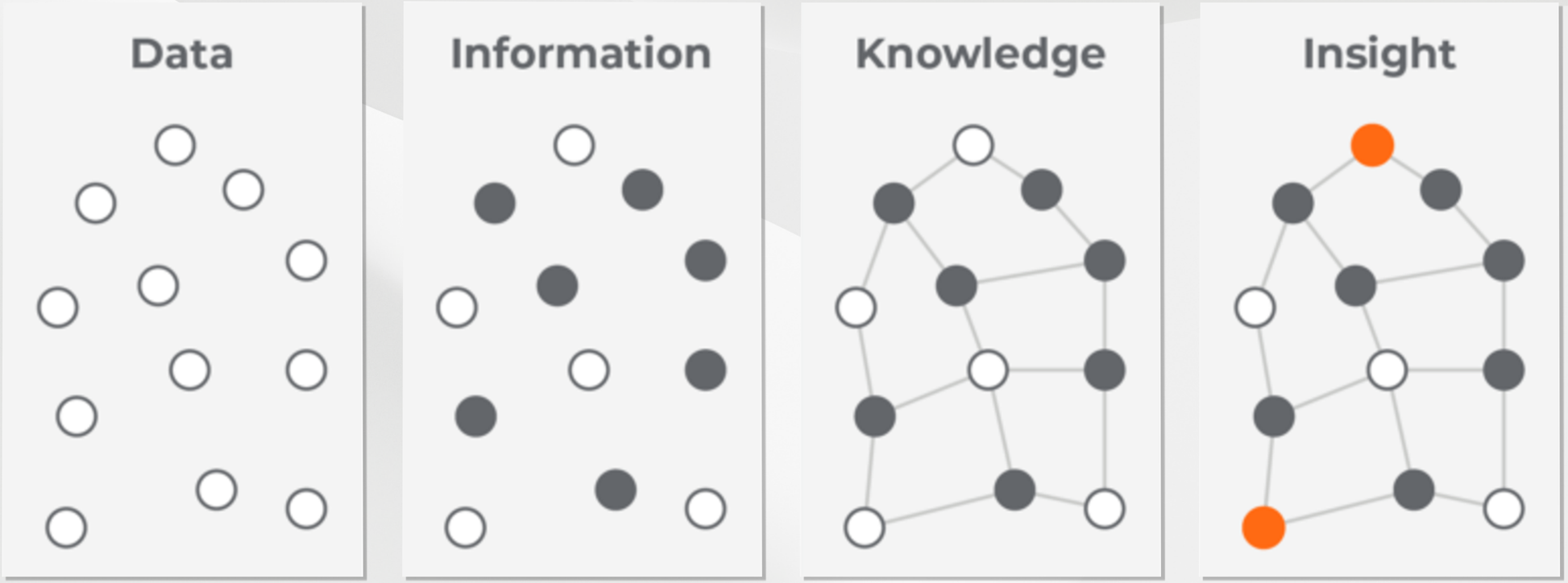


Image Credit: Welocalize, 2020

bio·in·for·mat·ics *noun*

the collection, classification, storage, and analysis of biochemical and biological information using computers, especially as applied to molecular genetics and genomics

THE PROBLEM

This negatively affects:

- Findability
- Storage
- Costs

SEQUENCE READ ARCHIVE

From genetic data
to actionable knowledge...

...and to a fundamental
understanding of biology

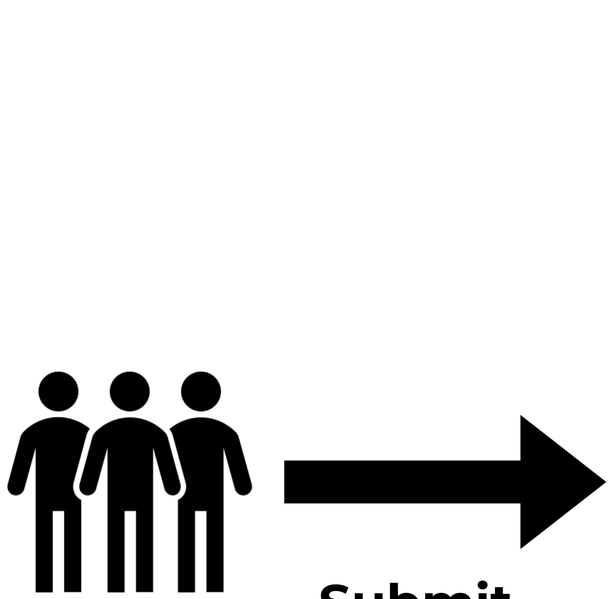
National Library of Medicine
National Center for Biotechnology Information

THE SEQUENCE READ ARCHIVE

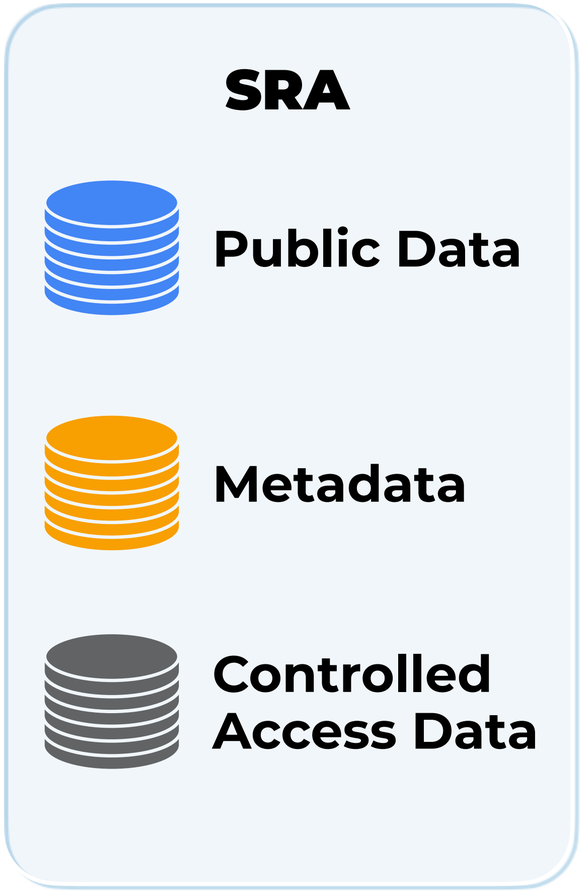
WHAT IS IT?

National Library of Medicine
National Center for Biotechnology Information

Sequence Read Archive



Submit Data

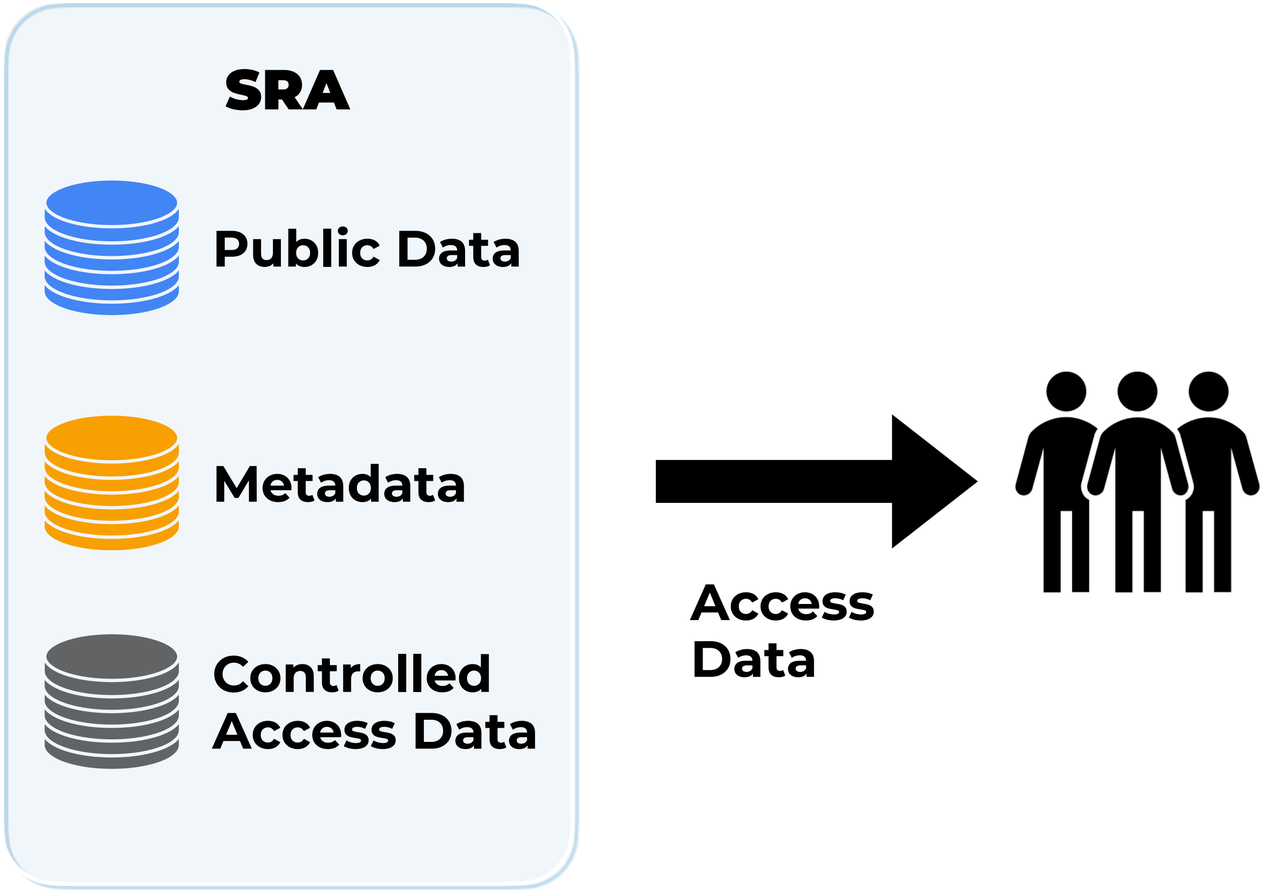


The SRA is a comprehensive, community-driven sequence repository

Sequence Read Archive

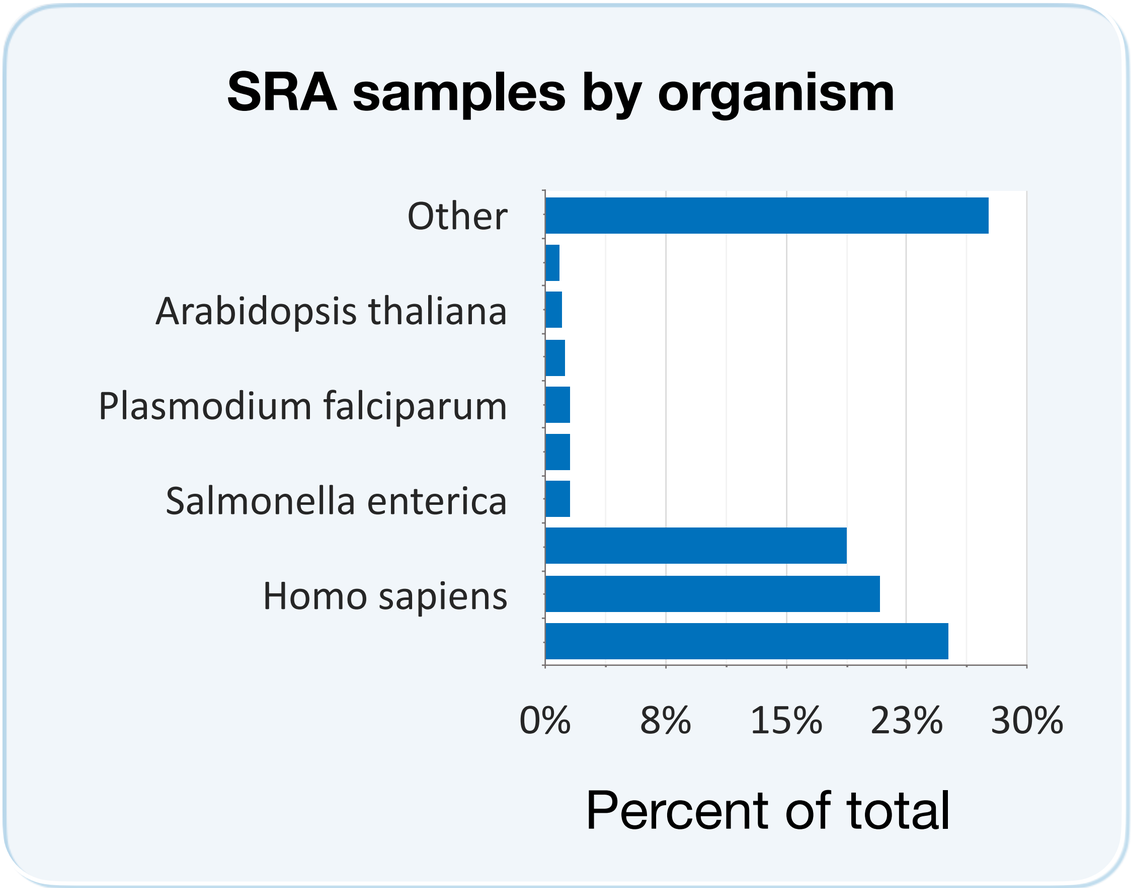
SRA supports
FAIR data principles

F indable
A ccessible
I nteroperable
R eusable

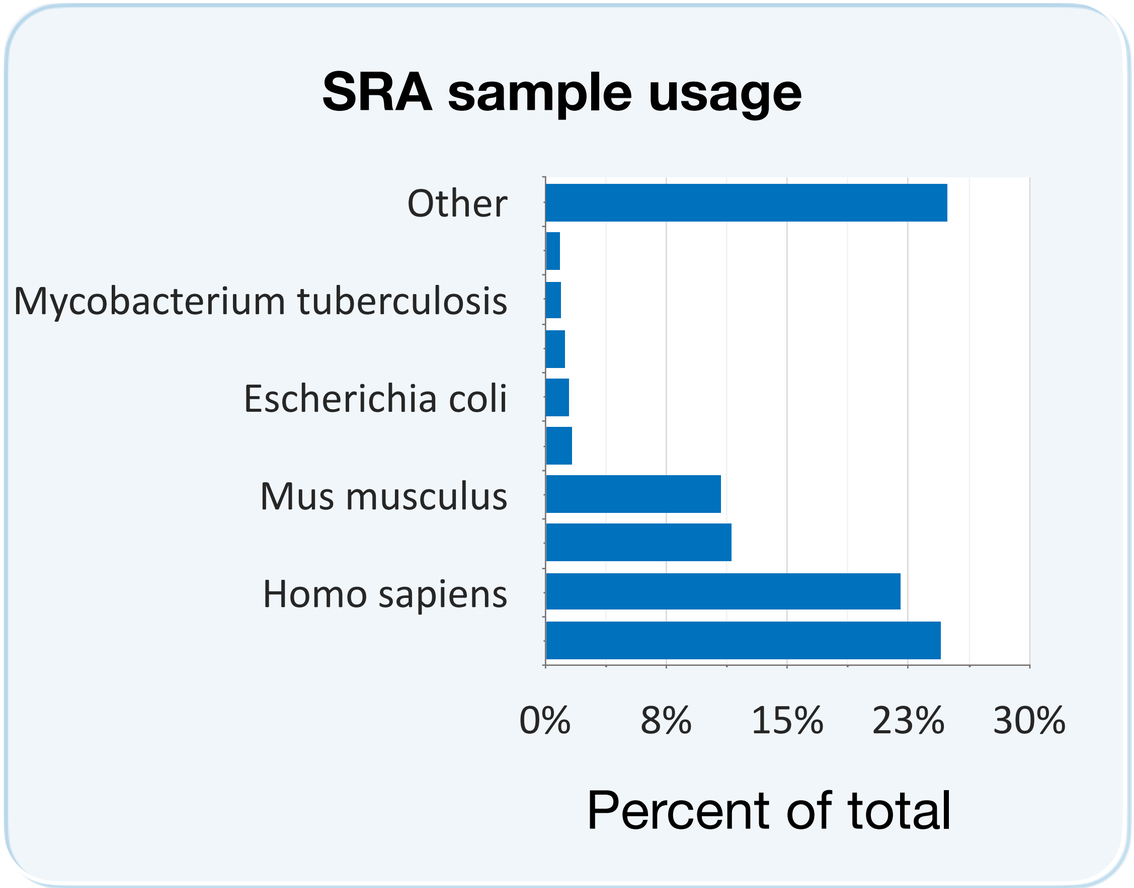


Sequence Read Archive

The SRA maintains a diverse dataset derived from all the kingdoms of life



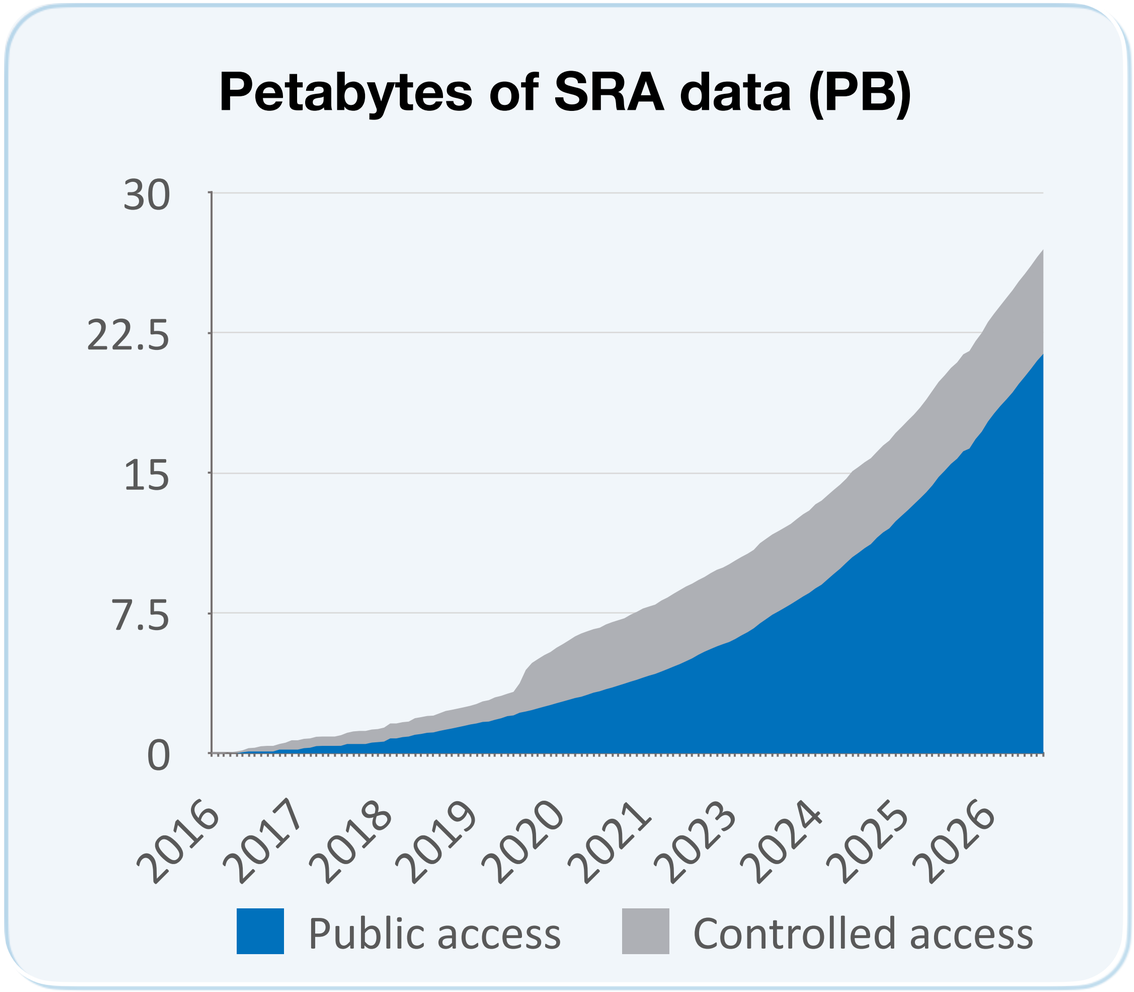
Sequence Read Archive



The SRA dataset is used across a diverse set of scientific use cases

Public Data, statistics collected 11/2019 through 03/2022

Sequence Read Archive



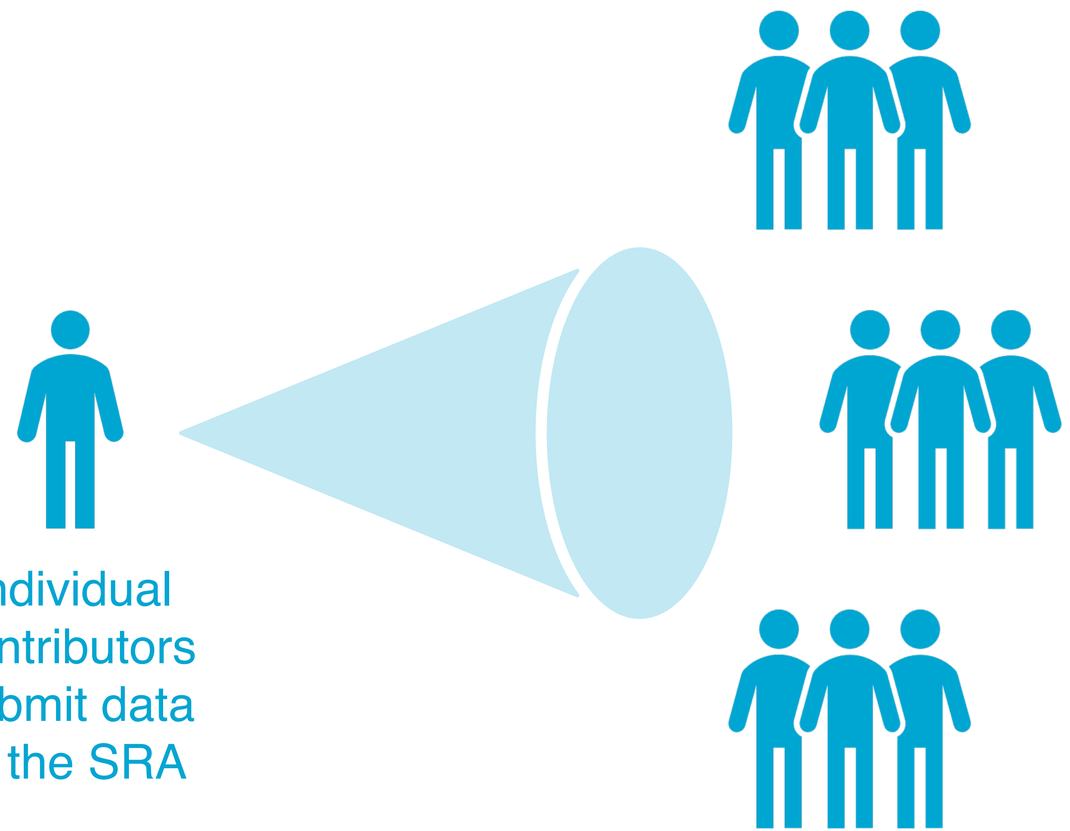
35 million sequence datasets

160+ Petabytes of sequence data available in three formats

Metadata describing biological samples and library prep

Sequence Read Archive

The SRA amplifies the impact of research investments



Individual contributors submit data to the SRA

Research communities reuse the data

Sequence Read Archive

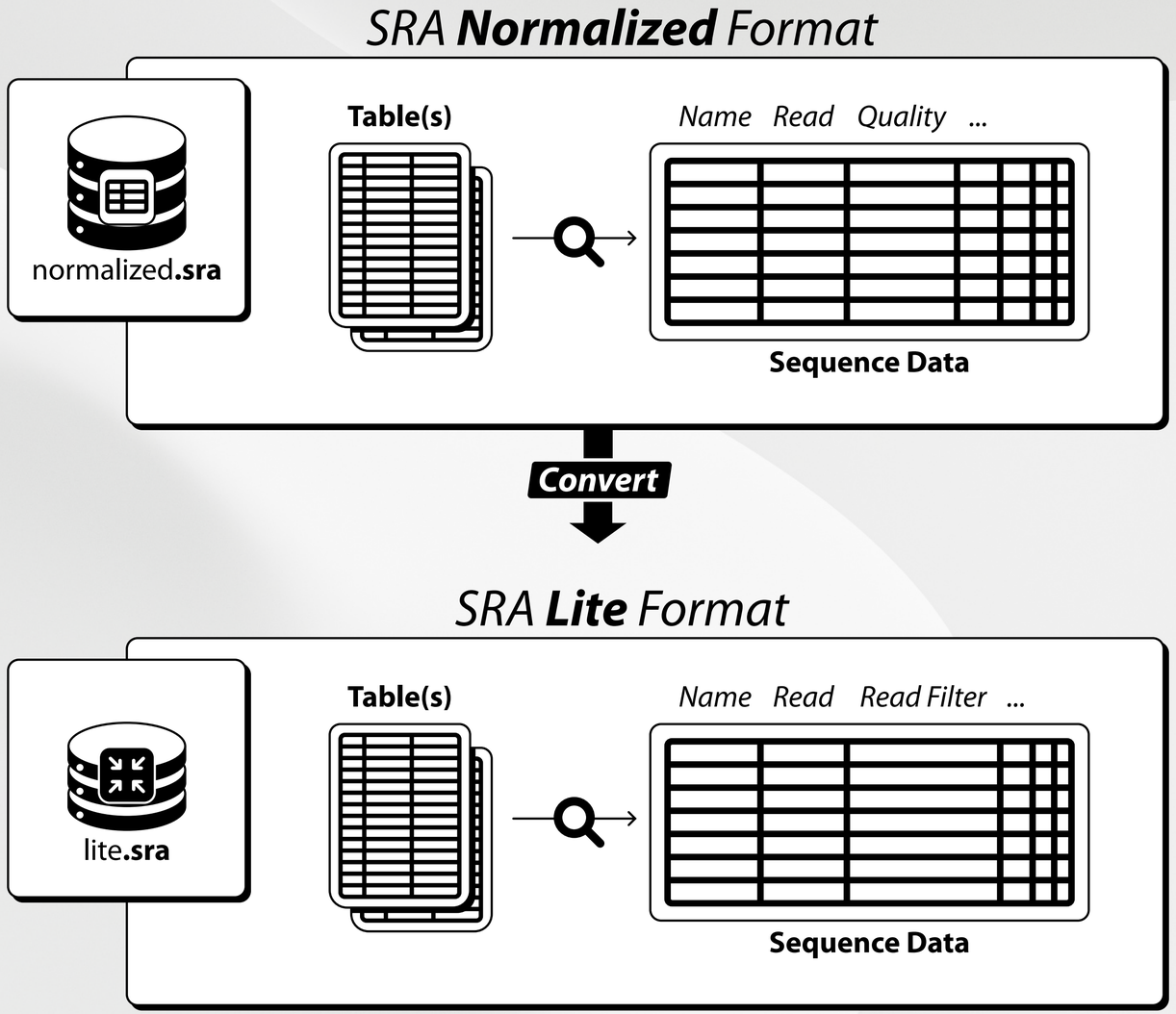
“Turns out, sequencing my genome was cheaper than storing it!”

THE SEQUENCE READ ARCHIVE

FILE FORMATS

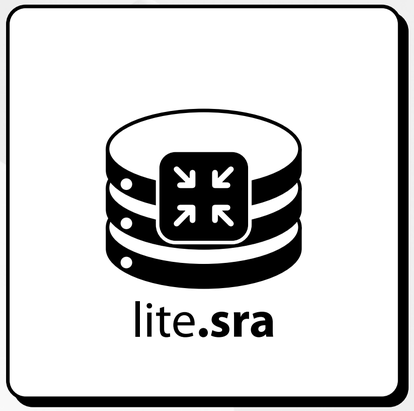
NIH National Library of Medicine
National Center for Biotechnology Information

SRA Lite Format



COMPARISON

VS.



Bigger and slower

Complete representation of the experiment

Smaller and faster

Includes Sequencing reads + base qualities.

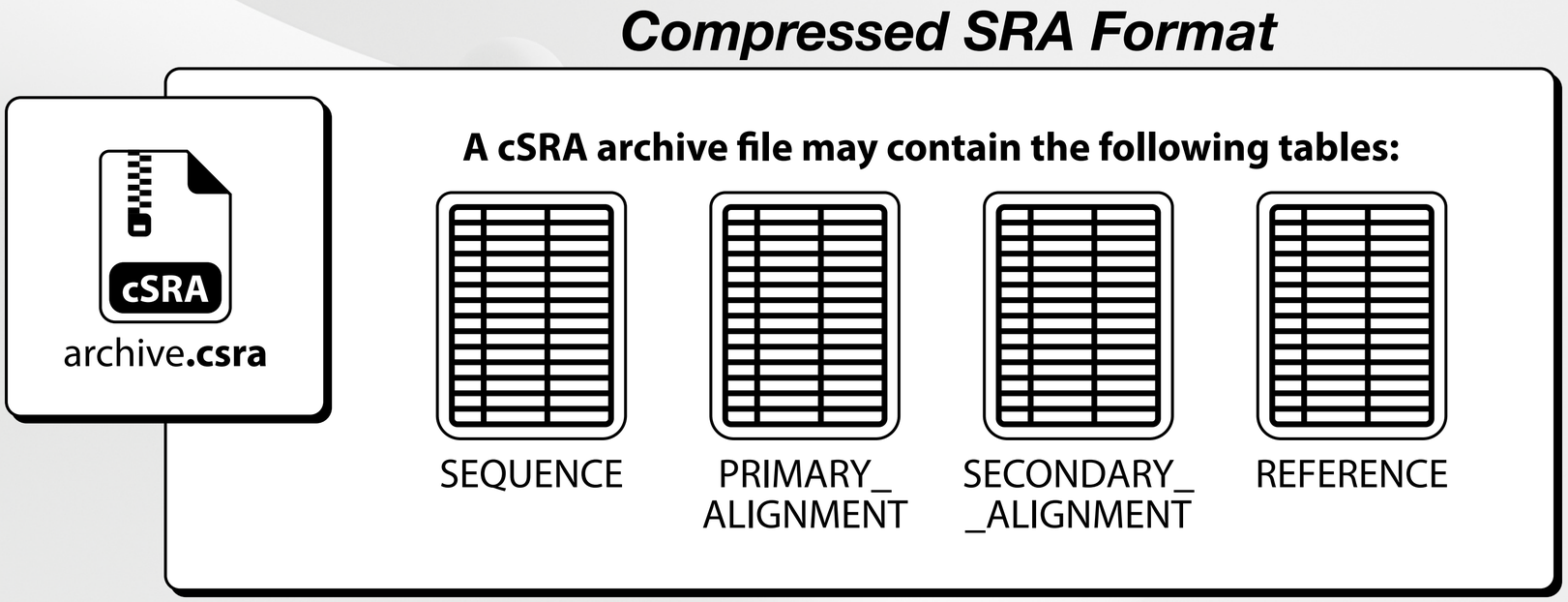
DATA DISTRIBUTION MODEL

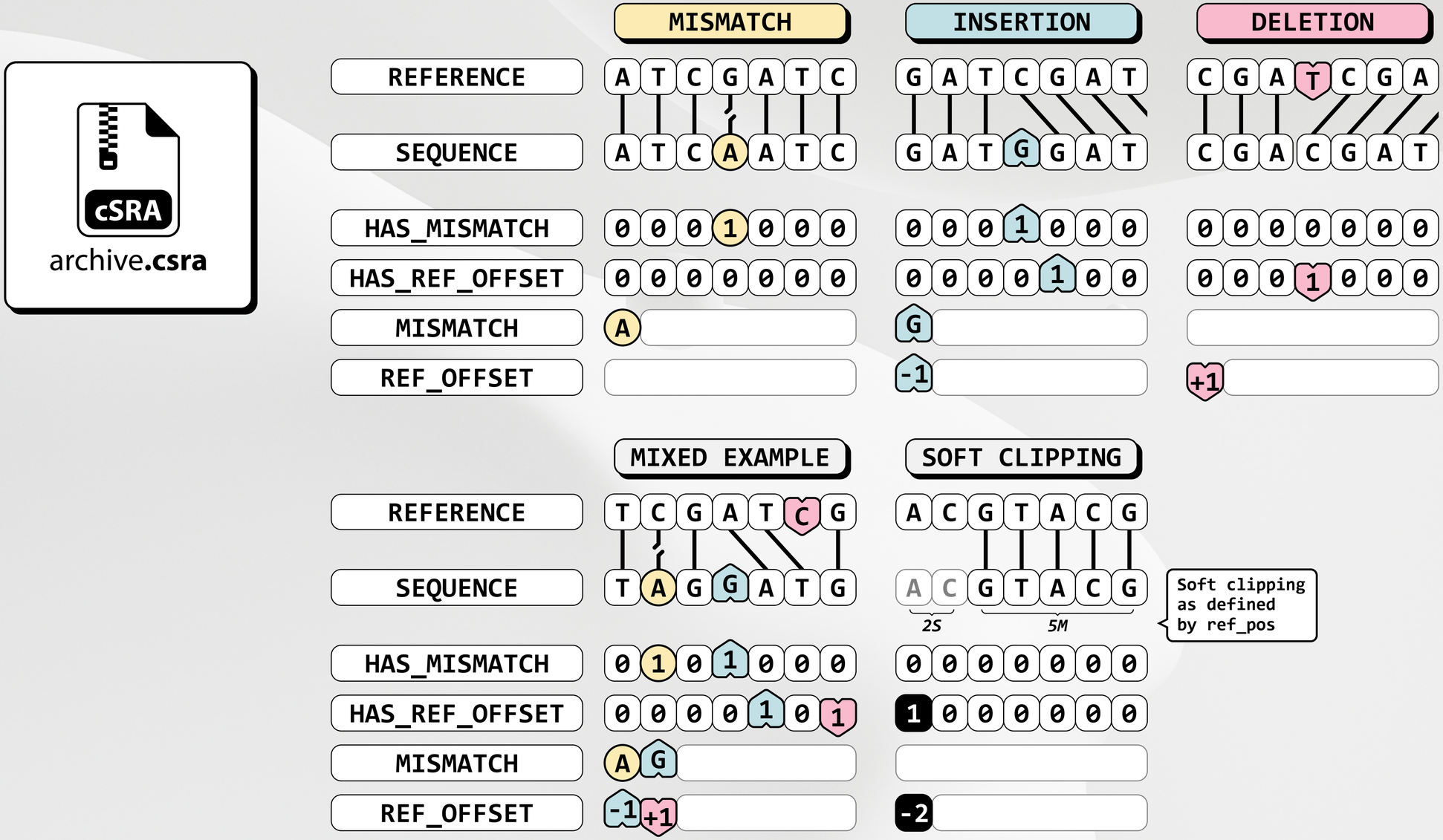
	Amazon Web Services			Google Cloud Platform		NCBI NLM
	Open Data Platform	Glacier (Instant Retrieval)	Glacier (Deep Archive)	Public Dataset Program	Coldline	On-Prem
SRA Lite		aws				NIH
SRA Normalized	aws					
Original (Submission)			aws			
VCF (COVID-19)	aws					

Open Access

At Cost

cSRA Format

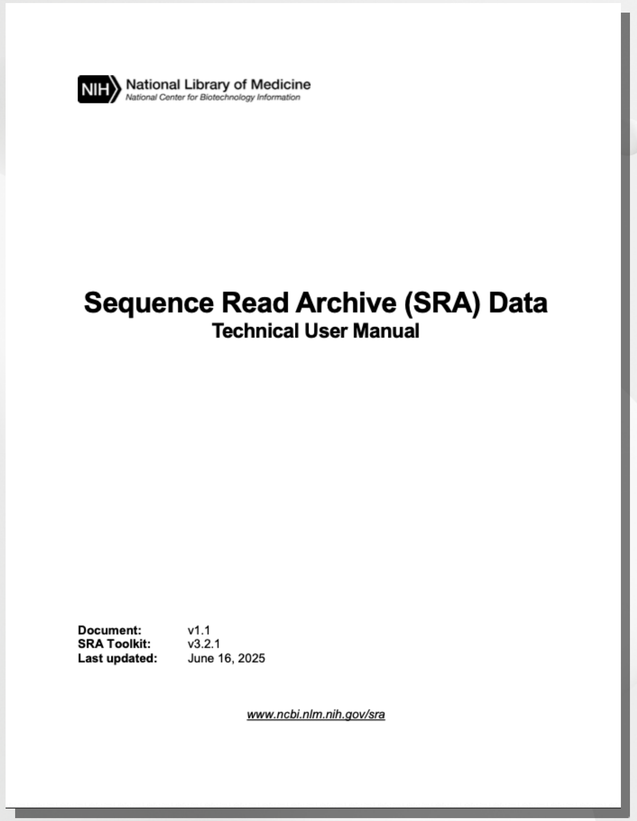

Compressed SRA Format
cSRA
archive.csra
A cSRA archive file may contain the following tables:
SEQUENCE
PRIMARY_ ALIGNMENT
SECONDARY_ _ALIGNMENT
REFERENCE


cSRA
archive.csra
MISMATCH
INSERTION
DELETION
REFERENCE
SEQUENCE
HAS_MISMATCH
HAS_REF_OFFSET
MISMATCH
REF_OFFSET
MIXED EXAMPLE
SOFT CLIPPING
2S
5M
Soft clipping as defined by ref_pos

GET MORE INFO

SRA Data Formats Manual

doi.org/10.5281/zenodo.15677383



NIH National Library of Medicine
National Center for Biotechnology Information

Sequence Read Archive (SRA) Data
Technical User Manual

Document: v1.1
SRA Toolkit: v3.2.1
Last updated: June 16, 2025

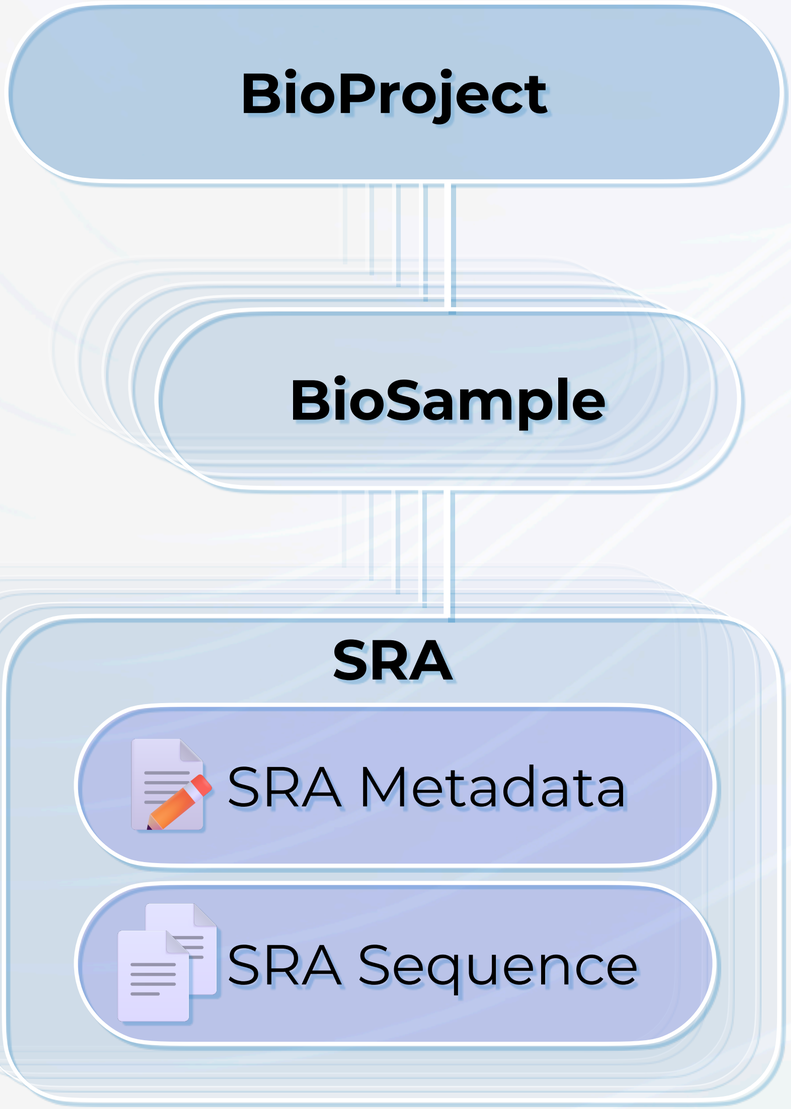
www.ncbi.nlm.nih.gov/sra

SEQUENCE READ ARCHIVE

RECORDS

NIH National Library of Medicine
National Center for Biotechnology Information

SRA SUBMISSION



Includes project goal and scope, including participants and grants

One BioProject can have multiple BioSamples

Biological specimen(s) where the sequence comes from

One specimen can have multiple sequences

Sequencing platform and library preparation

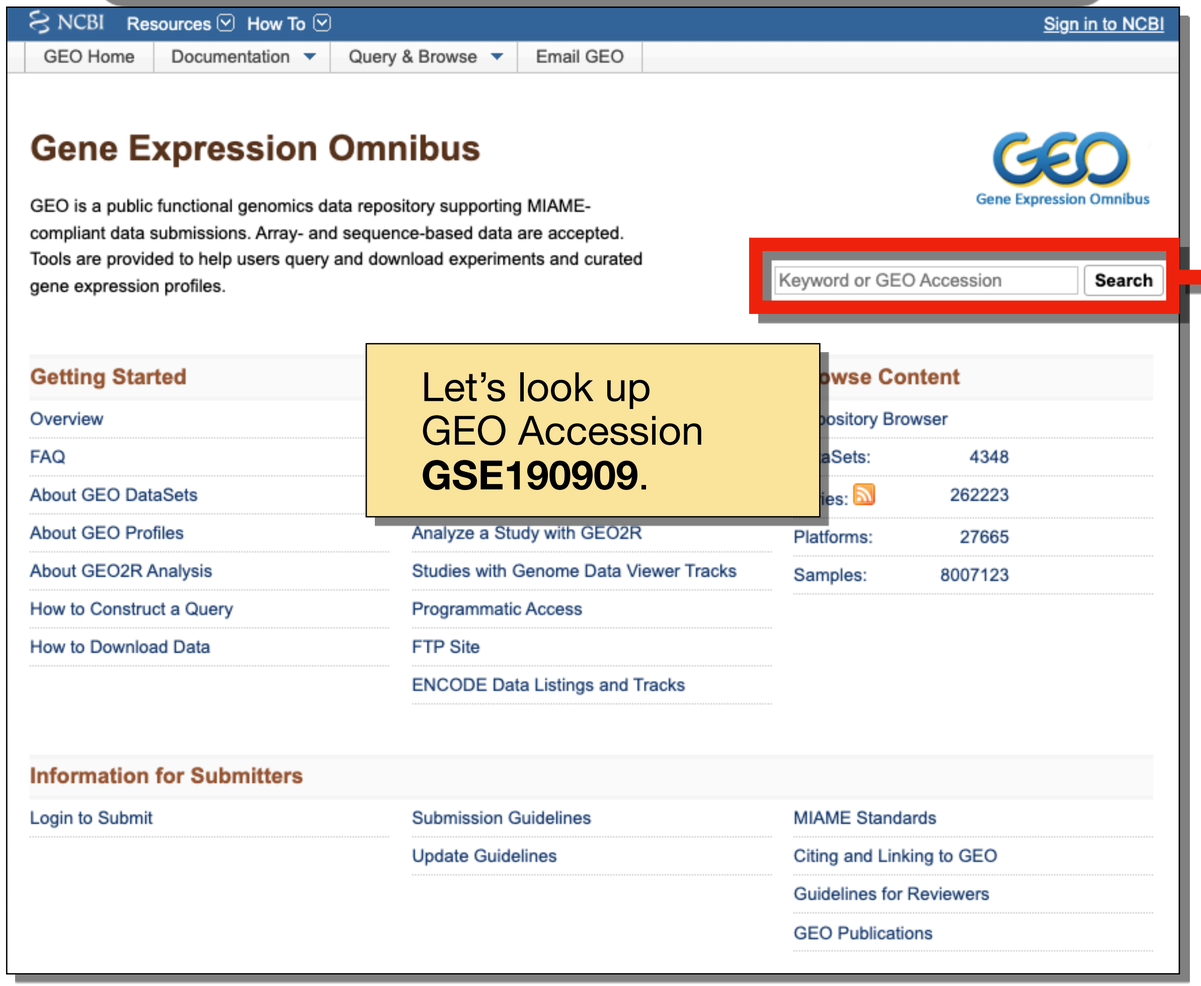
The individual sequencing files

SRA RECORDS

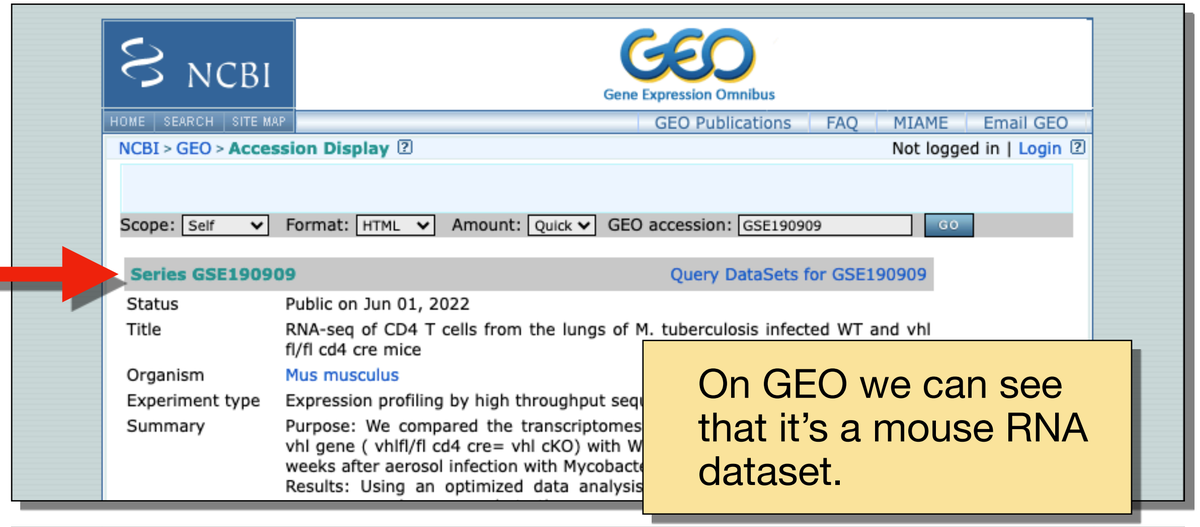
Project	SRP000000	*The research project or study*
Sample	SRS000000	*Sample info, organism*
Experiment	SRX000000	*Sequencer, library preparation*
Run	SRR000000	*Individual sequencing files*

Searching up records

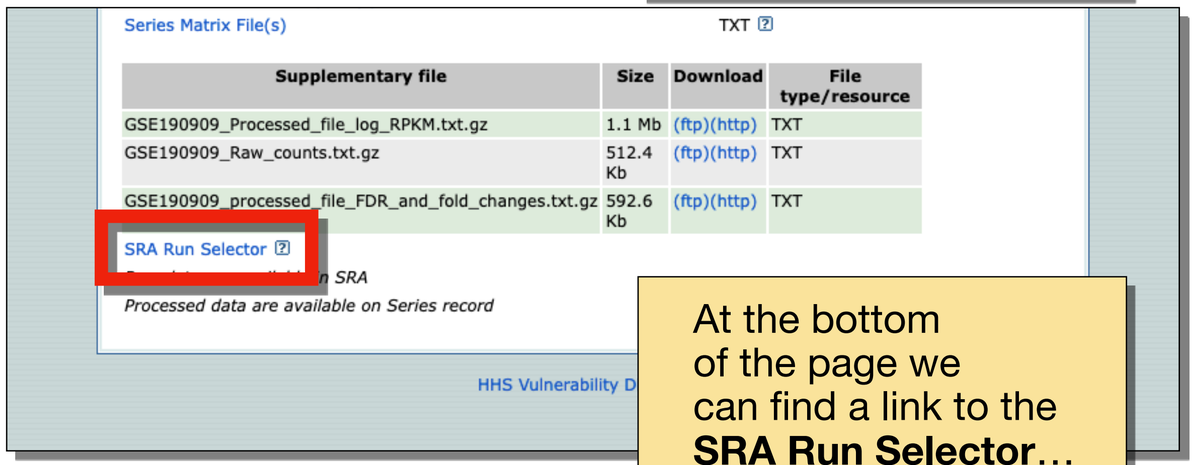
ncbi.nlm.nih.gov/geo



Let's look up GEO Accession **GSE190909**.



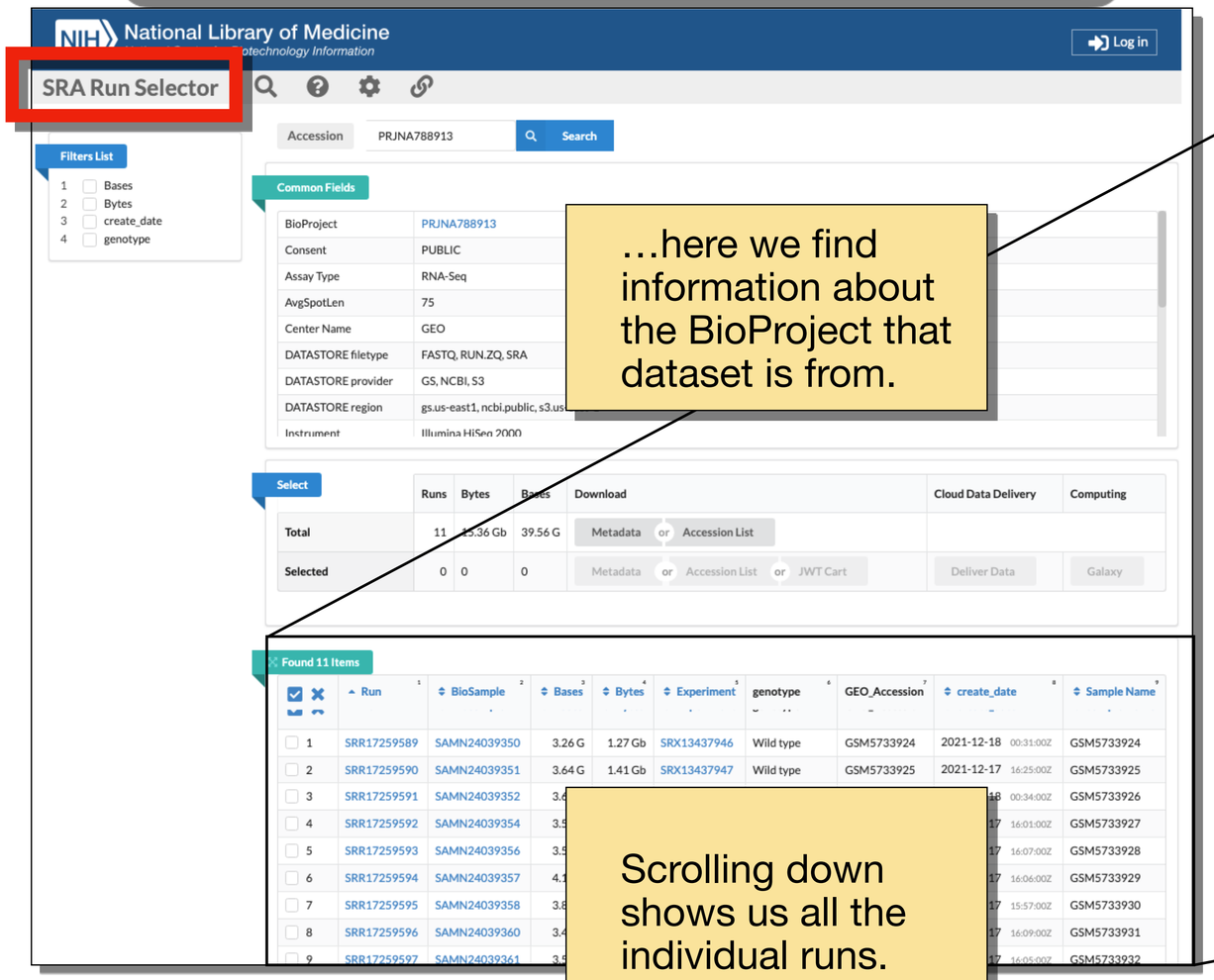
On GEO we can see that it's a mouse RNA dataset.



At the bottom of the page we can find a link to the **SRA Run Selector**…

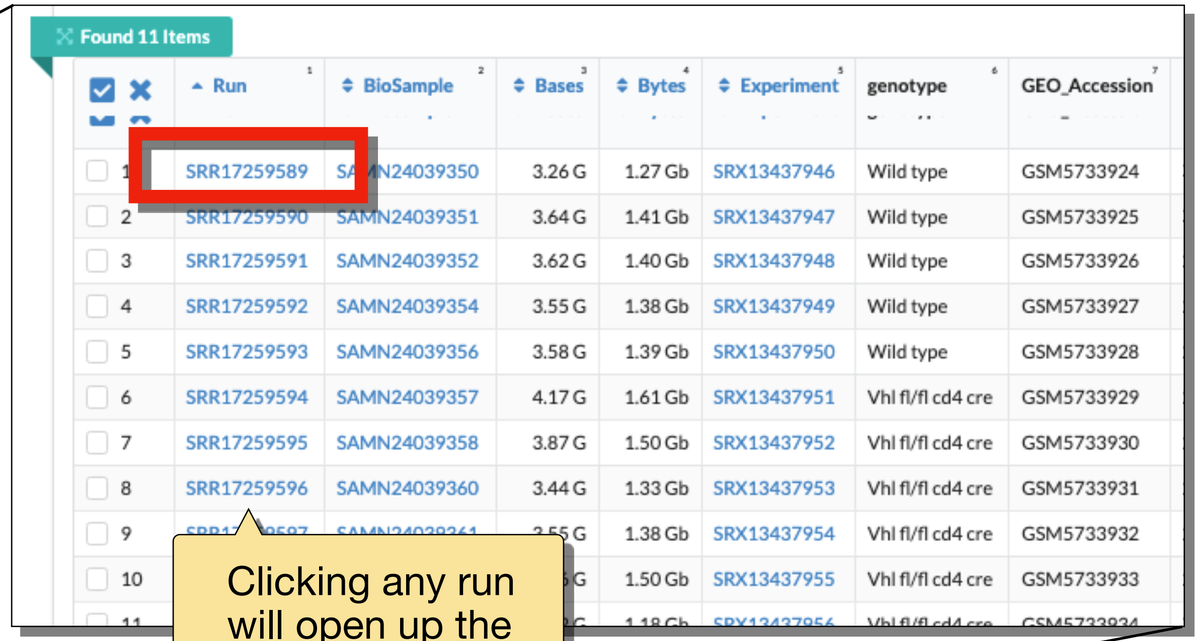
Searching up records

ncbi.nlm.nih.gov/Traces/study/?acc=PRJNA788913



…here we find information about the BioProject that dataset is from.

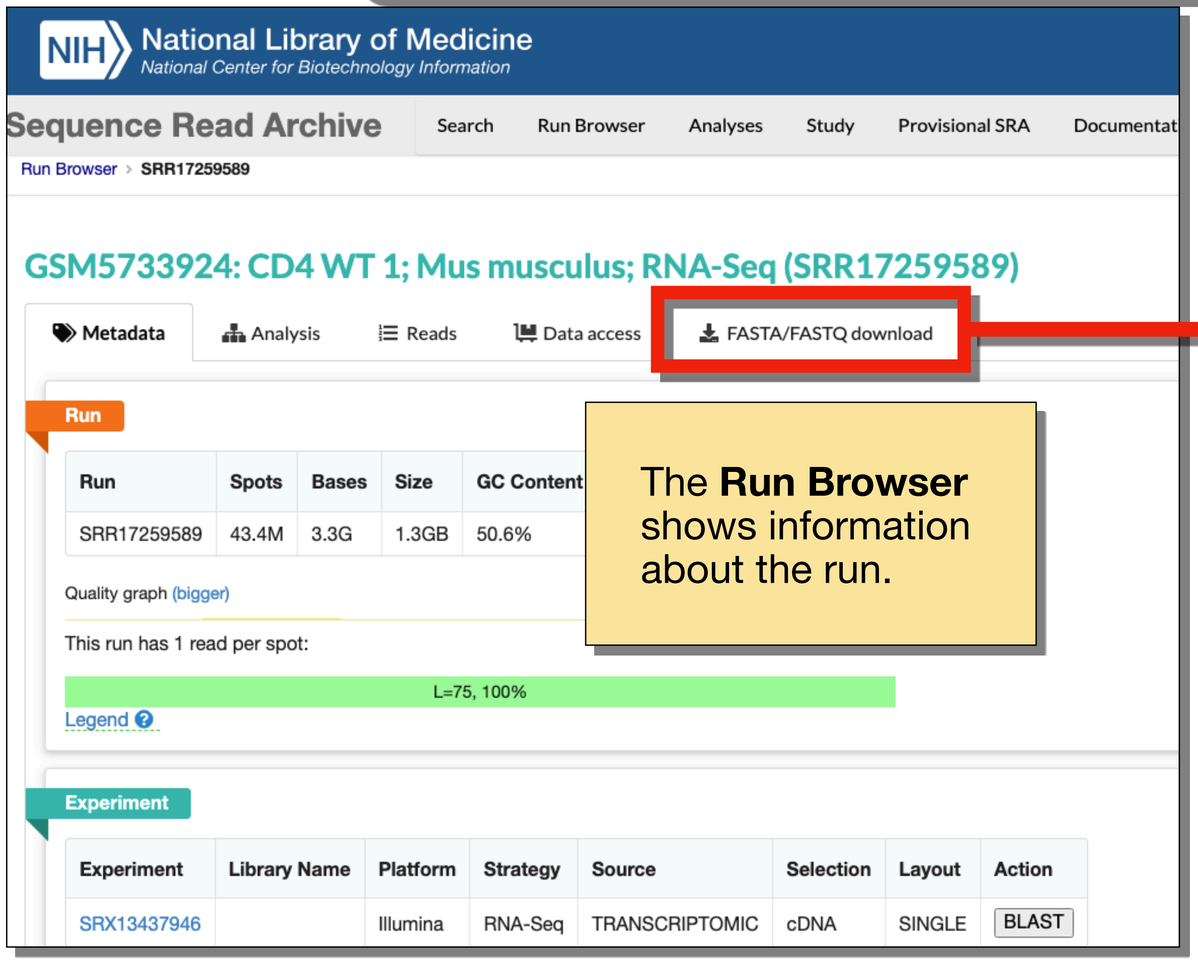
Scrolling down shows us all the individual runs.



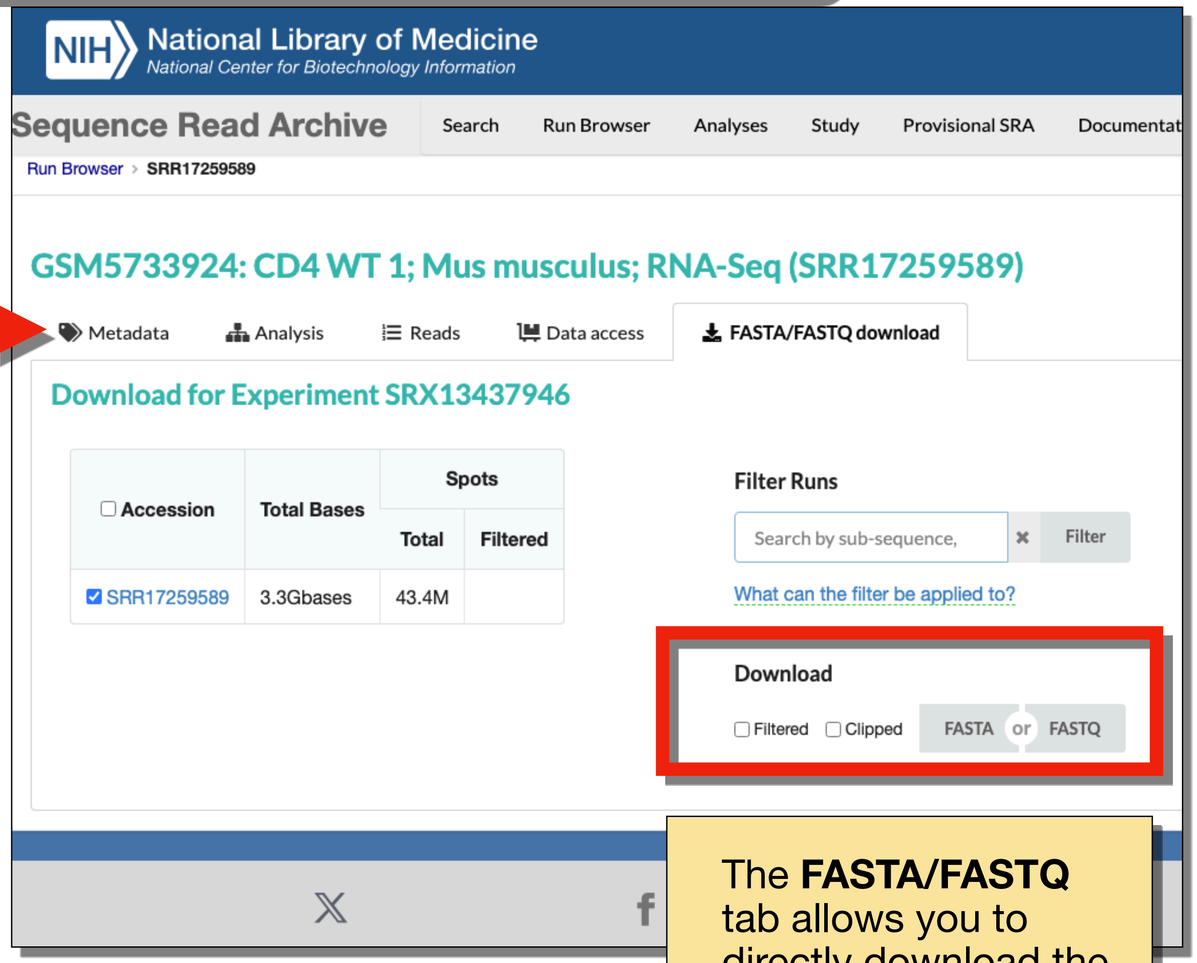
Clicking any run will open up the **Run Browser** for that run.

Searching up records

trace.ncbi.nlm.nih.gov/Traces/index.html?view=run_browser&acc=SRR17259589&display=download



The **Run Browser** shows information about the run.



The **FASTA/FASTQ** tab allows you to directly download the FASTA/FASTQ files.

Searching up records

There's a problem…

Found 11 Items

	Run	BioSample	Bases	Bytes	Experiment	genotype	GEO_Accession	create_date	Sample Name
1	SRR17259589	SAMN24039350	3.26 G	1.27 Gb	SRX13437946	Wild type	GSM5733924	2021-12-18 00:31:00Z	GSM5733924
	SRR17259590	SAMN24039351	3.64 G	1.41 Gb	SRX13437947	Wild type	GSM5733925	2021-12-17 16:25:00Z	GSM5733925
	SRR17259591	SAMN24039352	3.62 G				GSM5733926	2021-12-18 00:34:00Z	GSM5733926
	SRR17259592	SAMN24039354	3.55 G				GSM5733927	2021-12-17 16:01:00Z	GSM5733927
	SRR17259593	SAMN24039356	3.58 G				GSM5733928	2021-12-17 16:07:00Z	GSM5
	SRR17259594	SAMN24039357	4.17 G				GSM5733929	2021-12-17 16:06:00Z	GSM5
7	SRR17259595	SAMN24039358	3.87 G	1.50 Gb	SRX13437952	Vhl fl/fl cd4 cre	GSM5733930	2021-12-17 15:57:00Z	GSM5733930
8	SRR17259596	SAMN24039360	3.44 G	1.33 Gb	SRX13437953	Vhl fl/fl cd4 cre	GSM5733931	2021-12-17 16:09:00Z	GSM5733931
9	SRR17259597	SAMN24039361	3.55 G	1.38 Gb	SRX13437954	Vhl fl/fl cd4 cre	GSM5733932	2021-12-17 16:05:00Z	GSM5733932
10	SRR17259598	SAMN24039362	3.86 G	1.50 Gb	SRX13437955	Vhl fl/fl cd4 cre	GSM5733933	2021-12-17 16:07:00Z	GSM5733933

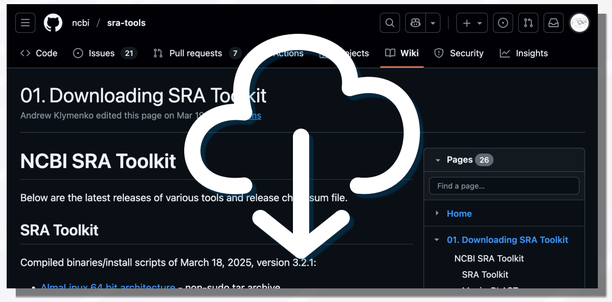
Downloading run files for ***10 runs*** might not be difficult…

…but what about projects with ***100 runs?***

…1,000 runs?

SRATOOLKIT

github.com/**ncbi/sra-tools**



Open source software to access, download, and work with SRA data.