

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



TODAY'S SESSION

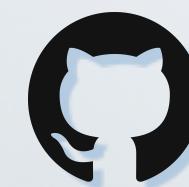
**Sequencing
Data**

**Sequence
Read Archive**

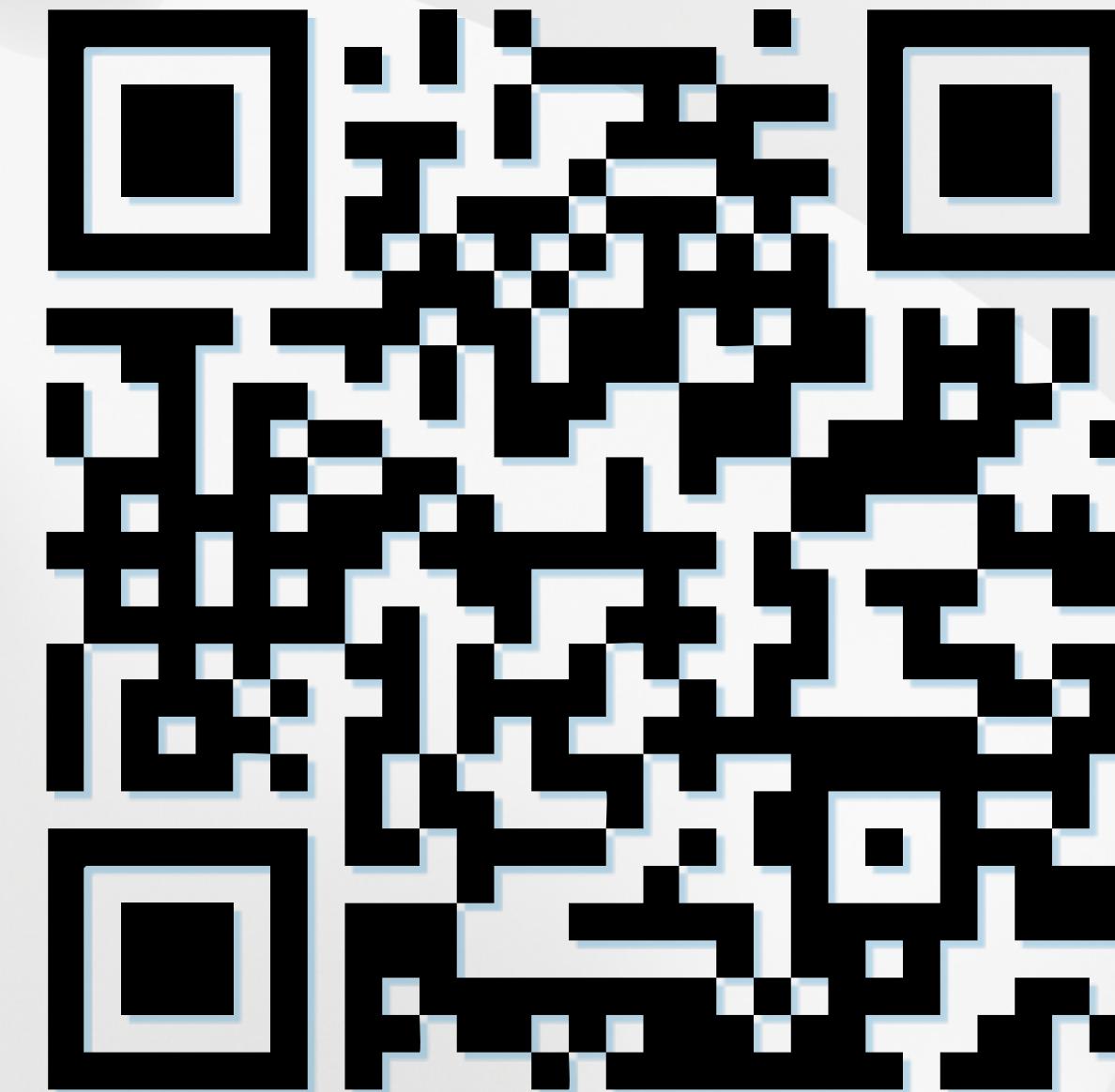
**SRA File
Formats**

**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?**



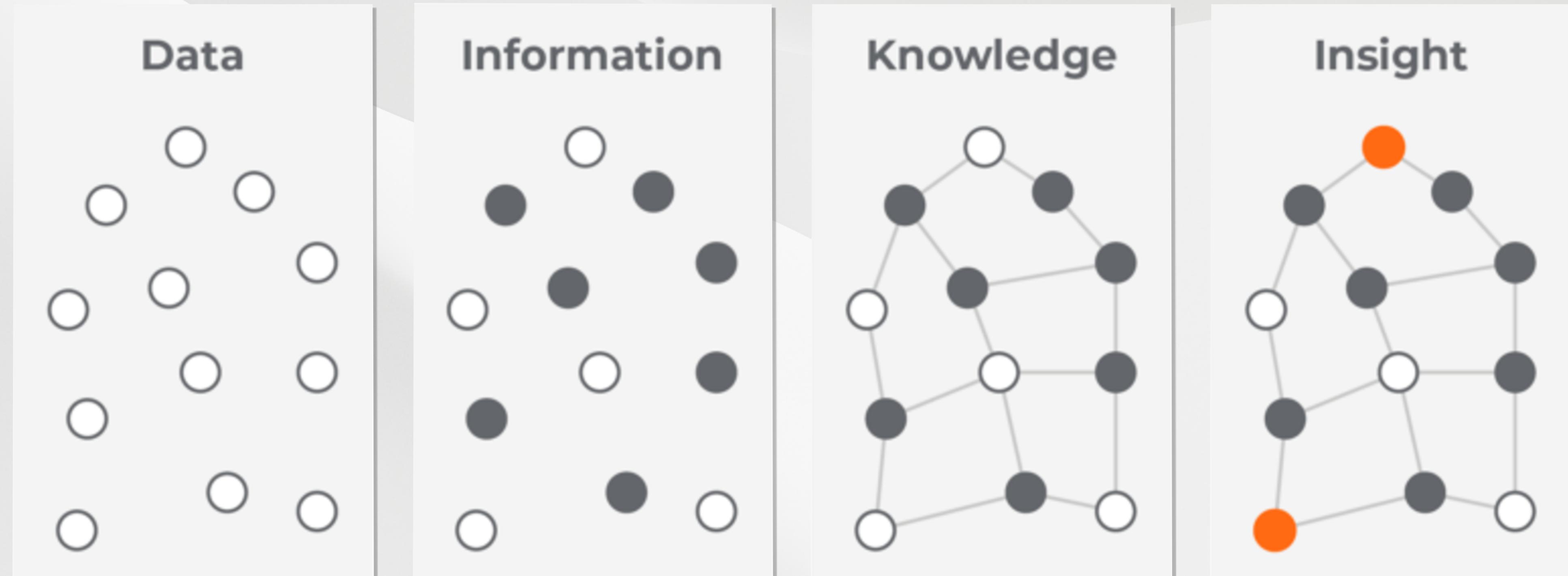
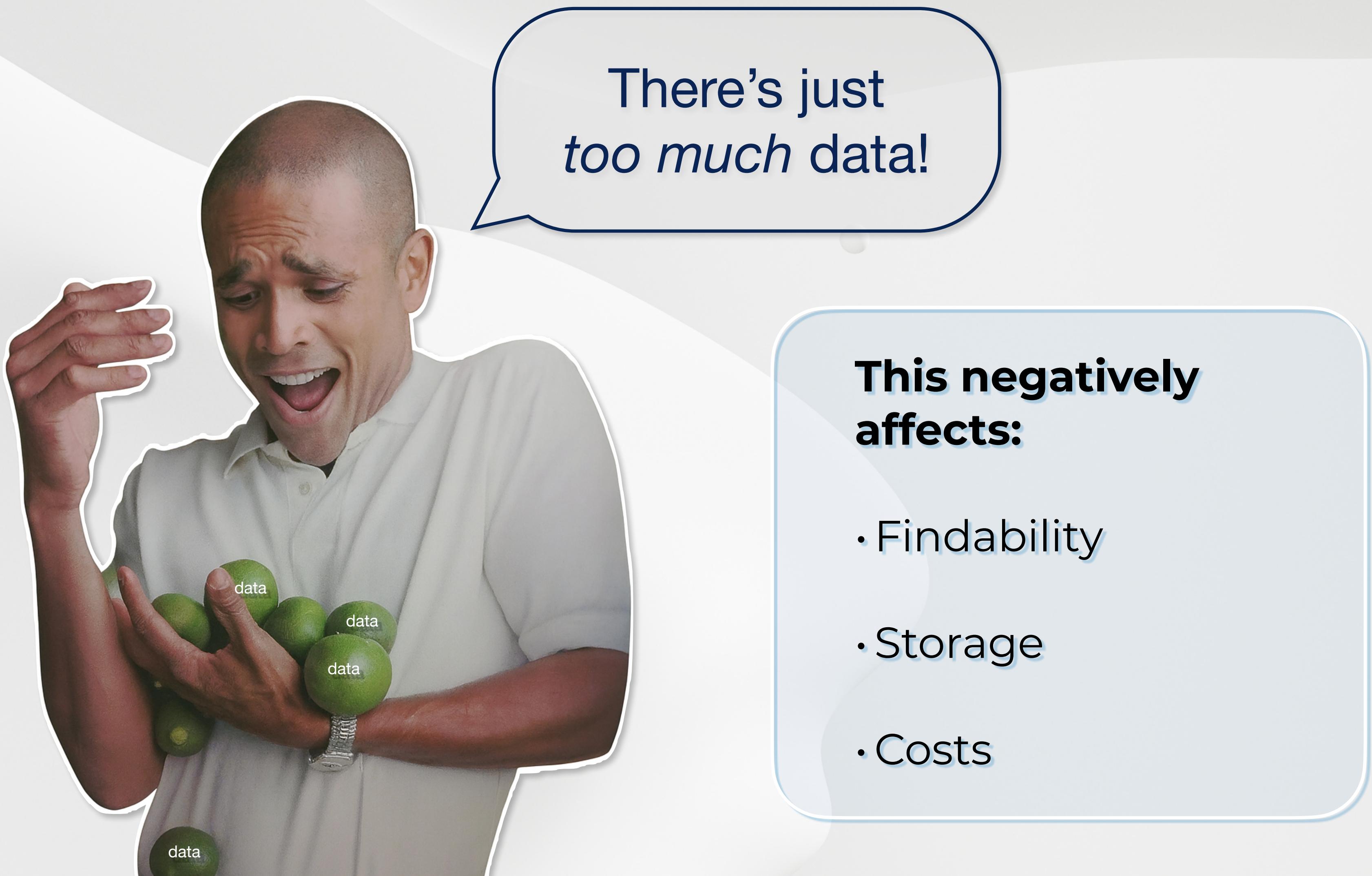


Image Credit: Welocalize, 2020

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



SEQUENCE READ ARCHIVE

From genetic data
to actionable knowledge...

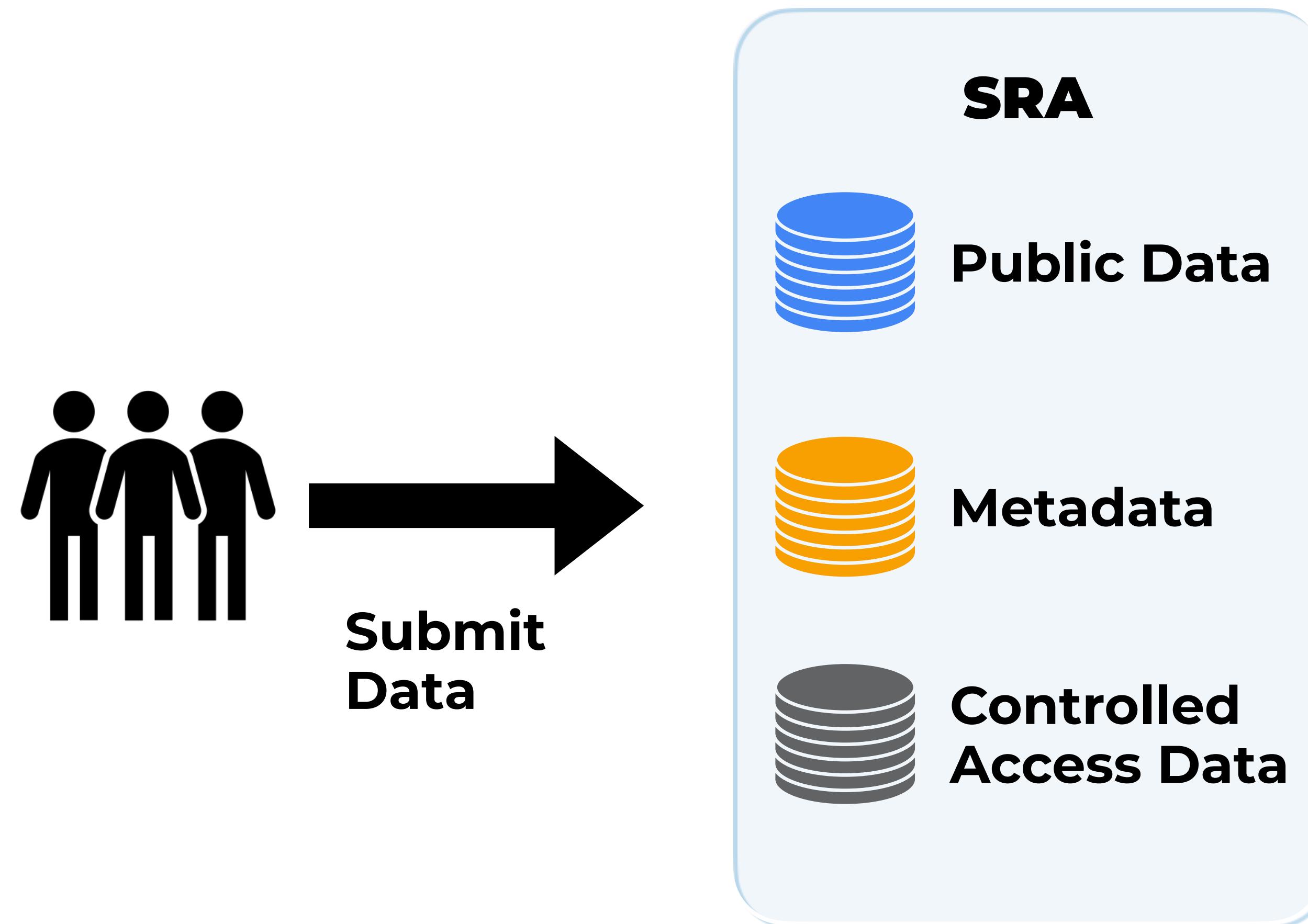
...and to a fundamental
understanding of biology

THE SEQUENCE READ ARCHIVE

WHAT IS IT?



Sequence Read Archive

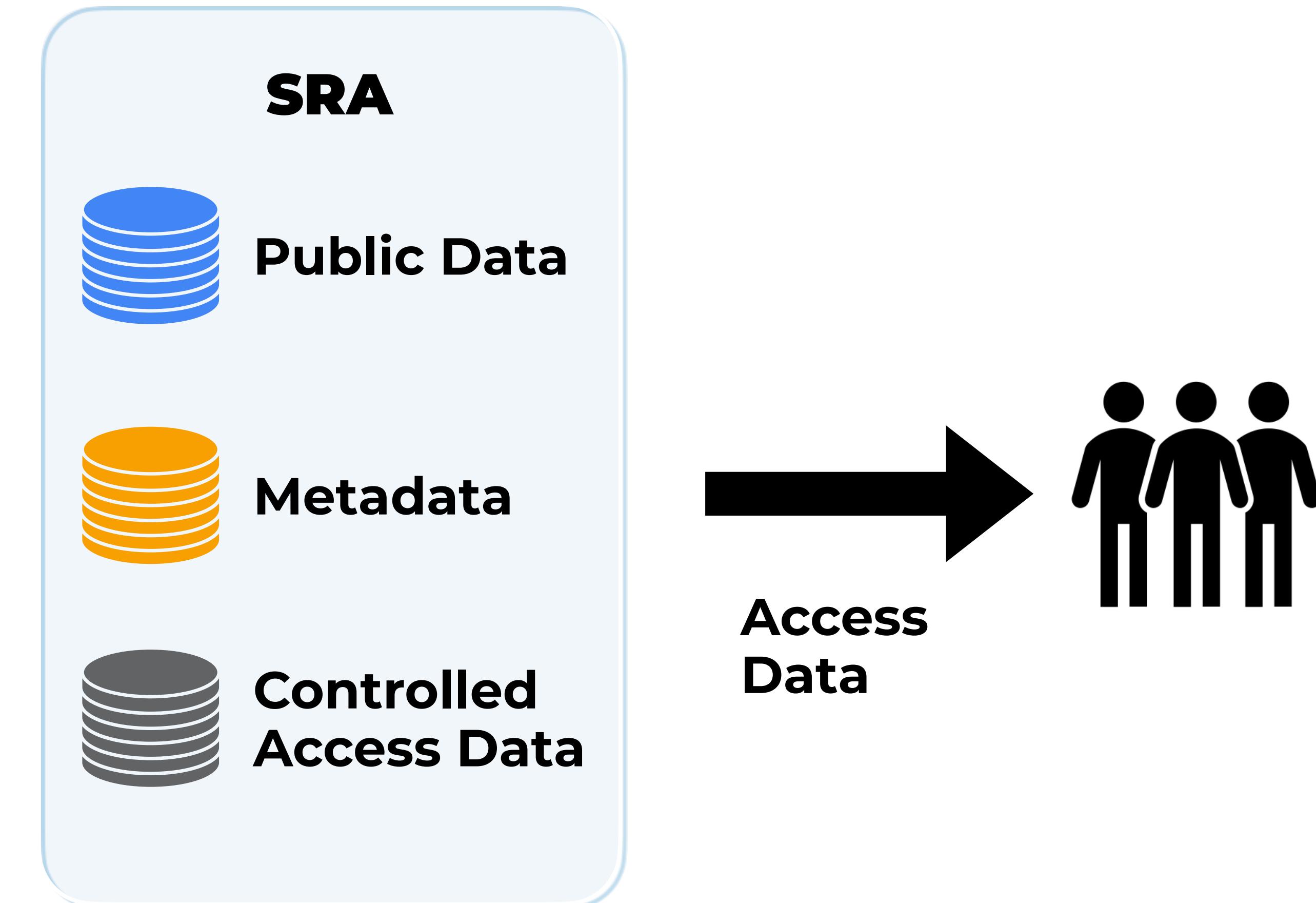


The SRA is a comprehensive, community-driven sequence repository

Sequence Read Archive

SRA supports
FAIR data principles

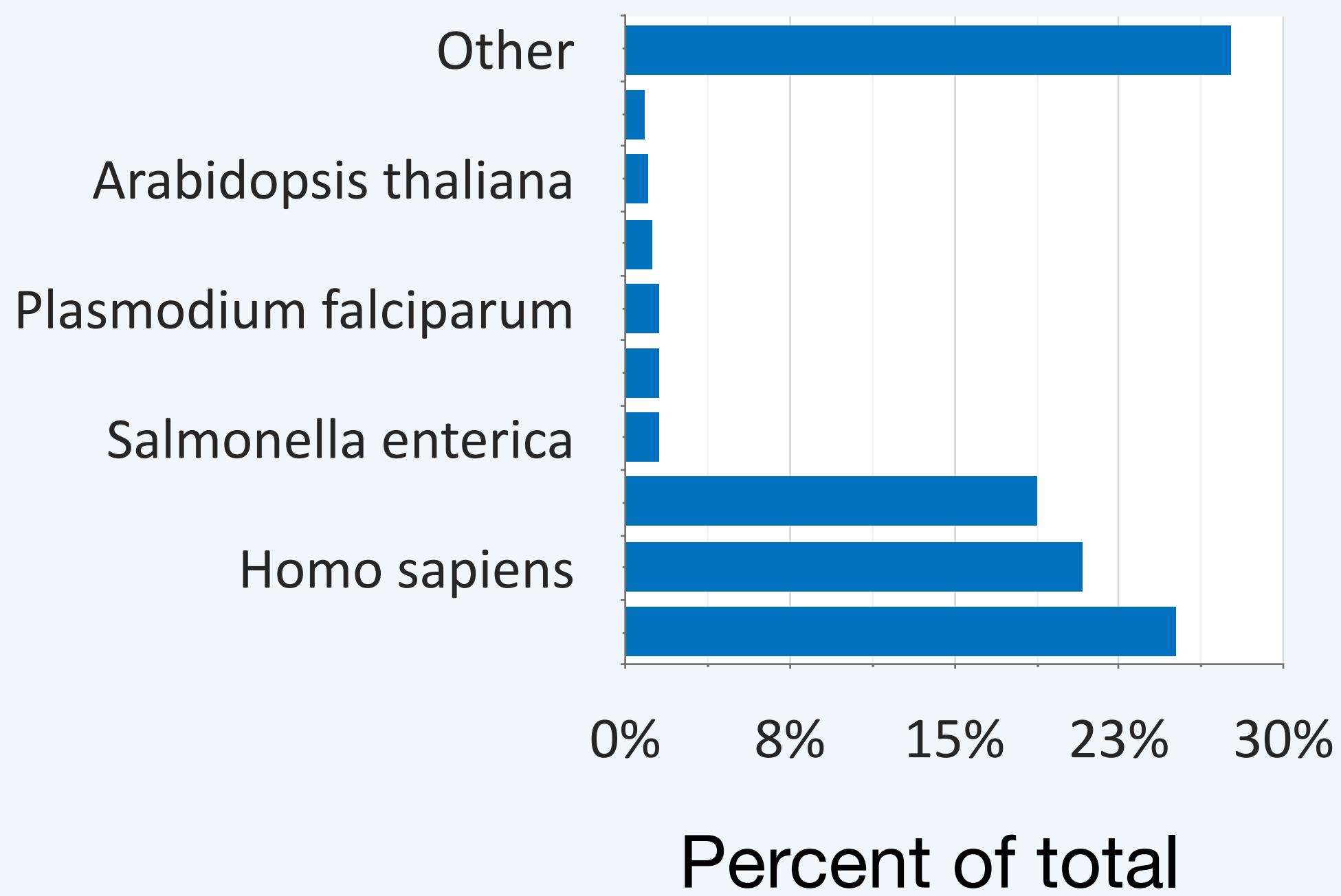
F indable
A ccessible
I nteroperable
R eusable



Sequence Read Archive

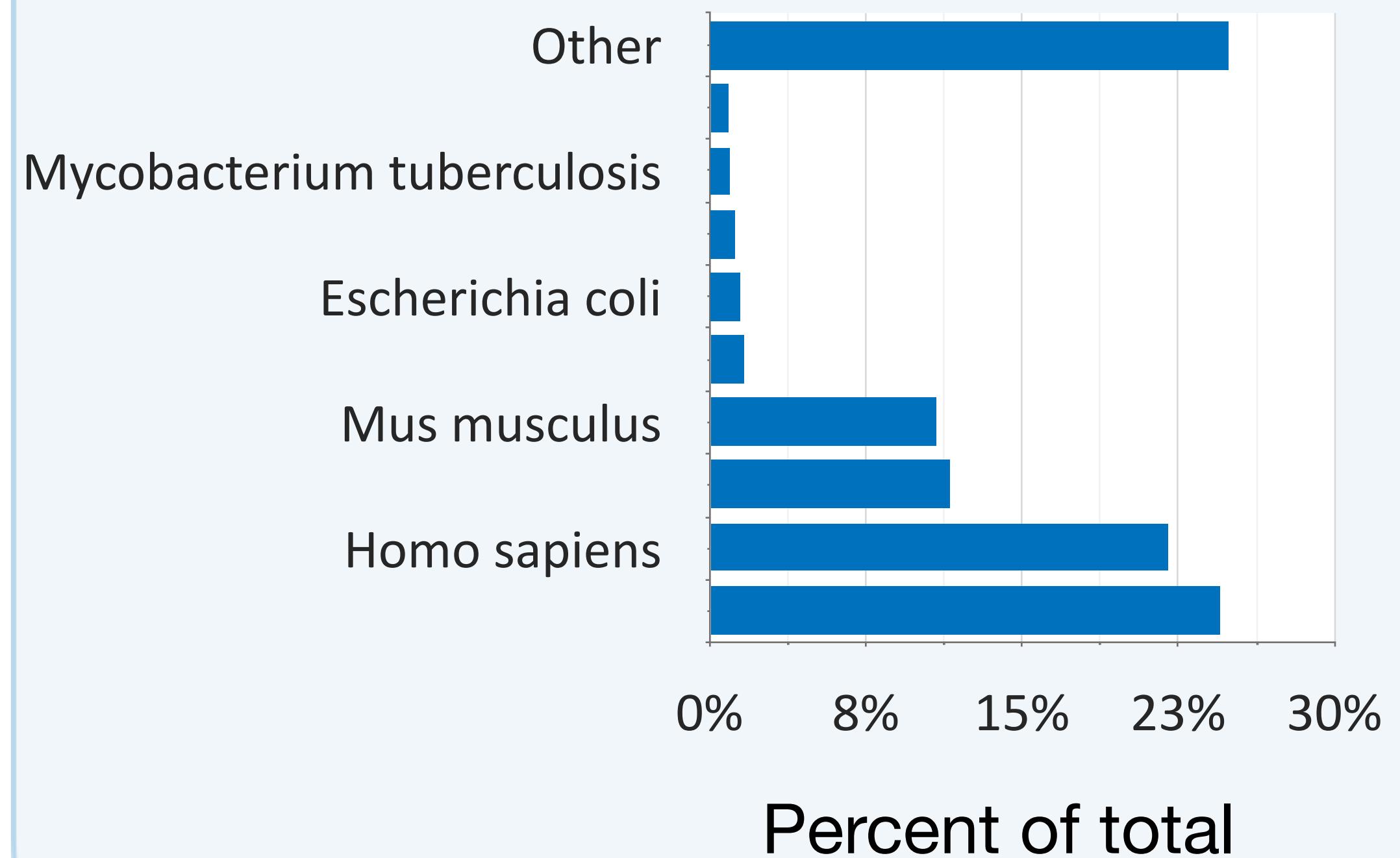
The SRA maintains a diverse dataset derived from all the kingdoms of life

SRA samples by organism



Sequence Read Archive

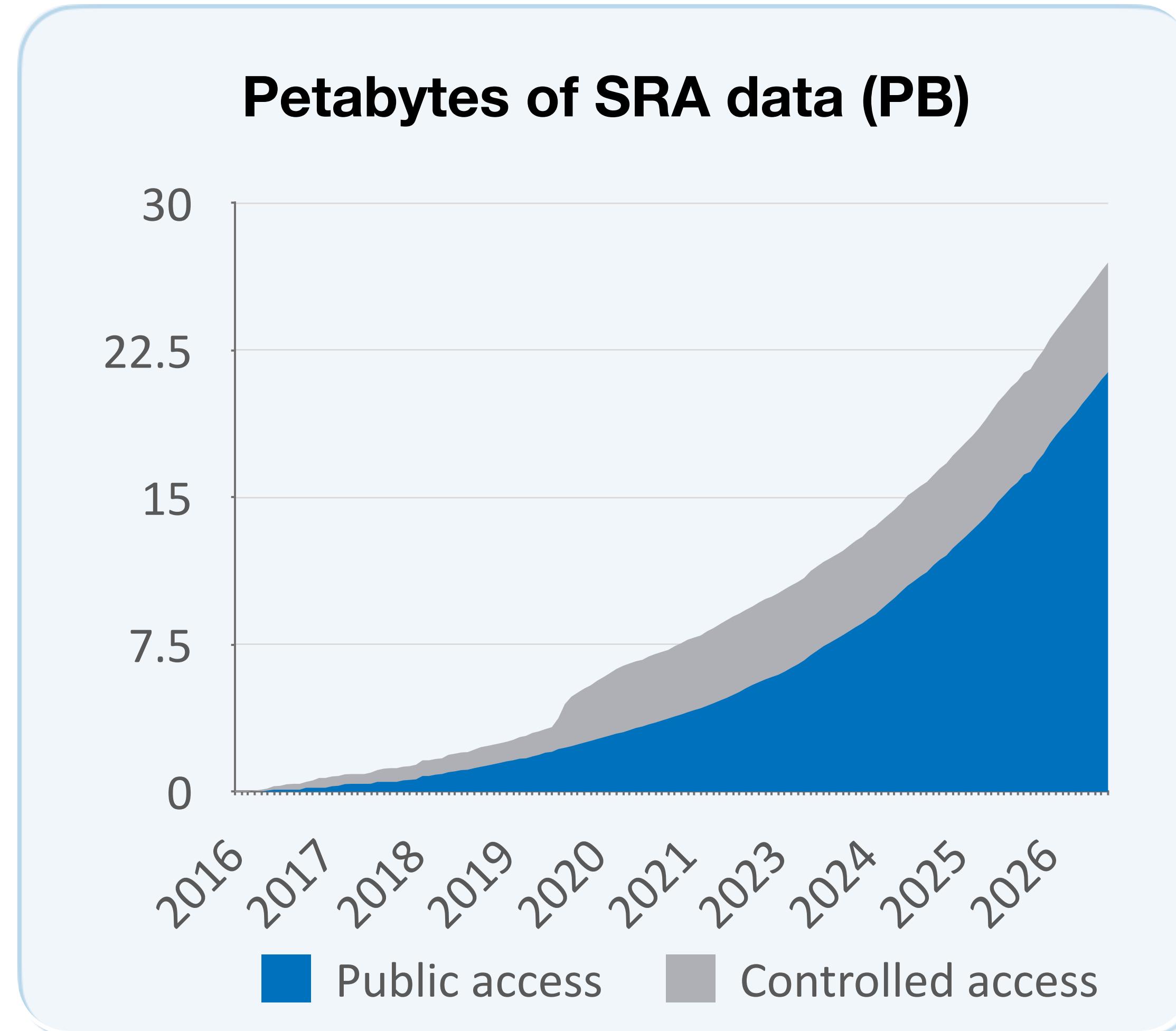
SRA sample usage



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



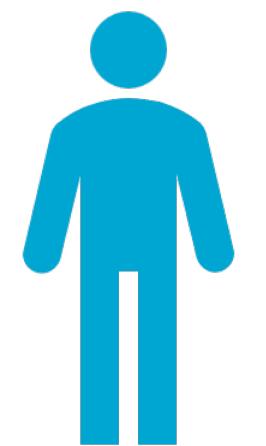
35 million sequence datasets

160+ Petabytes of sequence data available in three formats

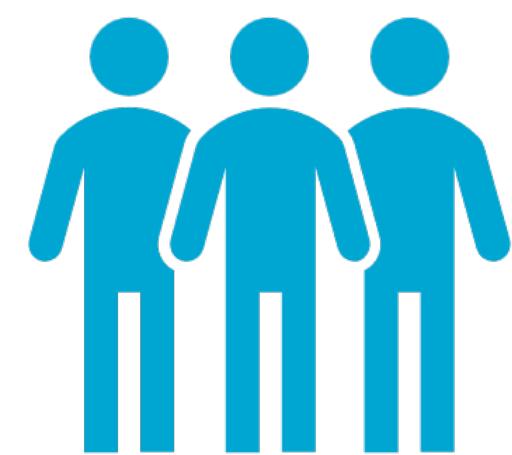
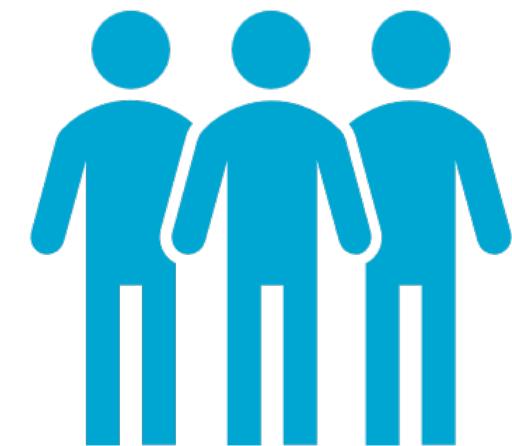
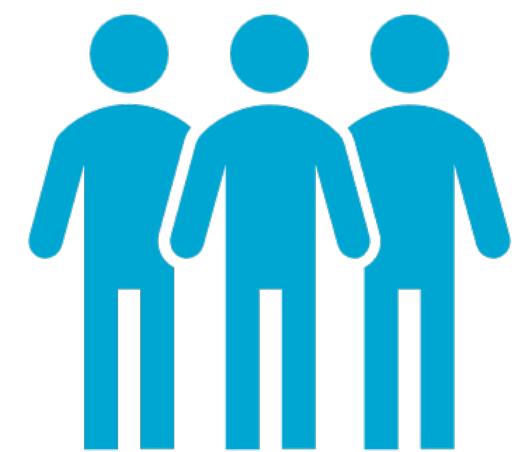
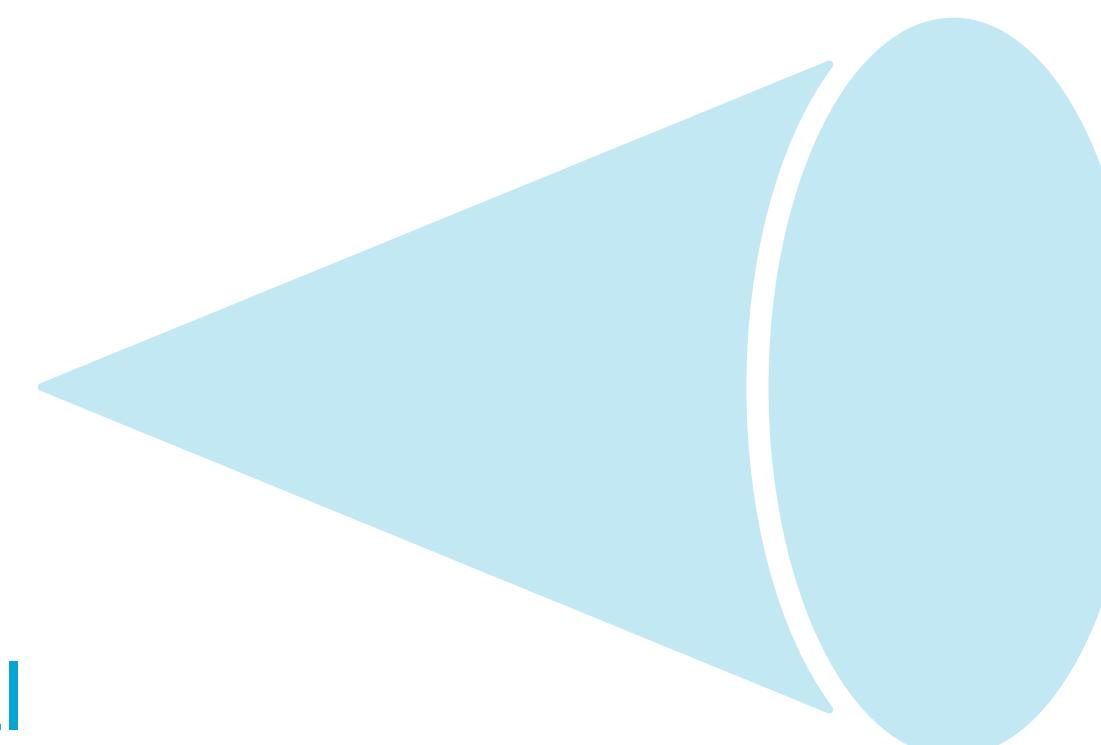
Metadata describing biological samples and library prep

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!”

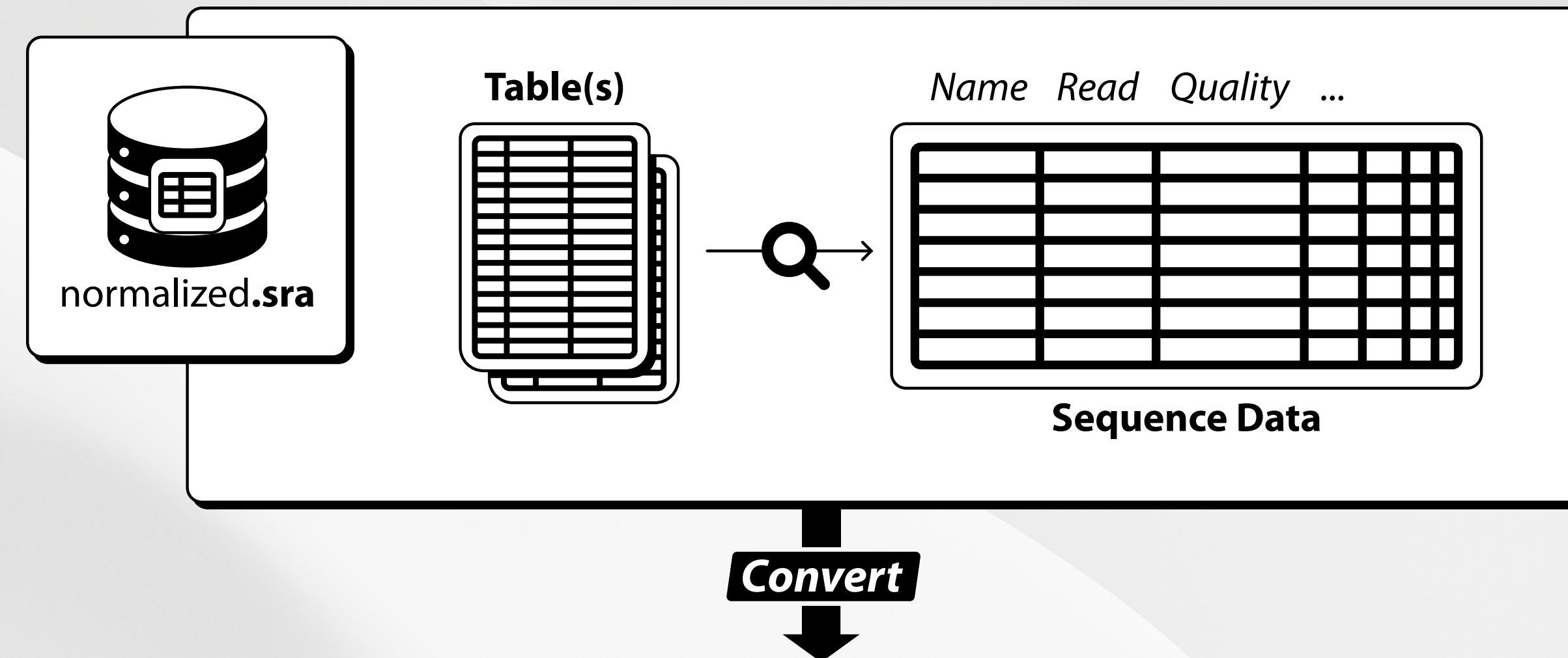
THE SEQUENCE READ ARCHIVE

FILE FORMATS

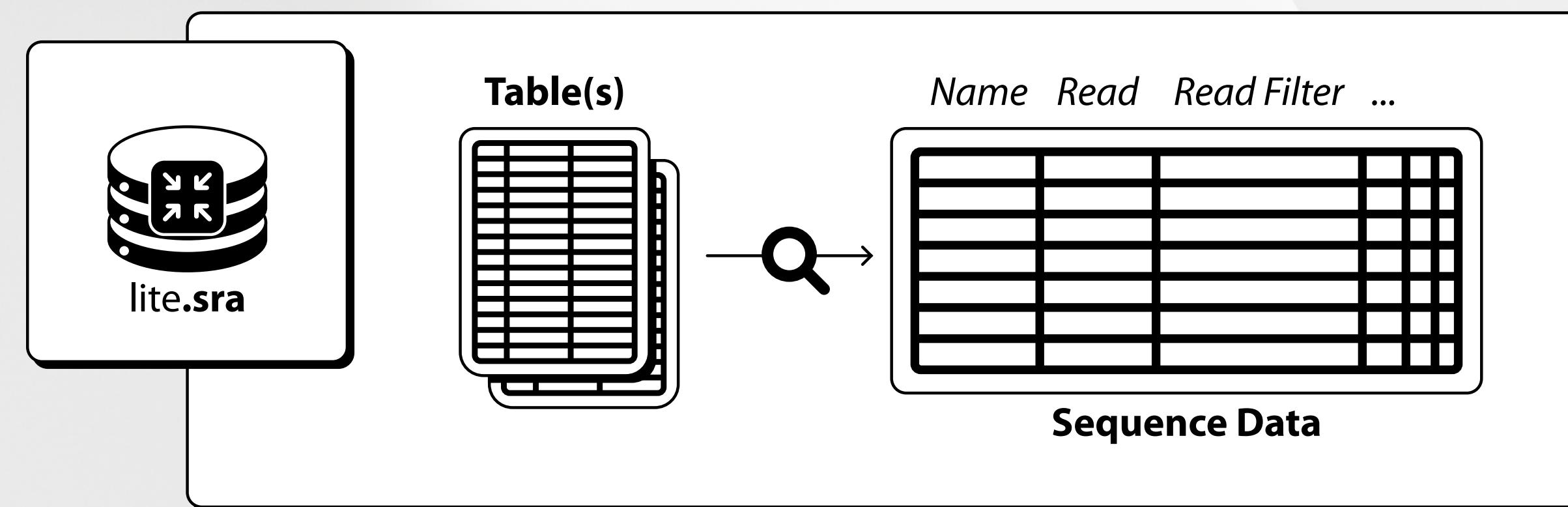


SRA Lite Format

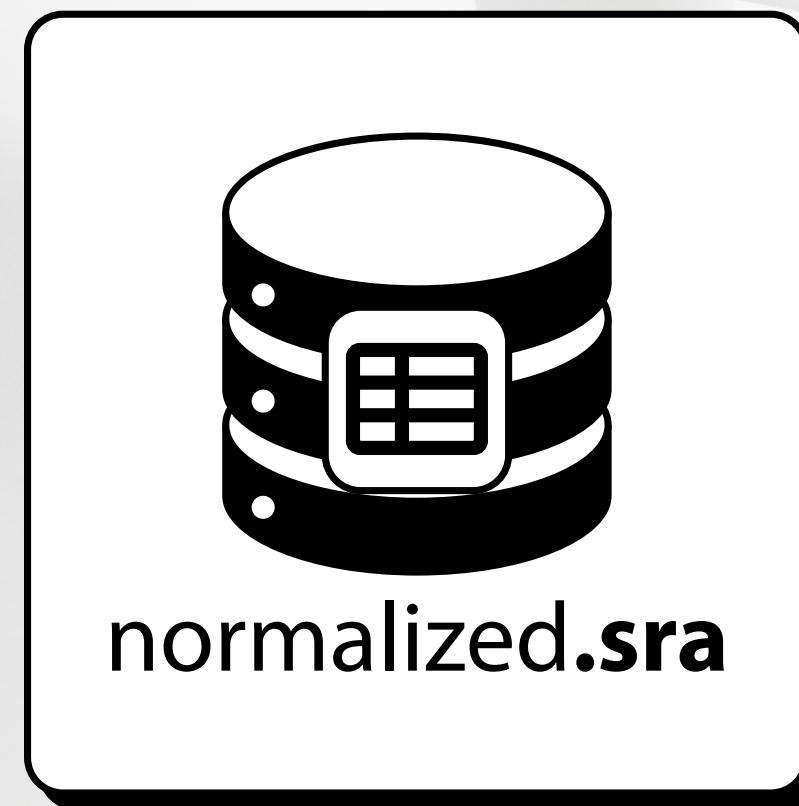
SRA Normalized Format



SRA Lite Format

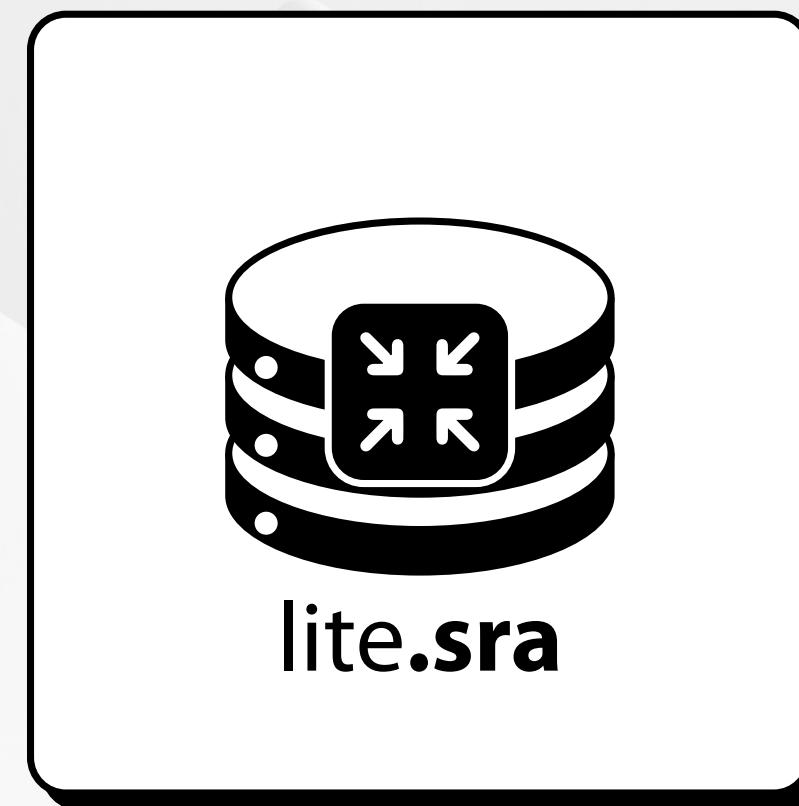


COMPARISON



normalized.sra

vs.



lite.sra

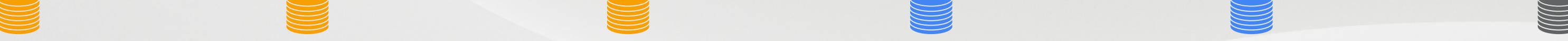
Bigger and slower

Complete representation
of the experiment

Smaller and faster

Includes Sequencing
reads + base qualities.

DATA DISTRIBUTION MODEL



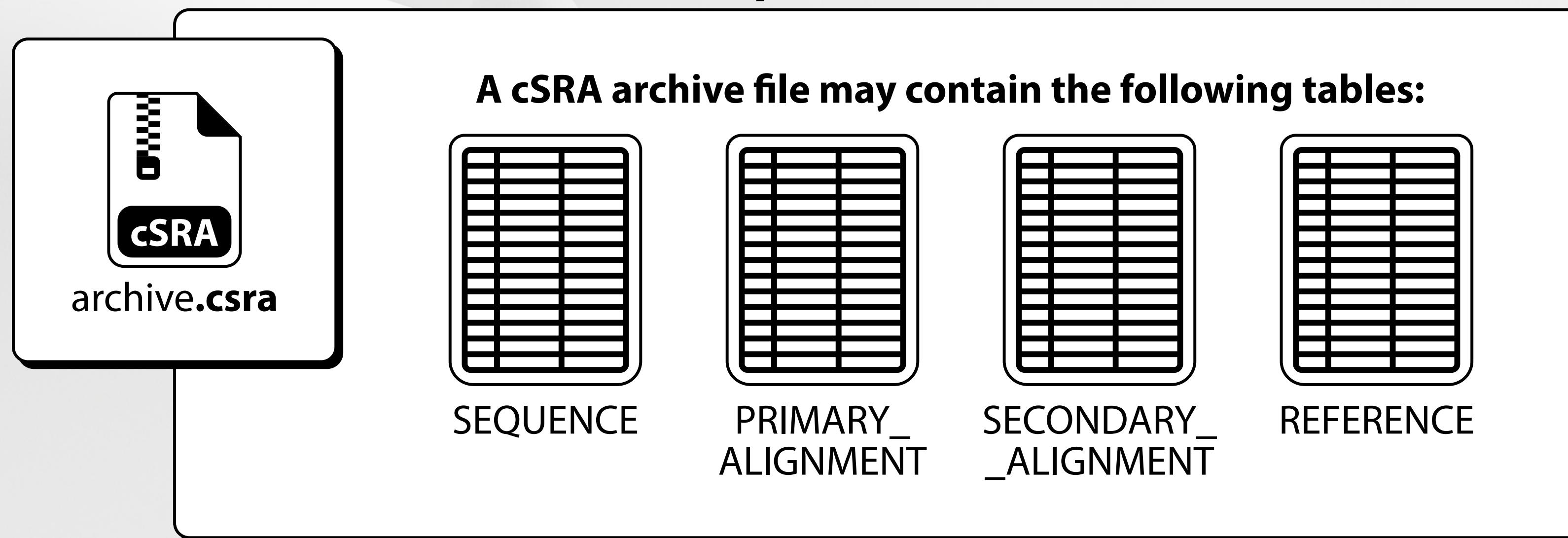
	Amazon Web Services		Google Cloud Platform		NCBI NLM	
	Open Data Platform	Glacier (Instant Retrieval)	Glacier (Deep Archive)	Public Dataset Program	Coldline	On-Prem
SRA Lite						
SRA Normalized						
Original (Submission)						
VCF (COVID-19)						

 Open Access

 At Cost

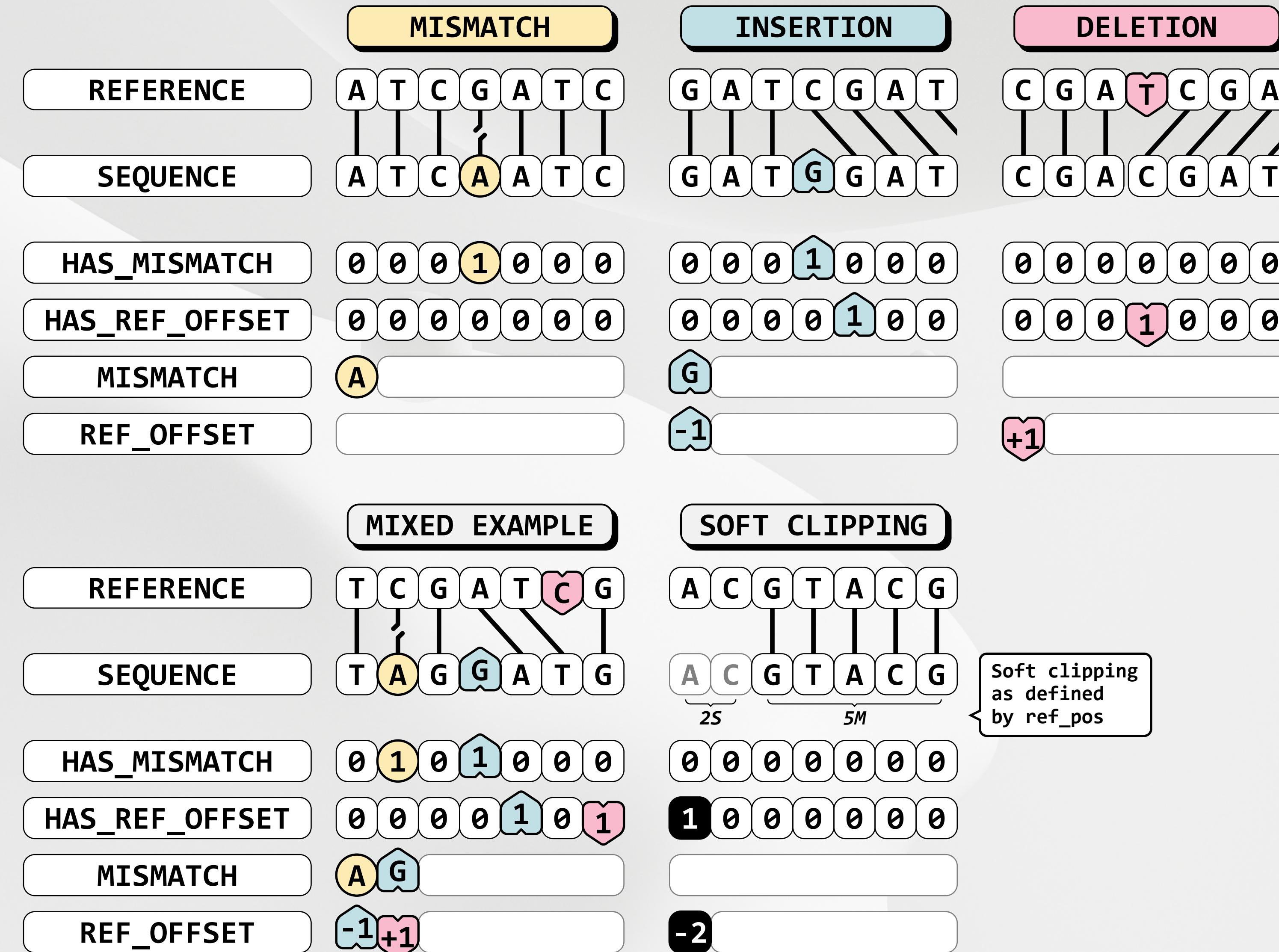
cSRA Format

Compressed SRA Format





archive.csra



Soft clipping
as defined
by ref_pos

GET MORE INFO

SRA Data Formats Manual

doi.org/10.5281/zenodo.15677383



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

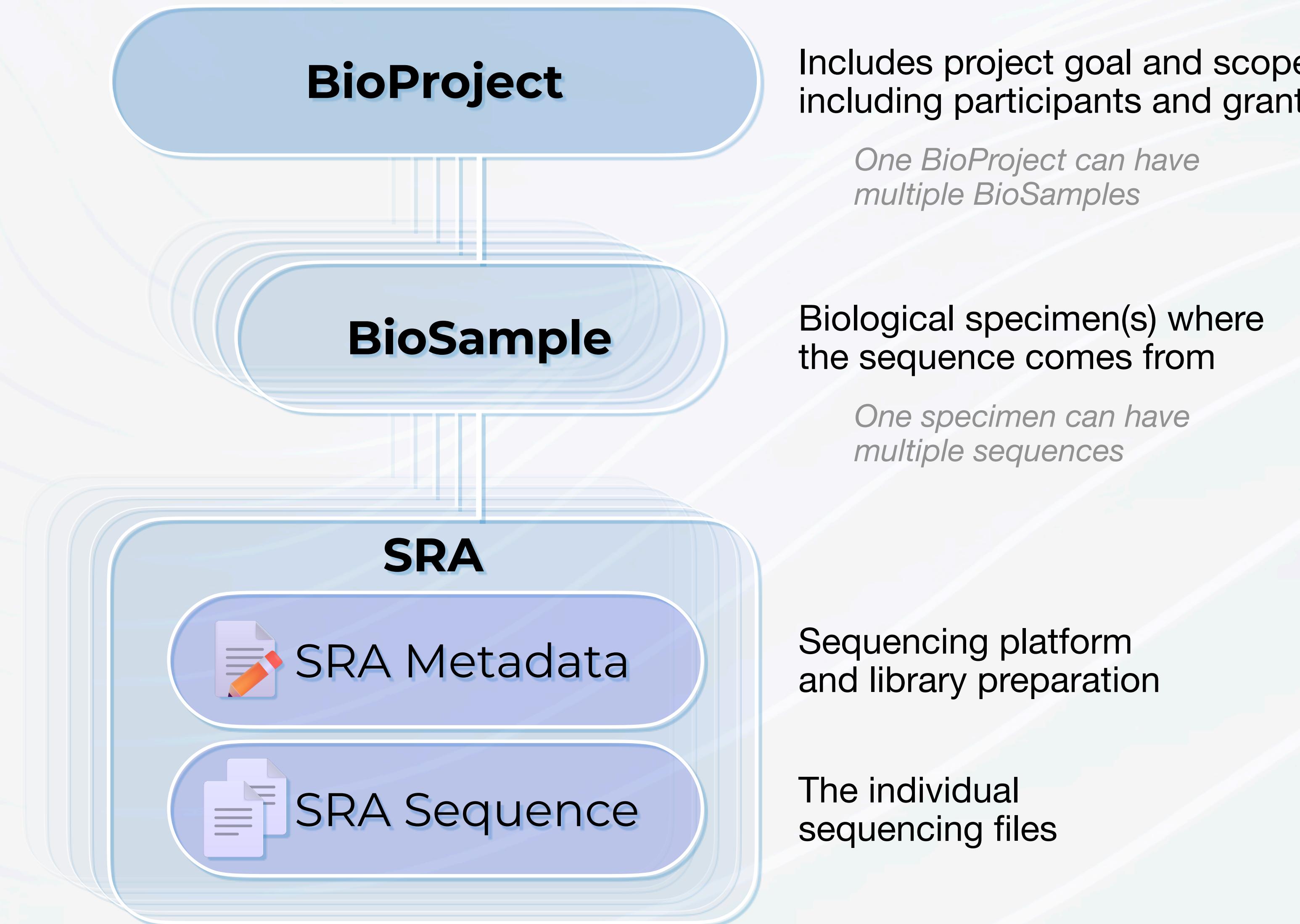
www.ncbi.nlm.nih.gov/sra

SEQUENCE READ ARCHIVE

RECORDS



SRA SUBMISSION



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

The screenshot shows the NCBI Gene Expression Omnibus (GEO) homepage at ncbi.nlm.nih.gov/geo. A yellow box highlights the search bar with the text "Let's look up GEO Accession GSE190909." A red arrow points from this box to the search results page on the right.

Search Results (Accession GSE190909):

- Series GSE190909**
- Status:** Public on Jun 01, 2022
- Title:** RNA-seq of CD4 T cells from the lungs of M. tuberculosis infected WT and vhl fl/fl cd4 cre mice
- Organism:** *Mus musculus*
- Experiment type:** Expression profiling by high throughput sequencing
- Summary:** Purpose: We compared the transcriptomes of vhl gene (vhlfl/fl cd4 cre= vhl cKO) with WT mice 4 weeks after aerosol infection with Mycobacterium tuberculosis. Results: Using an optimized data analysis...

Series Matrix File(s)

Supplementary file	Size	Download	File type/resource
GSE190909_Processed_file_log_RPKM.txt.gz	1.1 Mb	(ftp)(http)	TXT
GSE190909_Raw_counts.txt.gz	512.4 Kb	(ftp)(http)	TXT
GSE190909_processed_file_FDR_and_fold_changes.txt.gz	592.6 Kb	(ftp)(http)	TXT

SRA Run Selector

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

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

Searching up records

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

The screenshot shows the SRA Run Selector interface. At the top, there's a search bar with the accession number PRJNA788913. Below it, a red box highlights the "SRA Run Selector" header. A yellow callout box points to the "Common Fields" section, which displays various metadata about the BioProject, including BioProject ID (PRJNA788913), Assay Type (RNA-Seq), and Center Name (GEO). Another yellow callout box points to the "Run" section at the bottom, which lists 11 individual sequencing runs with their respective BioSample IDs, sizes, and experiment accessions.

Run	BioSample	Bases	Bytes	Experiment	genotype	GEO_Accession
SRR17259589	SAMN24039350	3.26 G	1.27 Gb	SRX13437946	Wild type	GSM5733924
SRR17259590	SAMN24039351	3.64 G	1.41 Gb	SRX13437947	Wild type	GSM5733925
SRR17259591	SAMN24039352	3.62 G	1.40 Gb	SRX13437948	Wild type	GSM5733926
SRR17259592	SAMN24039354	3.55 G	1.38 Gb	SRX13437949	Wild type	GSM5733927
SRR17259593	SAMN24039356	3.58 G	1.39 Gb	SRX13437950	Wild type	GSM5733928
SRR17259594	SAMN24039357	4.17 G	1.61 Gb	SRX13437951	Vhl fl/fl cd4 cre	GSM5733929
SRR17259595	SAMN24039358	3.87 G	1.50 Gb	SRX13437952	Vhl fl/fl cd4 cre	GSM5733930
SRR17259596	SAMN24039360	3.44 G	1.33 Gb	SRX13437953	Vhl fl/fl cd4 cre	GSM5733931
SRR17259597	SAMN24039361	3.55 G	1.38 Gb	SRX13437954	Vhl fl/fl cd4 cre	GSM5733932
SRR17259598	SAMN24039362	3.6 G	1.50 Gb	SRX13437955	Vhl fl/fl cd4 cre	GSM5733933
SRR17259599	SAMN24039363	3.5 G	1.18 Gb	SRV13437956	Vhl fl/fl cd4 cre	GSM5733934

The screenshot shows a list of 11 individual sequencing runs. A red box highlights the first run, SRR17259589. A yellow callout box points to this run, stating that clicking it will open the Run Browser. The table includes columns for Run ID, BioSample ID, Bases, Bytes, Experiment ID, genotype, and GEO Accession.

Run	BioSample	Bases	Bytes	Experiment	genotype	GEO_Accession
SRR17259589	SAMN24039350	3.26 G	1.27 Gb	SRX13437946	Wild type	GSM5733924
SRR17259590	SAMN24039351	3.64 G	1.41 Gb	SRX13437947	Wild type	GSM5733925
SRR17259591	SAMN24039352	3.62 G	1.40 Gb	SRX13437948	Wild type	GSM5733926
SRR17259592	SAMN24039354	3.55 G	1.38 Gb	SRX13437949	Wild type	GSM5733927
SRR17259593	SAMN24039356	3.58 G	1.39 Gb	SRX13437950	Wild type	GSM5733928
SRR17259594	SAMN24039357	4.17 G	1.61 Gb	SRX13437951	Vhl fl/fl cd4 cre	GSM5733929
SRR17259595	SAMN24039358	3.87 G	1.50 Gb	SRX13437952	Vhl fl/fl cd4 cre	GSM5733930
SRR17259596	SAMN24039360	3.44 G	1.33 Gb	SRX13437953	Vhl fl/fl cd4 cre	GSM5733931
SRR17259597	SAMN24039361	3.55 G	1.38 Gb	SRX13437954	Vhl fl/fl cd4 cre	GSM5733932
SRR17259598	SAMN24039362	3.6 G	1.50 Gb	SRX13437955	Vhl fl/fl cd4 cre	GSM5733933
SRR17259599	SAMN24039363	3.5 G	1.18 Gb	SRV13437956	Vhl fl/fl cd4 cre	GSM5733934

Scrolling down shows us all the individual runs.

Searching up records

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

National Library of Medicine
National Center for Biotechnology Information

Sequence Read Archive

Run Browser > SRR17259589

GSM5733924: CD4 WT 1; Mus musculus; RNA-Seq (SRR17259589)

Metadata Analysis Reads Data access **FASTA/FASTQ download**

Run

Run	Spots	Bases	Size	GC Content
SRR17259589	43.4M	3.3G	1.3GB	50.6%

Quality graph (bigger)

This run has 1 read per spot:

L=75, 100%

Legend ?

Experiment

Experiment	Library Name	Platform	Strategy	Source	Selection	Layout	Action
SRX13437946		Illumina	RNA-Seq	TRANSCRIPTOMIC	cDNA	SINGLE	BLAST

The Run Browser shows information about the run.

National Library of Medicine
National Center for Biotechnology Information

Sequence Read Archive

Run Browser > SRR17259589

GSM5733924: CD4 WT 1; Mus musculus; RNA-Seq (SRR17259589)

Metadata Analysis Reads Data access **FASTA/FASTQ download**

Download for Experiment SRX13437946

Accession	Total Bases	Spots	
		Total	Filtered
<input checked="" type="checkbox"/> SRR17259589	3.3Gbases	43.4M	

Filter Runs

Search by sub-sequence, Filter

What can the filter be applied to?

Download

Filtered Clipped **FASTA or FASTQ**

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	
<input type="checkbox"/>	1 SRR17259589	SAMN24039350	3.26 G	1.27 Gb	SRX13437946	Wild type	GSM5733924	2021-12-18 00:31:00Z	GSM5733924	
<input type="checkbox"/>	2 SRR17259590	SAMN24039351	3.64 G	1.41 Gb	SRX13437947	Wild type	GSM5733925	2021-12-17 16:25:00Z	GSM5733925	
<input type="checkbox"/>	3 SRR17259591	SAMN24039352	3.62 G				GSM5733926	2021-12-18 00:34:00Z	GSM5733926	
<input type="checkbox"/>	4 SRR17259592	SAMN24039354	3.55 G				GSM5733927	2021-12-17 16:01:00Z	GSM5733927	
<input type="checkbox"/>	5 SRR17259593	SAMN24039356	3.58 G				GSM5733928	2021-12-17 16:07:00Z	GSM5733928	
<input type="checkbox"/>	6 SRR17259594	SAMN24039357	4.17 G				GSM5733929	2021-12-17 16:06:00Z	GSM5733929	
<input type="checkbox"/>	7 SRR17259595	SAMN24039358	3.87 G	1.50 Gb	SRX13437952	Vhl fl/fl cd4 cre	GSM5733930	2021-12-17 15:57:00Z	GSM5733930	
<input type="checkbox"/>	8 SRR17259596	SAMN24039360	3.44 G	1.33 Gb	SRX13437953	Vhl fl/fl cd4 cre	GSM5733931	2021-12-17 16:09:00Z	GSM5733931	
<input type="checkbox"/>	9 SRR17259597	SAMN24039361	3.55 G	1.38 Gb	SRX13437954	Vhl fl/fl cd4 cre	GSM5733932	2021-12-17 16:05:00Z	GSM5733932	
<input type="checkbox"/>	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...

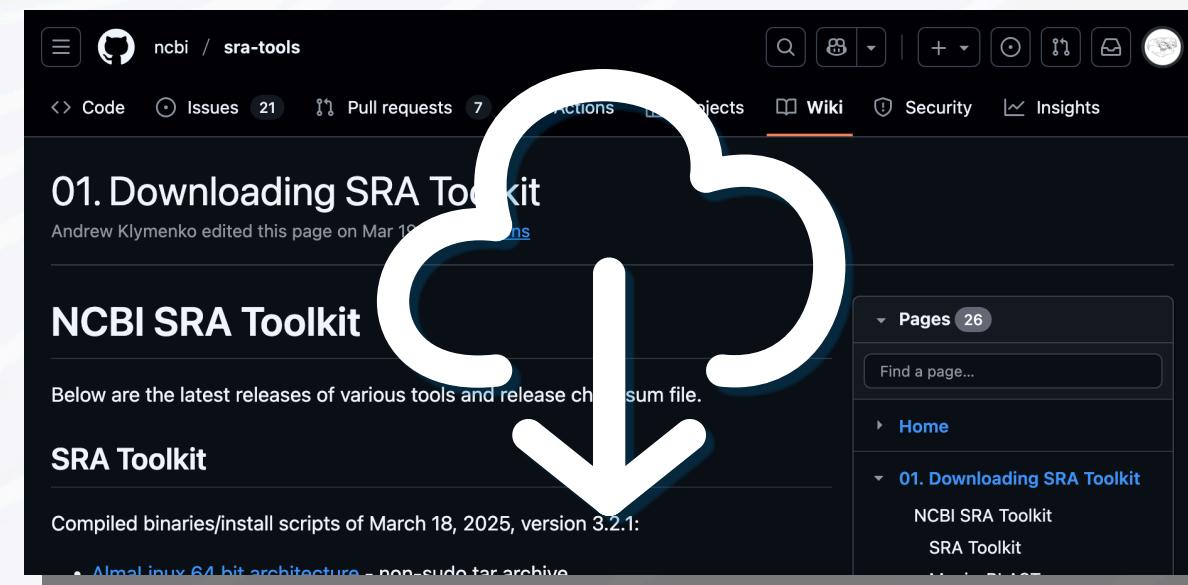
...but what about projects with **100 runs?**

...1,000 runs?

SRATOOLKIT



github.com/ncbi/sra-tools



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