Sequence data search and fetching metadata

Date etc

NC STATE UNIVERSITY

LANZAS

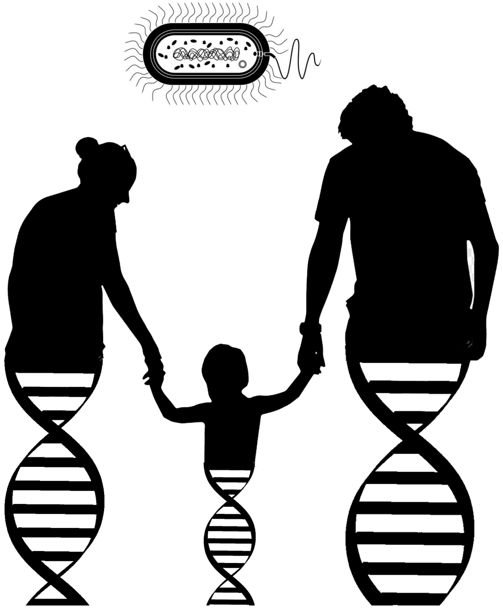
Infectious disease dynamics lab

Learning Outcomes - Day 1

- Learning to identify and download genomic data from NCBI.
  - Understand the differences between SRA, SRX, PRJ, SAM and SRR.
  - Identify the different IDs in NCBI website.

NC STATE UNIVERSITY

Gene sequencing



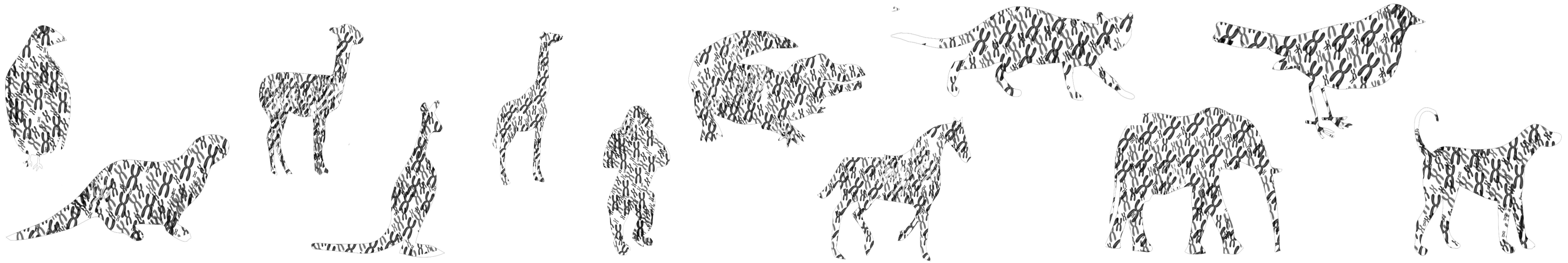
The Mendelian gene is a basic unit of heredity.

The molecular gene is a sequence of nucleotides in DNA, that is transcribed to produce a functional RNA and protein.

Gene detection in humans: Disease detection (i.e. cancer)

Gene detection in other animals: Disease detection, breeding.

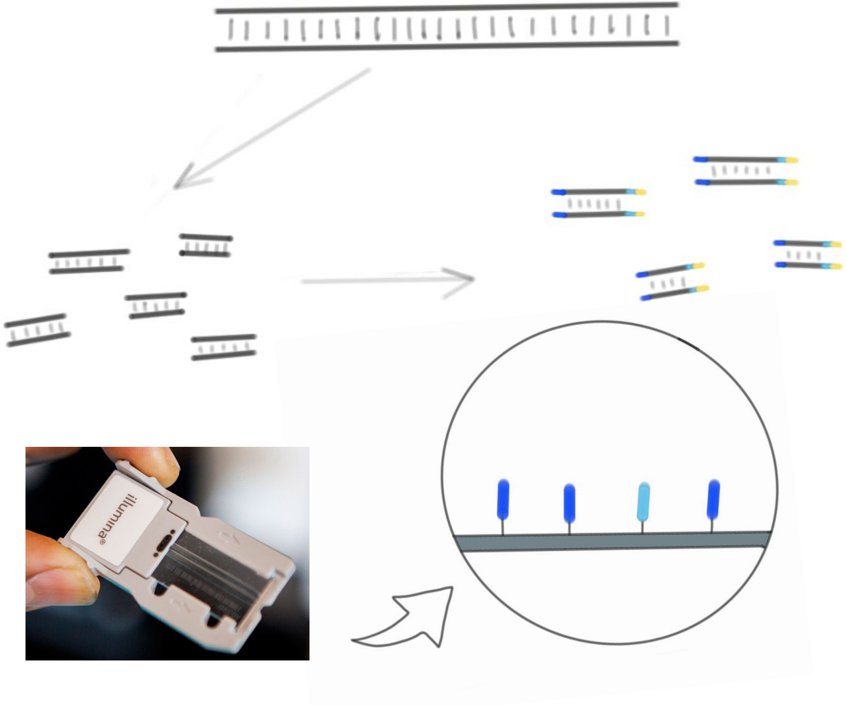
Gene detection in pathogens → AMR, virulence, other specific genes



NC STATE UNIVERSITY

Next Generation Sequencing (NGS)

DNA can be cut in small pieces and sequenced based on a **reference genome**.



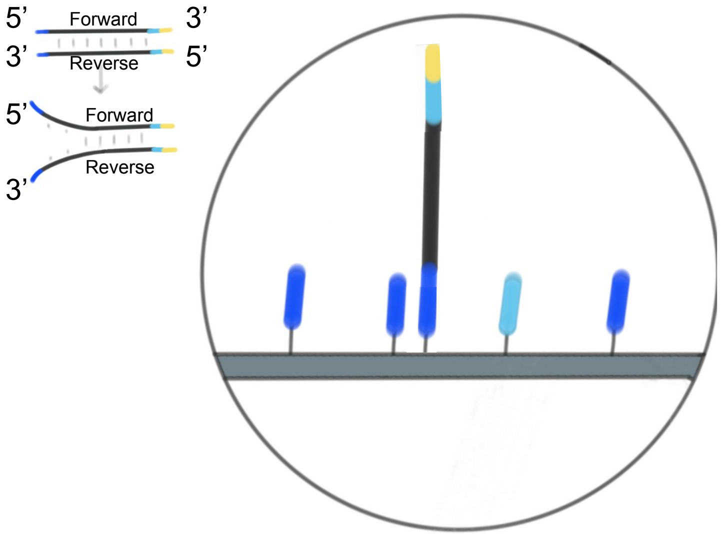
- **Library preparation**: Long strands of DNA are cut in small fragments, and specific sequences of DNA called adapters are added at each end of each fragment, including an index to identify the sample.

- **Sequencing by synthesis (Illumina)**: A glass surface called flow cell contains small fragments of DNA called oligonucleotides. These match the adapters in the library.

NC STATE UNIVERSITY

Next Generation Sequencing (NGS)

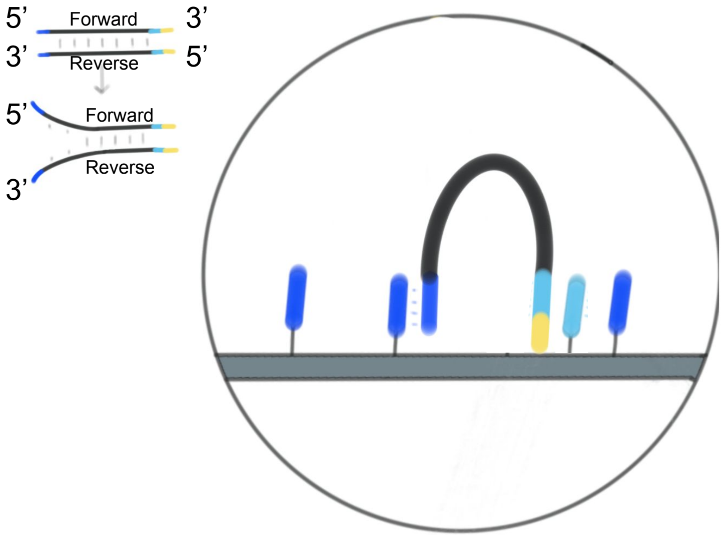
DNA can be cut in small pieces and sequenced based on a **reference genome**.



- One of your sequences (forward sense) attaches to one of the oligos by the adapter and it gets copied. This copy is the continuation of the oligo but it is the opposite sense (reverse). The forward sequences remaining are washed away.
- **Clonal amplification via PCR**: The reverse sequence bends and attaches to the next oligo, and it is copied. The result is the original sense of the sequence (forward sense). The reverse sense sequences are discarded.

Next Generation Sequencing (NGS)

DNA can be cut in small pieces and sequenced based on a **reference genome**.

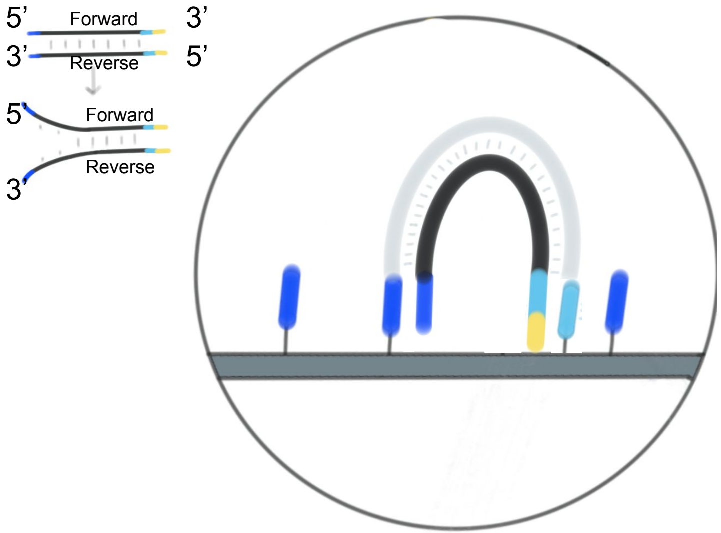


- One of your sequences (forward sense) attaches to one of the oligos by the adapter and it gets copied. This copy is the continuation of the oligo but it is the opposite sense (reverse). The forward sequences remaining are washed away.
- **Clonal amplification via PCR**: The reverse sequence bends and attaches to the next oligo, and it is copied. The result is the original sense of the sequence (forward sense). The reverse sense sequences are discarded.

NC STATE UNIVERSITY

Next Generation Sequencing (NGS)

DNA can be cut in small pieces and sequenced based on a **reference genome**.

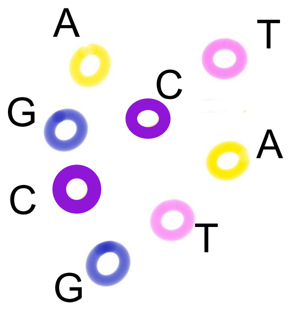


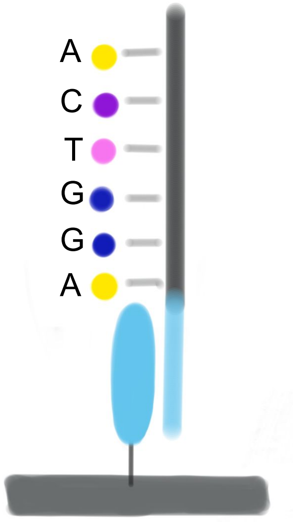
- One of your sequences (forward sense) attaches to one of the oligos by the adapter and it gets copied. This copy is the continuation of the oligo but it is the opposite sense (reverse). The forward sequences remaining are washed away.
- **Clonal amplification via PCR**: The reverse sequence bends and attaches to the next oligo, and it is copied. The result is the original sense of the sequence (forward sense). The reverse sense sequences are discarded.

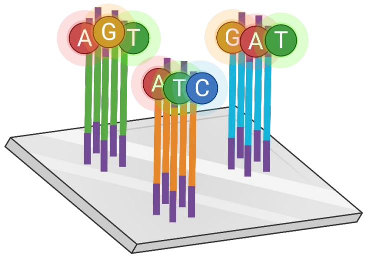
NC STATE UNIVERSITY

Next Generation Sequencing (NGS)

DNA can be cut in small pieces and sequenced based on a **reference genome**.





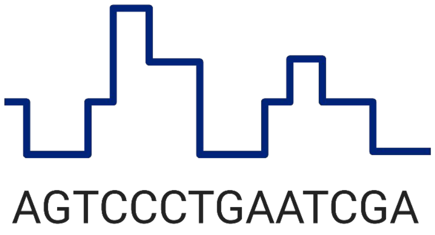


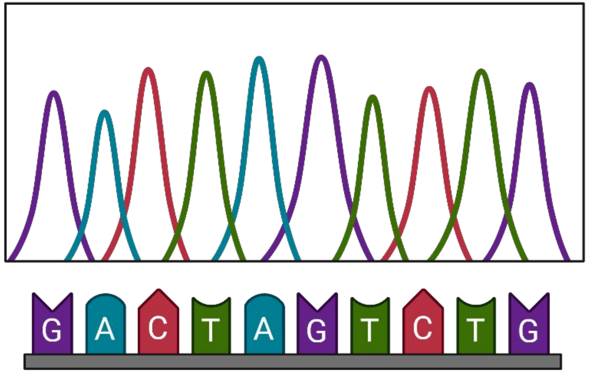
- Sequencing primers bind to the forward strand. In the sequencing medium, there are fluorescent nucleotides with one specific fluorescent tag, polymerase enzymes and a terminator.
- Only one nucleotide per cycle is attached to the sequence, and its color is detected and stored by the instrument. Finally, the index gets sequenced. **This is called single-end.**
- If a second index is sequenced, along with the reverse strand of the library, it is known as **paired-end sequencing**.

NC STATE UNIVERSITY

Next Generation Sequencing (NGS)

DNA can be cut in small pieces and sequenced based on a **reference genome**.

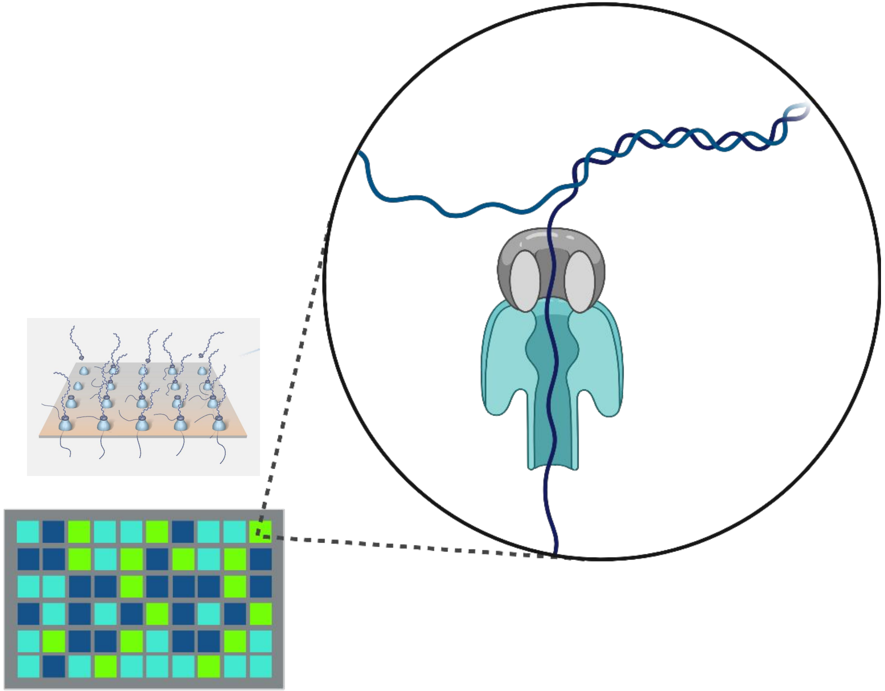




- After several filters within the instrument, the resulting sequences are **demultiplexed**, which means sort reads by sample based on the attached indexes.
- After demultiplexing, the sequences are **mapped to a reference genome.**
- This type of sequencing is **short read sequencing,** with fragments between 75-300bp

Nanopore sequencing

Long-read sequencing. Oxford Nanopore Technologies (ONT)

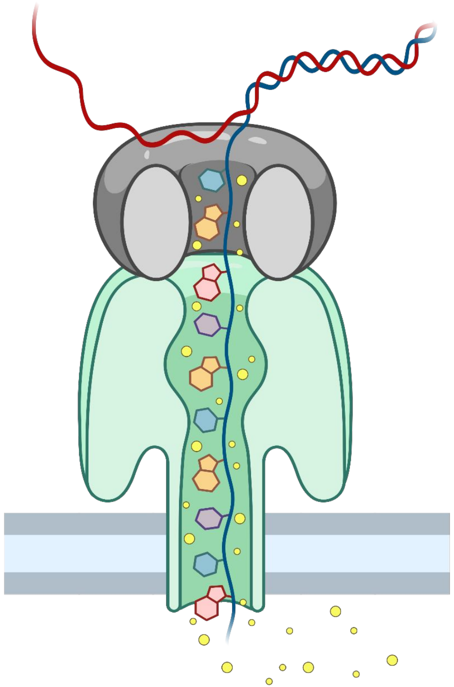


- Similar than with Illumina, libraries are prepared where an adapter is attached to each sequence.
- ONT sequencer flow cells are made of an electric-resistant bilayer created by a synthetic polymer with an array of nanopores (~1.8nm).
- A potential is applied across the membrane, resulting in a current flowing only through the nanopores.

NC STATE UNIVERSITY

Nanopore sequencing

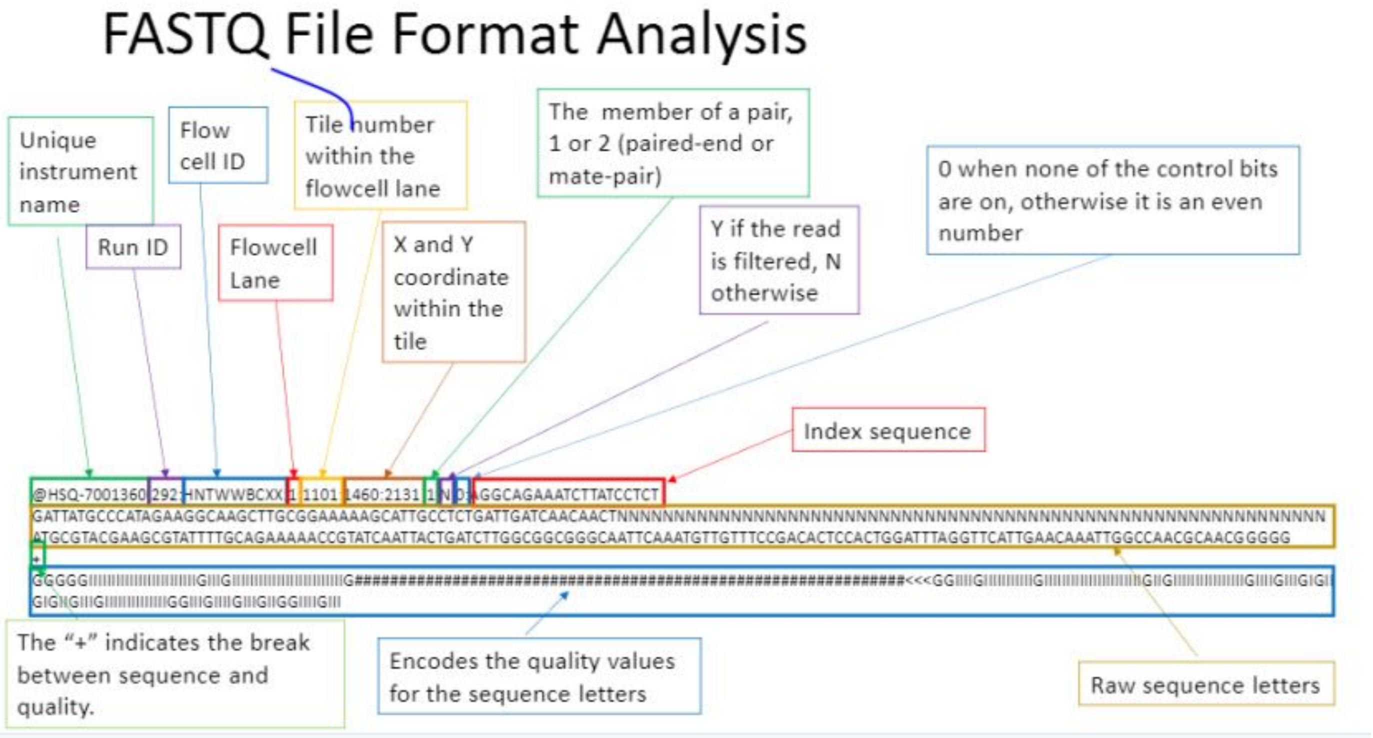
Long-read sequencing. Oxford Nanopore Technologies (ONT)



- Single molecules flowing through the nanopores cause characteristic disruptions in the electric current across the membrane.
- Measuring those disruptions, the molecules can be easily identified.
- ONT analyzes intact DNA strands that pass through the nanopores and analyzed in real time.
- By preparing the DNA to have a hairpin structure, ONT can produce pair-end data in one continuous read.
- Long-read sequencing still has higher error rate than short-read sequencing.

Raw sequences format

FASTQ format is a text-based format for storing both a biological sequence (usually nucleotide sequence) and its corresponding quality scores.



FASTQ File Format Analysis

Unique instrument name

Run ID

Flow cell ID

Flowcell Lane

Tile number within the flowcell lane

X and Y coordinate within the tile

The member of a pair, 1 or 2 (paired-end or mate-pair)

Y if the read is filtered, N otherwise

0 when none of the control bits are on, otherwise it is an even number

Index sequence

```
@HSQ-7001360:292:HNTWWBCXX:1:1101:1460:2131 1:N:0:AGGCAGAAATCTTATCCTCT
GATTATGCCCATAGAAGGCAAGCTTGCGGAAAAAGCATTGCCTCTGATTGATCAACAACTNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
ATGCGTACGAAGCGTATTTTGCAGAAAAACCGTATCAATTACTGATCTTGGCGGCGGGCAATTCAAATGTTGTTTCCGACACTCCACTGGATTTAGGTTCATTGAACAAATTGGCCAACGCAACGGGGG
+
GGGGGIIIIIIIIIIIIIIIIIIIIGIIGIIIIIIIIIIIIIIIIIIIIIIG##########################################################################<<<GGIIIIGIIIIIIIIIGIIIIIIIIIIIIIIIIIIIIGIIGIIIIIIIIIIIIIIIIGIIIIGIIIGIGI
GIGIIGIIIGIIIIIIIIIIIIIGGIIIGIIIIGIIIGIIGGIIIIGIII
```

The "+" indicates the break between sequence and quality.

Encodes the quality values for the sequence letters

Raw sequence letters

Genomic data and where to find it

NCBI

NIH National Library of Medicine

National Center for Biotechnology Information

BV-BRC

Bacterial and Viral Bioinformatics Resource Center

Bacterial Bioinformatics Using BV-BRC

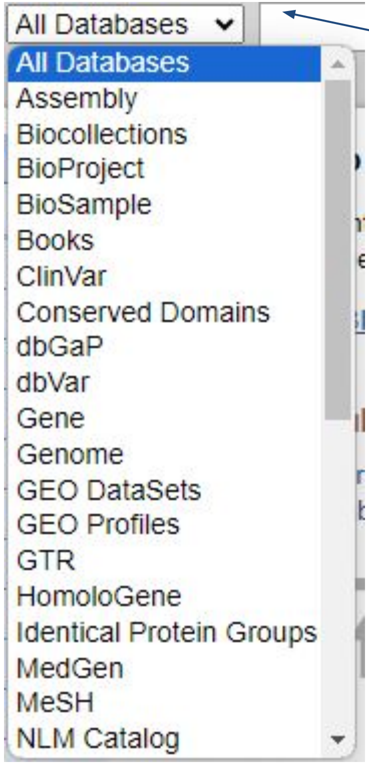
EnteroBase

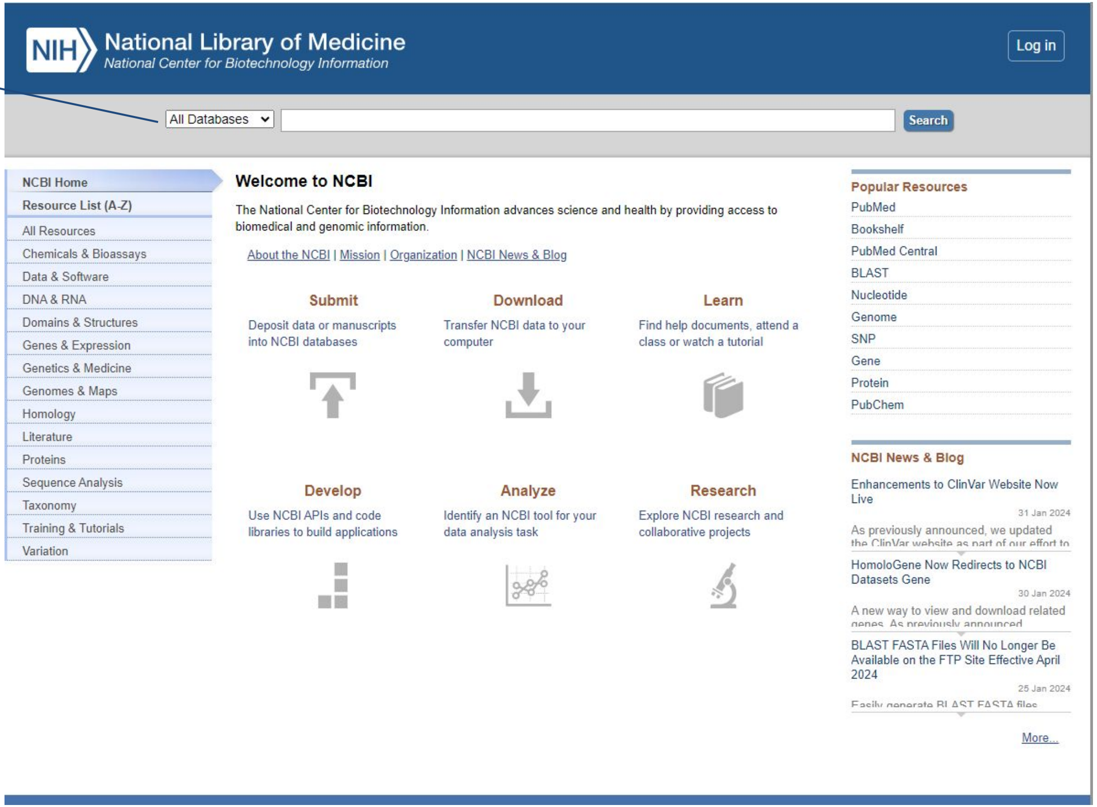
ENA

European Nucleotide Archive

NC STATE UNIVERSITY

Genomic data and where to find it





Welcome to NCBI

The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.

About the NCBI | Mission | Organization | NCBI News & Blog

Submit
Deposit data or manuscripts into NCBI databases

Download
Transfer NCBI data to your computer

Learn
Find help documents, attend a class or watch a tutorial

Develop
Use NCBI APIs and code libraries to build applications

Analyze
Identify an NCBI tool for your data analysis task

Research
Explore NCBI research and collaborative projects

NCBI News & Blog

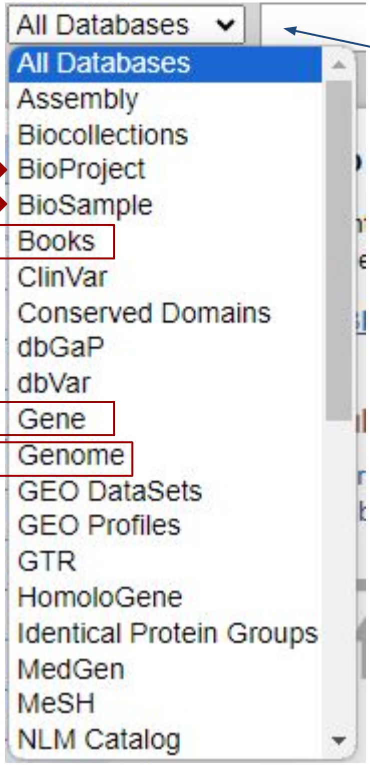
Enhancements to ClinVar Website Now Live
31 Jan 2024
As previously announced, we updated the ClinVar website as part of our effort to

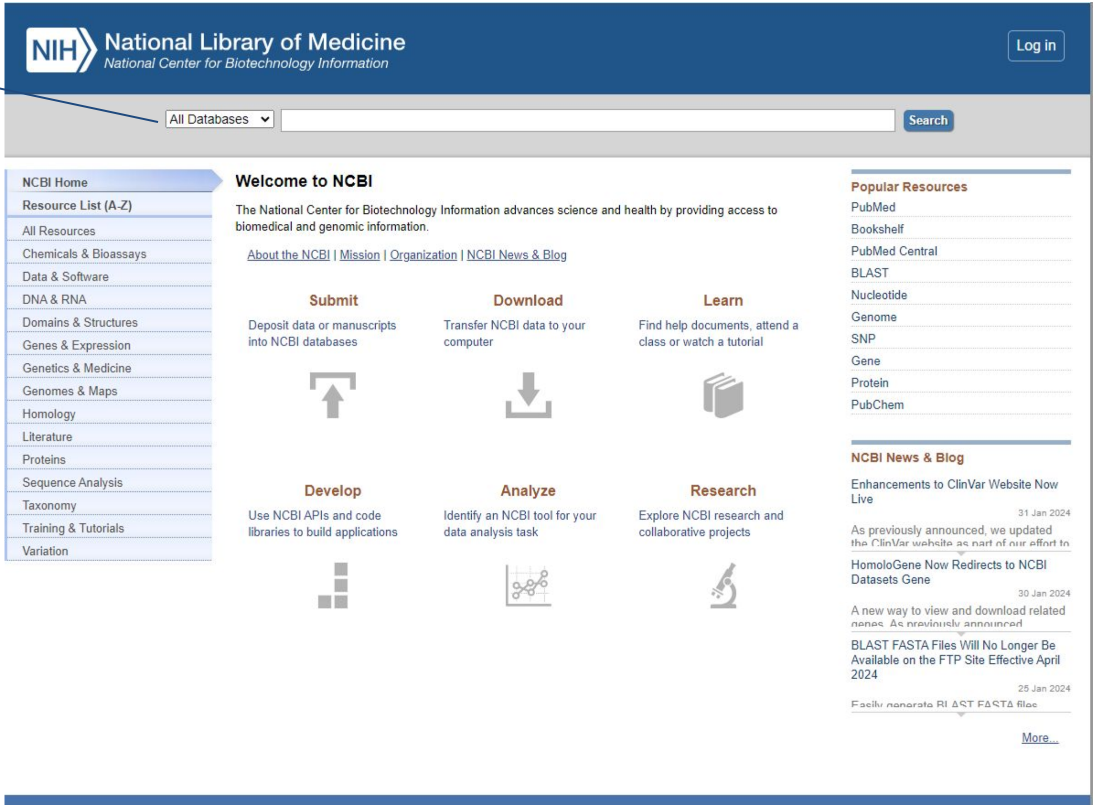
HomoloGene Now Redirects to NCBI Datasets Gene
30 Jan 2024
A new way to view and download related genes. As previously announced,

BLAST FASTA Files Will No Longer Be Available on the FTP Site Effective April 2024
25 Jan 2024
Easily generate BLAST FASTA files

More...

Genomic data and where to find it





Welcome to NCBI

The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.

About the NCBI | Mission | Organization | NCBI News & Blog

Submit
Deposit data or manuscripts into NCBI databases

Download
Transfer NCBI data to your computer

Learn
Find help documents, attend a class or watch a tutorial

Develop
Use NCBI APIs and code libraries to build applications

Analyze
Identify an NCBI tool for your data analysis task

Research
Explore NCBI research and collaborative projects

NCBI News & Blog

Enhancements to ClinVar Website Now Live
31 Jan 2024
As previously announced, we updated the ClinVar website as part of our effort to

HomoloGene Now Redirects to NCBI Datasets Gene
30 Jan 2024
A new way to view and download related genes. As previously announced

BLAST FASTA Files Will No Longer Be Available on the FTP Site Effective April 2024
25 Jan 2024
Easily generate BLAST FASTA files

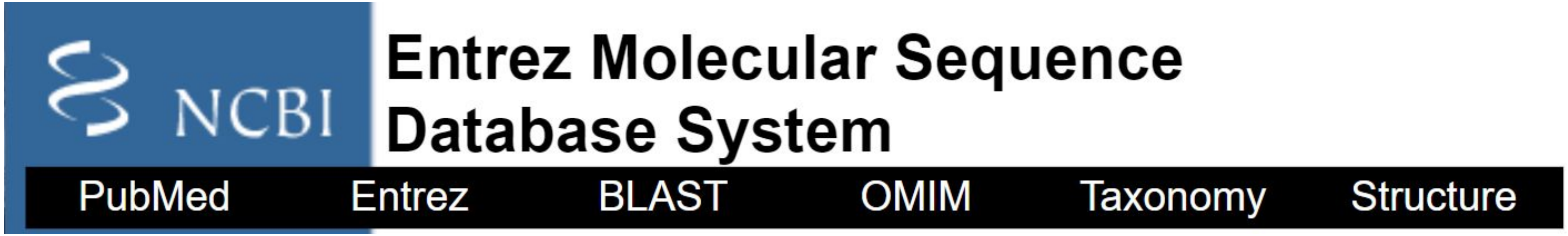
More...

SRAs and NCBI Entrez

Sequence Read Archive (**SRA**) data, available through multiple cloud providers and NCBI servers, is the largest publicly available repository of high throughput sequencing data.

Entrez is a molecular biology database system that provides integrated access to nucleotide and protein sequence data, gene-centered and genomic mapping information, 3D structure data, PubMed MEDLINE, and more.

Entrez covers over 20 databases including the complete protein sequence data from PIR-International, PRF, Swiss-Prot, and PDB and nucleotide sequence data from GenBank that includes information from EMBL and DDBJ.


Entrez Molecular Sequence Database System

NCBI Data Acronyms

Sequence Read Archive (SRA): repository that stores raw sequencing data. It includes the actual sequences obtained from the high-throughput sequencing machines before any processing or analysis.

PRJ: Refers to a genomic **project**. It represents a collection of related experiments and data associated with a particular scientific investigation.

SRX: Refers to a **sequencing experiment.** It is an organized set of runs, which are the actual sequencing data generated for a sample or set of samples. SRX[...], ERX[...], DRX[...].

SAM: Represents an individual **biological sample** that is part of a project. It is a specific instance of genetic material taken from an organism.

SRR: Represents a **specific run** of sequencing data. It is the raw data generated in a single sequencing run for a particular sample.

NCBI Data Acronyms

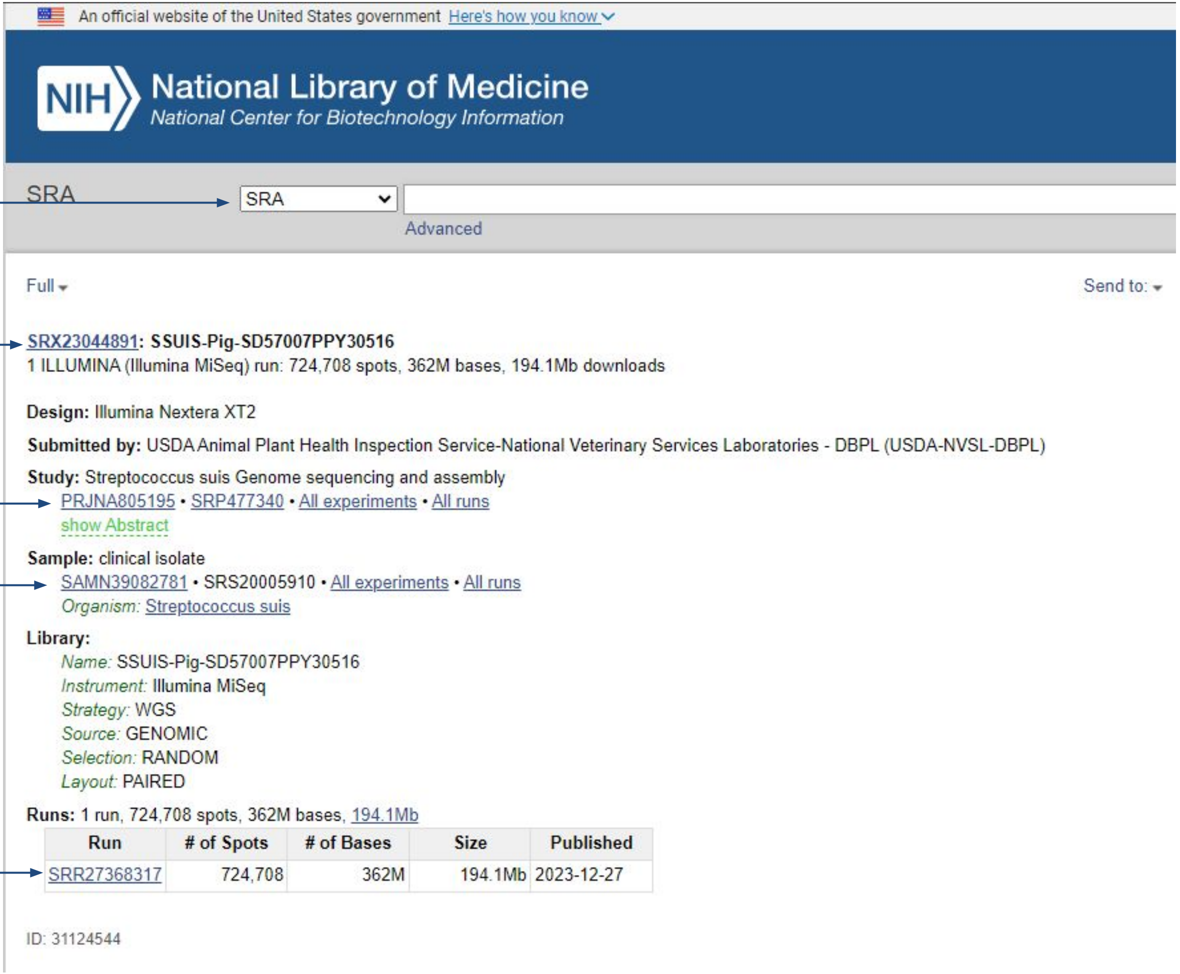
SRA

SRX

PRJ

SAM

SRR



An official website of the United States government Here's how you know
NIH
National Library of Medicine
National Center for Biotechnology Information
SRA
SRA
Advanced
Full
Send to:

SRX23044891: SSUIS-Pig-SD57007PPY30516
1 ILLUMINA (Illumina MiSeq) run: 724,708 spots, 362M bases, 194.1Mb downloads

Design: Illumina Nextera XT2

Submitted by: USDA Animal Plant Health Inspection Service-National Veterinary Services Laboratories - DBPL (USDA-NVSL-DBPL)

Study: Streptococcus suis Genome sequencing and assembly
PRJNA805195 • SRP477340 • All experiments • All runs
show Abstract

Sample: clinical isolate
SAMN39082781 • SRS20005910 • All experiments • All runs
Organism: Streptococcus suis

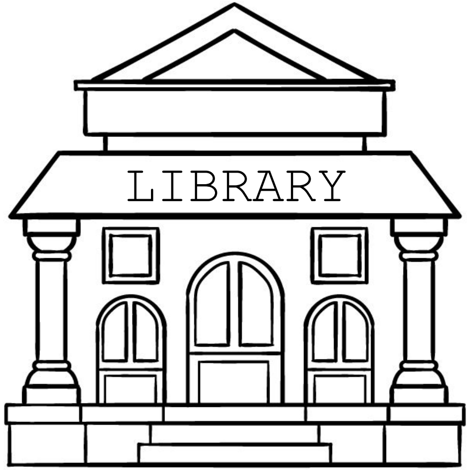
Library:
Name: SSUIS-Pig-SD57007PPY30516
Instrument: Illumina MiSeq
Strategy: WGS
Source: GENOMIC
Selection: RANDOM
Layout: PAIRED

Runs: 1 run, 724,708 spots, 362M bases, 194.1Mb

Run	# of Spots	# of Bases	Size	Published
SRR27368317	724,708	362M	194.1Mb	2023-12-27

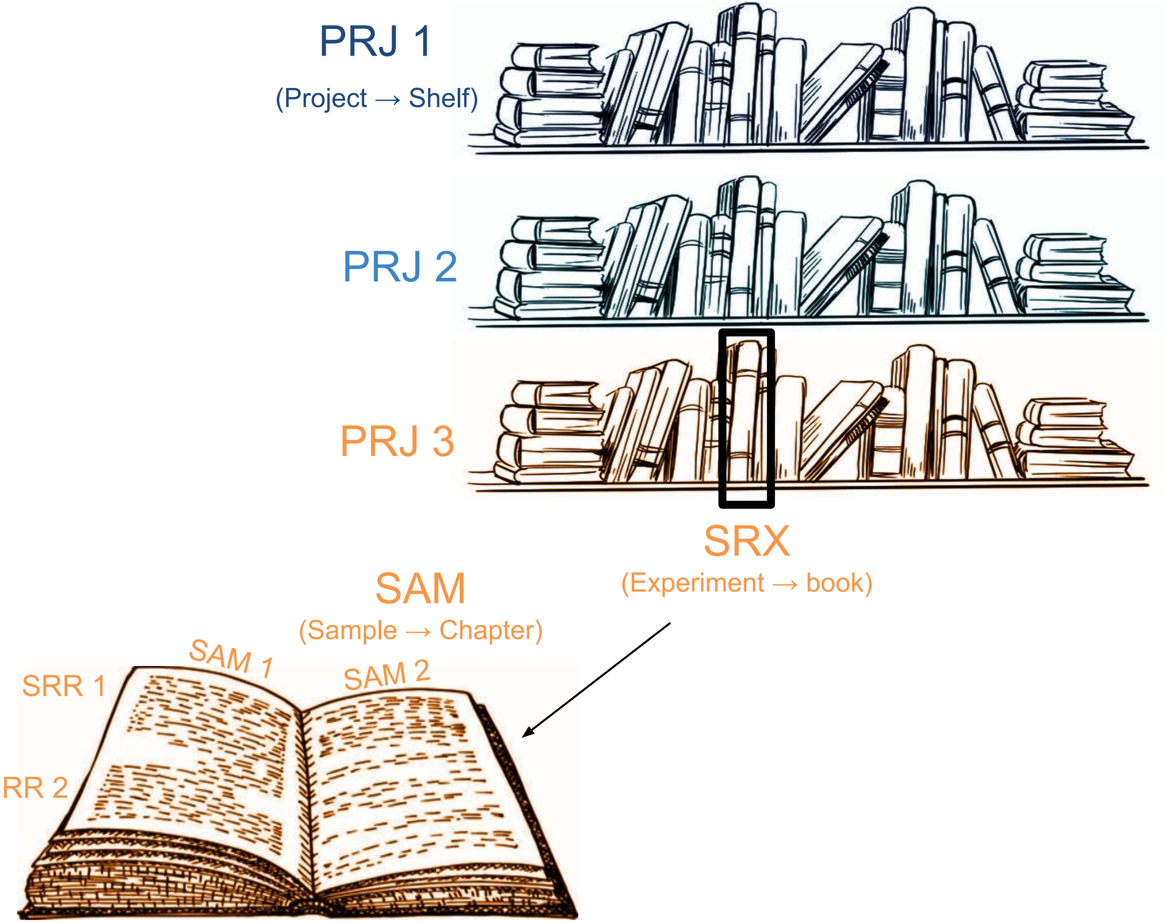
ID: 31124544

NCBI Data Acronyms



SRA

(Sequence Read Archive → Library)



PRJ 1

(Project → Shelf)

PRJ 2

PRJ 3

SRX

(Experiment → book)

SAM

(Sample → Chapter)

SAM 1

SAM 2

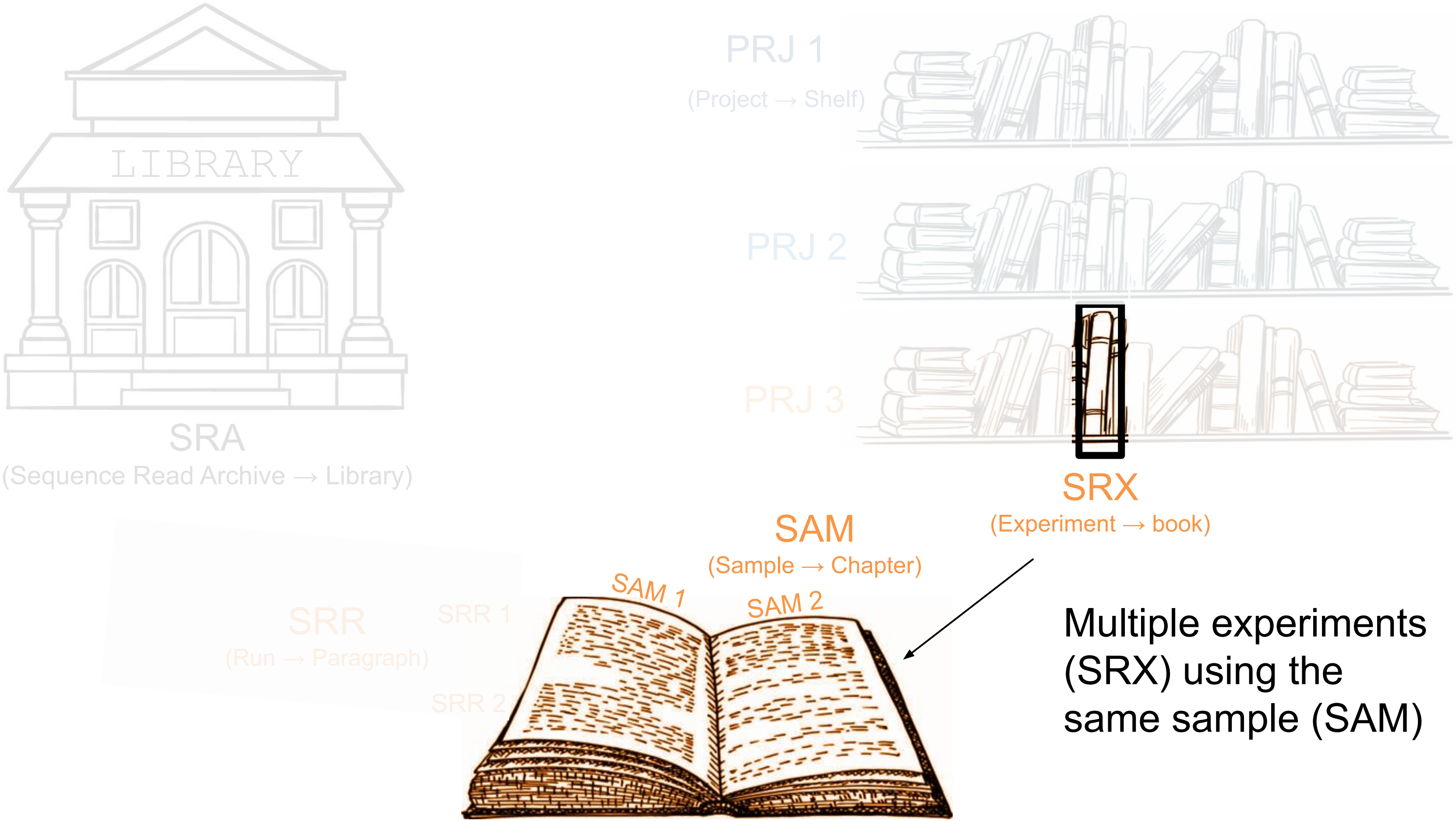
SRR

(Run → Paragraph)

SRR 1

SRR 2

NCBI Data Acronyms



Multiple experiments (SRX) using the same sample (SAM)

Take home messages

- Sequence details are important (i.e. single vs paired).
- Difference between long and short-read sequencing.
- FASTq files contain important information.
- Numerous public databases for genomic data.
- SRA → PRJ → SRX → SAM → SRR.

Questions