An NCBI Guide to Finding and Analyzing Metagenomic Data

Cooper J. Park, PhD

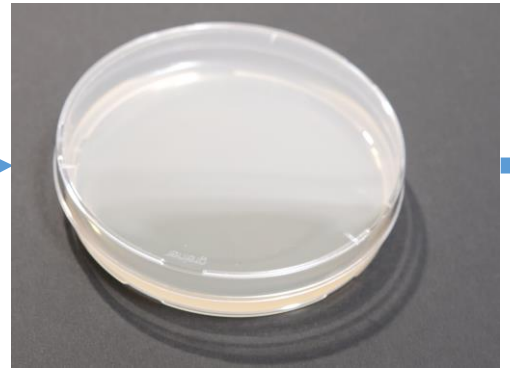
NIH National Library of Medicine
National Center for Biotechnology Information

Outline

- What is Metagenomics?
- How Does NCBI Support Metagenomics?
- Today's Case Study
  - Objective 1 – Find Metagenomic Reads in SRA and Explore Taxonomic Composition using STAT
  - Objective 2 – Use MagicBLAST to align metagenomic reads against reference sequences

Sequencing cultured organisms is not a perfect "science" - 1

Environmental Community



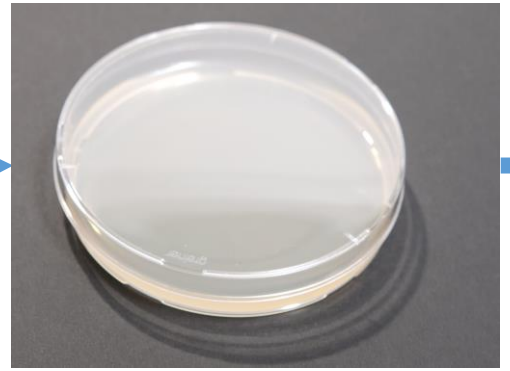
Laboratory Isolation

DNA Sequencing

Bioinformatic Analysis

Sequencing cultured organisms is not a perfect “science” - 2

Environmental Community



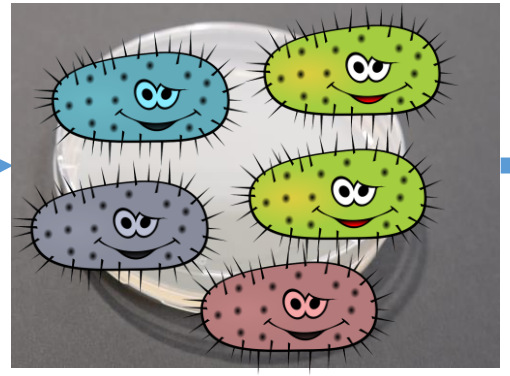
Laboratory Isolation

DNA Sequencing

Bioinformatic Analysis

Sequencing cultured organisms is not a perfect “science” - 3

Environmental Community



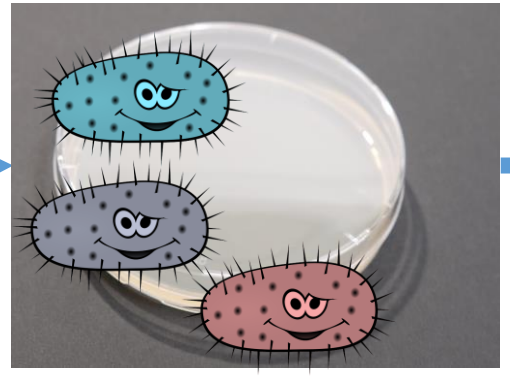
Laboratory Isolation

DNA Sequencing

Bioinformatic Analysis

Sequencing cultured organisms is not a perfect “science” - 4

Environmental Community



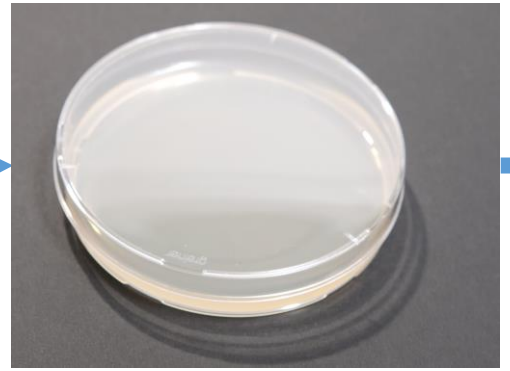
Laboratory Isolation

DNA Sequencing

Bioinformatic Analysis

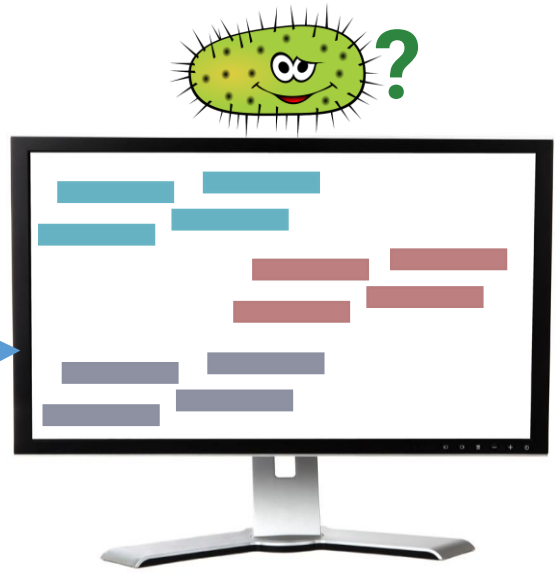
Sequencing cultured organisms is not a perfect “science” - 5

Environmental Community



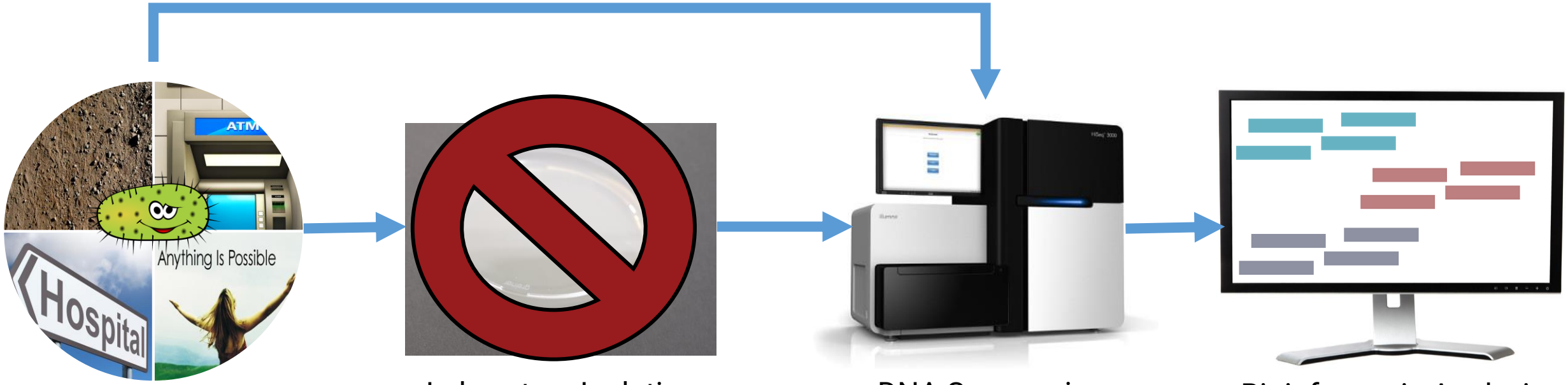
Laboratory Isolation

DNA Sequencing



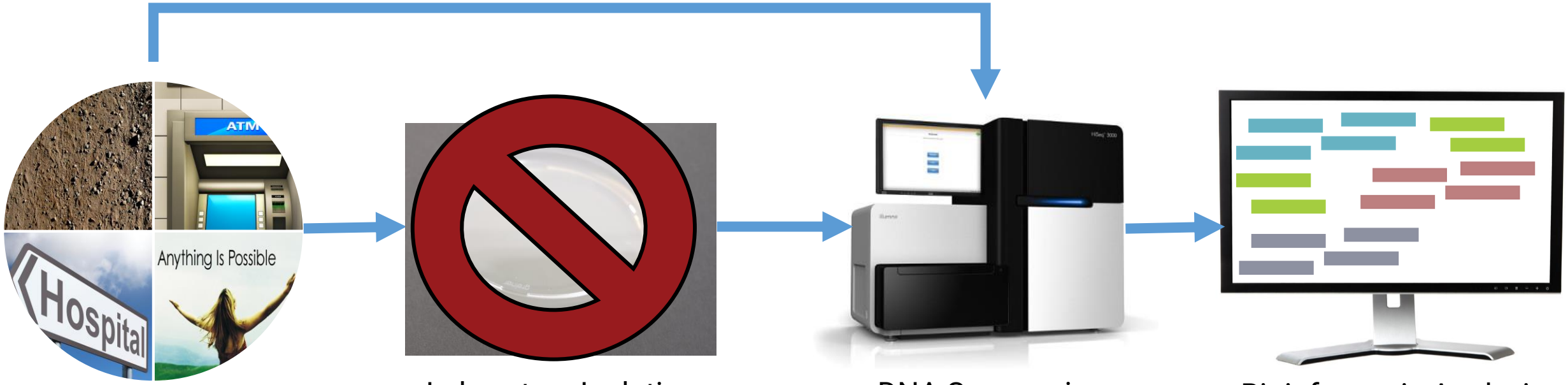
Bioinformatic Analysis

Sequencing cultured organisms is not a perfect "science" - 6



Metagenomics is sequencing without culturing

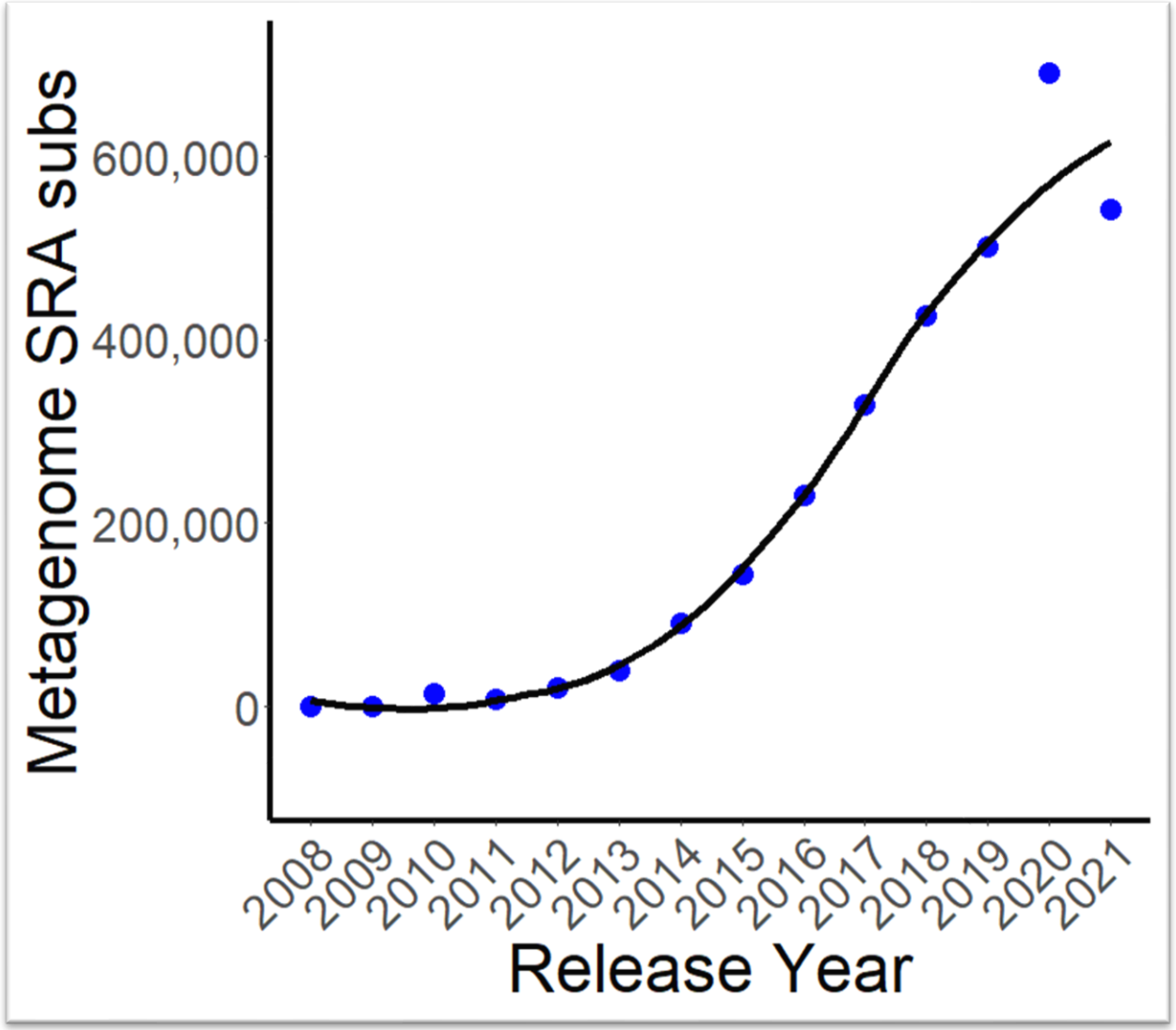
Sequencing cultured organisms is not a perfect "science" - 7



Metagenomics is sequencing without culturing

Current trends of metagenomic data production

- Over 3 million metagenomic read submissions in NCBI
- 675 Terabytes of read data
- Annual Rate of new data growing exponentially



Source: SRA AWS metadata tables queried on 09/05/2021

There’s still a catch…

Cultured Sequences

Metagenomic Sequences

NCBI Metagenomic Resources - 1

Data Storage

Sequence Read Archive

NCBI

Genbank & RefSeq

metagenome[ORGN]

Taxonomy Keyword

Data Analysis

SRA Taxonomy Analysis Tool

MagicBLAST

NCBI Metagenomic Resources - 2

Data Storage

SRA - Now available on the cloud

Sequence Read Archive (SRA) data, available through multiple clou
available repository of high throughput sequencing data. The archive
metagenomic and environmental surveys. SRA stores raw sequenci
and facilitate new discoveries through data analysis.

Sequence Read Archive

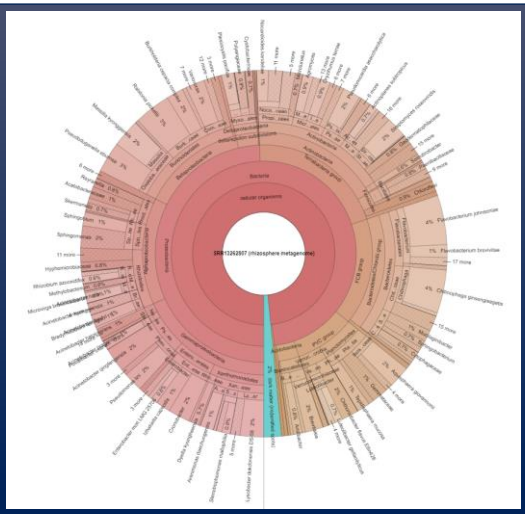
NCBI

Genbank & RefSeq

metagenome[ORGN]

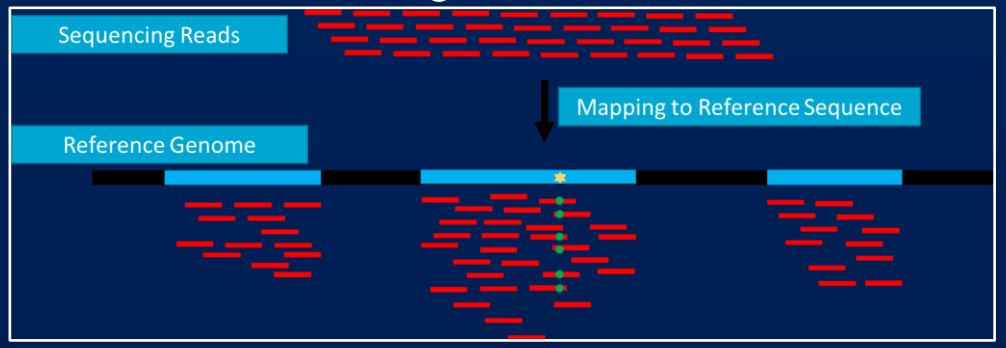
Taxonomy Keyword

Data Analysis



SRA Taxonomy Analysis Tool

MagicBLAST



Today's Case Study – Microbial Keratitis

Microbial Keratitis is a bacterial infection of the cornea (clear dome covering colored part of the eye)

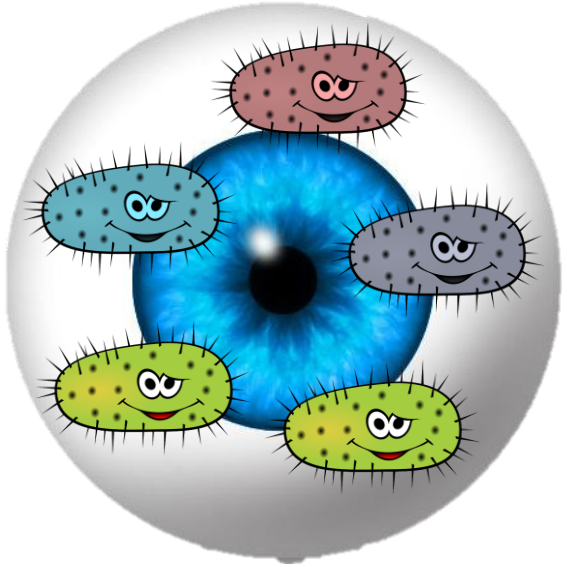
- Leading cause of preventable blindness worldwide
- Typically caused by *Pseudomonas aeruginosa, Staphylococcus aureus, Bacillus sp.*

Diagnosis is typically done via sampling and culturing of corneal samples

- Unreliable (~40% of cases are culture-negative)
- Time-consuming (~48hr turnaround + antibiotic resistance testing)

Source - https://peerj.com/articles/10778/

Today's Case Study – Methods



Corneal Scraping Community

16s rRNA Sequencing

Bioinformatic Analysis

Patient B has one infected eye, and one healthy eye.

A) Is the taxonomic distribution of each "cornea microbiome" different between eyes?

B) Do the taxonomic distributions of the eyes match our expectations for healthy and infected eyes?

Source - https://peerj.com/articles/10778/

Today's Case Study – Our objectives

<u>Objective 1 –</u> Find manuscript's original Metagenomic Reads in SRA

Bonus: Explore SRA-predicted taxonomic composition of submissions using STAT

<u>Objective 2 –</u> Use MagicBLAST to Align Reads to an NCBI Reference Database

Objective 1 - Find Metagenomic Reads in SRA and Explore Taxonomic Composition using STAT

NIH National Library of Medicine
National Center for Biotechnology Information

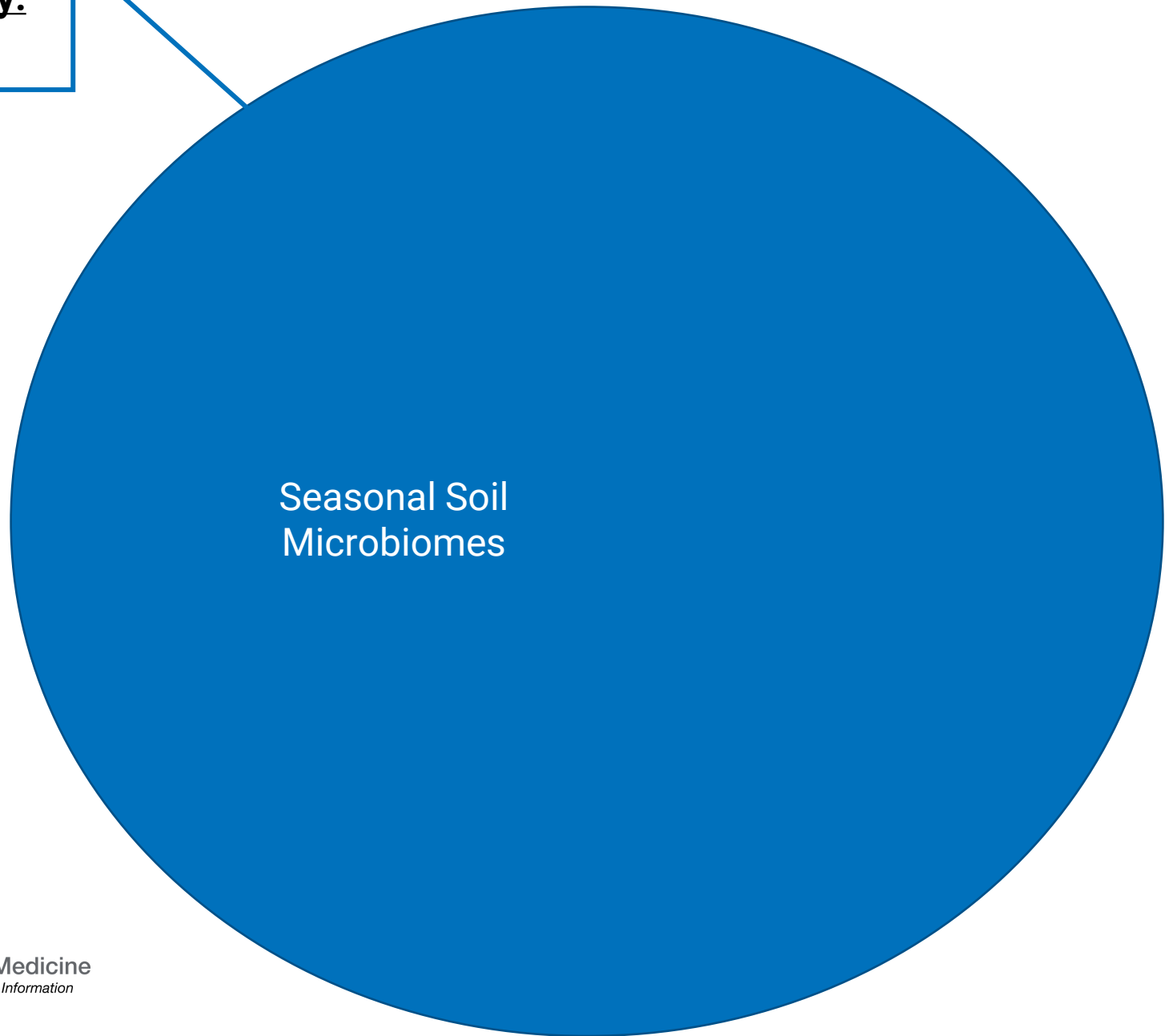
What is the Sequence Read Archive

https://www.ncbi.nlm.nih.gov/sra

- Collection of user-submitted nucleotide sequencing reads, most of which are publicly available to download
  - Current size = >10 petabytes
- You can search the data online using the URL above, or by exploring their metadata in the cloud

Structure of the SRA - 1

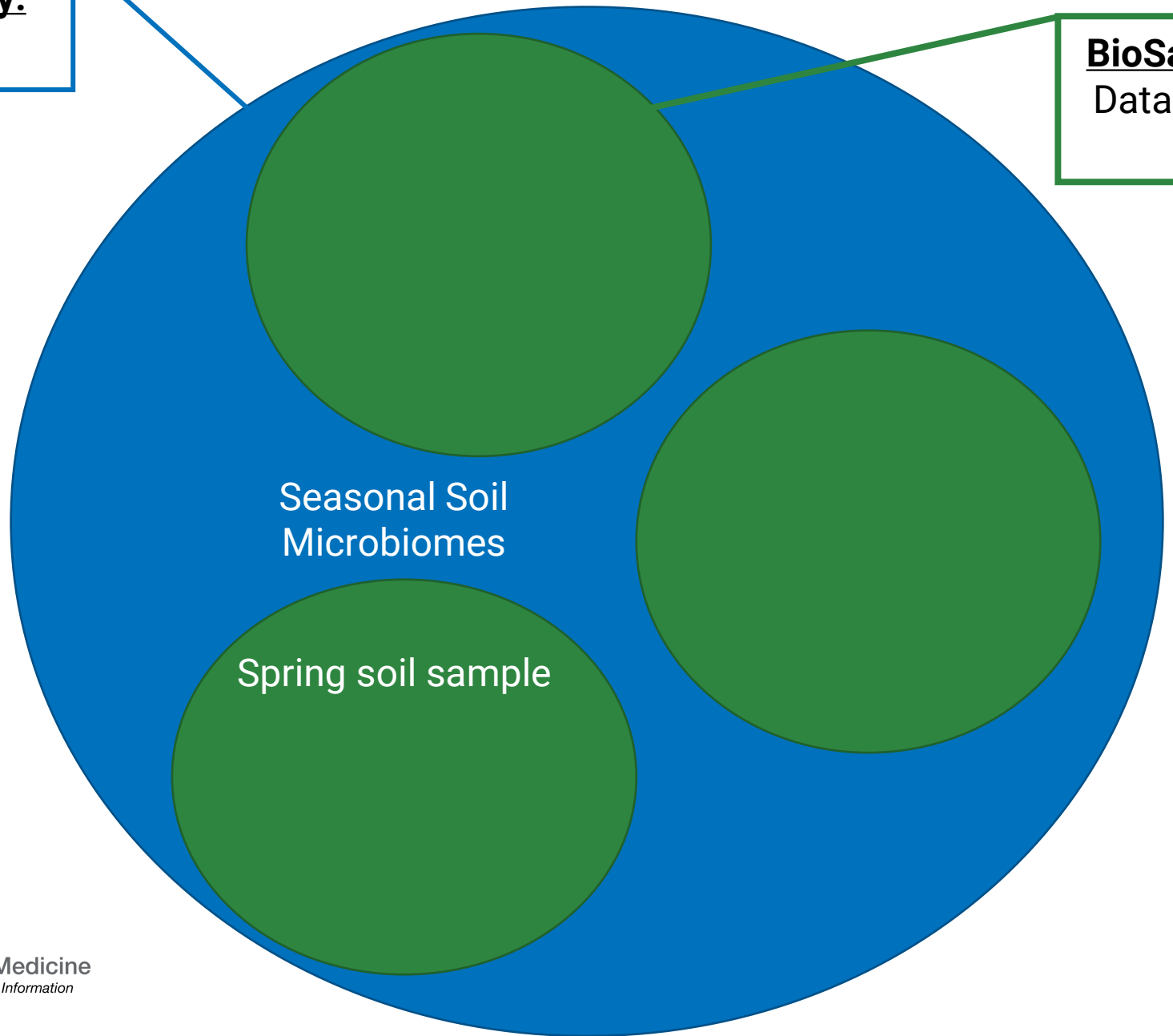
<u>BioProject / SRA Study:</u>
Data for a study



Structure of the SRA - 2

BioProject / SRA Study:
Data for a study

BioSample / SRA Sample:
Data for an individual in a study

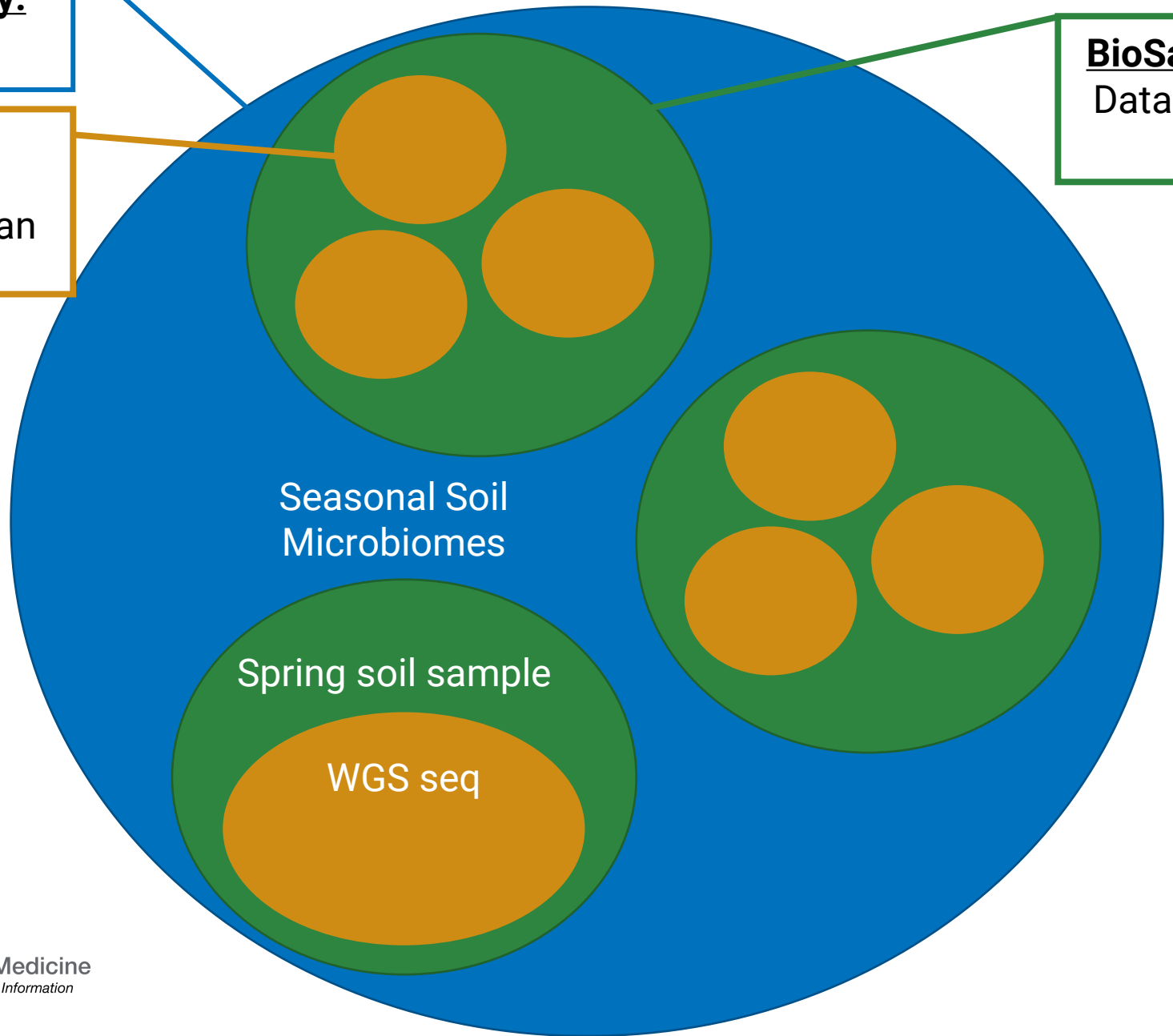


Structure of the SRA - 3

BioProject / SRA Study:
Data for a study

SRA Experiment:
Library Data for a sequencing project on an individual

BioSample / SRA Sample:
Data for an individual in a study



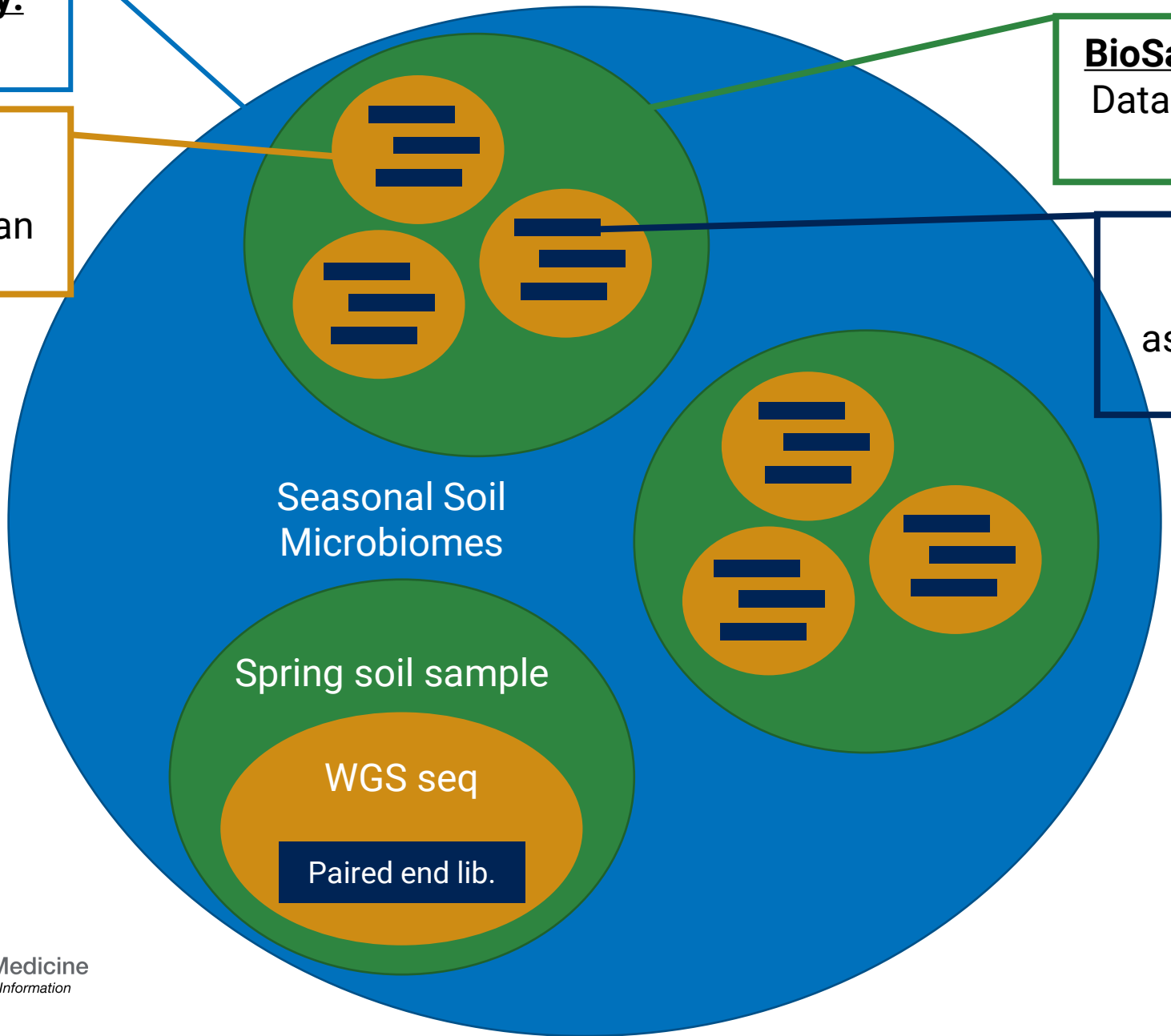
Structure of the SRA - 4

BioProject / SRA Study:
Data for a study

SRA Experiment:
Library Data for a sequencing project on an individual

BioSample / SRA Sample:
Data for an individual in a study

SRA Run:
Sequencing data associated with the SRA experiment



Finding the case study accessions - 1

*Letter depends on original collection group:

S = SRA (NCBI)
E = ERA (ENA)
D = DRA (DDBJ)

Evaluation of full-length nanopore 16S sequencing for detection of pathogens in microbial keratitis

Data Availability

The following information was supplied regarding data availability:

Bioinformatics scripts and the DNA sequencing (FASTQ) files are available at European Nucleotide Archive (PRJEB37709): SAMEA7573840, SAMEA7573841, SAMEA7573842, SAMEA7573843, SAMEA7573844, SAMEA7573845, SAMEA7573846, SAMEA7573847, SAMEA7573848, SAMEA7573849, SAMEA7573850, SAMEA7573851, SAMEA7573852, ERX4706745, ERX4706746, ERX4706747, ERX4706748, ERX4706749, ERX4706750, ERX4706751, ERX4706752, ERX4706753, ERX4706754, ERX4706755, ERX4706756, ERR4836967, ERR4836968, ERR4836969, ERR4836970, ERR4836971, ERR4836972, ERR4836973, ERR4836974, ERR4836975, ERR4836976, ERR4836977, ERR4836978, SAMEA7573853, ERX4706757, ERR4836979, SAMEA7573854, ERX4706758, ERR4836980, SAMEA7556110, ERX4692670, ERR4822680.

National Library of Medicine
National Center for Biotechnology Information

Source - https://peerj.com/articles/10778/

Finding the case study accessions - 2

*Letter depends on original collection group:

S = SRA (NCBI)
E = ERA (ENA)
D = DRA (DDBJ)

Evaluation of full-length nanopore 16S sequencing for detection of pathogens in microbial keratitis

Data Availability

The following information was supplied regarding data availability:

Bioinformatics scripts and the DNA sequencing (FASTQ) files are available at European Nucleotide Archive (PRJEB37709): SAMEA7573840, SAMEA7573841, SAMEA7573842, SAMEA7573843, SAMEA7573844, SAMEA7573845, SAMEA7573846, SAMEA7573847, SAMEA7573848, SAMEA7573849, SAMEA7573850, SAMEA7573851, SAMEA7573852, ERX4706745, ERX4706746, ERX4706747, ERX4706748, ERX4706749, ERX4706750, ERX4706751, ERX4706752, ERX4706753, ERX4706754, ERX4706755, ERX4706756, ERR4836967, ERR4836968, ERR4836969, ERR4836970, ERR4836971, ERR4836972, ERR4836973, ERR4836974, ERR4836975, ERR4836976, ERR4836977, ERR4836978, SAMEA7573853, ERX4706757, ERR4836979, SAMEA7573854, ERX4706758, ERR4836980, SAMEA7556110, ERX4692670, ERR4822680.

BioProject "PRJ*"

National Library of Medicine
National Center for Biotechnology Information

Source - https://peerj.com/articles/10778/

Finding the case study accessions - 3

*Letter depends on original collection group:

S = SRA (NCBI)
E = ERA (ENA)
D = DRA (DDBJ)

Evaluation of full-length nanopore 16S sequencing for detection of pathogens in microbial keratitis

Data Availability

The following information was supplied regarding data availability:

Bioinformatics scripts and the DNA sequencing (FASTQ) files are available at European Nucleotide Archive (PRJEB37709): SAMEA7573840, SAMEA7573841, SAMEA7573842, SAMEA7573843, SAMEA7573844, SAMEA7573845, SAMEA7573846, SAMEA7573847, SAMEA7573848, SAMEA7573849, SAMEA7573850, SAMEA7573851, SAMEA7573852, ERX4706745, ERX4706746, ERX4706747, ERX4706748, ERX4706749, ERX4706750, ERX4706751, ERX4706752, ERX4706753, ERX4706754, ERX4706755, ERX4706756, ERR4836967, ERR4836968, ERR4836969, ERR4836970, ERR4836971, ERR4836972, ERR4836973, ERR4836974, ERR4836975, ERR4836976, ERR4836977, ERR4836978, SAMEA7573853, ERX4706757, ERR4836979, SAMEA7573854, ERX4706758, ERR4836980, SAMEA7556110, ERX4692670, ERR4822680.

BioProject "PRJ*"

BioSample "SAM*"

National Library of Medicine
National Center for Biotechnology Information

Source - https://peerj.com/articles/10778/

Finding the case study accessions - 4

*Letter depends on original collection group:

S = SRA (NCBI)
E = ERA (ENA)
D = DRA (DDBJ)

Evaluation of full-length nanopore 16S sequencing for detection of pathogens in microbial keratitis

Data Availability

The following information was supplied regarding data availability:

Bioinformatics scripts and the DNA sequencing (FASTQ) files are available at European Nucleotide Archive (PRJEB37709): SAMEA7573840, SAMEA7573841, SAMEA7573842, SAMEA7573843, SAMEA7573844, SAMEA7573845, SAMEA7573846, SAMEA7573847, SAMEA7573848, SAMEA7573849, SAMEA7573850, SAMEA7573851, SAMEA7573852, ERX4706745, ERX4706746, ERX4706747, ERX4706748, ERX4706749, ERX4706750, ERX4706751, ERX4706752, ERX4706753, ERX4706754, ERX4706755, ERX4706756, ERR4836967, ERR4836968, ERR4836969, ERR4836970, ERR4836971, ERR4836972, ERR4836973, ERR4836974, ERR4836975, ERR4836976, ERR4836977, ERR4836978, SAMEA7573853, ERX4706757, ERR4836979, SAMEA7573854, ERX4706758, ERR4836980, SAMEA7556110, ERX4692670, ERR4822680.

BioProject "PRJ*"

BioSample "SAM*"

SRA Experiment "*RX"

NIH National Library of Medicine
National Center for Biotechnology Information

Source - https://peerj.com/articles/10778/

Finding the case study accessions - 5

*Letter depends on original collection group:

S = SRA (NCBI)
E = ERA (ENA)
D = DRA (DDBJ)

Evaluation of full-length nanopore 16S sequencing for detection of pathogens in microbial keratitis

Data Availability

The following information was supplied regarding data availability:

Bioinformatics scripts and the DNA sequencing (FASTQ) files are available at European Nucleotide Archive (PRJEB37709): SAMEA7573840, SAMEA7573841, SAMEA7573842, SAMEA7573843, SAMEA7573844, SAMEA7573845, SAMEA7573846, SAMEA7573847, SAMEA7573848, SAMEA7573849, SAMEA7573850, SAMEA7573851, SAMEA7573852, ERX4706745, ERX4706746, ERX4706747, ERX4706748, ERX4706749, ERX4706750, ERX4706751, ERX4706752, ERX4706753, ERX4706754, ERX4706755, ERX4706756, ERR4836967, ERR4836968, ERR4836969, ERR4836970, ERR4836971, ERR4836972, ERR4836973, ERR4836974, ERR4836975, ERR4836976, ERR4836977, ERR4836978, SAMEA7573853, ERX4706757, ERR4836979, SAMEA7573854, ERX4706758, ERR4836980, SAMEA7556110, ERX4692670, ERR4822680.

BioProject "PRJ*"

BioSample "SAM*"

SRA Experiment "*RX"

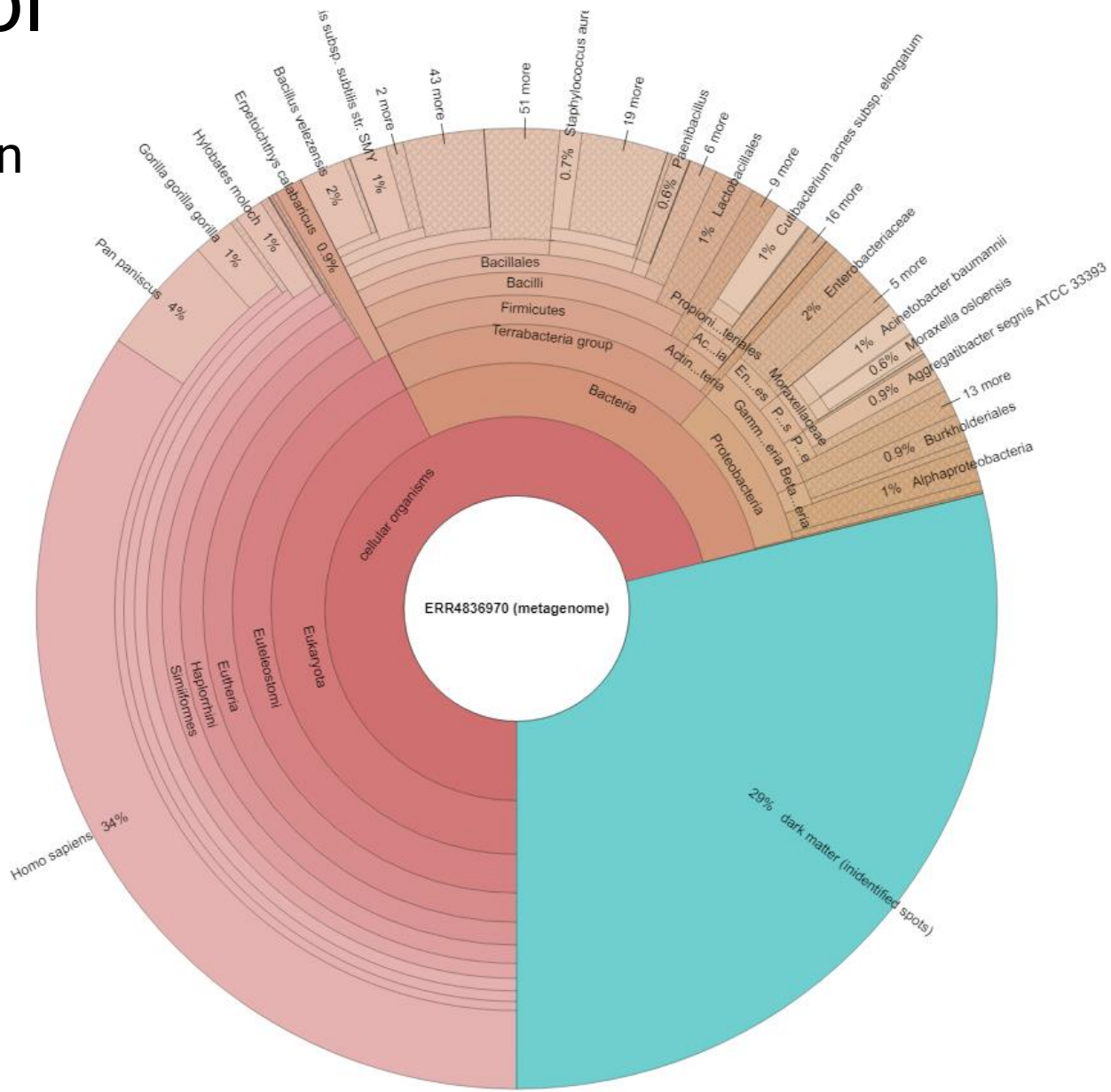
SRA Run "*RR"

National Library of Medicine
National Center for Biotechnology Information

Source - https://peerj.com/articles/10778/

SRA Taxonomy Analysis Tool

- Characterizes taxonomic distribution of reads in *every* SRA submission
  - Measured as a % of reads within the run
- Reads may be mapped to multiple taxa. If so, read is assigned to lowest common taxonomic group
  - e.g., two species share a genus, so the read is assigned to genus
- Under equal conditions, larger genomes naturally generate more reads
  - This should be considered when viewing results



Objective 1 - Goals

Practical:

- Search the NCBI website for SRA sequence data and subsequent BioSample metadata
- Use STAT to gain preliminary insights into sequence read taxonomic distribution

Case Study:

- Find sequence data associated with Patient B's unaffected and affected eye swabs
- Build a preliminary list of abundant species in each eye swab sample

Visit the “EXPLORING SRA” section of the Jupyter Notebook to get started!

Watch the chat box for the login link

Username: Email name (before the @)

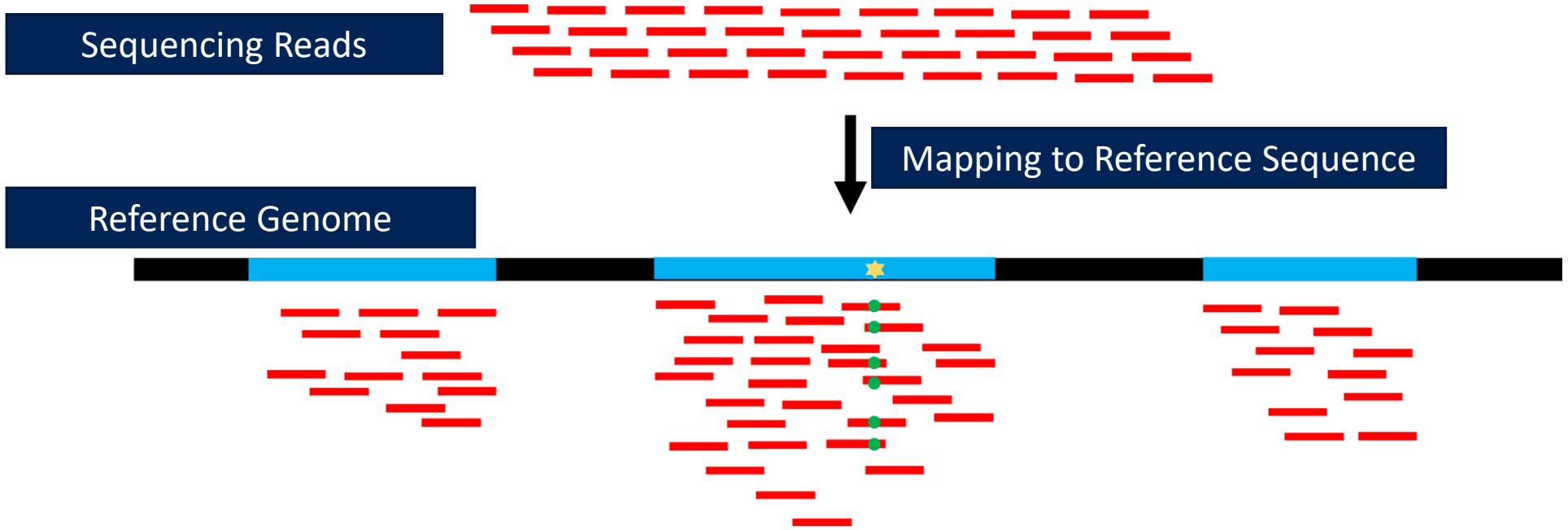
Password: <whatever you want>

Objective 2 - Use MagicBLAST to Align Metagenomic Reads Against Reference Sequences

NIH National Library of Medicine
National Center for Biotechnology Information

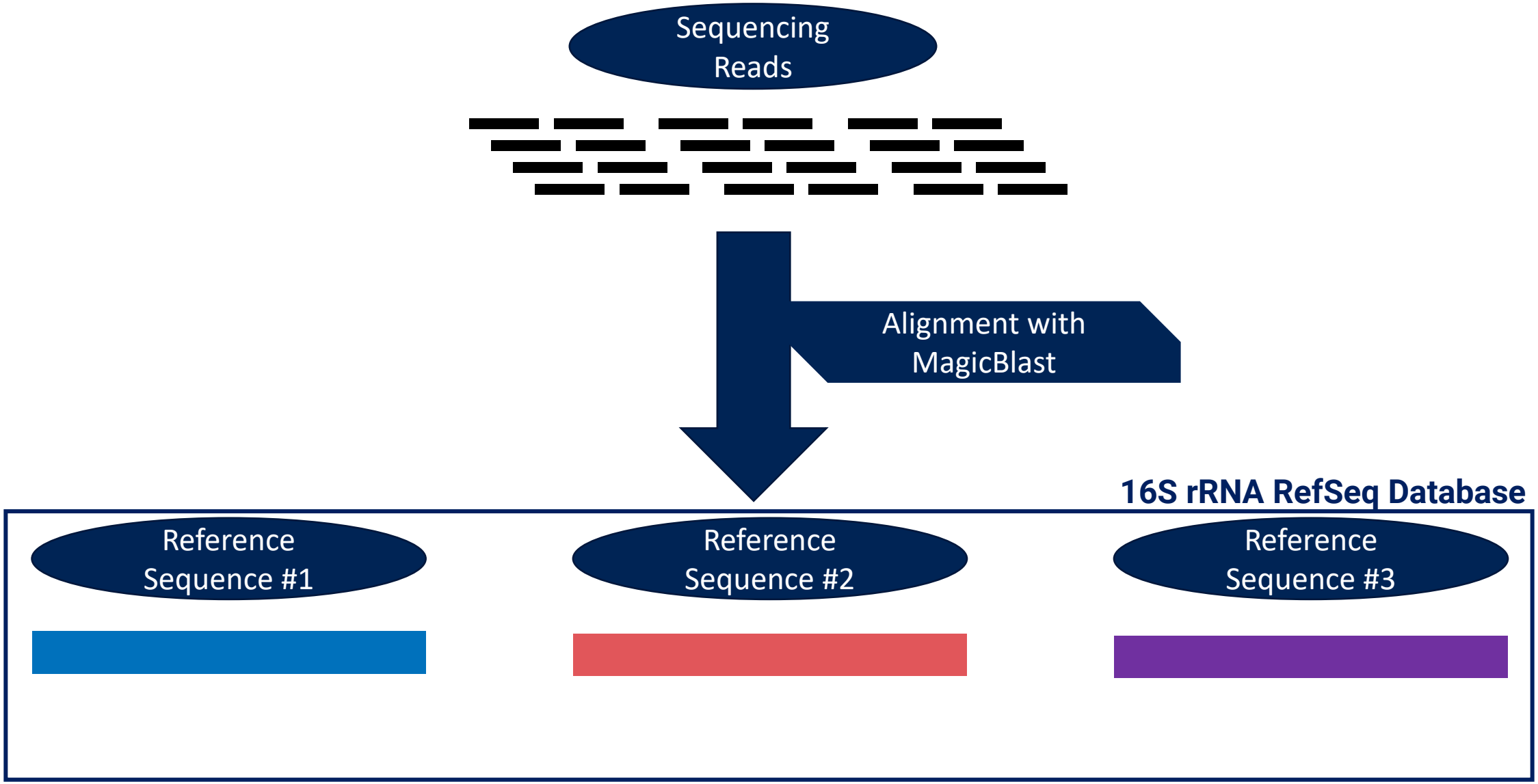
MagicBLAST

- A “flavor” of BLAST which aligns next-generation RNA or DNA sequencing reads against BLAST databases
  - Can use user-created custom databases OR NCBI maintained ones

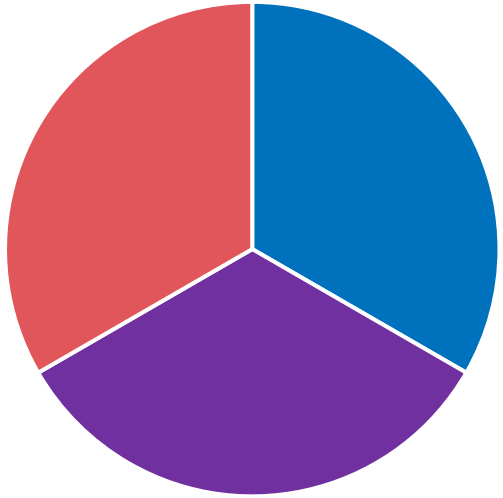


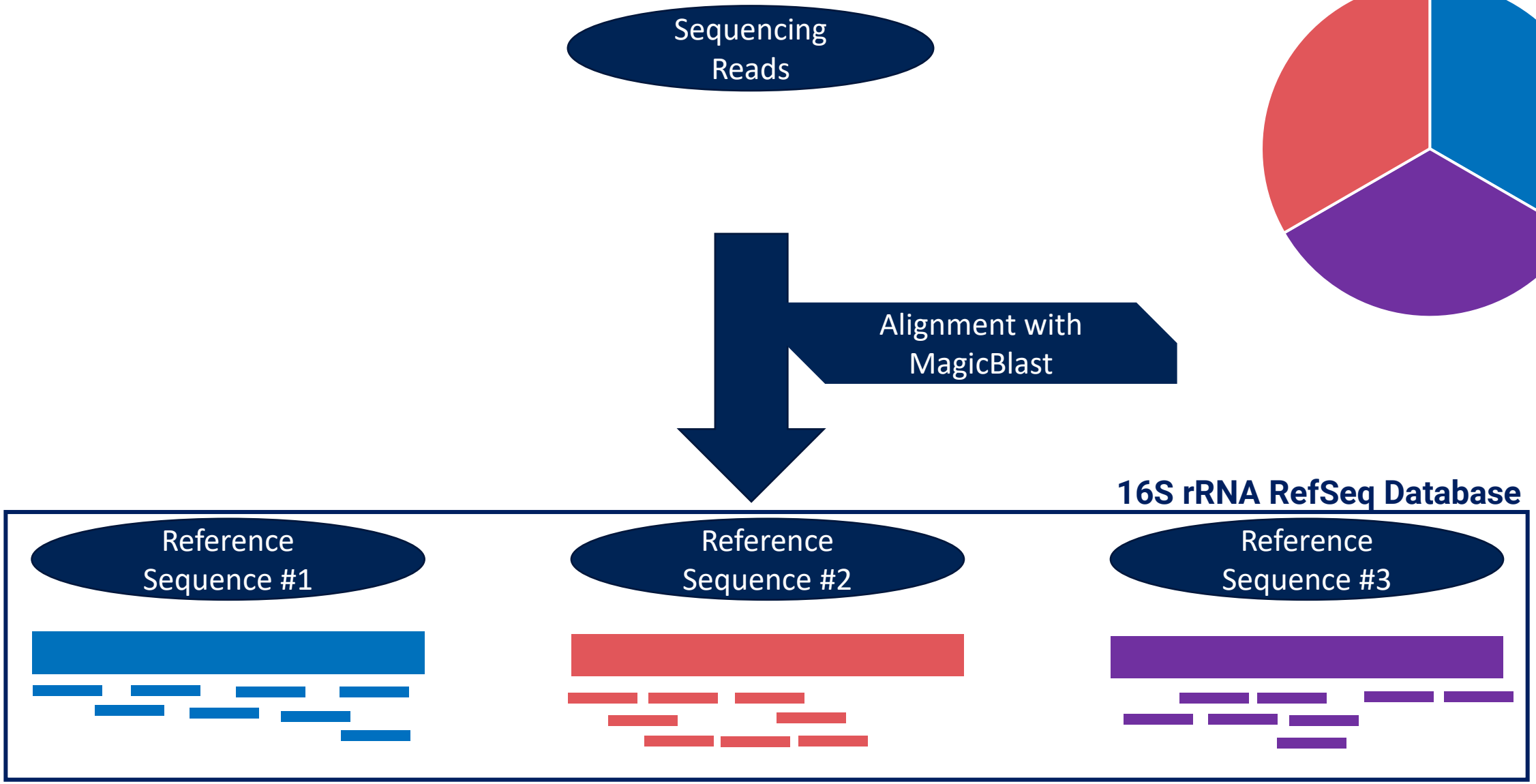
Documentation - https://ncbi.github.io/magicblast/

Metagenomes align against a collection of sequences - 1



Metagenomes align against a collection of sequences - 2





Objective 2 - Goals

Practical:

- Run MagicBLAST to align SRA reads against an NCBI database
- Compare species distribution from MagicBLAST to preliminary list gathered from STAT

Case Study:

- Characterize taxonomic content of both Patient B eye swabs using MagicBLAST
- Compare species content between unaffected and affected eyes
- Compare unaffected and affected eye species content to expected values

Visit the “ALIGNING SEQS WITH MAGICBLAST” section of the Jupyter Notebook to get started!

Advanced Metagenomics With NCBI

- Use MagicBLAST to align WGS metagenome datasets
  - Functional profiling
  - Higher accuracy taxonomic characterization
  - *Coming Soon: Clustered BLAST dbs for faster read mapping*
- Use STAT to filter SRA sequences to fit your next project
  - Explore in-depth STAT metadata in the cloud!
- Submit your sequences to SRA!
  - No excuse to provide little metadata!

Thank you!