

Introduction to metagenomics

Shirley (Xue) Li, PhD, Bioinformatician
Research Technology, TTS, Tufts University
xue.li37@tufts.edu
tts-research@tufts.edu



The Research Technology Team

- Consultation on Projects and Grants
- High Performance Cluster Support
- Workshops

<https://it.tufts.edu/bioinformatics>

<https://sites.tufts.edu/datalab/workshops/>

Tufts

Technology Services

Bioinformatics

Welcome to Bioinformatics

Tools for Life Science

We offer a range of services including bioinformatics tools on the HPC cluster, secondary analysis pipelines for NGS data including DNA-seq, RNA-seq, and CHIP-seq, data visualization, and training and consultation!

Overview

01. Introduction to Metagenomics

Defining Metagenomics

Contrast with Traditional Microbiological Techniques

02. Applications of Metagenomics

Human Health, food Industry

03. Technological Foundations

High-Throughput Sequencing Methods

04. Data Analysis in Metagenomics

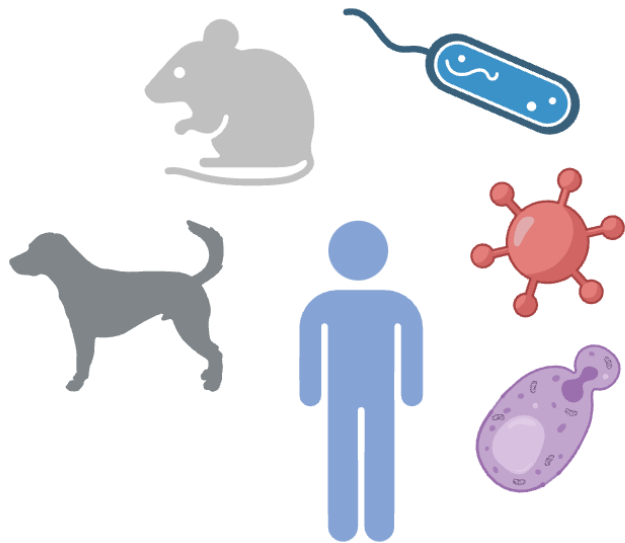
Bioinformatic Tools for Sequence Analysis

05. Metagenomics Hands-on session



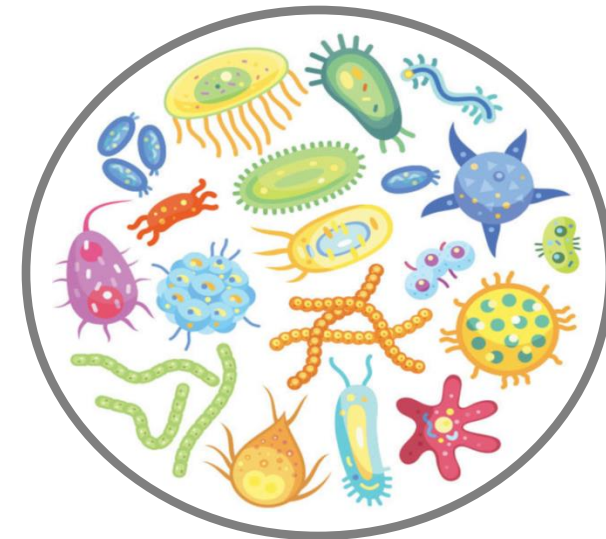
01. Introduction to Metagenomics

Genome



Genetic makeup of an individual organism

Metagenome



A collection of genomes from many individual microorganisms within a sample

Metagenomics

Metagenomics is the study of the structure and function of entire nucleotide sequences isolated and analyzed from all the organisms (typically microbes) in a bulk sample.

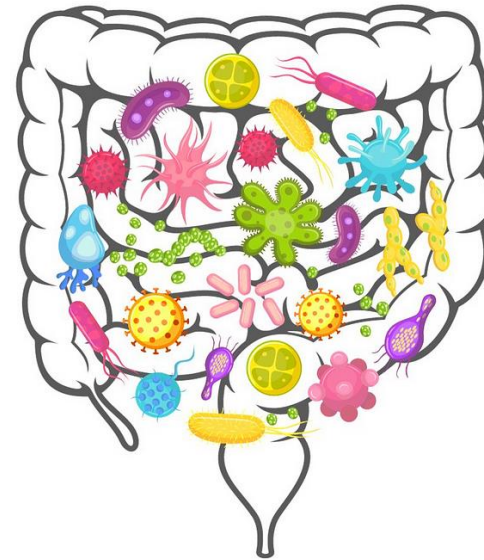
Metagenomics is often used to study a specific community of microorganisms, such as those residing on human skin, in the soil or in a water sample.

Microbiome

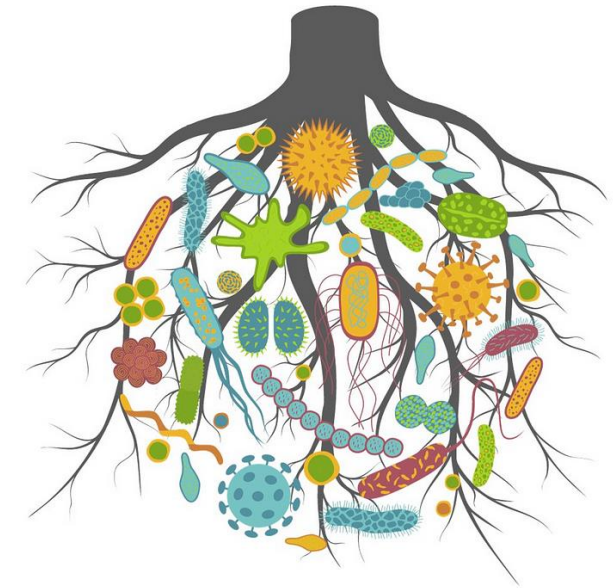
The microbiome refers to the entire habitat, including the **microorganisms** (bacteria, viruses, fungi, and archaea), their **genomes**, and the surrounding **environmental conditions**.

It's a broader term that encompasses the living organisms and their interactions with each other and with their environment.

Gut Microbiome



Root Microbiome



<https://medium.com/illumination/gut-microbiome-soil-microbiome-different-ecosystems-same-principles-2231ae0637a>

Metagenomics

Microbiome

Definition

Direct sequencing of the collective genome of all microorganisms present in an environmental sample.

The entire community of microorganisms living in a particular environment, including their genomes, interactions, and the environment itself.

Focus

On the genetic material itself, understand the diversity, function, and dynamics of microbial communities based on their DNA.

On the study of microbial communities, their composition, functions, and interactions within their host or in their natural **environment**.

Approaches

DNA sequencing, functional annotation, comparative study, ...

Metagenomics, metatranscriptomics, metaproteomics, metabolomics,...

Key Questions in Metagenomic Study

Who is present in the sample?

This involves determining the various microorganisms in the sample, including bacteria, viruses, fungi, etc.

What are their relative abundances?

This question addresses the quantification of different microbes, helping to understand the dominance or rarity of certain species within the community.

Why do abundance levels vary among species, and what are their functional roles?

Investigate the reasons behind the varying abundance of different bacteria. This includes exploring the functional genomics of the community to understand the ecological roles and interactions of these microbes.

Microbiome – How we Study Them

Traditionally, the microbiome was studied by collecting a sample and growing those microbes on a petri dish.

There we could assess what the community was composed of, observing the different types of colonies that formed, each representing a unique microorganism.



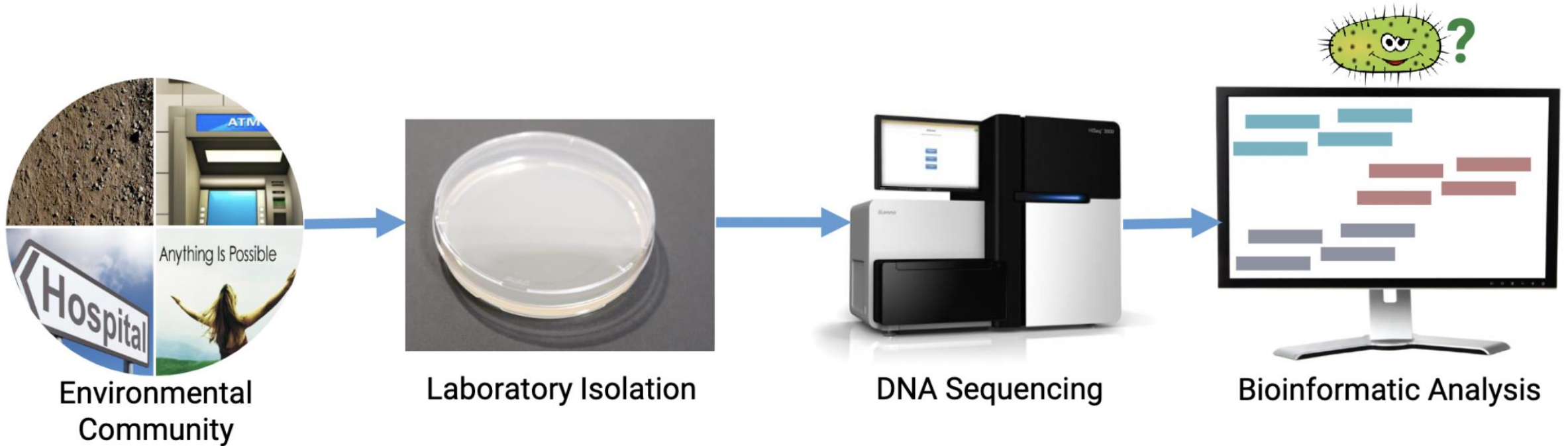
Microbes from a child's hand after playing outside illustrate our close connection with the microbial world within us, on us, and around us.

Source: Tasha Sturm at Cabrillo College via ASM's MicrobeWorld.

<https://asm.org/Articles/2019/March/Microbiomes-An-Origin-Story>

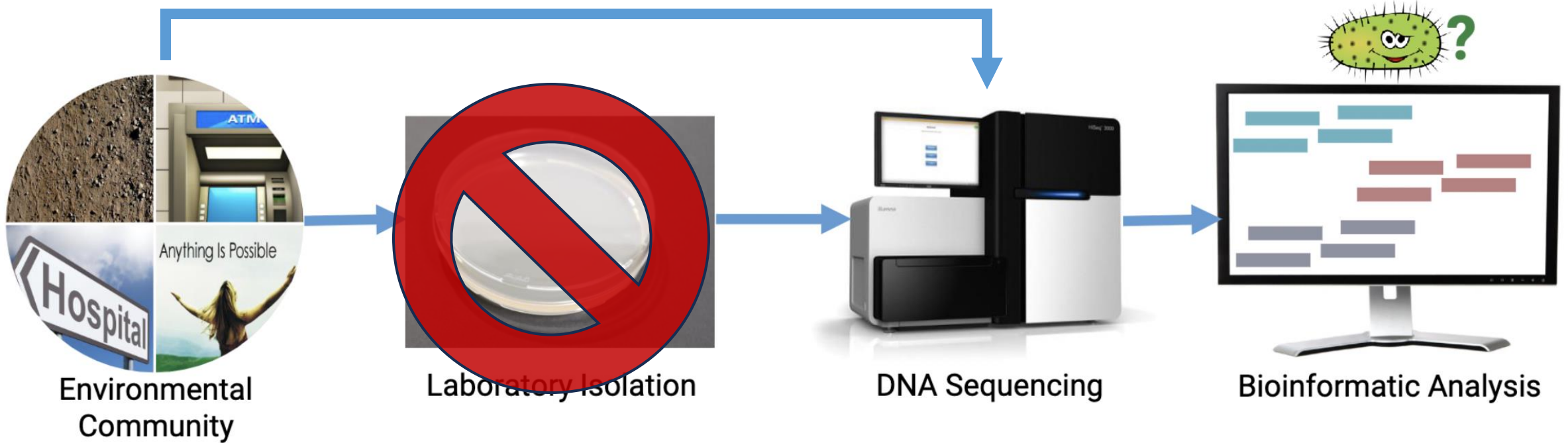
https://www.nlm.nih.gov/oet/ed/ncbi/2021_10_meta.html

Microbiome – How we Study Them



https://www.nlm.nih.gov/oet/ed/ncbi/2021_10_meta.html

Microbiome – How we Study Them



Metagenomics is sequencing without culturing

https://www.nlm.nih.gov/oet/ed/ncbi/2021_10_meta.html

Metagenomics vs. Traditional microbiological approaches

Culture Independent

Directly analyzes genetic material from environmental samples.

Comprehensive Community Analysis

Provides a broad overview of all organisms present, culturable or not.

Speed and scale

High-throughput sequencing technologies have made metagenomics a rapid method for analyzing microbial communities.

Functional potential

Offers insights into the metabolic capabilities and interactions within microbial communities.

02. Applications of Metagenomics


Example One

Metagenomics in human health

[nature](#) > [articles](#) > [article](#)

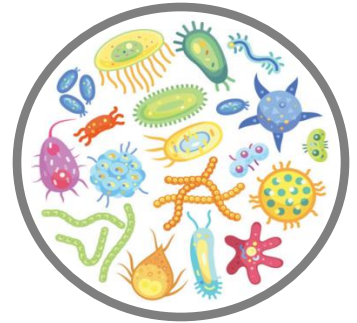
Article | [Published: 21 December 2006](#)

An obesity-associated gut microbiome with increased capacity for energy harvest

[Peter J. Turnbaugh](#), [Ruth E. Ley](#), [Michael A. Mahowald](#), [Vincent Magrini](#), [Elaine R. Mardis](#) & [Jeffrey I. Gordon](#) 

[Nature](#) **444**, 1027–1031 (2006) | [Cite this article](#)

121k Accesses | **8561** Citations | **1117** Altmetric | [Metrics](#)



<https://www.nature.com/articles/nature05414>

Example Two

Metagenomics in food industry

[nature](#) > [nature communications](#) > [articles](#) > [article](#)

Article | [Open access](#) | [Published: 21 December 2023](#)

Microbial interactions shape cheese flavour formation

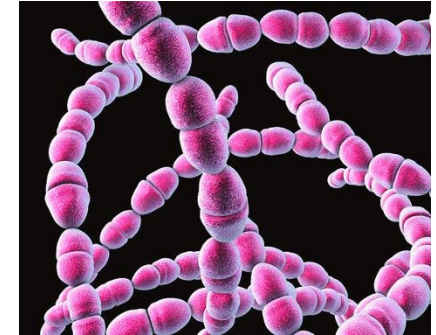
[Chrats Melkonian](#) ✉, [Francisco Zorrilla](#), [Inge Kjærbølling](#), [Sonja Blasche](#), [Daniel Machado](#), [Mette Junge](#), [Kim Ib Sørensen](#), [Lene Tranberg Andersen](#), [Kiran R. Patil](#) & [Ahmad A. Zeidan](#) ✉

[Nature Communications](#) **14**, Article number: 8348 (2023) | [Cite this article](#)

6588 Accesses | 1 Citations | 172 Altmetric | [Metrics](#)

<https://www.nature.com/articles/s41467-023-41059-2>

streptococcus thermophilus bacteria



Other Applications

Pharmaceuticals



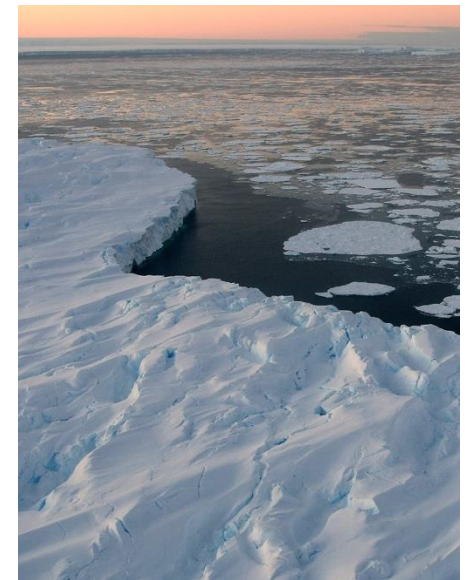
Marine biology



Agriculture

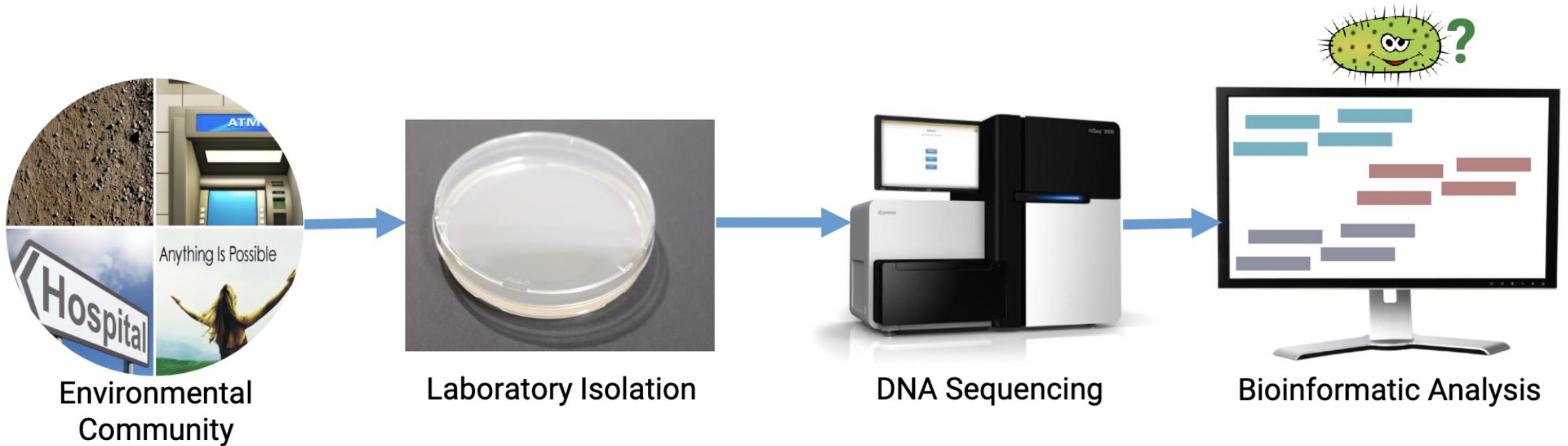


Climate change research



03. Technological Foundations

Microbiomes – How we Study Them



Metagenomics is sequencing without culturing

https://www.nlm.nih.gov/oet/ed/ncbi/2021_10_meta.html

Microbiomes – How we Study Them



DNA Sequencing

https://www.nlm.nih.gov/oet/ed/ncbi/2021_10_meta.html

Microbiomes – How we Study Them

Illumina HiSeq 3000



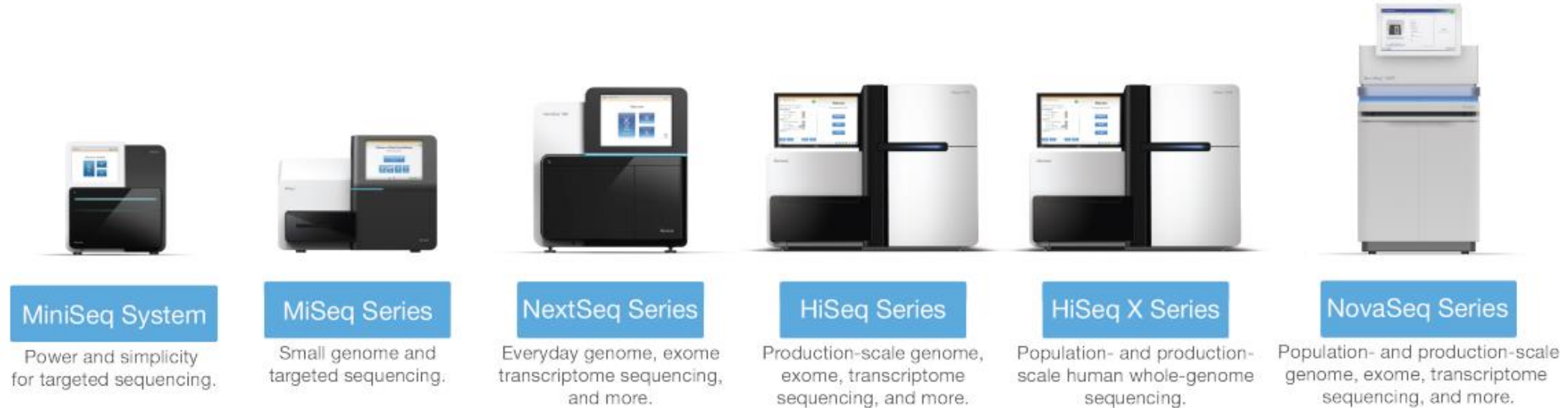
DNA Sequencing

https://www.nlm.nih.gov/oet/ed/ncbi/2021_10_meta.html

Overview of Illumina Sequencing Platforms

Next-generation sequencing (NGS)

$\leq 300\text{bp}$
Error rate: $< 0.1\%$



Second-generation sequencing, also known as **next-generation sequencing (NGS)**, revolutionized genomics with its **high-throughput** capabilities, allowing for the sequencing of large volumes of DNA fragments simultaneously, albeit with **shorter read lengths ($\leq 300\text{bp}$)**.

Third-generation sequencing



15-20kb
Error rate: <0.1%

PacBio Sequel System



< 4Mb
Error rate: ~1%

Nanopore product family

Third-generation sequencing uniquely enables the direct analysis of long DNA sequences and complex genomic regions, facilitating detailed studies of genomic structure, epigenetic modifications, and previously inaccessible aspects of genome biology.

The power of minION sequencing



Portability: The compact size of the MinION enables on-site sequencing in remote field studies, bringing genomic research capabilities directly to the source of samples.

Real-Time Sequencing: It delivers immediate sequencing results, crucial for rapid clinical diagnosis and timely decision-making in patient care.

Sequencing methods in metagenomics

The most commonly used platforms are from Illumina, such as the HiSeq and MiSeq systems, which are second-generation sequencers known for providing billions of short reads with high accuracy.



MiSeq Series

Small genome and targeted sequencing.



HiSeq Series

Production-scale genome, exome, transcriptome sequencing, and more.

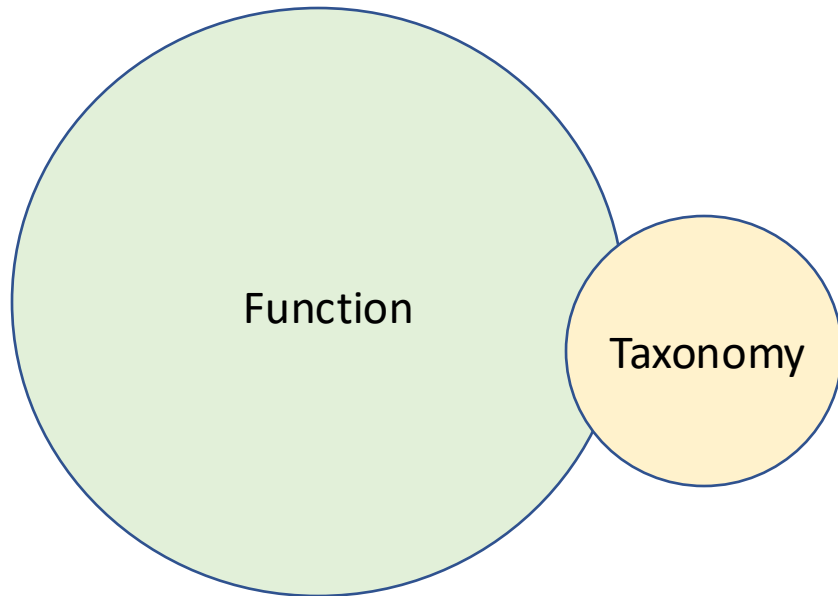
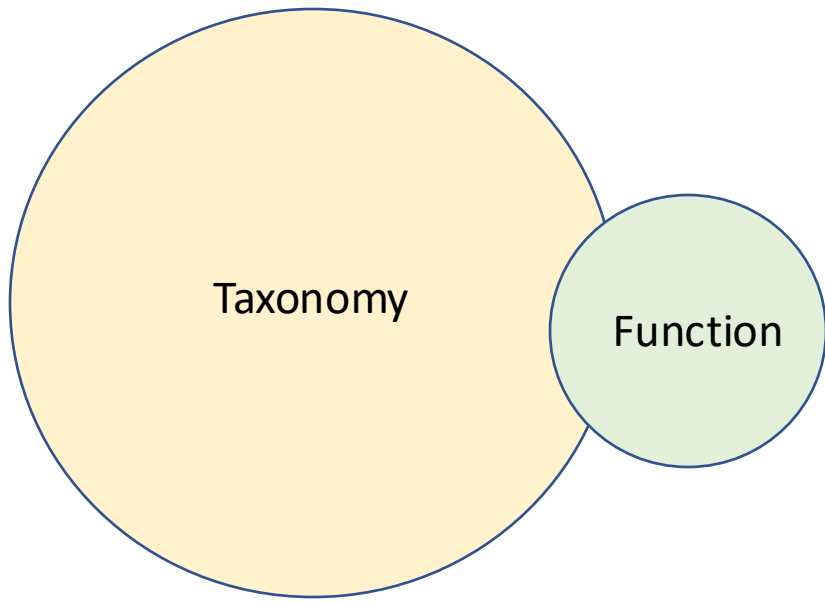
Two main approaches to profiling the microbiome

Metabarcoding/amplicon sequencing/16S rRNA sequencing/18S rRNA

- Any region that is conserved but also unique to each species.
- Conserved sequences (amplicons) that are highly informative on taxonomy
- Less expensive because only a small portion of the genome is sequenced

Metagenomics/whole-genome sequencing (WGS)

- Sequencing the entire genome to find what genes are present
- Becoming more common as sequencing costs decline
- More complex computationally



Metabarcoding

- Amplicon-based sequencing (primers designed to limit sequencing to a small section of the genome)
- Taxonomy inferred from databases of conserved sequences
- Function inferred from what is already known about the organism or type of organism in the sample

Whole genome sequencing

- Whole-genome shotgun (WGS) sequencing (everything)
- Function inferred from the list of genes present
- Taxonomy can be extracted from the sequence as well by finding the same conserved sequences identified for Metabarcoding

04. Data analysis in metagenomics

Types of analysis in metagenomics studies

Taxonomic assignment (Today's in-class exercise)

- **Identifying and classifying microorganisms within a sample to their respective taxa.**

Functional Annotation

- Assigning predicted genes and proteins to known functions.
- Analysis of metabolic pathways and potential biochemical activities within the microbial community.

Comparative Metagenomics

- Comparing the metagenomic profiles of different samples to understand differences in microbial community structure and function.
- Assessing the impact of environmental factors, host characteristics, or treatments on microbial communities.

Types of analysis in meta-omics studies

Metatranscriptomic Analysis

Studying the RNA transcripts to understand the active metabolic processes and functions being expressed by the community.

Metaproteomic Analysis

Analyzing the protein complement of the sample to get insights into active enzymes and pathways.

Metabolomic Analysis

Profiling the small molecule metabolites present in the community, providing functional evidence of metabolic activity.

Bioinformatics Tools to Analyze Metagenomics Data

Kraken 2

Taxonomic Sequence Classification System



Mothur

Kraken2, Mothur and Qiime2 on Tufts Cluster

Mothur

`mothur/1.46.0`

`mothur/1.47.0`

`mothur/1.48.0`

Qiime2

`qiime2/2023.2`

`qiime2/2023.5`

`qiime2/2023.7`

`qiime2/2023.9`

Kraken2

`/cluster/tufts/bio/tools/conda_envs/kraken/2.1.2/bin/kraken2`

Questions?

05. Metagenomics Hands-on session

https://github.com/shirleyxueli41/Tufts_workshops/blob/main/IGDH-1001_2024Feb/Hands-on%20session.md

<https://go.tufts.edu/idgh1001>

Exercise 1: Navigate the NCBI database to master the ability to find published raw datasets and get familiarize with BioProject pages, SRA experiment pages, and SRA runs.

Exercise 2: Navigate various sections of the database and applying your understanding of sequencing technologies to hypothetical research questions.

What is the Sequence Read Archive (SRA)

- Collection of user-submitted nucleotide sequencing reads, most of which are publicly available to download
- Link To SRA:
 - <https://www.ncbi.nlm.nih.gov/sra>
- Currently the SRA is over 23 petabytes
- These sequencing reads are stored within containers called BioProjects



BioProject

Stores the study data (e.g., Study of seasonal microbiome profile changes)

BioSample

Stores data for an individual in a study

Spring soil metagenome sample

SRA Experiment

Library data for a sequencing project on an individual

WGS Sequencing

Transcriptome Sequencing

SRA Run

Stores sequence data

WGS Run 1

WGS Run 2

RNAseq Run 1

RNAseq Run 2

Finding the case study accessions

*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](#).

Finding the case study accessions

*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

BioProject
"PRJ*"

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](#).

Finding the case study accessions

*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*"

Finding the case study accessions

*Letter depends on original collection group:

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

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.

BioSample
"SAM*"

BioProject
"PRJ*"

SRA Experiment
"*RX"

Finding the case study accessions

*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.

BioSample
"SAM*"

SRA Run
"*RX"

BioProject
"PRJ*"

SRA Experiment
"*RX"

Exercise 3: Taxonomy assignment and interpretation.
Taxonomy classification with Kraken2 tools on Tufts Galaxy.

Exercise 4: Taxonomy visualization with Krona plot.

Running Kraken2 on Tufts Galaxy and HPC Command Line

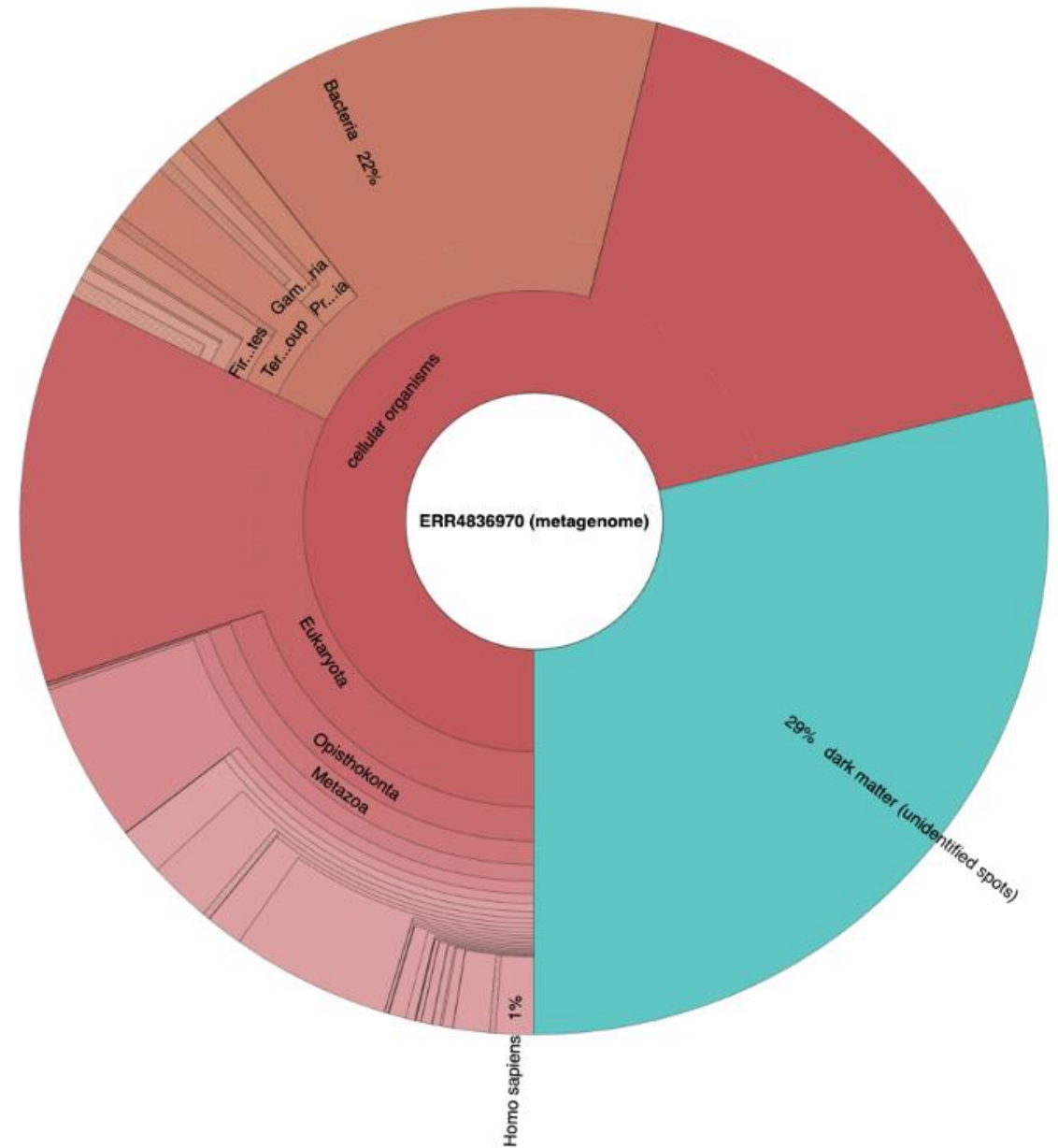
In exercise 3 & 4, we will run Kraken2 on Tufts Galaxy

To run kraken2 using HPC command line tool, check the previous workshops

https://github.com/tuftsdatalab/tuftsWorkshops/blob/main/docs/2023_workshops/metagenomeData/03_kraken.md

KRONA Plots

- KRONA plots are a way of visualizing taxonomic data in a sample
- Essentially it is a pie chart of taxonomic data
- Each “slice” represents a different taxa and you can click each slice to get the composition of organisms under that taxa
- For example, if we click on bacteria, we will see which bacteria are present in our sample



References

- <https://training.galaxyproject.org/training-material/topics/metagenomics/faqs/kraken.html>
- <https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1891-0>
- https://www.nlm.nih.gov/oet/ed/ncbi/2021_10_meta.html