Metagenome analysis

Jojy John
Post-Doctoral Fellow

Advisor: Prof. Barbara J Campbell

Campbell Lab
Department of Biological Sciences
Clemson University

Key Takeaways

Questions:

- How to analyze metagenomics data?
- What information can be extracted of metagenomics data?
- What is the difference between amplicon and shotgun data?
- What are the difference in the analyses of amplicon and shotgun data?

Source : https://www.youtube.com/watch?v=RcYXTpNS_XU

Objectives:

- Choosing the best approach to analyze metagenomics data
- Selection of tools to analyze amplicon data or shotgun data

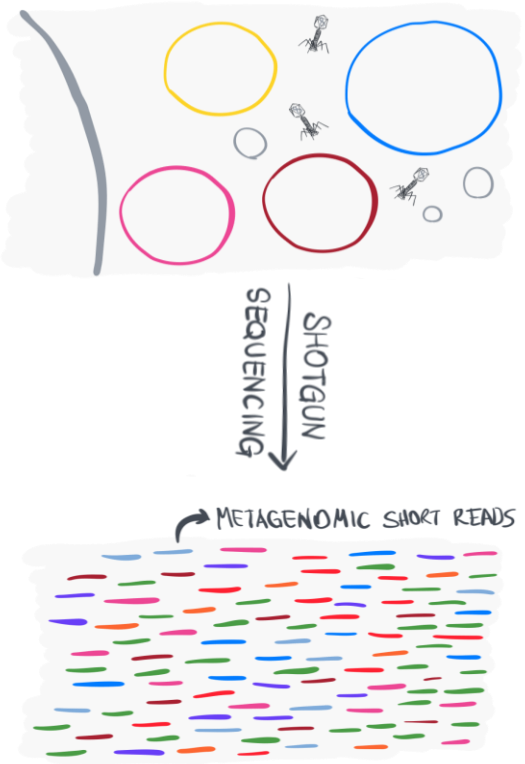
Source: https://training.galaxyproject.org/training-material/topics/microbiome/tutorials/general-tutorial/tutorial.html#shotgun-metagenomics-data

Short Introduction

Metagenome is the entire DNA content of an environment

In shotgun metagenome **full genomes of the micro-organisms** in the environment are sequenced.

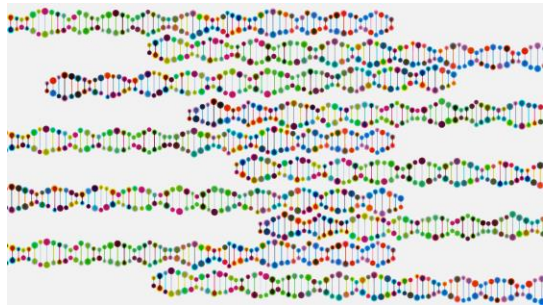
We can have **access to the all genes** of the micro-organisms.



Who is there?

What are they doing?

Metagenome Vocabulary



Sequencing Reads: Raw output of a sequencer. These are strings of the alphabet {A, C, T, G}, representing nucleotide sequences

Contig: A contiguous segment of DNA that is often 'assembled' from short reads or long reads

Metagenomic binning :A set of computational strategies that aims to identify and put together contigs that belong to the same population.

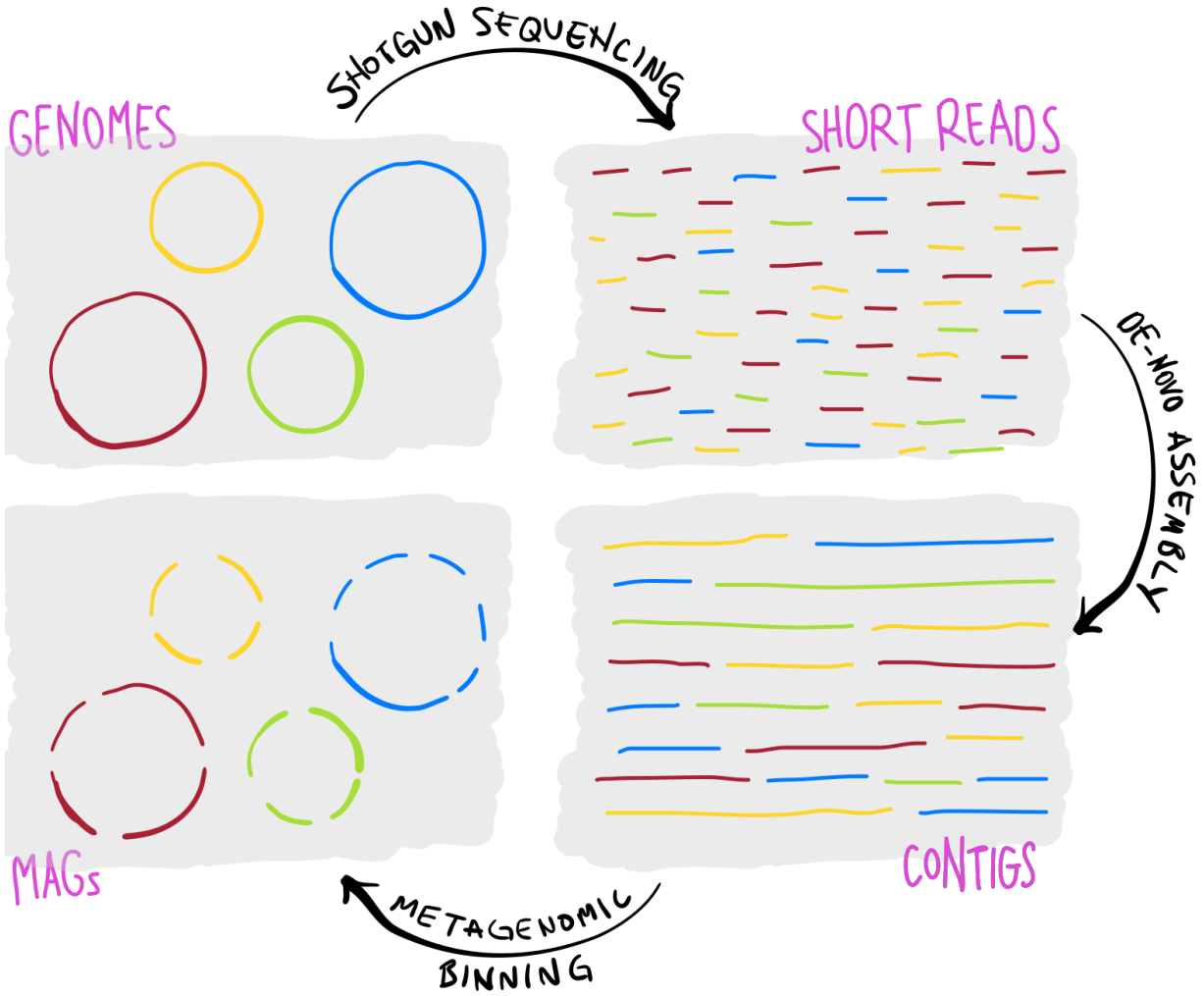
Metagenome-assembled genome (MAG): A genome that is reconstructed or recovered from a metagenome.A MAG can be a single contig or a collection of contigs that, collectively represent a single microbial organism

De novo assembly: The process of assembly aims to extend short reads into longer contiguous segments of DNA or RNA (i.e., contigs) to reconstruct the original sequence

Source: https://anvio.org/vocabulary/

Genome-resolved metagenomics

A family of strategies that generate genome-level insights from metagenomes

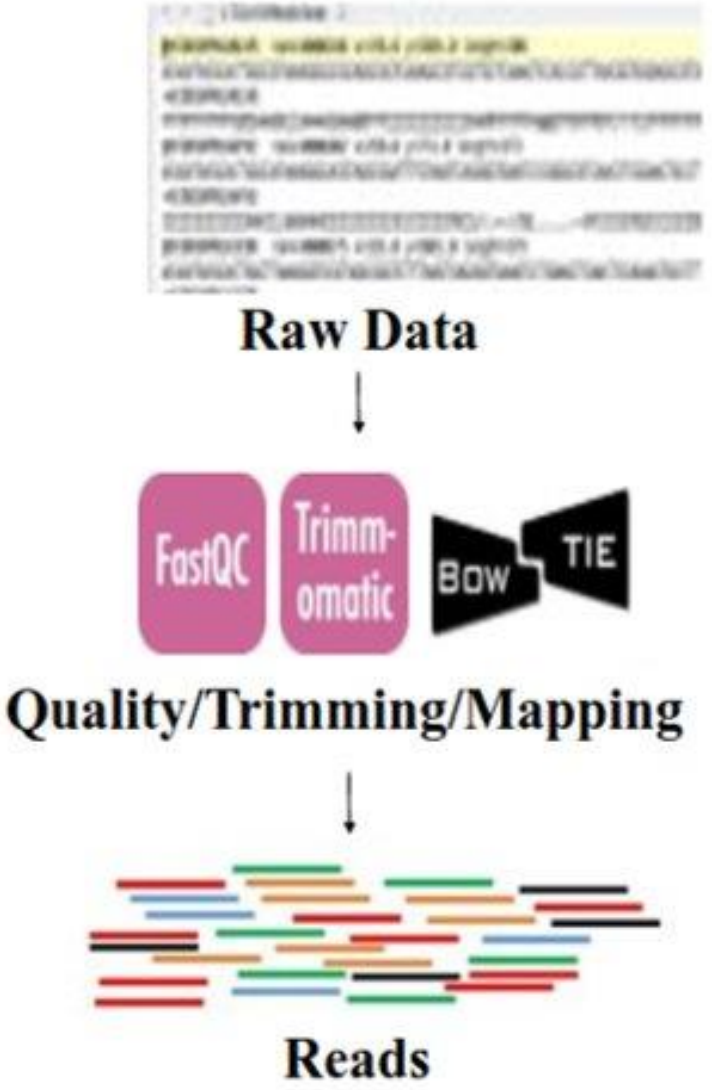


Quality Control

The first step in any analysis should be to check and improve the quality of our data.

Poor QC leads to:

- ❌ False taxa detection
- ❌ Inflated diversity
- ❌ Bad assemblies
- ❌ Wrong biological conclusions



Software Tools Covered

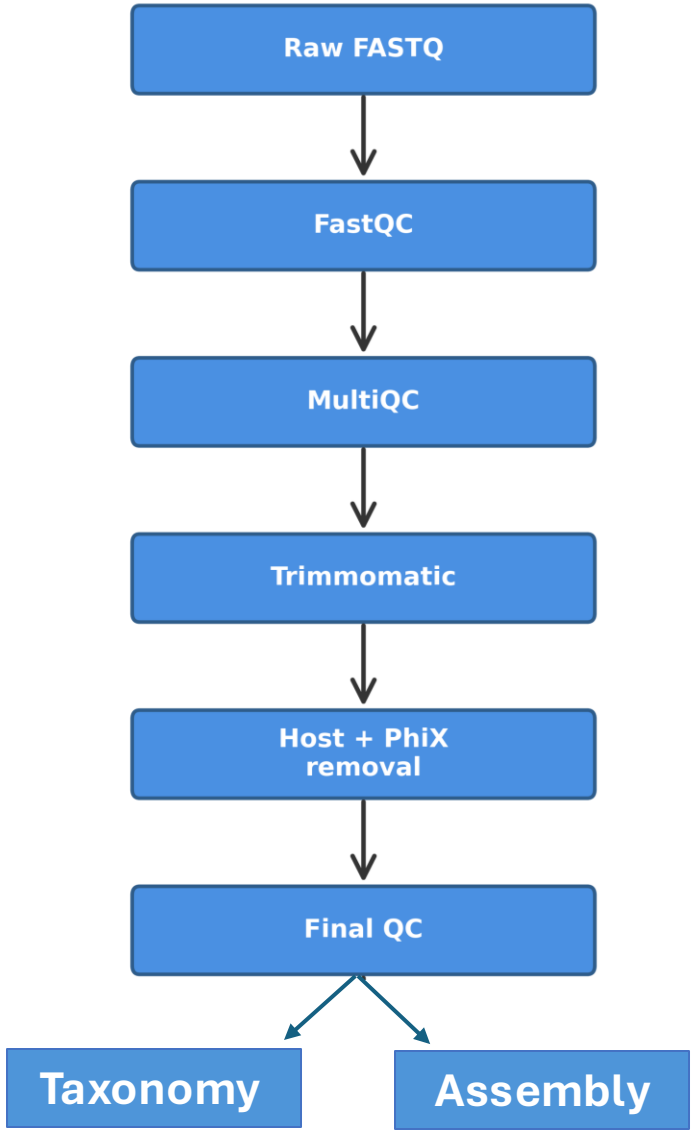
- **FastQC** – Quality control (practical session)
- **MultiQC** – QC report aggregation (theory only)
- **Trimmomatic** – Read trimming and filtering (practical session)
- **KneadData** – Host read removal (demonstration)

Quality Control-FastQC/MultiQC

FastQC :It runs a set of analyses on one or more raw sequence files in fastq or bam format and produces a report which summarises the results.

It checks:

- Per-base quality
- GC content
- Duplication levels
- Adapter contamination
- Overrepresented sequences



MultiQC: Aggregating Reports _theory only

- FastQC gives per-sample reports.
- MultiQC combines them.

Understanding FastQC: Per Base Quality

Phred Score Interpretation

Score	Meaning
>30	Excellent
20–30	Acceptable
<20	Poor

What to look for:

- Quality drops at 3’ end
- Red zone = trimming required

1 Per Sequence Quality

- Most reads should have mean >30

2 GC Content

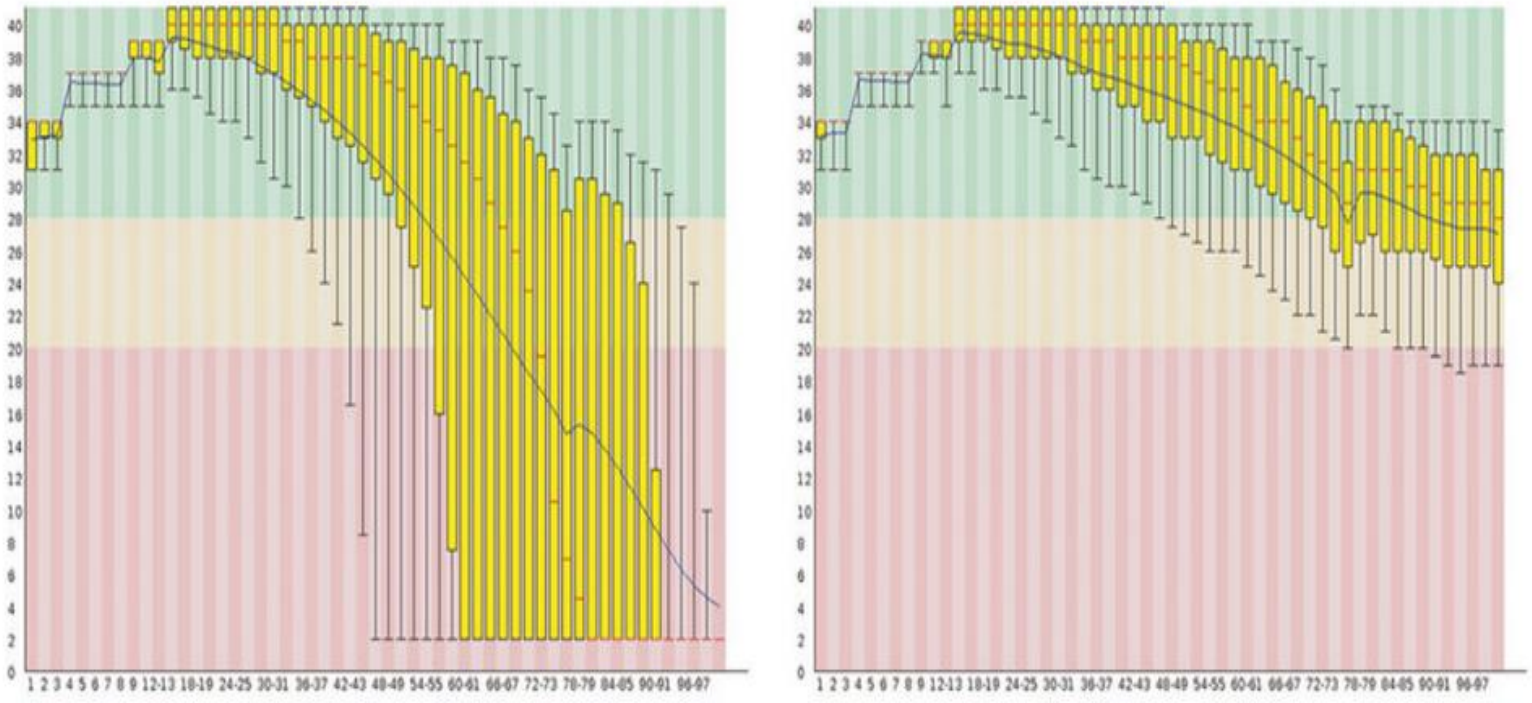
- Random library ≈ 25% each base
- Metagenomes: Some variation is normal

3 Duplication

- High duplication = PCR bias or dominant organisms

4 Adapter Content

- 5% → Must trim



Left: Raw reads with severe quality drop toward read ends
Right: Trimmed reads with stabilized base quality (These plots clearly show why trimming is required before assembly.)

- **X-axis:** Position in the read (base pairs)
- **Y-axis:** Phred quality score

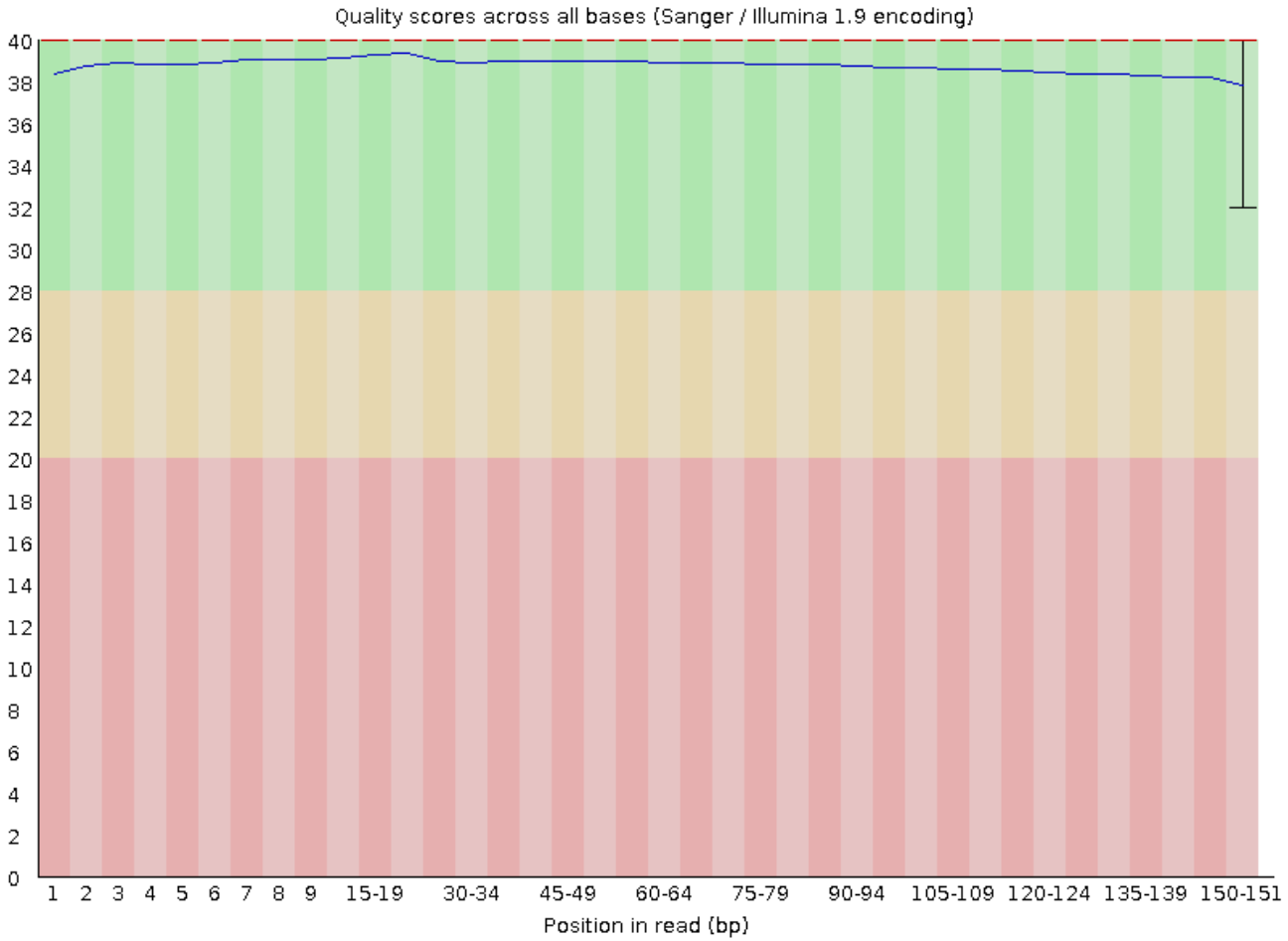
Boxplot Elements

- **Red line inside the box** → *Median quality score*
- **Yellow box** → *Interquartile range (25th–75th percentile)*
- **Whiskers** → *10th and 90th percentile values*
- **Blue line** → *Mean quality score across reads*

Background Colors

- 🟢 **Green zone (Q ≥ 28)** → High quality
- 🟡 **Orange zone (Q 20–28)** → Moderate quality
- 🔴 **Red zone (Q < 20)** → Poor quality

Now we will assess the quality of our metagenomic data. Open *metagenome_practical.md* and follow along.



- Quality remains in the **green zone across all positions**
- Median and mean are both ~Q38–39 (excellent quality)
- Very small spread (tight yellow boxes) → low variability
- Slight drop at the 3' end (normal for Illumina data)
- No bases fall into warning or fail zones

Adapter Trimming (Trimmomatic)

ŘƎÍĢĴĂĚĢĪĪĤÌJĂĨÌĂĖĢĨĨĤĦĄEÌĦĂĒĨĨ¦ÍÌIJD

1 Remove Adapters

- Sequencing machine adds adapters
- These must be removed before analysis

2 Trim Low-Quality Ends

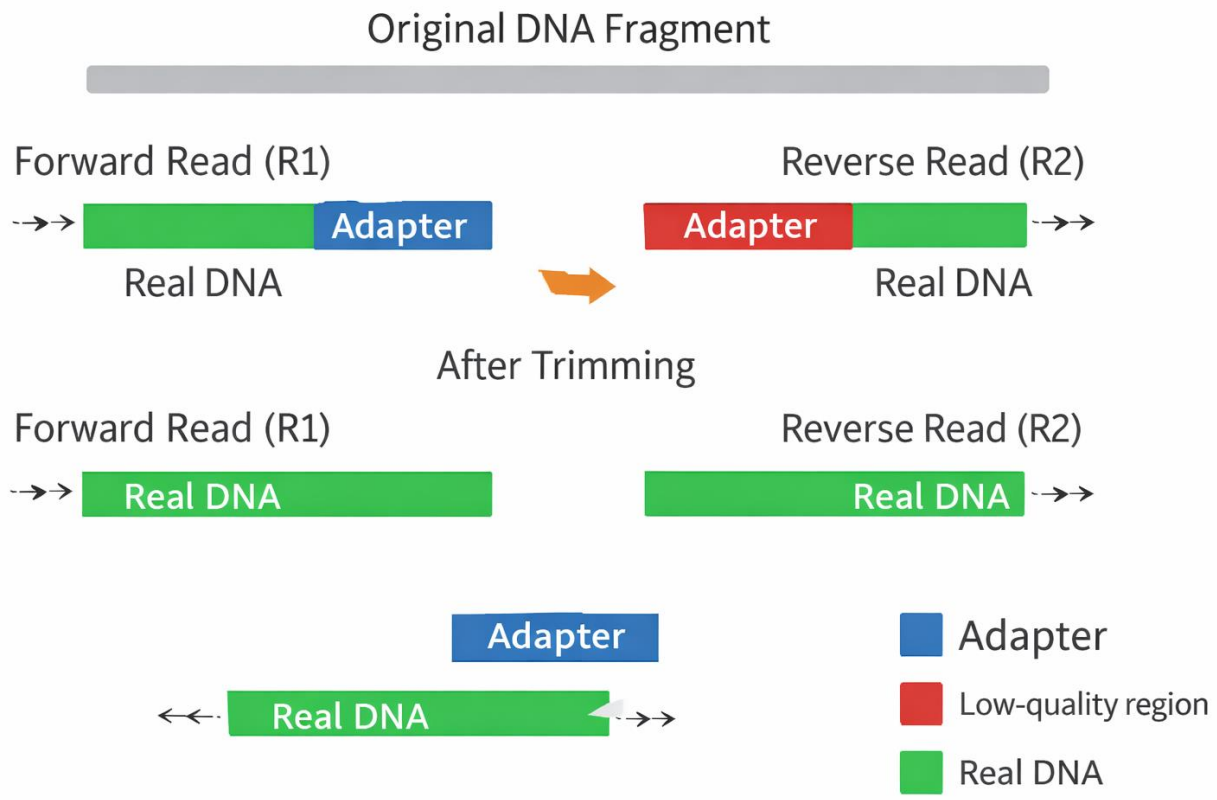
- Quality drops at 3' end
- Poor bases are cut off

3 Keep Paired Reads Together

- If both reads are good → stay paired
- If one becomes too short → may become unpaired

4 Remove Very Short Reads

- Reads shorter than MINLEN are discarded

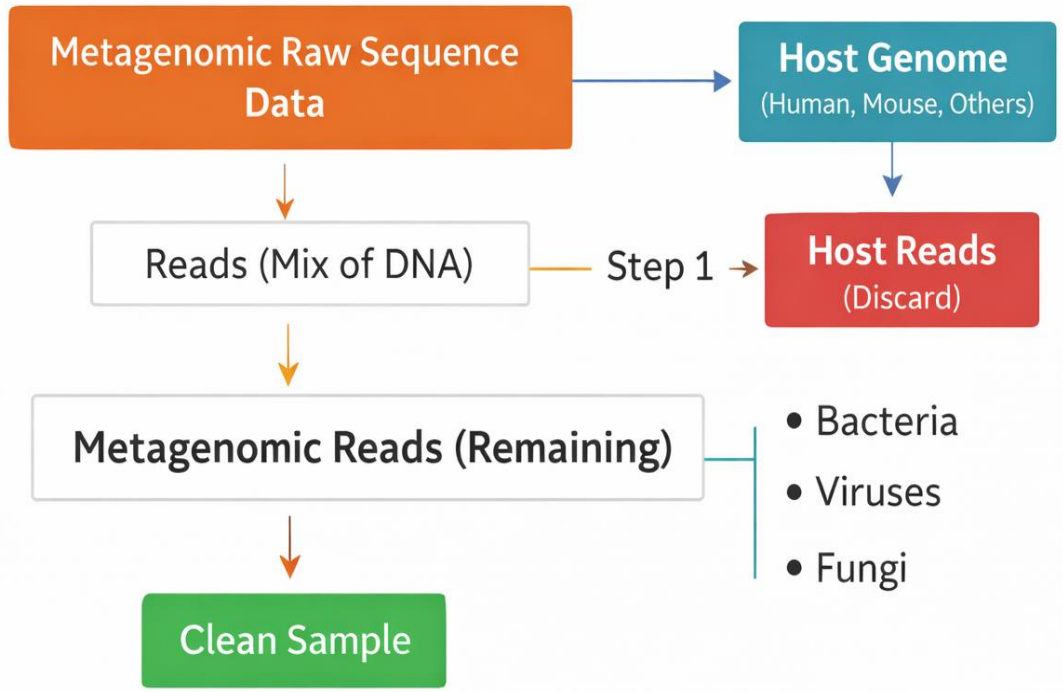


Host & Phix Removal

🧬 Metagenomes often contain:

- Human DNA
- Mouse DNA
- Plant DNA
- PhiX control

We remove host contamination before profiling.



🧬 **PhiX (PhiX174)** is a small viral genome (~5.3 kb) added by Illumina during sequencing as a:

- ✅ Control for run quality
- ✅ Calibration reference
- ✅ Base balance stabilizer (especially for low-diversity libraries)

It is **not part of your biological sample**.
So we must remove it before metagenomic analysis.

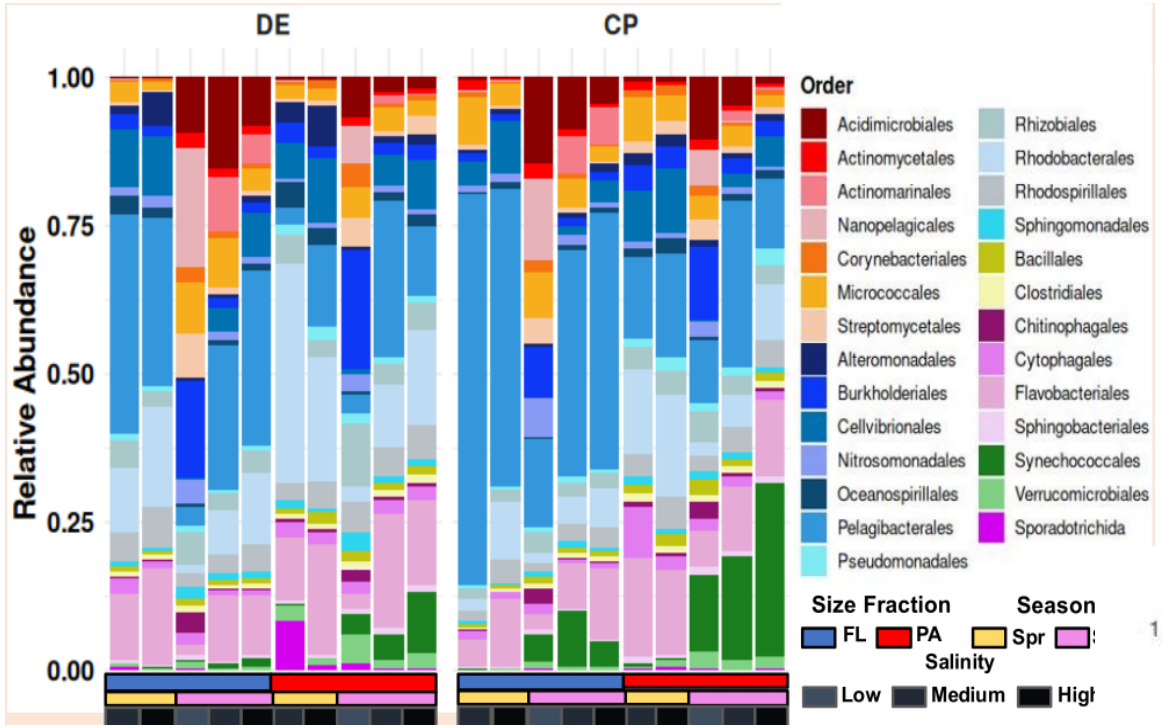
If PhiX is >5–10% of reads → ⚠ sequencing issue or low-diversity library

Now we will see how can we remove host/phix from our metagenomic data. Open *metagenome_practical.md* and follow along.

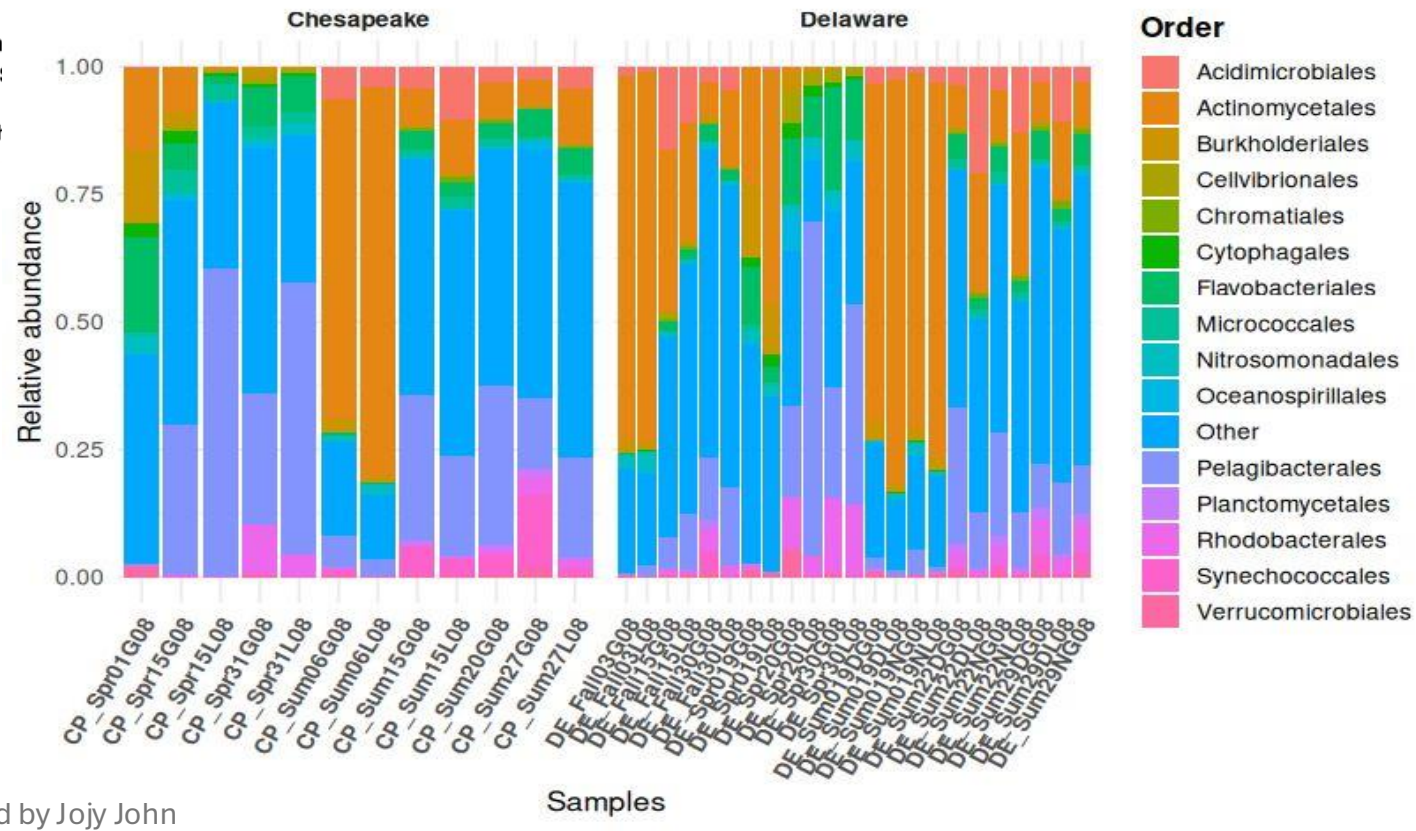
Read-Based Taxonomic Profiling

Now that we have clean reads, we can identify the organisms present in our samples. We'll use multiple tools to get a comprehensive view.

Tool	Database	Approach	Speed	Memory	Best For
Kaiju	NCBI nr (proteins)	Protein alignment	Fast	Medium	Comprehensive, novel organisms
Kraken2	Custom (DNA)	K-mer matching	Very Fast	High	Speed, abundance estimation
Bracken	Uses Kraken2	Statistical	Fast	Low	Abundance refinement
mOTUs	Marker genes	Gene alignment	Medium	Low	Species-level precision
MetaPhlAn	Marker genes	Marker alignment	Medium	Low	Species-level, publication-ready



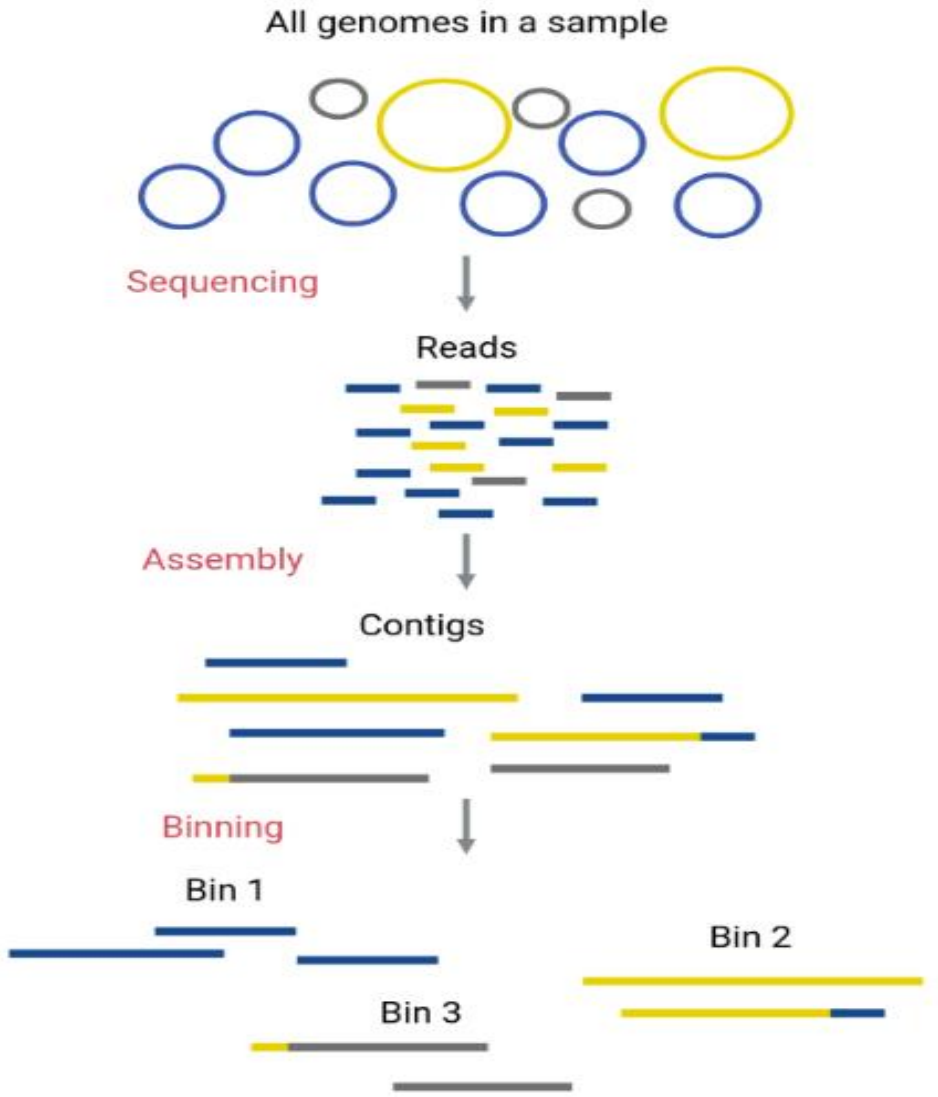
The exciting part—despite using fundamentally different approaches (protein-level classification with Kaiju vs marker-gene–based profiling with mOTUs), **both methods recovered highly similar dominant taxa,** especially when compared at the order level



Metagenome Assembly

Assembly is the process of reconstructing longer DNA sequences (contigs and scaffolds) from millions of short sequencing reads. Unlike single-genome assembly, **metagenomic assembly** is particularly challenging because:

- **Multiple genomes** present simultaneously
- **Varying abundance** (high to extremely low)
- **Strain variation** within species
- **Repetitive sequences** shared across organisms
- **Uneven sequencing coverage**



Why Assembly Matters

Assembly is crucial for:

- 📊 Recovering **complete or near-complete genomes** (MAGs)
- 🧬 Identifying **novel genes** and biosynthetic gene clusters
- 🔬 Enabling **functional annotation** beyond marker genes
- 🌐 Understanding **genomic context**
- 📈 Improving **resolution** of community structure

The De Bruijn Graph Approach

Modern metagenomic assemblers use **de Bruijn graphs** to construct contigs:

1. **K-mer decomposition**: Reads are broken into overlapping k-mers
2. **Graph construction**: K-mers become nodes, overlaps become edges
3. **Path finding**: Contigs are paths through the graph
4. **Graph simplification**: Remove bubbles, tips, and low-coverage paths

Key Concepts

Contig: A contiguous assembled sequence with no gaps
Scaffold: Multiple contigs connected by gap information (paired-end/mate-pair reads)
N50: The contig length at which 50% of the assembly is in contigs of this size or larger (higher is better)
Coverage: Average number of reads covering each base position
k-mer size: Length of subsequences used in assembly (critical parameter!)

Feature	metaSPAdes	MEGAHIT	Flye (Long-read)
Algorithm	Multi-k-mer SPAdes	Succinct de Bruijn	Repeat graph
Memory Usage	High (64-256 GB)	Low (16-64 GB)	Medium (32-128 GB)
Speed	Slow (hours-days)	Fast (30min-hours)	Medium (hours)
Contig Quality	Excellent	Very Good	Excellent
N50	Highest	High	Very High
Best For	High-quality assemblies	Large datasets, limited RAM	Long reads (ONT/PacBio)
Max k-mer	127 (default)	255	N/A (overlap-based)
Error Correction	Built-in BayesHammer	Basic	Consensus-based
Read Type	Illumina	Illumina	ONT/PacBio
Hybrid Mode	Yes (trusted contigs)	No	Yes (with Illumina polishing)
Typical RAM	100-200 GB	30-60 GB	50-120 GB
Typical Time	12-48 hours	2-8 hours	4-16 hours

Let's do the assembly.
Open *metagenome_practical.md* and follow along.

Assembly Quality Assessment

MetaQUAST (Quality Assessment Tool for Metagenome Assemblies) provides comprehensive quality metrics.

Statistics without reference	final_assembly
# contigs	22 437
# contigs (>= 0 bp)	22 437
# contigs (>= 1000 bp)	22 437
# contigs (>= 5000 bp)	2208
# contigs (>= 10000 bp)	687
# contigs (>= 25000 bp)	133
# contigs (>= 50000 bp)	34
Largest contig	176 373
Total length	63 261 614
Total length (>= 0 bp)	63 261 614
Total length (>= 1000 bp)	63 261 614
Total length (>= 5000 bp)	24 261 742
Total length (>= 10000 bp)	13 962 367
Total length (>= 25000 bp)	6 023 897
Total length (>= 50000 bp)	2 722 573
N50	3531
N75	1799
L50	3984

Basic Statistics:

- **# contigs**: Total number of contigs
- **Largest contig**: Length of longest contig
- **Total length**: Sum of all contig lengths
- **N50**: Contig length at 50% of assembly
- **L50**: Number of contigs comprising 50% of assembly

What Makes a Good Assembly?

Metric	Poor	Acceptable	Good	Excellent
N50	<1 kb	1-5 kb	5-20 kb	>20 kb
# contigs (>1kb)	>100k	50-100k	10-50k	<10k
Largest contig	<10 kb	10-50 kb	50-200 kb	>200 kb
Total length	Variable	Variable	Variable	Variable
L50	>50k	20-50k	5-20k	<5k
Misassemblies*	>1000	100-1000	10-100	<10

Let's check the quality of our assembly.
Open *metagenome_practical.md* and follow along.

Genome Binning - Recovering Individual Genomes (MAGs)

Key Features Used for Binning

1. Sequence Composition (Tetranucleotide Frequency)

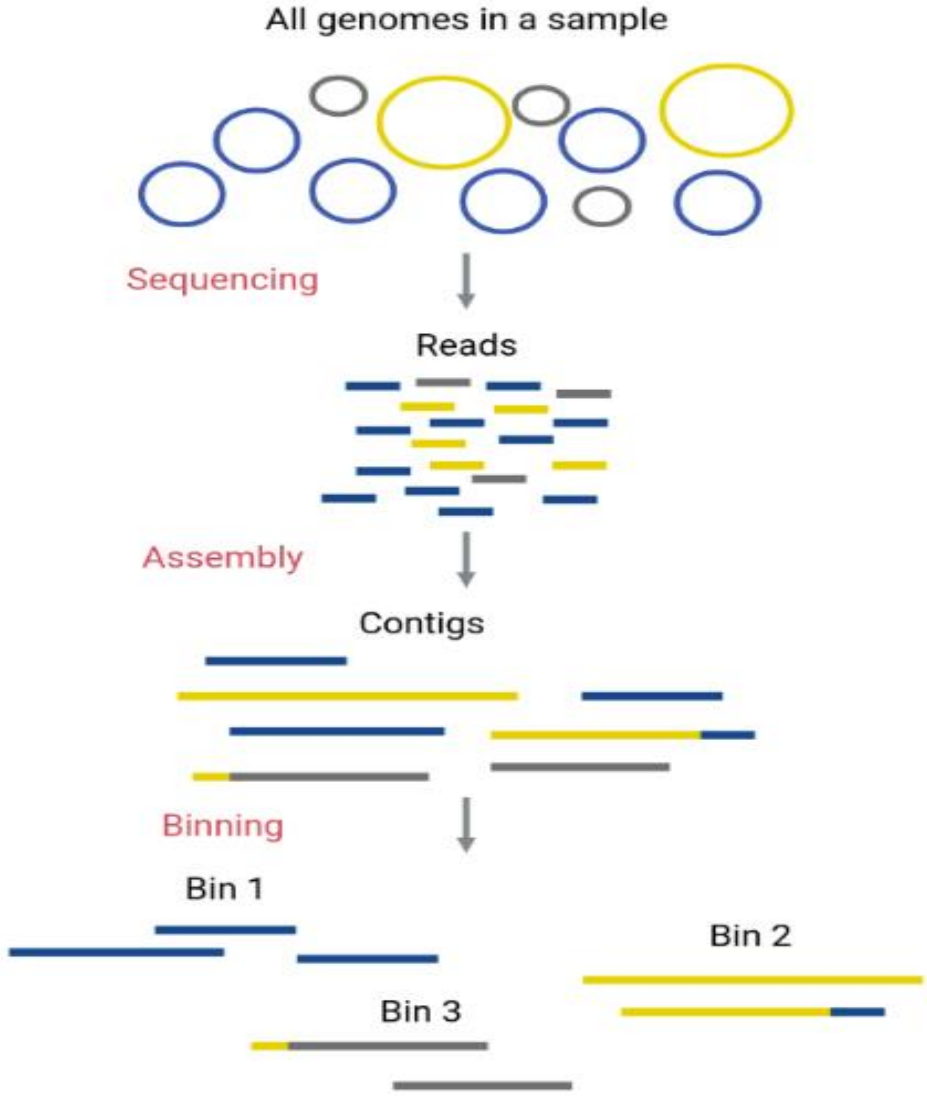
1. GC content patterns
2. Codon usage bias
3. k-mer frequencies
4. Different organisms have unique “genomic signatures”

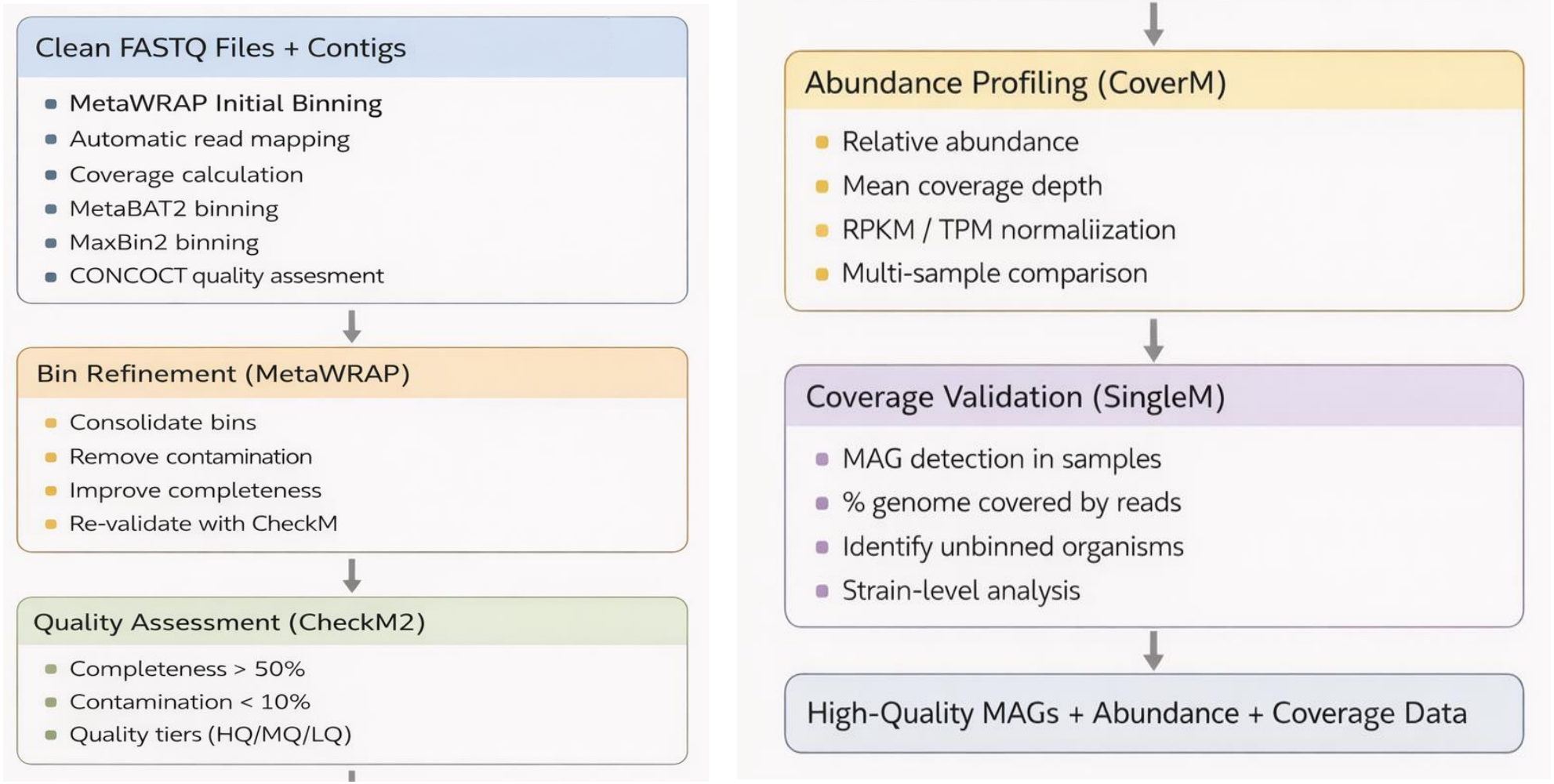
2. Coverage Patterns (Abundance)

1. Contigs from the same organism have similar coverage
2. Multi-sample binning: co-abundance across samples
3. More samples = better resolution

3. Taxonomic Markers

1. Single-copy marker genes
2. Phylogenetic placement
3. Reference database similarity





🤖**SemiBin2** is a modern, deep learning-based binner that often outperforms traditional methods.

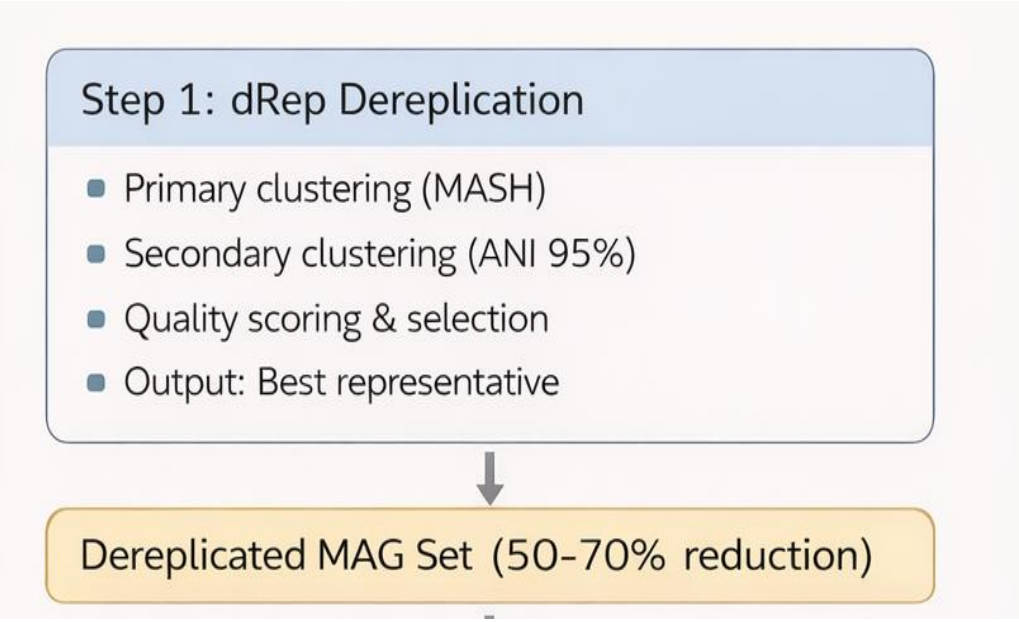
- ✅ **Deep learning-based** - Neural networks learn genomic patterns
- ✅ **Self-supervised** - No need for labeled training data
- ✅ **Multi-sample aware** - Excellent for time series/spatial data
- ✅ **Fast** - GPU acceleration available
- ✅ **High precision** - Lower contamination rates

Let's do it
Open *metagenome_practical.md* and follow along.

Genome Dereplication &Taxonomic Classification

Dereplication identifies and removes redundant genomes based on sequence similarity.

Why Dereplicate?
Problem: You recovered 200 MAGs across 10 samples, but:
•50 might be the same *E. coli* strain
•30 could be nearly identical *Bacteroides* species
•This redundancy inflates your dataset and wastes computational resources
Solution: Dereplication gives you:
- ✅ One representative genome per species
- ✅ Reduced dataset size (often 50-70% reduction)
- ✅ Faster downstream analysis
- ✅ Clearer biological interpretation



dRep uses a two-step clustering approach:

1. Primary clustering (MASH)

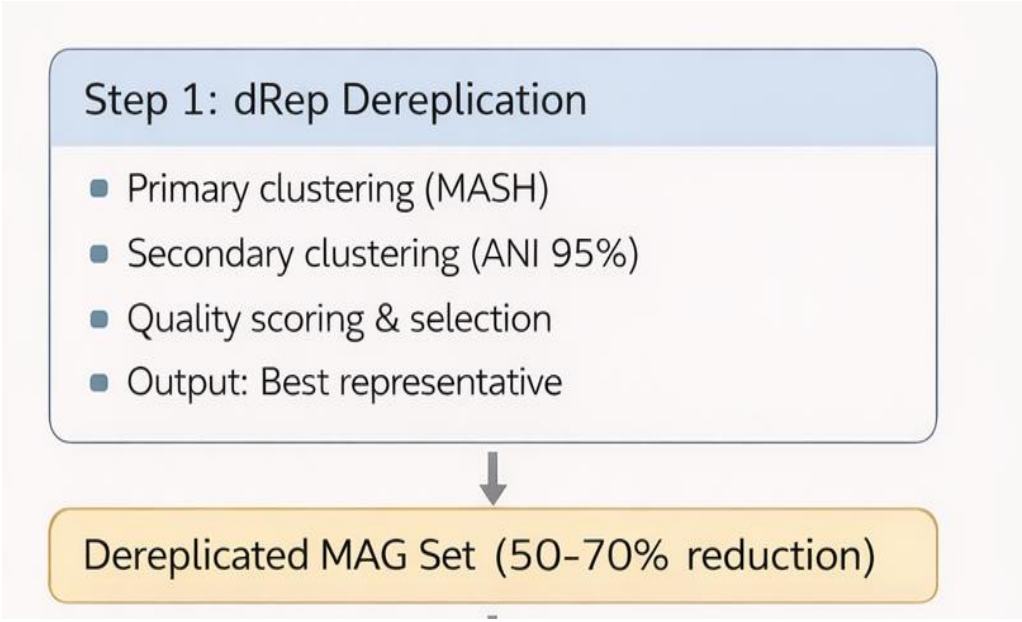
1. Fast, approximate sequence similarity
2. Groups genomes at ~90% ANI threshold
3. Creates initial clusters

2. Secondary clustering (ANI)

1. Precise average nucleotide identity (ANI)
2. Default: 95% ANI = same species
3. Adjustable for strain-level (99%) or genus-level (85%)

3. Scoring & Selection

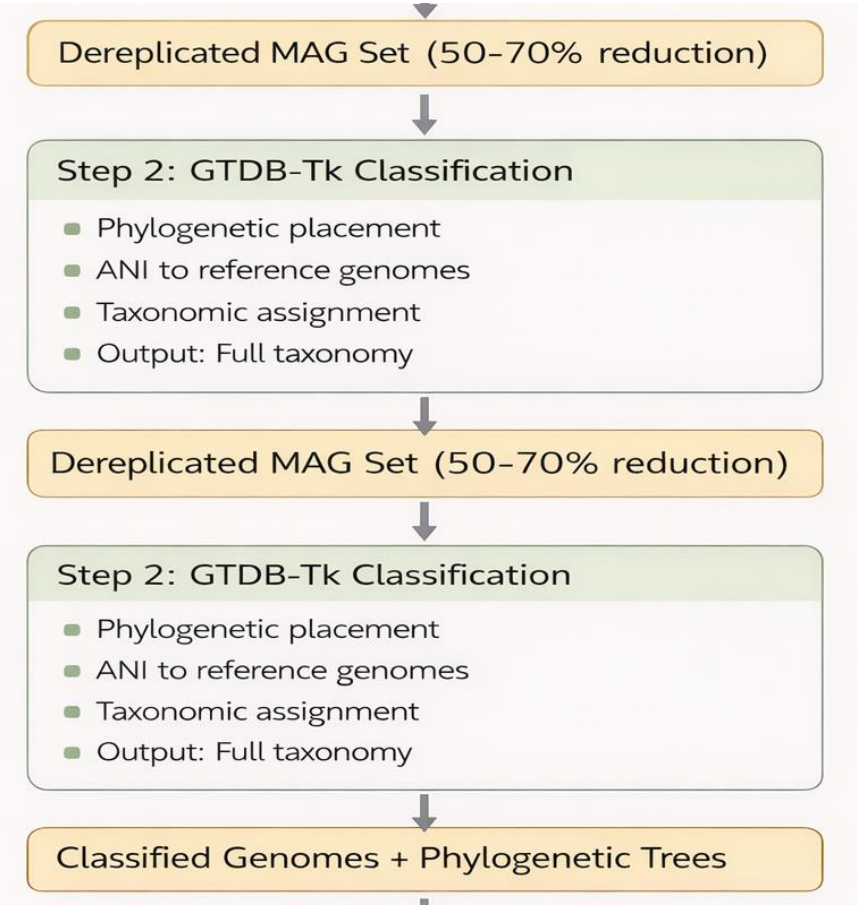
1. Ranks genomes by quality (completeness, contamination, N50)
2. Selects best representative per cluster
3. Outputs dereplicated genome set



Taxonomic Classification

GTDB (Genome Taxonomy Database) is the modern, standardized bacterial taxonomy system.

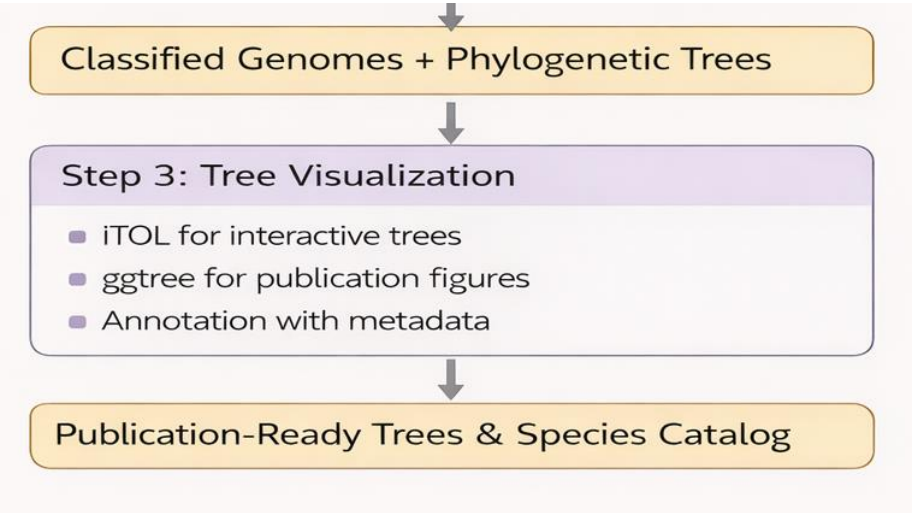
Feature	GTDB	NCBI Taxonomy
Standardization	Consistent, phylogeny-based	Inconsistent, literature-based
Genome count	402,709 genomes	Variable
Updates	Regular (every 6 months)	Irregular
ANI-based species	Yes (95% ANI)	No
Polyphyly	Resolved	Common issues

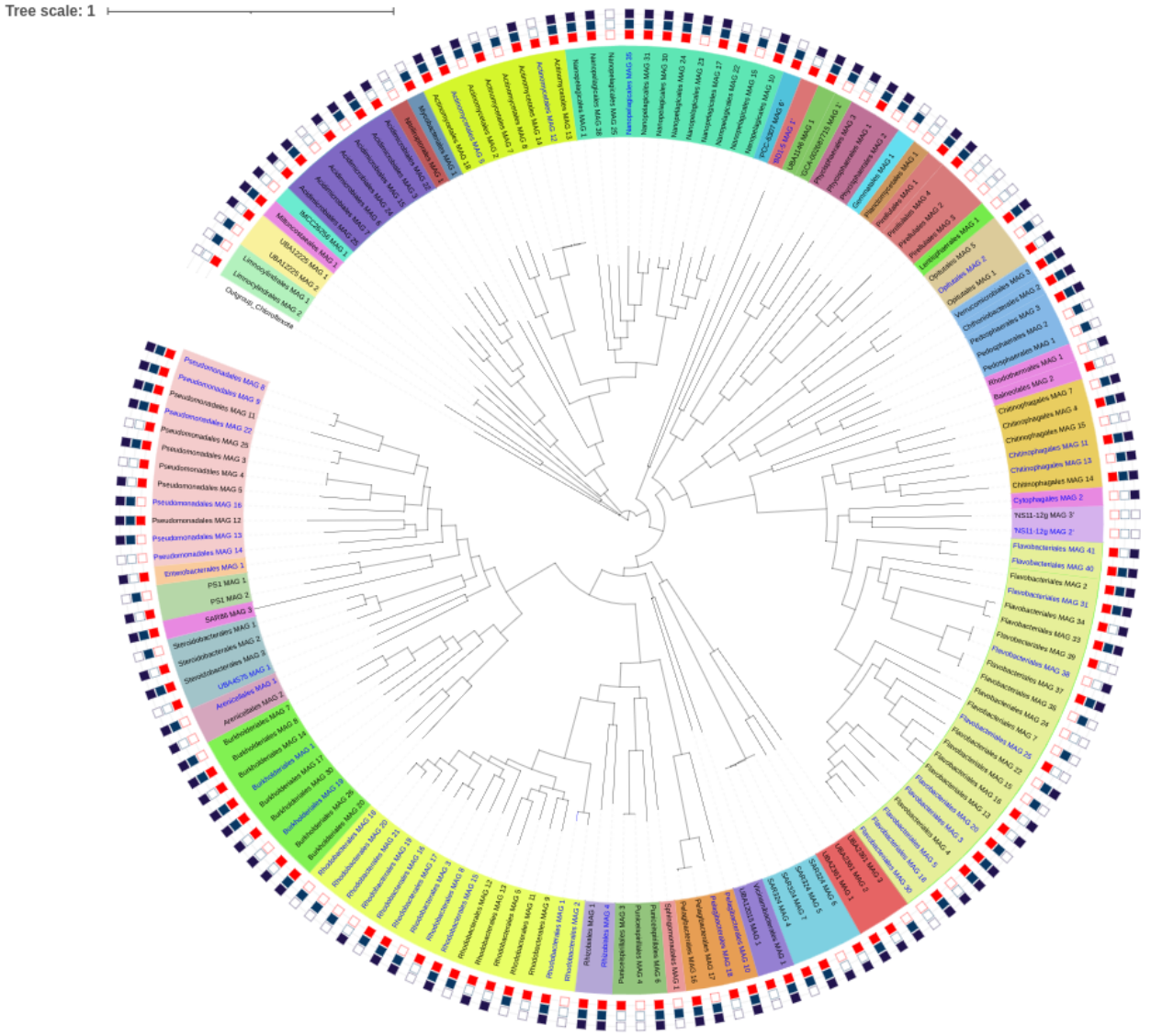


iTOL

iTOL is a web-based tool for beautiful, interactive phylogenetic trees.

Website: https://itol.embl.de/





prepared by Jojy John

Let's make a tree.

```
        }).done(function(response) {
            for (var i = 0; i < response.length; i++) {
                var layer = L.marker(
                    [response[i].latitude, response[i].longitude]
                    // ,{icon: myIcon}
                );
                layer.addTo(group);

                layer.bindPopup(
                    "<p>" + "Species: " + response[i].species + "<
                    "<p>" + "Description: " + response[i].descript
                    "<p>" + "Seen at: " + response[i].latitude + "
                    "<p>" + "On: " + response[i].sighted_at + "</p
                )

                $('select').change(function() {
                    species = this.value;
                });
        });
    }
$.ajax({
            url: queryURL,
            method: "GET"
        }).done(function(response) {
            for (var i = 0; i < response.length; i++) {
                var layer = L.marker(
                    [response[i].latitude, response[i].longitude]
                    // ,{icon: myIcon}
                );
                layer.addTo(group);
```

Annotation

Why Annotate?

- **Unannotated genome:** Just DNA sequences
- **Annotated genome:** Blueprint of an organism's capabilities

Answers you'll get:

- Can it fix nitrogen?
- Does it produce secondary metabolites?
- What carbon sources can it use?
- Does it have antibiotic resistance genes?
- Can it degrade pollutants?

The Annotation Hierarchy

```
Level 1: Gene Prediction
├─ Prodigal    (Fast, accurate)
└─ Prokka      (Gene prediction + basic annotation)

Level 2: Functional Annotation
├─ eggNOG-mapper  (Orthology-based)
└─ Prokka         (Swiss-Prot, Pfam)

Level 3: Metabolic Annotation
├─ DRAM          (Metabolic distillation)
└─ METABOLIC     (Comprehensive pathways)
```

Let's do the annotation.

```
        }).done(function(response) {
            for (var i = 0; i < response.length; i++) {
                var layer = L.marker(
                    [response[i].latitude, response[i].longitude]
                    // ,{icon: myIcon}
                );
                layer.addTo(group);

                layer.bindPopup(
                    "<p>" + "Species: " + response[i].species + "<
                    "<p>" + "Description: " + response[i].descript
                    "<p>" + "Seen at: " + response[i].latitude + "
                    "<p>" + "On: " + response[i].sighted_at + "</p
                )

                $('select').change(function() {
                    species = this.value;
                });
            });
    }
$.ajax({
            url: queryURL,
            method: "GET"
        }).done(function(response) {
            for (var i = 0; i < response.length; i++) {
                var layer = L.marker(
                    [response[i].latitude, response[i].longitude]
                    // ,{icon: myIcon}
                );
                layer.addTo(group);
```

Computational and Structural Biotechnology Journal 19 (2021) 6301–6314

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

A review of computational tools for generating metagenome-assembled genomes from metagenomic sequencing data

Chao Yang [a,1], Debajyoti Chowdhury [b,c,1], Zhenmiao Zhang [a], William K. Cheung [a], Aiping Lu [b,c], Zhaoxiang Bian [d,e], Lu Zhang [a,b,*]

[a] Department of Computer Science, Hong Kong Baptist University, Hong Kong Special Administrative Region
[b] Computational Medicine Lab, Hong Kong Baptist University, Hong Kong Special Administrative Region
[c] Institute of Integrated Bioinformedicine and Translational Sciences, School of Chinese Medicine, Hong Kong Baptist University, Hong Kong Special Administrative Region
[d] Institute of Brain and Gut Research, School of Chinese Medicine, Hong Kong Baptist University, Hong Kong Special Administrative Region
[e] Chinese Medicine Clinical Study Center, School of Chinese Medicine, Hong Kong Baptist University, Hong Kong Special Administrative Region

ARTICLE INFO

Article history:
Received 30 August 2021
Received in revised form 17 November 2021
Accepted 17 November 2021
Available online 23 November 2021

Keywords:
Metagenomic sequencing
Metagenome-assembled genomes
Genome assembly
Metagenome binning
Gene prediction
Gene functional annotation
Taxonomic classification
Microbial abundance profiling

ABSTRACT

Metagenomic sequencing provides a culture-independent avenue to investigate the complex microbial communities by constructing metagenome-assembled genomes (MAGs). A MAG represents a microbial genome by a group of sequences from genome assembly with similar characteristics. It enables us to identify novel species and understand their potential functions in a dynamic ecosystem. Many computational tools have been developed to construct and annotate MAGs from metagenomic sequencing, however, there is a prominent gap to comprehensively introduce their background and practical performance. In this paper, we have thoroughly investigated the computational tools designed for both upstream and downstream analyses, including metagenome assembly, metagenome binning, gene prediction, functional annotation, taxonomic classification, and profiling. We have categorized the commonly used tools into unique groups based on their functional background and introduced the underlying core algorithms and associated information to demonstrate a comparative outlook. Furthermore, we have emphasized the computational requisition and offered guidance to the users to select the most efficient tools. Finally, we have indicated current limitations, potential solutions, and future perspectives for further improving the tools of MAG construction and annotation. We believe that our work provides a consolidated resource for the current stage of MAG studies and shed light on the future development of more effective MAG analysis tools on metagenomic sequencing.

OXFORD

Briefings in Bioinformatics, 20(4), 2019, 1125–1139

doi: 10.1093/bib/bbx120
Advance Access Publication Date: 23 September 2017
Paper

A review of methods and databases for metagenomic classification and assembly

Florian P. Breitwieser, Jennifer Lu and Steven L. Salzberg

Corresponding author: Steven L. Salzberg, Center for Computational Biology, Johns Hopkins University, 1900 E. Monument St., Baltimore, MD, 21205, USA. E-mail: salzberg@jhu.edu

Abstract

Microbiome research has grown rapidly over the past decade, with a proliferation of new methods that seek to make sense of large, complex data sets. Here, we survey two of the primary types of methods for analyzing microbiome data: read classification and metagenomic assembly, and we review some of the challenges facing these methods. All of the methods rely on public genome databases, and we also discuss the content of these databases and how their quality has a direct impact on our ability to interpret a microbiome sample.

Key words: microbiome; microbial genomics; next-generation sequencing; bacteria; databases

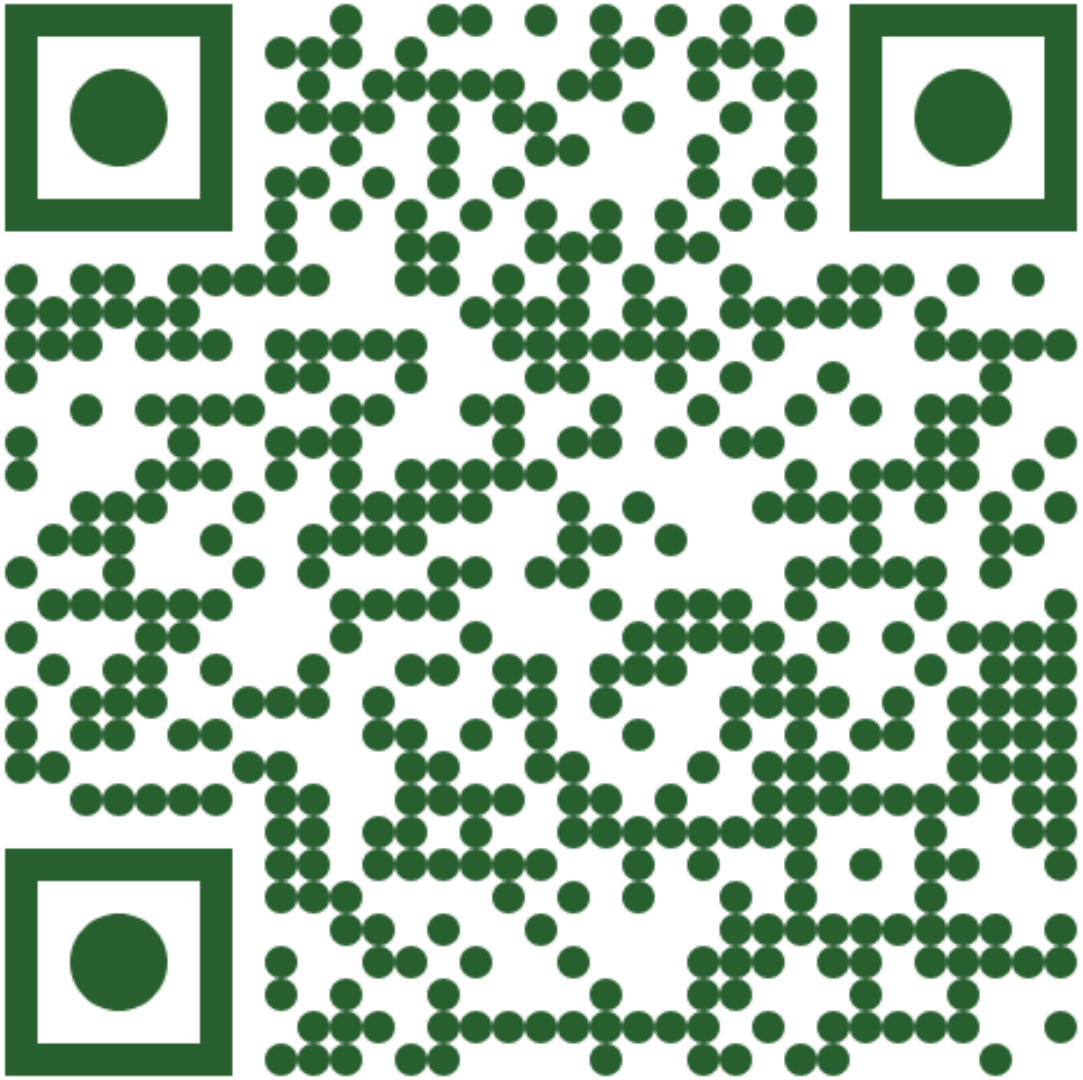
Introduction

Microbiome research has been expanding rapidly as a consequence of dramatic improvements in the efficiency of genome sequencing. As the variety and complexity of experiments has grown, so have the methods and databases used to analyze these experiments. Ever-larger data sets present increasing challenges for computational methods, which must minimize processing and memory requirements to provide fast turnaround and to avoid overwhelming the computational resources available to most research laboratories. The rapid increase in the number and variety of genomes also present many challenges, rising in part from the effort required to fit traditional taxonomic naming schemes onto a microbial world that we now know is vastly richer and more complex than scientists realized when

This review discusses the computational challenges of analyzing metagenomics data, focusing on methods but also including a discussion of microbial taxonomy and genome resources, which are rarely discussed in benchmark studies and tool reviews despite their critical importance. We begin with a review of terminology and a comparison of marker gene sequencing, shotgun metagenome sequencing and meta-transcriptome sequencing, all of which are sometimes included in the term metagenomics.

Metataxonomics, metagenomics, metatranscriptomics

The most widely used sequencing-based approaches for microbiome research are metataxonomics and metagenomics



Advanced level versions for each step available here, please visit