AMRITA VISHWA VIDYAPEETHAM

## 22BIO201 Intelligence of Biological Systems - 1

# **Meta Genomic** Data Analytics

**Team 7 :**
**G Prajwal Priyadarshan - 214**
**Kabilan K - 224**
**Kishore B - 227**
**Rahul L S - 248**

# Introduction

## Meta Genomics

Metagenomics is the study of all the genetic material (DNA) collected directly from environmental samples — like soil, water, or the human body — to identify the different microbes present and understand what they can do, without the need to grow them in a lab.

# Meta Genomic Data Analysis

->To analyze publicly available metagenomic datasets to identify the types of microbes present in a sample and understand their functional roles, using computational tools and AI techniques.

->Use publicly available metagenomic data (no lab work needed).
Cleans and processes the DNA data using bioinformatics tools.

->Find out:

Which microbes are present (species identification).
What they can do (functional roles).

->Can be applied in healthcare, environment, and agriculture.

# **Problem** Statement

Metagenomic datasets contain vast amounts of raw DNA sequences from mixed microbial communities, but this data is often unstructured, noisy, and difficult to interpret.

There is a need for efficient computational methods to clean, classify, and analyze these datasets to identify the diversity of microbes present and understand their functional roles.

Without proper data analytics, valuable biological insights remain hidden, limiting applications in healthcare, environmental monitoring, and agriculture.

# Objective

- **Processing raw metagenomic sequencing reads** obtained from public repositories (e.g., Tara Oceans, MG-RAST, NCBI SRA).
- **Performing taxonomic profiling** to classify reads into taxonomic groups (domain, phylum, class, order, family, genus, species).
- **Comparing performance** of multiple classification tools in terms of accuracy, computational efficiency, and resolution.
- **Deriving biological insights** about microbial diversity, relative abundance, and possible ecological roles in the sampled environments.

# Base Paper

The paper is a tutorial/review explaining how metagenomic analysis works, especially for crop soils (rhizosphere of rice, wheat, legumes, chickpea, sorghum).

## What Is KMAP & Paper's Role?

KMAP is a web-based, open-access software platform for exploring, analyzing, and comparing massive metagenomic datasets—especially shotgun sequencing datasets.

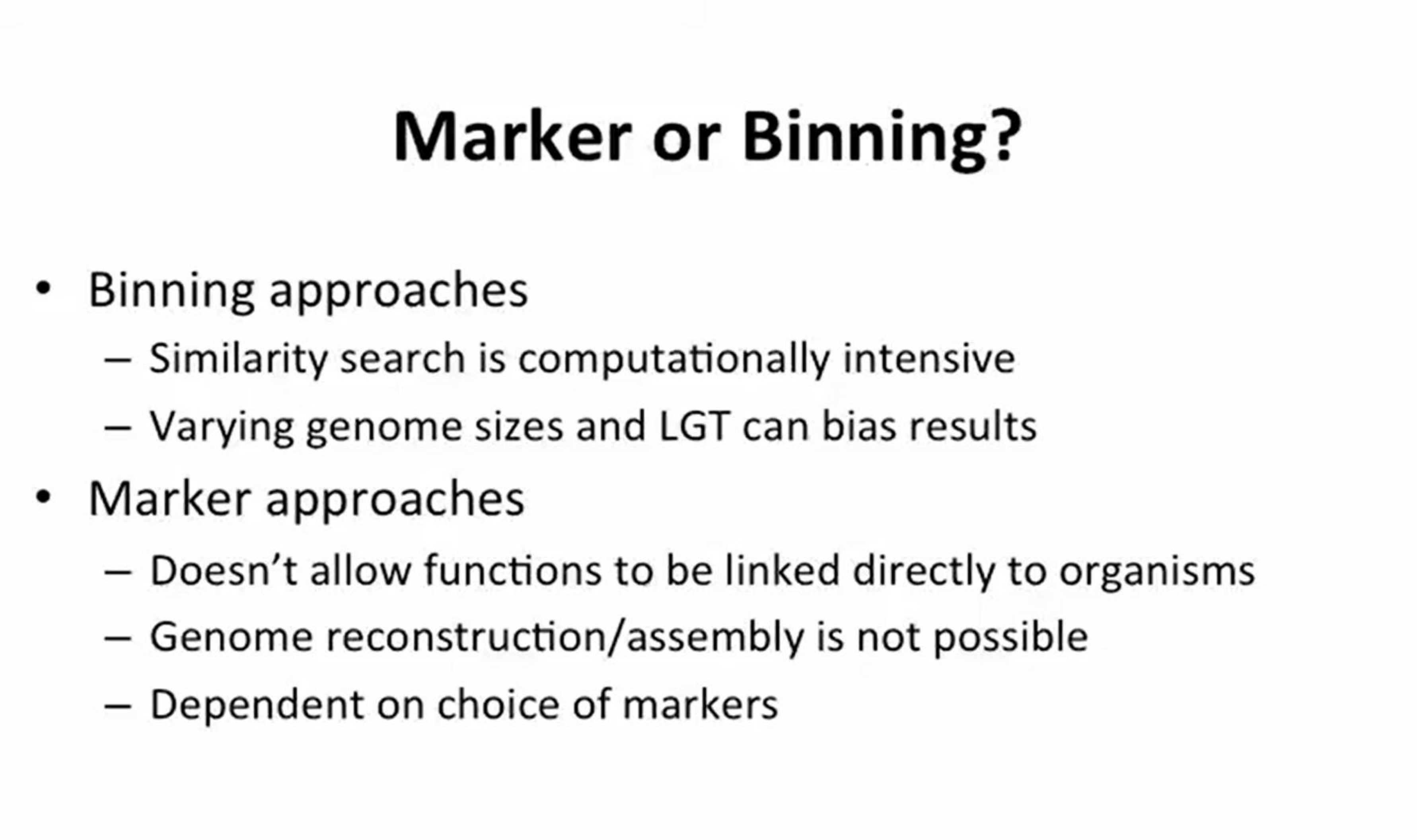
This paper's main goal: to offer researchers (even with limited compute resources) an easy way to access millions of publicly available metagenomes and their annotated genes—so you don't always need to generate raw data.

# Methodology

Metagenomics is the study of genetic material recovered directly from environmental samples without the need to culture individual microorganisms. It allows researchers to analyze the collective genomes of entire microbial communities, providing insights into:

- Community composition (Who are the microbes?)
- Functional potential (What can they do?)
- Ecological interactions (How do they interact?)
- Metabolic pathways and biochemical processes

The term was coined by Jo Handelsman in 1998, who defined it as "the cloning and functional analysis of collective genomes of soil microflora".



# Marker or Binning?

- Binning approaches
  - Similarity search is computationally intensive
  - Varying genome sizes and LGT can bias results
- Marker approaches
  - Doesn’t allow functions to be linked directly to organisms
  - Genome reconstruction/assembly is not possible
  - Dependent on choice of markers

# Methodology

MEGAN

MG-RAST

Kraken.

# Methodology

- Sample Collection and Processing
- DNA Extraction and Isolation
- Library Preparation and Sequencing
- Data Analysis Pipeline

AI & ML Applications :

- Taxonomic Classification
- Binning and Assembly
- Functional Annotation
- Novel Gene Discovery

# Expected outcome

- **A complete taxonomic profile** of microbial communities present in the analyzed metagenomic dataset, showing relative abundance from **domain to species level**.
- **Identification of dominant and rare taxa**, helping to understand community structure and ecological relationships in the sampled environment.
- **Performance comparison report** between classification tools (e.g., Kraken2, MetaPhlAn3, Kaiju) highlighting differences in speed, memory usage, and taxonomic resolution.
- **Visualization outputs** such as Krona charts, stacked bar plots, and heatmaps to intuitively represent microbial diversity.
- **Biological insights** into potential functional roles of identified organisms and their relevance to the environment studied (e.g., ocean microbiome patterns, pathogen detection, biogeochemical cycles).

# Conclusion

- Enables identification of microbial diversity and functional roles directly from environmental samples.
- Eliminates the need for culturing microbes in a lab.
- Uses computational and AI-driven bioinformatics tools for efficient analysis.
- Cleans, classifies, and interprets large volumes of raw sequencing data.
- Unlocks valuable insights for healthcare, environmental monitoring, and agriculture.
- Accelerates research and supports data-driven decision-making.

# Thank You !