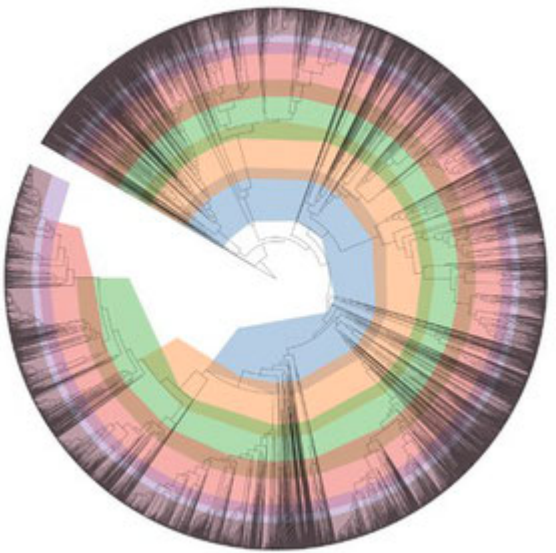
ENVIRONMENTAL METAGENOMICS

Physalia course, online, 11-15 November 2024

Introduction to metagenomics

Nikolay Oskolkov, Lund University, NBIS SciLifeLab
Luis Pedro Coelho, Queensland University of Technology



Physalia Courses

NB: original course material courtesy:

Dr. Antti Karkman, University of Helsinki

Dr. Igor Pessi, Finnish Environment Institute (SYKE)

What is a metagenome?

Marchesi and Ravel *Microbiome* (2015) 3:31
DOI 10.1186/s40168-015-0094-5

Microbiome

EDITORIAL Open Access

The vocabulary of microbiome research: a proposal

Julian R. Marchesi[1,2] and Jacques Ravel[3,4*]

Physalia Courses

A **metagenome** is a collection of genomes or genes from the members of a microbiota. A **microbiota** is an assemblage of microorganisms present in a defined environment. A **microbiome** refers to an entire habitat, including the microorganisms, their genomes, and the surrounding environmental conditions.

What is metagenomics?

"This collection is obtained through shotgun sequencing of DNA extracted from a sample (**metagenomics**) followed by mapping to a reference database or assembly, followed by annotation."

Marchesi & Ravel 2015, "The vocabulary of microbiome research"

JOURNAL OF BACTERIOLOGY, Feb. 1996, p. 591–599
0021-9193/96/$04.00+0

Vol. 178, No. 3

Characterization of Uncultivated Prokaryotes: Isolation and Analysis of a 40-Kilobase-Pair Genome Fragment from a Planktonic Marine Archaeon

JEFFEREY L. STEIN,[1]* TERENCE L. MARSH,[2] KE YING WU,[3] HIROAKI SHIZUYA,[4] AND EDWARD F. DELONG[3]*

Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products

Jo Handelsman[1], Michelle R Rondon[1], Sean F Brady[2], Jon Clardy[2] and Robert M Goodman[1]

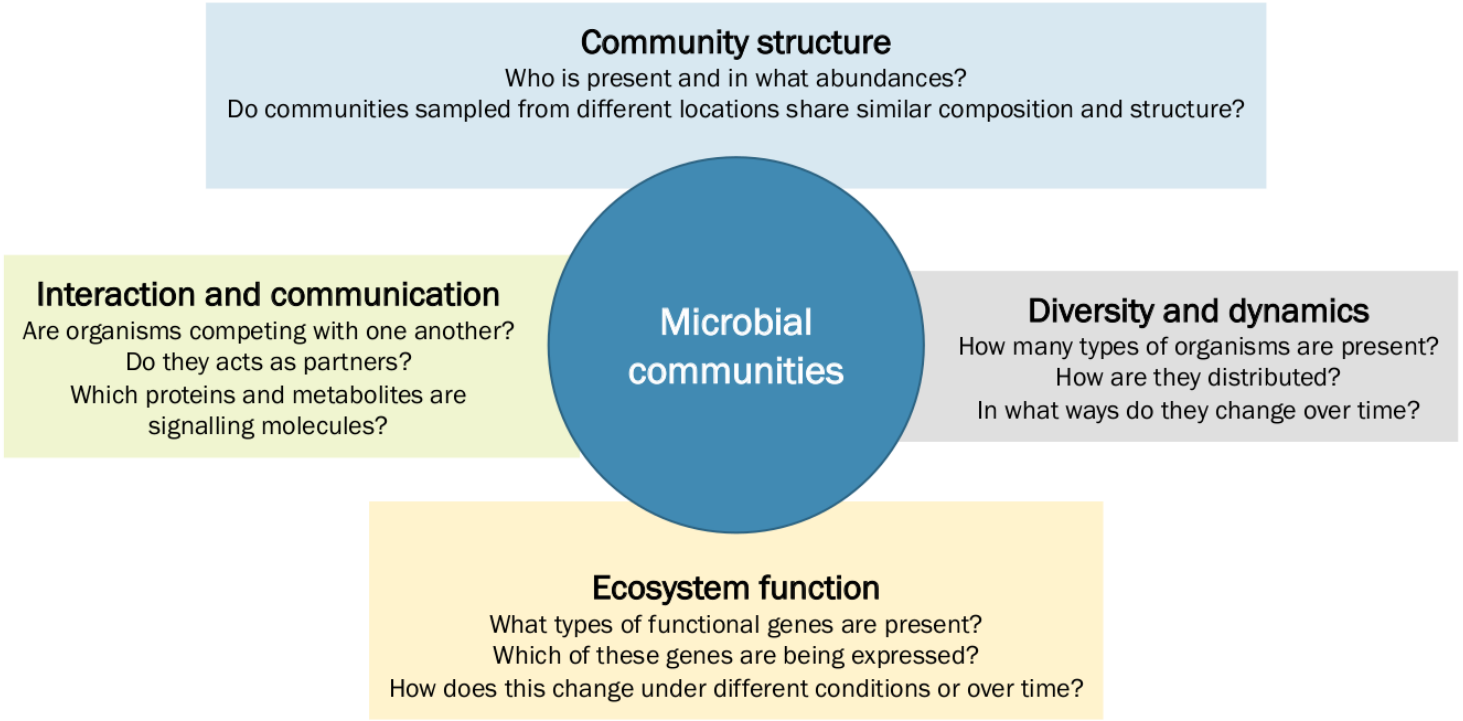
meta | genome

"beyond the genome"

"[cloning of environmental DNA into *E. coli* for phenotype screening] has been made possible by advances in molecular biology and Eukaryotic genomics, which have laid the groundwork for cloning and functional analysis of the collective genomes of soil microflora, which we term the metagenome of the soil"

Physalia
Courses

Metagenomics is the ultimate way to study microbial communities



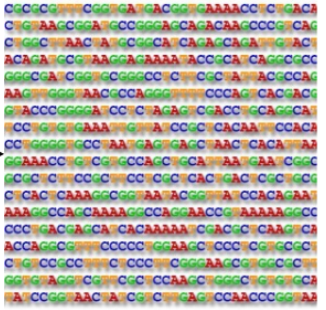
Physalia
Courses

What is metagenomics?

Environmental sample

DNA extraction

Sequencing



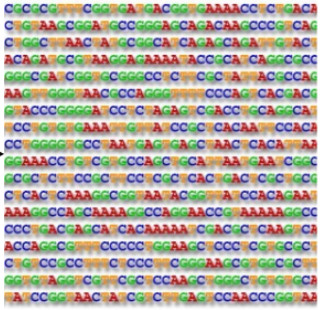
Physalia Courses

What is NOT metagenomics?

Environmental sample

DNA extraction

Sequencing

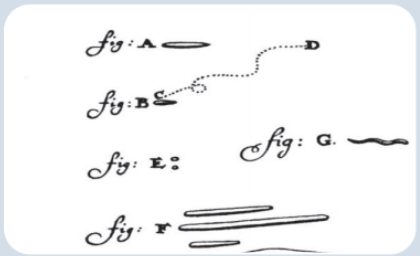


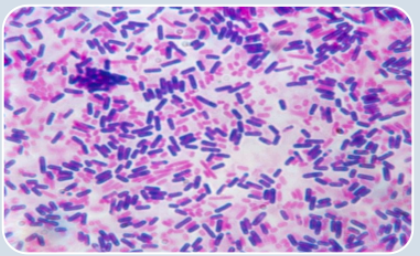
Sequencing

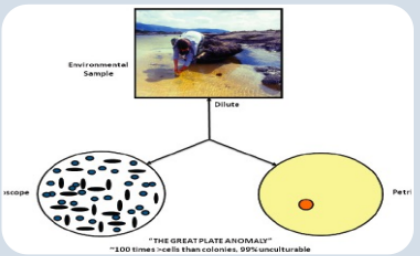
PCR
(e.g. 16S rRNA gene)

Physalia
Courses

Microbiology and technology go hand in hand







1670's

First observation of microbes under the microscope

1880's

Development of the Gram staining method

First isolation of a bacterium in solid media

1980's

The Great Plate Count Anomaly

First culture-independent studies

2000's

Advent of high-throughput sequencing

Physalia
Courses

Metagenomics *vs.* sequencing technologies

www.pacb.com/blog/the-evolution-of-dna-sequencing-tools/

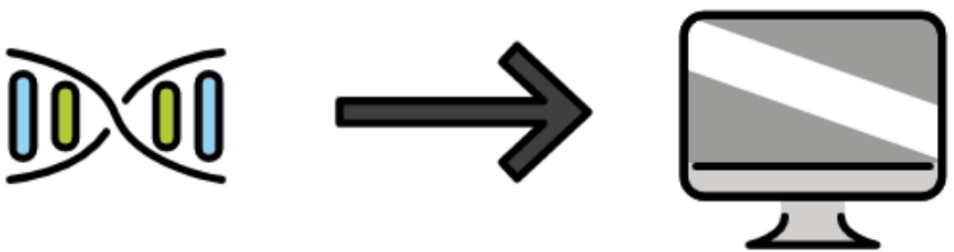
Physalia
Courses

What is Sequencing?

Converting the chemical nucleotides of a DNA molecule

to

ACTG on your computer screen

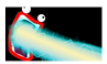


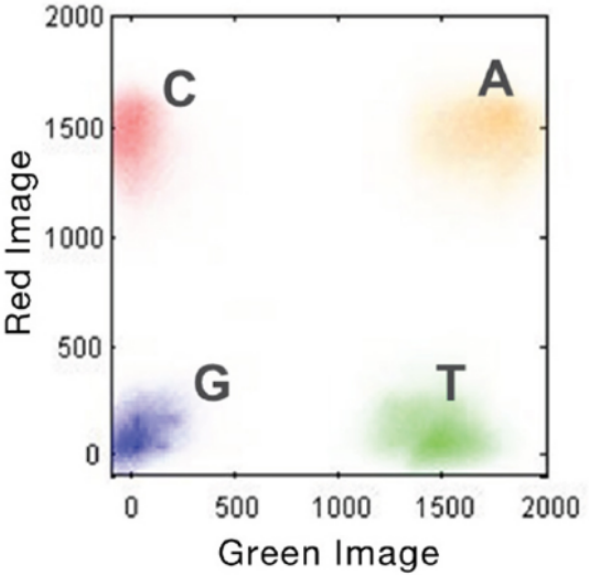
Physalia
Courses

How does it work?

Replicate a strand, but add complementary **fluorophore-modified nucleotide**, one colour per base

T C A G

Fire mah lazer  , and record the colour!
Rinse and repeat!



Physalia
Courses

FASTQ File

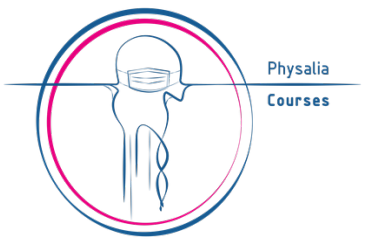
Example (files can be gigabytes in size!)

```
@K00233:37:HGHLYBBXX:3:1101:2646:1121 1:N:0:NACGCATC+NGCTGGTG
NCGCATGAGCCGCCTGTATCAGGCGCTGATCGGGCCGGGCATTGCAGTTGGGATAGATCGGGGGAGCACACGTCTG
+
#A7F<<GG<JFJFJJJJJJFFJJJJJJJAFFJFJJJJJJJFJAFFFJAJFJJ<FJJJJJFFF<FFA--FFFJJJJJ
@K00233:37:HGHLYBBXX:3:1101:4655:1121 1:N:0:NACGCATC+NGCTGGTG
NATGCATGACAGGAGGTGAGGGCATTTTCCAGATTTTCAGGCTGCGACCTTGAGCATCTTTCGCCGCTTCCAGCAC
+
#GG-<FFFF7JFF7JJJJJFJJ<JJJJJA7FJJJJJJJFF<JFF<J7-<FJJJJFJFFJJJGGGGFFJJ--AJAJJ
```

```
@ <read id, e.g. machine ID, location on flowcell> <extra metadata>
  <DNA sequence; Note: N = base couldn't be called!>
+ <a separator>
  <base quality scores for each nucleotide in sequence>
```

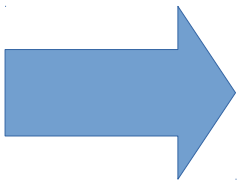
Quality score:

```
!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJ
0.2.....................26...31........41
```



Sequenced genome is a collection of bits of puzzle to be organized

From single genome to metagenome

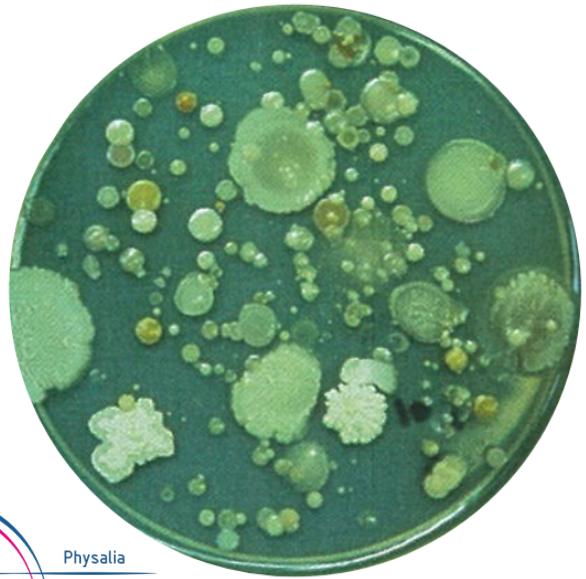


Who's there?

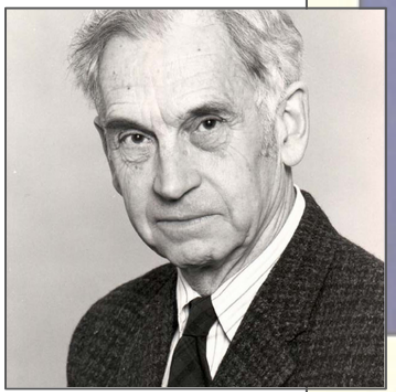
1970

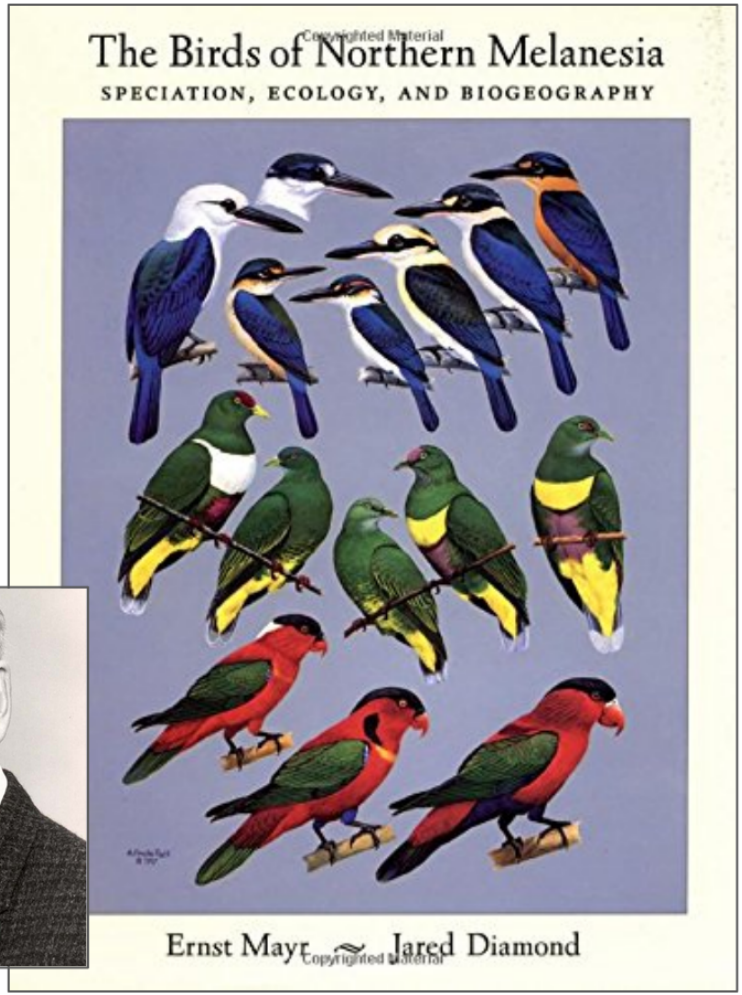
At a most basic level, the first question we usually ask in metagenomics is "**Who's there?**"

What is a microbial species?



Ernst Mayr
Biological Species Concept, 1942





Physalia Courses

Who's there?

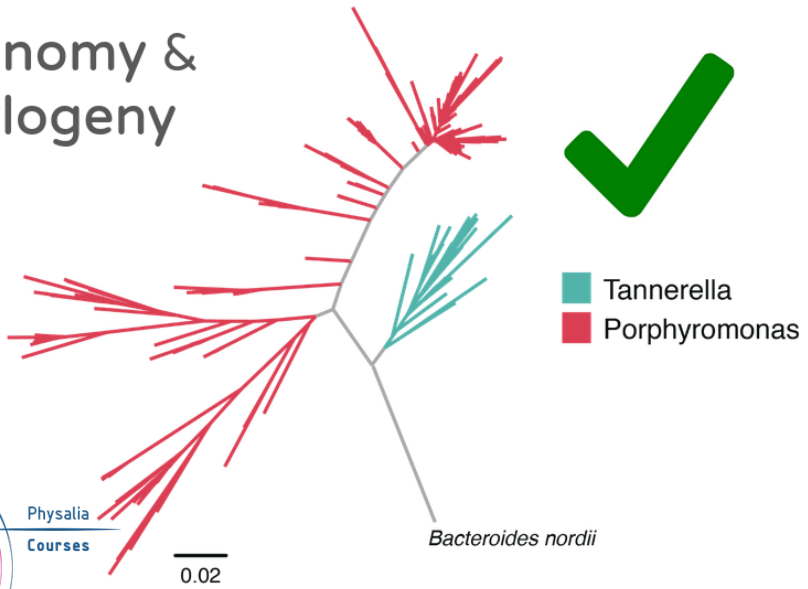
Taxonomy: classification or categorization of organisms into groups (taxa)

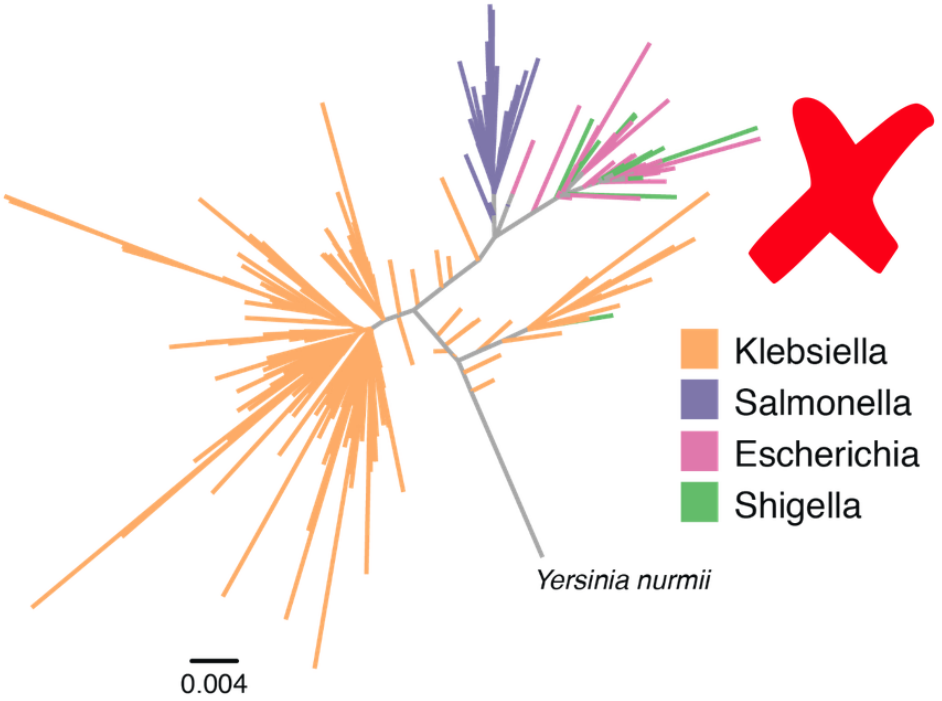
Phylogeny: evolutionary history of a set of taxa

At a most basic level, the first question we usually ask in metagenomics is "**Who's there?**"

What is a microbial species?

Taxonomy & Phylogeny





Physalia Courses

Typical analysis methods used in metagenomics

1) Alignment:

Bow Tie

BWA

stands for

Burrows Wheeler Aligner

Abbreviations.com

H O P S

2) Classification:

Centrifuge

MetaPhlan

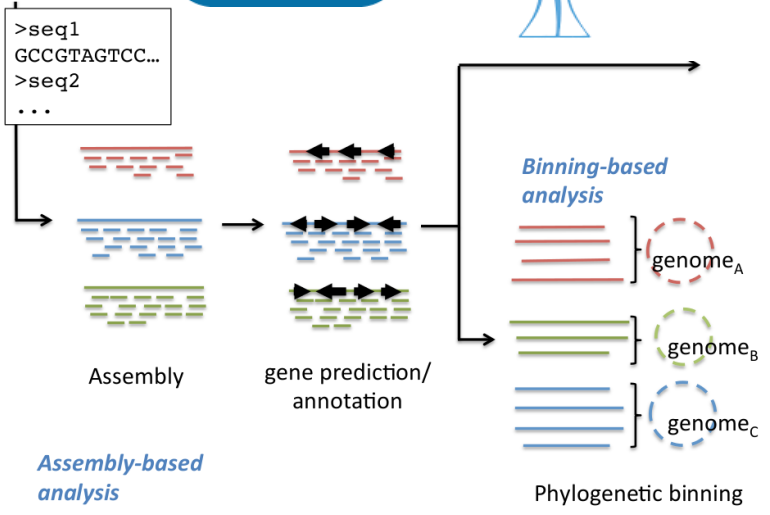
Clark

Reference based:
assume similarity to reference

3) De-novo assembly:

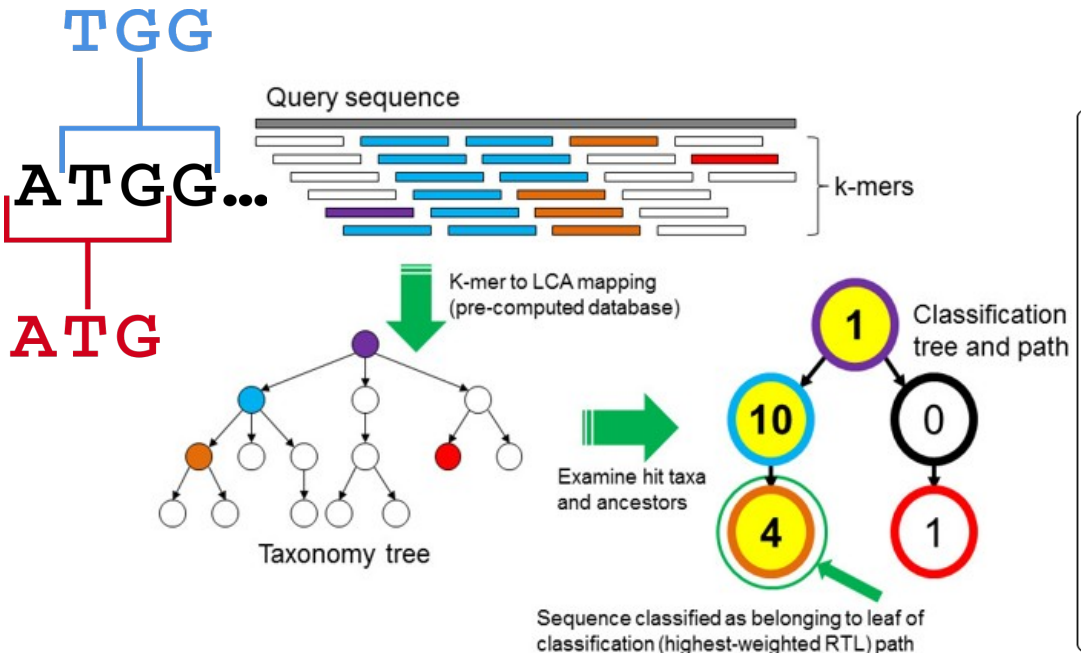
MEGA HIT

meta



Reference free:
unbiased but challenging

K-mer based taxonomic profiling: Kraken family of tools



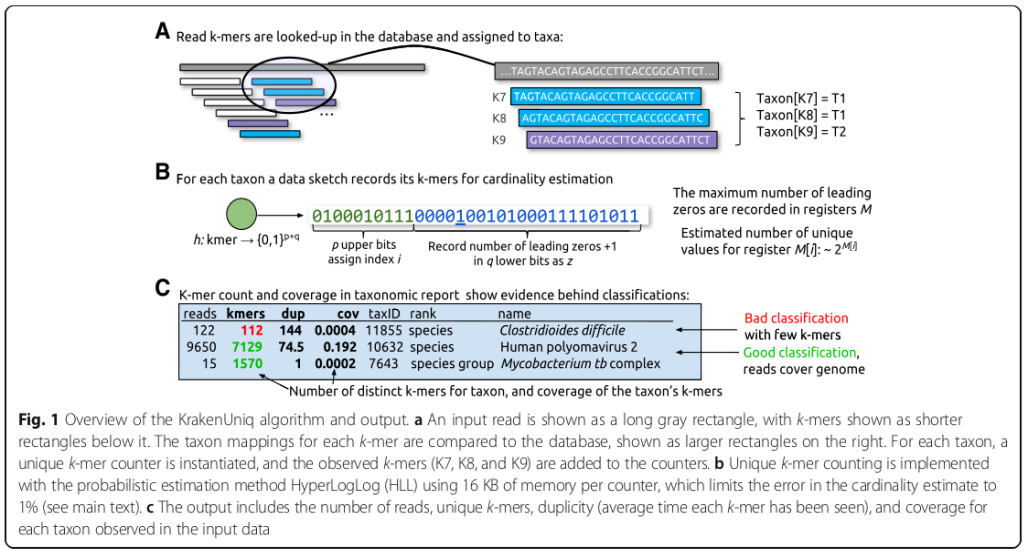


Fig. 1 Overview of the KrakenUniq algorithm and output. **a** An input read is shown as a long gray rectangle, with *k*-mers shown as shorter rectangles below it. The taxon mappings for each *k*-mer are compared to the database, shown as larger rectangles on the right. For each taxon, a unique *k*-mer counter is instantiated, and the observed *k*-mers (K7, K8, and K9) are added to the counters. **b** Unique *k*-mer counting is implemented with the probabilistic estimation method HyperLogLog (HLL) using 16 KB of memory per counter, which limits the error in the cardinality estimate to 1% (see main text). **c** The output includes the number of reads, unique *k*-mers, duplicity (average time each *k*-mer has been seen), and coverage for each taxon observed in the input data

Physalia Courses

Advantage of classification over alignment: speed, Kraken2 is very fast!

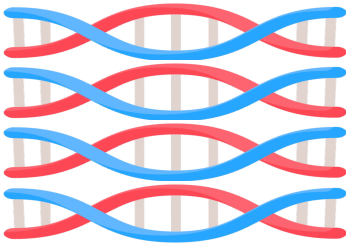
Coverage vs. depth vs. breadth of coverage

A)

Reference genome

$L_{genome} = 4 * L_{read}$

GCTACGATCTTTAGCTTAGCTGGGATCTGAATTCTCATCTCGGAT



L_{read}

Organism NOT detected

B)

Reference genome

$L_{genome} = 4 * L_{read}$

GCTACGATCTTTAGCTTAGCTGGGATCTGAATTCTCATCTCGGAT

L_{read} L_{read} L_{read} L_{read}

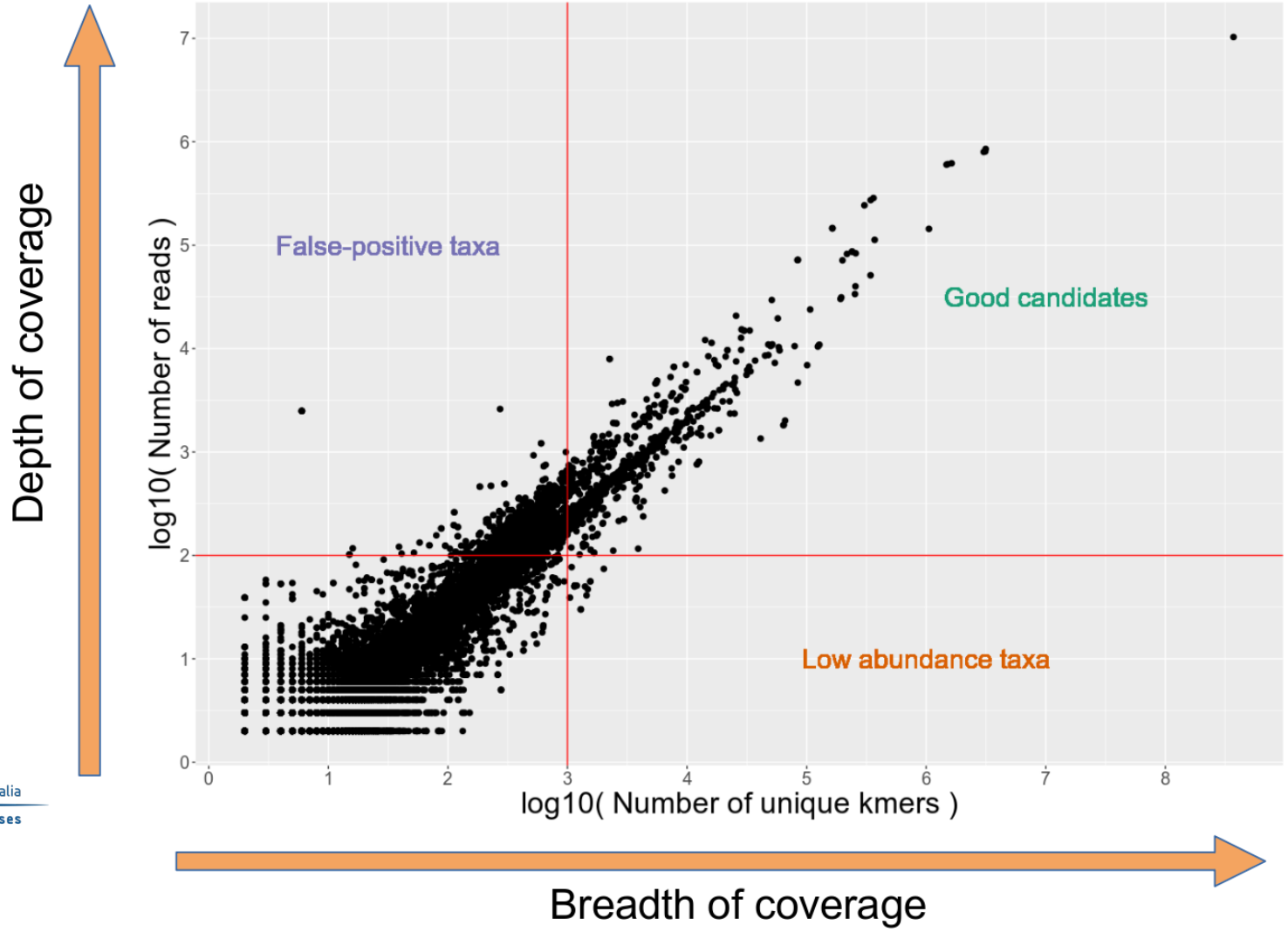
Organism detected

Physalia Courses

Both A) and B) have identical depth of coverage:

Coverage = $(N_{reads} * L_{read}) / L_{genome} = (4 * L) / (4 * L) = 1X$

Filtering Kraken output with respect to depth and breadth of coverage



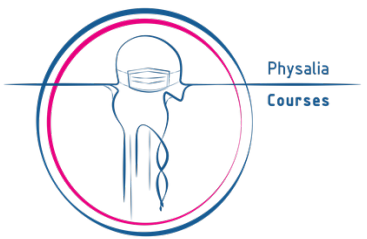
Physalia
Courses

DATABASES!

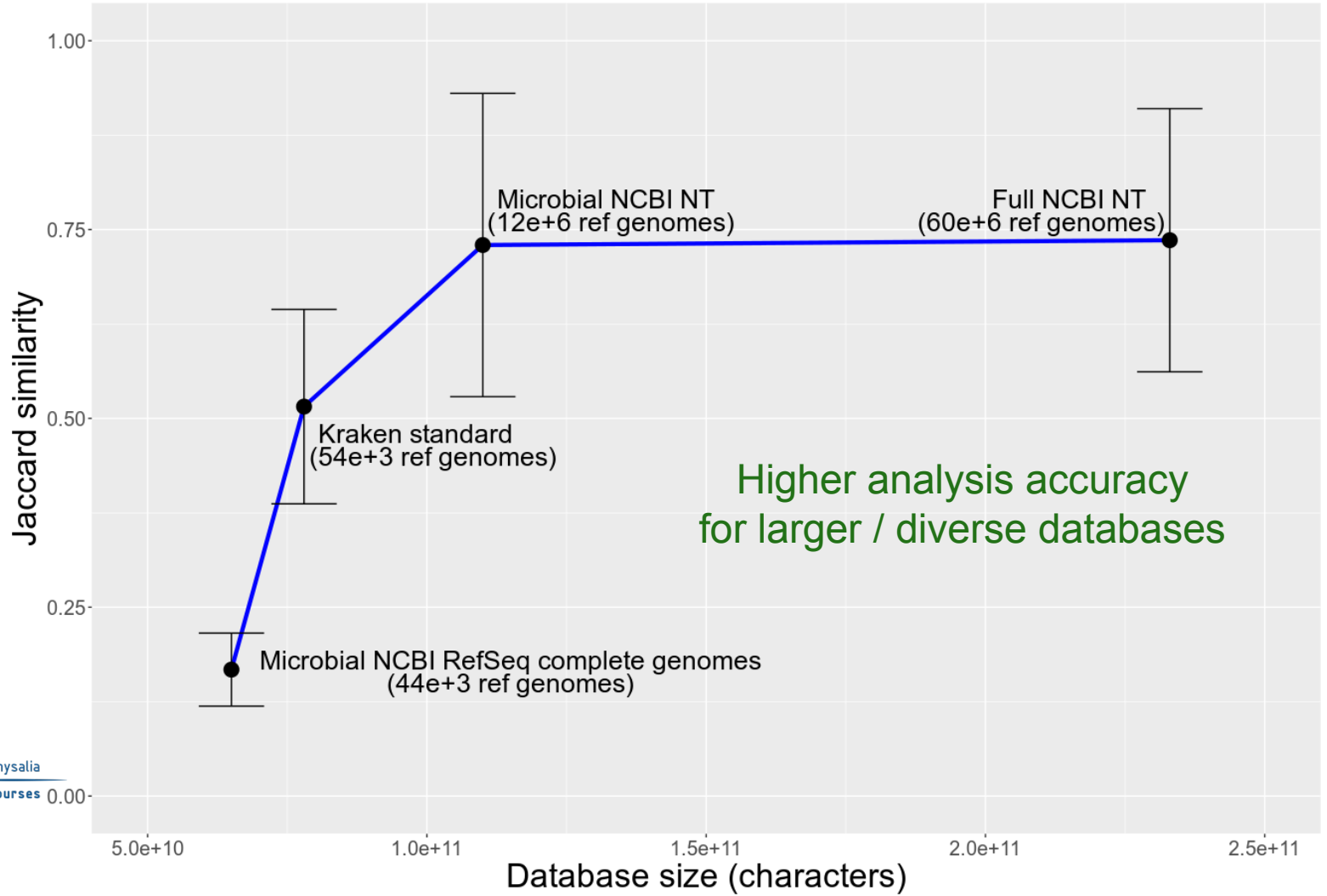
DATABASES!

DATABASES!

DATABASES!



Why large databases are important for accurate taxonomic profiling



Physalia Courses

Pochon et al.
Genome Biology 2023

aMeta: Pochon et al. Genome Biology 2023

Pochon *et al. Genome Biology* (2023) 24:242
https://doi.org/10.1186/s13059-023-03083-9

Genome Biology

METHOD Open Access

aMeta: an accurate and memory-efficient ancient metagenomic profiling workflow

Zoé Pochon[1,2†], Nora Bergfeldt[1,3,4†], Emrah Kırdök[5], Mário Vicente[1,2], Thijessen Naidoo[1,2,6,7], Tom van der Valk[1,4], N. Ezgi Altınışık[8], Maja Krzewińska[1,2], Love Dalén[1,3], Anders Götherström[1,2†], Claudio Mirabello[9†], Per Unneberg[10†] and Nikolay Oskolkov[11*†]

†Zoé Pochon, Nora Bergfeldt, Anders Götherström, Claudio Mirabello, Per Unneberg, and Nikolay Oskolkov shared authorship.

*Correspondence: Nikolay.Oskolkov@biol.lu.se

[11] Department of Biology, Science for Life Laboratory, National Bioinformatics Infrastructure Sweden, Lund University, Lund, Sweden
Full list of author information is available at the end of the article

Abstract

Analysis of microbial data from archaeological samples is a growing field with great potential for understanding ancient environments, lifestyles, and diseases. However, high error rates have been a challenge in ancient metagenomics, and the availability of computational frameworks that meet the demands of the field is limited. Here, we propose aMeta, an accurate metagenomic profiling workflow for ancient DNA designed to minimize the amount of false discoveries and computer memory requirements. Using simulated data, we benchmark aMeta against a current state-of-the-art workflow and demonstrate its superiority in microbial detection and authentication, as well as substantially lower usage of computer memory.

Keywords: Ancient metagenomics, Pathogen detection, Microbiome profiling, Ancient DNA

Background

Historically, ancient DNA (aDNA) studies have focused on human and faunal evolution and demography, extracting and analyzing predominantly eukaryotic aDNA [1–3]. With the development of next-generation sequencing (NGS) technologies, it was demonstrated that host-associated microbial aDNA from eukaryotic remains, which was previously treated as a sequencing by-product, can provide valuable information about ancient pandemics, lifestyle, and population migrations in the past [4–6]. Modern technologies have made it possible to study not only ancient microbiomes populating eukaryotic hosts, but also sedimentary ancient DNA (sedaDNA), which has rapidly become an independent branch of palaeogenetics, delivering unprecedented information about hominin and animal evolution without the need to analyze historical bones and teeth [7–12]. Previously available in microbial ecology, meta-barcoding methods lack validation and authentication power, and therefore, shotgun metagenomics has become the *de facto* standard in ancient microbiome research [13]. However, accurate detection,

README MIT license

aMeta

aMeta: an accurate and memory-efficient ancient Metagenomic profiling workflow

snakemake ≥6.10.0 Tests passing

About

aMeta is a Snakemake workflow for identifying microbial sequences in ancient DNA shotgun metagenomics samples. The workflow performs:

- trimming adapter sequences and removing reads shorter than 30 bp with Cutadapt
- quaity control before and after trimming with FastQC and MultiQC
- taxonomic sequence kmer-based classification with KrakenUniq
- sequence alignment with Bowtie2 and screening for common microbial pathogens
- deamination pattern analysis with MapDamage2
- Lowest Common Ancestor (LCA) sequence alignment with Malt
- authentication and validation of identified microbial species with MaltExtract

When using aMeta and / or pre-built databases provided together with the wokflow for your research projects, please cite our preprint: https://www.biorxiv.org/content/10.1101/2022.10.03.510579v1

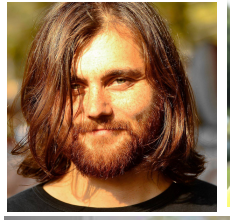
Authors

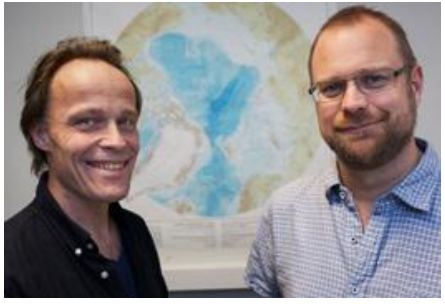
- Nikolay Oskolkov (@LeandroRitter) nikolay.oskolkov@scilifelab.se
- Claudio Mirabello (@clami66) claudio.mirabello@scilifelab.se
- Per Unneberg (@percyfal) per.unneberg@scilifelab.se

https://github.com/NBISweden/aMeta



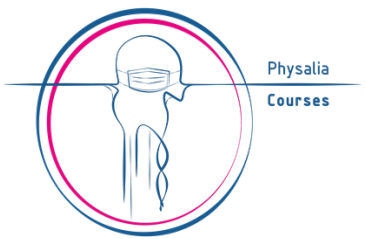




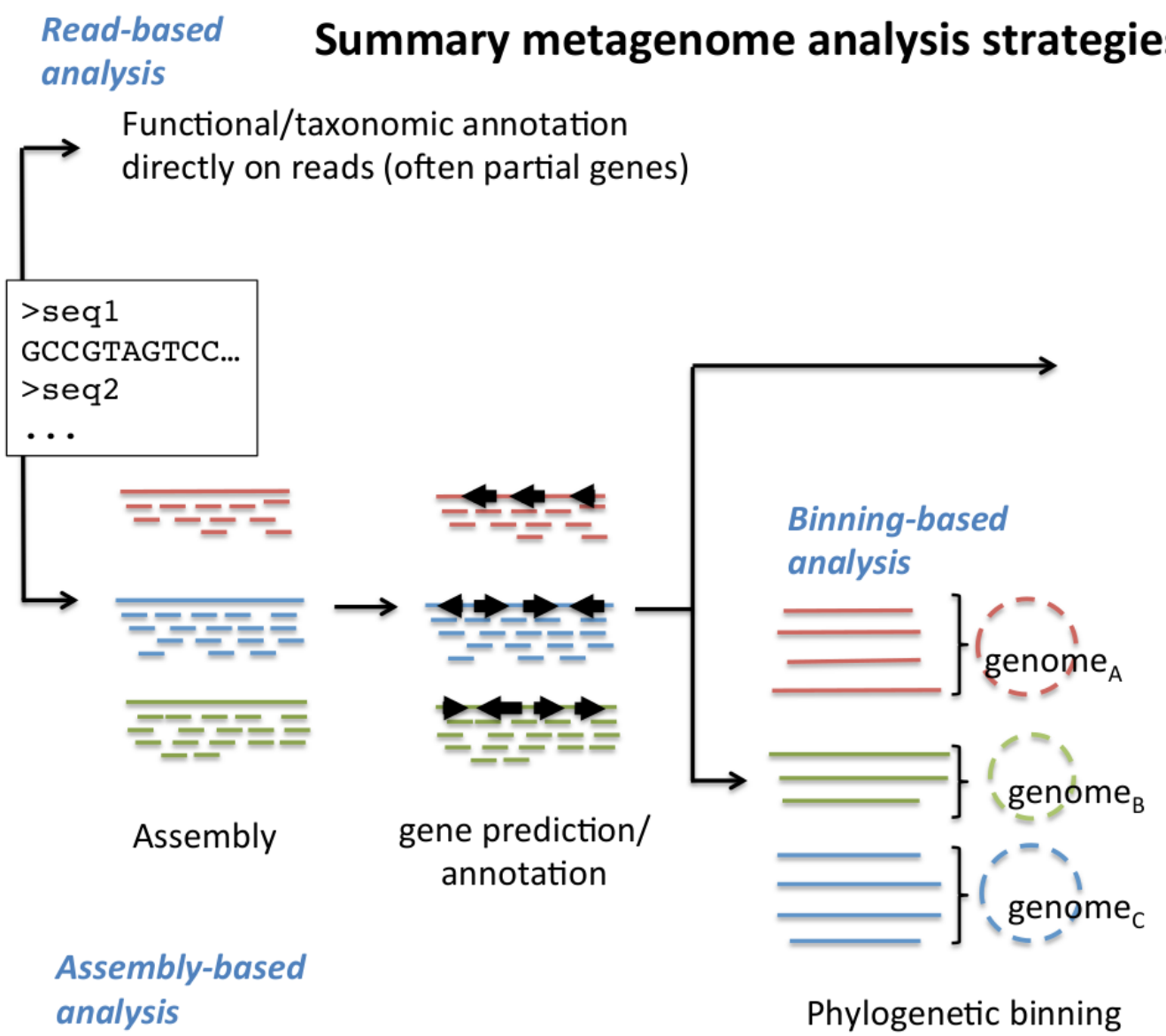


Strenghts and weaknesses of read-based metagenomics

Comprehensiveness	Can provide an aggregate picture of community function or structure, but is based only on the fraction of reads that map effectively to reference dbs
Community complexity	Can deal with communities of arbitrary complexity given sufficient sequencing depth and satisfactory reference database coverage
Novelty	Cannot resolve organisms for which genomes of close relatives are unknown
Computational burden	Can be performed efficiently, enabling large meta-analyses
Genome-resolved metabolism	Can typically resolve only the aggregate metabolism of the community, and links with phylogeny are only possible in the context of known reference genomes
Expert manual supervision	Usually does not require manual curation, but selection of reference genomes to use could involve human supervision
Integration with microbial genomics	Obtained profiles cannot be directly put into the context of genomes derived from pure cultured isolates



Summary metagenome analysis strategies



Physalia Courses

Strengths and weaknesses of assembly-based metagenomics

Comprehensiveness	Can construct multiple whole genomes, but only for organisms with enough coverage to be assembled and binned
Community complexity	In complex communities, only a fraction of the genomes can be resolved by assembly
Novelty	Can resolve genomes of entirely novel organisms with no sequenced relatives
Computational burden	Requires computationally costly assembly, mapping and binning
Genome-resolved metabolism	Can link metabolism to phylogeny through completely assembled genomes, even for novel diversity
Expert manual supervision	Manual curation required for accurate binning and scaffolding and for misassembly detection
Integration with microbial genomics	Assemblies can be fed into microbial genomic pipelines designed for analysis of genomes from pure cultured isolates