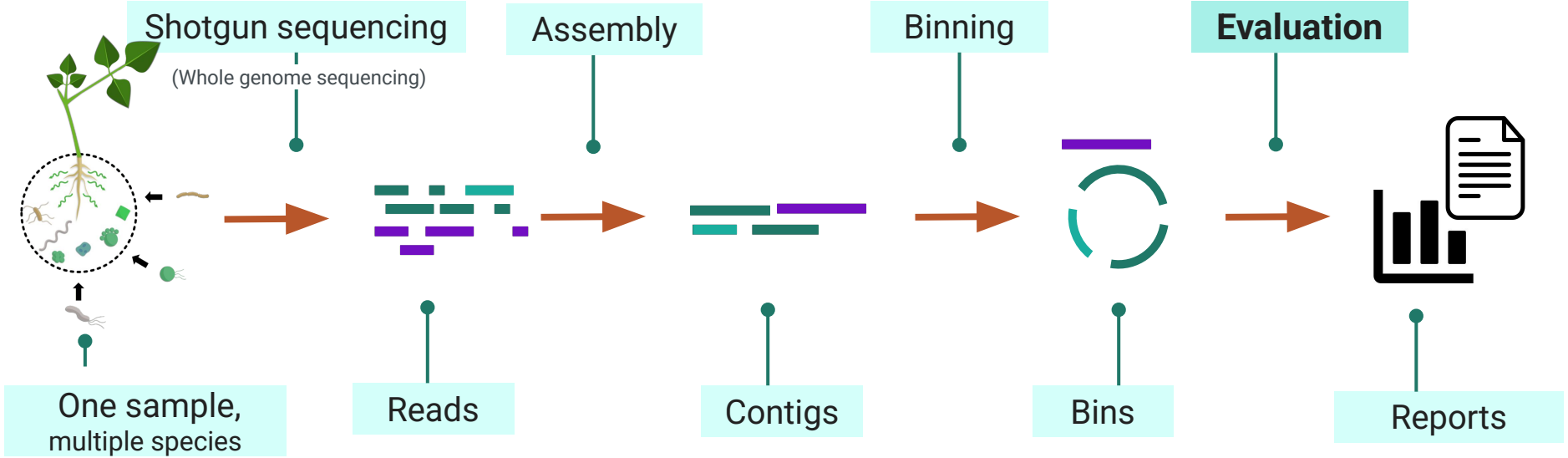
Mapler, a metagenome assembly and evaluation pipeline

Nicolas Maurice, Claire Lemaitre, Riccardo Vicedomini and Clémence Frioux

PEPI IBIS
October 2025

Introduction

Metagenome assembly



Introduction

Highly complex ecosystems

More species

- ➡ Lower coverage per species
- ➡ More shared sequences between species
- ➡ Lower quality assembly, less MAGs (Metagenome Assembled Genomes)

Introduction

Highly complex ecosystems

More species
- ➡ Lower coverage per species
- ➡ More shared sequences between species
- ➡ Lower quality assembly, less MAGs (Metagenome Assembled Genomes)

- ◆ Mock community[1] : 18,0 Gbp, 17 species ➡ ≈265X / species[5]
- ◆ Human gut[2] : 18,8 Gbp, >100[6] species ➡ < 50X / species
- ◆ Dead wood[3] : 16,1 Gbp, >1 000[7] species ➡ < 4X / species
- ◆ Soil[4] : 81,2 Gbp, >10 000[8] species ➡ < 2X / species

1 : ZymoD6331: SRR13128014
2 : Biocolective 139369 : SRR15275211
3 : Richy et al. (2024) : SRR28211698 to SRR28211701, coassembled
4 : Belliardo, Maurice et al. (2025)
5 : Assuming average genome ≈ 4 Mb
6 : Rowland et al. (2017)
7 : Pioli et al. (2023)
8 : Roesch et al. (2007)

Introduction

Metagenome assembly evaluation

Can its bin be exploitable as MAGs ?

Completeness, contamination, contiguity…

Near complete, high / medium / low quality

- Single bin assessment
- Single-sample multiple assembly comparison

Many evaluation tools already exist[1, 2]
Poorly suits complex ecosystems with few MAGs

1 : PacBio HiFi-MAG-Pipeline ; 2 : CheckM

Introduction

Metagenome assembly evaluation

Can its bin be exploitable as MAGs ?

Completeness, contamination, contiguity…

Near complete, high / medium / low quality

- Single bin assessment
- Single-sample multiple assembly comparison

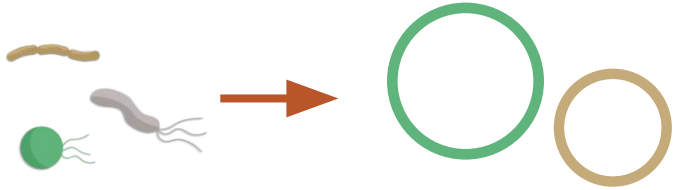
Many evaluation tools already exist[1, 2]
Poorly suits complex ecosystems with few MAGs

How representative of the sample is the assembly ?

How much of the diversity is captured by the assembly ?

What are the characteristics of that uncaptured diversity ?

Can such insight help us assemble better metagenomes ?



- Overall metagenome assessment i
- Multiple samples comparison

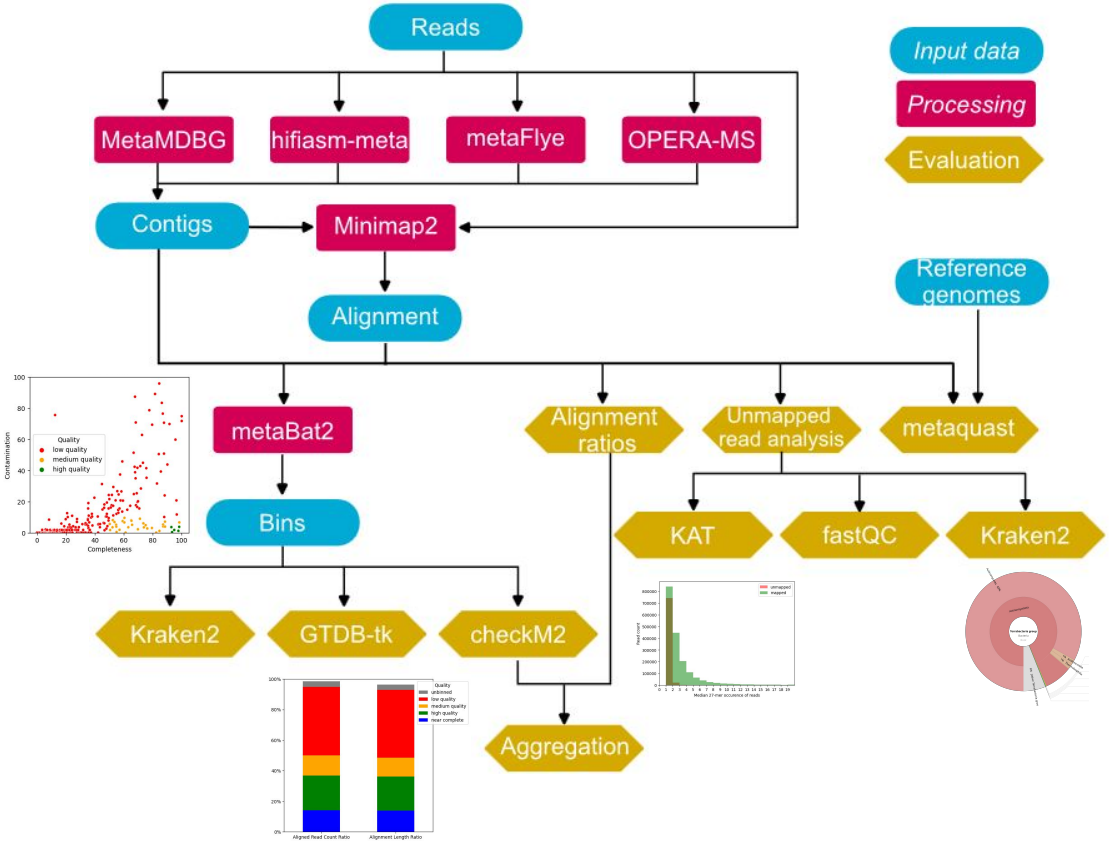
Lack of existing tools
Highly suits complex ecosystems with few MAGs

1 : PacBio HiFi-MAG-Pipeline ; 2 : CheckM

Mapler

Mapler: a pipeline for assessing assembly quality in taxonomically rich metagenomes sequenced with HiFi reads

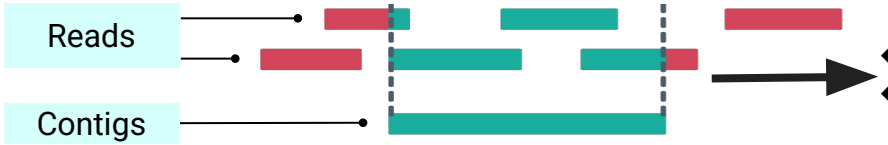
https://doi.org/10.1093/bioinformatics/btaf334



Bioinformatics, 2025

- Assembly & Binning
- **Evaluation**

Alignment ratios

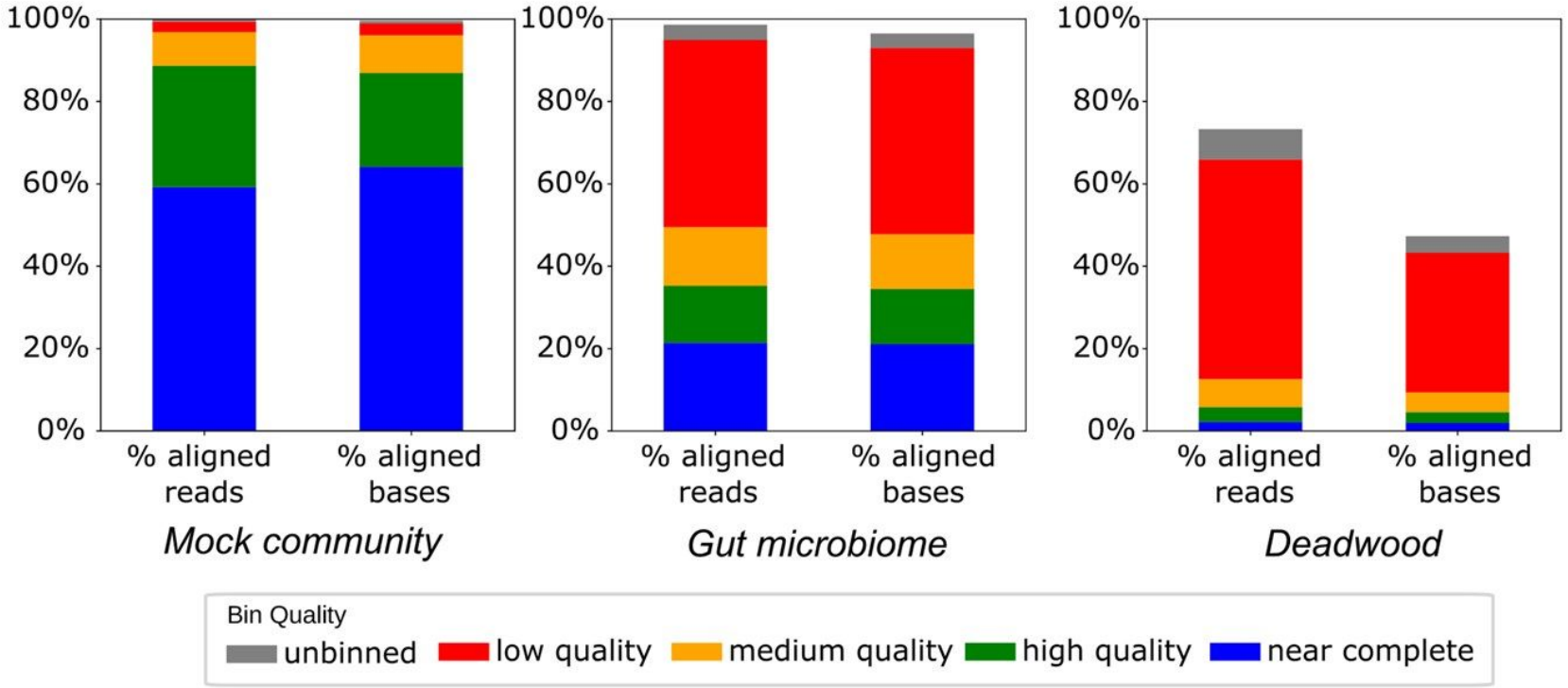


- Aligned reads ratio : Aligned read count / Total read count
- Aligned bases ratio : Aligned read bases / Total read length

Better alignment => Better representation of the sample diversity

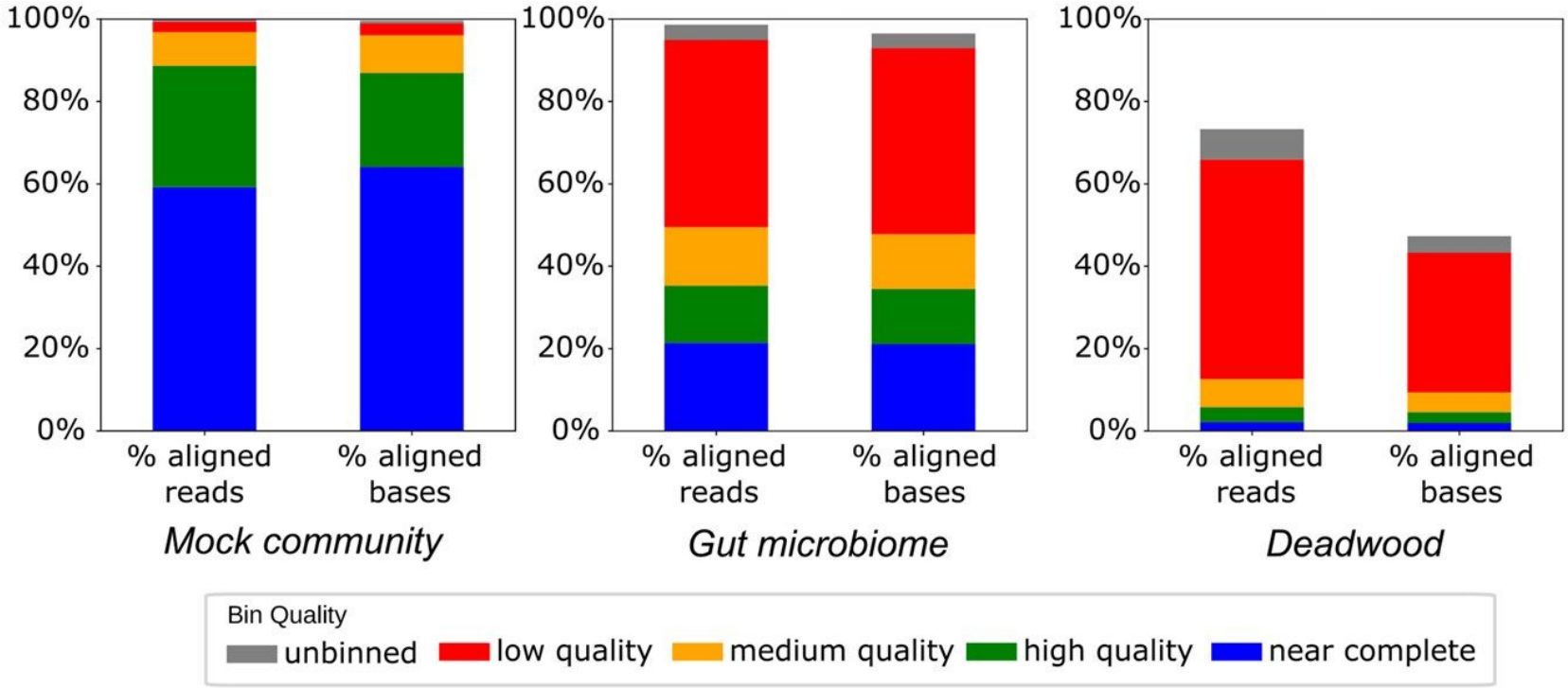
Aligned reads ratio >> Aligned bases ratio => Partially aligned reads

Alignment ratios of 3 metaMDBG[1] assemblies[2]



1 : Benoit et al. (2024) ; 2 : Maurice et al. (2025)

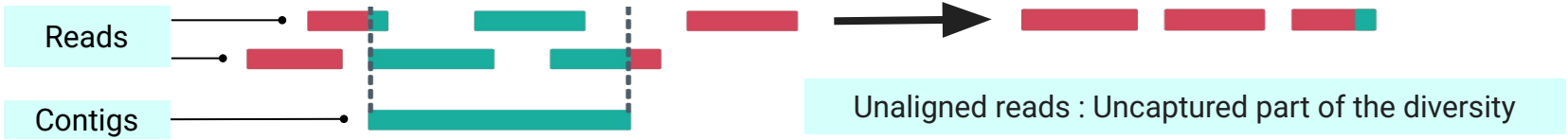
Alignment ratios of 3 metaMDBG[1] assemblies[2]



In highly complex ecosystem, most of the diversity is not captured by assembly

1 : Benoit et al. (2024) ; 2 : Maurice et al. (2025)

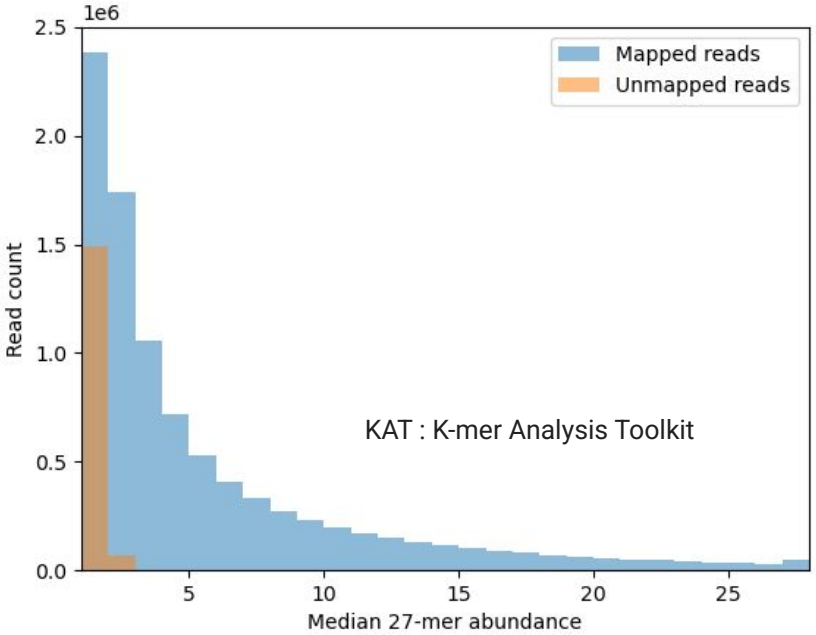
Uncaptured diversity



Unaligned reads : Uncaptured part of the diversity

Uncaptured diversity in the soil sample

Unmapped reads : Uncaptured part of the diversity



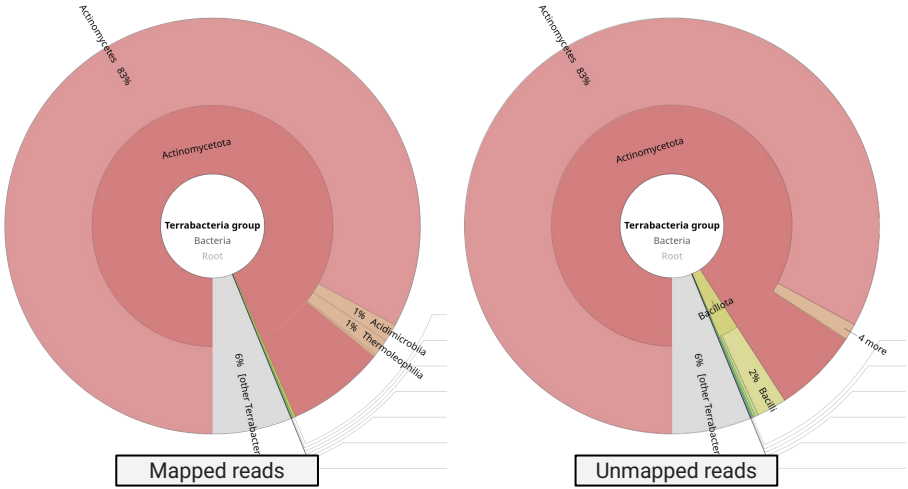
Belliardo, Maurice et al. (2025)

Compared to mapped reads, unmapped reads are :

- made up of rarer k-mers
  - Higher sequencing depth can improve assembly
  - k-mer abundance is not the only factor

Uncaptured diversity in the soil sample

Unmapped reads : Uncaptured part of the diversity

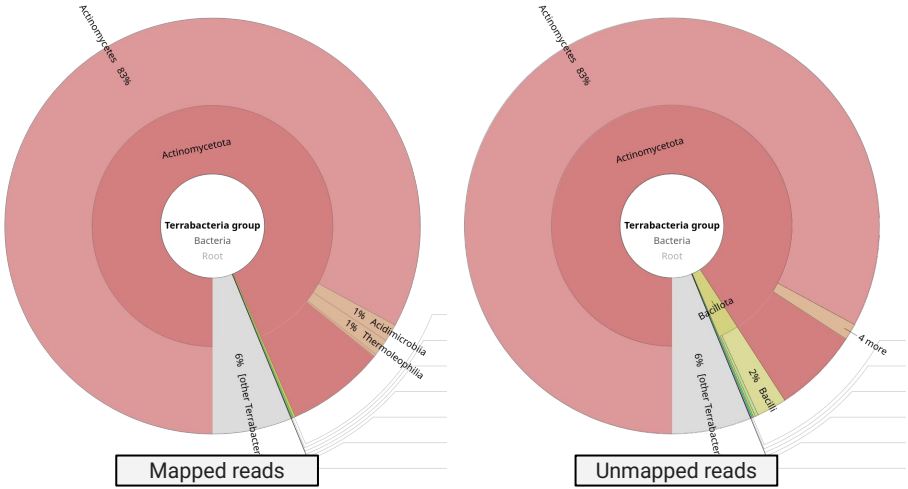


Compared to mapped reads, unmapped reads are :

- made up of rarer k-mers
  - ➡ Higher sequencing depth can improve assembly
  - ➡ k-mer abundance is not the only factor
- sometimes from under-represented taxa (Bacillota)

Uncaptured diversity in the soil sample

Unmapped reads : Uncaptured part of the diversity



Compared to mapped reads, unmapped reads are :

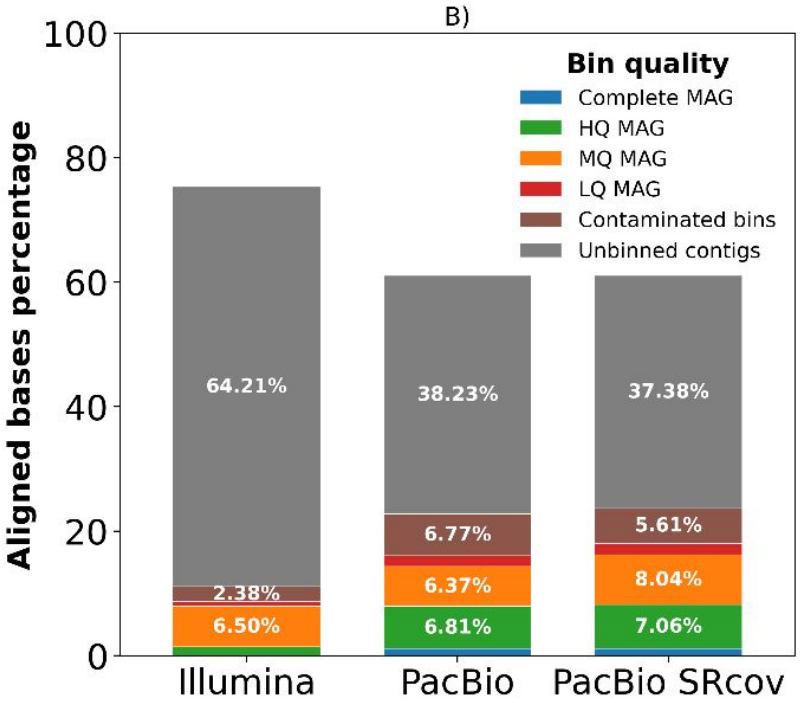
- made up of rarer k-mers
  - ➡ Higher sequencing depth can improve assembly
  - ➡ k-mer abundance is not the only factor
- sometimes from under-represented taxa (Bacillota)
- of slightly lower length (6.3 vs 7.1 Kbp average)
  - ➡ Higher read length can improve assembly, although not the most significant factor

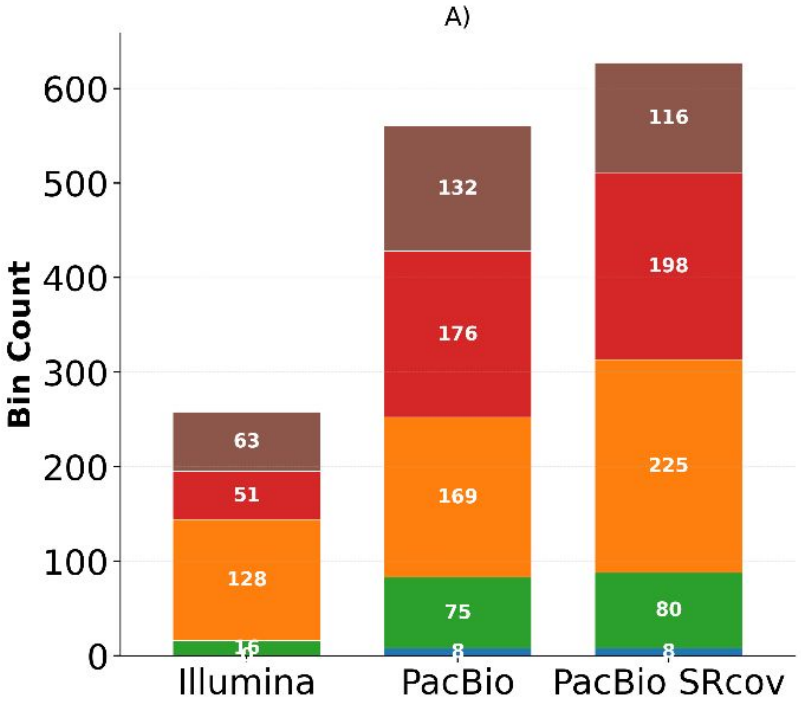
Application

Accurate MAG reconstruction from complex soil microbiome through combined short- and HiFi long-reads metagenomics

https://doi.org/10.1101/2025.09.12.675765

Mapler can highlight the diversity captured by Illumina assembly, even if poor contiguity prevents quality binning





Belliardo, Maurice et al. (2025)

Conclusion

In highly complex environments, most of the diversity cannot be captured

Mapler can assess the characteristics of that unassembled diversity :

- Low abundance populations
- Partially aligned reads

Which offer hindsight on how to improve assembly :

- Increase sequencing depth to reduce unique k-mers
  Although, 18Mb of mock community reads match 20 Gb of soil reads in term of assembly quality
- Improve DNA extraction and sequencing for longer reads
- Remap reads on contigs to improve alignment
- Use short-reads for binning if insufficient long-reads coverage

Many thanks to

PEPR "Agroécologie Numérique"
MISTIC project

Clémence Frioux
Claire Lemaitre
Riccardo Vicedomini
Carole Belliardo
Etienne Danchin
Marc Bailly-Bechet