Metagenomic assembly of complex ecosystems with highly accurate long-reads
Nicolas Maurice, Claire Lemaitre, Clémence Frioux, Riccardo Vicedomini

To cite this version:
Nicolas Maurice, Claire Lemaitre, Clémence Frioux, Riccardo Vicedomini. Metagenomic assembly of complex ecosystems with highly accurate long-reads. Journées 2024 du PEPR Agroécologie et Numérique, Jan 2024, Rennes, France. pp.1-1, 2024. hal-04425626

HAL Id: hal-04425626
https://inria.hal.science/hal-04425626
Submitted on 30 Jan 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution 4.0 International License
Metagenomic assembly of complex ecosystems with highly accurate long-reads

Nicolas Maurice\textsuperscript{1,2}, Claire Lemaitre\textsuperscript{1}, Clémence Frioux\textsuperscript{2}, Riccardo Vicedomini\textsuperscript{1}

\textsuperscript{1} Inria/IRISA Genscale, Campus de Beaulieu, 35042 Rennes, France
\textsuperscript{2} Inria/INRAE Pleiade, Campus de Talence, 33405 Bordeaux, France
nicolas.maurice@inria.fr

Context

Plant microbiota

Plants can recruit micro-organisms from the millions ($10^3$-$10^8$/g) found in soil, to regulate key functions such as:

- Nutrition and growth,
- Abiotic stress,
- Biotic stress.

Still, both ecosystems are poorly characterised, as cultureomics, transcriptomics and short-read metagenomics cannot recover enough complete genomes needed to modelise and characterise those communities.

MISTIC Project: project.inria.fr/mistic

Methodological development for microbial community dynamics modelling

- WP1: Top down approach from large-scale data on natural communities
- WP2: Bottom up approach, building dynamic models of increased complexity under controlled conditions.

Motivations

Emerging sequencing technologies

<table>
<thead>
<tr>
<th>Read type</th>
<th>Read length (bp)</th>
<th>Error rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Illumina</td>
<td>Short reads</td>
<td>151 bp</td>
</tr>
<tr>
<td>ONT R9.4</td>
<td>Long reads</td>
<td>$\sim$20kb</td>
</tr>
<tr>
<td>PacBio CLR</td>
<td>Highly accurate</td>
<td>$\sim$20kb</td>
</tr>
<tr>
<td>PacBio Hi-Fi</td>
<td>Long reads</td>
<td>$\sim$10kb</td>
</tr>
</tbody>
</table>

High-quality reads lead to higher quality assemblies.

How to exploit those highly accurate long-reads to recover high-quality genomes from complex environments such as soil and rhizosphere?

Mapler: gitlab.inria.fr/mistic/mapler

Pipeline

Built with Snakemake and Slurm integration.

Datasets

<table>
<thead>
<tr>
<th></th>
<th>Total size (Gb)</th>
<th>Median size (Mb)</th>
<th>Taxonomic complexity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zymo D6331</td>
<td>18.0</td>
<td>8.1</td>
<td>+</td>
</tr>
<tr>
<td>Human gut</td>
<td>15.2</td>
<td>8.4</td>
<td>++</td>
</tr>
<tr>
<td>Rhizosphere</td>
<td>19.5</td>
<td>5.8</td>
<td>++</td>
</tr>
<tr>
<td>Soil</td>
<td>13.4</td>
<td>6.0</td>
<td>+++</td>
</tr>
</tbody>
</table>

Evaluation metrics

- Aligning contigs on reference genomes, if available: NGA50 and genome fraction.
- Aligning reads on contigs: counting reads that map to the assembly.
- Evaluating the bins completeness and contamination with marker genes, then categorizing bins by level of quality.

Mapler: Results

Assembler comparison

On the soil sample

MetaMDBG captures:

- A greater proportion of reads,
- More high-quality bins,
- More medium-quality bins.

Sample comparison

Using metaMDBG

As taxonomic complexity increases:

- Quality of bins decreases,
- Proportion of captured taxa decreases.

Methodological avenues of improvement

Read partitioning

By partitioning the reads into smaller sets via:

- Taxonomic assignation,
- Composition-based clustering,
- Abundance-based clustering.

We could reduce the complexity of each subset, simplifying the assemblies.

Read recovery

By mapping the reads on contigs, we could extract unassembled reads, and assemble those separately with reduced complexity.

Hybrid assembly or binning

Highly accurate long reads could be complemented by a higher coverage of short reads, which could be used to better assess the abundance of each long-read, leading to a better assembly or binning.

References

[1] M.Kolmogorov et Al. metaFlye: scalable long-read metagenome assembly using repeat graphs, 2020
[2] X. Feng et Al. Metagenome assembly of high-fidelity long reads with hifiasm-meta, 2022