



HAL
open science

Reference guided genome assembly in metagenomic samples

Cervin Guyomar, Wesley Delage, Fabrice Legeai, Christophe Mougel,
Jean-Christophe Simon, Claire Lemaitre

► **To cite this version:**

Cervin Guyomar, Wesley Delage, Fabrice Legeai, Christophe Mougel, Jean-Christophe Simon, et al.. Reference guided genome assembly in metagenomic samples. RECOMB 2018 - 22nd International Conference on Research in Computational Molecular Biology, Apr 2018, Paris, France. pp.1. hal-01934823

HAL Id: hal-01934823

<https://hal.science/hal-01934823v1>

Submitted on 26 Nov 2018

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.

Reference guided genome assembly in metagenomic samples



Cervin Guyomar^{1,2}, Wesley Delage², Fabrice Legeai^{1,2},
Christophe Mougel¹, Jean-Christophe Simon¹, Claire Lemaitre²

1 : INRA, UMR 1349 IGEPP, le Rheu, France
2 : INRIA/IRISA GenScale, Campus de Beaulieu, Rennes, France



Motivations

Metagenomics = A mixture of reads :

- from different genomes
- with polymorphism (SNPs and large structural polymorphism)
- with no close reference genome (most of the time)



Objectives :

- Assemble an genome of interest
- from metagenomic reads,
- using a remote reference genome
- detecting structural polymorphism

Existing methods :

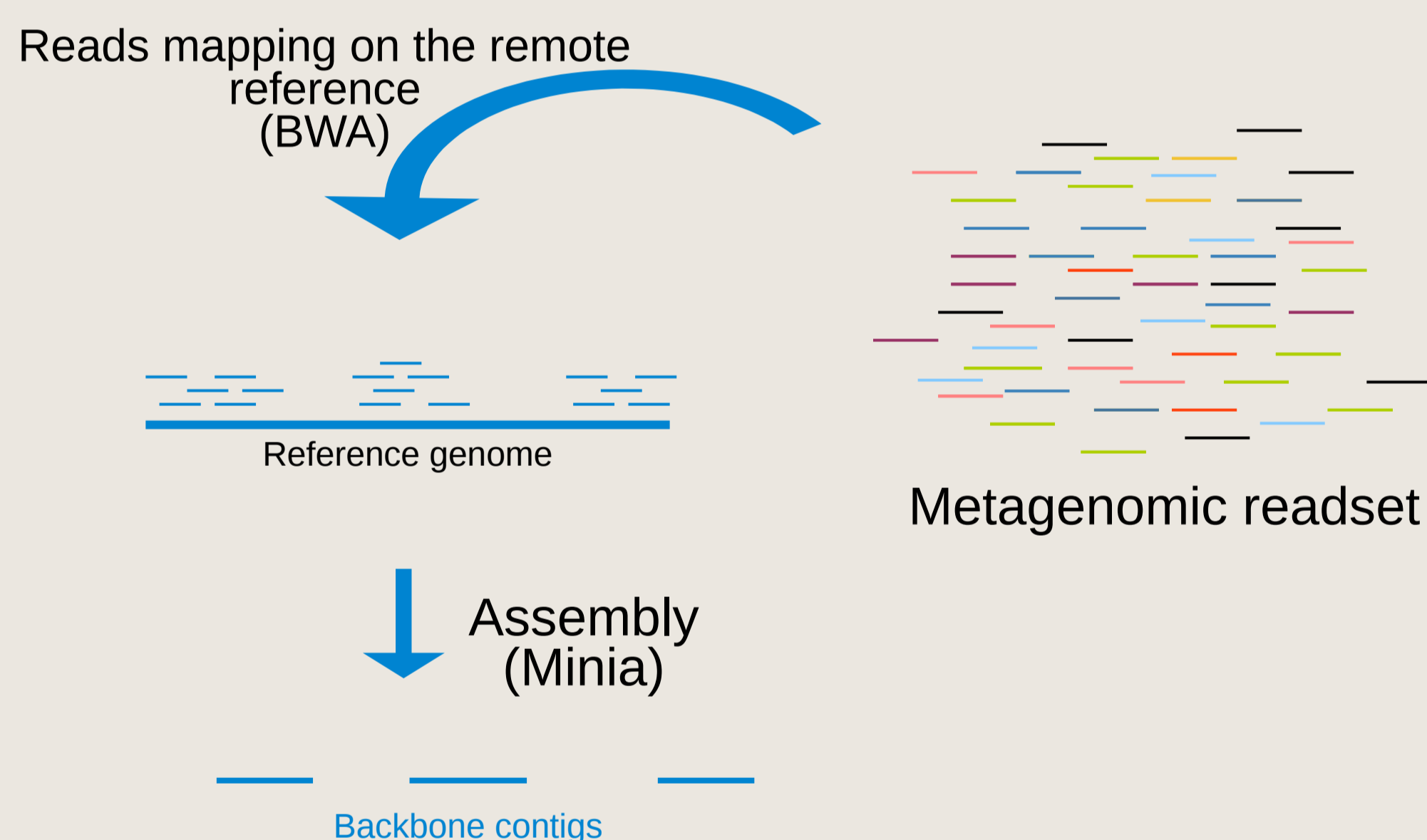
	Assembly first	Mapping first	Hybrid strategy
	De novo metagenomic assembly + contig taxonomic assignment Ex : MegaHit[1] + Blast	Assembly of reads Mapped on reference Ex : BWA+Minia[2]	Reference guided assembly Local assembly of the regions diverging from the reference Ex : MitoBim[3]
+	Assembly is reference-free	Assembly time reduced by read selection	- Conserved regions easily assembled using reference - Able to reconstruct diverging regions
-	- Time consuming and challenging assembly - Incomplete assembly if remote reference - Tricky contig filtering	- Requires a close reference genome - Incomplete assembly if remote reference	No tool suited for metagenomic data : - fail to detect structural polymorphism - No scale-up with metagenomic datasets

→ **Need for a tool dedicated to guided assembly in metagenomic context**

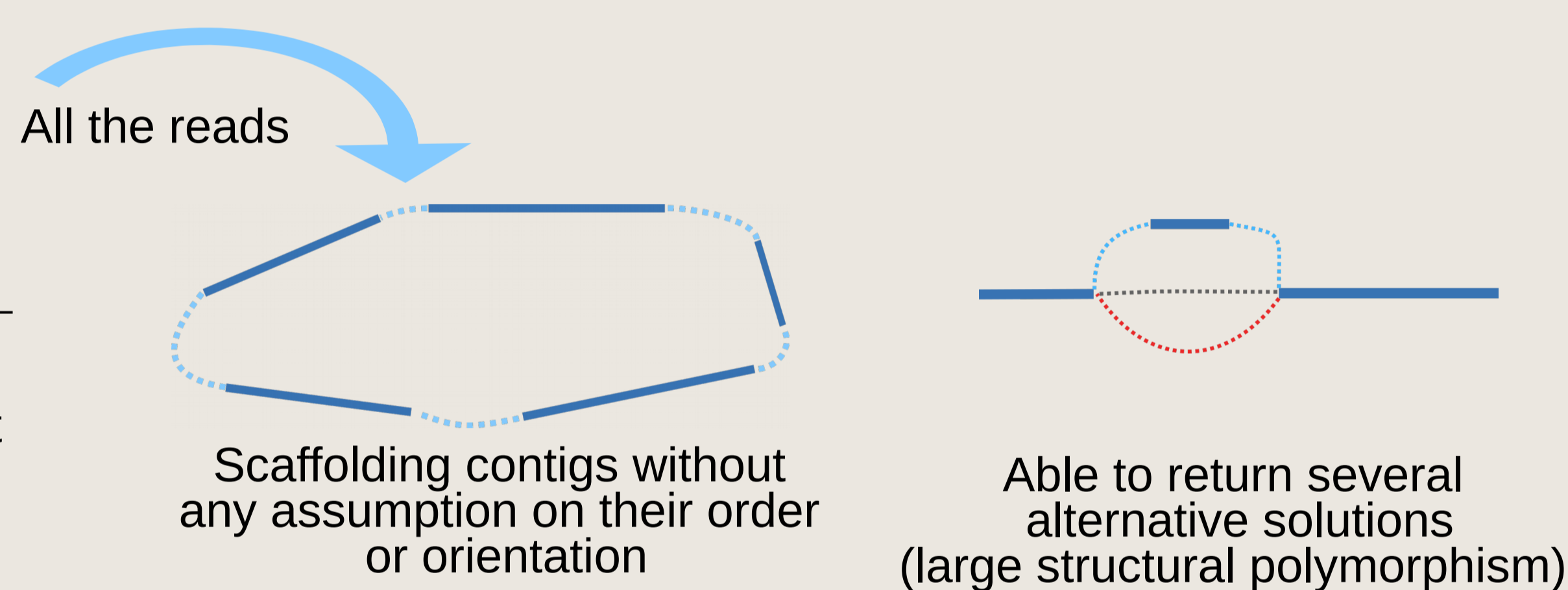
MindTheGap assembly workflow

- An hybrid approach :
- Mapping reads and assembly of conserved regions
 - Gapfilling of diverging regions with all reads

Step 1 : Reference based read recruiting and backbone contig assembly



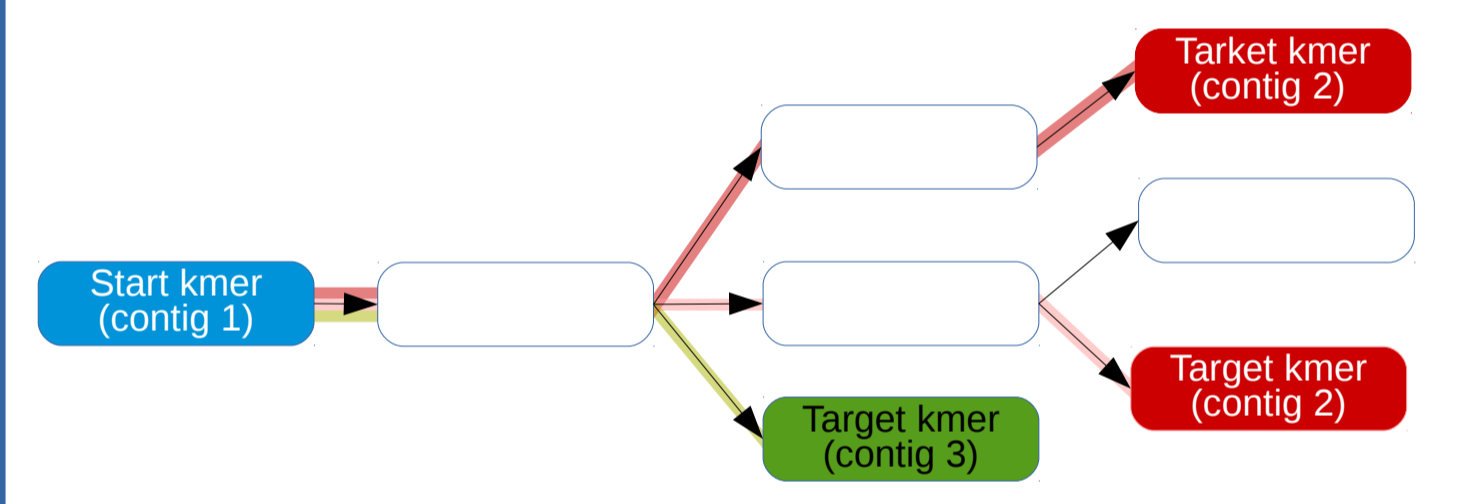
Step 2 : Reference free gapfilling between contigs using MindTheGap [4]



Output :
super-contigs (fasta) + gfa format

MindTheGap algorithm

- De Bruijn graph assembly starting from a contig end kmer
- Search target kmers in the contig graph



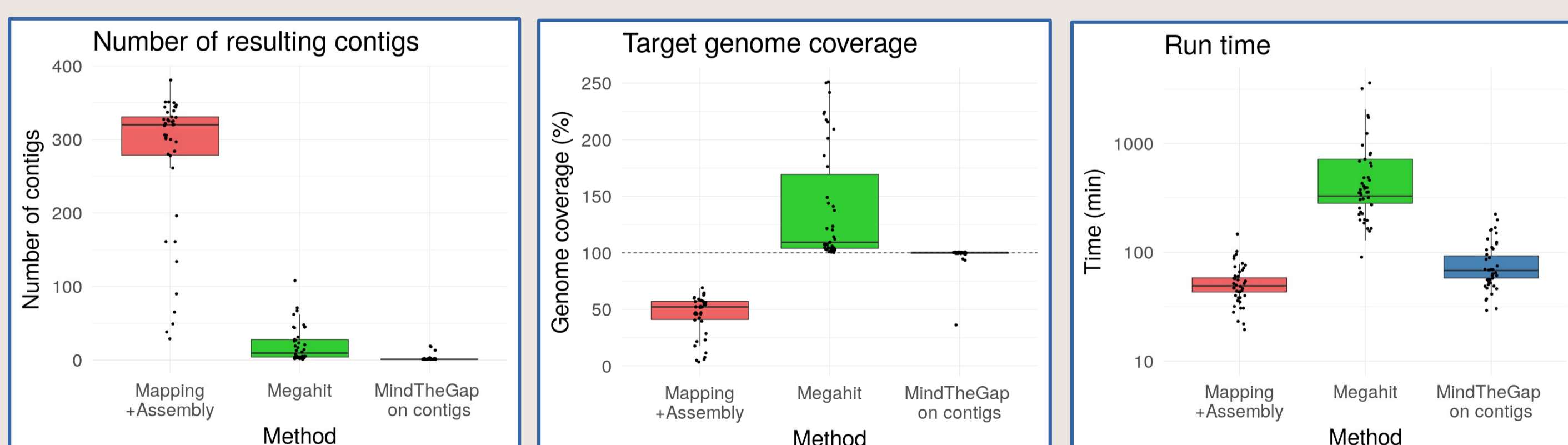
2 solutions between contig 1 and contig 2
1 solution between contig 1 and contig 3

Results in the pea aphid holobiont



Successful assembly of a bacterial genome in one circular contig

Assembly of *Buchnera aphidicola* from 42 pea aphid metagenomic samples [5] using *Buchnera* genome from another species

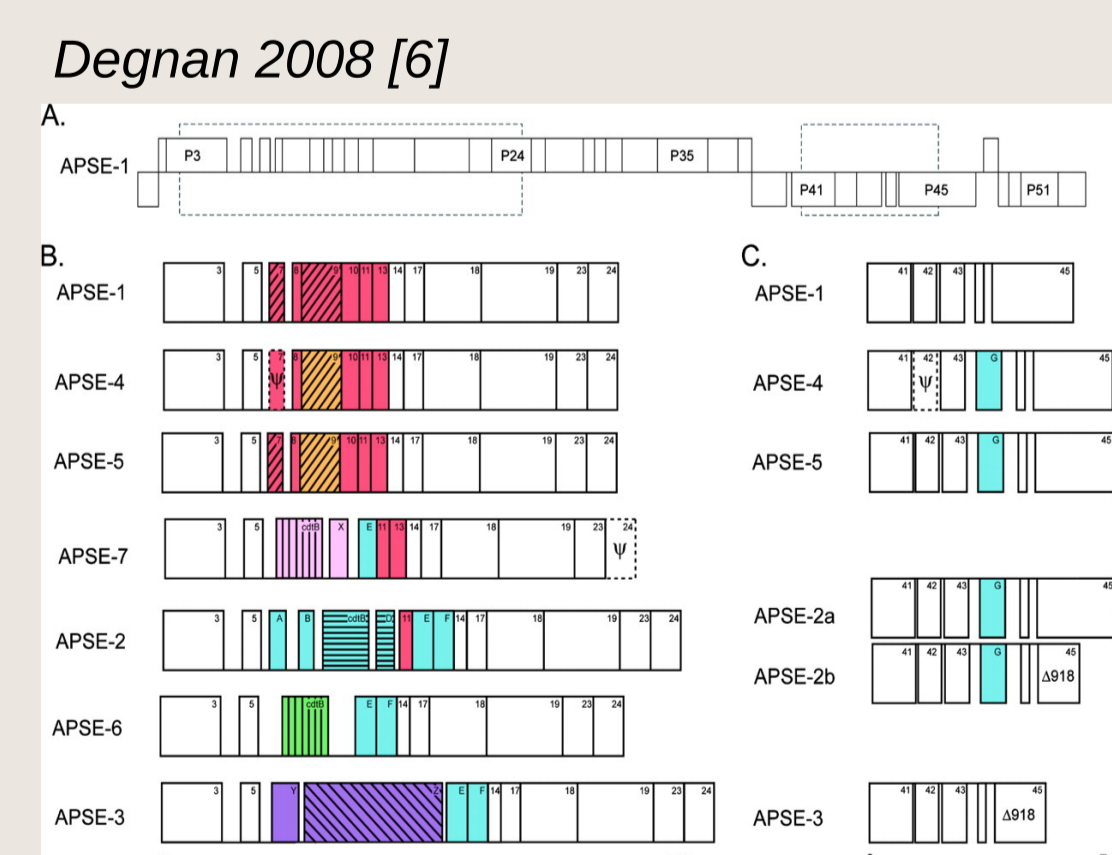


MindTheGap results

- One-contig assembly for 30 samples
- Genome length close to the real reference
- On average 7 times faster than Megahit

Discovery of unknown phage structural variants

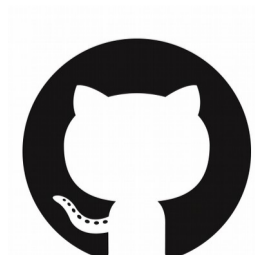
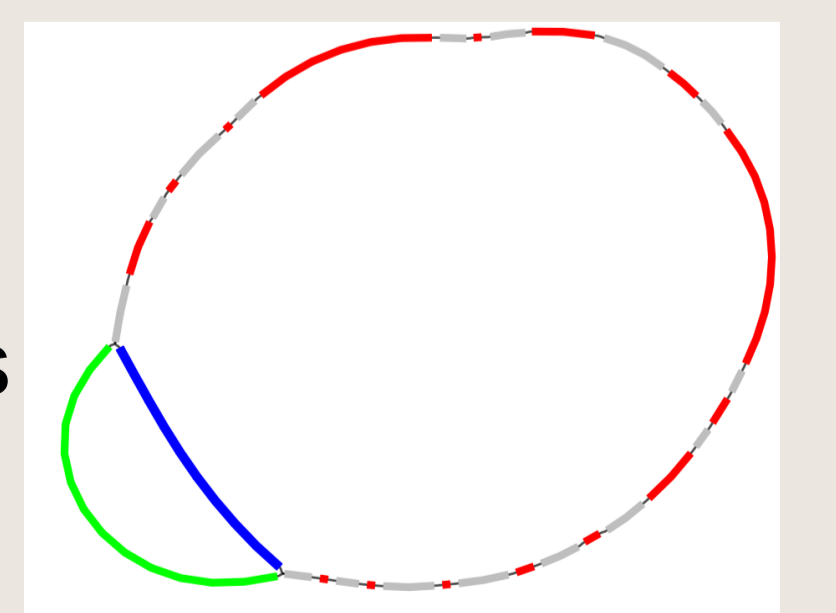
Assembly of the phage APSE from 42 pea aphid metagenomic samples



- 7 known variants known, differing by a ~5kb virulence cassette

Results :

- 3 new phage variants discovered in 5 samples
- Coabundant phage successfully assembled in 3 samples



In development on GitHub

https://github.com/GATB/MindTheGap/tree/contig_dev

References :

- [1] : Li et al (2015) MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics*
- [2] : Chikhi et al (2013) Space-efficient and exact de Bruijn graph representation based on a Bloom filter. *Algorithms for molecular biology : AMB*
- [3] : Hahn et al (2013) Reconstructing mitochondrial genomes directly from genomic next-generation sequencing reads - A baiting and iterative mapping approach. *Nucleic Acids Research*
- [4] : Rizk, G. et al., (2014) MindTheGap: integrated detection and assembly of short and long insertions. *Bioinformatics*,
- [5] : Guyomar et al (2018) Multi-scale characterization of symbiont diversity in the pea aphid complex through metagenomic approaches (*in revision - Microbiome*)
- [6] : Degnan and Moran (2008) Diverse Phage-Encoded Toxins in a Protective Insect Endosymbiont, *Applied and Environmental Microbiology*