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.archives-ouvertes.fr/hal-01934823
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

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,
- detecting stuctural polymorphism

Motivations

Existing methods:

| Step 1 : Reference based read recruiting and backbone contig assembly |
| Assembly first |
| De novo metagenomic assembly + contig taxonomic assignation Ex : MEGAhit[1] + Blast |
| Mapping first |
| Assembly time reduced by read selection Ex : BWA+Minia[2] |
| Hybrid strategy |
| Reference guided assembly Local assembly of the regions diverging from the reference Ex : MindTheGap[3] |

- Time consuming and challenging assembly
- Incomplete assembly if remote reference
- Tricky contig filtering

- Conserved regions easily assembled using reference
- Able to reconstruct diverging regions

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

MindTheGap assembly workflow

Step 1: Reference based read recruiting and backbone contig assembly

- Reads mapping on the remote reference (BWA)
- All the reads
- Metagenomic dataset
- Assembly (Minia)
- Scaffold contigs without any assumption on their order or orientation
- Able to return several alternative solutions (large structural polymorphism)


- Target kmer (contig 1)
- Target kmer (contig 2)
- Target kmer (contig 3)
- Reference genome

MindTheGap algorithm
- De Brujin 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

Output:
super-contigs (fasta) + gfa format

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

References:


In development on GitHub
https://github.com/GATB/MindTheGap/tree/contig_dev