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

Cervin Guyomar1,2, Wesley Delage2, Fabrice Legera1,2, Christophe Mouge1, 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 stuctural polymorphism

Existing methods :

<table>
<thead>
<tr>
<th>Assembly first</th>
<th>Mapping first</th>
<th>Hybrid stategy</th>
</tr>
</thead>
<tbody>
<tr>
<td>De novo metagenomic assembly + contig taxonomic assignation Ex : MegaHit[1] + Blast</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Assembly is reference-free Assembly time reduced by read selection</td>
<td></td>
<td></td>
</tr>
<tr>
<td>- Time consuming and challenging assembly - Incomplete assembly if remote reference - Tricky contig filtering</td>
<td></td>
<td></td>
</tr>
<tr>
<td>- Conserved regions easily assembled using reference - Able to reconstruct diverging regions</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Requires a close reference genome - Incomplete assembly if remote reference</td>
<td></td>
<td></td>
</tr>
<tr>
<td>No tool suited for metagenomic data : - fail to detect structural polymorphism - No scale-up with metagenomic datasets</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Step 1 : Reference based read recruiting and backbone contig assembly

Reads mapping on the remote reference (BWA)
Reference genome
Metagenomic readset
Assembly (Minia)
Backbone contig


All the reads
Scaffolding contigs without any assumption on their order or orientation
Able to return several alternative solutions (large structural polymorphism)

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

References :

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