Developing a MAGs recovery pipeline for large cohorts

Kateryna Pantiukh, Elin Org
Institute of Genomics, University of Tartu, Estonia

Research group of microbiome

Metagenome-assembled genomes (MAGs) can be recovered from metagenomic deep sequencing data. This genome sequence opens a wide range of possibilities for further functional and metabolic potential analysis. MAGs can be recovered using a variety of assembly methods.

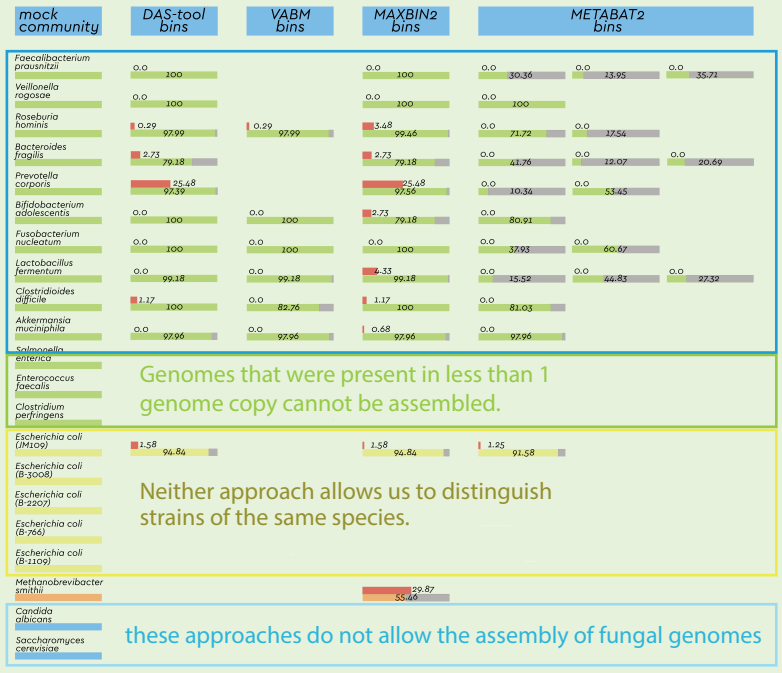
Human health and disease are strongly influenced by the gut microbiome. Large cohorts from Biobanks are often used to understand how microbes affect human health. But working with large cohorts requires a lot of computing power and time.

We set ourselves the goal of finding the optimal pipeline for MAGs recovery. Pipeline was evaluated in terms of the balance between assembly quality (MAGs completeness, contamination, fragmentation) and time and computational costs.

MAGs recovery process has three main steps:

Contigs assembly → Contigs binning → Bins refining & evaluation

Binning strategy. Single binner vs multi-binning with DAS-tool



Binning programs were checked with mock communnty data (ZymoBIOMICS® Gut Microbiome Standard Catalog No. D6331).

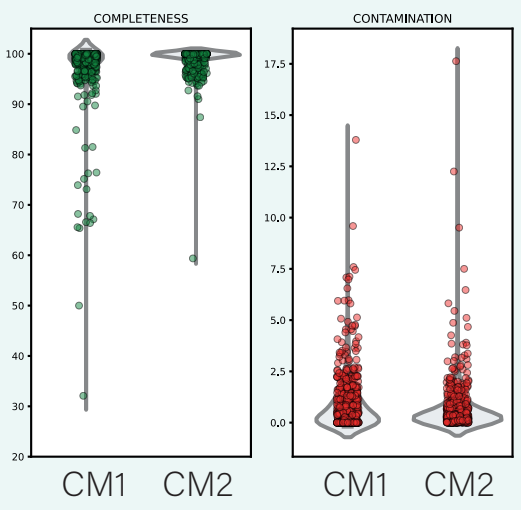
The multi-binning approach gives results much closer to the true community composition.

You can dowload full picture HERE

picture

Evaluation Bins quality. CheckM1 vs CheckM2

Programs quality were analyzed on complete circulated genomes downloaded from NCBI database. A total of 715 bacterial genomes from 33 main phylum were analyzed.

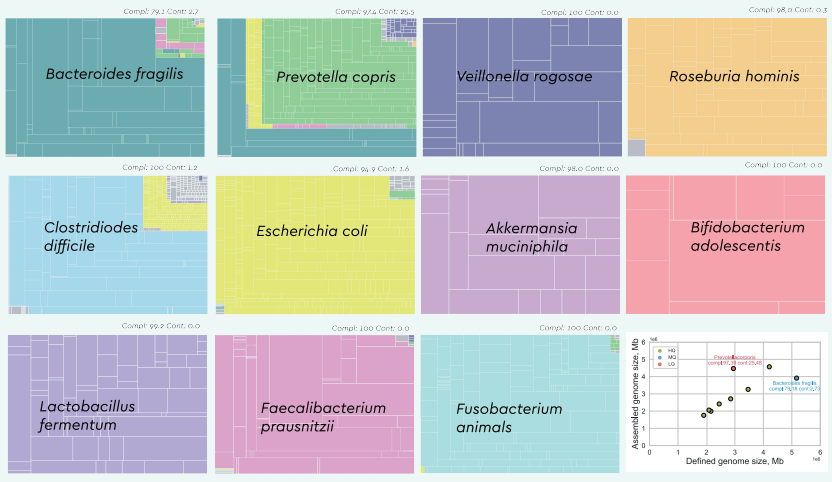


Assuming that the genomes loaded into the database are indeed complete and do not contain parts of other genomes, the completeness for all genomes should be 100 and contamination should be 0.

Results: the mean completeness for CheckM1 was 98,2 ± 5 (19 genomes had completeness <90). The mean completeness for CheckM2 was 99,4 ± 2 (only 2 genomes had completeness <90).

code&data

DAS-tool bins refining. Sourse of contaminations



sample: ZymoBIOMICS® Gut Microbiome Standard Catalog No. D6331

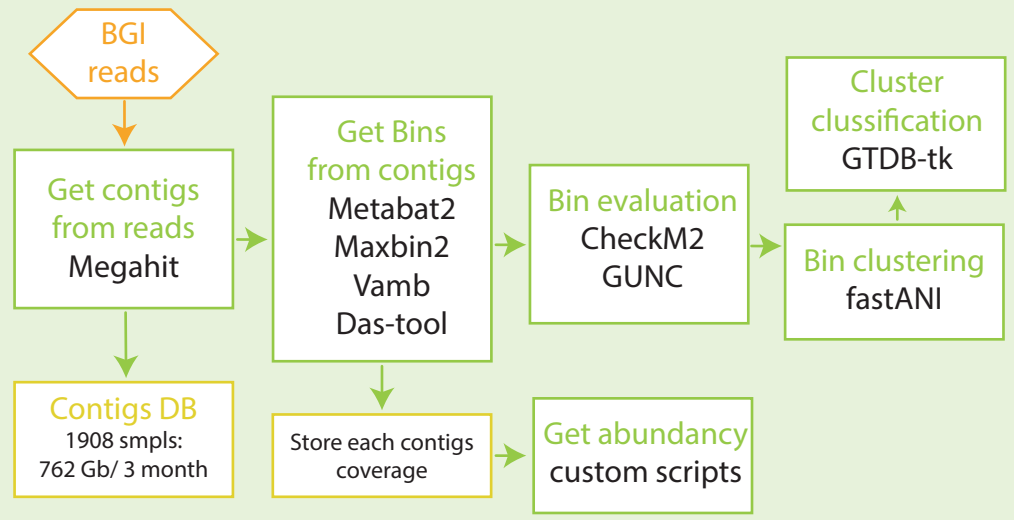
binning strategy: multi-binning (DAS-tool)

bins classification: GTDB-tk

bins contigs classification: GUNC

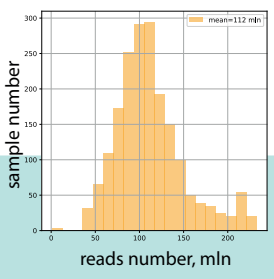
Each big box represents one final bin, the small cells inside the box represent contigs of these bins. The colors reflect the classification of the contigs.

Final pipeline for MAGs recovery

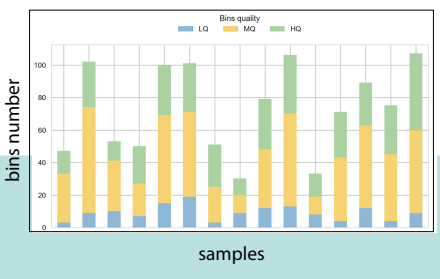


Check pipeline with a real human gut microbiome data (deep shotgun sequencing)

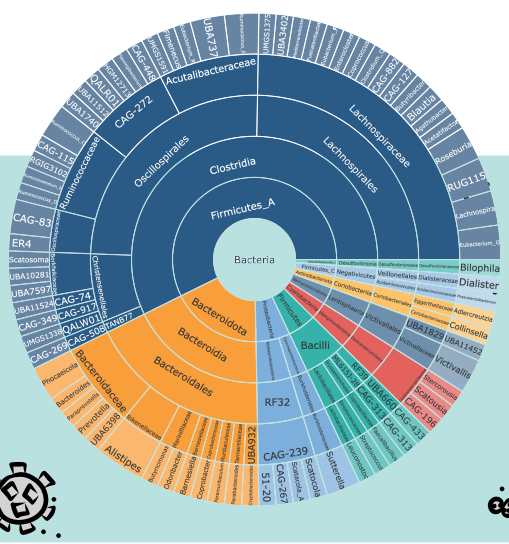
BGI reads data



Final Bins set



Final Bins set classification



Estonian Microbiome Cohort

N=2509

1764 745

Stool samples
oral samples
Plasma samples

my twitter

Estonian Research Council