BITS :: Call for Abstracts 2022 - Oral communication

Type	Oral communication
Session	Metagenomics
Title	New challenges in shotgun metagenomics with kMetaShot
All Authors	Defazio G(1), Fosso B(1,2), Pesole G(1,2)

Affiliation

(1) Department of Biosciences, Biotechnology and Biopharmacology, University of Bari, Bari
(2) Institute of Biomembranes, Bioenergetics and Molecular Biotechnologies, National Research Council, Bari

Motivation

Shotgun Metagenomics allows to unveil the composition of microbial communities colonizing several biological niches such as human gut and vagina or environments such as wastewater. The DNA extraction and direct sequencing via Next Generation Sequencing (NGS) platforms allow to generate genomic sequences by using meta-assembler and binning algorithms. Once “draft genomes” also called Metagenome Assembled Genomes (MAGs) are obtained, they can be classified in an appropriate taxonomy and annotated via gene prediction. To date, there are few tools able to taxonomically classify MAGs and some rely on alignment approaches as CheckM [1] or on multiple strategies such as CAMITAX [2]. Alignment-based algorithms assume collinearity among the analyzed sequences relying on a homology hypothesis but do not consider high mutation rate, genetic recombination, horizontal gene transfers and genes duplication [3]. Moreover, parameters and heuristics affect alignment-based algorithms reproducibility. In this scenario we present kMetaShot, an alignment-free taxonomic classifier based on k-mer/minimizer counting.

Methods

kMetaShot is made by a reference generator and a classifier module. The reference generator algorithm is based on the selection of relevant minimizers at genus and strain levels in a taxonomic division of interest (e.g. Bacteria). In brief, if the k-mer is a DNA sequence substring k nucleotides long, a minimizer is a k-mer substring n nucleotides long, chosen using a minimizing criterion (e.g. lexicographic) to synthetically represent a k-mer [4]. kMetaShot starts executing the minimizer counting in CDS (coding sequences) and ncRNAs (non-coding RNAs) from bacterial genome available in RefSeq. This operation produces a non-redundant set of minimizers for each genome. Then, by using the NCBI taxonomy the algorithm discards minimizers shared between genomes belonging to different genera and retains only minimizers exclusively represented in genome(s) of the same strain or the same genus. Retained minimizers and related taxonomy identifiers (taxid) are stored in a so-called storage matrix. Once the storage matrix is complete, kMetaShot is able to classify query sequences or bins/MAGs by applying a classification algorithm. It firstly decomposes the query sequences/MAGs in minimizers sets and then query the storage matrix to retrieve related taxonomic information if available. Finally, the classification algorithm applies a prevalence criterion to perform the taxonomic classification.
The kMetaShot reference generator was applied on 199,863 RefSeq Bacterial genomes (downloaded 1/10/2020) corresponding to 3,252 genera and 58,550 strains and the obtained reference was tested on 954 bacterial HMP (Human Microbiome Project [5]) genomes. In the testing phase ass2ref (a parameter weighting the assignment reliability) was finely tuned to avoid unreliable strain classifications. Also, a benchmark against CheckM and CAMITAX was performed by using a mock community containing 100 RefSeq Bacterial Genomes.

Results

The kMetaShot reference generator algorithm founds 665,757,464 relevant minimizers.
In the test performed on the HMP genomes, kMetaShot correctly assigned 489 (51.26%), 853 (89.41%), 909 (95.28%) genomes at strain, species and genus, respectively. The ass2ref parameter application improved precision from 51.26% to 76.86% for classification at strain level. Finally, a benchmark was performed to classify 98 MAGs obtained from meta-assembly and binning of the in silico generated mock community. 97, 77 and 69 MAGs were correctly assigned at species level respectively by kMetaShot, CheckM and CAMITAX. The False Positive Rate was 0.0002 for kMetaShot and 0.0001 for CheckM and CAMITAX at species level. kMetaShot was the only one able to correctly assign 68 bins at strain level.

Info

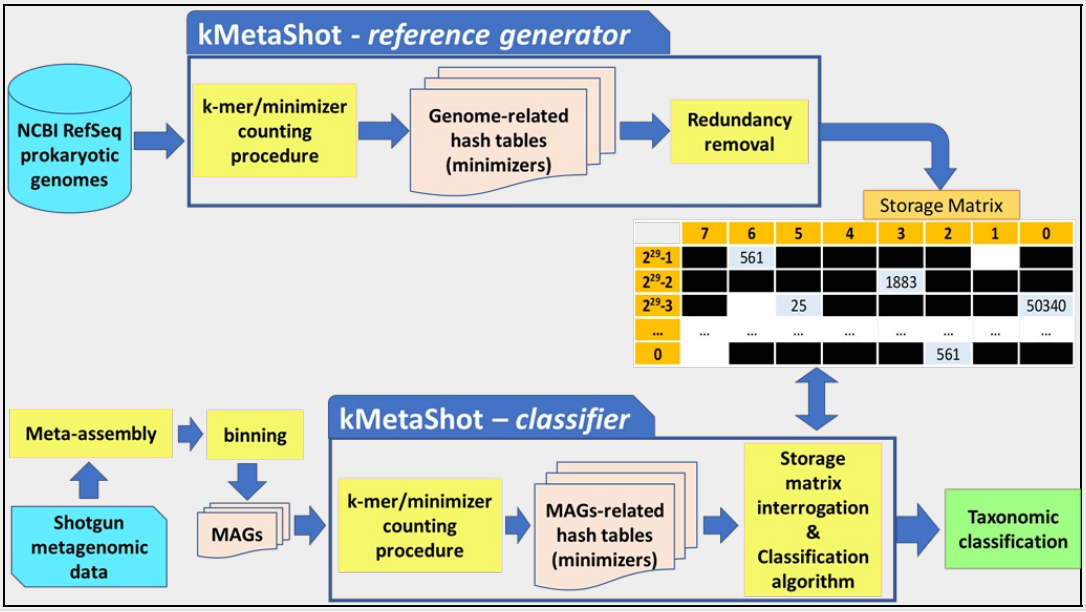
References
[1] D. H. Parks, M. Imelfort, C. T. Skennerton, P. Hugenholtz, and G. W. Tyson, “CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes,” Genome Res, vol. 25, no. 7, pp. 1043–1055, Jul. 2015, doi: 10.1101/gr.186072.114.
[2] A. Bremges, A. Fritz, and A. C. McHardy, “CAMITAX: Taxon labels for microbial genomes,” GigaScience, vol. 9, no. giz154, Jan. 2020, doi: 10.1093/gigascience/giz154.
[3] A. Zielezinski, S. Vinga, J. Almeida, and W. M. Karlowski, “Alignment-free sequence comparison: benefits, applications, and tools,” Genome Biol., vol. 18, no. 1, p. 186, 03 2017, doi: 10.1186/s13059-017-1319-7.
[4] M. Roberts, W. Hayes, B. R. Hunt, S. M. Mount, and J. A. Yorke, “Reducing storage requirements for biological sequence comparison,” Bioinformatics, vol. 20, no. 18, pp. 3363–3369, Dec. 2004, doi: 10.1093/bioinformatics/bth408.
[5] P. J. Turnbaugh, R. E. Ley, M. Hamady, C. M. Fraser-Liggett, R. Knight, and J. I. Gordon, “The

Human Microbiome Project," Nature, vol. 449, no. 7164, pp. 804–810, Oct. 2007, doi: 10.1038/nature06244.

_____

Image file represents the kMetaShot flow chart for "reference generator" and "classifier" modules.

filename	kMetaShot_flowchart_abstract.png
Figure	



Availability	-

Corresponding Author

Name, Surname	Graziano, Pesole
Email	graziano.pesole@uniba.it

Submitted on	27.04.2022