Bioinformatics, 33(11), 2017, 1730–1732
doi: 10.1093/bioinformatics/btx036
Advance Access Publication Date: 27 January 2017
Applications Note

OXFORD

Sequence analysis

MetaShot: an accurate workflow for taxon classification of host-associated microbiome from shotgun metagenomic data

B. Fosso[1], M. Santamaria[1], M. D'Antonio[2], D. Lovero[1], G. Corrado[3], E. Vizza[3], N. Passaro[4], A.R. Garbuglia[5], M.R. Capobianchi[5], M. Crescenzi[4], G. Valiente[6] and G. Pesole[1,7,*]

[1]Institute of Biomembranes and Bioenergetics, Consiglio Nazionale delle Ricerche, 70126 Bari, Italy, [2]CINECA, 00185 Rome, Italy, [3]Department of Oncological Surgery, Gynecologic Oncology Unit, "Regina Elena" National Cancer Institute, Rome, Italy, [4]Department of Cell Biology and Neurosciences, Italian National Institute of Health, Rome, Italy, [5]Lazzaro Spallanzani National Institute for Infectious Diseases, Rome 00149, Italy, [6]Algorithms, Bioinformatics, Complexity and Formal Methods Research Group, Technical University of Catalonia, E-08034 Barcelona, Spain and [7]Department of Biosciences, Biotechnology and Biopharmaceutics, University of Bari "A. Moro", Bari, Italy

*To whom correspondence should be addressed.

Associate Editor: John Hancock

Received on June 6, 2016; revised on October 24, 2016; editorial decision on January 18, 2017; accepted on January 19, 2017

Abstract

Summary: Shotgun metagenomics by high-throughput sequencing may allow deep and accurate characterization of host-associated total microbiomes, including bacteria, viruses, protists and fungi. However, the analysis of such sequencing data is still extremely challenging in terms of both overall accuracy and computational efficiency, and current methodologies show substantial variability in misclassification rate and resolution at lower taxonomic ranks or are limited to specific life domains (e.g. only bacteria). We present here MetaShot, a workflow for assessing the total microbiome composition from host-associated shotgun sequence data, and show its overall optimal accuracy performance by analyzing both simulated and real datasets.

Availability and Implementation: https://github.com/bfosso/MetaShot

Contact: graziano.pesole@uniba.it

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Shotgun metagenomics approaches are opening new amazing avenues for better understanding host-microbe interactions and related pathologies. However, the effective and accurate characterization of host-associated microbiomes is still a largely unsolved issue as different methodologies show substantial variability in classification accuracy, precision and computational resources consumption. Current methods for taxonomic binning of metagenomic shotgun reads may be classified as supervised or unsupervised, with the former having a substantial external dependency on reference databases (Santamaria *et al.*, 2012). Unsupervised methods are generally based on specifically taxon-associated compositional features of genome sequences (e.g. oligonucleotide composition, periodic sequence signals, etc.) (Koslicki *et al.*, 2014; Wood and Salzberg, 2014), whereas supervised methods are generally based on similarity data obtained by aligning sampled reads to reference databases. Although the accuracy of supervised methods is strongly dependent on the reliability of reference databases, they usually provide a deeper level of taxonomic classification, up to the species level which is unfeasible for k-mer based techniques which hardly distinguish viral genomes from bacterial and eukaryotic genomes (Bazinet and Cummings, 2012; Soueidan *et al.*, 2015). The microbial components of environmental or clinical samples include viruses and all three life domains,

including Archaea, Eubacteria, and Fungi and Protists among Eukaryotes. Furthermore, in the case of clinical samples data, cleaning from host reads is a crucial step for an appropriate and accurate microbiome assessment. We present here MetaShot, a novel analysis workflow for assessing the microbiome composition from host-associated shotgun sequence data, and show its overall better performance with respect to Kraken (Wood and Salzberg, 2014) and MetaPhlAn2 (Truong *et al.*, 2015), two state-of-the-art comparable tools.

2 Methods

The MetaShot workflow implements a two-step similarity-based approach to attain the best compromise between computational efficiency and assignment accuracy. Indeed, in consideration that the large majority of shotgun reads derive from the host we first carry out a fast similarity-based screening to detect candidate microbial reads, then a fine-grained taxonomic assessment of the much smaller set of putative microbial reads is carried out by using also an iterative taxon refinement procedure (see Supplementary Material for details). A software package implementing the MetaShot pipeline is freely available at https://github.com/bfosso/MetaShot and includes a utility tool for extracting all reads assigned to a specific NCBI taxonomic ID or all those left unassigned.

3 Results

In order to carry out a comparative assessment of MetaShot performance with respect to Kraken (Wood and Salzberg, 2014) and MetaPhlAn2 (Truong *et al.*, 2015), two state of the art tools for analyzing shotgun metagenomics data, we used ART (Huang *et al.*, 2012) to generate an *in silico* designed human microbiota with a composition resembling a typical human sample, containing human, bacterial and viral sequences (see Table 1A and Supplementary Material for a more detailed description).

The simulated dataset also included reads from PhiX phage, which have been shown to contaminate many assembled microbial genomes (Mukherjee *et al.*, 2015) and from human endogenous retroviruses (HERV), which escape detection by most tools designed for analyzing shotgun metagenomics data because they are simply labeled as host reads. Indeed, under specific conditions these viruses can be expressed and may play a role in disease pathogenesis (Agoni *et al.*, 2013; Li *et al.*, 2015). Moreover, in order to compare MetaShot, Kraken and MetaPhlAn2 on a controlled real dataset we analyzed a bacterial and viral mock community (Conceicao-Neto *et al.*, 2015) available in the NCBI-SRA archive (SRR3458569).

The results of the benchmark assessment displayed in Table 1 clearly show that MetaShot outperforms Kraken and MetaPhlAn2 in terms of the overall accuracy of reads assignment for the Prokaryotes and Viruses simulated datasets, at the Family, Genus and Species levels. In addition, MetaShot performs better that Kraken and MetaPhlAn2 also in terms of taxon assignment accuracy at Species and Genus levels at both qualitative (see Supplementary Tables S1–S4) and quantitative levels (See Supplementary Figs S2 and S3).

Finally, in order to test MetaShot on a real dataset we analyzed DNA-seq (528 034 456 100 bp x 2 PE reads) and RNA-Seq (61 318 866 100 bp x 2 PE reads) data from a sample of cervical squamous cell carcinoma of the uterus. While it is known that about 95% of these cancers harbor human papillomavirus (HPV) genomes, the specific serotype involved varies, the most common ones being HPV16, HPV18, and HPV31 (Growdon and Del Carmen, 2008). We previously established by PCR assessment that this test sample contained

Table 1. (A) Benchmark assessment of Kraken (KR) and MetaShot (MS) on a simulated dataset (see the Supplementary Material for details) consisting of 19 582 500 human (94.5%), 986 114 bacterial (4.8%) and 146 886 viral (0.7%) reads. (B) Precision (P), Recall (R), F-measure (F) and Unclassified reads (U) of Kraken (KR), MetaShot (MS) and MetaPhlAn2 (MP) on the same simulated dataset, at the Species level

(A)

	Assigned %[a]			Correctly Assigned %[b]		
	KR	MS	MP[c]	KR	MS	MP
Human (host)	100.00	99.18	0[c]	100.00	99.99	0[c]
Prokaryotes						
Family	57.41	97.91	5.16	96.77	98.37	97.59
Genus	55.01	98.14	4.96	95.92	98.17	98.02
Species	54.17	99.31	4.76	79.52	88.06	90.7
Viruses						
Family	74.78	97.74	49.32	99.16	98.53	98.48
Genus	101.88	97.39	66.85	99.37	99.75	99.30
Species	73.45	97.81	43.86	98.98	96.70	95.46

(B)

	Human (host)			Prokaryotes			Viruses		
	KR	MS	MP	KR	MS	MP	KR	MS	MP
P (%)	99.85	100.00	0	35.67	98.13	98.00	94.95	98.30	80.93
R (%)	100.00	99.97	0	35.16	84.52	87.31	92.77	98.19	79.32
F (%)	99.92	100.00	0	35.36	86.79	90.72	92.82	98.07	79.93
U (%)	0.00	1.04	99.99	55.28	2.44	94.50	4.25	3.94	30.74

[a]The percentage refers to the total number of reads assignable to the specific taxonomic rank.

[b]The percentage refers to the relevant assigned reads.

[c]MetaPhlAn2 assigns just the sequences containing specific taxon markers and does not search for human host sequences.

HPV31. Indeed, HPV31 was detected only by MetaShot in both DNA-Seq and RNA-Seq datasets (25 359 reads over 25 368 total viral reads in DNA-Seq data and 13 684 reads over 14 150 total viral reads in RNA-Seq data) whereas Kraken detected much fewer viral reads (2656 and 1565 in total for DNA-Seq and RNA-Seq data, respectively), notably not including HPV31 (see Supplementary Table S1) which also MetaPhlAn2 was unable to detect.

These results confirm the optimal performance of MetaShot with respect to Kraken and MetaPhlAn2 also in the case of real data analysis.

The MetaShot output consists of: (i) an HTML interactive table reporting for each node in the inferred taxonomy the taxon name, the NCBI taxonomy ID and the number of assigned reads; (ii) a CSV file containing the same information reported in the interactive table; (iii) a Krona graph (Ondov *et al.*, 2011) to graphically inspect the inferred microbiome.

A remarkable unique feature of MetaShot is the possibility to extract all unassigned reads or the set of reads assigned to a specific taxon, defined by the NCBI taxonomy ID. This feature is particularly useful for downstream analyses such as OTU generation, contig assembly for the characterization of unassigned reads, or functional annotation of the reads belonging to a specific species/strain. In addition, this feature may allow for shotgun mapping species-specific DNA-seq reads to their target genome, if available, to prevent the possibility of artifacts, usually associated with a strong positional mapping bias, due to chimeric contamination in GenBank reference sequences (Mukherjee *et al.*, 2015). Moreover,

in the case of shotgun RNA-Seq reads, mapping to their target genome may precisely assess their relevant expression profile.

The price for the overall better accuracy of MetaShot is a lower computational efficiency. MetaShot is about 2 and 3 times slower than Kraken and MetaPhlAn2, respectively, for the complete analysis of the simulated benchmark dataset (see Supplementary Material).

Acknowledgements

Computational resources were provided by ELIXIR-ITA and BioForIU (PONa3_00025).

Funding

This work was supported by MIUR-Italy; CNR Aging and Medicina Personalizzata programs 2012-2014; JPI Project ENPADASI; H2020 projects INMARE, EMBRIC and ELIXIR-EXCELERATE; Spanish Ministry of Economy and Competitiveness and European Regional Development Fund project DPI2015-67082-P (MINECO/FEDER); Grant no. PE - 2011 - 02346905 from the Italian Ministry of Health.

Conflict of Interest: none declared.

References

Agoni,L. *et al.* (2013) Detection of Human Endogenous Retrovirus K (HERV-K) transcripts in human prostate cancer cell lines. *Front. Oncol.*, **3**, 180.

Bazinet,A.L., and Cummings,M.P. (2012) A comparative evaluation of sequence classification programs. *BMC Bioinf.*, **13**, 92.

Conceicao-Neto,N. *et al.* (2015) Modular approach to customise sample preparation procedures for viral metagenomics: a reproducible protocol for virome analysis. *Sci. Rep.*, **5**, 16532.

Growdon,W.B., and Del Carmen,M. (2008) Human papillomavirus-related gynecologic neoplasms: screening and prevention. *Rev. Obstet. Gynecol.*, **1**, 154–161.

Huang,W. *et al.* (2012) ART: a next-generation sequencing read simulator. *Bioinformatics*, **28**, 593–594.

Koslicki,D. *et al.* (2014) WGSQuikr: fast whole-genome shotgun metagenomic classification. *Plos One*, **9**, e91784.

Li,W. *et al.* (2015) Human endogenous retrovirus-K contributes to motor neuron disease. *Sci. Transl. Med.*, **7**, 307ra153.

Mukherjee,S. *et al.* (2015) Large-scale contamination of microbial isolate genomes by Illumina PhiX control. *Stand Genomic Sci.*, **10**, 18.

Ondov,B.D. *et al.* (2011) Interactive metagenomic visualization in a Web browser. *BMC Bioinf.*, **12**, 385.

Santamaria,M. *et al.* (2012) Reference databases for taxonomic assignment in metagenomics. *Brief Bioinform.*, **13**, 682–695.

Soueidan,H. *et al.* (2015) Finding and identifying the viral needle in the metagenomic haystack: trends and challenges. *Front. Microbiol.*, **5**, 739.

Truong,D.T. *et al.* (2015) MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nat. Methods*, **12**, 902–903.

Wood,D.E., and Salzberg,S.L. (2014) Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.*, **15**, R46.