BMC Bioinformatics

SOFTWARE

Open Access

HOME-BIO (sHOtgun MEtagenomic analysis of BIOlogical entities): a specific and comprehensive pipeline for metagenomic shotgun sequencing data analysis

Carlo Ferravante[1,2,3†], Domenico Memoli[1†], Domenico Palumbo[1], Paolo Ciaramella[2], Antonio Di Loria[2], Ylenia D'Agostino[1], Giovanni Nassa[1], Francesca Rizzo[1], Roberta Tarallo[1], Alessandro Weisz[1,4*] and Giorgio Giurato[1*]

From Joint NETTAB/BBCC 2019 Meeting - Network Tools and Applications in Biology (NETTAB) & Bioinformatics and Computational Biology Conference Salerno, Italy. 11-13 November 2019

*Correspondence: aweisz@unisa.it; ggiurato@unisa.it
†Carlo Ferravante and Domenico Memoli have contributed equally to this work
[1] Laboratory of Molecular Medicine and Genomics, Department of Medicine, Surgery and Dentistry 'Scuola Medica Salernitana', University of Salerno, Via S. Allende, 1, 84081 Baronissi, SA, Italy
Full list of author information is available at the end of the article

Abstract

Background: Next-Generation-Sequencing (NGS) enables detection of microorganisms present in biological and other matrices of various origin and nature, allowing not only the identification of known phyla and strains but also the discovery of novel ones. The large amount of metagenomic shotgun data produced by NGS require comprehensive and user-friendly pipelines for data analysis, that speed up the bioinformatics steps, relieving the users from the need to manually perform complex and time-consuming tasks.

Results: We describe here HOME-BIO (sHOtgun MEtagenomic analysis of BIOlogical entities), an exhaustive pipeline for metagenomics data analysis, comprising three independent analytical modules designed for an inclusive analysis of large NGS datasets.

Conclusions: HOME-BIO is a powerful and easy-to-use tool that can be run also by users with limited computational expertise. It allows in-depth analyses by removing low-complexity/ problematic reads, integrating the analytical steps that lead to a comprehensive taxonomy profile of each sample by querying different source databases, and it is customizable according to specific users' needs.

Keywords: Shotgun metagenomics, Next-Generation-Sequencing, Pipeline

Background

Metagenomics is an interdisciplinary research field combining molecular genetics, microbial ecology, and data processing [1]. Nowadays, the advances in high-throughput sequencing technology allow analyzing the total genomic content of clinical or

© The Author(s) 2021. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

environmental samples, detecting a large number of organisms and viruses at once and allowing to correlate the taxonomic composition to the metabolic processes where they are involved [2, 3]. The main sequencing methods applied for the analysis of metagenome are, the so-called, 'marker gene' and 'shotgun' sequencing approaches. These are now widely used respect to the standard methods requiring isolation and cultivation of each single microbial species of interest, which can only provide a limited view, even considering that several microorganisms are not easily cultured in vitro [4]. Marker gene approach (e.g., amplicon sequencing of 16S rRNA gene of bacteria), focuses on a polymorphic region of the microbial genome, that is amplified and sequenced in order to infer at once the composition of the (nearly) entire microbial community present in a given sample [5]. This strategy foresees sequencing of PCR-amplified sequences corresponding to the 16S bacterial rRNA gene which shows some distinctive characteristics that make it suitable for taxonomic profiling and analysis in a single step of a large spectrum of bacteria [6]. In contrast, metagenomic shotgun sequencing (e.g. Whole Metagenome, WMG sequencing) targets all microbial nucleic acids isolated from a sample, followed by deep sequencing of the DNAs drawn randomly from the mixture. The main advantage of this procedure is represented by a large amount of data and information obtained. Contrary to 16S marker gene sequencing, this approach avoids PCR amplification, a major source of potential bias, and provides in-depth knowledge of the microbiome present in the sample, consisting not only of bacterial but also of fungal, protozoal, and viral entities [7, 8]. The presence of several microorganisms and their interaction with each other and the host, play an important role in many physio-pathological processes such as reproduction, immune system activity, cancer, and metabolic disorders [9–12]. Data produced by WMG shotgun sequencing consist of large sets of sequenced reads and represents a challenge for bioinformatics analysis and biological interpretation, particularly in clinical samples. WMG data analysis requires the implementation of multiple ad hoc tools designed, whose installation and configuration, including reference library and computational environment variables, is complex and troublesome in particular for investigators with limited computational expertise. Furthermore, most pipelines available for non-expert users are designed to measure the diversity of microbiota composition from environmental samples (e.g. soil, water) rather than their detection and characterization in biological matrices. Moreover, these pipelines do not provide a global view of sample metagenomic content, since generally they have been designed for specific entities [13–16]. In this context, Kodoja [17] is a workflow specifically developed for plant datasets, identifying only viral sequences from mixed RNAseq data. MetaFlow [18], instead, perform taxonomy profiling using genomic sequences (DNA). It is implemented in C++ and needs a specific input file format (LEMON's LGF): a graph-based representation of the alignments of metagenomic reads in a collection of reference genomes derived from a bacterial domain. Other pipelines such as nf-rnaSeqMetagen [19] are designed for RNAseq data, and therefore of limited use for metagenomics. Furthermore, this pipeline is implemented in Nextflow [20] and all the applications needed to execute the workflow are containerized in Singularity [21]. Nonetheless, it performs quality check step, host filtering, taxonomic profiling of unmapped reads, and, in order to make taxonomic classification faster, it also runs an assembly de novo step with Trinity [22]. Metaphlan2 [23] is another tool for profiling the composition of microbial communities but, in its present version, is unable to query protozoal databases. In addition, other freely available solutions, such as ASaiM [24] or Galaxy [25], are able to perform numerous steps in a metagenomic investigation, incorporating several exhaustive integrated tools. ASaiM, for example, is an Open-Source Galaxy-based framework dedicated to microbiota data analyses, distributed also via Docker and Conda. MGnify [26], instead, provides a platform for the assembly, analysis and archiving of microbiome data, it requires an internet connection and registration. A summary of common and specific features of several freely available metagenomic pipelines is reported in Additional file 1. Therefore, an easy to use, comprehensive and specific analysis workflow for shotgun sequencing from samples with variable content of host genetic material is still unavailable and much required.

Here we describe HOME-BIO (sHOtgun MEtagenomic annotation of BIOlogical entities), a dockerized solution for the analysis of WMG shotgun datasets that aims at addressing and solving the above mentioned limitations and problems. HOME-BIO is a modular and flexible pipeline that allows taxonomic profiling by allowing the operator to choose between two analytical approaches commonly used in WMG shotgun analysis. The first, defined Metagenomic shotgun module, performs taxonomic characterization and abundance estimation based on exact k-mer matches, to achieve high accuracy and fast classification speeds. The Assembly de novo module, instead, provides the assembly of sequences in contigs in order to perform an unbiased analysis of entire genomes sequenced in the sample. Furthermore, the performance of species classification in metagenomics applications can be improved using long reads, assembled from short reads [27]. The two main modules (Metagenomic shotgun and Assembly de novo module) are preceded by a quality control step, which combines quality check of the sequence reads, followed by filtering host and contaminant reads which may interfere with the results. Compared to the other pipelines mentioned above, HOME-BIO shows some common and specific features (Additional file 1). One of the features of HOME-BIO is the additional protein validation step, available in both modules, that makes more robust the taxonomic profiling analysis. Using protein-level classification it is possible to increase the study accuracy [28]. This characteristic makes HOME-BIO exhaustive and accurate, resulting in a flexible and ready-to-use tool in the hands of investigators interested in metagenomics. Indeed, HOME-BIO is platform-independent and does not require time-consuming and tedious installation or dependency issues due to the Docker implementation. Its modularity makes it flexible to users' needs. In this context, users can choose some input options (DNA or RNA, single or paired-end read sequencing protocols) and query various reference databases in order to obtain comprehensive taxonomic profiling of the metagenomic samples. We tested HOME-BIO on a public dataset deposited in the NCBI database under BioProject/BioSamples with accession number SRP040611 [29]. In their study, Mitra et al. collected atherosclerotic tissue samples from a group of 7 patients that underwent elective carotid endarterectomy following repeated transient ischemic attacks, or minor strokes, and asymptomatic atherosclerotic plaques from 5 controls. Outcomes are reported in a graphical that summarize all experiments analyzed and achieve a better comprehension of the biological entities identified in each sample.

Implementation

HOME-BIO is implemented using both tools mainly used in metagenomic analyses and custom python scripts to produce tables and charts for immediate and easier interpretation of the results. The central core of this pipeline is its modularity, being composed of three main blocks: 'Quality Control', 'Metagenomic Shotgun' and 'Assembly de novo' modules (Fig. 1). They can be run all together or separately, according to the user's needs. The "Quality Control" module allows to perform sequence read quality checks and includes FASTQC [30] and MultiQC [31], to perform the quality check and the summary of quality control, respectively, while the adapter trimming and removal of low-quality reads is performed by Cutadapt [32]. If required, HOME-BIO performs a filtering step to remove host and contaminant sequence fragments by mapping each of them on the corresponding genomes. This alignment is performed with bowtie [33], for input reads with length less than 50 bp and with bowtie2 [34] for those longer than 50 bp. The 'Metagenomic Shotgun' module performs taxonomic profiling by classifying unmapped reads with Kraken2 [35]. By default, the confidence score threshold of 0.5 is used to define the quality of taxonomic classification. To extract as much information as possible, bacterial, viral, and protozoal NCBI databases built by Kraken2 are provided. Nonetheless, if users intend to explore only one of the domains described before (or only

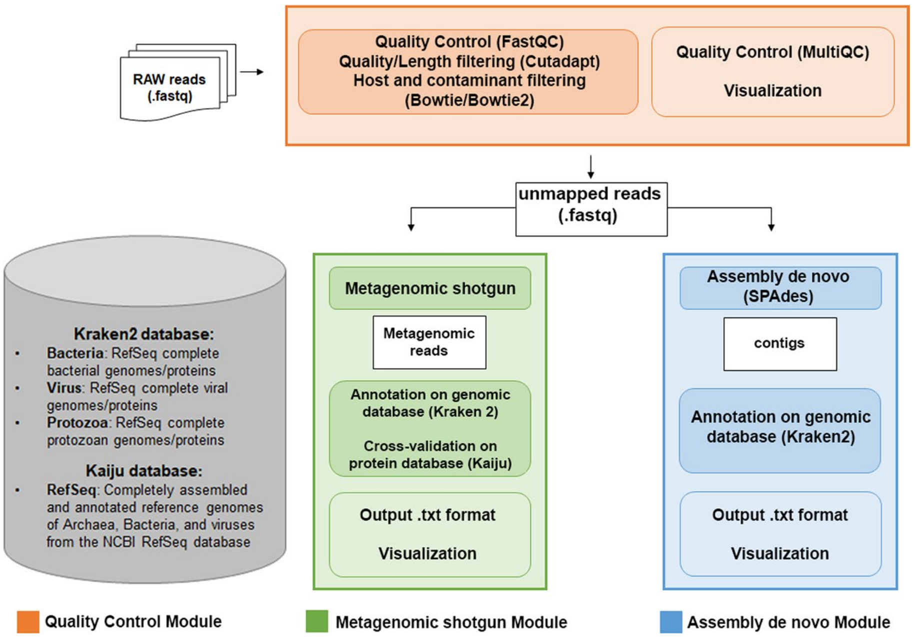
RAW reads (.fastq)
Quality Control (FastQC)
Quality/Length filtering (Cutadapt)
Host and contaminant filtering (Bowtie/Bowtie2)
Quality Control (MultiQC)
Visualization
unmapped reads (.fastq)
Kraken2 database:
Bacteria: RefSeq complete bacterial genomes/proteins
Virus: RefSeq complete viral genomes/proteins
Protozoa: RefSeq complete protozoan genomes/proteins
Kaiju database:
RefSeq: Completely assembled and annotated reference genomes of Archaea, Bacteria, and viruses from the NCBI RefSeq database
Metagenomic shotgun
Metagenomic reads
Annotation on genomic database (Kraken 2)
Cross-validation on protein database (Kaiju)
Output .txt format
Visualization
Assembly de novo (SPAdes)
contigs
Annotation on genomic database (Kraken2)
Output .txt format
Visualization
🟧 Quality Control Module
🟩 Metagenomic shotgun Module
🟦 Assembly de novo Module

Fig. 1 HOME-BIO workflow. The pipeline accepts NGS data as input and then proceeds to perform the analysis according to the different steps comprised in the three modules. The Quality Control module (in orange) performs QCs of the reads, low-complexity reads removal, and host and contaminant sequencing reads filtering (indicated in the dark orange rectangle), generating as output quality reports and graphical charts (indicated in the light orange rectangle). The sequence reads maintained after this step are processed using the Metagenomic shotgun module (in green) and/or the Assembly de novo module (in blue). In the first case, the reads are processed by Kraken2 and Kaiju (indicated in the dark green rectangle), producing as output text tables and charts. The Assembly de novo module processes the sequences by assembling with SPAdes and annotating with Kraken2 (dark blue rectangle), generating as output text tables and charts. For both modules, a custom database (indicates in the grey box) is used

viral entities), it is possible to set custom options to query only one (or more) databases and obtain the taxonomic information desired. Compared to other taxonomic profiling pipelines freely available, such as MicroPro [36], METAwrap [37], and Sunbeam [38], HOME-BIO performs an additional protein-validation step for non-eukaryotic entities. This implementation makes taxonomic classification outcomes more robust by using protein-level classification with Kaiju [28], thereby increasing the reliability and sensitivity of the analysis. Kaiju carries out a comparison to a reference database containing microbial and viral protein sequences. Unmapped metagenomic input reads are thus translated into amino acid sequences and then searched in the database using a modified backward search on a memory-efficient implementation of the Burrows-Wheeler transform, which finds maximum exact matches. As default, we set the evaluation in run mode *greedy*, with an e-value cutoff of 0.001. All the entities classified with Kaiju are then processed with Krona [39], to obtain a comprehensive graphic visualization. As outputs, the 'Metagenomic shotgun' module generates a table containing the Kranken2 taxonomy profile and related Kaiju protein-validation information. A given taxon is considered protein-validated when both tools classify and assign reads to it. In addition, HOME-BIO generates output pie-charts with the top 15 represented species, with an estimation of the relative abundance of each of them. In the 'Assembly de novo' module analysis, HOME-BIO uses SPAdes [40] in metagenomic mode (option *-meta*). It takes in input unmapped reads resulting from the 'Quality Control' module, generating sequence contigs that are classified with Kraken2, and protein-validated with Kaiju as described before. As mentioned above, HOME-BIO takes advantage of several reference databases. This allows users to investigate in depth the content of their biological samples. We linked NCBI taxonomy information from complete RefSeq bacterial, viral, and protozoal genomes/proteins. If it ran in end-to-end mode, the pipeline provides comprehensive profiling of specimens by querying the above-mentioned databases. This aspect makes HOME-BIO a powerful tool in the hands of users dealing with metagenomic data.

To make its installation easier, HOME-BIO is inserted in a Docker [41] image, with installed Ubuntu 18.04.4 LTS and Anaconda 3 (V. 02/2020) (https://anaconda.com). This allows that all the required tools are always compatible with each other, automatically updated, and handled by Anaconda itself. All the parameters are set using a configuration file that is given as input file when the pipeline is launched.

Results

HOME-BIO processes NGS data to characterize the composition of host-associated microbial communities. The pipeline is implemented to be used by biologists with limited computational experience who intend to explore the pathogens' content in their samples.

To test the performance of HOME-BIO, we considered the public dataset of Mitra et al. [29], comprising 12 samples: 7 from symptomatic atherosclerotic plaques and 5 matched controls from asymptomatic atherosclerotic plaques. The analysis was performed running all three modules with default parameters and using the databases provided with the pipeline [42], comprising complete RefSeq bacterial, viral, and protozoa genome sequences and a subset of NCBI BLAST non-redundant database containing all proteins belonging to Archaea, Bacteria, and Viruses. To extract more confident sequences to be

Table 1 Total number and percentage of reads assigned to Bacteria, Viruses, and Protozoa for each analyzed sample

Sample-name	Raw reads	Non-hg19%	HOME-BIO shotgun Bacteria reads classified (%)	HOME-BIO shotgun Virus reads classified (%)	HOME-BIO shotgun Protozoa reads classified (%)
SRR1205226	62.241.508	10.36	4617 (0.07)	331 (0.01%)	177713 (2.76)
SRR1205227	83.273.796	8.79	259578 (3.54)	595 (0.01%)	237811 (3.25)
SRR1205228	57.336.562	16.43	5955 (0.06)	677 (0.01%)	218329 (2.32)
SRR1205230	75.597.642	11.50	52529 (0.6)	2496 (0.03)	262034 (3.01)
SRR1205231	76.089.590	10.60	5316 (0.07)	697 (0.01)	217939 (2.7)
SRR1205232	70.667.046	20.82	7420 (0.05)	1152 (0.01)	278959 (1.9)
SRR1206003	76.651.484	12.06	139334 (1.51)	736 (0.01)	264054 (2.86)
SRR1206005	77.326.222	15.57	8235 (0.07)	720 (0.01%)	264974 (2.2)
SRR1206007	50.795.748	22.63	3713 (0.03)	684 (0.01)	174687 (1.52)
SRR1206009	49.963.912	21.03	4393 (0.04)	491 (0.00)	202508 (1.93)
SRR1206011	41.425.332	28.19	3434 (0.03)	732 (0.01%)	196978 (1.69)
SRR1206012	52.047.446	25.38	5146 (0.04)	748 (0.01)	230780 (1.75)

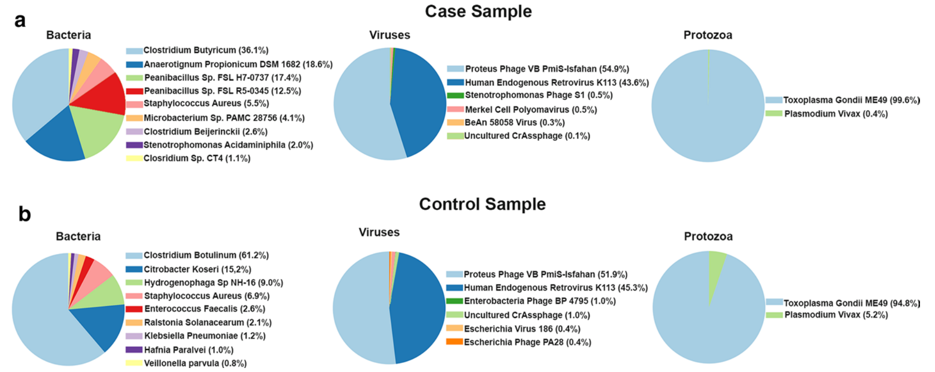

Fig. 2 Pie Charts regarding Bacteria, Viruses, and Protozoa detected in the samples. Pie-charts summarizing the percentage of Bacteria, Viruses, and Protozoa detected in samples SRR1206003 (Case) and SRR1202227 (Control) obtained by data analysis with the HOME-BIO pipeline

assigned to biological entities, the reads were mapped against the genome of the host (human genome assembly hg19). After running the Quality Control module, the number of unmapped reads, usable in the next computational step, ranged from 6.449.408 to 14.710.470, with an average value of 7.991.03720 for Controls and 11.841.55800 for Cases samples, representing 26.17% and 41.63% of the total reads in Control and Case, respectively. Subsequently, for each sample, the reads were analyzed with the Shotgun Metagenomic module, in order to obtain the taxonomy profile. HOME-BIO queried bacterial, viral, and protozoal databases. The Quality Control and Metagenomic results are summarized in Table 1 and Fig. 2 and are reported in Additional file 2: Tables S1–S24, Additional file 3: Tables S1–S24, and Additional file 4: Tables S1–S24.

Running the Assembly de novo module, 567.37957 contigs for Cases samples and 238.7888 contigs for Control samples were obtained. The results are summarized in the Additional file 5: Tables S1–S24, Additional file 6: Tables S1–S24, and Additional file 7: Tables S1–S24. We identified an average of 5.5752 and 12.9563 contigs classified as bacteria in Control and Cases samples, respectively. Among these, 4554 in Control and 816 in Cases were contigs classified as viruses, while 12.2198 contigs in Control and 18.4238 in Cases matched in the protozoal database. These results obtained on bacteria from the analysis of the 12 samples are in agreement with the ones described by Mitra et al. [29]. In addition, HOME-BIO allowed to identify also several other viruses and protozoa that could be related to the disease and are reported in Additional file 3: Tables S1–S24, Additional file 4: Tables S1–S24 comprised the reports from the Metagenomics module, and Additional file 6: Tables S1–S24 and Additional file 7: Tables S1–S24 those from the Assembly de novo module.

For a further test with a different species, HOME-BIO was used to analyze a small dataset derived in our lab by sequencing four samples from canine heart biopsies, two consisting of formalin-fixed, paraffin-embedded (FFPE) samples and two of snap-frozen biopsies (raw data available on Array Express with Accession n.er E-MTAB-9189 and E-MTAB-9191). Results confirmed the good performance of the tool in all these datasets, leading as expected to the identification of several microorganisms independent from the nature and pre-treatment of the tissue samples. The results from all these samples are included in Additional file 8: Tables S1–S2. A tutorial with figures and examples of usage is included in Additional file 9.

Conclusions

Here, we describe HOME-BIO, a user-friendly pipeline based on a dockerized solution, designed for analyzing shotgun metagenomic data avoiding time-consuming and error-prone installation and configuration steps. This modular pipeline provides a quality control step and the two main analysis approaches commonly used in metagenomic studies. By querying bacterial, viral, protozoal, and protein databases, HOME-BIO generates also an exhaustive taxonomic profiling of the biological entities in specimens.

Availability and requirements

Project name: HOME-BIO
Project home page: https://github.com/carlferr/HOME-BIO
Operating system(s): Any
Programming language: Python
Other requirements: Docker
License: GNU GPL
Any restrictions to use by non-academics: none

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-021-04004-y.

Additional file 1: Table S1. Characteristics of Shotgun Pipelines.

Additional file 2: Table S1–S24. HOME-BIO reports for bacteria identified with the metagenomic module in the twelve samples analyzed.

Additional file 3: Table S1–S24. HOME-BIO reports for viruses identified with the metagenomic module in the twelve samples analyzed.

Additional file 4: Table S1–S24. HOME-BIO reports for protozoa identified with the metagenomic module in the twelve samples analyzed.

Additional file 5: Table S1–S24. HOME-BIO reports for bacteria identified with the Assembly de novo module in the twelve samples analyzed.

Additional file 6: Table S1–S24. HOME-BIO reports for viruses identified with the Assembly de novo module in the twelve samples analyzed.

Additional file 7: Table S1–S24. HOME-BIO reports for protozoa identified with the Assembly de novo module in the twelve samples analyzed.

Additional file 8: Table S1–S2. HOME-BIO reports for bacteria, viruses, and protozoa identified Can_Fam samples.

Additional file 9. Tutorial and use examples of HOME-BIO.

Abbreviations

NGS: Next-Generation Sequencing; WMG: Whole Metagenome.

Acknowledgements

C.F. is a Ph.D. student of the Research Doctorates in Veterinary Sciences of the University of Napoli 'Federico II'.

About this supplement

This article has been published as part of BMC Bioinformatics Volume 22 Supplement 7 2021: Proceedings from the Joint NETTAB/BBCC 2019 Conference (part 2). The full contents of the supplement are available at https://bmcbioinformatics.biomedcentral.com/articles/supplements/volume-22-supplement-7.

Authors' contributions

AW, GG, ADL, and PC conceived the study. GG, CF, DM, and DP designed and implemented the pipeline. CF and DM performed tests and data analyses. YDA performed data curation and prepared the figures. AW and GG supervised the work. AW, FR, GG, GN and RT contributed to write and edit the manuscript. All authors read and approved the final manuscript.

Funding

Work was supported by: Regione Campania, Projects GENOMAeSALUTE (POR Campania FESR 2014/2020, azione 1.5; CUP: B41C17000080007) and RarePlatNet ('La Campania lotta contro il cancro', CUP: B63D18000380007), Genomix4life Srl and Ministero dell'Università e della Ricerca, programma PON R&I 2014-2020 'Dottorati innovativi con caratterizzazione industriale' XXXIV Ciclo (Fellowship DOT1318705, CUP: E66C18000940007 to C.F.). The funding bodies had no role in the design of the study and collection, analysis, and interpretation of data, or in writing the manuscript. Publication costs are funded by University of Salerno (FARB 2019).

Availability of data and materials

The data used in this study are available on NCBI with accession number SRP040611 and on ArrayExpress with accession number E-MTAB-9189 and E-MTAB-9191. Pre-indexed Kaiju and Kraken2 databases are available on Zenodo [42] (https://doi.org/10.5281/zenodo.4055180) and a test dataset is available on Zenodo [43] (https://doi.org/10.5281/zenodo.4061297).

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1] Laboratory of Molecular Medicine and Genomics, Department of Medicine, Surgery and Dentistry 'Scuola Medica Salernitana', University of Salerno, Via S. Allende, 1, 84081 Baronissi, SA, Italy. [2] Department of Veterinary Medicine and Animal Production, University of Naples Federico II, Via Delpino 1, 80137 Naples, Italy. [3] Genomix4Life, via S. Allende 43/L, 84081 Baronissi, SA, Italy. [4] CRGS - Genome Research Center for Health, University of Salerno Campus of Medicine, 84081 Baronissi, SA, Italy.

Received: 4 February 2021 Accepted: 8 February 2021
Published: 5 July 2021

References

1. Sudarikov K, Tyakht A, Alexeev D. Methods for the metagenomic data visualization and analysis. Vol. 24, Current issues in molecular biology. Curr Issues Mol Biol; 2017. p. 37–58.
2. Mendes LW, Braga LPP, Navarrete AA, de Souza DG, Silva GGZ, Tsai SM. Using metagenomics to connect microbial community biodiversity and functions. Curr Issues Mol Biol. 2017;24:103–18.

3. Quince C, Walker AW, Simpson JT, Loman NJ, Segata N. Shotgun metagenomics, from sampling to analysis. Vol. 35, Nature Biotechnology. Nature Publishing Group, Berlin; 2017. p. 833–44.
4. Lagier JC, Dubourg G, Million M, Cadoret F, Bilen M, Fenollar F, et al. Culturing the human microbiota and culturomics. Vol. 16, Nature Reviews Microbiology. Nature Publishing Group, Berlin; 2018. p. 540–50.
5. Turnbaugh PJ, Ley RE, Hamady M, Fraser-Liggett CM, Knight R, Gordon JI. The human microbiome project, vol. 449. Berlin: Nature Publishing Group; 2007. p. 804–10.
6. Amit Roy SR. Molecular markers in phylogenetic studies—a review. J Phylogenetics Evol Biol. 2014;02:02.
7. Marotz CA, Sanders JG, Zuniga C, Zaramela LS, Knight R, Zengler K. Improving saliva shotgun metagenomics by chemical host DNA depletion. Microbiome. 2018;6(1):42.
8. Conrads G, Abdelbary MMH. Challenges of next-generation sequencing targeting anaerobes, vol. 58. Cambridge: Academic Press; 2019. p. 47–52.
9. Benson AK, Kelly SA, Legge R, Ma F, Low SJ, Kim J, et al. Individuality in gut microbiota composition is a complex polygenic trait shaped by multiple environmental and host genetic factors. Proc Natl Acad Sci USA. 2010;107(44):18933–8.
10. Atreya CE, Turnbaugh PJ. Probing the tumor micro(b)environment. Science (80-). 2020;368(6494):938–9.
11. Nejman D, Livyatan I, Fuks G, Gavert N, Zwang Y, Geller LT, et al. The human tumor microbiome is composed of tumor type–specific intracellular bacteria. Science (80-). 2020;368(6494):973–80.
12. Zitvogel L, Ma Y, Raoult D, Kroemer G, Gajewski TF. The microbiome in cancer immunotherapy: diagnostic tools and therapeutic strategies, vol. 359. Washington: American Association for the Advancement of Science; 2018. p. 1366–70.
13. Ji Y, Huotari T, Roslin T, Schmidt NM, Wang J, Yu DW, et al. SPIKEPIPE: a metagenomic pipeline for the accurate quantification of eukaryotic species occurrences and intraspecific abundance change using DNA barcodes or mitogenomes. Mol Ecol Resour. 2020;20(1):256–67.
14. Milani C, Casey E, Lugli GA, Moore R, Kaczorowska J, Feehily C, et al. Tracing mother-infant transmission of bacteriophages by means of a novel analytical tool for shotgun metagenomic datasets: METAnnotatorX. Microbiome. 2018;6(1):145.
15. Piper AM, Batovska J, Cogan NOI, Weiss J, Cunningham JP, Rodoni BC, et al. Prospects and challenges of implementing DNA metabarcoding for high-throughput insect surveillance. Gigascience. 2019;8(8):giz092:1–22
16. Rampelli S, Soverini M, Turroni S, Quercia S, Biagi E, Brigidi P, et al. ViromeScan: a new tool for metagenomic viral community profiling. BMC Genomics. 2016;17(165):1–9.
17. Baizan-Edge A, Cock P, MacFarlane S, McGavin W, Torrance L, Jones S. Kodoja: A workflow for virus detection in plants using k-mer analysis of RNA-sequencing data. J Gen Virol. 2019;100(3):533–42. https://doi.org/10.1099/jgv.0.001210.
18. Sobih A, Tomescu AI, Mäkinen V. Metaflow: Metagenomic profiling based on whole-genome coverage analysis with min-cost flows. Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics). 2016;9649:111–21.
19. Mpangase PT, Frost J, Ramsay M, Hazelhurst S. nf-rnaSeqMetagen: a nextflow metagenomics pipeline for identifying and characterizing microbial sequences from RNA-seq data. Med Microecol. 2020;4:100011.
20. Di Tommaso P, Chatzou M, Floden EW, Barja PP, Palumbo E, Notredame C. Nextflow enables reproducible computational workflows. Nat Biotechnol. 2017;35(4):316–9.
21. Kurtzer GM, Sochat V, Bauer MW. Singularity: Scientific containers for mobility of compute. PLoS ONE. 2017;12(5):e0177459. https://doi.org/10.1371/journal.pone.0177459.
22. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol. 2011;29(7):644–52.
23. Segata N, Waldron L, Ballarini A, Narasimhan V, Jousson O, Huttenhower C. Metagenomic microbial community profiling using unique clade-specific marker genes. Nat Methods. 2012;9(8):811–4.
24. Batut B, Gravouil K, Defois C, Hiltemann S, Brugère J-F, Peyretaillade E, et al. ASaiM: a Galaxy-based framework to analyze microbiota data. Gigascience. 2018;7(6):giy057. https://doi.org/10.1093/gigascience/giy057/5001424.
25. Afgan E, Baker D, Batut B, van den Beek M, Bouvier D, Čech M, et al. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. Nucleic Acids Res. 2018;46(W1):W537–44.
26. Mitchell AL, Almeida A, Beracochea M, Boland M, Burgin J, Cochrane G, et al. MGnify: the microbiome analysis resource in 2020. Nucleic Acids Res. 2019. https://doi.org/10.1093/nar/gkz1035/5614179.
27. Tran Q, Phan V. Assembling reads improves taxonomic classification of species. Genes (Basel). 2020;11(8):1–11.
28. Menzel P, Ng KL, Krogh A. Fast and sensitive taxonomic classification for metagenomics with Kaiju. Nat Commun. 2016;7(1):1–9.
29. Mitra S, Drautz-Moses DI, Alhede M, Maw MT, Liu Y, Purbojati RW, et al. In silico analyses of metagenomes from human atherosclerotic plaque samples. Microbiome. 2015;3(1):38.
30. Andrews S. FastQC: A Quality Control Tool for High Throughput Sequence Data. 2010. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
31. Ewels P, Magnusson M, Lundin S, Käller M. MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics. 2016;32(19):3047–8.
32. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet J. 2011;17(1):10.
33. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 2009;10(3):R25.
34. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9(4):357–9.
35. Wood DE, Lu J, Langmead B. Improved metagenomic analysis with Kraken 2. Genome Biol. 2019;20(1):257.
36. Zhu Z, Ren J, Michail S, Sun F. MicroPro: Using metagenomic unmapped reads to provide insights into human microbiota and disease associations. Genome Biol. 2019;20(1):154.
37. Uritskiy GV, Diruggiero J, Taylor J. MetaWRAP—a flexible pipeline for genome-resolved metagenomic data analysis. Microbiome. 2018;6(1):158.

38. Clarke EL, Taylor LJ, Zhao C, Connell A, Lee JJ, Fett B, et al. Sunbeam: an extensible pipeline for analyzing metagenomic sequencing experiments. Microbiome. 2019;7(1):46.
39. Ondov BD, Bergman NH, Phillippy AM. Interactive metagenomic visualization in a Web browser. BMC Bioinform. 2011;12(385):1–9.
40. Nurk S, Bankevich A, Antipov D, Gurevich A, Korobeynikov A, Lapidus A, et al. Assembling genomes and mini-metagenomes from highly chimeric reads. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Springer, Berlin; 2013. p. 158–70.
41. Merkel D. Docker: lightweight Linux containers for consistent development and deployment. Linux J. 2014;2014:239.
42. Domenico P. Kraken2 & Kaiju pre-indexed databases. 2020 Sep 28 [cited 2020 Nov 17]; https://zenodo.org/record/4055180
43. Domenico P. Test Dataset for HOME-BIO. 2020 Oct 1 [cited 2020 Nov 17]; https://zenodo.org/record/4061297

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

BMC