Microb Ecol (2016) 72:443–447
DOI 10.1007/s00248-016-0785-x

METHODS

BMPOS: a Flexible and User-Friendly Tool Sets for Microbiome Studies

Victor S. Pylro[1] **· Daniel K. Morais**[1] **· Francislon S. de Oliveira**[1] **· Fausto G. dos Santos**[1] **·**
Leandro N. Lemos[2] **· Guilherme Oliveira**[3] **· Luiz F. W. Roesch**[4]

Received: 29 March 2016 / Accepted: 9 May 2016 / Published online: 24 May 2016

Abstract Recent advances in science and technology are leading to a revision and re-orientation of methodologies, addressing old and current issues under a new perspective. Advances in next generation sequencing (NGS) are allowing comparative analysis of the abundance and diversity of whole microbial communities, generating a large amount of data and findings at a systems level. The current limitation for biologists has been the increasing demand for computational power and training required for processing of NGS data. Here, we describe the deployment of the Brazilian Microbiome Project Operating System (BMPOS), a flexible and user-friendly Linux distribution dedicated to microbiome studies. The Brazilian Microbiome Project (BMP) has developed data analyses pipelines for metagenomic studies (phylogenetic marker genes), conducted using the two main high-throughput sequencing platforms (Ion Torrent and Illumina MiSeq). The BMPOS is freely available and possesses the entire requirement of bioinformatics packages and databases to perform all the pipelines suggested by the BMP team. The BMPOS may be used as a bootable live USB stick or installed in any computer with at least 1 GHz CPU and 512 MB RAM, independent of the operating system previously installed. The BMPOS has proved to be effective for sequences processing, sequences clustering, alignment, taxonomic annotation, statistical analysis, and plotting of metagenomic data. The BMPOS has been used during several metagenomic analyses courses, being valuable as a tool for training, and an excellent starting point to anyone interested in performing metagenomic studies. The BMPOS and its documentation are available at http://www.brmicrobiome.org.

Keywords Microbiome · BMP · Operating system · Linux · Bioinformatics

Introduction

The characterization of whole microbial communities via molecular methods has profoundly changed the way we conduct microbial ecology studies [1, 2]. The development of high-throughput sequencing technologies are allowing comparative analyses of diversity, abundance, and important ecosystem functional genes of whole microbial communities at far greater depths than ever before [3, 4]. However, the increasing amount of genomic information being produced is currently overcoming the analytical capacity of many laboratories. Creative solutions are required to empower researchers who have no access to a team of bioinformaticians or high end computational resources [5].

Useful solutions to these problems can be applied in two ways: firstly, computational tasks can be specified in terms of human-readable equations that are independent of the programming platform, and secondly, improved performance in execution of clearly defined tasks can be achieved [6]. However, understanding the code can impose a steep learning curve: consuming precious time between biologists and programmers in tasks of installation, configuration and maintenance of software and dependencies, necessary to create and execute pipelines. Moreover, the development and

✉ Victor S. Pylro
victor.pylro@brmicrobiome.org

[1] Genomics and Computational Biology Group, René Rachou Research Center (CPqRR-FIOCRUZ), Belo Horizonte, MG 30190-002, Brazil

[2] Programa de Pós-graduação Interunidades em Bioinformática, Instituto de Química - IQ, Universidade de São Paulo - USP, São Paulo, SP, Brazil

[3] Vale Institute of Technology – ITV, Belém, PA, Brazil

[4] Universidade Federal do Pampa - UNIPAMPA, São Gabriel, RS, Brazil

operation of tailored tools require dedicated staff, high expertise, and even the acquisition of powerful computational infrastructure. These requirements become a barrier for several research groups hindering science development especially in resource limited regions.

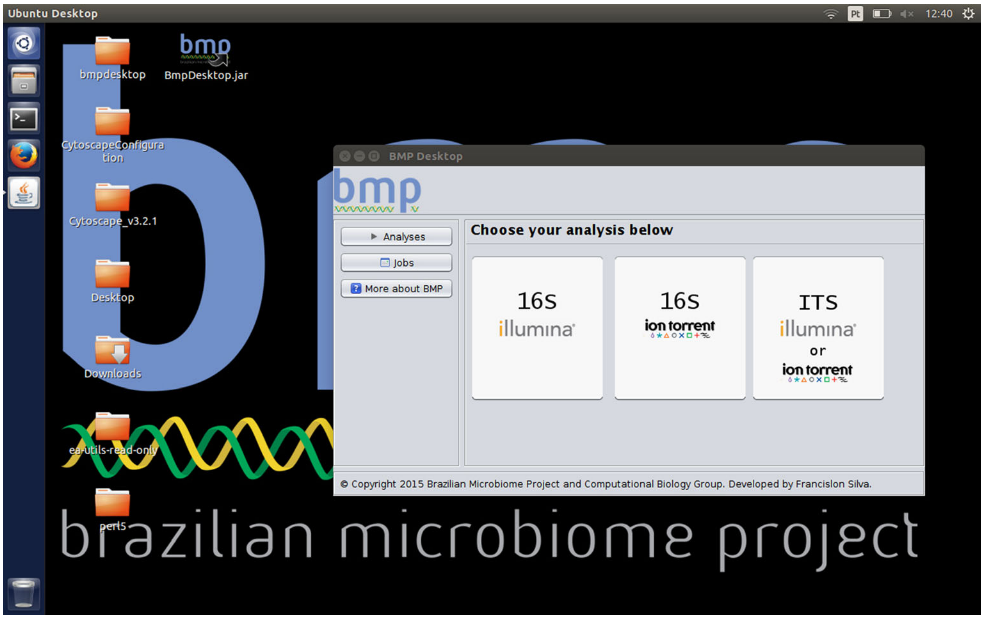
To facilitate the dissemination and use of bioinformatics tools, several Linux distributions such as BioLinux [7], Scibuntu [8], PhyLIS [9], LXtoo [10], and Mypro [11] have been created and represent an interesting and time saving option. All the cited OS built for bioinformatics fall into two categories, being either broad such as BioLinux, Scibuntu, and LXtoo or specific like PhyLIS and Mypro (aimed at phylogenetic analysis and prokaryotic genome assembly/annotation, respectively). None of the available OS addresses the need for bioinformatics tools to support the automated and user-friendly microbiome data analyses.

The Brazilian Microbiome Project (BMP) [12] addresses the importance of studying the huge biological diversity stored in Brazil in a resource limited setting for the analysis of microbiome data. One of the main challenges of the BMP resides on testing and/or creating bioinformatics pipelines to help researchers in handling next generation sequencing (NGS) data [13]. Metagenomic analyses have been described as one of the least reproducible NGS applications, mainly due the lack of integrated and standardized solutions for performing these kinds of studies [14]. Here, we describe the Brazilian Microbiome Project Operating System (BMPOS), a flexible and user-friendly Linux distribution available to help researchers handle the most frequently used bioinformatics packages dedicated to the study of microbial ecology. The BMPOS is valuable as a tool for end-to-end analysis, training, and an excellent starting point for anyone interested in performing microbiome studies based on NGS data.

Implementation

The BMPOS (Fig. 1) is based on the free GNU/Linux distribution Ubuntu 14.04 LTS. It allows users to reproduce all bioinformatics pipelines, as recommended by the BMP advisory board, available at [15]. The BMPOS contains the most widely used packages among microbial ecologists to analyze next generation sequencing (NGS) data. The packages are installed, configured, pre-compiled, and already defined at the system's path. An updated list of packages is maintained at [16] and is currently composed of software for sequence filtering and trimming, sequence clustering, sequence alignment, phylogenetic tree reconstruction, statistical analysis, data visualization, and database searching, besides of all BMP scripts created to make data compatible among different packages (Table 1).

Among all packages applied at the BMP pipelines, only USEARCH [18] has a restriction of use, being freely available only on its 32-bit version, i.e., not capable of handling files bigger than 4 Gb. In that case, it is necessary to acquire a license for the 64-bit version. The USEARCH package possess tools capable of database search, nominally hundreds of times faster than BLAST [36], possess algorithms for processing NGS reads like quality filtering, chimera detection, and dereplication, and being implemented in BMP pipelines because of their accuracy and speed. As an alternative to USEARCH, the BMPOS also provides VSEARCH [17] which supports most of the USEARCH functions, but as an open and free 64-bit multithreaded tool. Another open source bioinformatics package used in the initial steps of the BMP pipeline is QIIME (Quantitative Insights Into Microbial Ecology) [37]. QIIME has the advantage of containing a comprehensive suite of functions and procedures. It is easy to use, implement, and



Choose your analysis below

© Copyright 2015 Brazilian Microbiome Project and Computational Biology Group. Developed by Francislon Silva.

Fig. 1 The screenshot of the BMPOS, highlighting the BMP desktop application

Table 1 Packages, scripts, and databases available in the BMPOS

Package	Application	References
QIIME	Read processing, read alignment, OTU clustering, data analysis, and plotting	[17]
BMP scripts	File formatting and conversion	This study
USEARCH	Read processing, read alignment, OTU clustering, and data analysis	[18]
UPARSE scripts	File formatting and conversion	[18]
VSEARCH	Read processing, read alignment, and OTU clustering data analysis	[17]
R	Data analysis, processing, and clustering	[19]
ITSx	Sequences detection and filtering	[20]
HMMER	Database construction, database searching and sequence filtering	[21]
FastX	Sequences processing	[22]
BIOM	File format representing contingency tables and metadata	[23]
Oligotyping	Discriminates closely related taxa using the Shannong entropy method	[24]
Stamp	Data analyzing and plotting	[25]
Gephi	Interactive data visualization and exploration	[26]
SparCC	Data analysis using network inference tools	[27]
CytoScape	Interactive network visualization tool	[28]
SPADES	Genome assembler	[29]
QUAST	Genome assemblies evaluator	[30]
MaxBin	Tool for binning metagenomic sequences	[31]
Greengenes	Database of 16S rRNA gene sequence alignment	[32]
RDP	Aligned database of Bacterial and Archaeal 16S rRNA sequences and Fungal 28S rRNA sequences	[33]
Unite	Reference database for molecular identification of Fungi using the ITS regions	[34]
Gold	Chimera reference database from the Broad Microbiome Utilities	[35]
FastQC	A quality control tool for high-throughput sequence data	[39]

combine with other packages, and it has an extensive documentation [38]. Packages like ITSx [20] and HMMER [21] are used in combination to extract non-internal transcribed spacer (ITS) sequences and delivering a better taxonomic assignment for Fungi. For better handling and storage of contingency tables containing metadata, the BMPOS uses the BIOM package [23]. Moreover, our operating system also contains all needed databases for chimera detection and taxonomic assignment of Fungi, Bacteria and Archaea (Table 1).

The BMPOS may be used directly as a bootable live USB stick plugged in any computer with at least 1 GHz CPU and 512 MB RAM, without the need for package installation or any previous configuration. It can also be installed in the user's machine, independently of the operating system. This is the fastest way to spread bioinformatics packages among groups of collaborators.

Results and Discussion

The BMPOS is very effective for classes or courses and for research groups with limited human and computational resources. The Brazilian Microbiome Project has conducted successful courses, teaching nearly 100 students using the USB stick strategy. The courses included students that have never had any contact with command line and successfully concluded all analysis resulting in increased confidence and motivation. The use of the live USB strategy allowed the students to save the results obtained from the pipeline in the USB stick or their own computer, permitting the users to analyze the final data matrix in the software of their preference. Moreover, as they were given the USB stick, the students are able to further improve their own skills.

The execution of the BMP pipeline developed for metagenomic analyses using the 16S ribosomal RNA (rRNA) gene for Bacteria and Archaea, and ITS region for Fungi is a very simple approach for teaching how to conduct and understand each step of the pipeline that can itself be fully automated. The strategy adopted to automate these pipelines at the BMPOS relies on the so called BMP desktop application, a bash script coupled to a Java™ (SE Runtime Environment-build 1.8.0) interface that generates a better user-friendly experience, besides saving the effort of retyping that particular sequence of commands. For now, the java application only runs a default workflow, but in future versions, new improvements will allow users to adjust parameters, making it more flexible. On a personal computer with an Intel® Core™2 Duo CPU P8600 2.40 GHz (x2) and 4 GB of RAM memory, the BMPOS takes

~4 min to run the BMP recommended pipeline on a dataset of 166,931 16S rDNA paired-end Illumina reads (151 bp–120 Mb). The same computer takes ~5 min to run the BMP recommended pipeline on a dataset of 120,171 16S rDNA single-end Ion Torrent reads (300 bp–60Mb), and ~9 min. to run the BMP recommended pipeline on a dataset of 166,931 ITS single-end Illumina reads (251 bp–93.4 Mb). These datasets are distributed alongside the BMPOS, in a folder located at "usr/bmp/data_example," which allow users to run a preliminary test and evaluate their own machine. All the BMP recommended pipelines are available in the section "Standards and Protocols" of the BMP website (http://brmicrobiome.org).

Conclusions

The BMPOS presents as a useful and user-friendly starting point to anyone interested in metagenomic analyses of microbial communities. This strategy proved itself an effective way of settling an environment for bioinformatics training and routine analyses. We are open to suggestions regarding bug fixes, the addition of new packages or updates of the currently installed software and packages. Updates on the BMPOS will be made yearly or as soon as new analysis pipelines are developed. The BMPOS is available to download at http://brmicrobiome.org.

Acknowledgments We would like to thank Alysson Silvano and Thermo Fisher/Brazil, for providing the USB flash drives used to freely distribute the BMPOS (under request or during training). CNPq, FAPEMIG (PACCSS program), and CAPES (PROEX program) funded this work. LN Lemos receive research fellowship from the FAPESP (process number 13/05325-5). We also thank Melline Fontes Noronha for helping with editing some BMP scripts and Dr. Marc Redmile-Gordon (Centre for Sustainable Soils and Grassland Systems, Rothamsted Research, UK) for critical comments and review of the written English in the manuscript.

Authors Contributions VSP, FSO, LNL, and LFWR conceived the BMPOS and wrote the manuscript with contributions from DKM, FGS, and GO. FSO and FGS were responsible to install, implement, and test all packages and scripts in the BMPOS. VSP, FSO, and LNL wrote/edited the BMP scripts. FSO implemented the BMP desktop application (Java). All authors read and approved the final manuscript.

Compliance with Ethical Standards

Competing Interests The authors declare that they have no competing interests.

References

1. Franzosa EA, Hsu T, Sirota-Madi A, Shafquat A, Abu-Ali G, Morgan XC, Huttenhower C (2015) Sequencing and beyond: integrating molecular 'omics' for microbial community profiling. Nat Rev Microbiol 13(6):360–372
2. Siqueira JF Jr, Fouad AF, Rôças (2012) Pyrosequencing as a tool for better understanding of human microbiomes. J Oral Microbiol 4:10. doi:10.3402/jom.v4i0.10743
3. Roesch LF, Fulthorpe RR, Riva A, Casella G, Hadwin AK, Kent AD, Daroub SH, Camargo FA, Farmerie WG, Triplett EW (2015) Pyrosequencing enumerates and contrasts soil microbial diversity. ISME J 1(4):283–290
4. Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Huntley J, Fierer N, Owens SM, Betley J, Fraser L, Bauer M, Gormley N, Gilbert JA, Smith G, Knight R (2012) Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. ISME J 6(8):1621–1624
5. van Dijk EL, Auger H, Jaszczyszyn Y, Thermes C (2014) Ten years of next-generation sequencing technology. Trends Genet 30(9): 418–426
6. Schulthess TC (2015) Programming revisited. Nat Phys 11(5): 369–373
7. NERC Environmental Bioinformatics Centre. Bio-Linux. 2009. [http://environmentalomics.org/bio-linux/]
8. Anjar U (2006) Scibuntu: Ubuntu Linux for scientists. [http://scibuntu.sourceforge.net/]
9. Thomson RC (2009) phyLIs: a simple GnU/Linux distribution for phylogenetics and phyloinformatics. Evol Bioinforma 5:91–95
10. Yu G, Wang LG, Meng XH, He QY (2012) LXtoo: an integrated live Linux distribution for the bioinformatics community. BMC Res Notes 5(1):360
11. Liao YC, Lin HH, Sabharwal A, Haase EM, Scannapieco FA (2015) MyPro: a seamless pipeline for automated prokaryotic genome assembly and annotation. J Microbiol Methods 113:72–74
12. Pylro VS, Roesch LF, Ortega JM, do Amaral AM, Tótola MR, Hirsch PR, Rosado AS, Góes-Neto A, da Costa da Silva AL, Rosa CA, Morais DK, Andreote FD, Duarte GF, de Melo IS, Seldin L, Lambais MR, Hungria M, Peixoto RS, Kruger RH, Tsai SM, Azevedo V, Melo IS, Seldin L, Lambais MR, Hungria M, Peixoto RS, Kruger RH, Tsai SM, Azevedo V, Brazilian Microbiome Project Organization Committee (2014) Brazilian Microbiome Project: revealing the unexplored microbial diversity—challenges and prospects. Microb Ecol 67(2):237–241
13. Pylro VS, Roesch LF, Morais DK, Clark IM, Hirsch PR, Tótola MR (2014) Data analysis for 16S microbial profiling from different benchtop sequencing platforms. J Microbiol Methods 107:30–37
14. Goecks J, Nekrutenko A, Taylor J (2010) Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biol 11(8): R86. doi:10.1186/gb-2010-11-8-r86
15. Pylro VS (2014) BMP standards and protocols. http://www.brmicrobiome.org/#!standardsand-protocols/cpbw. Accessed 10 March 2016
16. Pylro VS (2014) BMP What is Included? http://www.brmicrobiome.org/#!what-is-included/c1for. Accessed 10 March 2016
17. Rognes T, Mahé F, Flouri T, Ijaz UZ, Nichols B, Quince C (2015) VSEARCH. https://zenodo.org/record/16153#.VfBYchFViko. Accessed 15 Dec 2015
18. Edgar RC (2010) Search and clustering orders of magnitude faster than BLAST. Bioinformatics 26(19):2460–2461
19. R Development Core Team. R (2008) A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna ISBN 3-900051-07-0. http://www.R-project.org. Accessed 15 Dec 2015
20. Bengtsson-Palme J, Ryberg M, Hartmann M, Branco S, Wang Z, Godhe A, De Wit P, Sánchez-García M, Ebersberger I, de Sousa F, Amend A, Jumpponen A, Unterseher M, Kristiansson E, Abarenkov K, Bertrand YJK, Sanli K, Eriksson KM, Vik U, Veldre V, Nilsson H (2013) Improved software detection and extraction of ITS1 and ITS2 from ribosomal ITS sequences of fungi

and other eukaryotes for analysis of environmental sequencing data. Methods Ecol Evol 4(10):914–919

21. Johnson LS, Eddy SR, Portugaly E (2010) Hidden Markov model speed heuristic and iterative HMM search procedure. BMC Bioinf 11(1):431–439
22. Gordon A, Hannon GJ (2010) Fastx-toolkit. FASTQ/A short-reads preprocessing tools. http://hannonlab.cshl.edu/fastx_toolkit. Accessed 15 Dec 2015
23. McDonald D, Clemente JC, Kuczynski J, Rideout JR, Stombaugh J, Wendel D, Wilke A, Huse S, Hufnagle J, Meyer F, Knight R, Caporaso JG (2012) The biological observation matrix (BIOM) format or: how I learned to stop worrying and love the ome-ome. GigaSci 1(1):7–13
24. Eren AM, Maignien L, Sul WJ, Murphy LG, Grim SL, Morrison HG, Sogin ML (2013) Oligotyping: differentiating between closely related microbial taxa using 16S rRNA gene data. Methods Ecol Evol 4(12):1111–1119
25. Parks DH, Beiko RG (2010) Identifying biologically relevant differences between metagenomic communities. Bioinformatics 26(6):715–721
26. Bastian M, Heymann S, Jacomy M (2008) Gephi: an open source software for exploring and manipulating networks. ICWSM 8:361–362
27. Friedman J, Alm EJ (2012) Inferring correlation networks from genomic survey data. PLoS Comput Biol 8(9):e1002687
28. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 13(11):2498–2504
29. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VA, Nikolenko SI, Pham S, Prjibelski AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA (2012) SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol 19(5):455–477
30. Gurevich A, Saveliev V, Vyahhi N, Tesler G (2013) QUAST: quality assessment tool for genome assemblies. Bioinformatics 29(8): 1072–1075
31. Wu YW, Tang YH, Tringe SG, Simmons BA, Singer SW (2014) MaxBin: an automated binning method to recover individual genomes from metagenomes using an expectation-maximization algorithm. Microbiome 2(1):1–18
32. DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, Huber T, Dalevi D, Hu P, Andersen GL (2006) Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl Environ Microbiol 72(7):5069–5072
33. Maidak BL, Olsen GJ, Larsen N, Overbeek R, McCaughey MJ, Woese CR (1996) The ribosomal database project (RDP). Nucleic Acids Res 24(1):82–85
34. Kõljalg U, Nilsson RH, Abarenkov K, Tedersoo L, Taylor AFS, Bahram M, Bates ST, Bruns TD, Bengtsson-Palme J, Callaghan TM, Douglas B, Drenkhan T, Eberhardt U, Dueñas M, Grebenc T, Griffith GW, Hartmann M, Kirk PM, Kohout P, Larsson E, Lindahl BD, Lücking R, Martín MP, Matheny PB, Nguyen NH, Niskanen T, Oja J, Peay KG, Peintner U, Peterson M, Põldmaa K, Saag L, Saar I, Schüßler A, Scott JA, Senés C, Smith ME, Suija A, Taylor DL, Telleria MT, Weiß M, Larsson K-H (2013) Towards a unified paradigm for sequence-based identification of fungi. Mol Ecol 22(21):5271–5277
35. Haas BJ, Gevers D, Earl AM, Feldgarden M, Ward DV, Giannoukos G, Ciulla D, Tabbaa D, Highlander SK, Sodergren E, Methé B, DeSantis TZ, The Human Microbiome Consortium, Petrosino JF, Knight R, Birren BW (2011) Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. Genome Res 21(3):494–504
36. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215(3):403–410
37. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Peña AG, Goodrich JK, Gordon JI, Huttley GA, Kelley ST, Knights D, Koenig JE, Ley RE, Lozupone CA, McDonald D, Muegge BD, Pirrung M, Reeder J, Sevinsky JR, Turnbaugh PJ, Walters WA, Widmann J, Yatsunenko T, Zaneveld J, Knight R (2010) QIIME allows analysis of high-throughput community sequencing data. Nat Methods 7(5):335–336
38. Nilakanta H, Drews KL, Firrell S, Foulkes MA, Jablonski KA (2014) A review of software for analyzing molecular sequences. BMC Res Notes 7(1):830
39. Andrews S. (2010). FastQC: a quality control tool for high throughput sequence data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc