A peer-reviewed version of this preprint was published in PeerJ on 30 July 2018.

View the peer-reviewed version (peerj.com/articles/5299), which is the preferred citable publication unless you specifically need to cite this preprint.

Morais D, Roesch LFW, Redmile-Gordon M, Santos FG, Baldrian P, Andreote FD, Pylro VS. 2018. BTW—Bioinformatics Through Windows: an easy-to-install package to analyze marker gene data. PeerJ 6:e5299 https://doi.org/10.7717/peerj.5299

BTW – Bioinformatics Through Windows: an easy-to-install package to analyze marker gene data

Daniel Morais [1] , **Luiz Roesch** [2] , **Marc Redmile-Gordon** [3] , **Fausto Santos** [4] , **Petr Baldrian** [1] , **Fernando Andreote** [5] , **Victor S Pylro** Corresp. [5]

[1] Institute of Microbiology, Czech Academy of Sciences, Prague, Czech Republic

[2] Centro para Pesquisa Interdisciplinar em Biotecnologia, Universidade Federal do Pampa, São Gabriel, Brazil

[3] Natural England, Natural England, Sheffield, United Kingdom

[4] Biosystems Informatics and Genomics, Instituto René Rachou, Belo Horizonte, Brazil

[5] Soil Department, “Luiz de Queiroz” College of Agriculture, Piracicaba, Brazil

Corresponding Author: Victor S Pylro
Email address: victor.pylro@usp.br

Recent advances in Next-Generation Sequencing (NGS) make comparative analyses of the composition and diversity of whole microbial communities possible at far greater depth than ever before. This brings new challenges, such as an increased dependence on computation to process these huge datasets. The demand on system resources usually requires migrating from Windows to Linux-based operating systems and prior familiarity with command-line interfaces. To overcome this barrier, we developed a fully automated and easy-to-install package as well as a complete, easy to follow pipeline for microbial metataxonomic analysis operating in the Windows Subsystem for Linux (WSL) - Bioinformatics Through Windows (BTW). BTW combines several open-access tools for processing marker gene data, including 16S rRNA, bringing the user from raw sequencing reads to diversity-related conclusions. It includes data quality filtering, clustering, taxonomic assignment and further statistical analyses, directly in WSL, avoiding the prior need of migrating from Windows to Linux. BTW is expected to boost the use of NGS amplicon data by facilitating rapid access to bioinformatics tools for Windows users. BTW is a Bash script and is available in GitHub (https://github.com/vpylro/BTW). The package is freely available for noncommercial users.

BTW – Bioinformatics Through Windows: an easy-to-install package to analyze marker gene data

Daniel Kumazawa Morais[1], Luiz Fernando Wurdig Roesch[2], Marc Redmile-Gordon[3], Fausto Gonçalves dos Santos[4], Petr Baldrian[1], Fernando Dini Andreote[5], Victor Satler Pylro[5]*

[1]Laboratory of Environmental Microbiology, Institute of Microbiology of the Czech Academy of Sciences, Prague, Czech Republic.
[2]Centro para Pesquisa Interdisciplinar em Biotecnologia, CIP-Biotec, Universidade Federal do Pampa, São Gabriel, Rio Grande do Sul, Brazil.
[3]Natural England, UK.
[4]Biosystems Informatics and Genomics Group, René Rachou Research Center, FIOCRUZ-MG. Belo Horizonte, MG, 30190-002. Brazil.
[5]Soil Microbiology Laboratory, Department of Soil Science, "Luiz de Queiroz" College of Agriculture, ESALQ/USP. Piracicaba, São Paulo, Brazil.

*To whom correspondence should be addressed: victor.pylro@brmicrobiome.org

Abstract

Recent advances in Next-Generation Sequencing (NGS) make comparative analyses of the composition and diversity of whole microbial communities possible at far greater depth than ever before. This brings new challenges, such as an increased dependence on computation to process these huge datasets. The demand on system resources usually requires migrating from Windows to Linux-based operating systems and prior familiarity with command-line interfaces. To overcome this barrier, we developed a fully automated and easy-to-install package as well as a complete, easy to follow pipeline for microbial metataxonomic analysis operating in the Windows Subsystem for Linux (WSL) - Bioinformatics Through Windows (BTW). BTW combines several open-access tools for processing marker gene data, including 16S rRNA, bringing the user from raw sequencing reads to diversity-related conclusions. It includes data quality filtering, clustering, taxonomic assignment and further statistical analyses, directly in WSL, avoiding the prior need of migrating from Windows to Linux. BTW is expected to boost the use of NGS amplicon data by facilitating rapid access to bioinformatics tools for Windows users. BTW is a Bash script and is available in GitHub (https://github.com/vpylro/BTW). The package is freely available for noncommercial users.

Keywords: 16S rRNA, metataxonomics, Windows, microbiome

Introduction

In April 2016, Microsoft announced the release of the Windows Sub-system for Linux (WSL), which is available to Windows 10 users. This distribution consists of a Linux environment compiled through Windows and enables most native command-line tools, utilities and binaries from Linux to run on Windows: the users can now run Bash scripts and all popular Linux command-line tools like *sed*, *awk*, *grep*, *sort*, *apt*, *ssh* and others. One anticipated outcome was that this effort would bring free software to a wider audience, since Windows is the native Operating System (OS) in ~85% of the Desktops and Laptops worldwide (per StatCounter for June 2017 - http://gs.statcounter.com/os-market-share/desktop/worldwide). However, a further but perhaps unconsidered benefit of expanding bioinformatics accessibility, is that Linux command line tools can be used to run several bioinformatics

applications on the hardware available without the need for a dedicated ma-chine, or the hassle of having multiple operating systems on a single machine (dual-boot or virtual machines).

Biologists have recently entered the world of big data (Marx, 2013), consisting of cross-referenced databases, ranging from DNA to metabolic pathways (Cook et al., 2015). However, it is still challenging to effectively perform data analyses using these databases and to manipulate high throughput sequencing data. This is largely because bioinformatics software is typically developed for Unix Shell (Seemann, 2013), and mastery of its command-line interface usually requires intensive training. As previously defined by Mushegian (2011), bioinformatics deals with a dual existence paradigm: as developing technology (the tools), and as a science that applies these tools. In the former, providing a user-friendly graphic interface for data analysis is not always feasible even though command-line gives users and developers flexibility to manipulate and sort data. Attempts have been made to help researchers outside the bioinformatics discipline to use tools dedicated to sequence analysis, for instance, MG-RAST (Meyer et al., 2008), SEED2 (Vetrovsky et al., 2018) and BMP desktop (Pylro et al., 2016). Even so, bioinformatics has been the cause of headache for many scientists, even those who grew up in the computer era.

Making bioinformatics accessible to everyone has been one of the main challenges of contemporary biology. Through the development of bioinformatics tools and training of users in biological data assessment, the Brazilian Microbiome Project (BMP http://brmicrobiome.org – Pylro et al., 2014a) and the Center for Systems Biology (http://c4sys.cz), we observed both students and adept professionals in biological sciences struggle when facing the command-line interfaces of the Unix-based OS (such as Linux and macOSX). The new WSL-Ubuntu feature presented here is aimed to help biologists to provide access to Next-Generation Sequencing (NGS) analysis tools without the above limitations. Although the native Bash tools are useful for manipulating biological data, this distribution come without specific bioinformatics packages, which require several steps of settings and installation before usage. To overcome this issue, we have created an easy to follow tutorial to installing WSL (available on http://brmicrobiome.org/tutorialbtw) and a Bash script (freely available on https://github.com/vpylro/BTW) that should be run through the command-line of WSL-Ubuntu, to set up all the necessary packages for running basic NGS data manipulations and the full microbial community metataxonomic analysis, as previously provided by the BMP to UNIX-based operating systems users (Pylro et al., 2014b).

Methods

Application

To demonstrate the functionality and performance of WSL-Ubuntu in running a complete 16S rRNA data analysis pipeline, we assessed the operation of Qiime 1.9 (Caporaso et al., 2010), VSEARCH 2.4.4 (Rognes et al., 2016) and BMP Scripts (Pylro et al., 2014b), and found that except for packages that require a graphical display (e.g. core_diversity.py from QIIME), all of them work just as in the pure Ubuntu installation. Programs with graphic output are not yet officially supported by WSL, but from our tests with Xming (http://straightrunning.com/XmingNotes) all the programs performed well. The complete pipeline for 16S rRNA data analysis on WSL is available on http://brmicrobiome/win16s (Figure 1). Briefly, 16S reads data of both forward and reverse amplicons are merged into contigs using the “fastq-join” method (Aronesty, 2013) in QIIME. The output file (.fastq) is then quality filtered,

trimmed to equal lengths, dereplicated, sorted and binned into operational taxonomic units (OTUs) using VSEARCH commands (Rognes et al., 2016). Taxonomy is assigned to each representative sequence using the RDP classifier against the GreenGenes (13_8) reference database. An OTU Table (biom format) containing both OTU abundance and taxonomy is constructed using QIIME. Finally, the .biom OTU table is fully compatible with the MicrobiomeAnalyst (Dhariwal et al., 2017), a user-friendly web-based platform for microbiome data analyses and visualizations, including taxonomy plots and estimates of α- and β-diversity (http://microbiomeanalyst.ca).

Results

Benchmarking

To evaluate the usability of such tool for 16S rRNA amplicon data analysis (Supplemental File), we performed a benchmark test using 166.931 reads of 16S rRNA gene data generated from paired-end 150 bp Illumina sequencing (accessible at http://www.brmicrobiome.org/16sillumina), in 3 different scenarios: (1) using the BTW package with the WSL in a Windows 10 machine; (2) using a virtual machine (VirtualBox v. 5.2.6 – Oracle) inside a Windows 10 containing Linux Ubuntu 16.4 (BMP OS [7]) ; and (3) using a dual boot system splitting the hard drive between two operating systems, Windows 10 and Linux Ubuntu 16.4. All tests were performed in a Desktop computer with an Intel® Core™ i7-4790k (4 cores; 8 logical processors - 4.0 GHz) and 32 GB of RAM memory. The time elapsed for processing the same data set in each scenario was 1:45 minutes for the WSL-Ubuntu/BTW, 1:30 minutes using a virtual machine and 1:10 minutes in a native Linux Ubuntu 16.4. The outputs were all the same, resulting in a final OTU table in the BIOM format.

Discussion

Our results show a similar performance among the tree tested approaches, corroborating the power of WSL-Ubuntu, together with BTW, to performing NGS data analyses. Despite being tested for 16S rRNA gene NGS data analysis, there is no limitation in using BTW for 18S rRNA and ITS amplicon analyses as well. One concern about the WSL adoption by the scientific community is that this feature is being distributed by a commercial company, who may stop supporting it at any time. Furthermore, WSL is still in the beta phase, meaning that some Bash scripts and tools currently used for bioinformatics will not work perfectly. For instance, we experienced fails [segmentation fault (core dumped)] while running QIIME Uclust-based commands and the USEARCH package (Edgar, 2010) which seems to be an issue related to the compilation process. On the other hand, if developers, scientists and general users show support through uptake, Microsoft may continue to develop and support the feature. In such an event, it is likely that developers of existing tools currently incompatible to WSL will be convinced to join the movement. Moreover, the bioinformatics users who already learned through this platform, would be able to migrate their knowledge to any full Unix operating system, but with the advantage of having avoided that most daunting first contact barrier with the command line system.

Conclusion

BTW has proved useful to facilitate rapid access to bioinformatics by Windows users, which will boost analytical capacity for NGS.

Funding

This work was supported by the Brazilian Microbiome Project (http://www.brmicrobiome.org) and the National Institute of Science and Technology: Microbiome (http://www.inct-microbiome.org). VSP receives fellowship from FAPESP (Process 14/50320-4 and 16/02219-8).

Conflict of Interest: none declared.

References

Aronesty E. 2013. Comparison of sequencing utility programs. The Open Bioinformatics Journal, 7:1-8.

Caporaso JG, Kuczynski J, Stombaugh J, et al. 2010. QIIME allows analysis of high-throughput community sequencing data. Nat Meth, 7(5): 335-336.

Cook CE, Bergman MT, Finn RD, Cochrane G, Birney, E, Apweiler R. 2015. The European Bioinformatics Institute in 2016: data growth and integration. Nucleic Acids Res, 44(D1), D20-D26.

Dhariwal A, Chong J, Habib S, et al. 2017. MicrobiomeAnalyst - a web-based tool for comprehensive statistical, visual and meta-analysis of microbiome data. Nucleic Acids Res, 45, W180-188.

Edgar R.C. 2010. Search and clustering orders of magnitude faster than BLAST. Bioinformatics, 26, 2460 - 2461.

Marx V. 2013. Biology: The big challenges of big data. Nature, 498(7453), 255-260.

Meyer F, Paarmann D, D'Souza M et al. 2008. The metagenomics RAST server—a public resource for the automatic phylogenetic and functional analysis of metagenomes. BMC Bioinformatics, 9, 386.

Mushegian A. 2011. Grand Challenges in Bioinformatics and Computational Biology. Front Genet, 2, 60.

Pylro VS, Morais DK, de Oliveira FS, et al. 2016. BMPOS: A flexible and user-friendly tool sets for microbiome studies. Microb Ecol, 72(2), 443-447.

Pylro VS, Roesch LFW, Morais DK, et al. 2014b. Data analysis for 16S microbial profiling from different benchtop sequencing platforms. J Microbiol Methods, 107, 30-37.

Pylro VS, Roesch LFW, Ortega JM, et al. 2014a. Brazilian microbiome project: revealing the unexplored microbial diversity—challenges and prospects. Microb Ecol, 67(2), 237-241.

Rognes T, Flouri T, Nichols B, et al. 2016. VSEARCH: a versatile open source tool for metagenomics. PeerJ, 4:e2584.

Seemann T. 2013. Ten recommendations for creating usable bioinformatics command line software. GigaScience, 2(1), 15.

Vetrovsky T, Baldrian P, Morais D. 2018. SEED 2: a user-friendly platform for amplicon high-throughput sequencing data analyses. Bioinformatics, bty071.

Figure 1

Figure 1. Flowchart demonstrating the 16S rRNA profiling data analysis pipeline on Windows (WSL).