Bioinformatics, 34(13), 2018, 2292–2294
doi: 10.1093/bioinformatics/bty071
Advance Access Publication Date: 14 February 2018
Applications Note

OXFORD

Sequence analysis

SEED 2: a user-friendly platform for amplicon high-throughput sequencing data analyses

Tomáš Větrovský, Petr Baldrian and Daniel Morais*

Institute of Microbiology of the CAS, 14220 Prague 4, Czech Republic

*To whom correspondence should be addressed.
Associate Editor: Bonnie Berger

Received on September 26, 2017; revised on January 26, 2018; editorial decision on February 5, 2018; accepted on February 13, 2018

Abstract

Motivation: Modern molecular methods have increased our ability to describe microbial communities. Along with the advances brought by new sequencing technologies, we now require intensive computational resources to make sense of the large numbers of sequences continuously produced. The software developed by the scientific community to address this demand, although very useful, require experience of the command-line environment, extensive training and have steep learning curves, limiting their use. We created SEED 2, a graphical user interface for handling high-throughput amplicon-sequencing data under Windows operating systems.

Results: SEED 2 is the only sequence visualizer that empowers users with tools to handle amplicon-sequencing data of microbial community markers. It is suitable for any marker genes sequences obtained through Illumina, IonTorrent or Sanger sequencing. SEED 2 allows the user to process raw sequencing data, identify specific taxa, produce of OTU-tables, create sequence alignments and construct phylogenetic trees. Standard dual core laptops with 8 GB of RAM can handle ca. 8 million of Illumina PE 300 bp sequences, ca. 4 GB of data.

Availability and implementation: SEED 2 was implemented in Object Pascal and uses internal functions and external software for amplicon data processing. SEED 2 is a freeware software, available at http://www.biomed.cas.cz/mbu/lbwrf/seed/ as a self-contained file, including all the dependencies, and does not require installation. Supplementary data contain a comprehensive list of supported functions.

Contact: daniel.morais@biomed.cas.cz

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

1 Background

High-throughput sequencing technologies have significantly increased our capability to describe microbial communities. These modern molecular techniques can detect fine compositions of communities or subtle changes in the proportions of specific taxa in any environment (Halfvarson *et al.*, 2017; Patin *et al.*, 2017). However, to convert raw amplicon sequencing data into interpretable tables and figures, it is necessary to remove errors introduced by sequencing and library preparation, isolate informative sequences, cluster and count similar sequences and finally identify sequences and determine their taxonomic affiliation.

There are many different pipelines, tools and approaches to execute the steps required to profile microbial communities based on PCR amplicons (Amir *et al.*, 2017; Callahan *et al.*, 2016; Caporaso *et al.*, 2010; Edgar, 2010), although most of them are exclusive to Unix-based operating systems (Linux and IOS; Altschul *et al.*, 2013). Windows-based operating systems are used by ca. 85% of computer users worldwide (Statcounter. Global stat, http://gs.statcounter.com/os-market-share/desktop/worldwide, 2017) due to their more intuitive and user-friendly interface. Unix systems might be more versatile and more resourceful than Windows-based systems (Mangul *et al.*, 2017), but a platform like SEED 2 provides

Windows-users, who are not familiar with Unix systems, with access to sequence-processing technology. Moreover, offering scientists easy-to-use tools not only saves them time but also allows them to spend more energy interpreting their data and planning experiments.

There have been several efforts to facilitate access to bioinformatic tools for users without experience with command-line approaches by offering a GUI (graphical user interface), for instance mcaGUI (Copeland *et al.*, 2012), Clovr-ITS (White *et al.*, 2013), BMP desktop (Pylro *et al.*, 2016) and PipeCraft (Anslan *et al.*, 2017). Nevertheless, none of those tools supplies the users with sequence visualization functions and are not available for Windows users. Sequence viewing and editing tools such as Seqotron (Fourment and Holmes, 2016), UGENE (Okonechnikov *et al.*, 2012) and Jalview (Waterhouse *et al.*, 2009) offers a GUI approach, but are focused on sequence alignment, secondary structure visualization and phylogenetic reconstruction. Furthermore, none of these GUI-based tools support sequence clustering, taxonomy assignment and OTU table construction, functions which are fundamental in microbial community analyses.

Here we present SEED 2, an intuitive graphical user interface for batch processing of fasta and fastq files specific for amplicon sequencing studies. It further facilitates clustering, quality filtering/trimming, taxonomic identification, creation and description of molecular taxa and their phylogenetic placements and for quick assessment of basic microbial community statistics.

2 Materials and methods

SEED 2 works through a graphical interface (Supplementary Fig. S1) to process data from Illumina, Ion Torrent and Sanger sequencing. It accepts fasta, fastq and text formats as input files. For Illumina-generated data, users can join paired-end reads through the graphical interface and use them as input. Upon selecting an input file, SEED 2 loads this file into the computer's memory at which time sequences can be visualized and edited. Through the sequence editor/visualizer, it is possible to remove or trim low-quality sequences, search for specific sequence domains such as sequencing barcodes or primer sequences, including degenerate oligonucleotides, and to group and label sequences containing specified domains in a process known as demultiplexing. This process is typically utilized in metabarcoding studies when multiple samples, individually labelled with artificial barcodes are sequenced together (Caporaso *et al.*, 2012). To reduce computational time, it is possible to dereplicate the sequences after filtering and labelling and work with a concise set of unique sequences, saving a mapping file of the dereplication step. To perform clustering of similar sequences into molecular taxa or OTUs (Operational Taxonomic Units), SEED 2 offers the use of two external algorithms implemented in Vsearch (Rognes *et al.*, 2016) and Usearch (Edgar, 2010). These two software tools perform open-reference clustering, ranking sequences by their abundances, using the most abundant ones as the clustering starting points, and subsequently grouping sequences by an arbitrary (user defined) level of identity. Chimera removal is also possible through these two software tools. After clustering, it is possible to create OTU-tables with counts across samples, to filter out singletons or OTUs at any minimum/maximum abundance threshold and assign taxonomy to the OTUs using the alignment software BLAST (Altschul *et al.*, 1997), which allows the retrieval of any number of best hits for each query. BLAST searches can be performed with the user's favourite database at their own computer or remotely through the NCBI API (internet connection required). SEED 2 even makes it possible to create a user-tailored searchable database from a loaded fasta file. To generate a custom BLAST database, SEED 2 uses the makeblastdb command where the input is a fasta file, and the output is the blastdb itself. With this taxonomical information, nonspecific sequences can be removed, diversity indexes and rarefaction curves can be created and experimental data can be further explored. SEED 2 is one of the very few amplicon-processing tools that has the option to create and visualize phylogenetic trees and allows them to be edited, manipulated and exported as Newick trees. Finally, SEED 2 exports all tables and summaries of the sequence data as 'txt', 'tab' delimited files or they can be directly copied to the clipboard. All command-line software used within SEED 2 are accessed through the graphical interface to ensure maximal usability.

3 Implementation

SEED 2 was written in Object Pascal and is available to all 64-bit Windows platforms from Windows 7 onward. The software implements functions to find sequence domains, group sequences and edit sequence labels. Data is loaded into the computer's RAM and allows users to apply quality filters, trim and manipulate sequences in batch. SEED 2 makes use of **'hash table'** structures (called Dictionary in Pascal language), which is shared by all compiled binaries used during the running of SEED 2. Data is cached into the RAM where it remains for all the processing steps, while in software built on scripting language pipelines, data is erased and reloaded from the RAM at every processing step. This makes SEED 2 not limited to a recommended pipeline, but rather a whole platform for data processing, faster and more memory efficient than scripting-based pipelines. This comes at a cost that the maximum number of sequences to be processed is limited by the user's computer RAM. However, a standard laptop computer with 8 GB RAM can handle at least 8 million Illumina sequencing reads comprising tens of thousands of OTUs which amounts to ca. 4 GB of data. All steps taken during data processing are stored in a workflow manager for automatizing functions and improvement of reproducibility.

As a benchmark, in a Windows 8.1 computer, with 4 cores i7-6700 3.4 GHz and 16 GB of RAM we processed 100 000 16S amplicon Illumina PE 300 bp sequences in 157 min. For the ITS amplicon marker, we processed 100 000 Illumina PE 300 bp sequences in 79 min. All the steps performed, including the time consumed for these analyses and a list of external software and default commands used for each function are reported in the Supplementary Doc1-benchmark and Doc2-list_of_commands, respectively. Moreover, SEED 2 requires 185 MB of HD space.

4 Conclusions

SEED 2 is a fast, intuitive and memory efficient sequence-processing tool. It is applicable to any study using fasta or fastq data from all current high-throughput sequencing platforms. The graphical interface supplies users with tools necessary to quickly analyze meta-taxonomic data.

Acknowledgements

We thank Zander Human (Institute of Microbiology of the CAS/Czech Republic) for his critical comments and review of the written English in the manuscript.

Funding

This work was supported by the Ministry of Education, Youth and Sports of the Czech Republic [LM2015055 to PB].

Conflict of Interest: none declared.

References

Altschul,S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Altschul,S.F. *et al.* (2013) The anatomy of successful computational biology software. *Nat. Biotechnol.*, **31**, 894–897.

Amir,A. *et al.* (2017) Deblur rapidly resolves single-nucleotide community sequence patterns. *mSystems*, **2**, e00191-16.

Anslan,S. *et al.* (2017) PipeCraft: flexible open-source toolkit for bioinformatics analysis of custom high-throughput amplicon sequencing data. *Mol. Ecol. Resour.*, **17**, e234.

Callahan,B.J. *et al.* (2016) DADA2: high-resolution sample inference from Illumina amplicon data. *Nat. Methods*, **13**, 581–583.

Caporaso,J.G. *et al.* (2010) QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods*, **7**, 335–336.

Caporaso,J.G. *et al.* (2012) Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. *ISME J.*, **6**, 1621.

Copeland,W.K. *et al.* (2012) mcaGUI: microbial community analysis R-Graphical User Interface (GUI). *Bioinformatics*, **28**, 2198–2199.

Edgar,R.C. (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, **26**, 2460–2461.

Fourment,M. and Holmes,E.C. (2016) Seqotron: a user-friendly sequence editor for Mac OS X. *BMC Res. Notes*, **9**, 106.

Halfvarson,J. *et al.* (2017) Dynamics of the human gut microbiome in inflammatory bowel disease. *Nat. Microbiol.*, **2**, 17004.

Mangul,S. *et al.* (2017) Addressing the digital divide in contemporary biology: lessons from teaching UNIX. *Trends Biotechnol.*, **35**, 901–903.

Okonechnikov,K. *et al.* (2012) Unipro UGENE: a unified bioinformatics toolkit. *Bioinformatics*, **28**, 1166–1167.

Patin,N.V. *et al.* (2017) Effects of actinomycete secondary metabolites on sediment microbial communities. *Appl. Environ. Microb.*, **83**, e02676-16.

Pylro,V.S. *et al.* (2016) BMPOS: a flexible and user-friendly tool sets for microbiome studies. *Microb. Ecol.*, **72**, 443–447.

Rognes,T. *et al.* (2016) VSEARCH: a versatile open source tool for metagenomics. *PeerJ*, **4**, e2584.

Statcounter Global Stat. (2017) Desktop Operating System Market Share Worldwide (17 July 2017, date last accessed).

Waterhouse,A.M. *et al.* (2009) Jalview Version 2—a multiple sequence alignment editor and analysis workbench. *Bioinformatics*, **25**, 1189–1191.

White,J.R. *et al.* (2013) CloVR-ITS: automated internal transcribed spacer amplicon sequence analysis pipeline for the characterization of fungal microbiota. *Microbiome*, **1**, 6.