Received: 30 December 2016 | Revised: 24 April 2017 | Accepted: 18 May 2017

DOI: 10.1111/1755-0998.12692

RESOURCE ARTICLE

WILEY MOLECULAR ECOLOGY RESOURCES

PipeCraft: Flexible open-source toolkit for bioinformatics analysis of custom high-throughput amplicon sequencing data

Sten Anslan[1] | Mohammad Bahram[1,2] | Indrek Hiiesalu[1] | Leho Tedersoo[3]

[1]Institute of Ecology and Earth Sciences, University of Tartu, Tartu, Estonia

[2]Department of Organismal Biology, Evolutionary Biology Centre, Uppsala University, Uppsala, Sweden

[3]Natural History Museum, University of Tartu, Tartu, Estonia

Correspondence
Sten Anslan, Institute of Ecology and Earth Sciences, University of Tartu, Tartu, Estonia.
Email: sten.anslan@gmail.com

Funding information
Estonian Science Foundation, Grant/Award Number: PUT1317, PUT1399, IUT20-30; MOBERC; ECOLCHANGE

Abstract

High-throughput sequencing methods have become a routine analysis tool in environmental sciences as well as in public and private sector. These methods provide vast amount of data, which need to be analysed in several steps. Although the bioinformatics may be applied using several public tools, many analytical pipelines allow too few options for the optimal analysis for more complicated or customized designs. Here, we introduce PipeCraft, a flexible and handy bioinformatics pipeline with a user-friendly graphical interface that links several public tools for analysing amplicon sequencing data. Users are able to customize the pipeline by selecting the most suitable tools and options to process raw sequences from Illumina, Pacific Biosciences, Ion Torrent and Roche 454 sequencing platforms. We described the design and options of PipeCraft and evaluated its performance by analysing the data sets from three different sequencing platforms. We demonstrated that PipeCraft is able to process large data sets within 24 hr. The graphical user interface and the automated links between various bioinformatics tools enable easy customization of the workflow. All analytical steps and options are recorded in log files and are easily traceable.

KEYWORDS
high-throughput sequencing, metabarcoding, pipeline, sequencing data analysis, software

1 | INTRODUCTION

The development of high-throughput molecular identification methods has greatly improved our understanding about microbial communities and functioning. The so-called metabarcoding approach (c.f. Taberlet, Coissac, Pompanon, Brochmann, & Willerslev, 2012) is commonly used to identify selected groups of micro- and macroorganisms based on a single or a few taxonomic markers in combination. Information about the distribution of organisms is often used in environmental (Tedersoo et al., 2014; Wilson, Sing, Lee, & Wee, 2016) and palaeoecological surveys (Capo et al., 2016; Epp et al., 2012), monitoring diseases (Lohan, Fleischer, Carney, Holzer, & Ruiz, 2016), forensics (Pechal et al., 2014), etc.

The overall throughput of sequencing platforms has been exponentially rising since their introduction to the market >10 years ago (Heather & Chain, 2016). Since then, high-throughput sequencing technologies have been integrated into the working routine of many nonresearch governmental institutions and private sector. Thus, there is an enormous public demand for more accurate data at lower analytical costs as rapidly as possible. As bioinformatics expertise is relatively costly, in-house data analysis by laboratory personnel or assistants would strongly reduce the overall analytical expenses, but this would require user-friendly sequencing data analysis tools. The current bioinformatics workflows are often based on command-line and/or optimized for analysis of specific taxonomic groups of organisms based on a single specific marker. For example, bacteriologists use most widely the 16S ribosomal RNA gene, whereas mycologists and zoologists use the internal transcribed spacer (ITS) of rRNA genes and cytochrome 1 oxidase of mitochondrial DNA (CO1), respectively. Applying the tools that are developed for processing metabarcoding data from prokaryotes is not straightforward for eukaryote metabarcoding data (Gweon et al., 2015; Ramirez-Gonzalez et al., 2013).

There are multiple analytical tools for processing HTS data, most of which have been optimized for prokaryote 16S rRNA gene. Of these, the most widely used routines include mothur (Schloss et al., 2009), QIIME (Caporaso et al., 2010) and usearch (Edgar, 2010). For fungal ITS region, UNITE 454 pipeline (Tedersoo et al., 2010), CLOTU (Kumar et al., 2011), SCATA (https://scata.mykopat.slu.se/), CloVR-ITS (White, Maddox, White, Angiuoli, & Fricke, 2013), SEED (Větrovský & Baldrian, 2013) and several other tools have been developed with an initial focus on the Roche 454 pyrosequencing platform. In most of these analytical tools, the code has been updated and adjusted for the Illumina platform with paired-end reads (see also Balint, Schmidt, Sharma, Thines, & Schmitt, 2014). More recently, bioinformatics tools such as LOTUS (Hildebrand, Tadeo, Voigt, Bork, & Raes, 2014), PIPITS (Gweon et al., 2015), MICCA (Albanese, Fontana, De Filippo, Cavalieri, & Donati, 2015) and Bio-MaS (Fosso et al., 2015) have been developed for analysing Illumina data (also Roche 454). However, none of the above mentioned tools allow analysis of third-generation sequencing data generated by Pacific Biosciences (PacBio) and Oxford Nanopore platforms.

It is important to carefully consider the bioinformatics workflow before the application (Majaneva, Hyytiainen, Varvio, Nagai, & Blomster, 2015), especially for more complicated designs (several markers, multiplex primers, variable tags, etc.), because many analytical pipelines allow too few options for optimal analysis. Additionally, switching between different original programs may suffer from problems in data conversion, especially when the output of one program is unsuitable input for another. For more pronounced taxonomic resolution, trimming of the long barcode to a shorter but more variable region (variable “V” regions of the 16S/18S rRNA genes, the ITS region, variable “D” regions of the 28S rRNA gene and introns of plastid genes and other functional genes) is often important, but un-supported by the majority of programs. For example, in fungal metabarcoding studies, it is crucial to remove the conserved rRNA gene regions flanking the ITS1 and ITS2 subregions, which may greatly distort species-level taxonomic resolution and bias the clustering step (Bengtsson-Palme et al., 2013). So far, only the workflow of Balint et al. (2014) and PIPITS (Gweon et al., 2015) have implemented the extraction step for ITS sequences.

This study introduces PipeCraft, a user-friendly and flexible program with a graphical user interface (Figure S1). PipeCraft incorporates several publicly available tools (Table 1) for analysing amplicon sequencing data from Illumina, PacBio, Ion Torrent and Roche 454 sequencing platforms. All the incorporated programs have been widely used in metabarcoding studies (e.g., Riit et al., 2016; Tedersoo et al., 2014; de Vargas et al., 2015) and their relative performance has been discussed in several papers (e.g., Flynn, Brown, Chain, MacIsaac, & Cristescu, 2015; Forster, Dunthorn, Stoeck, & Mahe, 2016; Westcott & Schloss, 2015). The benefits of PipeCraft over other pipelines include the user-friendly interface, flexibility with barcoding regions, options of choosing HTS platforms and algorithms as well as computation-efficient optimizations for ultra-large data sets. PipeCraft is available through PlutoF system (Abarenkov, Tedersoo, et al., 2010; download link https://plutof.ut.ee/#/datacite/10.15156%2FBIO%2F587450).

2 | SOFTWARE DESIGN AND DESCRIPTION

PipeCraft is an open-source software built using Gambas (3.8.4), Python (2.7) and bash programming language on Ubuntu 14.04. It links a number of third-party applications (Table 1) in which the users may choose the most suitable options for their specific needs. PipeCraft has been built on the Docker container (created with Docker engine v1.11.1) in which all the software and dependencies have already been installed and configured. Considering the dependencies, one exception is usearch (Edgar, 2010). Due to licence restriction, usearch has to be downloaded by users (PipeCraft is compatible with usearch for Linux platform, v8.1.1861). However, the presence of usearch is not crucial for functioning of PipeCraft. The Docker engine itself has to be installed prior to loading the PipeCraft image (instructions available at https://docs.docker.com/engine/installation/linux). To run PipeCraft, users have to download the image file and follow the simple configuration instructions. Currently, PipeCraft is supported only for Linux distributions, but is accessible to other operation systems over the Virtual Box. The overall workflow of PipeCraft is illustrated in Figure 1.

PipeCraft is able to handle several raw data formats produced by Illumina (paired-end fastq files; both ASCII 33 and 64 encoded Phred scales are supported), Pacific Biosciences (bax.h5, bam, fastq), Ion Torrent (bam, fastq) and Roche 454 (sff) sequencing platforms. Depending on a sequencing platform, initial analyses will include creating circular consensus sequences (CCS) for PacBio, or assembling the paired-end reads for Illumina, followed by quality filtering. Working with bam or sff files, the process starts with the conversion of raw data to fastq which is required for the subsequent analyses. For the Oxford Nanopore platform, PipeCraft 1.0 is able to handle fasta and fastq files but not the raw output files. A simple fasta file may be filtered with obiclean (OBITools; Boyer et al., 2016) for the detection of potential PCR/sequencing errors.

For demultiplexing (allocating sequences to samples), users have to specify the primers and tags to process input fastq or fasta file. Because raw data from all sequencing platforms contains both 5′–3′ and 3′–5′ orientated reads, recognition of primers is used to reorient all reads to 5′–3′. For both tags and primers, degenerate positions are allowed. Users can also specify whether to allow any mismatches to primers or tags. PipeCraft allows both one-sided and double-sided tag recognition. In the demultiplexing panel, users may systematically rename the reads and set the minimum unique read size to discard rare groups of reads (e.g., full-length unique sequences that have less than two identical matches in the data set).

Putative chimeric sequences are recognized and discarded by *de novo* and/or reference-based filtering options using uchime (Edgar, Haas, Clemente, Quince, & Knight, 2011) as implemented in vsearch (https://github.com/torognes/vsearch). Additionally, users may filter

TABLE 1 List of third-party applications that are incorporated in PipeCraft

Analysis step	Program	Version	Source
Create circular consensus sequences	pbccs	2.0.2	github.com/PacificBiosciences/unanimity
	bax2bam	0.08	github.com/PacificBiosciences/pitchfork
Assemble paired-end sequences	PANDAseq	2.10	Masella, Bartram, Truszkowski, Brown, and Neufeld (2012)
	FLASH	1.2.11	Magoc and Salzberg (2011)
	vsearch	1.11.1	github.com/torognes/vsearch
Quality filtering	mothur	1.36.1	Schloss et al. (2009)
	OBItools	1.2.9	Boyer et al. (2016)
	vsearch	1.11.1	github.com/torognes/vsearch
Demultiplexing	OBItools	1.2.9	Boyer et al. (2016)
	mothur	1.36.1	Schloss et al. (2009)
Chimera filtering	vsearch	1.11.1	github.com/torognes/vsearch
Gene extraction	ITS extractor	1.0.11	Bengtsson-Palme et al. (2013)
	V-Xtractor	2.1	Hartmann et al. (2010)
	Metaxa2	2.1	Bengtsson-Palme et al. (2015)
	mothur *(cut sequences)*	1.36.1	Schloss et al. (2009)
Clustering	CD-HIT	4.6	Fu, Niu, Zhu, Wu, and Li (2012)
	vsearch	1.11.1	github.com/torognes/vsearch
	swarm	2.1.8	Mahé, Rognes, Quince, De Vargas, and Dunthorn (2015)
	mothur	1.36.1	Schloss et al. (2009)
	usearch[a]	8.1.1861	Edgar (2010)
Taxonomy assignment	BLAST+	2.2.28+	Camacho et al. (2009)
	mothur	1.36.1	Schloss et al. (2009)
Extras	python *with biopython*	2.7	www.python.org
	fqgrep	0.4.4	github.com/indraniel/fqgrep
	fastqc	0.10.1	Andrews (2010)
	fastq-splitter	0.1.2	kirill-kryukov.com/study/tools/fastq-splitter
	fastx toolkit	0.0.14	hannonlab.cshl.edu/fastx_toolkit
	samtools	1.3.1	Li et al. (2009)
	BBMap	36.02	github.com/BioInfoTools/BBMap

[a] Usearch has to be downloaded by the user.

out reads in which the full-length sequencing primer string is detected within the read (so-called multiprimer artefact), which is particularly common in Illumina sequencing data (Balint et al., 2014).

PipeCraft includes tools that enable to extract specific barcoding regions amongst the full read (ITS1, ITS2, full ITS, “V”-regions of 16S/18S and “D”-regions of 28S rRNA genes) using the programs ITSx (Bengtsson-Palme et al., 2013), V-Xtractor (Hartmann, Howes, Abarenkov, Mohn, & Nilsson, 2010) and Metaxa2 (Bengtsson-Palme et al., 2015). As implemented in these original programs, users can choose the groups of organisms to be recognized by the extraction module. Because barcode extraction is a relatively slow process, only unique sequences are subjected to extraction after which all the sequence abundance data will be restored. Additionally, users may precluster the sequences using more relaxed similarity thresholds for further reduction in computing time. Besides barcode extraction, it is possible to trim all reads to a specified length and to remove specified number of bases from the beginning or from the end of the reads.

PipeCraft contains several programs to cluster reads into operational taxonomic units (OTUs; Table 1). It extends the use of implemented clustering programs (except usearch) by allowing the selection of the minimum number of sequences per OTU. This simplifies the common practice of discarding global singletons (i.e., OTUs represented by a single sequence across the entire data) to improve the quality of the data set (Balint et al., 2016; Majaneva et al., 2015; Tedersoo et al., 2010). Users can use the options implemented in individual programs, including specifying the sequence similarity threshold and collapsing homopolymers of specific length. The latter may be required in 454 (Lindahl et al., 2013) and PacBio data sets. PipeCraft is also capable of merging and differentiating several sequencing runs that may contain identical primers and tags that should be clustered together. The OTU by sample table may be created at the same phase. This is based on the cluster- and groups files that are generated in the clustering- and demultiplexing phase, respectively. This tab-delimited text file includes the abundance of each OTU per sample.

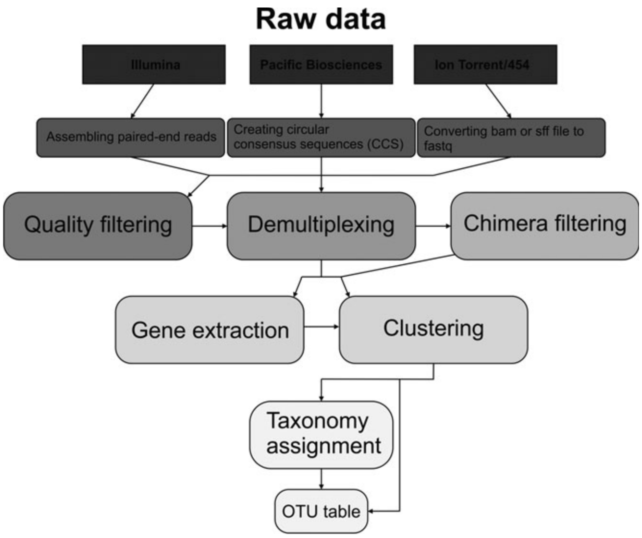

FIGURE 1 Illustration of the PipeCraft workflow

Users may define generation of representative sequences for taxonomy assignment from the following options: i) the longest sequence, ii) the most abundant unique sequence type per OTU; iii) representative nominated by the clustering program by default. These sequences are queried against user-specified reference sequence database(s) for taxonomy assignment. The BLAST algorithm for these searches can be modified by the user to compromise between time and precision by specifying word length, match score, gap creation and extension penalties. The output of this procedure contains two files that contain the best hit and 10 best hits. It is possible to use also the Naïve Bayesian Classifier (Wang, Garrity, Tiedje, & Cole, 2007) as implemented in mothur. Taxonomic assignments of the OTUs are aligned accordingly with the OTU by sample table so that the user may straightforwardly copy the taxonomic information to the OTU table.

Alongside with the analysis pipeline, PipeCraft implements certain other tools for processing fasta, fastq and cluster files. Several data converting options are also available including conversions of bam or sff files to fastq, split fastq and merge files, reformatting the quality scores, sorting paired-end reads and creating BLAST databases from fasta files.

3 | EVALUATION

To evaluate the software, we analysed data obtained from three different sequencing platforms using Ubuntu 14.04 (Intel® Core™ i7-4770 CPU @ 3.40 GHz, 9.5 Gb RAM) via VirtualBox (v5.0.10) on Windows machine. The time scale for the analyses processes is outlined in Table 2. For the test run, we used our yet unpublished data from Illumina MiSeq (2 × 300) and PacBio RS II platforms that consisted of 11,594,152 (14.2 Gb) and 2,422,466 (23.3 Gb) raw- and subreads, respectively. These data represent fungal ITS region amplicons from environmental samples. The Illumina data set was amplified with plant/fungal primers gITS7 (5′-GTGARTCATCGARTCTTTG-3′; Ihrmark et al., 2012) and ITS4ngs (5′-TCCTSCGCTTATTGATATGC-3′; Tedersoo et al., 2014) for obtaining the ITS2 subregion. The PacBio data set was amplified with all-eukaryote primers 1389f (5′-TTGTACACACCGCCC-3′; Amaral-Zettler, McCliment, Ducklow, & Huse, 2009) and ITS4ngsUni (5′-CCTCCSCTTANTDATATGC-3′; Tedersoo & Lindahl, 2016) for obtaining the 18S V9 and full ITS region. Third, we used bacterial 16S (SRA054360) and mitochondrial COI (DRX051481-051490) data produced by the 454 sequencing platform from the study of Hildebrand et al. (2013) and Saitoh et al. (2016) that consisted of 393,070 (892 Mb) and 59,008 reads (44.4 Mb), respectively (Table 2).

For Illumina data, we first assembled the paired-end reads using vsearch (options: fastq_minovlen = 15, fastq_minlen = 50, fastq_maxee = 1, fastq_maxns = 0, allowmergestagger = T) that discarded low-quality sequences at that stage. Raw data from PacBio was subjected to CCS generation with default settings of pbccs (https://github.com/PacificBiosciences/unanimity). The 16S and COI data from 454 pyrosequencing platforms were subjected to quality filtering using mothur (qwindowaverage = 30, qwindowsize = 50, maxambig = 0, maxhomop = 12 and qwindowaverage = 25, qwindowsize = 50, maxambig = 0, maxhomop = 12, minlength = 200, respectively). Initial processes for different platforms generated 4,352,160; 49,289 (CCS reads); 362,548 and 28,335 high-quality reads for Illumina, PacBio and 454, respectively (Table 2). These reads (except COI data) were subjected to demultiplexing, allowing no mismatches in the primer and barcode strings. The COI data were demultiplexed allowing two mismatches in the primers. To exclude putative chimeric sequences, we performed uchime *de novo* filtering. For Illumina and PacBio data, an additional reference-based chimera filtering (against UNITE reference data set v7.0) was performed. Filtered fasta files were subjected to ITSx analysis to obtain only fungal ITS2 subregion for the Illumina data set and full ITS region for the PacBio data set. Prior to clustering, the filtered 16S rRNA gene sequences from the 454 platform were subjected to V-Xtractor for extracting the V4-V5 regions of bacterial 16S. The processed reads were clustered using 97% sequence similarity threshold using CD-HIT (Illumina data) or usearch (PacBio data). The 454 16S (V4-V5) and COI sequence data were clustered using swarm (with $d = 2$) and vsearch (90% similarity threshold), respectively. Singleton OTUs were discarded by setting the minimum cluster size to consist of at least of two sequences. For the final taxonomy assignment step, we chose the representative sequences based on the default options of clustering programs and compared those against the UNITE fungal database (v7.1; Abarenkov, Nilsson, et al., 2010) with the default BLASTn options (word size = 10, reward = 2, penalty = −3, gap open = 5, gap extend = 2) for Illumina and PacBio data. Using the SILVA 16S database (release 123; Quast et al., 2013), taxonomy was assigned to 16S rRNA gene data sing the Naïve Bayesian classifier as implemented in mothur. The taxonomy assignment step for COI data was performed with the default BLASTn values using the BOLD database (local database consisting of sequences downloaded from BOLD; Ratnasingham & Hebert, 2007) as reference. The entire process took 17 hr 18 min,

TABLE 2 Time scale of the data analysis

	Workfolw; time; remaining number of reads/OTUs			
Step	**Illumina; 11,594,152 reads; 14.2 Gb**	**PacBio; 2,533,136 subreads; 23.3 Gb**	**454; 393,070 reads; 892 Mb**	**454; 59,008 reads; 44.4 Mb**
Assembling paired-end reads	vsearch; 19 min *4,352,160 reads*			
Circular consensus sequences (CCS)		pbccs; 8 hr 59 min *49,289 CCS reads*		
Quality filtering	*in assembling step* *4,352,160 reads*	*in CCS step* *49,289 reads*	mothur; 7 min *362,548 reads*	mothur; 1 min *28,335 reads*
Demultiplexing	mothur; 20 min *3,549,518 reads*	mothur; 2 min *22,362 reads*	mothur; 2 min *338,229 reads*	obitools; 1 min *7457 reads*
Chimera filtering	vsearch; 1 h *3,414,596 reads*	vsearch; 2 min *21,450 reads*	vsearch; 5 min *333,384 reads*	vsearch; 1 min *7452 reads*
Gene extraction	ITSx; 15 h *2,964,515 reads*	ITSx; 26 min *21,215 reads*	V-Xtractor; 6 h 39 min *260,752 reads*	
Clustering	CDHIT; 12 min *2143 OTUs (ITS2)*	usearch (UPARSE); 1 min *594 OTUs (full ITS)*	swarm; 2 min *608 OTUs (V4-V5)*	vsearch; 1 min *213 OTUs*
Taxonomy assignment	Database size 443 Mb blast+; 27 min	Database size 443 Mb blast+; 2 hr 27 min	Database size 2.6 Gb Naïve Bayesian classifier; 3 hr 34 min	Database size 1.5 Gb blast+; 55 min
Total time	17 hr 18 min	11 hr 57 min	10 hr 29 min	59 min
Amplicon	*ITS2*	*full ITS*	*16S V4-V5 regions*	*partial COI*

Text in italic represent the remaining number of reads after the process and formed number of OTUs after clustering.

11 hr 57 min, 10 hr 29 min and 59 min for Illumina, PacBio, 454 16S and 454 COI data, respectively. The taxonomic assignments were transferred to the OTU by sample table to obtain a final table with taxonomy, sample and sequence abundance information. However, to detect remaining artefact OTUs that have passed the bioinformatics filtering, it is important to perform additional manual OTU filtering considering the taxonomy assignment values (Nguyen, Smith, Peay, & Kennedy, 2014).

4 | DISCUSSION

The rising popularity of high-throughput sequencing technologies leads to the need for user-friendly sequencing data analysis platforms that do not require strong bioinformatics expertise. Here, we introduce PipeCraft, which enables to develop a suitable pipeline to process HTS amplicons. We demonstrated that PipeCraft is able to efficiently process large data sets in a relatively short time. PipeCraft implements a graphical user interface for easy selection of the available analysis options instead of a command-line-based approach. Because the software and all the dependencies are compiled in a single downloadable image, users are free of struggle to set up multiple tools for bioinformatics analyses. PipeCraft supports bioinformatics analysis of sequence data from various platforms, including third-generation sequencing platforms such as PacBio (raw files of RSII, fastq and fasta files of the Sequel model) and Oxford Nanopore (only fastq and fasta files). Users have multiple choices for quality filtering, specific gene extraction, chimera filtering, clustering and taxonomic assignment that are common to all sequencing platforms, albeit with specific technical nuances. Additionally, PipeCraft offers multiple choices for specific platforms such as options for read assembly (Illumina paired-end data) and circular consensus sequence generation (PacBio data). The final output consists of an OTU by sample table and a text file of reference-based taxonomic assignments. Users may perform analyses as a sequential process from raw input to the final output, or run selected routines only. The implemented third-party software packages should be cited appropriately if they are used. PipeCraft is an open-source package available through PlutoF system (https://plutof.ut.ee/#/datacite/10.15156%2FBIO%2F587450) without the need of creating user account. PipeCraft will be maintained regularly by upgrading the dependencies and/or adding new tools. Detailed instructions on how to get started are provided in the user manual that is included in the package (also provided in Supporting Information).

ACKNOWLEDGEMENTS

The authors would like to thank the bioinformaticians who have created the open-source tools that are incorporated in PipeCraft. We are grateful to Kessy Abarenkov for supporting the PipeCraft distribution through the PlutoF system. We also thank anonymous reviewers and MER Editor Dr. Eric Coissac for feedback on our manuscript. This project is funded by the Estonian Science

Foundation grants (PUT1317, PUT1399), IUT20-30, MOBERC and ECOLCHANGE.

AUTHOR CONTRIBUTIONS

L.T, M.B and S.A conceived the ideas. S.A programmed the software with contributions from I.H and M.B. S.A performed the tests to validate the software using data provided by L.T. S.A wrote the manuscript with contributions from all authors who have approved the final manuscript.

DATA ACCESSIBILITY

The program (https://doi.org/10.15156/bio/587450), user manual and example dataset are freely available through PlutoF system: https://plutof.ut.ee/#/datacite/10.15156%2FBIO%2F587450.

REFERENCES

Abarenkov, K., Nilsson, R. H., Larsson, K.-H., Alexander, I. J., Eberhardt, U., Erland, S., . . . Kõljalg, U. (2010). The UNITE database for molecular identification of fungi - recent updates and future perspectives. *New Phytologist*, *186*, 281–285.

Abarenkov, K., Tedersoo, L., Nilsson, R. H., Vellak, K., Saar, I., Veldre, V., . . . Kõljalg, U. (2010). PlutoF-a Web Based Workbench for Ecological and Taxonomic Research, with an Online Implementation for Fungal ITS Sequences. *Evolutionary Bioinformatics*, *6*, 189–196.

Albanese, D., Fontana, P., De Filippo, C., Cavalieri, D., & Donati, C. (2015). MICCA: a complete and accurate software for taxonomic profiling of metagenomic data. *Scientific Reports*, *5*, 9743.

Amaral-Zettler, L. A., McCliment, E. A., Ducklow, H. W., & Huse, S. M. (2009). A method for studying protistan diversity using massively parallel sequencing of V9 hypervariable regions of small-subunit ribosomal RNA genes. *PLoS ONE*, *4*, e6372.

Andrews, S. (2010). FastQC: A quality control tool for high throughput sequence data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Balint, M., Bahram, M., Eren, A. M., Faust, K., Fuhrman, J. A., Lindahl, B., . . . Tedersoo, L. (2016). Millions of reads, thousands of taxa: microbial community structure and associations analyzed via marker genes. *FEMS Microbiology Reviews*, *40*, 686–700.

Balint, M., Schmidt, P.-A., Sharma, R., Thines, M., & Schmitt, I. (2014). An Illumina metabarcoding pipeline for fungi. *Ecology and Evolution*, *4*, 2642–2653.

Bengtsson-Palme, J., Hartmann, M., Eriksson, K. M., Pal, C., Thorell, K., Larsson, D. G. J., & Nilsson, R. H. (2015). metaxa2: improved identification and taxonomic classification of small and large subunit rRNA in metagenomic data. *Molecular Ecology Resources*, *15*, 1403–1414.

Bengtsson-Palme, J., Ryberg, M., Hartmann, M., Branco, S., Wang, Z., Godhe, A., . . . Nilsson, R. 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 in Ecology and Evolution*, *4*, 914–919.

Boyer, F., Mercier, C., Bonin, A., Le Bras, Y., Taberlet, P., & Coissac, E. (2016). OBITOOLS: a UNIX-inspired software package for DNA metabarcoding. *Molecular Ecology Resources*, *16*, 176–182.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST plus: architecture and applications. *BMC Bioinformatics*, *10*, 421.

Capo, E., Debroas, D., Arnaud, F., Guillemot, T., Bichet, V., Millet, L., . . . Pignol, C. (2016). Long-term dynamics in microbial eukaryotes communities: a paleolimnological view based on sedimentary DNA. *Molecular Ecology*, *25*, 5925–5943.

Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., . . . Gordon, J. I. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nature Methods*, *7*, 335–336.

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

Edgar, R. C., Haas, B. J., Clemente, J. C., Quince, C., & Knight, R. (2011). UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics*, *27*, 2194–2200.

Epp, L. S., Boessenkool, S., Bellemain, E. P., Haile, J., Esposito, A., Riaz, T., . . . Brochmann, C. (2012). New environmental metabarcodes for analysing soil DNA: potential for studying past and present ecosystems. *Molecular Ecology*, *21*, 1821–1833.

Flynn, J. M., Brown, E. A., Chain, F. J. J., MacIsaac, H. J., & Cristescu, M. E. (2015). Toward accurate molecular identification of species in complex environmental samples: testing the performance of sequence filtering and clustering methods. *Ecology and Evolution*, *5*, 2252–2266.

Forster, D., Dunthorn, M., Stoeck, T., & Mahe, F. (2016). Comparison of three clustering approaches for detecting novel environmental microbial diversity. *PeerJ*, *4*, e1692.

Fosso, B., Santamaria, M., Marzano, M., Alonso-Alemany, D., Valiente, G., Donvito, G., . . . Pesole, G. (2015). BioMaS: a modular pipeline for Bioinformatic analysis of Metagenomic AmpliconS. *BMC Bioinformatics*, *16*, 203.

Fu, L., Niu, B., Zhu, Z., Wu, S., & Li, W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, *28*, 3150–3152.

Gweon, H. S., Oliver, A., Taylor, J., Booth, T., Gibbs, M., Read, D. S., . . . Schonrogge, K. (2015). PIPITS: an automated pipeline for analyses of fungal internal transcribed spacer sequences from the Illumina sequencing platform. *Methods in Ecology and Evolution*, *6*, 973–980.

Hartmann, M., Howes, C. G., Abarenkov, K., Mohn, W. W., & Nilsson, R. H. (2010). V-Xtractor: An open-source, high-throughput software tool to identify and extract hypervariable regions of small subunit (16 S/18 S) ribosomal RNA gene sequences. *Journal of Microbiological Methods*, *83*, 250–253.

Heather, J. M., & Chain, B. (2016). The sequence of sequencers: The history of sequencing DNA. *Genomics*, *107*, 1–8.

Hildebrand, F., Nguyen, T. L. A., Brinkman, B., Yunta, R. G., Cauwe, B., Vandenabeele, P., . . . Raes, J. (2013). Inflammation-associated enterotypes, host genotype, cage and inter-individual effects drive gut microbiota variation in common laboratory mice. *Genome Biology*, *14*, R4.

Hildebrand, F., Tadeo, R., Voigt, A. Y., Bork, P., & Raes, J. (2014). LotuS: an efficient and user-friendly OTU processing pipeline. *Microbiome*, *2*, 30.

Ihrmark, K., Bodeker, I. T. M., Cruz-Martinez, K., Friberg, H., Kubartova, A., Schenck, J., . . . Lindahl, B. D. (2012). New primers to amplify the fungal ITS2 region - evaluation by 454-sequencing of artificial and natural communities. *FEMS Microbiology Ecology*, *82*, 666–677.

Kumar, S., Carlsen, T., Mevik, B.-H., Enger, P., Blaalid, R., Shalchian-Tabrizi, K., & Kauserud, H. (2011). CLOTU: An online pipeline for processing and clustering of 454 amplicon reads into OTUs followed by taxonomic annotation. *BMC Bioinformatics*, *12*, 1–9.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., . . . Durbin, R. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics*, *25*, 2078–2079.

Lindahl, B. D., Nilsson, R. H., Tedersoo, L., Abarenkov, K., Carlsen, T., Kjoller, R., . . . Kauserud, H. (2013). Fungal community analysis by high-throughput sequencing of amplified markers - a user's guide. *New Phytologist*, *199*, 288–299.

Lohan, K. M. P., Fleischer, R. C., Carney, K. J., Holzer, K. K., & Ruiz, G. M. (2016). Amplicon-based pyrosequencing reveals high diversity of protistan parasites in Ships' Ballast Water: implications for biogeography and infectious diseases. *Microbial Ecology*, *71*, 530–542.

Magoc, T., & Salzberg, S. L. (2011). FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics*, *27*, 2957–2963.

Mahé, F., Rognes, T., Quince, C., De Vargas, C., & Dunthorn, M. (2015). Swarm v2: highly-scalable and high-resolution amplicon clustering. *PeerJ*, *3*, e1420.

Majaneva, M., Hyytiainen, K., Varvio, S. L., Nagai, S., & Blomster, J. (2015). Bioinformatic amplicon read processing strategies strongly affect eukaryotic diversity and the taxonomic composition of communities. *PLoS ONE*, *10*, e0130035.

Masella, A. P., Bartram, A. K., Truszkowski, J. M., Brown, D. G., & Neufeld, J. D. (2012). PANDAseq: PAired-eND Assembler for Illumina sequences. *BMC Bioinformatics*, *13*, 31.

Nguyen, N. H., Smith, D., Peay, K., & Kennedy, P. (2014). Parsing ecological signal from noise in next generation amplicon sequencing. *New Phytologist*, *205*, 1389–1393.

Pechal, J. L., Crippen, T. L., Benbow, M. E., Tarone, A. M., Dowd, S., & Tomberlin, J. K. (2014). The potential use of bacterial community succession in forensics as described by high throughput metagenomic sequencing. *International Journal of Legal Medicine*, *128*, 193–205.

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., . . . Glöckner, F. O. (2013). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Research*, *41*, D590–D596.

Ramirez-Gonzalez, R., Yu, D. W., Bruce, C., Heavens, D., Caccamo, M., & Emerson, B. C. (2013). PyroClean: Denoising Pyrosequences from Protein-Coding Amplicons for the Recovery of Interspecific and Intraspecific Genetic Variation. *PLoS ONE*, *8*, e57615.

Ratnasingham, S., & Hebert, P. D. N. (2007). BOLD: The Barcode of Life Data System (www.barcodinglife.org). *Molecular Ecology Notes*, *7*, 355–364.

Riit, T., Tedersoo, L., Drenkhan, R., Runno-Paurson, E., Kokko, H., & Anslan, S. (2016). Oomycete-specific ITS primers for identification and metabarcoding. *MycoKeys*, 17–30.

Saitoh, S., Aoyama, H., Fujii, S., Sunagawa, H., Nagahama, H., Akutsu, M., . . . Nakamori, T. (2016). A quantitative protocol for DNA metabarcoding of springtails (Collembola). *Genome*, *59*, 705–723.

Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., . . . Weber, C. F. (2009). Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and Environmental Microbiology*, *75*, 7537–7541.

Taberlet, P., Coissac, E., Pompanon, F., Brochmann, C., & Willerslev, E. (2012). Towards next-generation biodiversity assessment using DNA metabarcoding. *Molecular Ecology*, *21*, 2045–2050.

Tedersoo, L., Bahram, M., Polme, S., Kõljalg, U., Yorou, N. S., Wijesundera, R., . . . Abarenkov, K. (2014). Global diversity and geography of soil fungi. *Science*, *346*, 1256688.

Tedersoo, L., & Lindahl, B. (2016). Fungal identification biases in microbiome projects. *Environmental Microbiology Reports*, *8*, 774–779.

Tedersoo, L., Nilsson, R. H., Abarenkov, K., Jairus, T., Sadam, A., Saar, I., . . . Koljalg, U. (2010). 454 Pyrosequencing and Sanger sequencing of tropical mycorrhizal fungi provide similar results but reveal substantial methodological biases. *New Phytologist*, *188*, 291–301.

de Vargas, C., Audic, S., Henry, N., Decelle, J., Mahé, F., Logares, R., . . . Probert, I. (2015). Eukaryotic plankton diversity in the sunlit ocean. *Science*, *348*, 1261605.

Větrovský, T., & Baldrian, P. (2013). Analysis of soil fungal communities by amplicon pyrosequencing: current approaches to data analysis and the introduction of the pipeline SEED. *Biology and Fertility of Soils*, *49*, 1027–1037.

Wang, Q., Garrity, G. M., Tiedje, J. M., & Cole, J. R. (2007). Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Applied and Environmental Microbiology*, *73*, 5261–5267.

Westcott, S. L., & Schloss, P. D. (2015). De novo clustering methods outperform reference-based methods for assigning 16S rRNA gene sequences to operational taxonomic units. *PeerJ*, *3*, e1487.

White, J. R., Maddox, C., White, O., Angiuoli, S. V., & Fricke, W. F. (2013). CloVR-ITS: Automated internal transcribed spacer amplicon sequence analysis pipeline for the characterization of fungal microbiota. *Microbiome*, *1*, 6.

Wilson, J. J., Sing, K. W., Lee, P. S., & Wee, A. K. S. (2016). Application of DNA barcodes in wildlife conservation in Tropical East Asia. *Conservation Biology*, *30*, 982–989.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

How to cite this article: Anslan S, Bahram M, Hiiesalu I, Tedersoo L. PipeCraft: Flexible open-source toolkit for bioinformatics analysis of custom high-throughput amplicon sequencing data. *Mol Ecol Resour*. 2017;17:e234–e240. https://doi.org/10.1111/1755-0998.12692