Great differences in performance and outcome of high-throughput sequencing data analysis platforms for fungal metabarcoding

Sten Anslan[1*], R. Henrik Nilsson[2], Christian Wurzbacher[3], Petr Baldrian[4], Leho Tedersoo[5], Mohammad Bahram[5,6,7*]

[1]Braunschweig University of Technology, Zoological Institute, Mendelssohnstr. 4, 38106 Braunschweig, Germany. [2]Gothenburg Global Biodiversity Centre, Department of Biological and Environmental Sciences, University of Gothenburg, Box 461, 405 30 Gothenburg, Sweden. [3]Technical University of Munich, Am Coulombwall 3, 85748 Garching, Germany. [4]Institute of Microbiology of the Czech Academy of Sciences, Videnska 1083, 14220 Praha 4, Czech Republic. [5]Natural History Museum of Tartu University, 14a Ravila, 50411 Tartu, Estonia. [6]Department of Ecology, Swedish University of Agricultural Sciences, Ulls väg 16, 756 51 Uppsala, Sweden. [7]Department of Ecology, Swedish University of Agricultural Sciences, Ulls väg 16, 756 51 Uppsala, Sweden.

Correspondence: s.anslan@tu-braunschweig.de, +372 58372084; bahram@ut.ee, +372 5160487

Abstract

Along with recent developments in high-throughput sequencing (HTS) technologies and thus fast accumulation of HTS data, there has been a growing need and interest for developing tools for HTS data processing and communication. In particular, a number of bioinformatics tools have been designed for analysing metabarcoding data, each with specific features, assumptions and outputs. To evaluate the potential effect of the application of different bioinformatics workflow on the results, we compared the performance of different analysis platforms on two contrasting high-throughput sequencing data sets. Our analysis revealed that the computation time, quality of error filtering and hence output of specific bioinformatics process largely depends on the platform used. Our results show that none of the bioinformatics workflows appear to perfectly filter out the accumulated errors and generate Operational Taxonomic Units, although PipeCraft, LotuS and PIPITS perform better than QIIME2 and Galaxy for the tested fungal amplicon data

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27019v1 | CC BY 4.0 Open Access | rec: 3 Jul 2018, publ: 3 Jul 2018

set. We conclude that the output of each platform require manual validation of the OTUs by examining the taxonomy assignment values.

Key words: Microbial communities, microbiome, mycobiome, fungal biodiversity, metagenomics, amplicon sequencing.

Introduction

Fungi are major ecological and functional players in terrestrial ecosystems. The full diversity of fungi remains largely uncharted due to their largely unculturable nature, the lack of tangible morphological manifestations and shortcomings of the mycological community to sample beyond traditional habitats and substrates (Grossart et al., 2016; Hibbett et al., 2017). As a result, identification of fungi has come to rely mainly on direct DNA sequencing of material containing fungal hyphae or spores. In this regard, several DNA barcoding regions have been evaluated, and the current consensus is that the nuclear ribosomal internal transcribed spacer (ITS) region is the best region for delimiting fungal taxa at the species level across a variety of fungal groups (Schoch et al., 2012). Recent advances in high-throughput sequencing (HTS) have made it possible to sequence millions of reads and identify thousands of fungal taxa from a single sample. Handling this enormous amount of data is often complicated and requires extensive bioinformatics expertise.

Multiple analysis platforms have been introduced to facilitate bioinformatics treatment of HTS data. However, most of these software suites were developed for the prokaryotic 16S rRNA gene and may therefore perform poorly with other markers and other organisms, in particular ITS sequences due to their length variation and unalignability across taxonomic expanses. To accommodate for this, several tailored platforms have recently been developed to specifically address fungal ITS datasets (e.g. Hildebrand et al., 2014; Gweon et al., 2015; Anslan et al., 2017; Větrovský et al., 2018). These platforms cover multiple steps of the analysis procedure, including quality control, clustering, taxonomic assignment and generating Operation Taxonomic Unit (OTU) abundance tables. Many of these platforms cover all these analysis steps, whereas others do not.

The application of different bioinformatics workflows may introduce variation in the data quality and output OTU table (Majaneva et al., 2015; Sinha et al., 2017). However, to date there

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27019v1 | CC BY 4.0 Open Access | rec: 3 Jul 2018, publ: 3 Jul 2018

are no data on the relative performance of the available tools for fungal HTS data analysis. In this study, we report on the relative performance of the most popular software pipelines on two contrasting HTS datasets.

Methods

We compared the performance of bioinformatics analysis platforms on two fungal ITS data sets with contrasting properties. Tested data sets include Illumina MiSeq paired-end ITS2 amplicons from arthropod substrates (Anslan et al., 2018), and full ITS circular consensus sequences from Pacific Biosciences (PacBio) Sequel machine, amplified from soil samples (unpublished data).

For bioinformatics analyses, we used multiple platforms that support all steps in the analysis of HTS-based metabarcoding datasets: QIIME2 (v2018.2; Caporaso et al., 2010), LotuS (v1.59; Hildebrand et al., 2014), Galaxy (v.2.1.1; Afgan et al., 2016), PipeCraft (v1.0; Anslan et al., 2017), and PIPITS (v2.0; Gweon et al., 2015). Quality filtering was performed using VSEARCH (Rognes et al., 2016), trimmomatic (Bolger et al., 2014), DADA2 (Callahan et al., 2016), sdm (Hildebrand et al., 2014) and fastx (http://hannonlab.cshl.edu/fastx_toolkit). Quality filtered sequences were passed through chimeric reads removal algorithms as implemented in USEARCH (Edgar et al., 2011; Edgar, 2013) and VSEARCH. Using PipeCraft, LotuS and PIPITS, reads were also subjected to ITS extraction using ITSx (Bengtsson-Palme et al., 2013) to remove conservative flanking genes of the ITS region. OTU formation (clustering) was performed using USEARCH and/or VSEARCH as outlined below. For each platform, we relied on *de-novo* single linkage clustering, which is the most popular approach in fungal community studies, knowing that reference based clustering methods can provide similar results (Cline et al., 2017). Taxonomic affiliations were assigned to OTUs using DP Naive Bayesian rRNA Classifier (Wang et al., 2007) (RDP classifier v2.11) with the Warcup Fungal ITS trainset 2 (Deshpande et al., 2016) (confidence threshold: 80%) as well as BLAST+ (Camacho et al., 2009) search (e-value = 0.001, word size = 7, reward = 1, penalty = -1, gap open = 1, gap extend = 2) against the UNITE v7.2 reference database.

Using QIIME2, reads were assembled (Illumina data) and quality filtered using DADA2 (Callahan et al., 2016) with default options, except --p-trunc-len = 0, --p-max-ee = 1 and --p-chimera-method = none (with chimera-method = consensus, QIIME2 reported error for our data). Clustering was performed with VSEARCH cluster-features-de-novo (--p-perc-identity

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27019v1 | CC BY 4.0 Open Access | rec: 3 Jul 2018, publ: 3 Jul 2018

0.97). In LotuS pipline, data was assembled (Illumina data), quality filtered (minimum length = 170, minAvgQuality = 27, TruncateSequenceLength = 170, maxAccumulatedError = 0.75) and demultiplexed with sdm (pdiffs = 1, bdiffs = 1). Chimera filtering was done using USEARCH *de novo* chimera filtering (abundance annotation = 0.97, abskew = 2), and USEARCH reference-based chimera filtering using UNITE v7.2 (Kõljalg et al., 2013) as reference database. Flanking genes of the ITS region were discarded using ITSx (v1.0.11; default options). ITS reads were clustered to OTUs with USEARCH/UPARSE algorithm (-id = 3, -minsize = 2). Using web-based Galaxy pipeline, Illumina data was assembled with Fastq joiner (Galaxy Version 2.0.1.1; Blankenberg et al., 2010) with default options. Quality filtering was performed with Trimmomatic (Galaxy Version 0.36.3; Bolger et al., 2014) – SLIDINGWINDOW; number of bases to average across = 15, average quality required = 30, minimum length of kept reads = 45. Fastq files were converted to FASTA files using FASTQ to FASTA converter (Galaxy Version 1.0.0). Fasta files were demultiplexed using mothur (Galaxy Version 1.39.5.0; Schloss et al., 2009) – pdiffs=2, bdiffs=1. Because sequences were of mixed orientation in the files (5'-3' and 3'-5'), demultiplexing step was repeated for reverse oriented sequences (reads were reversed using mothur reverse.seqs). Chimera filtering was done using VSEARCH chimera detection (Galaxy Version 1.9.7.0) with default settings (abundance annotation = 97% similarity threshold) and using the UNITE v7.2 database as reference. Clustering was performed using VSEARCH (--cluster-fast –id 0.97). In PipeCraft platform reads were assembled (Illumina data) and quality filtered using VSEARCH (minimum overlap = 15, minimum length = 100, E max = 1, max ambiguous = 0, allowstagger = T). Demultiplexing was done using mothur (pdiffs=2, bdiffs=1). In this step sequences are also reoriented into the 5'-3' orientation based on primers (2 mismatches allowed).

Chimeric sequences were removed using VSEARCH *de novo* (abundance annotation = 0.97, abskew = 2) and reference-based (UNITE v7.2 as reference) chimera filtering algorithms. In chimera filtering step, PipeCraft supported option for "primer artefact" removal was also used (sequences where primer strings were found in the middle of the sequence were removed). ITS reads were extracted using ITSx (default options). Clustering was done using USEARCH/UPARSE algorithm (id = 3, minsize = 2). Using PIPITS, sequences were assembled with VSEARCH and quality-filtering was done with fastx through the PIPITS command pispino_createreadpairslist. The ITSx was executed through the PIPITS command pipits_funits.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27019v1 | CC BY 4.0 Open Access | rec: 3 Jul 2018, publ: 3 Jul 2018

Chimera filtering and clustering was done using VSEARCH through the PIPITS command pipits_process.

The manual OTU table filtering was based on the BLAST similarity scores when run against UNITE (v7.2) reference database. Any OTUs that had no BLAST hit or that were not classified to the kingdom Fungi were discarded from the OTU table. Remaining OTUs were filtered based on BLAST e-value and query coverage. OTUs with higher e-value than $1e^{-25}$ and query coverage less than 70% were excluded from the dataset (as putative artefacts or non-fungal OTUs). Additionally, OTUs with low numbers of sequences per sample were removed (less than 10 sequences per sample; Brown et al. (2015)). Finally, the LULU (Frøslev et al., 2017) algorithm was applied (minimum_ratio_type = "min", minimum_match = 97) to merge consistently co-occurring 'daughter' OTUs.

To detect the effect of analysis platform choice on the OTU composition, we pooled sequences originating from different platforms and applied common clustering method to generate a single OTU table. Filtered reads from PipeCraft, LotuS, and PIPITS were pooled and clustered using CD-HIT at 97% sequence similarity (-id 0.97; Fu et al., 2012).

We used PERMANOVA analysis (Anderson and Walsh, 2013) (Type III SS, 4,999 permutations) on Bray-Curtis distances of Hellinger-transformed OTU matrices, using PRIMER6 (Clarke and Gorley, 2006). The numbers of sequences per sample were included in the analysis as covariates. Rarefaction curves were generated based on OTU abundance matrices for each dataset using the RTK package (Saary et al., 2017) of R (R-Core-Team, 2015).

Results and Discussion

Properties of bioinformatics analysis platforms

All tested bioinformatics platforms offer straightforward installation. While Galaxy provides a freely available online platform, the benefits of PipeCraft and QIIME2 include easy-to-use graphical user interfaces and multiple options for data analysis. These platforms bundle many tools for diverse tasks (Figure 1). LotuS and PIPITS represent command-line based platforms. PIPITS offers a limited number of tools, but data analysis is easily performed with a straightforward pipeline. LotuS has been developed to minimize computational time and memory requirements. Specifically for accuracy of ITS-based analyses of fungi and other eukaryotes, PipeCraft, LotuS and PIPITS implement the ITSx tool (Bengtsson-Palme et al., 2013), which

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27019v1 | CC BY 4.0 Open Access | rec: 3 Jul 2018, publ: 3 Jul 2018

removes the fragments of conservative flanking genes for precise clustering purposes. There is no such option in QIIME2 and Galaxy.

Bioinformatics platforms differ by specific requirements to the input data, with the options being a raw multiplexed file (a single file containing all sequences from one run) and multiple demultiplexed files (reads split into separate files based on indexes). PipeCraft and Galaxy use raw multiplexed data, whereas QIIME2 and PIPITS require demultiplexed files. Only LotuS allows both, multiplexed and demultiplexed files as input. As the raw data files are multiplexed by default, QIIME2 and PIPITS platforms required additional steps of analyses outside these tool to meet the input requirements. Using a Python script, we demultiplexed the raw Illumina data, allowing 2 and 1 mismatches to primer and index strings, respectively. However, PacBio data analysis was dropped for QIIME2 and PIPITS as the present versions of these platforms are limited to analysis of short read (Illumina) data.

Performance of bioinformatics platforms on sequence data

For both the Illumina and PacBio datasets, the final OTU richness (singleton OTUs excluded) differed considerably among the tested workflows (Table 1; Figure 2). Compared with the other platforms, the Galaxy workflow produced a substantially larger number of OTUs, which was most likely due to the effect of inadequate error filtering. In particular, for Illumina data, this was illustrated by the QIIME2 workflow that generated much less OTUs using the same clustering method but different error-filtering algorithm. None of these platforms included the ITS extraction step. Pipelines that produced roughly comparable numbers of total OTUs (QIIME2, PipeCraft, PIPITS, and LotuS for Illumina data) still exhibited large variation in OTU richness per sample (Figure 2,3). By performing joint *de-novo* clustering for filtered sequences from different pipelines, we observed a weak but significant effect of pipeline choice on overall OTU composition for the Illumina data set (PERMANOVA: pseudo-$F_{2,868}$ = 5.88, R^2_{adj} = 0.012, $P < 0.001$). For PacBio data set, differences among platforms were slightly stronger (pseudo-$F_{2,512}$ = 9.174; R^2_{adj} = 0.033, $P < 0.001$).

Taxonomic annotation tools differed in the ability to classify OTUs. In general, BLAST searches revealed many cases of high-quality matches to non-fungal organisms (in some cases for hundreds of OTUs), while RDP as combined with the Warcup Fungal ITS trainset optimistically classified all OTUs to Fungi (100% confidence). Numerous papers have evaluated

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27019v1 | CC BY 4.0 Open Access | rec: 3 Jul 2018, publ: 3 Jul 2018

the performance of different methods on the accuracy of taxonomic assignment, and performance inevitably hinges on the completeness of the reference database used (e.g. Gdanetz et al., 2017; Richardson et al., 2017). In spite of its relatively rapid performance, the RDP Fungal ITS trainset does not include any non-fungal data, which explains its shortcomings in detecting non-fungal OTUs. However, the confidence score of an RDP classifier did not exceed 64% for non-fungal OTUs, mostly overestimating the group of unclassified fungi.

We also observed that the quality-filtered datasets included up to ~10% of obvious erroneous/chimeric OTUs that produced matches with low query coverage and confidence scores. A Long tail of satellite OTUs, assigned to a single species hypothesis with 99-100% BLAST identity and RDP classifier confidence level, were also common - especially in the results where relatively a high number of OTUs was observed (Galaxy platform). After filtering the spurious OTUs manually (see Methods), we found that richness estimates per sample became more homogeneous across pipelines (Illumina data: Figure 3). When OTU table filtering was applied to jointly clustered reads from different pipelines, the significant effect of pipeline choice on the community composition diminished (Illumina data: pseudo-$F_{2,837}$ = 0.955, R^2_{adj} = 0.007, P = 0.779).

In conclusion, our results indicate that bioinformatics analysis pipelines greatly differ in their relative performance on ITS data sets targeting fungi, although roughly similar quality-oriented settings were implemented. Overall, our recommended Illumina data workflow would be PipeCraft, PIPITS or LotuS, which provide a good balance between speed, mycological accuracy (including support for ITS Extractor) and technical quality. For PacBio, the tools implemented in PipeCraft were most suitable for the long-read analysis. Conversely, the widely used platform in prokaryote 16S-based studies, Galaxy, performed relatively poorly on the ITS data. While QIIME2 implements accurate quality filtering algorithm of DADA2, the lack of ITS region extraction lowers the accuracy for mycological studies. Of classification tools, BLAST searches against the UNITE database provided more accurate results on the kingdom and phylum levels compared with the RDP and Warcup ITS trainset combined. We emphasize that none of the tested bioinformatics workflows are able to fully filter out the errors that accumulated during sample preparation and sequencing, even when using the most elaborate error-filtering options. Therefore, manual curation of OTU tables continues to be an important step in obtaining robust datasets, although semi-automatic tools to assist evaluation are becoming available (Frøslev et

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27019v1 | CC BY 4.0 Open Access | rec: 3 Jul 2018, publ: 3 Jul 2018

al., 2017). It is also important to rely on high-coverage reference databases to be able to recognize non-target organisms and metagenomic reads.

Acknowledgments

We thank Falk Hildebrand for advice on bioinformatics analysis. This study was supported by the Estonian Research Council (grant no. PUT1317).

References

Afgan, E., Baker, D., Van den Beek, M., Blankenberg, D., Bouvier, D., Čech, M., Chilton, J., Clements, D., Coraor, N., Eberhard, C., 2016. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. Nucleic Acids Research 44, W3-W10.

Anderson, M.J., Walsh, D.C.I., 2013. PERMANOVA, ANOSIM, and the Mantel test in the face of heterogeneous dispersions: What null hypothesis are you testing? Ecological Monographs 83, 557-574.

Anslan, S., Bahram, M., Hiiesalu, I., Tedersoo, L., 2017. PipeCraft: flexible open-source toolkit for bioinformatics analysis of custom high-throughput amplicon sequencing data. Molecular Ecology Resources 17, e234-e240.

Anslan, S., Bahram, M., Tedersoo, L., 2018. Seasonal and annual variation in fungal communities associated with epigeic springtails (Collembola spp.) in boreal forests. Soil Biology and Biochemistry 116, 245-252.

Bengtsson-Palme, J., Ryberg, M., Hartmann, M., Branco, S., Wang, Z., Godhe, A., De Wit, P., Sanchez-Garcia, M., Ebersberger, I., de Sousa, F., Amend, A.S., Jumpponen, A., Unterseher, M., Kristiansson, E., Abarenkov, K., Bertrand, Y.J.K., Sanli, K., Eriksson, K.M., Vik, U., Veldre, V., 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.

Blankenberg, D., Gordon, A., Von Kuster, G., Coraor, N., Taylor, J., Nekrutenko, A., Team, G., 2010. Manipulation of FASTQ data with Galaxy. Bioinformatics 26, 1783-1785.

Bolger, A.M., Lohse, M., Usadel, B., 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30, 2114-2120.

Brown, S.P., Veach, A.M., Rigdon-Huss, A.R., Grond, K., Lickteig, S.K., Lothamer, K., Oliver, A.K., Jumpponen, A., 2015. Scraping the bottom of the barrel: are rare high throughput sequences artifacts? Fungal Ecology 13, 221-225.

Callahan, B.J., McMurdie, P.J., Rosen, M.J., Han, A.W., Johnson, A.J.A., Holmes, S.P., 2016. DADA2: high-resolution sample inference from Illumina amplicon data. Nature Methods 13, 581.

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

Caporaso, J.G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F.D., Costello, E.K., Fierer, N., Peña, A.G., Goodrich, J.K., Gordon, J.I., 2010. QIIME allows analysis of high-throughput community sequencing data. Nature Methods 7, 335-336.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27019v1 | CC BY 4.0 Open Access | rec: 3 Jul 2018, publ: 3 Jul 2018

Clarke, K., Gorley, R., 2006. PRIMER V6: User manual/tutorial. Primer-E Ltd. Plymouth, 192pp.

Cline, L.C., Song, Z., Al- Ghalith, G.A., Knights, D., Kennedy, P.G., 2017. Moving beyond de novo clustering in fungal community ecology. New Phytologist.

Deshpande, V., Wang, Q., Greenfield, P., Charleston, M., Porras-Alfaro, A., Kuske, C.R., Cole, J.R., Midgley, D.J., Tran-Dinh, N., 2016. Fungal identification using a Bayesian classifier and the Warcup training set of internal transcribed spacer sequences. Mycologia 108, 1-5.

Edgar, R.C., 2013. UPARSE: highly accurate OTU sequences from microbial amplicon reads. Nature Methods 10.

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.

Frøslev, T.G., Kjøller, R., Bruun, H.H., Ejrnæs, R., Brunbjerg, A.K., Pietroni, C., Hansen, A.J., 2017. Algorithm for post-clustering curation of DNA amplicon data yields reliable biodiversity estimates. Nature communications 8, 1188.

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.

Gdanetz, K., Benucci, G.M.N., Pol, N.V., Bonito, G., 2017. CONSTAX: a tool for improved taxonomic resolution of environmental fungal ITS sequences. BMC Bioinformatics 18, 538.

Grossart, H.-P., Wurzbacher, C., James, T.Y., Kagami, M., 2016. Discovery of dark matter fungi in aquatic ecosystems demands a reappraisal of the phylogeny and ecology of zoosporic fungi. Fungal Ecology 19, 28-38.

Gweon, H.S., Oliver, A., Taylor, J., Booth, T., Gibbs, M., Read, D.S., Griffiths, R.I., 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.

Hibbett, D., Abarenkov, K., Koljalg, U., Opik, M., Chai, B., Cole, J.R., Wang, Q., Crous, P.W., Robert, V.A.R.G., Helgason, T., Herr, J., Kirk, P., Lueschow, S., O'Donnell, K., Nilsson, H., Oono, R., Schoch, C.L., Smyth, C., Walker, D., Porras-Alfaro, A., Taylor, J.W., Geiser, D.M., 2017. Sequence-based classification and identification of Fungi. Mycologia 108, 1049-1068.

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

Kõljalg, U., Nilsson, R.H., Abarenkov, K., Tedersoo, L., Taylor, A.F., Bahram, M., et al., 2013. Towards a unified paradigm for sequence- based identification of fungi. Molecular ecology, 22(21), pp.5271-5277.

Majaneva, M., Hyytiäinen, 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.

R-Core-Team, 2015. R Foundation for Statistical Computing; 2015. R: A language and environment for statistical computing.

Richardson, R.T., Bengtsson-Palme, J., Johnson, R.M., 2017. Evaluating and optimizing the performance of software commonly used for the taxonomic classification of DNA metabarcoding sequence data. Molecular Ecology Resources 17, 760-769.

Rognes, T., Flouri, T., Nichols, B., Quince, C., Mahé, F., 2016. VSEARCH: a versatile open source tool for metagenomics. PeerJ 4, e2584.

Saary, P., Forslund, K., Bork, P., Hildebrand, F., 2017. RTK: efficient rarefaction analysis of large datasets. Bioinformatics 33, 2594-2595.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27019v1 | CC BY 4.0 Open Access | rec: 3 Jul 2018, publ: 3 Jul 2018

Schloss, P.D., Westcott, S.L., Ryabin, T., Hall, J.R., Hartmann, M., Hollister, E.B., Lesniewski, R.A., Oakley, B.B., Parks, D.H., Robinson, C.J., Sahl, J.W., Stres, B., Thallinger, G.G., Van Horn, D.J., 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.

Schoch, C.L., Seifert, K.A., Huhndorf, S., Robert, V., Spouge, J.L., Levesque, C.A., Chen, W., et al., 2012. Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. Proceedings of the National Academy of Sciences of the United States of America 109, 6241-6246.

Sinha, R., Abu-Ali, G., Vogtmann, E., Fodor, A.A., Ren, B., Amir, A., Schwager, E., Crabtree, J., Ma, S., Abnet, C.C., 2017. Assessment of variation in microbial community amplicon sequencing by the Microbiome Quality Control (MBQC) project consortium. Nature Biotechnology.

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. Appl Env Microbiol 73.

Větrovský, T., Baldrian, P., Morais, D. 2018. SEED 2: a user-friendly platform for amplicon high-throughput sequencing data analyses. Bioinformatics 34: 2292-2294.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27019v1 | CC BY 4.0 Open Access | rec: 3 Jul 2018, publ: 3 Jul 2018

Table 1. Used software, sequence and OTU counts (values in bold) by **a**) Illumina and **b**) PacBio analysis platforms.

a

	LotusS	Qiime2	PipeCraft	Galaxy	PIPITS
Raw reads	7,981,812[a]	7,335,838[b]	7,981,812[a]	7,981,812[a]	7 335 838[b]
Assembly	FLASH/ NA	DADA2/ NA	VSEARCH/ 7,511,274	FASTQ joiner/ 7,911,554	VSEARCH/ 7,198,094
Quality filtering	sdm/ NA	DADA2/ 5,428,563	VSEARCH/ 7,511,274	trimmomatic/ 7,879,960	fastqx/ 7,142,354
Demultiplexing	sdm/ 6,727,631	NP	mothur/ 6,558,772	mothur/ 1,643,879	NP
Chimera filtering	USEARCH/ 6,486,802	NP	VSEARCH/ 6,300,085	VSEARCH/ 1,621,330	VSEARCH/ NA
ITS extractor	5,919,084	NP	6,262,000	NP	6,401,097
Clustering (OTUs)	UPARSE/ **8,659**	VSEARCH/ **7,477**	UPARSE/ **7,598**	VSEARCH/ **106,245**	VSEARCH/ **7,887**

b

	LotusS	PipeCraft	Galaxy
CCS[c] reads	720,222[a]	720,222[a]	720,222[a]
Quality filtering	sdm/ NA	VSEARCH/ 462,010	trimmomatic/ 672,292
Demultiplexing	sdm/ 258,085	mothur/ 380,722	mothur/ 457,173
Chimera filtering	USEARCH/ 255,746	VSEARCH/ 341,154	VSEARCH/ 405,025
ITS extraction	192,485	338,150	NP
Clustering (OTUs)	UPARSE/ **942**	UPARSE/ **4,176**	VSEARCH/ **8,854**

[a]multiplexed input data; [b]demultiplexed input data; [c]circular consensus sequences; NA: indicate not available; NP: not performed. Singleton OTUs were excluded from the counts.

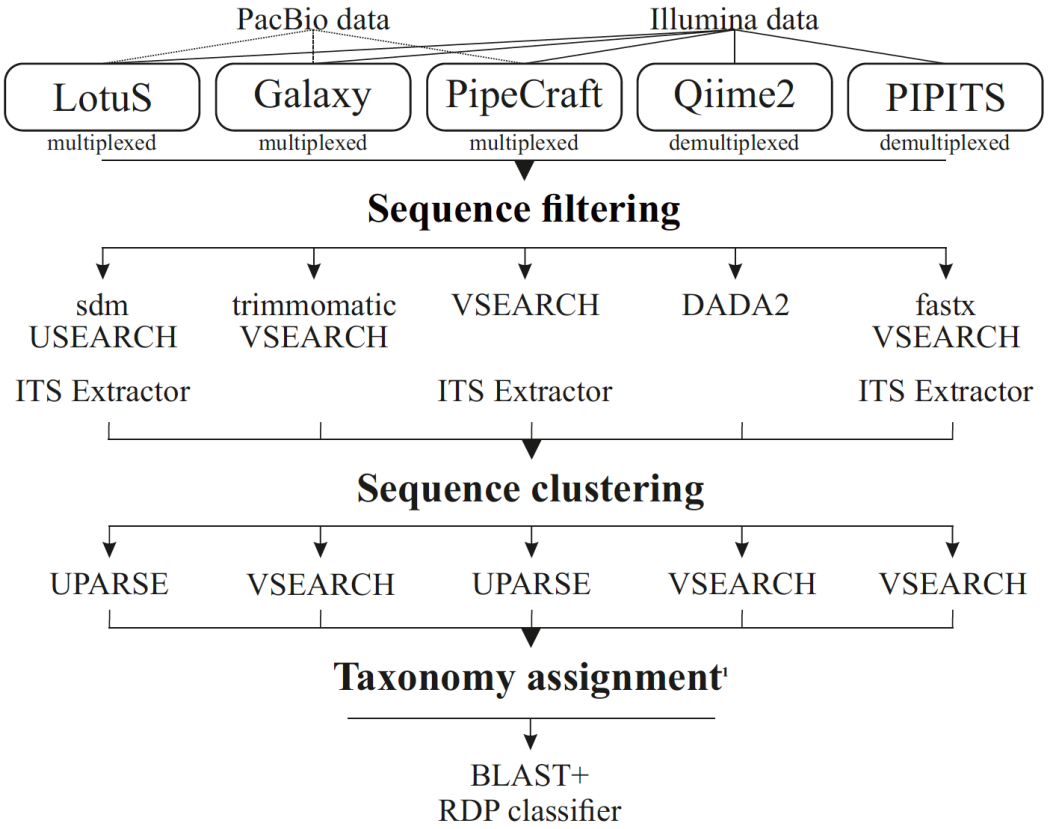
PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27019v1 | CC BY 4.0 Open Access | rec: 3 Jul 2018, publ: 3 Jul 2018

358 **Figures**

359

360

361



362

363 **Figure 1.** Outline of workflow in different analysis pipelines. [1]Taxonomy assignment was

364 performed outside the listed pipelines.

365

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27019v1 | CC BY 4.0 Open Access | rec: 3 Jul 2018, publ: 3 Jul 2018

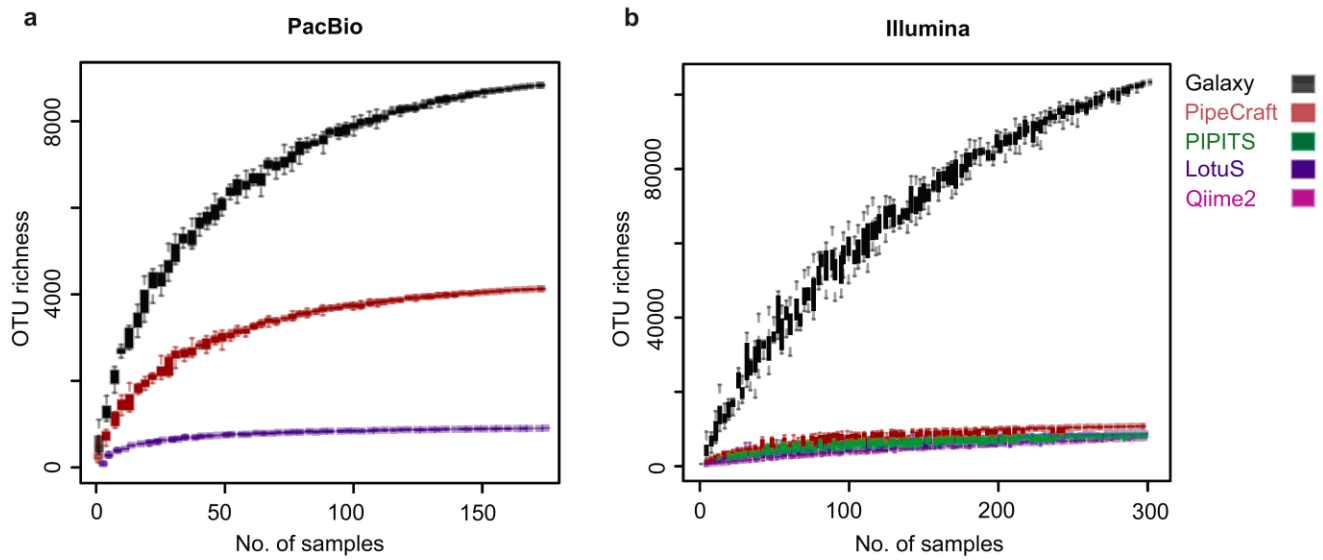


Figure 2. OTU accumulation curves of the evaluated pipelines for a) PacBio and b) Illumina data sets.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27019v1 | CC BY 4.0 Open Access | rec: 3 Jul 2018, publ: 3 Jul 2018

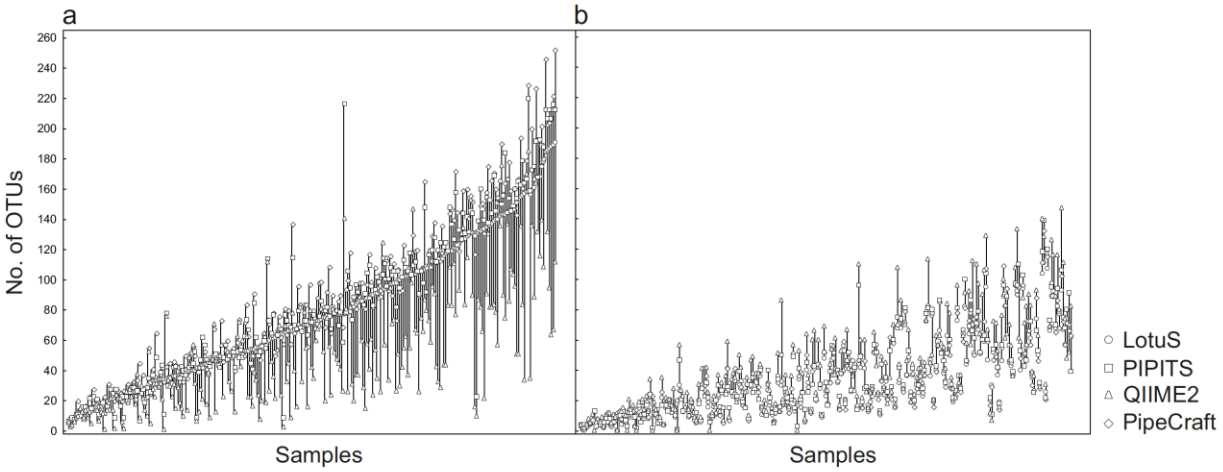


Figure 3. Number of OTUs per sample for Illumina data recorded from a) pipeline-generated OTU tables (median differences = 38 OTUs) and from b) filtered OTU tables (median differences = 12 OTUs). The Galaxy workflow was excluded because of the several orders of magnitude higher number of generated OTUs.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27019v1 | CC BY 4.0 Open Access | rec: 3 Jul 2018, publ: 3 Jul 2018