Hundo: a Snakemake workflow for microbial community sequence data

Joseph Brown[1], **Nicole Zavoshy**[1], **Colin J Brislawn**[1], **and Lee Ann McCue**[1]

[1]**Earth and Biological Sciences Directorate, Pacific Northwest National Laboratory, Richland, WA USA**

Corresponding author:
Joseph Brown[1]

Email address: joseph.brown@pnnl.gov

ABSTRACT

Hundo is a software package that performs quality control and annotation of ribosomal RNA and internal transcribed spacer sequence reads. It provides a comprehensive suite of tools and options that aim to reduce the effort of performing robust sequence annotation while obtaining the maximum amount of data provenance to ensure replicability. The software package performing annotation is implemented in Python, the analytical workflow is implemented in Snakemake, and dependencies are installed via Bioconda. Extensive documentation and the full source code are available under the MIT license at: https://github.com/pnnl/hundo.

INTRODUCTION

Many packages exist for processing 16S ribosomal RNA (rRNA) and internal transcribed spacer (ITS) amplicon sequence reads, each with nuances pertaining to quality control, annotation, output, and importantly, usage and install requirements. Existing packages like QIIME (Caporaso et al., 2010), QIIME 2 (QQIIME 2 Development Team, 2018), and mothur (Schloss et al., 2009) offer vast toolsets to carry out analyses, often resulting in over-complication and workflow fragility arising from incompatible file formats required at various stages of complex and ad hoc, user-specified workflows.

When designing a workflow, one of the primary considerations is how to account for user choice and how much flexibility to allow. Existing annotation and analysis packages like QIIME 2 account for user preferences by providing both pipelines and individual methods within those pipelines. In addition, these tools, including QIIME 2 and mothur, provide many additional tools for downstream analysis after sequences have been annotated. By accounting for all use cases and by adding additional analytical and visualization tools, the commands and subcommands for each of these packages increase in number and complexity. For instance, QIIME lists 150 individual commands, QIIME 2 has 20 sub-commands and 119 sub-sub-commands, and mothur lists 146 commands.

The documentation for QIIME2 and mothur are both exceptionally detailed; include comprehensive installation instructions; offer multiple tutorials, each walking the reader through a slew of subcommands where the inputs for the current task are dependent upon the outputs of the previous; and provide example datasets for users to test workflows on their own installation. While detailed documentation is useful for experts, it can be daunting for first-time users as they may not know which settings are important and how altering parameters may affect the outcome.

We introduce Hundo, a 16S, 18S, and ITS annotation workflow. It incorporates a minimal set of tools to carry out quality control and annotation while adding important aspects of environment and data provenance that existing software packages lack. By reducing the toolset and streamlining processing from raw sequence data to annotated (representative) sequences down to a single command, we lessen the burden of understanding how and where to get started, what was performed to the sequence reads, and what outputs were generated. User flexibility is offered through command line options that affect individual commands or change commands within the workflow, like swapping VSEARCH (Rognes et al., 2016) for alignment rather than BLAST (Altschul et al., 1997). We provide documentation, example data,

outputs, and a sample report, to facilitate rapid evaluation and adoption of our software. By restricting use cases to quality control and annotation, and by taking advantage of cloud services and a workflow manager, Hundo represents a well-documented, lightweight software package capable of performing robust rRNA sequence annotation.

APPROACH

Hundo is a combination of a command line application written in Python 3 and Snakemake (Köster and Rahmann, 2012). The command line interface (CLI) works to parse options and format the sub-command to Snakemake and only requires specifying a directory containing paired-end FASTQ files.

Hundo provides environment definitions which are used by Snakemake to install tested versions of binaries from Bioconda (Grüning et al., 2018) before starting to execute sequence processing tasks defined in the workflow. In addition to handling prerequisite software installs, Hundo's version controlled reference annotations (Brown, 2017) are downloaded during execution from Zenodo (Zenodo, 2018).

All output files generated from the workflow have associated provenance data including attributes like the command used in their generation, the input file, and the date and time of execution. Additionally, Hundo builds upon this metadata by compiling a comprehensive HTML report that includes quality control statistics, interactive plots with taxonomic breakdowns across samples, the workflow's configuration, the environment definitions, and explanations of output file contents. Providing comprehensive provenance increases replicability, which is why both Hundo and QIIME2 include it by default.

RESULTS

Hundo is comprised of custom methods as well as wrappers for existing software. Sequence processing is initiated by calling `hundo annotate`. Command line options are parsed, validated, and sent to the accompanying workflow. Paired-end sequence reads are trimmed by quality and are optionally filtered of sequence content like PhiX, a known contaminant among bacterial assemblies (Mukherjee et al., 2015), using BBduk of the BBMap toolset (Bushnell, 2018). Passing reads are merged using VSEARCH then aggregated into a single FASTA file with headers describing the origin and count of the sequences. Merged sequence reads are evaluated by their expected error rates and filtered a final time before clustering into operational taxonomic units (OTU).

Within Hundo, clusters are created from de-replicated merged sequences that have passed quality control using VSEARCH. Sequences are pre-clustered into centroids using VSEARCH to accelerate chimera filtering. Chimera filtering is completed in two steps: *de novo* and then reference-based. Following chimera filtering, sequences are placed into clusters using distance-based, greedy clustering with VSEARCH based on the allowable percent difference of the configuration.

After OTU sequences have been determined, BLAST or VSEARCH is used to align sequences to the reference database. Sequence alignments are filtered relative to the best hit before the least common ancestor (LCA) is calculated. Reference databases for 16S were curated by the CREST team and Hundo incorporates the CREST LCA method (Lanzén et al., 2012) for 16S. ITS reference databases and taxonomies are maintained by UNITE (Kõljalg et al., 2013).

For downstream processing, counts are assigned to OTUs using the global alignment method of VSEARCH, which outputs the final feature-abundance table as a tab-delimited text file. The Biom tool (McDonald et al., 2012) is used to convert the tab-delimited table to biom. Multiple alignment of sequences is completed using MAFFT (Nakamura et al., 2018) in order to generate a tree based on the OTU sequences using FastTree2 (Price et al., 2010).

Hundo finalizes processing by generating a comprehensive HTML report (Figure 1) with summary tables, interactive plots, explanation of the workflow, details of output files, and an embedded archive containing the most relevant output files. The summary table includes counts per sample input, quality controlled count, merged, and finally the count assigned to OTUs. Shannon, Simpson, and inverse Simpson indexes are included alongside the counts. Sequence quality plots report qualities per base per sequence before and after filtering. Observed taxonomies per sample are plotted by absolute count and relative abundance across phylum, class, and order to facilitate a cursory analysis of taxonomic composition across an experiment. To ease replication, the report includes the user's configuration and environment setup with absolute versions of applications used in their protocol. The report concludes by

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27272v1 | CC BY 4.0 Open Access | rec: 10 Oct 2018, publ: 10 Oct 2018

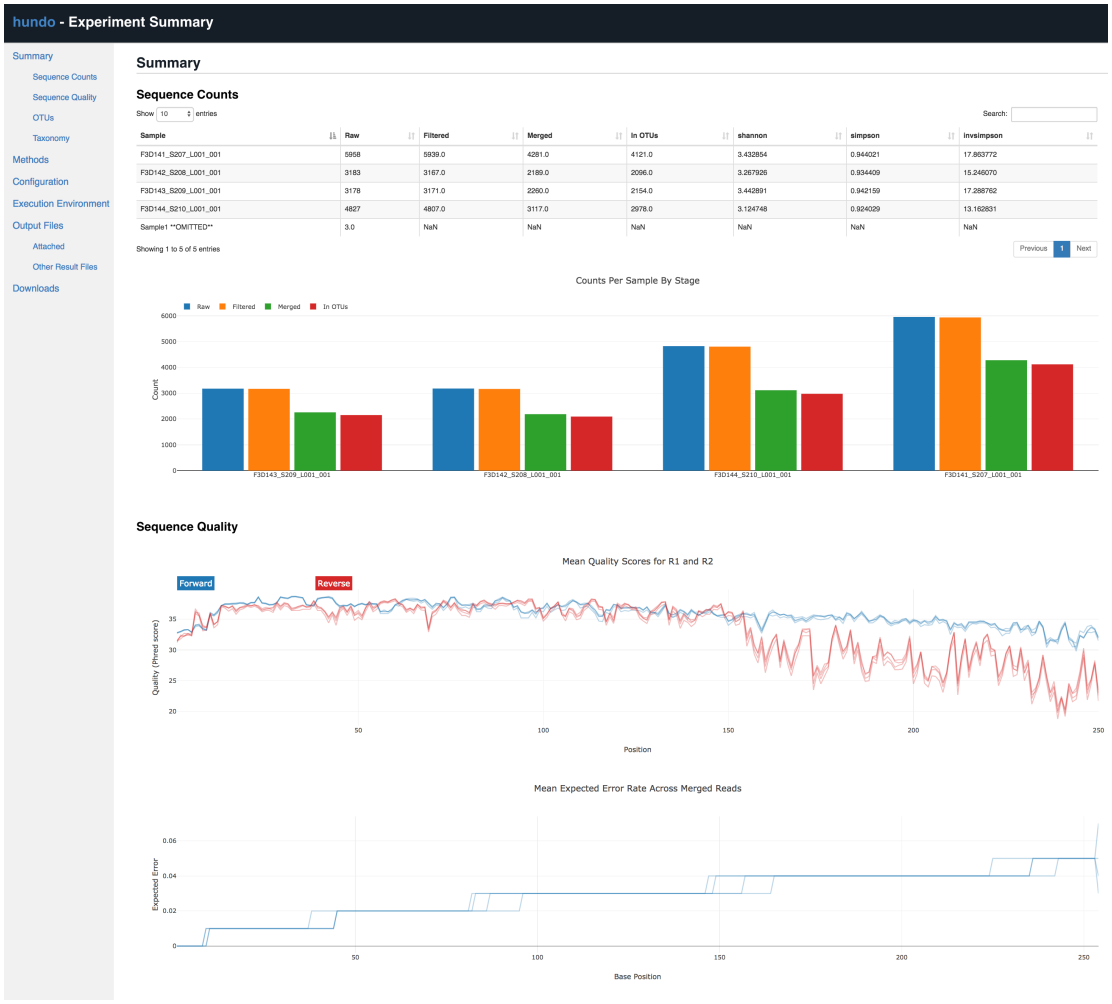

Figure 1. Example HTML report generated by Hundo. An interactive version is available at: https://pnnl.github.io/hundo/.

defining output files and providing a link to download a subset of output files that are most relevant for use in dedicated analysis packages like VEGAN (Dixon, 2003) or phyloseq (McMurdie and Holmes, 2013).

DISCUSSION

Using 16S rRNA sequence data from Dohnalkova et al. (Dohnalkova et al., 2017), we aimed to reproduce their findings among rhisospheric microbial communities of *Pinus resinosa* grown in controlled, laboratory environments, using Hundo and Phyloseq. After observing complex extracellular soil organic matter forming associations with minerals, Dohnalkova et al. performed 16S amplicon sequencing on 30 samples to identify microbes capable of contributing to the mineral weathering process and those capable of nitrogen fixation.

Hundo was installed and executed as `hundo annotate` across paired-end sequence data with filtering for Illumina adapters and PhiX contamination, SILVA rRNA reference database for taxonomic annotation and chimera removal, and both `--jobs` and `--threads` set to 24. Hundo generated the feature abundance table and phylogenetic tree necessary for downstream processing in 12 minutes with progress messages, a clear message of a successful run, and no additional user input. The feature abundance table and taxonomic tree were imported into Phyloseq to plot abundances and their phylogeny.

Our analysis of the microbial communities replicates and confirms taxonomic composition of microbial communities among the soil rhizosphere includes high percentages of *Burkholderiales* and *Rhizobiales* as well as unidentified microbes phylogenetically similar to *Leptospiraceae*. Relative to the published work,

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27272v1 | CC BY 4.0 Open Access | rec: 10 Oct 2018, publ: 10 Oct 2018

our replication study shows that among OTUs classified as *Rhizobiales*, we see an increase in genus level assignments which could allow finer interpretation among these community members. These data, our work, and recreated figures are publicly available in a version controlled code and data archive (Brislawn, 2018).

CONCLUSION

Hundo efficiently processes raw paired-end rRNA and ITS sequence data from a single command into standardized outputs. By tailoring its functionality to a well-defined workflow and by using Snakemake to handle dependency installs and track provenance, Hundo serves as a tool for replicable 16S, 18S, and ITS sequence processing.

FUNDING

This research was supported by the Microbiomes in Transition Initiative LDRD Program at the Pacific Northwest National Laboratory, a multi-program national laboratory operated by Battelle for the DOE under Contract DE-AC06-76RL01830.

REFERENCES

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research*, 25(17):3389–3402.

Brislawn, C. J. (2018). Minimal replication of Dohnalkova 2017 using Hundo. *Open Science Framework*. DOI: 10.17605/OSF.IO/T4NC7.

Brown, J. (2017). 16S and ITS Reference Annotation Databases [Data set]. *Zenodo*. DOI: 10.5281/zenodo.1043977.

Bushnell, B. (2018). BBMap. https://sourceforge.net/projects/bbmap/, last accessed on 09/08/2018.

Caporaso, J Gregoand Huttley, G. A., Kelley, S. T., Knights, D., Koenig, J. E., Ley, R. E., Lozupone, C. A., McDonald, D., Muegge, B. D., Pirrung, M., Reeder, J., Sevinsky, J. R., Turnbaugh, P. J., Walters, W. A., Widmann, J., Yatsunenko, T., Zaneveld, J., and Knight, R. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nature methods*, 7(5):335–336.

Dixon, P. (2003). VEGAN, a package of R functions for community ecology. *Journal of Vegetation Science*, 14(6):927–930.

Dohnalkova, A. C., Tfaily, M. M., Smith, A. P., Chu, R. K., Crump, A. R., Brislawn, C. J., Varga, T., Shi, Z., Thomashow, L. S., Harsh, J. B., et al. (2017). Molecular and microscopic insights into the formation of soil organic matter in a red pine rhizosphere. *Soils*, 1(1):4.

Grüning, B., Dale, R., Sjödin, A., Chapman, B. A., Rowe, J., Tomkins-Tinch, C. H., Valieris, R., Köster, J., and Bioconda Team (2018). Bioconda: sustainable and comprehensive software distribution for the life sciences. *Nature methods*, 15(7):475–476.

Kõljalg, U., Nilsson, R. H., Abarenkov, K., Tedersoo, L., Taylor, A. F. S., Bahram, M., Bates, S. T., Bruns, T. D., Palme, J. B., Callaghan, T. M., Douglas, B., Drenkhan, T., Eberhardt, U., Dueñas, M., Grebenc, T., Griffith, G. W., Hartmann, M., Kirk, P. M., Kohout, P., Larsson, E., Lindahl, B. D., Lücking, R., Martín, M. P., Matheny, P. B., Nguyen, N. H., Niskanen, T., Oja, J., Peay, K. G., Peintner, U., Peterson, M., Põldmaa, K., Saag, L., Saar, I., Schüßler, A., Scott, J. A., Senés, C., Smith, M. E., Suija, A., Taylor, D. L., Telleria, M. T., Weiss, M., and Larsson, K. H. (2013). Towards a unified paradigm for sequence-based identification of fungi. *Molecular Ecology*, 22(21):5271–5277.

Köster, J. and Rahmann, S. (2012). Snakemake–a scalable bioinformatics workflow engine. *Bioinformatics (Oxford, England)*, 28(19):2520–2522.

Lanzén, A., Jørgensen, S. L., Huson, D. H., Gorfer, M., Grindhaug, S. H., Jonassen, I., Øvreås, L., and Urich, T. (2012). CREST–classification resources for environmental sequence tags. *PloS one*, 7(11):e49334.

McDonald, D., Clemente, J. C., Kuczynski, J., Rideout, J. R., Stombaugh, J., Wendel, D., Wilke, A., Huse, S., Hufnagle, J., Meyer, F., Knight, R., and Caporaso, J. G. (2012). The Biological Observation Matrix (BIOM) format or: how I learned to stop worrying and love the ome-ome. *GigaScience*, 1(1):7.

McMurdie, P. J. and Holmes, S. (2013). phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PloS one*, 8(4):e61217.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27272v1 | CC BY 4.0 Open Access | rec: 10 Oct 2018, publ: 10 Oct 2018

Mukherjee, S., Huntemann, M., Ivanova, N., Kyrpides, N. C., and Pati, A. (2015). Large-scale contamination of microbial isolate genomes by Illumina PhiX control. *Standards in genomic sciences*, 10:18.

Nakamura, T., Yamada, K. D., Tomii, K., and Katoh, K. (2018). Parallelization of MAFFT for large-scale multiple sequence alignments. *Bioinformatics (Oxford, England)*, 34(14):2490–2492.

Price, M. N., Dehal, P. S., and Arkin, A. P. (2010). FastTree 2–approximately maximum-likelihood trees for large alignments. *PloS one*, 5(3):e9490.

QQIIME 2 Development Team (2018). QIIME 2. https://qiime2.org/, last accessed on 09/08/2018.

Rognes, T., Flouri, T., Nichols, B., Quince, C., and Mahé, F. (2016). VSEARCH: a versatile open source tool for metagenomics. *PeerJ*, 4(17):e2584.

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., and Weber, C. F. (2009). Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and Environmental Microbiology*, 75(23):7537–7541.

Zenodo (2018). Zenodo - Research. Shared. https://zenodo.org/, last accessed on 09/08/2018.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27272v1 | CC BY 4.0 Open Access | rec: 10 Oct 2018, publ: 10 Oct 2018