MetaWorks: A flexible, scalable bioinformatic pipeline for high-throughput multi-marker biodiversity assessments

Teresita M. Porter[1*], Mehrdad Hajibabaei[1]

[1] Centre for Biodiversity Genomics @ Biodiversity Institute of Ontario & Department of Integrative Biology, University of Guelph, Guelph, ON, CANADA

* Corresponding author:

T.M. Porter

terrimporter@gmail.com

Abstract

Multi-marker metabarcoding is increasingly being used to generate biodiversity information across different domains of life from microbes to fungi to animals such as for molecular ecology and biomonitoring applications in different sectors from academic research to regulatory agencies and industry. Current popular bioinformatic pipelines support microbial and fungal marker analysis, while ad hoc methods are often used to process animal metabarcode markers from the same study. MetaWorks provides a harmonized processing environment, pipeline, and taxonomic assignment approach for demultiplexed Illumina reads for all biota using a wide range of metabarcoding markers such as 16S, ITS, and COI. A Conda environment is provided to quickly gather most of the programs and dependencies for the pipeline. Several workflows are provided such as: taxonomically assigning exact sequence variants, provides an option to generate operational taxonomic units, and facilitates single-read processing. Pipelines are automated using Snakemake to minimize user intervention and facilitate scalability. All pipelines use the RDP classifier to provide taxonomic assignments with confidence measures. We extend the functionality of the RDP classifier for taxonomically assigning 16S (bacteria), ITS (fungi), and 28S (fungi), to also support COI (animals), rbcL (eukaryotes, land plants, diatoms), 12S (fish), 18S (eukaryotes, diatoms) and ITS (fungi, plants). MetaWorks properly handles ITS by trimming flanking conserved rRNA gene regions as well as protein coding genes by providing two options for removing obvious pseudogenes. MetaWorks is available at https://github.com/terrimporter/MetaWorks

along with quick-start instructions using test data, detailed workflow descriptions, and a tutorial for new users.

Keywords

Metabarcoding, Conda, Snakemake, ITS extraction, pseudogene removal, rRNA, ITS, COI, rbcL, 12S

Background

Marker gene sequencing, metabarcoding, or metasystematics are interrelated techniques that involves extracting DNA from bulk samples such as soil, water, or mixtures of individuals collected from traps, without having to isolate or identify individual specimens, followed by enrichment of a signature DNA region to identify biological community composition using bioinformatics [1–3]. In microbial ecology to animal biodiversity studies, different signature DNA regions are targeted for their ability to identify target taxa. For example, in prokaryotes, the 16S small subunit (SSU) ribosomal RNA (rRNA) region is often used for genus level taxonomic assignments [4,5]. Other popular markers include cytochrome c oxidase (COI) for animals; ribulose bisphosphate large subunit (rbcL) for plants and diatoms; the internal transcribed spacer (ITS) for fungi and plants; 18S SSU for eukaryotes, arbuscular mycorrhizal fungi, and diatoms; and 12S mitochondrial SSU for fish [6–14].

Existing pipelines such as QIIME2 and DADA2 were initially developed to support the microbial ecology community [15,16]. In comprehensive, multi-trophic,

multi-marker studies, there is a need for a pipeline that can handle rRNA genes, spacer regions, as well as protein-coding markers in a single harmonized environment [17,18]. For the ITS region, we needed a pipeline that could remove the conserved flanking rRNA genes as this has been shown to improve taxonomic assignment accuracy [19]. For protein-coding regions, we needed a pipeline that could remove putative pseudogenes [20–23]. We also wanted the ability to generate high quality exact sequence variants (ESVs) for any marker (not just 16S or ITS) for the additional level of genetic and taxonomic resolution SVs can provide [24–26]. For taxonomic assignment, we wanted to use a classifier that would provide a measure of confidence for assignments to reduce false-positive assignments [27–29].

As multi-marker studies are carried out on phylogenetically divergent taxa, such as in biodiversity or trophic studies, there is a need for more generic pipelines where different markers can be analyzed using similar dataflows with 3rd party programs instead of being limited to database-specific pipelines and tools [17,30]. We developed MetaWorks with the following objectives: 1) reproducibility with respect to the computational environment used as well as the pipeline itself, 2) scalability to leverage high performance computer clusters to speed up the analysis of large datasets, 3) naive Bayes classifier support for popular metabarcode markers; and 4) to support marker-specific processing steps such as ITS extraction and pseudogene-removal for protein-coding markers. MetaWorks was designed for data analysts who are comfortable using Linux command-line tools but would like a single harmonized environment and pipeline to process multi-marker metabarcode datasets.

Implementation & Workflow

Implementation

MetaWorks is a multi-marker 'meta'-barcode pipeline that does 'the works' by supporting the bioinformatic processing of popular markers including rRNA genes, spacers, and protein coding genes generating taxonomically assigned sequence variants or operational taxonomic units (OTUs). To facilitate reproducibility, scalability, and shareability of workflows we use the Conda package manager to facilitate the download of most programs and dependencies and the Snakemake workflow manager to automate pipelines and utilize computational resources efficiently [31–33]. Snakemake supports re-entrancy and automatic deployment of multiple parallel jobs, both ideal for high performance computing environments where many cores are available to speed up the analysis of large datasets.

We provided instructions on how to install and use Conda in the online documentation. One additional program not available as a Conda package, ORFfinder, may need to be downloaded separately if pseudogene-filtering will be conducted and instructions are provided in the online documentation. MetaWorks is available from https://github.com/terrimporter/MetaWorks and a suite of trained classifiers for taxonomic assignment are also available from GitHub (Table 1). Depending on the DNA metabarcode marker(s) the user will be processing, these can be individually downloaded from GitHub and instructions are provided in the online documentation.

Table 1. RDP-trained reference sets that can be used with MetaWorks.

Marker	Target taxa	Classifier availability	Number of	Number of	Source data

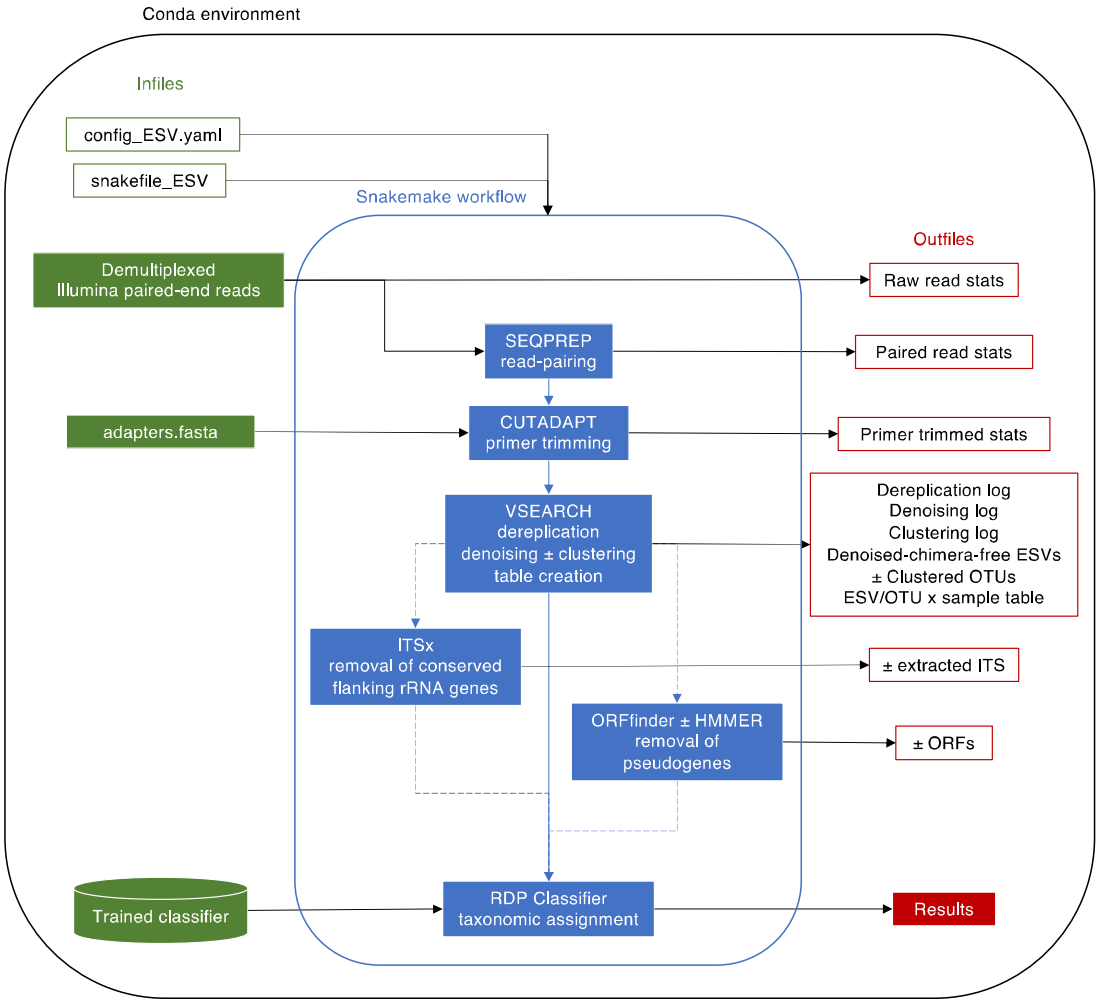
			included sequences	**included taxa at all ranks (species)**	
COI	Eukaryotes	https://github.com/terrimporter/CO1Classifier	1,221,528	154,351 (114,687)	BOLD [34], INSDC [35]
rbcL	Diatoms	https://github.com/terrimporter/rbcLdiatomClassifier	3,504	1,432 (1,023)	Diat.barcode [36]
rbcL	Land plants	https://github.com/terrimporter/rbcL_landPlant_Classifier	148,258	61,398 (50,778)	INSDC [35]
rbcL	Eukaryotes	https://github.com/terrimporter/rbcLClassifier	164,454	65,742 (53,344)	INSDC [35]
12S	Fish	https://github.com/terrimporter/12SfishClassifier	2,853	4,751 (2,833)	MitoFish [37]
12S	Vertebrates	https://github.com/terrimporter/12SvertebrateClassifier	10,654	15,007 (9,564)	INSDC [35] and MitoFish [37]
SSU (18S)	Diatoms	https://github.com/terrimporter/SSUdiatomClassifier	2,962	1,198 (828)	Diat.barcode [36]
SSU (16S)	Vertebrates	https://github.com/terrimporter/16SvertebrateClassifier	72,195	21,282 (15,155)	INSDC [35]
SSU (18S)	Eukaryotes	https://github.com/terrimporter/18SClassifier	42,301	7,504 (5,440 genera)	SILVA [38]
SSU (16S)	Prokaryotes	Built-in to the RDP classifier*	13,212	3,247 (2,506 genera)	RDP [5]
ITS	Fungi (Warcup)	Built-in to the RDP classifier	17,878	10,621 (8,551)	Deshpande et al., 2016 [39]
ITS	Fungi (UNITE 2014)	Built-in to the RDP classifier	145,019	23,222 (20,337)	Abarenkov et al., 2010 [40]
ITS	Fungi (UNITE 2021)	https://github.com/terrimporter/UNITE_ITSClassifier	1,393,203	376,167 (352,588)	UNITE [40]
ITS	Plants	https://github.com/terrimporter/PLANiTS_ITSClassifier	104,387	72,632 (61,693)	PLANiTS [41] and UNITE [40]
LSU	Fungi	Built-in to the RDP classifier	11,442	2,633 (1,895)	Liu et al., 2012 [42]

Workflow

The pipeline begins with demultiplexed Illumina paired-end reads as this is the format most often provided by sequencing centres to their clients. Several workflows are available as Snakemake pipelines such as taxonomic assignment of sequence

variants (Fig 1), clustering of sequence variants into OTUs, or for processing single reads. For each of these workflows described below, parameter settings for each bioinformatic step can be customized in the config.yaml file. The user also needs to provide a file of primer sequences so we provide a template for the adapters.fasta file as well as a small set of raw Illumina sequences for the COI amplicon that can be used to test the installation. The online documentation provides a 'quick-start' example using the provided COI test data. We also provide a more detailed tutorial in the online documentation that walks users through the steps necessary to set up their environment to run the pipeline for the first time, assuming the user has never worked with Conda or Snakemake before.

Figure 1. MetaWorks workflow to produce taxonomically assigned sequence variants. To aid reproducibility, a Conda environment is provided (environment.yml). Although multiple Snakemake workflows are provided in MetaWorks, here we show the workflow to produce taxonomically assigned sequence variants. User input is shown in green on the left, the ESV workflow is shown in the middle in blue, and outfiles are shown in red on the right. The results are provided in a comma-separated value (CSV) file and shows each ESV/OTU per sample with read counts and taxonomic assignments. Abbreviations: internal transcribed spacer (ITS) region, open reading frame FASTA-formatted sequences (ORFs).



132

133

134 *Generation and taxonomic assignment of sequence variants*

135 The ESV workflow will run the pipeline shown in Fig 1. In the config_ESV.yaml
136 file, users indicate the path to the directory that contains the demultiplexed Illumina
137 paired-end reads, specify the unique part of filenames to distinguish between samples
138 and reads, and specify the name of the directory that will contain the outfiles. Default

settings for each program are provided in the config_ESV.yaml file but these can be customized by the user. SEQPREP was initially chosen for pairing the forward and reverse reads, because the program comes with the option to output the alignments for visual inspection, an option that most read-pairing programs do not have [43]. For SEQPREP read-pairing, users can specify a Phred score quality cutoff, the minimum overlap between the forward and reverse reads, the maximum fraction of mismatches allowed in the overlap region, and the minimum fraction of matching bases in the overlap region. CUTADAPT was chosen for primer-trimming because it is fast and already widely used in the metabarcoding community for this purpose, so most users will likely already be familiar with how this program works [44]. For CUTADAPT, users need to provide a FASTA-formatted primer sequence file (adapters.fasta), they can also specify the minimum sequence length to retain after primer-trimming, a Phred quality score cutoff, the maximum error rate, minimum adapter overlap, and maximum number of ambiguous bases allowed. VSEARCH was chosen to dereplicate reads (retain unique reads) and remove artefactual sequences using the UNOISE3 and UCHIME3 algorithms [45,46]. We chose the open-source VSEARCH program over alternatives because the program can utilize all the available memory on a system, facilitating the analysis of large datasets on high performance computer systems. We prefer the UNOISE3 method for denoising because it performs up to 1,200 faster and uses less memory than other denoising programs [47]. To map read counts to the newly generated denoised-chimera ESVs to create an ESV x sample table, we use the 'search_exact' method because it is faster and optimized to find exact matches

compared with the 'usearch_global' command with the 'id 1.0' parameter, but this is just an intermediate step and further filtering of this table is performed by MetaWorks.

If the internal transcribed spacer (ITS) region is analyzed, then the pipeline uses the ITSx program to trim away the flanking conserved rRNA gene sequences so that taxonomic assignment is based solely on the variable spacer region sequences (ITS1 or ITS2) [19]. This step has been shown to improve sensitivity of clustering and taxonomic assignments [19].

If a protein coding marker is being processed, the user can select a pseudogene-removal method in the config_ESV.yaml file. We have previously described two methods for removing putative pseudogenes from DNA barcode and metabarcoding datasets [21]. The NCBI ORFfinder program is used to translate reads into all possible open reading frames (ORFs). The first pseudogene removal method retains the longest ORF for each read, calculates a distribution of ORF lengths, and removes reads with outlier lengths as putative ORFs. The second pseudogene removal method can be used if a hidden Markov model is available and is provided for processing COI arthropods. The longest ORFs are compared to the profile using HMMER available from http://hmmer.org . MetaWorks calculates a distribution of bit scores and removes reads with short outlier bit scores as putative ORFs. Removing noise caused by the sequencing of pseudogenes in metabarcode datasets can help users avoid over-estimating richness in subsequent analyses, yet this step is not included in the most popular metabarcode pipelines as they were developed to support the analysis of rRNA genes where this is not a problem.

One of the features of MetaWorks, is the use of a single taxonomic assignment method for any metabarcode marker that provides a measure of confidence for taxonomic assignments. We chose the RDP Classifier for this task as this method has a long-history of use in the microbial ecology literature, additionally the classifier can be customized and validated for any metabarcode marker [5]. The RDP classifier calculates k-mer frequencies and uses a naive Bayes method to taxonomically assign unknown query sequences. Bootstrapping is used to provide a measure of statistical support, or repeatability, for each assignment at each rank. We have previously described how this method works compared to the top BLAST hit method [28]. In that comparison, we showed how the RDP classifier is faster than the top BLAST hit method and helps to reduce the rate of false-positive assignments. In studies where erroneously identifying a metabarcode sequence as a potential invasive species or pathogen could lead to alarm, reducing the false-positive assignment rate is critical. We provide a suite of trained classifiers, ready for use with MetaWorks (Table 1). Additionally, we provide the training files so that users can check that key target taxa are present in the reference database, and users are free to use the FASTA-formatted sequence files to create custom BLAST databases for similarity-based searches for data exploration or to build reference sets for subsequent phylogenetic analysis. The final file is a comma-separated value file (results.csv) where the taxonomic assignment for each sequence variant is provided for each sample along with read counts. If an rRNA marker was processed, then the sequence variant sequence is provided in this file; and if a protein coding region was processed, then longest ORF is provided.

Operational taxonomic units

This pipeline supports the analysis of sequence variants for the additional genetic and taxonomic resolution provided by this level of analysis [24]. Though this method of analysis was initially used to process 16S rRNA genes, studies using ITS and COI have also shown that the analysis of SVs improves the detection of genetic diversity and richness, when assessing beta diversity, both SVs and OTUs tend to recover similar gradients in multivariate analyses [25,26]. Although it has been shown that for many clustering methods sequence order matters and OTU composition can change form one analysis to the next making reproducibility an issue, there are a number of reasons why a user would want to analyze OTUs. For example, it may be more advantageous to work with OTUs instead of sequence variants for network analysis to detect more co-occurrences, for legacy reasons to compare results to previous studies that used OTUs, or to approximate 'species' units [48].

After processing raw reads using the snakefile_ESV workflow described in the previous section, users cluster sequence variants into OTUs, a method that combines the benefits of denoising with clustering using a 97% sequence similarity cutoff using the snakefile_OTU workflow [26,49]. This method uses VSEARCH 'cluster_smallmem' method to cluster SVs using a 97% sequence similarity cutoff. Settings can be adjusted in the in the config_OTU.yaml file such as pointing to the directory that contains the sequence variants and choosing a classifier for the OTUs.

Results and Discussion

MetaWorks has already been used in several publications for the Canadian STREAM biomonitoring program, the Government of Canada, Genomics Research and Development Initiative, Metagenomics-based ecosystem biomonitoring (Ecobiomics) project, and by Natural Resources Canada [18,50,51]. The benefits of using an automated, scalable, versioned pipeline for biomonitoring are many-fold, from the ability to share reproducible workflows with collaborators to facilitate the re-analysis of data as more samples are collected from year-to-year. Below, we describe three MetaWorks use-cases in more detail below.

Use case 1: As a part of the Canadian STREAM biomonitoring Initiative, the MetaWorks pipeline has been used to process macroinvertebrates COI metabarcodes surveyed from stream sites across Canada [50]. One feature of this project is the quick 1-2 month turn-around time from sampling through to the production of watershed biodiversity reports. This is an improvement over reports generated using conventional morphology-based methods that would normally take 6 - 12 months to produce. The use of a consistent bioinformatics workflow to process metabarcodes has played a key role in the reproducibility, scalability, and throughput to facilitate timely reporting [52]. Generally, samples are processed in batches of 96 per sequencing run then later split into custom reports for stakeholders, processing about 500 samples per year. One feature of these reports are the taxonomic assignments made using the naive Bayesian classifier that can be filtered for identifications to the species, genus, and family ranks with 99% accuracy assuming the query is represented in the underlying sequence database [28]. This is in contrast with the use of more traditional methods for taxonomic

assignment, where taxa are routinely missed during subsampling and taxa detected by primary analysts and auditors may differ by up to 30% [53]. This use-case shows how MetaWorks can be used to create taxon lists for large-scale biodiversity monitoring of streams across Canada.

Use case 2: Also as a part of the STREAM project described above, MetaWorks results were to analyze sequence variants from diatoms (rbcL) and arthropods (COI) sampled within and across sites of varying water quality [54]. Using the MetaWorks pipeline, two different protein-coding markers were bioinformatically processed in two runs. The first run processed the rbcL marker, using a mixture of 5 different primers in a single adapters.fasta file, and pseudogenes were removed from this dataset using the simple ORFfinder method [21]. The second run processed three COI amplicons, each targeting an approximately 200 bp length of the COI barcoding region using 6 different primers in a single adapters.fasta file, and pseudogenes were removed from this dataset using the ORFfinder+HMMER method since a COI arthropod HMM model was available [21]. The study reported a diversity assessment across sites of varying water quality using richness, effective richness, and beta diversity. Additionally, the taxonomic assignments generated from MetaWorks were used to obtain resource-consumer relationships from a global database of biotic interactions (GloBI) so that community stability using trophic and network measures could be assessed across sites with varying water quality [55]. This use-case shows how MetaWorks can handle a variety of protein-coding markers for trophic and network analyses to facilitate ecological assessments of freshwater condition.

Use case 3: As a part of collaborative work with Environment and Climate Change Canada, MetaWorks was used to assess macroinvertebrate and (non-macroinvertebrate) eukaryote taxa in an urban harbour using COI and 18S rRNA [56]. Using the MetaWorks pipeline, COI metabarcodes were identified down to species rank with 99% accuracy and 18S metabarcodes were identified to genus rank with 80% accuracy using a custom-trained classifier based on the SILVA 18S release 138 [38]. In this study, conventional macroinvertebrate sampling for assessing water quality in Toronto Harbour was compared with metabarcoding methods. COI metabarcoding was found to detect more diversity at a finer level of taxonomic resolution compared with conventional approaches and was able to distinguish sites with particularly high levels of sediment contaminants. Additionally, the use of a multi-marker approach allowed microscopic eukaryote diversity to be sampled at the same time from the same samples, producing indicators that responded to gradients in both sediment contaminants and water physical-chemical features. This use-case illustrates how MetaWorks can facilitate the application of multi-marker metabarcoding approaches that target different domains of life.

As demonstrated in the above examples, MetaWorks supports a wide range of analysis scenarios from metabarcoding data. We envision that MetaWorks will aid broader user communities and fill a need in multi-marker metabarcoding studies that target taxa from multiple different domains of life, to provide a unified processing environment, pipeline, and taxonomic assignment approach for each marker from ribosomal RNA genes, spacers, or protein coding genes. QIIME2 is perhaps the most

297 popular and comprehensive platform for such work, but to date, focuses on processing
298 mainly prokaryote and fungal datasets [16]. As of yet, MetaWorks is the only
299 bioinformatic pipeline that can handle rRNA genes but that also integrates special
300 processing steps to handle ITS spacers as well as filter out obvious pseudogenes in
301 protein coding markers such as COI.

302 There has been a lot of activity with respect to building new bioinformatic tools to
303 handle COI metabarcodes. Recent work, such as the BOLDigger program, has
304 attempted to make the BOLD identification engine more suitable for identifying large
305 batches of COI metabaracodes and has both GUI and command-line interfaces for
306 efficient sample processing [57]. A new program, called NUMTdumper, has been
307 developed as a stand-alone program meant to be incorporated into bioinformatic
308 pipelines [20]. NUMTdumper provides a method to screen for NuMTs based on read
309 counts while acknowledging the trade-offs between removing all possible NuMTs while
310 erroneously removing genuine reads. An R package called 'coil' has also recently been
311 developed that will place COI barcode and metabarcode sequences in frame using
312 profile HMM analysis [58]. MetaWorks aims to extend the COI metabarcode toolkit that
313 provides a harmonized environment where data from other organismal markers in multi-
314 marker, multi-trophic studies can also be analyzed.

315

316 **Conclusion**

317

318 MetaWorks is provided as free and open software that is versioned, can be easily
319 deployed in a Conda environment, and is supported by a suite of classifiers for popular

metabarcoding markers. The software comes with a small set of raw data and a quick-start pipeline to help users gain experience quickly. There is extensive online documentation available at the link above included detailed explanations of the pipeline, available workflows, and a tutorial for new users who have never used Conda or Snakemake before. MetaWorks generates a CSV file that lists all sequence clusters, for each sample, with associated read counts, taxonomic assignments, and bootstrap support values. Numerous statistics and log files are also provided so that users can track the number of reads that pass each major bioinformatic step. Given the current use of MetaWorks by large-scale national initiatives such as STREAM and Ecobiomics, we foresee additional developments and enhancements. Future improvements include the development of additional HMM models for pseudogene filtering, updated and additional classifiers for taxonomic assignment, and support for processing larger jobs both on HPCs and in a cloud environment. We welcome suggestions and potential collaborative work to further advance this pipeline for the scientific community.

Availability and requirements

MetaWorks runs at the command-line in a Conda environment on a Linux operating system. Miniconda can be downloaded from https://repo.continuum.io/miniconda/Miniconda3-latest-Linux-x86_64.sh. If pseudogene filtering will be performed ORFfinder can be downloaded from the NCBI ftp site at ftp://ftp.ncbi.nlm.nih.gov/genomes/TOOLS/ORFfinder/linux-i64/ORFfinder.gz. MetaWorks workflows, quick-start example, and a tutorial for new users is available from GitHub at https://github.com/terrimporter/MetaWorks. Additional trained classifiers

that can be used by MetaWorks for taxonomic assignment are available from GitHub at https://github.com/terrimporter.

Funding

MH received funding from Genome Canada and Ontario Genomics for the Sequencing the Rivers for Environmental Assessment and Monitoring (STREAM) project. TMP received funding from the Government of Canada through the Genomics Research and Development Initiative (GRDI), Metagenomics-based ecosystem biomonitoring (Ecobiomics) project.

Authors' contributions

TP and MH conceived of the work. TP designed and wrote the pipeline. All the authors helped to write the manuscript, approve the submitted version, and agree to be personally accountable for the author's own contributions and to ensure that questions related to the accuracy or integrity of any part of the work, even ones in which the author was not personally involved, are appropriately investigated, resolved, and the resolution documented in the literature.

Acknowledgements

We would like to thank Josip Rudar, Katie M. McGee, Chloe V. Robinson, Victoria C. Maitland, and Michael T.G. Wright from the Hajibabaei lab for helpful discussions and testing the pipeline with various datasets.

References

1. Pace NR. A Molecular View of Microbial Diversity and the Biosphere. Science. 1997;276: 734–740. doi:10.1126/science.276.5313.734

2. Hajibabaei M. The golden age of DNA metasystematics. Trends in genetics. 2012;28: 535–537.

3. Taberlet P, Coissac E, Pompanon F, Brochmann C, Willerslev E. Towards next-generation biodiversity assessment using DNA metabarcoding. Molecular ecology. 2012;21: 2045–2050.

4. Gilbert JA, Jansson JK, Knight R. The Earth Microbiome project: successes and aspirations. BMC biology. 2014;12: 69.

5. Wang Q, Garrity GM, Tiedje JM, Cole JR. Naive Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy. Applied and Environmental Microbiology. 2007;73: 5261–5267. doi:10.1128/AEM.00062-07

6. Schüßler A. Glomales SSUrRNA gene diversity. New Phytologist. 1999;144: 205–207. doi:10.1046/j.1469-8137.1999.00526.x

7. Hebert PDN, Cywinska A, Ball SL, deWaard JR. Biological identifications through DNA barcodes. Proceedings of the Royal Society B: Biological Sciences. 2003;270: 313–321. doi:10.1098/rspb.2002.2218

8. James TY, Kauff F, Schoch CL, Matheny PB, Hofstetter V, Cox CJ, et al. Reconstructing the early evolution of Fungi using a six-gene phylogeny. Nature. 2006;443: 818–822. doi:10.1038/nature05110

9. Hibbett DS, Binder M, Bischoff JF, Blackwell M, Cannon PF, Eriksson OE, et al. A higher-level phylogenetic classification of the Fungi. Mycological Research. 2007;111: 509–547. doi:10.1016/j.mycres.2007.03.004

10. CBOL Plant Working Group, Hollingsworth PM, Forrest LL, Spouge JL, Hajibabaei M, Ratnasingham S, et al. A DNA barcode for land plants. Proceedings of the National Academy of Sciences. 2009;106: 12794–12797.

11. Schoch CL, Seifert KA, Huhndorf S, Robert V, Spouge JL, Levesque CA, et al. Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. Proceedings of the National Academy of Sciences. 2012;109: 6241–6246. doi:10.1073/pnas.1117018109

12. Zimmermann J, Abarca N, Enk N, Skibbe O, Kusber W-H, Jahn R. Taxonomic Reference Libraries for Environmental Barcoding: A Best Practice Example from Diatom Research. Schierwater B, editor. PLoS ONE. 2014;9: e108793. doi:10.1371/journal.pone.0108793

13. Sato Y, Miya M, Fukunaga T, Sado T, Iwasaki W. MitoFish and MiFish Pipeline: A Mitochondrial Genome Database of Fish with an Analysis Pipeline for Environmental DNA Metabarcoding. Kumar S, editor. Molecular Biology and Evolution. 2018;35: 1553–1555. doi:10.1093/molbev/msy074

14. Ahmed M, Back MA, Prior T, Karssen G, Lawson R, Adams I, et al. Metabarcoding of soil nematodes: the importance of taxonomic coverage and availability of reference sequences in choosing suitable marker(s). MBMG. 2019;3: e36408. doi:10.3897/mbmg.3.36408

15. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP. DADA2: High-resolution sample inference from Illumina amplicon data. Nature Methods. 2016;13: 581–583. doi:10.1038/nmeth.3869

16. Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet C, Al-Ghalith GA, et al. Reproducible, interactive, scalable, and extensible microbiome data science using QIIME 2. Nature Biotechnology. 2019;37: 852–857. doi:10.7287/peerj.preprints.27295v1

17. Drummond AJ, Newcomb RD, Buckley TR, Xie D, Dopheide A, Potter BC, et al. Evaluating a multigene environmental DNA approach for biodiversity assessment. GigaSci. 2015;4: 46. doi:10.1186/s13742-015-0086-1

18. Edge TA, Baird DJ, Bilodeau G, Gagné N, Greer C, Konkin D, et al. The Ecobiomics project: Advancing metagenomics assessment of soil health and freshwater quality in Canada. Science of The Total Environment. 2020;710: 135906. doi:10.1016/j.scitotenv.2019.135906

19. Bengtsson-Palme J, Ryberg M, Hartmann M, Branco S, Wang Z, Godhe A, et al. Improved software detection and extraction of ITS1 and ITS2 from ribosomal ITS sequences of fungi and other eukaryotes for analysis of environmental sequencing data. Bunce M, editor. Methods in Ecology and Evolution. 2013;4: 914–919. doi:10.1111/2041-210X.12073

20. Andújar C, Creedy TJ, Arribas P, López H, Salces-Castellano A, Pérez-Delgado A, et al. NUMT dumping: validated removal of nuclear pseudogenes from mitochondrial metabarcode data. Evolutionary Biology; 2020 Jun. doi:10.1101/2020.06.17.157347

21. Porter TM, Hajibabaei M. Profile hidden Markov model sequence analysis can help remove putative pseudogenes from DNA barcoding and metabarcoding datasets. BMC Bioinformatics. 2021;22: 256. doi:10.1186/s12859-021-04180-x

22. Song H, Buhay JE, Whiting MF, Crandall KA. Many species in one: DNA barcoding overestimates the number of species when nuclear mitochondrial pseudogenes are coamplified. PNAS. 2008;105: 13486–13491. doi:10.1073/pnas.0803076105

23. Moulton MJ, Song H, Whiting MF. Assessing the effects of primer specificity on eliminating numt coamplification in DNA barcoding: a case study from Orthoptera (Arthropoda:

Insecta): DNA BARCODING. Molecular Ecology Resources. 2010;10: 615–627. doi:10.1111/j.1755-0998.2009.02823.x

24. Callahan BJ, McMurdie PJ, Holmes SP. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. The ISME Journal. 2017;11: 2639–2643.

25. Glassman SI, Martiny JB. Ecological patterns are robust to use of exact sequence variants versus operational taxonomic units. mSphere. 2018;3: e00148-18. doi:10.1101/283283

26. Porter TM, Hajibabaei M. Putting COI Metabarcoding in Context: The Utility of Exact Sequence Variants (ESVs) in Biodiversity Analysis. Front Ecol Evol. 2020;8: 248. doi:10.3389/fevo.2020.00248

27. Porter TM, Gibson JF, Shokralla S, Baird DJ, Golding GB, Hajibabaei M. Rapid and accurate taxonomic classification of insect (class Insecta) cytochrome c oxidase subunit 1 (COI) DNA barcode sequences using a naïve Bayesian classifier. Mol Ecol Resour. 2014;14: 929–942. doi:10.1111/1755-0998.12240

28. Porter TM, Hajibabaei M. Automated high throughput animal CO1 metabarcode classification. Scientific Reports. 2018;8: 4226. doi:https://doi.org/10.1038/s41598-018-22505-4

29. Virgilio M, Backeljau T, Nevado B, De Meyer M. Comparative performances of DNA barcoding across insect orders. BMC bioinformatics. 2010;11: 206.

30. Adamowicz SJ, Boatwright JS, Chain F, Fisher BL, Hogg ID, Leese F, et al. Trends in DNA barcoding and metabarcoding. Genome. 2019;62: v–viii. doi:10.1139/gen-2019-0054

31. Koster J, Rahmann S. Snakemake--a scalable bioinformatics workflow engine. Bioinformatics. 2012;28: 2520–2522. doi:10.1093/bioinformatics/bts480

32. Anaconda. Anaconda Software Distribution. 2016. Available: https://anaconda.com

33. Wratten L, Wilm A, Göke J. Reproducible, scalable, and shareable analysis pipelines with bioinformatics workflow managers. Nat Methods. 2021 [cited 24 Sep 2021]. doi:10.1038/s41592-021-01254-9

34. Ratnasingham S, Hebert PD. BOLD: The Barcode of Life Data System (http://www. barcodinglife. org). Molecular ecology notes. 2007;7: 355–364.

35. Cochrane G, Karsch-Mizrachi I, Takagi T, Sequence Database Collaboration IN. The International Nucleotide Sequence Database Collaboration. Nucleic Acids Research. 2016;44: D48–D50. doi:10.1093/nar/gkv1323

36. Rimet F, Gusev E, Kahlert M, Kelly MG, Kulikovskiy M, Maltsev Y, et al. Diat.barcode, an open-access curated barcode library for diatoms. Sci Rep. 2019;9: 15116. doi:10.1038/s41598-019-51500-6

37. Iwasaki W, Fukunaga T, Isagozawa R, Yamada K, Maeda Y, Satoh TP, et al. MitoFish and MitoAnnotator: A Mitochondrial Genome Database of Fish with an Accurate and Automatic Annotation Pipeline. Molecular Biology and Evolution. 2013;30: 2531–2540. doi:10.1093/molbev/mst141

38. Pruesse E, Quast C, Knittel K, Fuchs BM, Ludwig W, Peplies J, et al. SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. Nucleic Acids Research. 2007;35: 7188–7196. doi:10.1093/nar/gkm864

39. Deshpande V, Wang Q, Greenfield P, Charleston M, Porras-Alfaro A, Kuske CR, et al. Fungal identification using a Bayesian classifier and the Warcup training set of internal transcribed spacer sequences. Mycologia. 2016;108: 1–5.

40. Abarenkov K, Nilsson RH, Larsson K-H, Alexander IJ, Eberhardt U, Erland S, et al. The UNITE database for molecular identification of fungi – recent updates and future perspectives. New Phytologist. 2010;186: 281–285.

41. Banchi E, Ametrano CG, Greco S, Stanković D, Muggia L, Pallavicini A. PLANiTS: a curated sequence reference dataset for plant ITS DNA metabarcoding. Database. 2020;2020: baz155. doi:10.1093/database/baz155

42. Liu K-L, Porras-Alfaro A, Kuske CR, Eichorst SA, Xie G. Accurate, Rapid Taxonomic Classification of Fungal Large-Subunit rRNA Genes. Appl Environ Microbiol. 2012;78: 1523–1533. doi:10.1128/AEM.06826-11

43. St. John J. SeqPrep. Downloaded 2016. Available: https://github.com/jstjohn/SeqPrep/releases

44. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet journal. 2011;17: pp-10.

45. Edgar RC. UNOISE2: improved error-correction for Illumina 16S and ITS amplicon sequencing. bioRxiv. 2016 [cited 28 Jun 2018]. doi:10.1101/081257

46. Edgar R. UCHIME2: improved chimera prediction for amplicon sequencing. bioRxiv. 2016; 074252.

47. Nearing JT, Douglas GM, Comeau AM, Langille MGI. Denoising the Denoisers: an independent evaluation of microbiome sequence error-correction approaches. PeerJ. 2018;6: e5364. doi:10.7717/peerj.5364

48. He Y, Caporaso JG, Jiang X-T, Sheng H-F, Huse SM, Rideout JR, et al. Stability of operational taxonomic units: an important but neglected property for analyzing microbial diversity. Microbiome. 2015;3. doi:10.1186/s40168-015-0081-x

49. Antich A, Palacin C, Wangensteen OS, Turon X. To denoise or to cluster, that is not the question: optimizing pipelines for COI metabarcoding and metaphylogeography. BMC Bioinformatics. 2021;22: 177. doi:10.1186/s12859-021-04115-6

50. Robinson CV, Baird DJ, Wright MTG, Porter TM, Hartwig K, Hendriks E, et al. Combining DNA and people power for healthy rivers: Implementing the STREAM community-based approach for global freshwater monitoring. Perspectives in Ecology and Conservation. 2021;19: 279–285. doi:10.1016/j.pecon.2021.03.001

51. Smenderovac E, Emilson C, Porter T, Morris D, Hazlett P, Diochon A, et al. Forest soil biotic communities show few responses to wood ash applications at multiple sites across Canada. Sci Rep. 2022;12: 4171. doi:10.1038/s41598-022-07670-x

52. Porter TM, Hajibabaei M. Scaling up: A guide to high-throughput genomic approaches for biodiversity analysis. Molecular Ecology. 2018;27: 313–338. doi:10.1111/mec.14478

53. Haase P, Pauls SU, Schindehütte K, Sundermann A. First audit of macroinvertebrate samples from an EU Water Framework Directive monitoring program: human error greatly lowers precision of assessment results. Journal of the North American Benthological Society. 2010;29: 1279–1291. doi:10.1899/09-183.1

54. Robinson CV, Porter TM, Maitland VC, Wright MT, Hajibabaei M. Multi-marker metabarcoding resolves subtle variations in freshwater condition: Bioindicators, ecological traits, and trophic interactions. Ecology; 2021 Nov. doi:10.1101/2021.11.14.468533

55. Poelen JH, Simons JD, Mungall CJ. Global biotic interactions: An open infrastructure to share and analyze species-interaction datasets. Ecological Informatics. 2014;24: 148–159. doi:10.1016/j.ecoinf.2014.08.005

56. Robinson CV, Porter TM, McGee KM, McCusker M, Wright MTG, Hajibabaei M. Multi-marker DNA metabarcoding detects suites of environmental gradients from an urban harbour. BioRxiv. 2022; 35.

57. Buchner D, Leese F. BOLDigger – a Python package to identify and organise sequences with the Barcode of Life Data systems. MBMG. 2020;4: e53535. doi:10.3897/mbmg.4.53535

58. Nugent CM, Elliott TA, Ratnasingham S, Adamowicz SJ. coil: an R package for cytochrome C oxidase I (COI) DNA barcode data cleaning, translation, and error evaluation. bioRxiv. 2019; 35. doi:doi: http://dx.doi.org/10.1101/2019.12.12.865014