THAPBI PICT - a fast, cautious, and accurate metabarcoding analysis pipeline

Peter J. A. Cock[1], David E. L. Cooke[1], Peter Thorpe[2], and Leighton Pritchard[3]

[1]**James Hutton Institute, Dundee, DD2 5DA, UK**
[2]**Data Analysis Group, School of Life sciences, University of Dundee, Dundee, DD1 5EH, UK**
[3]**Strathclyde Institute of Pharmacy & Biomedical Sciences, Glasgow, G4 0RE, UK**

Corresponding author:
First Author[1]

Email address: peter.cock@hutton.ac.uk

ABSTRACT

THAPBI PICT is an open source software pipeline for metabarcoding analysis of Illumina paired-end reads, including cases of multiplexing where more than one amplicon is amplified per DNA sample. Initially a *Phytophthora* ITS1 Classification Tool (PICT), we demonstrate using worked examples with our own and public data sets how, with appropriate primer settings and a custom database, it can be applied to other amplicons and organisms, and used for reanalysis of existing datasets. The core dataflow of the implementation is (i) data reduction to unique marker sequences, often called amplicon sequence variants (ASVs), (ii) dynamic thresholds for discarding low abundance sequences to remove noise and artifacts (rather than error correction by default), before (iii) classification using a curated reference database. The default classifier assigns a label to each query sequence based on a database match that is either perfect, or a single base pair edit away (substitution, deletion or insertion). Abundance thresholds for inclusion can be set by the user or automatically using per-batch negative or synthetic control samples. Output is designed for practical interpretation by non-specialists and includes a read report (ASVs with classification and counts per sample), sample report (samples with counts per species classification), and a topological graph of ASVs as nodes with short edit distances as edges. Source code available from `https://github.com/peterjc/thapbi-pict/` with documentation including installation instructions.

INTRODUCTION

Metabarcoding of DNA is a sensitive and powerful method to detect, identify, and potentially quantify the diversity of biological taxa present in any given environmental sample. It is based on PCR amplification of a “barcode” region diagnostic for the groups of organisms of interest followed by high-throughput sequencing of the amplimers, and is often applied to environmental DNA (eDNA) samples (Deiner et al., 2017). This method is revolutionising areas of research including wildlife conservation, ecological processes and microbiology, by highly-sensitive detection of biodiversity across many taxa simultaneously (Arulandhu et al., 2017). Metabarcoding enables early detection of invasive threats to plant and human health in support of biosecurity (Batovska et al., 2021; Green et al., 2021), and is applicable to many complex and intractable systems, such as soil (Ahmed et al., 2019), in which standard methods of microbial isolation and characterisation are impractical or costly.

Our motivating use case is metabarcoding in which multiple environmental samples are multiplexed for high-throughput sequencing on the Illumina platform using paired-end reads, and for which the expected PCR amplification product is short enough to be fully covered by the overlapping paired reads. Each sample is expected to yield taxon-specific marker sequences that can be matched to a high-quality database of marker sequences with known taxonomic identity, to give a taxonomic breakdown reflecting the community composition. One of our goals was to minimise false positive identification of the presence of any taxon on the basis of small or disputable quantities of physical evidence. Metabarcoding is prone to generation of artefactual sequence variation and sufficiently highly sensitive to register low-abundance sample reads at the same level as such sequences, and sequences originating from cross-sample contamination and "splashover" in even a careful laboratory. We therefore chose to prioritise accurate reporting of taxonomic assignment for high abundance sequences over sensitive detection of low-abundance marker sequences.

This manuscript was initially released as a preprint (Cock et al., 2023). We describe THAPBI PICT v1.0.0, a metabarcoding tool developed as part of the UKRI-funded Tree Health and Plant Biosecurity Initiative (THAPBI) Phyto-Threats project, which focused on identifying *Phytophthora* species in commercial forestry and horticultural plant nurseries (Green et al., 2021). *Phytophthora* (from Greek meaning plant-destroyer) is an economically important genus of oomycete plant pathogens that causes severe losses and damage to plants in agricultural, forest and natural ecosystems. The Phyto-Threats project's metabarcoding protocol used nested PCR primers designed to target the Internal Transcribed Spacer 1 marker sequence (ITS1; a genomic region located between 18S and 5.8S rRNA genes in eukaryotes) of *Phytophthora* and related plant pathogenic *oomycetes* (Scibetta et al., 2012). This approach is the current *de facto* standard within the oomycete community (Robideau et al., 2011), and these primers have been used in conjunction with THAPBI PICT in recent *Phytophthora* surveys (Vélez et al., 2020; La Spada et al., 2022). PICT was short for *Phytophthora* ITS1 Classification Tool.

We describe the implementation, operation, performance and output of THAPBI PICT using datasets from the Phyto-Threats project, and public metabarcoding datasets. Although originally designed as a *Phytophthora* ITS1 Classification Tool (PICT), we show that with appropriate primer settings and a custom database of genus/species distinguishing markers, THAPBI PICT is an effective tool for analysis of short read amplicon sequencing data with barcode marker sequences from other organisms.

WORKFLOW OVERVIEW

The THAPBI PICT core workflow comprises (i) data reduction to unique marker sequences, often called amplicon sequence variants (ASVs) (ii) discard of low abundance sequences to remove noise and artifacts (rather than attempting error correction by default), and (iii) classification using a curated reference database. This approach contrasts with commonly-used operational taxonomic unit (OTU) clustering approaches (as implemented, for example, in QIIME (Caporaso et al., 2010), UPARSE (Edgar, 2013), and MOTHUR (Schloss et al., 2009)), which can be sensitive to changes in the input data resulting in unpredictable clustering behaviour (Callahan et al., 2017) and overestimate population diversity (Nearing et al., 2018).

THAPBI PICT's approach of reducing amplicons to ASVs is similar to that of DADA2 (Callahan et al., 2016) but, by contrast, THAPBI PICT does not by default attempt to correct sequencing errors with a denoising model. Our approach is instead to discard low-abundance sequences because we consider that they are likely not to represent meaningful biological information in the sequenced sample. We observe using synthetic control sequences that the abundance of such controls accidentally transferred between samples tends to exceed by no small margin the abundance of amplicons whose sequence variation might constitute "noise" in the amplicon sequence data. We consider the observed abundance of (e.g. synthetic) control sequences, which could not have been present in the biological sample, to be a lower bound for the abundance of reads we can confidently claim derive from that sample. Consequently, ASVs with much lower total abundance cannot confidently be determined to derive from the analysed sample, and so are discarded. In general, we consider that proper use of negative and synthetic controls, to account for alternative sources of experimental error, such as accidental transfer or "splashover" from one well to another, should be considered best practice in metabarcoding.

Figure 1 gives an overview of the workflow. Paired raw Illumina FASTQ files for each sample are merged by overlap, trimmed to remove primers, and reduced to a list of observed unique marker sequences (labelled by MD5 checksum) with abundance counts. Discarding low abundance sequences further reduces the data volume - unique reads alone may represent half the data (and 90% of the ASVs), but may not derive from the sequenced sample. The remaining higher abundance sequences are then classified by matching them to a curated database. By default a species-level assignment is made when a database entry is identical or different by at most one base pair (1bp; algorithm `onebp`) to the query. The matching algorithm can be chosen to adjust sensitivity for taxonomic classification (Table 1)

Following read preparation and ASV classification, the pipeline generates two tables describing (i) taxon presence/absence for each sample, and (ii) ASV presence/absence for each sample (Figure 2 (a)

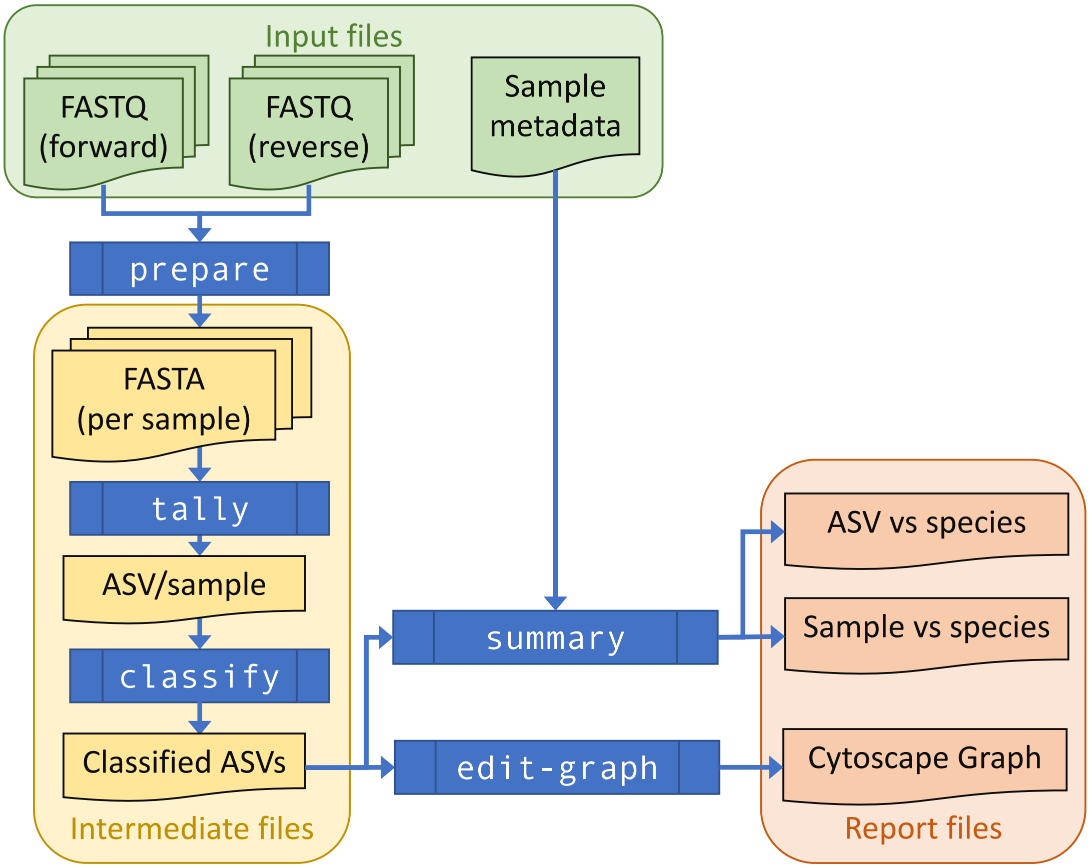
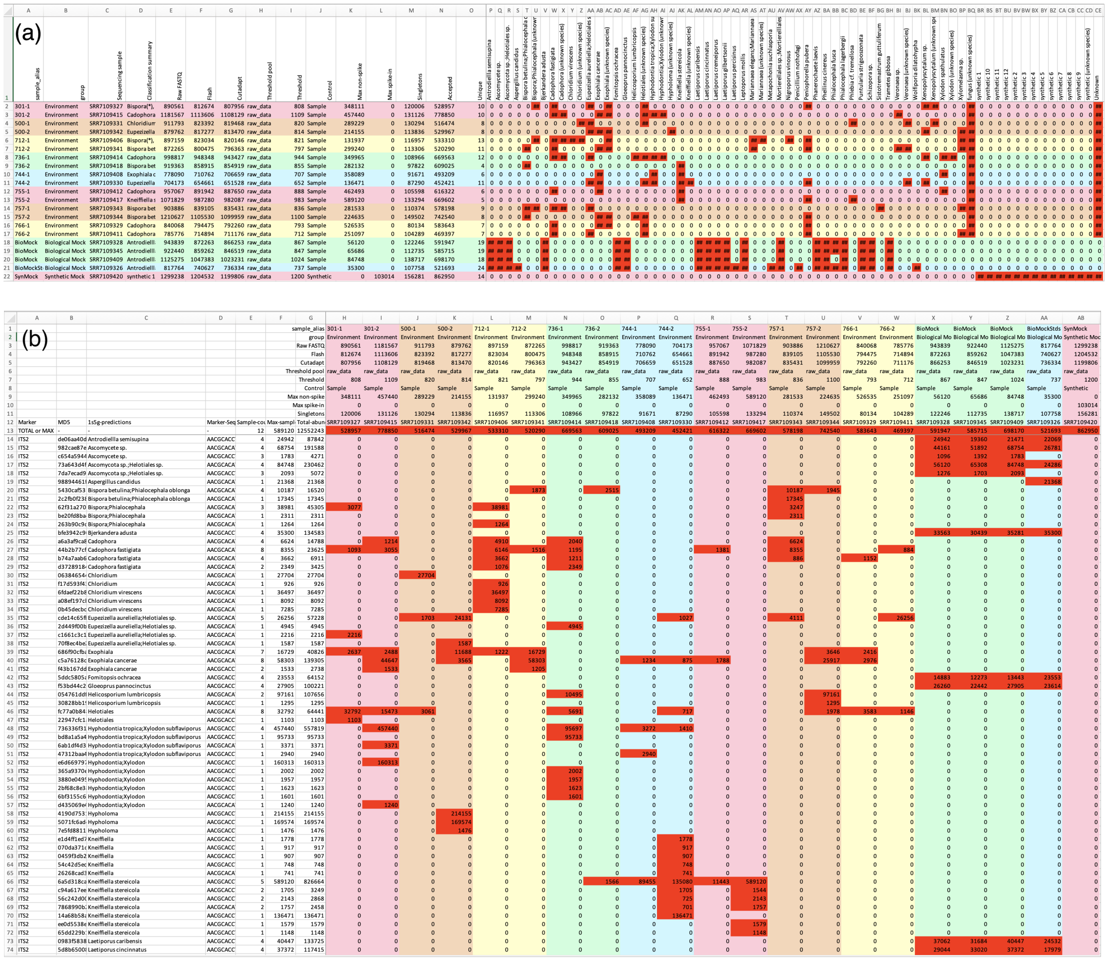


Figure 1. THAPBI PICT workflow overview. Raw paired FASTQ input data is transformed (commands `prepare-reads`, `sample-tally`, `classify`) into intermediate FASTA and TSV (tab-separated value) format files recording tallies of ASV counts and ASV classifications, using a local marker sequence database. Summary report generation (command `summary`) produces output in reproducible (TSV, TXT file) and user-focused (e.g. colour-coded Excel spreadsheet) formats. The stages of THAPBI PICT can be run individually, or as a single `pipeline` command that incorporates the `prepare-reads`, `sample-tally`, `classify` and `summary` commands. Sample metadata can optionally be incorporated into report output, and used to sort reports and support downstream interpretation. In addition BIOM format output can be requested. An ASV edit graph for the samples can be generated (command `edit-graph`) to aid in diagnosis and interpretation.

Name	Description
`identity`	Perfect match in database (strictest)
`substr`	Perfect match or perfect substring of a database entry
`onebp`	Perfect match, or one bp away (default)
`1s2g`	As `onebp` but falling back on up to 2bp away for a genus only match.
`1s3g`	As `onebp` but falling back on up to 3bp away for a genus only match.
`1s4g`	As `onebp` but falling back on up to 4bp away for a genus only match.
`1s5g`	As `onebp` but falling back on up to 5bp away for a genus only match.
`blast`	Best NCBI blastn alignment covering at least 85% of the query, and 95% identity.

Table 1. Taxonomic classifier algorithms in THAPBI PICT. Names constructed as `XsYg` reflect an edit distance of up to and including `X`bp for species classification, and `Y`bp for genus-level classification. Genus-level classification does not attempt to assign a species-level identity to the sequence.



(a)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	sample_alias	group	Sequencing sample	Classification summary	Raw FASTQ	Flash	Cutadapt	Threshold pool	Threshold	Control	Max non-spike	Max spike-in	Singletons	Accepted	Unique
2	301-1	Environment	SRR7109327	Bispora(*),	890561	812674	807956	raw_data	808	Sample	348111	0	120006	528957	10
3	301-2	Environment	SRR7109415	Cadophora	1181567	1113606	1108129	raw_data	1109	Sample	457440	0	131126	778850	10
4	500-1	Environment	SRR7109331	Chloridium	911793	823392	819468	raw_data	820	Sample	289229	0	130294	516474	8
5	500-2	Environment	SRR7109342	Eupezizella	879762	817277	813470	raw_data	814	Sample	214155	0	113836	529967	8
6	712-1	Environment	SRR7109406	Bispora(*),	897159	823034	820146	raw_data	821	Sample	131937	0	116957	533310	13
7	712-2	Environment	SRR7109341	Bispora bet	872265	800475	796363	raw_data	797	Sample	299240	0	113306	520290	11
8	736-1	Environment	SRR7109414	Cadophora	998817	948348	943427	raw_data	944	Sample	349965	0	108966	669563	12
9	736-2	Environment	SRR7109418	Bispora bet	919363	858915	854919	raw_data	855	Sample	282132	0	97822	609025	4
10	744-1	Environment	SRR7109408	Exophiala c	778090	710762	706659	raw_data	707	Sample	358089	0	91671	493209	6
11	744-2	Environment	SRR7109330	Eupezizella	704173	654661	651528	raw_data	652	Sample	136471	0	87290	452421	11
12	755-1	Environment	SRR7109412	Cadophora	957067	891942	887650	raw_data	888	Sample	462493	0	105598	616322	6
13	755-2	Environment	SRR7109417	Kneiffiella s	1071829	987280	982087	raw_data	983	Sample	589120	0	133294	669602	5
14	757-1	Environment	SRR7109343	Bispora bet	903886	839105	835431	raw_data	836	Sample	281533	0	110374	578198	9
15	757-2	Environment	SRR7109344	Bispora bet	1210627	1105530	1099959	raw_data	1100	Sample	224635	0	149502	742540	8
16	766-1	Environment	SRR7109329	Cadophora	840068	794475	792260	raw_data	793	Sample	526535	0	80134	583643	7
17	766-2	Environment	SRR7109411	Cadophora	785776	714894	711176	raw_data	712	Sample	251097	0	104289	469397	7
18	BioMock	Biological Mock	SRR7109328	Antrodielll	943839	872263	866253	raw_data	867	Sample	56120	0	122246	591947	19
19	BioMock	Biological Mock	SRR7109345	Antrodielll	922440	859262	846519	raw_data	847	Sample	65686	0	112735	585715	19
20	BioMock	Biological Mock	SRR7109409	Antrodielll	1125275	1047383	1023231	raw_data	1024	Sample	84748	0	138717	698170	18
21	BioMockSt	Biological Mock	SRR7109326	Antrodielll	817764	740627	736334	raw_data	737	Sample	35300	0	107758	521693	24
22	SynMock	Synthetic Mock	SRR7109420	synthetic 1	1299238	1204532	1199806	raw_data	1200	Synthetic	0	103014	156281	862950	14

	A	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF	AG	AH	AI	AJ	AK	AL	AM	AN	AO	AP	AQ	AR
1	sample_alias	Antrodiella semisupina	Ascomycete sp.	Ascomycota sp.;Helotiales sp.	Aspergillus candidus	Bispora betulina;Phialocephala c	Bispora;Phialocephala (unknowr	Bjerkandera adusta	Cadophora fastigiata	Cadophora (unknown species)	Chloridium virescens	Chloridium (unknown species)	Eupezizella aureliella;Helotiales s	Exophiala cancerae	Exophiala (unknown species)	Fomitopsis ochracea	Gloeoprus pannocinctus	Helicosporium lumbricopsis	Helotiales (unknown species)	Hyphodontia tropica;Xylodon su	Hyphodontia;Xylodon (unknowr	Hypholoma (unknown species)	Kneiffiella stereicola	Kneiffiella (unknown species)	Laetiporus caribensis	Laetiporus cincinnatus	Laetiporus cremeiporus	Laetiporus gilbertsonii	Laetiporus percisinus	Leptoporus mollis
2	301-1	0	0	0	0	0	##	0	##	0	0	0	##	0	##	0	0	0	##	0	0	0	0	0	0	0	0	0	0	0
3	301-2	0	0	0	0	0	0	0	##	##	0	0	0	##	##	0	0	0	##	##	##	0	0	0	0	0	0	0	0	0
4	500-1	0	0	0	0	0	0	0	0	0	0	##	##	0	0	0	0	0	##	0	0	0	0	0	0	0	0	0	0	0
5	500-2	0	0	0	0	0	0	0	0	0	0	0	##	##	##	0	0	0	0	0	0	##	0	0	0	0	0	0	0	0
6	712-1	0	0	0	0	0	##	0	##	##	##	##	0	0	##	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	712-2	0	0	0	0	##	0	0	##	0	0	0	0	##	##	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	736-1	0	0	0	0	0	0	0	##	##	0	0	##	0	0	0	0	##	##	##	##	0	0	0	0	0	0	0	0	0
9	736-2	0	0	0	0	##	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	##	0	0	0	0	0	0	0
10	744-1	0	0	0	0	0	0	0	0	0	0	0	0	##	0	0	0	0	0	##	0	0	##	0	0	0	0	0	0	0
11	744-2	0	0	0	0	0	0	0	0	0	0	0	##	##	0	0	0	0	##	##	0	0	##	##	0	0	0	0	0	0
12	755-1	0	0	0	0	0	0	0	##	0	0	0	0	##	0	0	0	0	0	0	0	0	##	0	0	0	0	0	0	0
13	755-2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	##	0	0	0	0	0	0	0
14	757-1	0	0	0	0	##	##	0	##	##	0	0	##	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
15	757-2	0	0	0	0	##	0	0	0	0	0	0	0	##	##	0	0	##	##	0	0	0	0	0	0	0	0	0	0	0
16	766-1	0	0	0	0	0	0	0	##	0	0	0	0	##	##	0	0	0	##	0	0	0	0	0	0	0	0	0	0	0
17	766-2	0	0	0	0	0	0	0	##	0	0	0	##	0	0	0	0	0	##	0	0	0	0	0	0	0	0	0	0	0
18	BioMock	##	##	##	0	0	0	##	0	0	0	0	0	0	0	##	##	0	0	0	0	0	0	0	##	##	##	##	0	##
19	BioMock	##	##	##	0	0	0	##	0	0	0	0	0	0	0	##	##	0	0	0	0	0	0	0	##	##	##	##	0	##
20	BioMock	##	##	##	0	0	0	##	0	0	0	0	0	0	0	##	##	0	0	0	0	0	0	0	##	##	##	##	0	##
21	BioMockSt	##	##	##	##	0	0	##	0	0	0	0	0	0	0	##	##	0	0	0	0	0	0	0	##	##	##	##	##	##
22	SynMock	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

	A	AS	AT	AU	AV	AW	AX	AY	AZ	BA	BB	BC	BD	BE	BF	BG	BH	BI	BJ	BK	BL	BM	BN	BO	BP	BQ	BR	BS	BT	BU	BV	BW	BX	BY	BZ	CA	CB	CC	CD	CE
1	sample_alias	Mariannaea elegans;Mariannaea	Mariannaea (unknown species)	Metapochonia suchlasporia	Mortierellales sp.;Mortierellales	Nigroporus vinosus	Penicillium nothofagi	Peniophorella pubera	Phanerochaete laevis	Phellinus cinereus	Phialocephala fusca	Phialocephala lagerbergii	Phlebia cf. tremellosa	Puntularia strigoosozonata	Schizopora sp.	Sistotremastrum guttuliferum	Trametes gibbosa	Veronaea sp.	Veronaea (unknown species)	Wolfiporia dilatohypha	Xenopolyscytalum sp.	Xenopolyscytalum (unknown spe	Xylodon spathulatus	Xylodon (unknown species)	Xylomelasma sp.	fungus (unknown species)	synthetic 1	synthetic 10	synthetic 11	synthetic 12	synthetic 2	synthetic 3	synthetic 4	synthetic 5	synthetic 6	synthetic 7	synthetic 8	synthetic 9	synthetic (unknown species)	Unknown
2	301-1	0	0	0	0	0	0	##	0	0	0	0	0	0	0	0	0	0	0	0	##	##	0	0	0	##	0	0	0	0	0	0	0	0	0	0	0	0	0	##
3	301-2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	##	0	0	0	0	0	0	0	##	0	0	0	0	0	0	0	0	0	0	0	0	0	##
4	500-1	0	0	0	0	0	0	0	0	0	0	0	##	0	0	0	0	0	##	0	0	##	0	0	0	##	0	0	0	0	0	0	0	0	0	0	0	0	0	##
5	500-2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	##	0	0	0	##	##	0	0	0	0	0	0	0	0	0	0	0	0	0	##
6	712-1	##	##	0	0	##	0	##	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	##	##	0	0	0	0	0	0	0	0	0	0	0	0	0	##
7	712-2	##	0	0	0	0	0	##	0	0	0	0	0	0	0	0	0	##	##	0	0	0	0	0	##	##	0	0	0	0	0	0	0	0	0	0	0	0	0	##
8	736-1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	##	0	##	##	0	##	0	0	0	0	0	0	0	0	0	0	0	0	0	##
9	736-2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	##	0	0	0	0	0	0	0	0	0	0	0	0	0	##
10	744-1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	##	0	0	##	0	0	0	0	0	0	0	0	0	0	0	0	0	##
11	744-2	0	0	0	0	0	0	##	0	0	0	0	0	0	0	0	0	0	##	0	##	0	0	0	0	##	0	0	0	0	0	0	0	0	0	0	0	0	0	##
12	755-1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	##	0	0	##	0	0	0	0	0	0	0	0	0	0	0	0	0	##
13	755-2	0	0	0	0	0	0	##	0	0	0	0	##	0	0	0	0	0	0	0	0	0	0	0	0	##	0	0	0	0	0	0	0	0	0	0	0	0	0	##
14	757-1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	##	0	0	0	0	0	0	0	0	##	##	0	0	0	0	0	0	0	0	0	0	0	0	0	##
15	757-2	0	0	0	0	0	0	##	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	##	0	0	0	0	0	0	0	0	0	0	0	0	0	##
16	766-1	0	0	0	0	0	0	##	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	##	0	0	0	0	0	0	0	0	0	0	0	0	0	##
17	766-2	0	0	0	0	0	0	##	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	##	##	0	0	0	0	0	0	0	0	0	0	0	0	0	##
18	BioMock	0	0	0	##	0	0	0	##	##	##	##	0	##	##	0	##	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
19	BioMock	0	0	0	##	0	0	0	##	##	##	##	0	##	##	0	##	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
20	BioMock	0	0	0	##	0	0	0	##	##	0	##	0	##	##	0	##	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
21	BioMockSt	0	0	##	##	0	##	0	##	##	##	##	0	##	##	0	##	0	0	##	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
22	SynMock	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	##	##	##	##	##	##	##	##	##	##	##	##	##	##

(b)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB
1							sample_alias	301-1	301-2	500-1	500-2	712-1	712-2	736-1	736-2	744-1	744-2	755-1	755-2	757-1	757-2	766-1	766-2	BioMock	BioMock	BioMock	BioMockStds	SynMock
2							group	Environment	Environment	Environment	Environment	Environment	Environment	Environment	Environment	Environment	Environment	Environment	Environment	Environment	Environment	Environment	Environment	Biological Mo	Biological Mo	Biological Mo	Biological Mo	Synthetic Moc
3							Raw FASTQ	890561	1181567	911793	879762	897159	872265	998817	919363	778090	704173	957067	1071829	903886	1210627	840068	785776	943839	922440	1125275	817764	1299238
4							Flash	812674	1113606	823392	817277	823034	800475	948348	858915	710762	654661	891942	987280	839105	1105530	794475	714894	872263	859262	1047383	740627	1204532
5							Cutadapt	807956	1108129	819468	813470	820146	796363	943427	854919	706659	651528	887650	982087	835431	1099959	792260	711176	866253	846519	1023231	736334	1199806
6							Threshold pool	raw_data	raw_data	raw_data	raw_data	raw_data	raw_data	raw_data	raw_data	raw_data	raw_data	raw_data	raw_data	raw_data	raw_data	raw_data	raw_data	raw_data	raw_data	raw_data	raw_data	raw_data
7							Threshold	808	1109	820	814	821	797	944	855	707	652	888	983	836	1100	793	712	867	847	1024	737	1200
8							Control	Sample	Sample	Sample	Sample	Sample	Sample	Sample	Sample	Sample	Sample	Sample	Sample	Sample	Sample	Sample	Sample	Sample	Sample	Sample	Sample	Synthetic
9							Max non-spike	348111	457440	289229	214155	131937	299240	349965	282132	358089	136471	462493	589120	281533	224635	526535	251097	56120	65686	84748	35300	0
10							Max spike-in	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	103014
11							Singletons	120006	131126	130294	113836	116957	113306	108966	97822	91671	87290	105598	133294	110374	149502	80134	104289	122246	112735	138717	107758	156281
12	Marker	MD5	1s5g-predictions	Marker-Seq	Sample-cou	Max-sampl	Total-abun	SRR7109327	SRR7109415	SRR7109331	SRR7109342	SRR7109406	SRR7109341	SRR7109414	SRR7109418	SRR7109408	SRR7109330	SRR7109412	SRR7109417	SRR7109343	SRR7109344	SRR7109329	SRR7109411	SRR7109328	SRR7109345	SRR7109409	SRR7109326	SRR7109420
13	TOTAL or MAX	-	-	-	12	589120	12552243	528957	778850	516474	529967	533310	520290	669563	609025	493209	452421	616322	669602	578198	742540	583643	469397	591947	585715	698170	521693	862950
14	ITS2	de06aa40d	Antrodiellla semisupina	AACGCACC	4	24942	87842	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	24942	19360	21471	22069	0
15	ITS2	982cae87e	Ascomycete sp.	AACGCACA	4	68754	191588	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	44161	51892	68754	26781	0
16	ITS2	c654a5944	Ascomycete sp.	AACGCACC	3	1783	4271	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1096	1392	1783	0	0
17	ITS2	73a643d4f	Ascomycota sp.;Helotiales sp.	AACGCACA	4	84748	230462	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	56120	65308	84748	24286	0
18	ITS2	7da7ecad9	Ascomycota sp.;Helotiales sp.	AACGCACC	3	2093	5072	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1276	1703	2093	0	0
19	ITS2	98894461	Aspergillus candidus	AACGCACA	1	21368	21368	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	21368	0
20	ITS2	5430caf53	Bispora betulina;Phialocephala oblonga	AACGCACA	4	10187	16520	0	0	0	0	0	1873	0	2515	0	0	0	0	10187	1945	0	0	0	0	0	0	0
21	ITS2	2c2fb0f23	Bispora betulina;Phialocephala oblonga	AACGCACA	1	17345	17345	0	0	0	0	0	0	0	0	0	0	0	0	17345	0	0	0	0	0	0	0	0
22	ITS2	62f31a270	Bispora;Phialocephala	AACGCACA	3	38981	45305	3077	0	0	0	38981	0	0	0	0	0	0	0	3247	0	0	0	0	0	0	0	0
23	ITS2	be20fd8ba	Bispora;Phialocephala	AACGCACA	1	2311	2311	0	0	0	0	0	0	0	0	0	0	0	0	2311	0	0	0	0	0	0	0	0
24	ITS2	263b90c9c	Bispora;Phialocephala	AACGCACA	1	1264	1264	0	0	0	0	1264	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
25	ITS2	bfe3942c9	Bjerkandera adusta	AACGCACC	4	35300	134583	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	33563	30439	35281	35300	0
26	ITS2	a6a3af9ca	Cadophora	AACGCACA	4	6624	14788	0	1214	0	0	4910	0	2040	0	0	0	0	0	6624	0	0	0	0	0	0	0	0
27	ITS2	44b2b77cf	Cadophora fastigiata	AACGCACA	8	8355	23625	1093	3055	0	0	6146	1516	1195	0	0	0	1381	0	8355	0	0	884	0	0	0	0	0
28	ITS2	b74a7aab6	Cadophora fastigiata	AACGCACA	4	3662	6911	0	0	0	0	3662	0	1211	0	0	0	0	0	886	0	1152	0	0	0	0	0	0
29	ITS2	d3728918	Cadophora fastigiata	AACGCACA	2	2349	3425	0	0	0	0	1076	0	2349	0	0	0	0	0	0	0	0	0	0	0	0	0	0
30	ITS2	06384654	Chloridium	AACGCACA	1	27704	27704	0	0	27704	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
31	ITS2	f17d593f4	Chloridium	AACGCACA	1	926	926	0	0	0	0	926	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
32	ITS2	6fdaef22b	Chloridium virescens	AACGCACA	1	36497	36497	0	0	0	0	36497	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
33	ITS2	a08ef197c	Chloridium virescens	AACGCACA	1	8092	8092	0	0	0	0	8092	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
34	ITS2	0b45decbc	Chloridium virescens	AACGCACA	1	7285	7285	0	0	0	0	7285	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
35	ITS2	cde14c65f	Eupezizella aureliella;Helotiales sp.	AACGCACA	5	26256	57228	0	0	1703	24131	0	0	0	0	0	1027	0	0	4111	0	0	26256	0	0	0	0	0
36	ITS2	2d449f00b	Eupezizella aureliella;Helotiales sp.	AACGCACA	1	4945	4945	0	0	0	0	0	0	4945	0	0	0	0	0	0	0	0	0	0	0	0	0	0
37	ITS2	c1661c3c1	Eupezizella aureliella;Helotiales sp.	AACGCACA	1	2216	2216	2216	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
38	ITS2	70f8ec4be	Eupezizella aureliella;Helotiales sp.	AACGCACA	1	1587	1587	0	0	0	1587	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
39	ITS2	686f90cfb	Exophiala	AACGCACA	7	16729	40826	2637	2488	0	11688	1222	16729	0	0	0	0	0	0	0	3646	2416	0	0	0	0	0	0
40	ITS2	c5a76128c	Exophiala cancerae	AACGCACA	8	58303	139305	0	44647	0	3565	0	58303	0	0	1234	875	1788	0	0	25917	2976	0	0	0	0	0	0
41	ITS2	f43b167dd	Exophiala cancerae	AACGCACC	2	1533	2738	0	1533	0	0	0	1205	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
42	ITS2	5ddc5805	Fomitopsis ochracea	AACGCACC	4	23553	64152	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	14883	12273	13443	23553	0
43	ITS2	f53bd44c2	Gloeoprus pannocinctus	AACGCACC	4	27905	100221	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	26260	22442	27905	23614	0
44	ITS2	054761dd	Helicosporium lumbricopsis	AACGCACA	2	97161	107656	0	0	0	0	0	0	10495	0	0	0	0	0	0	97161	0	0	0	0	0	0	0
45	ITS2	30828bb1	Helicosporium lumbricopsis	AACGCACC	1	1295	1295	0	0	0	0	0	0	0	0	0	0	0	0	0	1295	0	0	0	0	0	0	0
46	ITS2	fc77a0b84	Helotiales	AACGCACA	8	32792	64441	32792	15473	3061	0	0	0	5691	0	0	717	0	0	0	1978	3583	1146	0	0	0	0	0
47	ITS2	22947cfc1	Helotiales	AACGCACC	1	1103	1103	1103	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
48	ITS2	736336f31	Hyphodontia tropica;Xylodon subflaviporus	AACGCACC	4	457440	557819	0	457440	0	0	0	0	95697	0	3272	1410	0	0	0	0	0	0	0	0	0	0	0
49	ITS2	bd8a1a5a4	Hyphodontia tropica;Xylodon subflaviporus	AACGCACC	1	95733	95733	0	0	0	0	0	0	95733	0	0	0	0	0	0	0	0	0	0	0	0	0	0
50	ITS2	6ab1df4d3	Hyphodontia tropica;Xylodon subflaviporus	AACGCACA	1	3371	3371	0	3371	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
51	ITS2	47312baa4	Hyphodontia tropica;Xylodon subflaviporus	AACGCACC	1	2940	2940	0	0	0	0	0	0	0	0	2940	0	0	0	0	0	0	0	0	0	0	0	0
52	ITS2	e6d66979	Hyphodontia;Xylodon	AACGCACC	1	160313	160313	0	160313	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
53	ITS2	365a9370	Hyphodontia;Xylodon	AACGCACC	1	2002	2002	0	0	0	0	0	0	2002	0	0	0	0	0	0	0	0	0	0	0	0	0	0
54	ITS2	3880e049	Hyphodontia;Xylodon	AACGCACC	1	1957	1957	0	0	0	0	0	0	1957	0	0	0	0	0	0	0	0	0	0	0	0	0	0
55	ITS2	2bf68c8e3	Hyphodontia;Xylodon	AACGCACC	1	1623	1623	0	0	0	0	0	0	1623	0	0	0	0	0	0	0	0	0	0	0	0	0	0
56	ITS2	6bf3155c6	Hyphodontia;Xylodon	AACGCACC	1	1601	1601	0	0	0	0	0	0	1601	0	0	0	0	0	0	0	0	0	0	0	0	0	0
57	ITS2	d435069e	Hyphodontia;Xylodon	AACGCACA	1	1240	1240	0	1240	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
58	ITS2	4190d753	Hypholoma	AACGCACC	1	214155	214155	0	0	0	214155	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
59	ITS2	5071fc6ad	Hypholoma	AACGCACC	1	169574	169574	0	0	0	169574	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
60	ITS2	7e5fd8811	Hypholoma	AACGCACC	1	1476	1476	0	0	0	1476	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
61	ITS2	e1d4ff1ed7	Kneiffiella	AACGCACC	1	1778	1778	0	0	0	0	0	0	0	0	0	1778	0	0	0	0	0	0	0	0	0	0	0
62	ITS2	070da371	Kneiffiella	AACGCACC	1	917	917	0	0	0	0	0	0	0	0	0	917	0	0	0	0	0	0	0	0	0	0	0
63	ITS2	0459f3db2	Kneiffiella	AACGCACC	1	907	907	0	0	0	0	0	0	0	0	0	907	0	0	0	0	0	0	0	0	0	0	0
64	ITS2	54c42d5ec	Kneiffiella	AACGCACC	1	748	748	0	0	0	0	0	0	0	0	0	748	0	0	0	0	0	0	0	0	0	0	0
65	ITS2	26268cad3	Kneiffiella	AACGCACA	1	741	741	0	0	0	0	0	0	0	0	0	741	0	0	0	0	0	0	0	0	0	0	0
66	ITS2	6a5d318ca	Kneiffiella stereicola	AACGCACC	5	589120	826664	0	0	0	0	0	0	0	1566	89455	135080	11443	589120	0	0	0	0	0	0	0	0	0
67	ITS2	c94a617ee	Kneiffiella stereicola	AACGCACC	2	1705	3249	0	0	0	0	0	0	0	0	0	1705	0	1544	0	0	0	0	0	0	0	0	0
68	ITS2	56c242d0	Kneiffiella stereicola	AACGCACC	2	2143	2868	0	0	0	0	0	0	0	0	0	725	0	2143	0	0	0	0	0	0	0	0	0
69	ITS2	7868990b	Kneiffiella stereicola	AACGCACA	2	1757	2458	0	0	0	0	0	0	0	0	0	701	0	1757	0	0	0	0	0	0	0	0	0
70	ITS2	14a68b58	Kneiffiella stereicola	AACGCACC	1	136471	136471	0	0	0	0	0	0	0	0	0	136471	0	0	0	0	0	0	0	0	0	0	0
71	ITS2	ee0d5538e	Kneiffiella stereicola	AACGCACC	1	1579	1579	0	0	0	0	0	0	0	0	0	0	0	1579	0	0	0	0	0	0	0	0	0
72	ITS2	65dd229b	Kneiffiella stereicola	AACGCACC	1	1148	1148	0	0	0	0	0	0	0	0	0	0	0	1148	0	0	0	0	0	0	0	0	0
73	ITS2	0983f5838	Laetiporus caribensis	AACGCACC	4	40447	133725	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	37062	31684	40447	24532	0
74	ITS2	5d8b6500	Laetiporus cincinnatus	AACGCACC	4	37372	117415	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	29044	33020	37372	17979	0

Figure 2. Screenshots of the (a) sample and (b) read reports, using the "m6" ITS2 MiSeq run from Palmer et al. (2018), also used in Figures 3 and 4. Both tables show cells with read counts in the lower right section, using conditional formatting to apply a red background for non-zero entries. Excel shows read counts as "##" where the count is too wide for the column width, as in (a) where the default sample report layout prioritises showing an overview. The column widths in the sample report have been adjusted for display in (b), and the bottom of the table cropped. In this example two fields of user-supplied metadata (sample alias and group) are included in both reports, which have been used for sample sorting and the automatic use of a rainbow of five pastel background colours to visually show the sample groupings. In this case the environmental samples are in pairs. The next fields are from the data itself, reads counts in the samples as raw FASTQ, after read merging with Flash, primer trimming with Cutadapt, the abundance threshold applied, the maximum ASV read count for non-spike-in or spike-in sequences, number of singletons, number of unique ASVs accepted, and the total number of reads for the accepted ASVs. These fields were used to generate Figure 4. The read report also includes the full ASV sequence and its MD5 checksum which is used internally as an identifier, and a concatenation of all the species present in the classifier output as a single field.

and (b) respectively), in both plain text and Excel format. If the user provides suitably formatted sample metadata, cross-referenced by the filename stem, this can be incorporated into the report to make for easier interpretation. Additionally, an edit-graph showing the distances between the ASVs recorded in the sample can be exported (e.g. Figure 3).

Read preparation

The first and slowest stage of the workflow is read preparation. Paired raw Illumina FASTQ files are processed into intermediate FASTA files per amplicon marker containing the ASV sequences and their abundances. It is simplest to run the pipeline on all input data sequentially, but with large projects or for most efficient usage of a computer cluster it is advisable to run the read preparation step in batches, for example by MiSeq plate or sample batch, as separate jobs.

The first step is merging the overlapping FASTQ read pairs, currently done using Flash (Magoč and Salzberg, 2011). This is invoked with the allow “outies” option and maximum overlap increased from the default 65 to 300bp, which was especially important when working with smaller fragments. Initially we used Pear (Zhang et al., 2014), but open source development ended with Pear v0.9.6, and Flash was faster with equivalent output. The merged sequences for each sample are tallied (discarding the per-base quality scores), which avoids re-processing repeated sequences in each sample.

Next, we use cutadapt (Martin, 2011) to identify each amplicon sequence using the primer sequences, which are then removed. These shorter unique sequences in each sample are re-tallied, and unique reads appearing only once in a sample (singletons) are discarded at this point. This gives a list of ASVs with counts per marker per sample.

Earlier versions of the tool and the pre-cursor `metapy` pipeline (Riddell et al., 2019) started by removing the Illumina adapter sequences using Trimmomatic (Bolger et al., 2014), before merging the overlapping reads. Flash was developed before tools like Trimmomatic, and does not require this. Skipping adapter trimming at the start was faster, and made minimal difference to the output, especially since any residual adapter sequence is removed when primer trimming.

Collectively our dataset for the Phyto-Threats project (Green et al., 2021) and related work including natural ecosystems (Riddell et al., 2019), is now over 30 MiSeq plates, with several thousand sequenced samples. To balance performance versus complexity we run the read-preparation by plate. In a typical run on HPC nodes with 2nd-Gen Xeon Scalable (Cascade Lake; 2019) processors preparing the slowest plate took 12.5 minutes, while global tallying through to reporting (see below) added a further 7.5 minutes, giving a total elapsed time of approximately 20 minutes.

Sample tallying and optional read-correction

Once all the FASTQ sample files have been prepared (which is the slowest part of the pipeline), the unique ASVs are tallied per marker per sample. This workflow accommodates large projects where new plates of MiSeq data are sequenced over time, and exploring the effect of adjusting settings like the abundance thresholds.

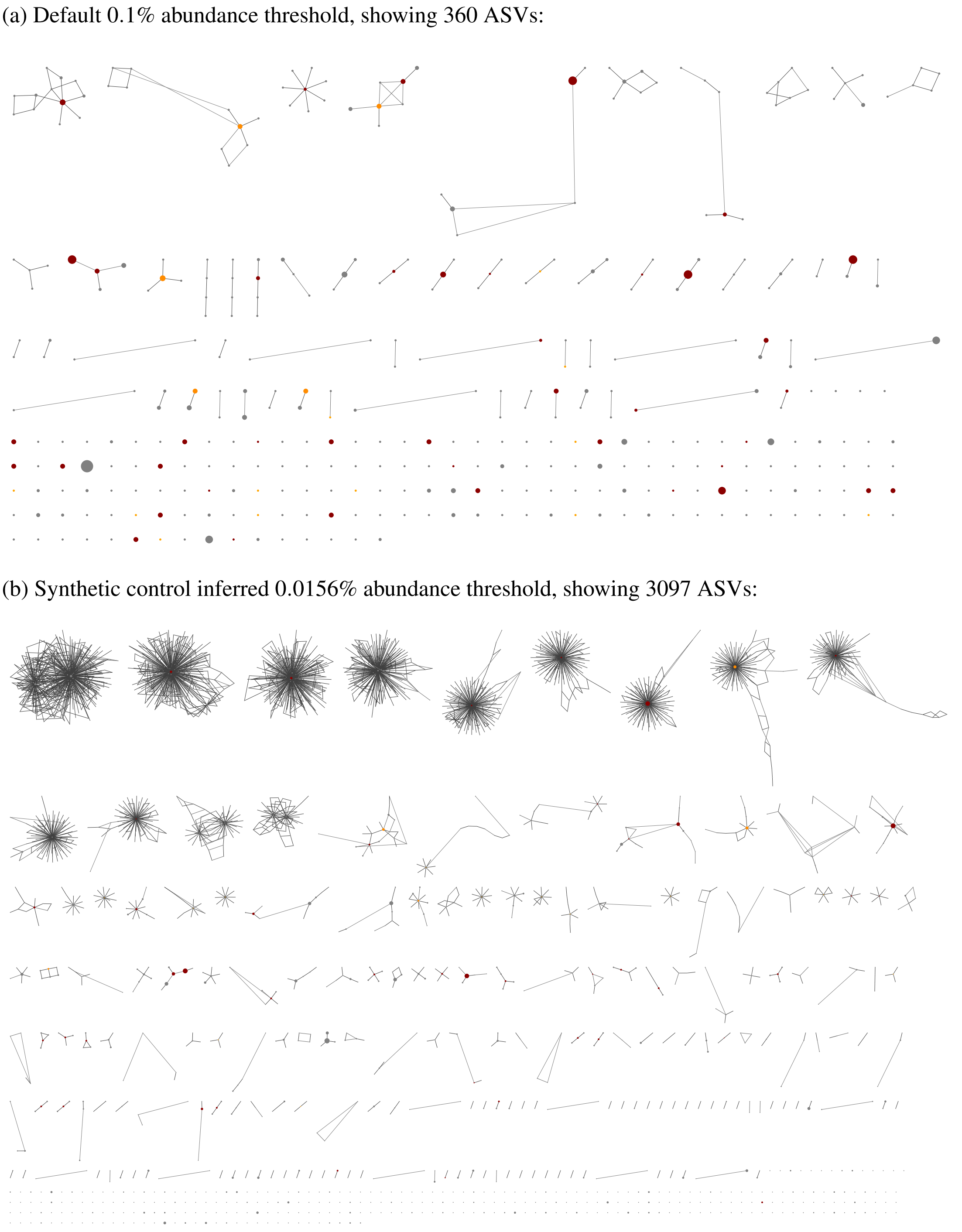
At this point, before applying abundance thresholds (see below), optional read-correction can be applied. This can use our re-implementation of the original UNOISE2 read-correction method as described in Edgar (2016) using the Levenshtein distance as implemented in the Rapid Fuzz library (Bachmann et al., 2022). Alternatively, it can call the later UNOISE3 algorithm via Edgar’s command line tool `usearch`, or as reverse engineered in `vsearch` (Rognes et al., 2016).

The ASV sample tally table is output as a plain text tab-separated variable (TSV) file, and optionally in the Biological Observation Matrix (BIOM) format facilitating use with alternative classifiers (McDonald et al., 2012).

Abundance thresholds

There are two compelling reasons to impose abundance thresholds. Firstly, most rare ASVs including singletons are generated via errors in either the PCR amplification or sequencing steps (Edgar (2016); Figure 4), and their removal improves the signal to noise ratio and results in a marked improvement in computational efficiency. Secondly it plays a key role in dealing with cross-sample contamination, including Illumina tag-switching (Schnell et al., 2015).

The tool implements both an absolute minimum abundance threshold defaulting to 100 copies (based on examination of our own datasets), and a fractional threshold defaulting to the widely used value of 0.1% (Muri et al., 2020) of the paired reads in a sample which passed primer trimming for that marker.



(a) Default 0.1% abundance threshold, showing 360 ASVs:

(b) Synthetic control inferred 0.0156% abundance threshold, showing 3097 ASVs:

Figure 3. Example ASV edit-graph, exported as an XGMML format graph, then opened in Cytoscape v3.9.1 (Shannon et al., 2003) showing ITS2 sequences from the same Palmer et al. (2018) MiSeq run shown in Figures 2 and 4. Each node represents an ASV, orange if matched to a synthetic control, dark red for a matched genus, grey otherwise. The node circles are scaled according to the number of samples it was in, and shown here without labels for clarity. The edges are solid for a one base pair edit distance, dashed for a two base pair edit distance, and dotted for a three base pair edit distance. The nodes were arranged in CytoScape using edge weighted prefuse force directed layout, and their placement is not consistent between (a) and (b). As the abundance threshold is lowered from (a) to (b), the number of nodes increases roughly ten-fold. The more common ASV nodes become the centre of a halo of 1bp variants, typically each seen in a single sample, which we attribute to PCR noise and/or sequencing error.

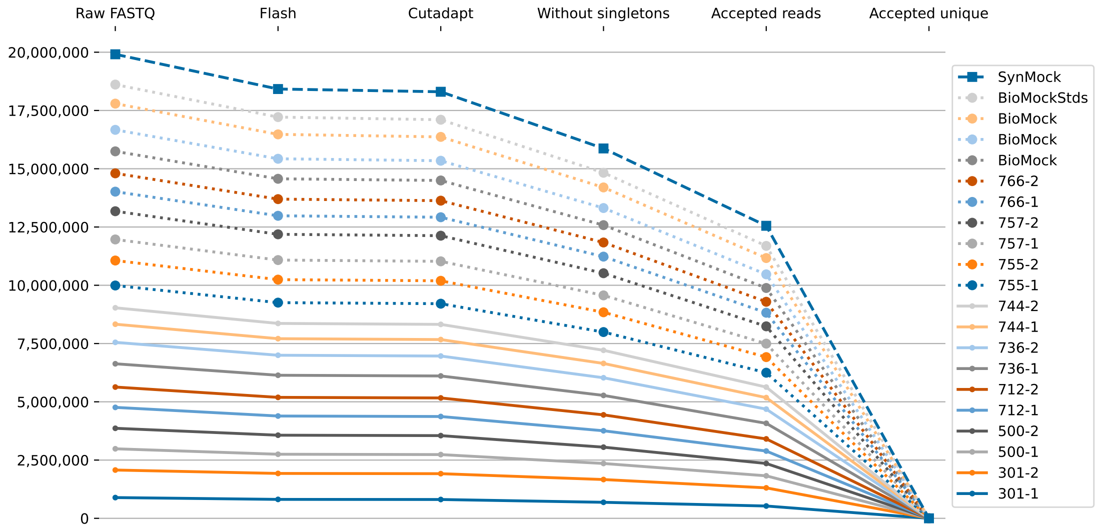


Figure 4. Stacked line graph illustration of how the raw FASTQ read counts are reduced to ASV tallies, showing reads counts from ITS2 sequences from the same Palmer et al. (2018) MiSeq run shown in Figures 2 and 3. Starting from raw FASTQ files, over 90% could be merged into overlapping reads, most of which could be primer trimmed. At this point the data is already held as ASV tally tables internally. The next drop represents removing singletons, leaving about 80% of the starting reads. Applying the default minimum abundance thresholds drops this to just over 60% of the original reads. The final drop off shown, from millions to hundreds of sequences, is to illustrate switching from counting reads to counting unique sequences (ASV) as a tally table. The samples are the synthetic control, biological mocks, and then numerical codes for environmental samples.

These are applied to each sample. The fractional threshold is more appropriate than an absolute threshold if the sampling depth varies dramatically between samples. The default absolute threshold may be too high for low yield runs like the Illumina Nano Kit protocol, or if the focus is maximising sensitivity. An ASV supported by a single read is known as a singleton, and for efficiency these are always automatically excluded. In most cases singletons are a single base pair away from a more dominant sequence, and are presumed to originate from amplification or sequencing errors, resulting in a halo effect when visualised as an edit-graph (see Figure 3). In such cases, read correction would map them to that central node, but this is not always clear cut as there can be multiple high abundance high occurrence adjacent nodes. Unlike the tools DADA2 (Callahan et al., 2016), obiclean (De Barba et al., 2014; Boyer et al., 2016), or UNOISE2 (Edgar, 2016), we default to simply excluding these reads via the abundance threshold.

Another source of unwanted low abundance sequences comes from Illumina tag-switching (Schnell et al., 2015). Using metabarcoding synthetic controls, Palmer et al. (2018) quantified the effective rate as under 0.02%, consistent with between 0.01% and 0.03% of reads in earlier work reviewed by Deiner et al. (2017). However, while excluding low abundance variants from PCR noise and tag-switching is important, as in Muri et al. (2020) we use a higher default of 0.1% for excluding most contamination. The tool supports a data-driven minimum abundance threshold using (unwanted) amplification in negative control samples, a widely used strategy (Sepulveda et al., 2020). The control samples are processed before the non-controls, in order to infer and apply a potentially higher control-driven threshold to the other samples in that batch. Sample batches are defined by providing input data in sub-folders, which could be MiSeq runs, or reflect samples amplified together.

Simple blank negative controls should contain no sequences at all, so the highest abundance sequence present can be used as an inferred absolute abundance threshold (if higher than the default), and applied to all the samples in that batch. Massart et al. (2022) caution however that trace levels of DNA in an otherwise empty control may amplify very efficiently, overestimating contamination, and so recommend a spike-in or positive control approach.

If the experiment uses synthetic sequences spiked into a negative control, it is possible to distinguish the expected spike-in sequences (subject only to technical noise and artifacts from PCR and sequencing) from biological contamination from laboratory practices (Palmer et al., 2018). In principle a biological out-group or “alien control” could be used as the spike-in (Massart et al., 2022), but unique novel synthetic control sequences will provide the greatest confidence. Provided the tool can identify and thus ignore the spike-in sequences, any remaining reads in those controls can be used to raise the absolute threshold. However, the percentage of the most abundant non-spike-in sequence can be taken as an inferred fractional abundance threshold (if higher than the default). Palmer et al. (2018) takes a more optimistic approach in their tool AMPtk by applying ASV specific thresholds, assuming the other biological sequences not observed as cross contaminants are well behaved. THAPBI PICT takes the more pessimistic approach of taking the worse case as indicative of cross contamination rates for that sample batch in general.

In our own data, (cross-)sample contamination appears to be more of an issue than Illumina tag-switching. At the time of writing we have completed 30 *Phytophthora* ITS1 MiSeq sequencing runs with synthetic control samples, covering plant nurseries (Green et al., 2021) and environmental samples. One run was discarded after finding 1 to 5% non-synthetic reads in all the controls, traced to contamination of the PCR master mix. Another problematic run saw 4 of the 6 controls in a 96-sample amplification plate with over 2% non-synthetic reads. These had an identical ASV profile, suggesting a single contamination event after pipetting the first two controls. The dominant contaminant here was a rare *Phytophthora* species not seen on the rest of the samples being sequenced, making the most likely contamination source DNA from an older sample previously processed in the laboratory. Thalinger et al. (2021) has a number of recommendations on the laboratory side for minimising contamination events. By using the worst control non-synthetic read fractions as thresholds for those plates we reduce the chances of false positives, at the cost of false negatives for minority community members. This is not unprecedented - for example guided by their mock community controls, Hänfling et al. (2016) used thresholds of 0.3% and 1% for their 12S and cytB amplicons respectively (and an absolute threshold of at least 3 reads per ASV).

CLASSIFIERS AND DATABASES

Classifier implementations

All of the classifiers in THAPBI PICT are based on independent comparisons of each ASV to the sequences in the database as strings of letters. There is no clustering, meaning the classification can be performed on a global listing of all ASV, without considering the context of what other sequences were present in the same samples.

Technically the `identity` classifier does the matching with an SQL query within SQLite. For performance the `substr` classifier is done in Python after loading all the database sequences into memory. The edit distance based classifiers also load all the sequences into memory, and then use the Levenshtein metric as implemented in the Rapid Fuzz library (Bachmann et al., 2022), where a one base-pair insertion, deletion, or substitution is considered an edit distance of one. All our distance classifiers accept a species level match at most one base pair away, equivalent to about a 99.5% identity threshold (assuming amplicons around 200bp long). This may seem high, but historic thresholds like 97% for the 16S marker are too relaxed (Edgar, 2018). The least stringent classifier currently implemented (`blast`) assigns the species of the best BLAST nucleotide match within the database Camacho et al. (2009), ranked by bit-score subject to a minimum alignment length and score intended to exclude the most obvious false positives. In objective assessment (see below), this does over-predict (assigning sometimes tenuous species matches). This BLAST based classifier should only be used for preliminary analyses like exploring a new dataset with an uncurated database.

Database and classifier interactions

The tool has been designed as a framework which can be applied to multiple biological contexts, demonstrated in the worked examples discussed below. In each case, a relevant reference database will need to be compiled.

Applied to environmental samples, some primer pairs will amplify a much wider sequence space than others, either reflecting a more diverse genome region, or simply from having longer amplicons. Related to this, the fraction of observed sequences with a published reference will also vary - a problem particularly in understudied organisms, or with novel barcoding amplicons. This means the density of the references in experimentally observed sequence space is context dependant, and thus so is the most appropriate classifier algorithm.

The default classifier allows perfect matches, or a single base pair (bp) difference (substitution, insertion or deletion). This requires good database coverage with unambiguous sequences trimmed to the amplicon only, which we have been able to achieve for the *Phytophthora* ITS1 region targeted. This classifier can still be used with reference sequences containing a single IUPAC ambiguity code (which will count as the single allowed mismatch), but more than that and the reference could only be used with a less stringent classifier (such as the best BLAST nucleotide match).

Default ITS1 database and conflict resolution

Our chosen ITS1 primers target a region of eukaryote genomes between the 18S and 5.8S rRNA genes, with nested PCR primers to selectively target *Phytophthora* (Scibetta et al., 2012), related paraphyletic genera of downy mildews and the sister taxa *Nothophytophthora*. They have been observed to occasionally amplify related genera, such as *Pythium* and *Phytopythium*, especially when *Phytophthora* levels in the sample are very low. Our curated database initially focused on *Phytophthora*, building on the work in Català et al. (2015) and Riddell et al. (2019). Published ITS1 sequences are often truncated to the start of the ITS1 region, and thus omit our left primer and the highly conserved 32bp section of the 18S region at the start of our amplicon of interest, which handicapped building a reference set. In addition to using public sequences, we also performed additional Sanger capillary sequencing. Also, given that *Phytophthora* rRNA is known to be present in variable numbers of copies in a tandem array with potential variability between copies, we also ran some single isolates from culture collections through the MiSeq pipeline which determined that many species were uniform but others revealed secondary ITS1 variants. The primary goal was classification of the genus *Phytophthora*, but widening coverage to downy mildews and related genera such as *Nothophytophthora* and the rarely amplified *Pythium* created two additional challenges. First, there are fewer published sequences available, and thus the default classifier becomes too strict to assign many species. The Phyto-Threats project therefore uses a more relaxed classifier which falls back on a genus level classification based on the closest database entries up to 3bp edits away. Second, the taxonomic annotation becomes less consistent, particularly within the former *Pythium* genus that was subject to taxonomic revision that generated new genera such as *Globisporangium* or *Phytopythium*. This led to many conflicts with database accessions of (near) identical ITS1 sequences having different genus names. These direct conflicts, and similar cases of apparent misannotation, were resolved manually by excluding the unwanted accessions in the database build script.

With any amplicon marker, it is possible that distinct species will share the exact same sequence. For example, this happens with model organism *Phytophthora infestans* and sister species such as *P. andina* and *P. ipomoeae*. In such cases the classifier reports *all* equally valid taxonomic assignments. The database author could instead record a single assignment like *Phytophthora infestans*-complex. Conversely, some *Phytophthora* genomes are known to contain multiple copies of our target marker ITS1 through tandem repeats of the rDNA ITS region. In such cases the recognised variant forms should be added to the reference database. Despite their shortcomings, the ITS1 region has remained the de-facto standard within the oomycete community (Robideau et al., 2011), but alternatives are being explored (Foster et al., 2022).

CLASSIFICATION ASSESSMENT

In assessing classification performance, it is the combination of both classification method (algorithm) and marker database which matters. Settings like the abundance threshold are also important, and the tool default settings partly reflect one of the original project goals being to avoid false positives.

To objectively assess a metabarcoding classifier we require sequenced samples of known composition, which generally means single isolates (where a single marker sequence is typically expected), or mock communities of known species (the bulk of our examples). Carefully controlled environmental samples may also be used, such as Muri et al. (2020) in our worked examples. Here a lake was drained to collect and identify all the individual fish, but this is problematic as the lake was large enough that DNA from each fish could not be expected at all the sampling points, giving an inflated false negative count.

Our tool includes a presence/absence based assessment framework based on supplying expected species lists for control samples, from which the standard true positive (TP), false positive (FP), true negative (TN), and false negative (FN) counts can be computed for each species. These are the basis of standard metrics like sensitivity (recall), specificity, precision, F-score (F-measure, or F1), and Hamming Loss. It is simple but not overly helpful to apply metrics like this to each species, rather the overall performance is more informative.

However, some scores like the Hamming Loss are fragile with regards to the TN count when comparing databases. The Hamming Loss is given by the total number of mis-predicted class entries divided by the number of class-level predictions, thus $(FP+FN)/(TP+FP+FN+TN)$. Consider a mock community of ten species, where the classifier made 11 predictions which break down as 9 TP and 2 FP, meaning $10-9=1$ FN. Suppose the database had a hundred species (including all ten in the mock community), that leaves $100-9-1-2=88$ TN, and a Hamming Loss of $3/100=0.03$. Now suppose the database was extended with additional references not present in this mock community, perhaps expanding from European *Phytophthora* species to include distinct entries for tropical species, or a sister group like *Peronospora*. The denominator would increase, reducing the Hamming Loss, but intuitively the classifier performance on this mock community has not changed. To address this, the classifier assessment also includes a modified *ad hoc* loss metric calculated as the total number of mis-predicted class entries divided by the number of class-level predictions ignoring TN, or $(FP+FN)/(TP+FP+FN)$ which in this example would give $3/12=0.25$ regardless of the number of species in the database. This is an intuitive measure weighting FP and FN equally (smaller is better, zero is perfect), a potential complement to the F-score.

Note that the assessment framework only considers species level predictions, ignoring genus only predictions and unknowns, and thus will not distinguish between the default `onebp` classifier and variants like `1s3g` (see Table 1).

As a benchmark of the default classifier and *Phytophthora* focused database, we used the 10 and 15 species mixes in Riddell et al. (2019), see Table 2. This was originally analysed with the `metapy` pipeline with a high stringency classifier using `bowtie` to find perfect alignments, and a more relaxed classifier using `swarm` for clustering. In both samples and both classifiers, *Phytophthora boehmeriae* was not found, and this was attributed to uncompetitive amplification in a mixed DNA sample due to poor PCR primer binding. That being so, the best classifier results would be either 14 TP and 9 TP respectively, with 0 FP if the markers were unique, and 1 FN. Note however not all the markers are unique, both mixes contain species known to share their ITS1 marker with other species, giving unavoidable technical FP, also discussed in Riddell et al. (2019).

Using F1 score or our *ad hoc* loss ranking, THAPBI PICT is clearly performing best on the 10 species mix (and better than metapy did). However, with default settings it does worse on the 15 species mix, due the high FN count where the default ASV abundance threshold of 100 reads is excluding expected species. In this MiSeq dataset the Illumina Nano Kit was used giving lower yields, making the default 100 read threshold overly harsh. Optimising on maximising the F1 score and minimising ad-hoc-loss, and weighting the two communities equally, suggests running THAPBI PICT with an ASV read abundance threshold of around 50 reads performs best overall here. This is a fundamental problem however, low abundance community members can be indistinguishable from background noise/contamination, meaning without controls the best threshold is arbitrary.

REPORTING

The pipeline produces two tabular reports (which can also be requested directly with the `summary` command), output as both tab-separated plain text, and Excel format with colouring and conditional formatting (Figure 2). These include information on read counts from the preparation stage (as used in Figure 4), information on the abundance thresholds, and foremost the species classification from the chosen method. The user may provide a table of metadata cross referenced by the sample FASTQ filename stem, which will be used for sorting the samples and if possible colouring inferred sample groupings (e.g. sample source, or replicates) to ease interpretation. This allows quick visual comparison of replicates as adjacent rows/columns.

The read report by default sorts the ASVs by their taxonomic classification, and then by abundance. This makes it easy to identify the most common unknowns or genus-only predictions for manual review (using the ASV sequence). This sorting also means that when the thresholds are low enough to let through noise, the grey halo effect shown in the edit graph (see Figure 3) is also visually distinct as highly abundant rows followed by less abundance variants. This read report can also be exported in BIOM format.

For many of the worked examples the sample metadata on the NCBI Short Read Archive (SRA) or European Nucleotide Archive (ENA) had to be supplemented by information in the associated publication. Providing such metadata to the archives using an approved ontology based checklist is non-trivial, but adds greatly to the reuse potential (Tedersoo et al., 2015). We provide an `ena-submit` command which

Mock community	Method	TP	FP	FN	F1	*Ad hoc* loss
15 species mix	metapy/bowtie	11	1	4	0.81	0.333
15 species mix	metapy/swarm	14	4	1	0.85	0.263
15 species mix	THAPBI PICT (defaults)	8	2	7	0.64	0.529
15 species mix	THAPBI PICT (50 reads)	11	3	4	0.76	0.389
15 species mix	Theoretical best	14	0	1	0.97	0.067
10 species mix	metapy/bowtie	7	6	3	0.61	0.563
10 species mix	metapy/swarm	9	10	1	0.62	0.550
10 species mix	THAPBI PICT (defaults)	8	2	2	0.80	0.333
10 species mix	THAPBI PICT (50 reads)	8	2	2	0.80	0.333
10 species mix	Theoretical best	9	0	1	0.95	0.100
Combined	metapy/bowtie	18	7	7	0.72	0.438
Combined	metapy/swarm	23	14	2	0.74	0.410
Combined	THAPBI PICT (defaults)	16	4	9	0.71	0.448
Combined	THAPBI PICT (50 reads)	19	5	6	0.78	0.367
Combined	Theoretical best	23	0	2	0.96	0.080

Table 2. Species level classifier assessment on the Riddell et al. (2019) mock communities, with TP and FP counts from their Table 1, and FN counts from their text. THAPBI PICT using default settings has an abundance threshold of 100 reads, also shown using just 50 reads. The theoretical best assumes everything except *Phytophthora boehmeriae* could be found, and ignores that some of the ITS1 amplicons are ambiguous at species level. F1 score or F-measure calculated as $2TP/(2TP+FP+FN)$, given to 2dp. *Ad hoc* loss defined as $(FP+FN)/(TP+FP+FN)$, given to 3dp.

facilitates using the interactive ENA upload step for matching FASTQ filenames to previously entered sample information.

The tool's repository includes a number of helper scripts, including a pooling script written for the Phyto-Threats project for preparing plant nursery specific summary reports. This simplifies the sample report by combining replicate samples into a single row, and can either use the read count sum, or just "Y" (present) or "N" (absent).

The other noteworthy report from the tool is an edit graph, invoked via the `edit-graph` command, as shown in Figure 3. By default this outputs the edit graph in XGMML format which can then be visualised in a tool like Cytoscape (Shannon et al., 2003), with a choice of node layouts and representations (e.g. customising node size by sample count, or colour by genus). The graph can help guide the choice of minimum abundance threshold (as discussed above), and the choice of classifier. In the example shown with a 3bp maximum edit-distance shown, the cliques formed are for the most part clearly distinct species, with a single central node. With the default ITS1 marker used for *Phytophthora* we find greater sequence variation and therefore more diverse non-simple clusters for species like *Phytophthora nicotianae* and *P. gonapodyides*, but most species show a single central ITS1 sequence.

DEVELOPMENT PRACTICES

THAPBI PICT is released as open source software under the MIT licence. It is written in Python, a free open source language available on all major operating systems. Version control using git hosted publicly on GitHub at `https://github.com/peterjc/thapbi-pict/` is used for the source code, documentation, and database builds including tracking the hand curated reference set of *Phytophthora* etc ITS1 sequences. Continuous integration of the test suite is currently run on both CircleCI (Linux) and AppVeyor (Windows). Software releases are to the Python Packaging Index (PyPI) as standard for the Python ecosystem, and additionally packaged for Conda via the BioConda channel (Grüning et al., 2018). This offers simple installation of the tool itself and all the command line dependencies on Linux or macOS. Installation on Windows requires manual installation of some dependencies. The documentation is currently hosted on Read The Docs, updated automatically from the GitHub repository.

WORKED EXAMPLES

In this section we briefly discuss the application of THAPBI PICT to public data sets from several published papers, covering a range of organisms and markers. The selection has prioritised examples including mock communities and negative controls, and have been included in the tool documentation as worked examples. These worked examples generally are highly concordant with the published analyses, with differences largely down to the exact choice of thresholds.

The example scripts first-run times range from a few minutes with under 1GB of raw FASTQ data (Bakker, 2018; Riddell et al., 2019; Walker et al., 2019; Muri et al., 2020), to a few hours with the larger datasets like Ahmed et al. (2019) with 12GB of input. These times are dominated by the merging the paired reads during read preparation stage, and as discussed earlier, running the read preparation stage in parallel across a cluster is advised on larger projects.

The first worked example is a simple one using the provided *Phytophthora* ITS1 database we have generated for this work to reexamine Riddell et al. (2019). This example does not include the synthetic controls introduced later, but does have blanks as negative controls and simple mock communities as DNA mixes (discussed above for classifier assessment). The second example uses Redekar et al. (2019) but focuses on how to build a database, including how species names can optionally be validated against the NCBI taxonomy.

The example based on Muri et al. (2020) is a single 12S marker for fish, with a custom database including numerous off-target matches like humans and sheep. In this case the lake contents were determined by draining the lake and collecting the fish, but this did not determine which of the sampling sites any given fish might have visited, complicating interpretation compared to an artificial mock community. Another single marker example based on Walker et al. (2019) uses COI in simple mock communities of bats, and shows the importance of the database content with the default classifier. The most interesting single marker example is based on Palmer et al. (2018), fungal ITS2 markers with mock biological fungal communities and *synthetic control* sequences. This has been discussed above in the context of setting abundance thresholds.

There are examples with multiple markers which were sequenced *separately* in Klymus et al. (2017), two different 16S mitochondrial markers with mock communities, and Ahmed et al. (2019), four different markers in mock nematode communities. The example in Batovska et al. (2021) uses three markers together, while Arulandhu et al. (2017) sequences over a dozen markers together. Here the primer sequences themselves are non-overlapping and so serve to separate out the amplicons for each sample, allowing them to be matched to the relevant reference set. Note currently a primer cocktail as used for the COI example in this data set is not supported. This paper is also noteworthy as an inter-laboratory replication study of metabarcoding.

Datasets from some potentially useful publications could not be used directly, generally for technical reasons. Many used custom multiplexing (Elbrecht and Leese, 2015; Port et al., 2016; Elbrecht et al., 2016, 2017, 2019; Braukmann et al., 2019), and thus would require separate de-multiplexing before use. Some like Braukmann et al. (2019) and Duke and Burton (2020), use an amplicon too long to span with overlapping Illumina MiSeq paired reads. Sometimes however papers did not provide the *raw* FASTQ files. For instance, Blanckenhorn et al. (2016) did not share the raw FASTQ files at all, while Hänfling et al. (2016) and Zaiko et al. (2022) provided primer trimmed FASTQ files. Some older papers (also) used the Roche 454 or Ion Torrent platforms, which would require re-engineering mainly around the different error profile, which is potentially unsuited for our default strict classifier.

DISCUSSION

Here we present a novel and flexible pipeline for the objective analysis of metabarcode data, not just of single markers but also pooled markers where the amplicons can be separated via their primer sequences. Some of the design choices and default settings reflect the initial use case being *Phytophthora* ITS1 sequence markers in a context where specificity was favoured over sensitivity. In general, appropriate abundance thresholds and classifier algorithm will be experiment and/or marker specific, with the quality of the reference database a key factor. All amplicon barcoding experiments should be designed with suitable controls to assess the limits of quantification versus presence/absence (Lamb et al., 2019), including the effects of the PCR (Thielecke et al., 2017) and contamination (Thalinger et al., 2021).

By design, the tool currently reports lists of genus/species for each ASV, without attempting anything like a most recent common ancestor analysis. This limitation can be a handicap with some use-cases where the markers may not readily resolve at species level, and/or an ASV is often shared between genera. See for example, the Brassicaceae discussed in Arulandhu et al. (2017), and fish examples in Muri et al. (2020). Moreover, it makes the tool unsuited to markers like regions of the bacterial 16S rRNA gene which are typically used at phylum level with environmental datasets (Straub et al., 2020). Rather it is appropriate for comprehensive analyses of better defined taxonomic markers such as the plant pathogenic oomycete ITS1 marker used primarily for *Phytophthora*, where it is proving valuable for the ongoing characterisation of a comprehensive set of several thousand samples from plant nurseries in the Phyto-Threats project (Green et al., 2021) and in natural ecosystems (Riddell et al., 2019).

Our pipeline supports using negative or synthetic spike-in controls to set an abundance threshold on groups of samples (such as each sequencing run). Rather than ASV-specific thresholds as in Palmer et al. (2018), THAPBI PICT takes the more cautious approach of interpreting the worst case as indicative of cross contamination rates for that sample batch in general.

The pipeline does not currently explicitly attempt to find and remove chimera sequences beyond the use of abundance filters. As discussed in Edgar (2016), chimeras which are also only 1bp away from a reference sequence cannot be distinguished from a point error, and would be matched to that reference by all but our strictest identity classifier. Apart from this special case, any high abundance chimera would likely appear in our reports as an unknown, and would most likely be only in a single sample. Regular manual inspection of the high abundance unknown reads appearing in multiple samples was part of the ongoing quality assurance during the Phyto-Threats project, primarily to identify any gaps in the database coverage. The only clear chimeras identified were from our synthetic controls, where part of our non-biological sequence was fused to some unexpected sequence. Potentially more complex mock communities of synthetic sequences could be used to generate a gold standard for identifying chimeras which might serve as a benchmark dataset for testing chimera algorithms.

Another important difference from other ASV based tools like DADA2 (Callahan et al., 2016), obitools (Boyer et al., 2016) and UNOISE2 (Edgar, 2016), is THAPBI PICT does not by default attempt read correction. From the halo pattern of PCR induced variants seen from synthetic inputs as viewed on an edit-graph, like Figure 4(b), there is usually a natural central node to which a variant can be attributed. However, the situation is not always clear cut, with some species like *Phytophthora gonapodyides* showing a range of known ITS1 sequences. Rather our approach is to exclude most PCR noise through the abundance filters, and allow a modest amount of variation when matching the higher abundance sequences to the reference database. As an option however, the pipeline can apply our re-implementation of the original UNOISE2 Levenstein distance based read-correction described in Edgar (2016), or invoke the UNOISE3 algorithm implemented in the USEARCH or VSEARCH tools. Read-correction seems most appropriate where the reference sequences are well separated, unlike our default *Phytophthora* ITS1 amplicon where a single base pair can distinguish known species, and thus read correction can mask lower abundance species under their more abundant neighbours.

Examination of mock community samples of our synthetic spike-in sequences showed ASV abundance to be at best semi-quantitative, as found in other work (Palmer et al., 2018; Lamb et al., 2019). For the Phyto-Threats project reports sent to plant nursery owners we therefore only use species presence/absence (above or below the abundance threshold, and pooled replicates). However, the raw abundances are in the main tool reports, and can be used for plots or a quantitative interpretation where appropriate. The nested primer protocol with two rounds of PCR may be a factor in undermining quantitative interpretation, and increasing the risk of cross-sample or other sample contamination.

CONCLUSION

Here we present a novel and flexible pipeline for the objective analysis of metabarcode data, with user friendly reports including ASV read counts enabling custom graphs, as well as summary species lists per sample. While initially designed for *Phytophthora* ITS1 sequence markers, the THAPBI PICT tool can be applied more generally, including to samples containing multiple marker regions. It is best suited to markers where ASV are at least genus specific. Care should be taken picking appropriate abundance thresholds, which can be set using negative and/or synthetic controls, and in applying read-correction for de-noising. While high-throughput amplicon sequencing does give read counts per species (or per ASV), we and others caution against treating these as quantitative (Palmer et al., 2018; Lamb et al., 2019). The most suitable classifier algorithm will be marker specific, with the quality and coverage of the reference database a key factor. Including mock communities in your experiment allows the performance of classifier and database to be evaluated objectively.

ACKNOWLEDGMENTS

We thank our colleagues at the James Hutton Institute including Eva Randall, Beatrix Keillor, Pete Hedley, and Jenny Morris, and at Forest Research including Sarah Green, Debra Frederickson Matika and Carolyn Riddell. The reviewers are thanked for their constructive feedback, Celine Mercier, Alexander Piper, and the anonymous Reviewer 3.

The authors acknowledge the Research/Scientific Computing teams at The James Hutton Institute and NIAB for providing computational resources and technical support for the “UK’s Crop Diversity Bioinformatics HPC” (BBSRC grant BB/S019669/1), use of which has contributed to the results reported within this paper.

REFERENCES

Ahmed, M., Back, M. A., Prior, T., Karssen, G., Lawson, R., Adams, I., and Sapp, M. (2019). Metabarcoding of soil nematodes: the importance of taxonomic coverage and availability of reference sequences in choosing suitable marker(s). *Metabarcoding and Metagenomics*, 3:e36408.

Arulandhu, A. J., Staats, M., Hagelaar, R., Voorhuijzen, M. M., Prins, T. W., Scholtens, I., Costessi, A., Duijsings, D., Rechenmann, F., Gaspar, F. B., Barreto Crespo, M. T., Holst-Jensen, A., Birck, M., Burns, M., Haynes, E., Hochegger, R., Klingl, A., Lundberg, L., Natale, C., Niekamp, H., Perri, E., Barbante, A., Rosec, J.-P., Seyfarth, R., Sovová, T., Van Moorleghem, C., van Ruth, S., Peelen, T., and Kok, E. (2017). Development and validation of a multi-locus DNA metabarcoding method to identify endangered species in complex samples. *GigaScience*, 6:gix080.

Bachmann, M., layday, Fihl-Pearson, J., Schreiner, H., and Sherman, M. (2022). maxbachmann/rapidfuzz: Release 2.11.1.

Bakker, M. G. (2018). A fungal mock community control for amplicon sequencing experiments. *Molecular Ecology Resources*, 18:541–556.

Batovska, J., Piper, A. M., Valenzuela, I., Cunningham, J. P., and Blacket, M. J. (2021). Developing a non-destructive metabarcoding protocol for detection of pest insects in bulk trap catches. *Scientific Reports*, 11:7946.

Blanckenhorn, W. U., Rohner, P. T., Bernasconi, M. V., Haugstetter, J., and Buser, A. (2016). Is qualitative and quantitative metabarcoding of dung fauna biodiversity feasible? *Environmental Toxicology and Chemistry*, 35(8):1970–1977.

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, 30:2114–2120.

Boyer, F., Mercier, C., Bonin, A., Le Bras, Y., Taberlet, P., and Coissac, E. (2016). obitools: a unix-inspired software package for DNA metabarcoding. *Molecular Ecology Resources*, 16(1):176–182.

Braukmann, T. W. A., Ivanova, N. V., Prosser, S. W. J., Elbrecht, V., Steinke, D., Ratnasingham, S., de Waard, J. R., Sones, J. E., Zakharov, E. V., and Hebert, P. D. N. (2019). Metabarcoding a diverse arthropod mock community. *Molecular Ecology Resources*, 19(3):711–727.

Callahan, B. J., McMurdie, P. J., and Holmes, S. P. (2017). Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *The ISME Journal*, 11:2639–2643.

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., and Holmes, S. P. (2016). DADA2: High-resolution sample inference from Illumina amplicon data. *Nature Methods*, 13:581–583.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and 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., 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.

Català, S., Pérez-Sierra, A., and Abad-Campos, P. (2015). The use of genus-specific amplicon pyrosequencing to assess *Phytophthora* species diversity using eDNA from soil and water in northern spain. *PLOS ONE*, 10(3):1–14.

Cock, P. J. A., Cooke, D. E. L., Thorpe, P., and Pritchard, L. (2023). THAPBI PICT - a fast, cautious, and accurate metabarcoding analysis pipeline. *bioRxiv*.

De Barba, M., Miquel, C., Boyer, F., Mercier, C., Rioux, D., Coissac, E., and Taberlet, P. (2014). Dna metabarcoding multiplexing and validation of data accuracy for diet assessment: application to omnivorous diet. *Molecular Ecology Resources*, 14(2):306–323.

Deiner, K., Bik, H. M., Mächler, E., Seymour, M., Lacoursière-Roussel, A., Altermatt, F., Creer, S., Bista, I., Lodge, D. M., de Vere, N., Pfrender, M. E., and Bernatchez, L. (2017). Environmental DNA metabarcoding: Transforming how we survey animal and plant communities. *Molecular Ecology*, 26(21):5872–5895.

Duke, E. M. and Burton, R. S. (2020). Efficacy of metabarcoding for identification of fish eggs evaluated with mock communities. *Ecology and Evolution*, 10(7):3463–3476.

Edgar, R. C. (2013). UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nat. Methods*, 10(10):996–998.

Edgar, R. C. (2016). UNOISE2: improved error-correction for Illumina 16S and ITS amplicon sequencing. *bioRxiv*.

Edgar, R. C. (2018). Updating the 97% identity threshold for 16S ribosomal RNA OTUs. *Bioinformatics*, 34(14):2371–2375.

Elbrecht, V., Braukmann, T. W. A., Ivanova, N. V., Prosser, S. W. J., Hajibabaei, M., Wright, M., Zakharov, E. V., Hebert, P. D. N., and Steinke, D. (2019). Validation of COI metabarcoding primers for terrestrial arthropods. *PeerJ*, 7:e7745.

Elbrecht, V. and Leese, F. (2015). Can DNA-based ecosystem assessments quantify species abundance? testing primer bias and biomass – sequence relationships with an innovative metabarcoding protocol. *PLOS ONE*, 10(7):1–16.

Elbrecht, V., Peinert, B., and Leese, F. (2017). Sorting things out: Assessing effects of unequal specimen biomass on DNA metabarcoding. *Ecology and Evolution*, 7(17):6918–6926.

Elbrecht, V., Taberlet, P., Dejean, T., Valentini, A., Usseglio-Polatera, P., Beisel, J.-N., Coissac, E., Boyer, F., and Leese, F. (2016). Testing the potential of a ribosomal 16s marker for DNA metabarcoding of insects. *PeerJ*, 4:e1966.

Foster, Z. S. L., Albornoz, F. E., Fieland, V. J., Larsen, M. M., Jones, F. A., Tyler, B. M., Nguyen, H. D. T., Burgess, T. I., Riddell, C., Voglmayr, H., Martin, F. N., and Grünwald, N. J. (2022). A new oomycete metabarcoding method using the rps10 gene. *Phytobiomes Journal*, 6(3).

Green, S., Cooke, D. E. L., Dunn, M., Barwell, L., Purse, B., Chapman, D. S., Valatin, G., Schlenzig, A., Barbrook, J., Pettitt, T., Price, C., Pérez-Sierra, A., Frederickson-Matika, D., Pritchard, L., Thorpe, P., Cock, P. J. A., Randall, E., Keillor, B., and Marzano, M. (2021). Phyto-threats: Addressing threats to UK forests and woodlands from *Phytophthora*; identifying risks of spread in trade and methods for mitigation. *Forests*, 12:1617.

Grüning, B., Dale, R., Sjödin, A., Chapman, B. A., Rowe, J., Tomkins-Tinch, C. H., Valieris, R., Caprez, A., Batut, B., Haudgaard, M., Cokelaer, T., Beauchamp, K. A., Pedersen, B. S., Hoogstrate, Y., Bretaudeau, A., Ryan, D., Corguillé, G. L., Yusuf, D., Luna-Valero, S., Kirchner, R., Brinda, K., Wollmann, T., Raden, M., Heeringen, S. J. v., Soranzo, N., Pantano, L., Charlop-Powers, Z., Unneberg, P., Smet, M. D., Martin, M., Kuster, G. V., Antao, T., Miladi, M., Thornton, K., Brueffer, C., Beek, M. v. d., Maticzka, D., Blank, C., Will, S., Gravouil, K., Wolff, J., Holtgrewe, M., Fallmann, J., Piro, V. C., Shlyakhter, I., Yousif, A., Mabon, P., Zhang, X.-O., Shen, W., Cabral, J., Thomas, C., Enns, E., Brown, J., Boekel, J., Hollander, M. d., Kelleher, J., Turaga, N., Ruiter, J. R. d., Bouvier, D., Gladman, S., Choudhary, S., Harding, N., Eggenhofer, F., Kratz, A., Fang, Z., Kleinkauf, R., Timm, H., Cock, P. J. A., Seiler, E., Brislawn, C., Nguyen, H., Stovner, E. B., Ewels, P., Chambers, M., Johnson, J. E., Hägglund, E., Ye, S., Guimera, R. V., Pruesse, E., Dunn, W. A., Parsons, L., Patro, R., Koppstein, D., Grassi, E., Wohlers, I., Reynolds, A., Cornwell, M., Stoler, N., Blankenberg, D., He, G., Bargull, M., Junge, A., Farouni, R., Freeberg, M., Singh, S., Bogema, D. R., Cumbo, F., Wang, L.-B., Larson, D. E., Workentine, M. L., Devisetty, U. K., Laurent, S., Roger, P., Garnier, X., Agren, R., Khan, A., Eppley, J. M., Li, W., Stöcker, B. K., Rausch, T., Taylor, J., Wright, P. R., Taranto, A. P., Chicco, D., Sennblad, B., Baaijens, J. A., Gopez, M., Abdennur, N., Milne, I., Preussner, J., Pinello, L., Srivastava, A., Chande, A. T., Kensche, P. R., Pirola, Y., Knudsen, M., Bruijn, I. d., Blin, K., Gonnella, G., Enache, O. M., Rai, V., Waters, N. R., Hiltemann, S., Bendall, M. L., Stahl, C., Miles, A., Boursin, Y., Perez-Riverol, Y., Schmeier, S., Clarke, E., Arvai, K., Jung, M., Domenico, T. D., Seiler, J., Rasche, E., Kornobis, E., Beisser, D., Rahmann, S., Mikheyev, A. S., Tran, C., Capellades, J., Schröder, C., Salatino, A. E., Dirmeier, S., Webster, T. H., Moskalenko, O., Stephen, G., and Köster, J. (2018). Bioconda: A sustainable and comprehensive software distribution for the life sciences. *Nature Methods*, 15:475–476.

Hänfling, B., Lawson Handley, L., Read, D. S., Hahn, C., Li, J., Nichols, P., Blackman, R. C., Oliver, A., and Winfield, I. J. (2016). Environmental DNA metabarcoding of lake fish communities reflects long-term data from established survey methods. *Molecular Ecology*, 25(13):3101–3119.

Klymus, K. E., Marshall, N. T., and Stepien, C. A. (2017). Environmental DNA (eDNA) metabarcoding assays to detect invasive invertebrate species in the great lakes. *PLOS ONE*, 12(5):1–24.

La Spada, F., Cock, P. J. A., Randall, E., Pane, A., Cooke, D. E. L., and Cacciola, S. O. (2022). DNA metabarcoding and isolation by baiting complement each other in revealing *Phytophthora* diversity in anthropized and natural ecosystems. *Journal of Fungi*, 8(4).

Lamb, P. D., Hunter, E., Pinnegar, J. K., Creer, S., Davies, R. G., and Taylor, M. I. (2019). How quantitative is metabarcoding: A meta-analytical approach. *Molecular Ecology*, 28:420–430.

Magoč, T. and Salzberg, S. L. (2011). FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics*, 27(21):2957–2963.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet Journal*, 17:10–12.

Massart, S., Adams, I., Al Rwahnih, M., Baeyen, S., Bilodeau, G. J., Blouin, A. G., Boonham, N., Candresse, T., Chandellier, A., De Jonghe, K., Fox, A., Gaafar, Y. Z., Gentit, P., Haegeman, A., Ho, W., Hurtado-Gonzales, O., Jonkers, W., Kreuze, J., Kutjnak, D., Landa, B. B., Liu, M., Maclot, F., Malapi-Wight, M., Maree, H. J., Martoni, F., Mehle, N., Minafra, A., Mollov, D., Moreira, A. G., Nakhla, M., Petter, F., Piper, A. M., Ponchart, J. P., Rae, R., Remenant, B., Rivera, Y., Rodoni, B., Botermans, M., Roenhorst, J., Rollin, J., Saldarelli, P., Santala, J., Souza-Richards, R., Spadaro, D., Studholme, D. J., Sultmanis, S., van der Vlugt, R., Tamisier, L., Trontin, C., Vazquez-Iglesias, I., Vicente, C. S., van de Vossenberg, B. T., Westenberg, M., Wetzel, T., Ziebell, H., and Lebas, B. S. M. (2022). Guidelines for the reliable use of high throughput sequencing technologies to detect plant pathogens and pests. *Peer Community Journal*, 2:e62.

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(7).

Muri, C. D., Handley, L. L., Bean, C. W., Li, J., Peirson, G., Sellers, G. S., Walsh, K., Watson, H. V., Winfield, I. J., and Hänfling, B. (2020). Read counts from environmental DNA (eDNA) metabarcoding reflect fish abundance and biomass in drained ponds. *Metabarcoding and Metagenomics*, 4:e56959.

Nearing, J. T., Douglas, G. M., Comeau, A. M., and Langille, M. G. I. (2018). Denoising the denoisers: an independent evaluation of microbiome sequence error-correction approaches. *PeerJ*, 8(6):e5364.

Palmer, J. M., Jusino, M. A., Banik, M. T., and Lindner, D. L. (2018). Non-biological synthetic spike-in controls and the AMPtk software pipeline improve mycobiome data. *PeerJ*, 6:e4925.

Port, J. A., O'Donnell, J. L., Romero-Maraccini, O. C., Leary, P. R., Litvin, S. Y., Nickols, K. J., Yamahara, K. M., and Kelly, R. P. (2016). Assessing vertebrate biodiversity in a kelp forest ecosystem using environmental DNA. *Molecular Ecology*, 25(2):527–541.

Redekar, N. R., Eberhart, J. L., and Parke, J. L. (2019). Diversity of *Phytophthora*, *Pythium*, and *Phytopythium* species in recycled irrigation water in a container nursery. *Phytobiomes Journal*, 3(1):31–45.

Riddell, C. E., Frederickson-Matika, D., Armstrong, A. C., Elliot, M., Forster, J., Hedley, P. E., Morris, J., Thorpe, P., Cooke, D. E. L., Pritchard, L., Sharp, P. M., and Green, S. (2019). Metabarcoding reveals a high diversity of woody host-associated *Phytophthora* spp. in soils at public gardens and amenity woodlands in britain. *PeerJ*, 7:e6931.

Robideau, G. P., de Cock, A. W., Coffey, M. D., Voglmayr, H., Brouwer, H., Bala, K., Chitty, D. W., Désaulniers, N. L., Eggertson, Q. A., Gachon, C. M. M., Hu, C.-H., Küpper, F. C., Rintoul, T. L., Sarhan, E. A. D., Verstappen, E., Zhang, Y., Bonants, P., Ristaino, J. B., and Lévesque, C. A. (2011). DNA barcoding of oomycetes with cytochrome c oxidase subunit I and internal transcribed spacer. *Molecular Ecology Resources*, 11:1002 – 1011.

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

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

Schnell, I. B., Bohmann, K., and Gilbert, M. T. (2015). Tag jumps illuminated–reducing sequence-to-sample misidentifications in metabarcoding studies. *Molecular ecology resources*, 15(6):1289–1303.

Scibetta, S., Schena, L., Chimento, A., Cacciola, S. O., and Cooke, D. E. (2012). A molecular method to assess *Phytophthora* diversity in environmental samples. *Journal of Microbiological Methods*, 88(3):356–368.

Sepulveda, A. J., Hutchins, P. R., Forstchen, M., Mckeefry, M. N., and Swigris, A. M. (2020). The elephant in the lab (and field): Contamination in aquatic environmental DNA studies. *Frontiers in Ecology and Evolution*, 8.

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Research*, 13(11):2498–2504.

Straub, D., Blackwell, N., Langarica-Fuentes, A., Peltzer, A., Nahnsen, S., and Kleindienst, S. (2020). Interpretations of environmental microbial community studies are biased by the selected 16S rRNA (gene) amplicon sequencing pipeline. *Frontiers in Microbiology*, 11.

Tedersoo, L., Ramirez, K. S., Nilsson, R. H., Kaljuvee, A., Kõljalg, U., and Abarenkov, K. (2015). Standardizing metadata and taxonomic identification in metabarcoding studies. *GigaScience*, 4:34.

Thalinger, B., Deiner, K., Harper, L. R., Rees, H. C., Blackman, R. C., Sint, D., Traugott, M., Goldberg, C. S., and Bruce, K. (2021). A validation scale to determine the readiness of environmental DNA assays for routine species monitoring. *Environmental DNA*, 3(4):823–836.

Thielecke, L., Aranyossy, T., Dahl, A., Tiwari, R., Roeder, I., Geiger, H., Fehse, B., Glauche, I., and Cornils, K. (2017). Limitations and challenges of genetic barcode quantification. *Scientific Reports*, 7:43249.

Vélez, M. L., La Manna, L., Tarabini, M., Gomez, F., Elliott, M., Hedley, P. E., Cock, P. J. A., and Greslebin, A. (2020). *Phytophthora austrocedri* in Argentina and co-inhabiting phytophthoras: Roles of anthropogenic and abiotic factors in species distribution and diversity. *Forests*, 11(11).

Walker, F. M., Tobin, A., Simmons, N. B., Sobek, C. J., Sanchez, D. E., Chambers, C. L., and Fofanov, V. Y. (2019). A fecal sequel: Testing the limits of a genetic assay for bat species identification. *PLOS ONE*, 14:1–16.

Zaiko, A., Greenfield, P., Abbott, C., von Ammon, U., Bilewitch, J., Bunce, M., Cristescu, M. E., Chariton, A., Dowle, E., Geller, J., Ardura Gutierrez, A., Hajibabaei, M., Haggard, E., Inglis, G. J., Lavery, S. D., Samuiloviene, A., Simpson, T., Stat, M., Stephenson, S., Sutherland, J., Thakur, V., Westfall, K., Wood, S. A., Wright, M., Zhang, G., and Pochon, X. (2022). Towards reproducible metabarcoding data: Lessons from an international cross-laboratory experiment. *Molecular Ecology Resources*, 22(2):519–538.

Zhang, J., Kobert, K., Flouri, T., and Stamatakis, A. (2014). PEAR: a fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics*, 30:614–620.