Received: 3 April 2021 | Revised: 24 May 2021 | Accepted: 25 May 2021

DOI: 10.1002/ece3.7788

ORIGINAL RESEARCH

Ecology and Evolution Open Access WILEY

Method for quick DNA barcode reference library construction

Yanlei Liu[1,2] | Chao Xu[1] | Yuzhe Sun[1,2] | Xun Chen[1,3] | Wenpan Dong[1,4] | Xueying Yang[5] | Shiliang Zhou[1,2]

[1]State Key Laboratory of Systematic and Evolutionary Botany, Institute of Botany, Chinese Academy of Sciences, Beijing, China

[2]College of Life Sciences, University of Chinese Academy of Sciences, Beijing, China

[3]College of Landscape Architecture, Northeast Forestry University, Harbin, China

[4]Laboratory of Systematic Evolution and Biogeography of Woody Plants, School of Ecology and Nature Conservation, Beijing Forestry University, Beijing, China

[5]National Engineering Laboratory for Forensic Science, Key Laboratory of Forensic Genetics, Institute of Forensic Science, Ministry of Public Security, Beijing, China

Correspondence
Shiliang Zhou, State Key Laboratory of Systematic and Evolutionary Botany, Institute of Botany, Chinese Academy of Sciences, Beijing, 100093, China.
Email: slzhou@ibcas.ac.cn

Xueying Yang, National Engineering Laboratory for Forensic Science, Key Laboratory of Forensic Genetics, Institute of Forensic Science, Ministry of Public Security, Beijing, 100038, China.
Email: yxystyhhp@163.com

Funding information
the National Key R&D Program of China, Grant/Award Number: 2017YFC0803803; the open project of Institute of Forensic Science, Ministry of Public Security, Grant/Award Number: 2018FGKFKT04; Strategic Priority Research Program of the Chinese Academy of Sciences, Grant/Award Number: XDA19050303 and XDA23080204; National Natural Science Foundation of China, Grant/Award Number: NSFC31872679; the fundamental research fund of the Central Public Service Research Institute, Grant/Award Number: 2018JB001

Abstract

DNA barcoding has become one of the most important techniques in plant species identification. Successful application of this technology is dependent on the availability of reference database of high species coverage. Unfortunately, there are experimental and data processing challenges to construct such a library within a short time. Here, we present our solutions to these challenges. We sequenced six conventional DNA barcode fragments (ITS1, ITS2, *matK*1, *matK*2, *rbcL*1, and *rbcL*2) of 380 flowering plants on next-generation sequencing (NGS) platforms (Illumina Hiseq 2500 and Ion Torrent S5) and the Sanger sequencing platform. After comparing the sequencing depths, read lengths, base qualities, and base accuracies, we conclude that Illumina Hiseq2500 PE250 run is suitable for conventional DNA barcoding. We developed a new "Cotu" method to create consensus sequences from NGS reads for longer output sequences and more reliable bases than the other three methods. Step-by-step instructions to our method are provided. By using high-throughput machines (PCR and NGS), labeling PCR, and the Cotu method, it is possible to significantly reduce the cost and labor investments for DNA barcoding. A regional or even global DNA barcoding reference library with high species coverage is likely to be constructed in a few years.

KEYWORDS

Cotu, data processing method, DNA barcode, next-generation sequencing

Yanlei Liu and Chao Xu contributed equally to this work.

1 | INTRODUCTION

Since the term "DNA barcode" was proposed (Hebert et al., 2003), DNA barcoding has soon become a routine technology in molecular identification of organisms and served as a new tool for biologists to understand biota (Kress, 2017). This technology has extensive applications for the identification of microorganisms (Barberán et al., 2015), dietary composition of animals (Kartzinel et al., 2015), components in processed foods or drugs (Nithaniyal et al., 2017), cryptic species discovery (Tyagi et al., 2019), invasive species monitor (Xu et al., 2018), rare and endangered species conservation (Giovino et al., 2016; Hosein et al., 2017), etc.

Reliable molecular identification depends on the resolution of molecular markers (or DNA barcodes) and the species coverage of the reference library. Considerable efforts have been made to find the ideal DNA barcodes for plants (CBOL Plant Working Group, 2009; Dong et al., 2014, 2015; Kress & Erickson, 2007; Li et al., 2011) as well as to develop new technical improvements (Giovino et al., 2020a; Hollingsworth et al., 2011; Xu et al., 2015; Yu et al., 2011). Unlike animals which COI (mitochondrial cytochrome oxidase I) is a nearly sole DNA barcode, plant DNA barcoding is much more complicated and no ideal DNA barcodes for plants have yet been discovered, or perhaps they do not exist at all (Giovino et al., 2020b). A well-curated reference sequence library with high species coverage remains to be constructed for extensive applications of this technology. The good news is that some ambitious projects have been launched in the past few years [e.g., BARCODE 500K (https://ibol.org), BIOSCAN (Hobern & Hebert, 2019), and ISHAM-ITS (Irinyi et al., 2016)]. Even so, sequence data deposited in public databases are still rather small. Taking 258,650 flowering plants as an example (Thorne, 2002), 51,132 (19.8%) species have *matK* sequences and 46,130 (17.8%) species have *rbcL* sequences deposited in GenBank (accessed on 1 May 2020).

In order to construct such a reference library in a relatively short time, we have to use cost-efficient next-generation sequencing (NGS) platforms and acquire the ability of manipulating the NGS data. Although NGS platforms have been applied for this purpose for a decade (Boyer et al., 2016; Piry et al., 2012; Richardson et al., 2017; Shi et al., 2018; Shokralla et al., 2012; Toju, 2015), no consensus has been reached concerning the platforms themselves and the data processing methods owing to rapid replacements or upgrades of sequencing machines. Roche 454, which was one of the most suitable choices for DNA barcoding (Guo et al., 2019; Hajibabaei et al., 2011; Shokralla et al., 2014), is no longer available. The third-generation sequencing (TGS) or single molecule sequencing (SMS) platforms, such as PacBio and Nanopore, are now available for DNA super barcodes (such as Zhang et al., 2020). Illumina systems (Hiseq and Miseq) and Ion Torrent systems are currently the mainstream NGS platforms for conventional DNA barcode sequencing. The paired-end 250 (PE250) Illumina Hiseq recovers half the sequence lengths (ca 400 bp after removal of prefixes such as primers) of conventional DNA barcodes of 600–800 bp, whereas the Ion Torrent S5 series has a capacity of generating ca. 600-bp sequences. Both platforms have been used on DNA metabarcoding of environmental samples (Deagle et al., 2013; Evans et al., 2016; Fantini et al., 2015; Schmidt et al., 2013). The operational taxonomic units (OTUs) from the environmental samples were much more than actual situations, and it is one major concern whether these two platforms are suitable for conventional DNA barcodes or not (Lahens et al., 2017; Marine et al., 2020; Quail et al., 2012; Speranskaya et al., 2018).

Although researchers have made some efforts on applying NGS platforms to conventional DNA barcoding (Akankunda et al., 2020; Creedy et al., 2020; de Kedrel et al., 2020; Srivathsan et al., 2018), the ease of data analyses is still the other major concern in DNA barcoding using NGS. Several software packages have been developed for DNA metabarcoding, such as Vsearch, Usearch, Qiime2, OBITools, PipeCraft, and Mothur (Anslan et al., 2017; Boyer et al., 2016; Callahan et al., 2016; Edgar, 2013, 2016; Rognes et al., 2016; Schloss et al., 2009). Although these software packages can also be adopted to conventional DNA barcodes, there are some gaps to be bridged in the NGS data analysis pipelines. For example, a DNA metabarcoding sample contains many species, whereas in conventional DNA barcoding, a sample usually contains only one species. For cost efficiency consideration, many gene fragments of multiple samples are DNA-labeled, mixed, and sequenced in a single NGS run. Demultiplexing is necessary, and usually, only one or a few sequences need to be generated for each gene fragment in each sample.

One of the outstanding features of NGS is its ability to generate large data and different software packages with different functions have to be used. The NGS data processing major include (a) quality control to find and remove reads of low quality; (b) assembling read1 and read2 when paired-end sequencing method is used; (c) demultiplexing data to assign data to genes of samples according to primer and label sequences; and (d) creating correct consensus sequences when using NGS for conventional DNA barcode creation. Unfortunately, most scientists who devote themselves to DNA barcoding do not have the skills to handle such kind of data.

There are two prevailing strategies for processing amplicon sequencing data from NGS platforms. The first one applies an arbitrary cutoff (say, a minimum similarity of 0.97) for lumping reads into OTUs. Software packages, such as Usearch, Vsearch, Mothur, and PipeCraft, use this strategy. The second strategy resolves amplicon sequence variants (ASVs) to proofread the final sequences based on the sequencing depth of exact sequence variants (ESVs, usually one nucleotide difference; Knight et al., 2018). Software packages DADA2 and Unoise3 belong to this strategy (Callahan et al., 2017; Edgar, 2018). This strategy avoids imposing the arbitrary similarity thresholds but introduces random sequencing errors because not all single base differences are real. For the final sequence generation, Usearch, Mothur, Swarm, DADA2, etc., pick up a representative sequence while Vsearch, Geneious, Mothur, etc., create a consensus sequence. The representative sequence strategy has a risk of introducing sequencing errors and length variations to the final sequences. The consensus sequence strategy retains indel errors in homopolymeric regions (Srivathsan et al., 2018) and sequence length variations at both sequence ends. Current software is still imperfect for creating DNA barcodes, and they were not designed specially for conventional DNA barcode data analysis. In order to improve the accuracy and length of output consensus sequences,

we developed the new Cotu method under the majority rule (the letter "C" stands for consensus, Cotu means consensus-based OTU creation method, and the majority rule means only the majority base will be treated as the right base in each certain position in an alignment file).

Although a few studies have made performance comparisons between Illumina and Ion Torrent sequencing platforms (Lahens et al., 2017; Salipante et al., 2014; Speranskaya et al., 2018), it is still difficult for researchers to determine with certainty which one is more suitable for DNA barcoding. Similarly, due to imperfections of current data processing software, it is still unknown to researchers which software is most reliable in creating final sequences. In this study, we tested the suitability of two popular NGS platforms, Illumina Hiseq2500 and Ion Torrent S5, based on (a) the base quality; (b) read length variation; (c) sequencing depth bias among samples; and (d) genetic distance. For data processing methods, we compared the reliability of final sequences created by three most widely used data analysis methods (Otu, Zotu, and DADA2) to the standard sequences obtained from Sanger sequencing platform and provided our solution Cotu method in creating final sequences. The parameters we used are (a) sequence recovery; (b) sequence length; (c) genetic distance; and (d) sequence reliability. Among the four methods for clustering reads, Otu and Cotu use cutoff parameter, and Zotu and DADA2 use ESV strategy. Otu, Zotu, and DADA2 methods generate sequences using the representative sequence principle while Cotu adopts a majority consensus strategy. We aim to provide a better methodological solution to the collection of reliable sequences using NGS for the construction of a DNA barcode reference library in a short time with investments as small as possible.

2 | MATERIALS AND METHODS

Our experimental workflow is depicted in Figure 1. Gene fragments were sequenced by the Sanger sequencing method on ABI 3730xl, and NGS method on Illumina Hiseq2500 PE250 and Ion Torrent S5xl platforms. The clean reads from the Illumina and Ion Torrent platforms were analyzed using Otu, Zotu, DADA2, and Cotu methods. Sequences from ABI 3730xl were used as "gold standards" to test the results of the four methods using the reads from the two NGS platforms (details shown below).

2.1 | Plant material sampling and DNA extraction

All materials were collected from Beijing Botanical Garden of the Institute of Botany, Chinese Academy of Sciences. Fresh leaf materials of 380 samples (Table S1) belonging to 253 species, 139 genera, and 60 families were collected and immediately oven-dried at 65°C for 2–3 hr. Most species were collected during their flowering period. One to five individuals were sampled for each species. All voucher specimens were deposited in the herbarium of Institute of Botany, the Chinese Academy of Sciences (PE). Total DNA was extracted using the mCTAB method (Li et al., 2013), and the concentration was adjusted to 10 ng/μl for subsequent PCR experiments according to the measurements of the Nanodrop 2000c spectrophotometer (Thermo Fisher Scientific Inc.).

2.2 | PCR for Sanger sequencing

The nuclear internal transcribed spacer (ITS), chloroplast maturase K (*matK*), and ribulose-1,5-bisphosphate carboxylase/oxygenase (*rbcL*) were amplified using universal primers ITS-P5 + ITS-U4, matK-472F + matK-1248R, and rbcLbF + rbcLbR (Table 1) for Sanger sequencing (Dong et al., 2015). *psbA-trnH* was not used due to a long homopolymeric region frequently existing in many samples. Fragments were sequenced on ABI 3730xl DNA Analyzer (Applied Biosystems, USA) at the Majorbio Company in Beijing, China.

2.3 | PCR for next-generation sequencing (NGS)

In order to meet the read length limitations of Illumina and Ion torrent S5 sequencing platforms, we amplified two fragments about

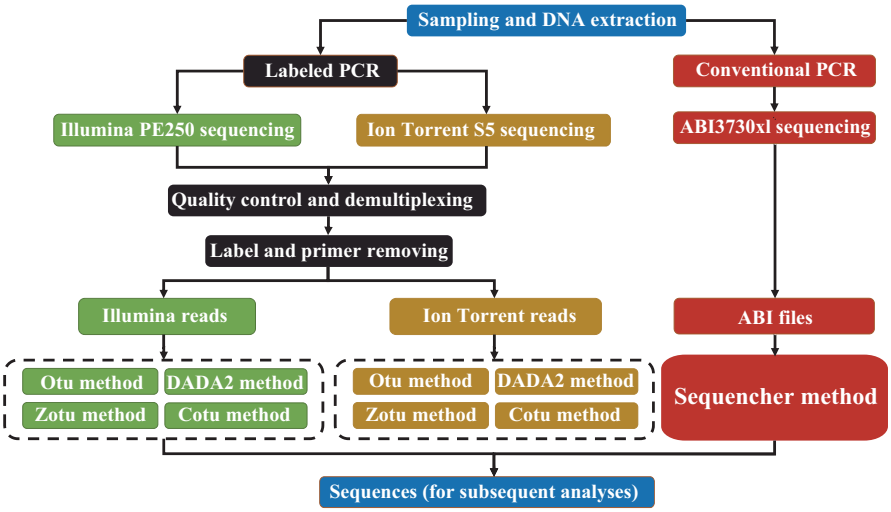

FIGURE 1 Experimental workflow. Gene fragments are sequenced on ABI 3730xl, Illumina Hiseq2500, and Ion Torrent S5 platforms. Data generated by the latter two platforms are processed with Cotu, Dotu, Otu, and Zotu methods. The sequences from ABI 3730xl serve as references and those from the two NGS platforms are queries in analyses

TABLE 1 The first-round PCR primers for amplifying DNA fragments

Barcode	Primer name	Primer sequence (5'–3')	Tm(°C)	GC%	Expected length
ITS1	ITS-P5	CCTTATCAYTTAGAGGAAGGAG	68.16	47.22	370–380 bp
	ITS-U2	GCGTTCAAAGAYTCGATGRTTC	68.73	47.22	
ITS2	ITS-P3	YGACTCTCGGCAACGGATA	69.70	54.55	440–450 bp
	ITS-U4	RGTTTCTTTTCCTCCGCTTA	67.15	47.06	
*matK*1	*matK*–472F	CCCRTYCATCTGGAAATCTTGGTTC	70.54	48.72	380–390 bp
	matK–821R	TTTCCTTGATATCTAACATAATG	64.20	37.84	
*matK*2	*matK*–821F	CATTATGTTAGATATCAAGGAAA	64.20	37.84	420–430 bp
	matK–1248R	GCTRTRATAATGAGAAAGATTTCTGC	67.32	44.00	
*rbcL*1	*rbcL*bF	AGACCTWTTTGAAGAAGGTTCWGT	67.50	44.74	420–430 bp
	*rbcL*717R	CATGTACCTGCAGTAGCATTCAAGT	69.49	48.72	
*rbcL*2	*rbcL*717F	ACTTGAATGCTACTGCAGGTACATG	69.49	48.72	430–440 bp
	*rbcL*bR	TCGGTYAGAGCRGGCATRTGCCA	72.51	56.76	

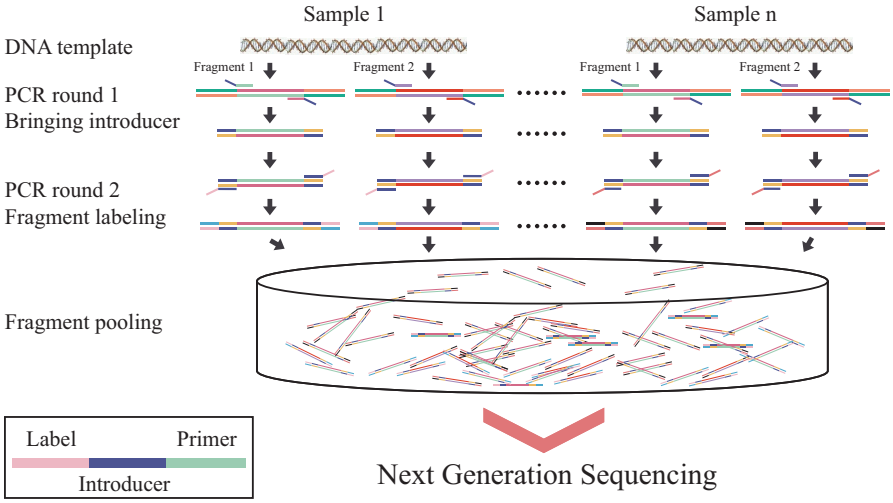

FIGURE 2 Labeling gene fragments using sample-specific oligoes by two rounds of PCR for multiplexed sequencing on the Illumina Hiseq2500 and Ion Torrent S5 platforms. An "introducer" is attached to the ends of fragments during the first round of PCR. The introducer serves as the priming site during the second round of PCR, and sample-unique oligoes were added to the ends of the fragments

400 bp for each conventional barcode (middle primers were used and listed in Table 1 and the primer annealing positions are displayed in Figure S1). When designing the middle primers, overlap region was considered for the whole ITS assembling. The length of DNA barcode *matK* and *rbcL* is about 800 bp each, and they are very variable among angiosperm plants. It is hard to find two possible positions to design primer pairs forming overlaps in the middle of the DNA barcodes. In order to improve the primer universality, we only find one position for *matK* and *rbcL* each to design the overlapped primers. Therefore, two parts of each DNA barcode can also be assembled through the overlap positions of the middle primers. For multiplexing on NGS platforms, gene fragments from the same sample were labeled with a unique DNA oligo by two rounds of PCR (Figure 2). In the first round of PCR, primers attached by an oligo (called introducer, 5'-GTAGACTGCGTACC-3') at the 5' end were used to amplify gene fragments. The PCR procedures were the same as Dong et al. (2015), except that the primer concentration was only 10% of that in conventional PCR and that the number of PCR cycles was increased to 40 for using up all primer molecules. In the second round of PCR, products from the first-round PCR served as templates, and the introducer attached by a sample-specific oligo of ten bases (called DNA label) at the 5' end was used as a primer for each sample. In the present study, 380 such primers were synthesized and used to label 380 samples (Table S1). Different Gene fragments from the same sample were amplified individually with the same labeling primer. The PCR program was the same as Dong et al. (2015).

The PCR products of the same gene were mixed, gel-purified, and quantified on a Nanodrop 2000c spectrophotometer. The PCR products of different genes were combined at nearly equal molar ratios according to their concentrations.

2.4 | Library construction for next-generation sequencing

The final PCR mixture for NGS was divided into two equal parts. One part was used for Illumina library construction using NEBNext® Ultra™ DNA Library Prep Kit for Illumina® (New England BioLabs) and sequenced at BerryGenomics, Beijing, China,

on Illumina Hiseq2500 (pair end, PE250). The other part was used for Ion Torrent S5 platform library construction using NEBNext® Fast DNA Library Prep Set for Ion Torrent (New England BioLabs) and sequenced at Maize Research Center, Beijing Academy of Agriculture and Forestry Sciences for Ion Torrent S5 Chip400 sequencing.

2.5 | ABI 3730xl sequencing data processing

Sequences were assembled by combining forward and reverse strands of the same fragments according to the overlap region, and they were edited using Sequencher v5.4.5 (Gene Codes Corporation) based on files from ABI 3730xl Analyzer. Base-calling mistakes if any were corrected according to the chromatograms. If the major base chromatogram peak is obviously bigger than other little chromatograms, we consider this base is correct. If there are two major chromatograms, we consider this base as a degenerate base. If there are no major chromatogram and the chromatogram peak is low, we consider this base is not credible.

2.6 | Illumina Hiseq2500 and Ion Torrent S5 sequencing data processing

2.6.1 | Quality control

Illumina Hiseq2500 PE250 data were quality-controlled using the NGS QC toolkit v2.3.3 with the default parameters (Patel & Jain, 2012). After quality control, the lengths of the reads longer than 200 bp were categorized by genes and by platforms using the default parameters in FASTX-Toolkit v0.0.13 (http://hannonlab.cshl.edu/fastx_toolkit/) and statistics was done with Excel 2016.

2.6.2 | Demultiplexing and data cleaning

The quality-controlled reads were merged using Flash v1.2.11 (Magoc & Salzberg, 2011) with default settings (Ion Torrent S5 data processing did not include this step). The merged reads were demultiplexed using FASTX-Toolkit v0.0.13 (http://hannonlab.cshl.edu/fastx_toolkit/) according to the sample labels and primers. In order to find out which platform had less sequencing bias among samples, the number of reads was transformed into percentages, and the significance of sequencing depth bias variations among samples was tested using double factor variance analysis (two-way ANOVA) in SPSS v19 between the two platforms gene by gene and as a whole.

Unlabeled sequences and sequences shorter than 200 bp were discarded using NGS QC toolkit v2.3.3. Artificially added regions, such as sample labels, introducers, and primers, were trimmed off using Cutadapt v2.7 (https://cutadapt.readthedocs.io/en/stable/).

2.6.3 | Sequencing error estimation

The demultiplexed clean reads of each sample were mapped to the corresponding reference sequences obtained from Sanger sequencing using Geneious Prime 2019.2.3 by the highest sensitivity. The frequencies of the bases along the whole length were calculated using python script *base-counter.py* (https://github.com/Mycroft-maker/Base-counter). Site-by-site comparisons were carried out between the final sequences from both platforms and the references from ABI. If there was a mismatch of the highest frequency to the reference, the secondly highest base was considered, and so forth until an exact match was found. The mismatches are an estimate of sequencing errors.

2.6.4 | Sequence creation

Four methods (Otu, Zotu, DADA2, and Cotu) were used to create final sequences. For Otu method, we followed the UPARSE protocol (http://www.drive5.com/uparse) (Edgar, 2013; Edgar & Flyvbjerg, 2015). For Zotu method, we followed the Unoise3 protocol (http://www.drive5.com/usearch/manual/unoise_algo) (Edgar, 2013; Edgar & Flyvbjerg, 2015). For DADA2 method, we used the DADA2 (Callahan et al., 2016) plugin in Qiime2 2019.04 (Bolyen et al., 2019) and we call the sequence generated by DADA2 method "Dotu" in this study.

The major features of the Cotu method are elimination of insertions caused by occasional reads under majority rule and sequence length extensions at both ends of alignments. The qualities of the beginning and ending bases are usually low and frequently trimmed off, which causes uneven alignments. If the majority rule was used for this situation, shortened consensus sequences would likely be created. We, therefore, apply a threshold of bases on a site (e.g., 20%) at both ends. In order to reduce gaps in an alignment, base proportion with less than 10% of total bases in a certain position is automatically removed before the application of majority rule. The program calculates the number of bases in three consecutive base positions and sets a position range based on the first appearance of three consecutive base positions with both bases number larger than 50% from each ends automatically. If a site in the beginning and ending regions has the number of bases less than the threshold, the site is considered an incorrect insertion; if a site has bases of more than the threshold number, the site is considered normal and the majority rule is applied. For this method, demultiplexed single copy reads are accurately aligned with Mafft v7.467 (Nakamura et al., 2018) and a consensus sequence is created using Cotu-Generator.py (https://github.com/YanleiLiu1989/Cotu-master). For data consisting of multiple copies (such as some nuclear genes and allotetraploids) or multiple species (such as DNA metabarcoding), the clean reads are first sorted into gene copies or species using Vsearch V2.4.3. The sorted reads are then accurately aligned with Mafft v7.467 and a consensus sequence is created using Cotu method. User's manual (Supporting Document S1) and step-by-step instructions (Supporting Document S2) are provided for correct use of the method.

2.7 | Comparative analyses between NGS platforms and among data processing methods

The base quality, read length variation, sequencing depth bias among samples, and base accuracy were used to judge the suitability of Illumina Hiseq2500 and Ion Torrent S5 for conventional DNA barcoding. Base quality was quantified by base scores (values from 1 to 45 given by sequencing machine) using FASTX v0.0.13 and averaged over whole length. The sequencing depths were transformed into relative sequencing depths using percentages of the number of reads each sample to the total reads. Base accuracy was evaluated site by site by comparing to the reference sequences and summarized with Excel.

Sequence recovery, sequence length, and genetic distance were used to test the reliability of Otu, Zotu, DADA2, and Cotu data processing methods. Sequence recovery was the percentage of the number of sequences created by each method compared with the total number of samples with data. The sequence length variations were considered by using the whole-sequence length.

Genetic distance was parameterized using Kimura two-parameter genetic distance between sequences created by each method (queries) and the corresponding reference sequence of the same sample using Mega 7.0.18(Kumar et al., 2018). All ambiguous sites were ignored for each sequence pair. Only samples with all four methods results and Sanger reference were adopted for genetic distance comparison.

All significance tests of difference between sequencing platforms and among data processing methods were simplified to be one-factor analyses of variance (ANOVA) in SPSS v19. The original data were transformed into percentages or genetic distances. The percentages or genetic distances were further transformed into square root values to meet the statistical requirement of normal distribution for ANOVA.

3 | RESULTS

3.1 | Reference sequences

Among the 380 samples, ITS, *matK*, and *rbcL* fragments were successfully amplified and sequenced in 304, 367, and 369 samples, respectively, with high base quality. The sequence lengths were from 450 bp to 784 bp for ITS (including 5.8S ribosomal RNA sequences), from 652 bp to 754 bp for *matK*, and 785 bp for *rbcL*.

3.2 | Differences between Illumina Hiseq and Ion Torrent S5 platforms

3.2.1 | Sequencing depth

The average sequencing depth of the samples on the Illumina Hiseq platform was 572× (× represents number of reads) for ITS1, 288× for ITS2, 2,850× for *matK*1, 2,234× for *matK*2, 547× for *rbcL*1, and 321× for *rbcL*2 (Figure S2a). The average sequencing depth per sample was 1,135×. The average sequencing depth of the samples on the Ion Torrent S5 platform was 335× for ITS1, 131× for ITS2, 417× for *matK*1, 1,034× for *matK*2, 523× for *rbcL*1, and 737× for *rbcL*2 (Figure S2b). The average sequencing depth per sample was 530×.

TABLE 2 Statistical *F* test of sequencing depth bias, read length, and base accuracy between sequencing platforms

	Sequencing depth bias	Read length	Base accuracy
Mean	0.0026	383.7094	0.9612
Variance	0.0006	6.86247E−05	0.0013
F-value	58.5202	47.3708	81.2693
p	<0.001	<0.001	<0.001

Note: Sequencing depth bias is proportion of reads each sample to the total number of reads from a platform. (Paired end) Read length was number of nucleotides. Base accuracy is proportion of correct bases to the total bases.

3.2.2 | Sequencing depth bias among samples

Illumina Hiseq2500 and Ion Torrent S5 exhibited sequencing depth bias among samples. The variances of sequencing depths among samples on Illumina Hiseq2500 were smaller than on Ion Torrent S5 for all six gene fragments (Figure S3). The difference of sequencing depth bias between the two platforms was tested to be significant ($p < .001$, Table 2).

3.2.3 | Read length

For Illumina PE250 run (maximum fragment length 250 bp), the average lengths of reads were 248 bp and 99.1% of read1 and 98.8% of read2 had minimum lengths of 200 bp (Figure S4a). For Ion Torrent S5 400 chip, the average length of reads was 350 bp and 65.4% of the reads had lengths longer than 320 bp (Figure S4b). The read length difference between the two platforms was tested to be significant ($p < .001$, Table 2).

Considerable read length variations were observed in ITS1, ITS2, *rbcL*1, and *rbcL*2 fragments on both platforms with standard deviation from 101.64 to 147.35 (Figure S5, paired-end reads for Illumina Hiseq2500). The percentages of mean values to the expected lengths ranged from 62% to 93% for Illumina Hiseq2500 and from 73% to 84% for Ion Torrent S5. The averages of the percentages were nearly the same, 78% for Illumina Hiseq2500 and 77% for Ion Torrent S5.

3.2.4 | Base quality

The average base quality scores of read1 and read2 from Illumina platform were 38.76 ($SD = 0.9132$) and 38.05 ($SD = 1.2667$),

respectively (Figure S6a,b). The quality of read1 was better than that of read2. The average base quality score of Ion Torrent S5 sequences was 25.64 (SD = 5.8194, Figure S6c). Base quality decreased with the progress of sequencing on both platforms.

3.2.5 | Base accuracy

Both platforms had over 96% matches for *matK*1, *matK*2, *rbcL*1, and *rbcL*2 fragments and relatively poor matches for ITS1 and ITS2 regions (Figure S7). The base accuracy seemed gene fragment-dependent. Illumina platform performed better than Ion Torrent for ITS1, ITS2, and *rbcL*1, but worse for *matK*1, *matK*2, and *rbcL*2. Mismatches occurred in the more variable regions of ITS1 and ITS2 and in the beginning parts of *matK*1, *matK*2, and *rbcL*1 (Figure S8). The difference of base accuracy between the two platforms was tested to be significant ($p < .001$, Table 2).

3.3 | Differences among Otu, Zotu, Dotu, and Cotu methods

The clean reads from Illumina Hiseq2500 and Ion Torrent S5 platforms were analyzed using four methods, Cotu, Dotu, Otu, and Zotu, and the outcomes are listed in Result S1.zip (https://github.com/YanleiLiu1989/Cotu-master). The reliabilities of these four methods were parameterized by sequence recovery, sequence length, and genetic distance.

3.3.1 | Sequence recovery

Next-generation sequencing reads have random sequencing errors. Sequencing depth determines the accuracy of output sequences. We set a minimum depth of 10× for both platforms and a similarity of 97% for ITS and 99% for *matK* and *rbcL*. With these restrictions, the number of sequences recovered by the four methods varied slightly for the data from Illumina platform (Figure S9a) but remarkably for the data from Ion Torrent platform (Figure S9b). For the data from Illumina platform, all methods except Dotu recovered more than 350 (92.1%) sequences of five gene fragments of 380 samples. However, for the data from Ion Torrent platform, only Cotu recovered 350 sequences of four gene fragments. In general, Cotu recovered the highest number of sequences, whereas Dotu recovered the lowest number of sequences (Figure S9).

3.3.2 | Sequence length

Different sequence creation methods are based on different principles and use slightly different reads from the same sample, and therefore, the length of the sequences created by different methods varies. The sequences created by Cotu method were the longest for all DNA fragments (Table 3). The length differences of sequences created by different methods were tested significant in ANOVA ($p < .01$, Table 3).

3.3.3 | Genetic distance

The more accurate the sequences, the smaller the genetic distances between the output sequences and the reference sequences. Again, the sequences created by Cotu method were the most similar to the reference sequences with the smallest distances for all DNA fragments (Table 3). Likewise, the genetic distance difference of sequences was tested significant among the four methods ($p < .01$, Table 3).

4 | DISCUSSION

It is believed that nearly two million species we know today are only a small fraction of total species diversity in the world (Mora et al., 2011). Nowadays, discovery of new species, especially microorganisms, is technique-dependent. DNA (meta)barcoding is one of the most effective methods for species identification, and it has been used for evaluating species diversity (Chen et al., 2016), monitoring changes in microorganism composition in the environment (Barberán et al., 2015), identifying species in processed food or drug materials (Chin et al., 2016), etc. However, the DNA (meta) barcoding technology is heavily dependent on the species coverage of the reference library. Since the publication of the paper by Hebert et al. (2003), seventeen years have passed and very few such libraries have been constructed. In order to construct a reference library of high species coverage in a relatively short period of time, we have to overcome several major challenges.

4.1 | The first challenge is high costs in raw data collections

A major investment for DNA barcode reference library construction is in DNA sequencing. Conventional Sanger sequencing is costly and of low efficiency. With minor technical modification in this study (Figure 2), different gene fragments of multiple samples can be sequenced simultaneously on NGS platforms, which significantly lowers sequencing costs. For example, a mixture containing 16 gene fragments of 384 samples can be sequenced in a sequencing library and data of 10G give an average sequencing depth of 3,255× theoretically. 10G NGS data cost less than $1,000, and the average cost of per final barcode sequence is about $0.15. Compared with about $2.8 cost of each Sanger sequence, the sequencing cost in NGS is only about 5% of that in the Sanger sequencing.

TABLE 3 Statistical *F* test of length and accuracy of sequences created with four data processing methods using reads from Illumina HiSeq 2500 and Ion Torrent S5

	Illumina								Ion Torrent							
	Mean								Mean							
	Cotu	Dotu	Otu	Zotu	*df*	MS	*F*	*p*	Cotu	Dotu	Otu	Zotu	*df*	MS	*F*	*p*
Sequence length					3	3.153	262.36	0.000					2	7.005	444.68	0.000
ITS1	360.3	338.2	354.5	352.8					362.4		346.4	343.5				
ITS2	395.0	305.7	380.2	379.0					401.5		376.1	374.4				
matK1	338.8	338.7	332.2	332.2					338.4		332.7	329.3				
matK2	380.6	378.5	368.2	380.4					380.5		370.4	373.0				
rbcL1	375.2	347.4	374.5	375.2					377.0		361.5	357.3				
rbcL2	410.9	326.4	391.3	389.4					414.7		379.7	366.2				
Genetic distance					3	0.209	15.09	0.000					2	0.11	4.752	0.009
ITS1	0.024	0.037	0.035	0.034					0.084		0.138	0.139				
ITS2	0.020	0.030	0.023	0.023					0.065		0.070	0.071				
matK1	0.007	0.010	0.009	0.009					0.012		0.015	0.015				
matK2	0.006	0.007	0.007	0.010					0.012		0.018	0.018				
rbcL1	0.008	0.008	0.008	0.011					0.004		0.005	0.005				
rbcL2	0.003	0.010	0.011	0.010					0.003		0.004	0.004				
Standard deviation																
ITS1	0.090	0.117	0.114	0.111					0.163		0.261	0.262				
ITS2	0.078	0.115	0.078	0.077					0.165		0.161	0.161				
matK1	0.042	0.060	0.054	0.054					0.062		0.076	0.076				
matK2	0.033	0.036	0.036	0.039					0.059		0.089	0.089				
rbcL1	0.026	0.026	0.027	0.026					0.018		0.023	0.023				
rbcL2	0.017	0.039	0.047	0.045					0.009		0.018	0.018				

Note: Genetic distance is measured by genetic distances between created sequences and the reference sequence of the same sample. DADA2 was not applicable to the data from Ion Torrent platform due to extraordinary variability of sequence lengths.

Abbreviations: *df*, degree of freedom; *F*, value of F-statistics; MS, mean square.

Standard deviation is calculated based on genetic distance.

4.2 | The second challenge is the perplexity in choosing NGS platform

Scientists are always trying to get more results and better done a research with less money. Wise selection of an NGS platform is crucial for obtaining high-quality results and saving money. Base quality, read length, data sizes, sequencing depth, and cost efficiency should be taken into consideration when selecting an NGS platform. Owing to the relatively high base quality and low cost, Illumina sequencing platform and Ion torrent S5 platform are currently the most suitable platforms for conventional DNA barcoding compared with other sequencing platforms (PacBio, Nanopore, Sanger, and so on). For Illumina Hiseq PE250 and Ion Torrent S5 suitable for fragments of 400 bp, the former performs better than the latter in this study in terms of base quality (Figure S3), read length (Table 2), sequencing depth (Figure S2), and sequence accuracy (Table 3).

NGS platforms have been successfully used in DNA metabarcoding of microorganisms, such as bacteria and viruses (Krehenwinkel et al., 2019), but are not very commonly used for DNA barcode reference library construction. Both Illumina and Ion Torrent S5 platforms meet the requirement for conventional DNA barcodes in half or full length. Although the first few bases are prone to be wrongly sequenced and need to be treated with caution, the average genetic distances (0.014, 0.007, and 0.003 separately for ITS, *matK*, and *rbcL* output from Cotu) between the queries and the references are small enough, indicating the reliability of NGS platforms for conventional DNA barcoding.

4.3 | The third challenge is the complexity of data processing

The Cotu method has several advantages over other methods. (a) It separates contaminants by sorting the reads into groups in combination with Vsearch and creates sequences for every group. (b) It eliminates PCR and sequencing errors using consensus sequences under the majority rule. And (c) it maximizes lengths of output sequences using a user-given threshold and avoids misuse of the majority rule. We compared this new method with other methods and found that the new method performed best in terms of both sequence length and sequence accuracy (Figure 3).

Unfortunately, as we have mentioned before, researchers have to face some software packages when doing different treatments of NGS data. In order to facilitate researchers who are not good at bioinformatics for Cotu generation, we packaged core Cotu steps together and named it "Cotu Master" (https://github.com/YanleiLiu1989/Cotu-master) which is also provided in the Supporting Software. A few fastq datasets (Supporting data.zip) are provided for testing the program together with the expected results without (Result S2) or with 500 reads limit (Result S3) in https://github.com/YanleiLiu1989/Cotu-master. Researchers can get their data just by entering a simple command following the step-by-step instructions (Supporting Document S2). Besides, in order to reduce the computing burden using Cotu method with an ordinary computer, an option of a maximum data usage was provided without lowering the quality of results. If the read number is larger than the maximum value (500×, for example), only the first 500 reads will be used.

4.4 | The fourth challenge is the difficulties in determining thresholds

Most software packages are developed flexibly for users to input thresholds for NGS data processing. The similarity of reads to be grouped is one of the most important parameters to be determined before analyses. In the analyses of DNA metabarcoding data of microorganisms, a sequence similarity of 0.97 was often adopted for 16S, 18S, or COI fragments (Berry et al., 2017; Bremond et al., 2017; Yamamoto et al., 2017). The chloroplast plant DNA barcodes *matK* and *rbcL* are not so variable as nuclear barcode ITS (Figure S7 and Figure S8) and a different threshold of sequence similarity had better

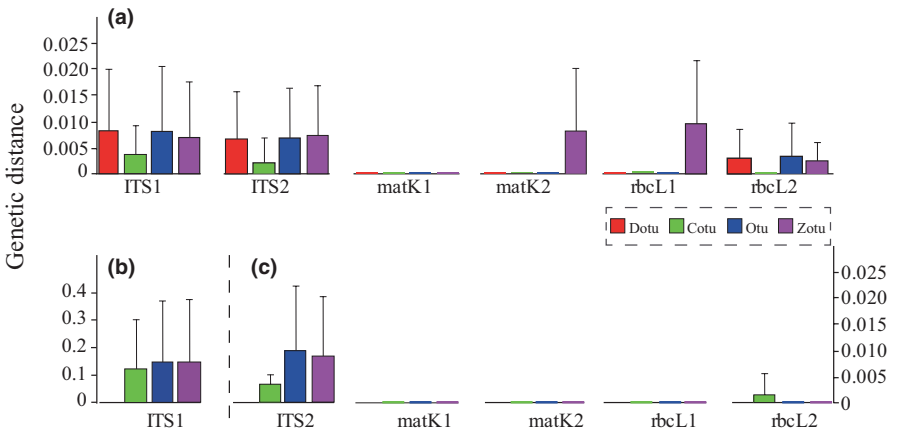

FIGURE 3 Accuracies of sequences created by Cotu (green), Dotu (red), Otu (blue), and Zotu (purple) methods with the data from Illumina Hiseq2500 (a) and Ion Torrent S5 (b & c) platforms. Average Kimura two-parameter genetic distances between the queries and corresponding references were calculated with MEGA7. The Dotu method was not applicable to Ion Torrent S5 platform because read lengths were too variable to create reliable sequences. Ambiguous sites were not considered

be used for situations of mixed samples or multiple gene copies. A sequence similarity of 0.99 is suitable for *matK* and *rbcL*, but lower similarity 0.97 may be appropriate for ITS. If the similarity value were set too high, the OTU diversity would be inflated. On the contrary, if the similarity value were set too low, differences between OTUs would be overwhelmed by the majority. For the Cotu method, no arbitrary similarity is necessary for reads of single copy fragments in a sample.

5 | CONCLUSION

In order to support accurate molecular identification of organisms by means of DNA barcoding, a reference library with high species coverage needs to be constructed as quick and cheap as possible. To reach this goal, high-throughput sequencing platforms are indispensable to speed up the processes and lower the costs. In this study, we show that the Illumina Hiseq PE250 is currently the right platform for conventional DNA barcodes. After comparing the newly developed data processing Cotu method to the existing Dotu, Otu, and Zotu methods, we conclude that the Cotu method is simpler, more accurate, and reliable. The packaged program Cotu master for creating consensus sequences had been uploaded to github (https://github.com/YanleiLiu1989/Cotu-master). Besides, the user's manual (Supporting Document S1), step-by-step instructions (Supporting Document S2) are also provided for getting familiar to Cotu method more quickly. By using high-throughput machines (PCR and NGS), labeling PCR, and the Cotu method, it is possible to significantly reduce the cost and labor investments for DNA barcoding. A regional or even global DNA barcoding reference library with high species coverage is likely to be constructed in a few years. As an example, a DNA reference library of seed plants in China is constructing using these methods and will soon be constructed with an investment of a few million dollars.

ACKNOWLEDGMENTS

We thank Xinqiang Yu for writing python scripts and Mingyu Yin for AMOVAs. This study was partly supported by the funds from the National Key R&D Program of China (2017YFC0803803), the open project of Institute of Forensic Science, Ministry of Public Security (2018FGKFKT04), Strategic Priority Research Program of the Chinese Academy of Sciences (XDA23080204, XDA19050303), National Natural Science Foundation of China (NSFC31872679), and the fundamental research fund of the Central Public Service Research Institute (2018JB001).

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTION

Yanlei Liu: Conceptualization (equal); Data curation (equal); Formal analysis (equal); Investigation (equal); Supervision (equal); Writing-original draft (equal); Writing-review & editing (equal). **Chao Xu:** Conceptualization (equal); Data curation (equal); Formal analysis (equal); Investigation (equal); Supervision (equal); Writing-review & editing (equal). **Yuzhe Sun:** Formal analysis (equal); Writing-review & editing (equal). **Xun Chen:** Formal analysis (equal); Investigation (equal). **Wenpan Dong:** Formal analysis (equal); Investigation (equal). **Xueying Yang:** Conceptualization (equal); Data curation (equal); Writing-original draft (equal); Writing-review & editing (equal). **Shi-Liang Zhou:** Conceptualization (equal); Data curation (equal); Writing-original draft (equal); Writing-review & editing (equal).

DATA AVAILABILITY STATEMENT

Sanger sequences of six gene fragments have been deposited in GenBank. The accession numbers are listed in Table S2. Original NGS data from Ion Torrent S5 and Illumina Hiseq2500 platforms have been submitted to NCBI with the accession numbers SRR11183118 and SRR11183119, respectively.

ORCID

Yanlei Liu https://orcid.org/0000-0002-8160-0141

REFERENCES

Akankunda, T., To, H., Lopez, C. R., Leijs, R., & Hogendoorn, K. (2020). A method to generate multilocus barcodes of pinned insect specimens using miseq. *Molecular Ecology Resources*, *20*, 692–705. https://doi.org/10.1111/1755-0998.13143

Anslan, S., Bahram, M., Hiiesalu, I., & Tedersoo, L. (2017). Pipecraft: Flexible open-source toolkit for bioinformatics analysis of custom high-throughput amplicon sequencing data. *Molecular Ecology Resources*, *17*, e234–e240. https://doi.org/10.1111/1755-0998.12692

Barberán, A., Ladau, J., Leff, J. W., Pollard, K. S., Menninger, H. L., Dunn, R. R., & Fierer, N. (2015). Continental-scale distributions of dust-associated bacteria and fungi. *Proceedings of the National Academy of Sciences*, *112*, 5756–5761. https://doi.org/10.1073/pnas.1420815112

Berry, T. E., Osterrieder, S. K., Murray, D. C., Coghlan, M. L., Richardson, A. J., Grealy, A. K., Stat, M., Bejder, L., & Bunce, M. (2017). DNA metabarcoding for diet analysis and biodiversity: A case study using the endangered Australian sea lion (*Neophoca cinerea*). *Ecology and Evolution*, *7*, 5435–5453. https://doi.org/10.1002/ece3.3123

Bolyen, E., Rideout, J. R., Dillon, M. R., Bokulich, N., Abnet, C. C., Al-Ghalith, G. A., Alexander, H., Alm, E. J., Arumugam, M., Asnicar, F., Bai, Y., Bisanz, J. E., Bittinger, K., Brejnrod, A., Brislawn, C. J., Brown, C. T., Callahan, B. J., Caraballo-Rodriguez, A. M., Chase, J., … Caporaso J. G. (2019). Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nature Biotechnology*, *37*, 852–857. https://doi.org/10.1038/s41587-019-0209-9

Boyer, F., Mercier, C., Bonin, A., Le Bras, Y., Taberlet, P., & Coissac, E. (2016). Obitools: A unix-inspired software package for DNA metabarcoding. *Molecular Ecology Resources*, *16*, 176–182. https://doi.org/10.1111/1755-0998.12428

Bremond, L., Favier, C., Ficetola, G. F., Tossou, M. G., Akouegninou, A., Gielly, L., Giguet-Covex, C., Oslisly, R., & Salzmann, U. (2017). Five thousand years of tropical lake sediment DNA records from Benin. *Quaternary Science Reviews*, *170*, 203–211. https://doi.org/10.1016/j.quascirev.2017.06.025

Callahan, B. J., McMurdie, P. J., & Holmes, S. P. (2017). Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *Isme Journal*, *11*, 2639–2643. https://doi.org/10.1038/ismej.2017.119

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., & Holmes, S. P. (2016). DADA 2: High-resolution sample inference from Illumina amplicon data. *Nature Methods*, *13*, 581. https://doi.org/10.1038/nmeth.3869

CBOL Plant Working Group (2009). A DNA barcode for land plants. *Proceedings of the National Academy of Sciences*, *106*(31), 12794–12797. https://doi.org/10.1073/pnas.0905845106

Chen, R., Jiang, L. Y., Chen, J., & Qiao, G. X. (2016). DNA barcoding reveals a mysterious high species diversity of conifer-feeding aphids in the mountains of southwest china. *Scientific Reports*, *6*, 20123. https://doi.org/10.1038/srep20123

Chin, T. C., Adibah, A., Hariz, Z. D., & Azizah, M. S. (2016). Detection of mislabelled seafood products in Malaysia by DNA barcoding: Improving transparency in food market. *Food Control*, *64*, 247–256. https://doi.org/10.1016/j.foodcont.2015.11.042

Creedy, T. J., Norman, H., Tang, C. Q., Chin, K. Q., Andujar, C., Arribas, P., O'Connor, R. S., Carvell, C., Notton, D. G., & Vogler, A. P. (2020). A validated workflow for rapid taxonomic assignment and monitoring of a national fauna of bees (Apiformes) using high throughput DNA barcoding. *Molecular Ecology Resources*, *20*, 40–53. https://doi.org/10.1111/1755-0998.13056

de Kerdrel, G. A., Andersen, J. C., Kennedy, S. R., Gillespie, R., & Krehenwinkel, H. (2020). Rapid and cost-effective generation of single specimen multilocus barcoding data from whole arthropod communities by multiple levels of multiplexing. *Scientific Reports*, *10*, 78. https://doi.org/10.1038/s41598-019-54927-z

Deagle, B. E., Thomas, A. C., Shaffer, A. K., Trites, A. W., & Jarman, S. N. (2013). Quantifying sequence proportions in a DNA-based diet study using ion torrent amplicon sequencing: Which counts count? *Molecular Ecology Resources*, *13*, 620–633. https://doi.org/10.1111/1755-0998.12103

Dong, W. P., Cheng, T., Li, C. H., Xu, C., Long, P., Chen, C., & Zhou, S. L. (2014). Discriminating plants using the DNA barcode *rbcL*b: An appraisal based on a large data set. *Molecular Ecology Resources*, *14*, 336–343. https://doi.org/10.1111/1755-0998.12185

Dong, W. P., Xu, C., Li, C. H., Sun, J. H., Zuo, Y. J., Shi, S., Cheng, T., Guo, J. J., & Zhou, S. L. (2015). *ycf1*, the most promising plastid DNA barcode of land plants. *Scientific Reports*, *5*, 8348. https://doi.org/10.1038/srep08348

Edgar, R. C. (2013). UPARSE: Highly accurate OTU sequences from microbial amplicon reads. *Nature Methods*, *10*, 996. https://doi.org/10.1038/nmeth.2604

Edgar, R. C. (2016). UNOISE 2: Improved error-correction for Illumina 16S and ITS amplicon sequencing. *BioRxiv*, 081257. https://doi.org/10.1101/081257

Edgar, R. C. (2018). Updating the 97% identity threshold for 16s ribosomal rna otus. *Bioinformatics*, *34*, 2371–2375. https://doi.org/10.1093/bioinformatics/bty113

Edgar, R. C., & Flyvbjerg, H. (2015). Error filtering, pair assembly and error correction for next-generation sequencing reads. *Bioinformatics*, *31*, 3476–3482. https://doi.org/10.1093/bioinformatics/btv401

Evans, N. T., Olds, B. P., Renshaw, M. A., Turner, C. R., Li, Y., Jerde, C. L., Mahon, A. R., Pfrender, M. E., Lamberti, G. A., & Lodge, D. M. (2016). Quantification of mesocosm fish and amphibian species diversity via environmental DNA metabarcoding. *Molecular Ecology Resources*, *16*, 29–41. https://doi.org/10.1111/1755-0998.12433

Fantini, E., Gianese, G., Giuliano, G., & Fiore, A. (2015). Bacterial metabarcoding by 16s rrna gene ion torrent amplicon sequencing. *Bacterial Pangenomics*, *1231*, 77–90. https://doi.org/10.1007/978-1-4939-1720-4_5

Giovino, A., Carrubba, A., Lazzara, S., Napoli, E., & Domina, G. (2020a). An integrated approach to the study of Hypericum occurring in Sicily. *Turkish Journal of Botany*, *44*(3), 309–321. http://hdl.handle.net/10447/418310

Giovino, A., Marino, P., Domina, G., Scialabba, A., Schicchi, R., Diliberto, G., Rizza, C., & Scibetta, S. (2016). Evaluation of the DNA barcoding approach to develop a reference data-set for the threatened flora of Sicily. *Plant Biosystems*, *150*(4), 631–640. https://doi.org/10.1080/11263504.2014.989285

Giovino, A., Martinelli, F., & Perrone, A. (2020b). The technique of Plant DNA Barcoding: potential application in floriculture. *Caryologia*, *73*(2), 27–37. https://doi.org/10.13128/caryologia-730

Guo, J. J., Cheng, T., Xu, H., Li, Y., & Zeng, J. (2019). An efficient and cost-effective method for primer-induced nucleotide labeling for massive sequencing on next-generation sequencing platforms. *Scientific Reports*, *9*, 3125. https://doi.org/10.1038/s41598-019-38996-8

Hajibabaei, M., Shokralla, S., Zhou, X., Singer, G. A. C., & Baird, D. J. (2011). Environmental barcoding: A next-generation sequencing approach for biomonitoring applications using river benthos. *PLoS One*, *6*(4), e17497. https://doi.org/10.1371/journal.pone.0017497

Hebert, P. D., Cywinska, A., Ball, S. L., & Dewaard, J. R. (2003). Biological identifications through DNA barcodes. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *270*, 313–321. https://doi.org/10.1098/rspb.2002.2218

Hobern, D., & Hebert, P. (2019). Bioscan-revealing eukaryote diversity, dynamics, and interactions. *Biodiversity Information Science and Standards*, *3*, e37333. https://doi.org/10.3897/biss.3.37333

Hollingsworth, P. M., Graham, S. W., & Little, D. P. (2011). Choosing and Using a Plant DNA Barcode. *PLoS One*, *6*(5), e19254. https://doi.org/10.1371/journal.pone.0019254

Hosein, F. N., Austin, N., Maharaj, S., Johnson, W., Rostant, L., Ramdass, A. C., & Rampersad, S. N. (2017). Utility of DNA barcoding to identify rare endemic vascular plant species in Trinidad. *Ecology and Evolution*, *7*, 7311–7333. https://doi.org/10.1002/ece3.3220

Irinyi, L., Lackner, M., De Hoog, G. S., & Meyer, W. (2016). DNA barcoding of fungi causing infections in humans and animals. *Fungal Biology*, *120*, 125–136. https://doi.org/10.1016/j.funbio.2015.04.007

Kartzinel, T. R., Chen, P. A., Coverdale, T. C., Erickson, D. L., Kress, W. J., Kuzmina, M. L., Rubenstein, D. I., Wang, W., & Pringle, R. M. (2015). DNA metabarcoding illuminates dietary niche partitioning by African large herbivores. *Proceedings of the National Academy of Sciences of the United States of America*, *112*, 8019–8024. https://doi.org/10.1073/pnas.1503283112

Knight, R., Vrbanac, A., Taylor, B. C., Aksenov, A., Callewaert, C., Debelius, J., Gonzalez, A., Kosciolek, T., McCall, L. I., McDonald, D., Melnik, A. V., Morton, J. T., Navas, J., Quinn, R. A., Sanders, J. G., Swafford, A. D., Thompson, L. R., Tripathi, A., Xu, Z. J. Z., … Dorrestein, P. C. (2018). Best practices for analysing microbiomes. *Nature Reviews Microbiology*, *16*, 410–422. https://doi.org/10.1038/s41579-018-0029-9

Krehenwinkel, H., Pomerantz, A., & Prost, S. (2019). Genetic biomonitoring and biodiversity assessment using portable sequencing technologies: Current uses and future directions. *Genes*, *10*, 858. https://doi.org/10.3390/genes10110858

Kress, W. J. (2017). Plant DNA barcodes: Applications today and in the future. *Journal of Systematics Evolution*, *55*, 291–307. https://doi.org/10.1111/jse.12254

Kress, W. J., & Erickson, D. L. (2007). A two-locus global DNA barcode for land plants: The coding *rbcL* gene complements the non-coding trnH-psbA spacer region. *PLoS One*, *2*(6), e508. https://doi.org/10.1371/journal.pone.0000508

Kumar, S., Stecher, G., Li, M., Knyaz, C., & Tamura, K. (2018). MEGA X: Molecular evolutionary genetics analysis across computing platforms. *Molecular Biology and Evolution*, *35*, 1547–1549. https://doi.org/10.1093/molbev/msy096

Lahens, N. F., Ricciotti, E., Smirnova, O., Toorens, E., Kim, E. J., Baruzzo, G., Hayer, K. E., Ganguly, T., Schug, J., & Grant, G. R. (2017). A comparison of Illumina and Ion Torrent sequencing platforms in the context of differential gene expression. *BMC Genomics*, *18*, 602. https://doi.org/10.1186/s12864-017-4011-0

Li, D. Z., Gao, L. M., Li, H. T., Wang, H., Ge, X. J., Liu, J. Q., Chen, Z. D., Zhou, S. L., Chen, S. L., Yang, J. B., Fu, C. X., Zeng, C. X., Yan, H. F., Zhu, Y. J., Sun, Y. S., Chen, S. Y., Zhao, L., Wang, K., Yang, T., & Duan, G. W. (2011). Comparative analysis of a large dataset indicates

that internal transcribed spacer (ITS) should be incorporated into the core barcode for seed plants. *Proceedings of the National Academy of Sciences*, *108*, 19641–19646. https://doi.org/10.1073/pnas.1104551108

Li, J. L., Wang, S., Yu, J., Wang, L., & Zhou, S. L. (2013). A modified CTAB protocol for plant DNA extraction. *Chinese Bulletin of Botany*, *48*, 72–78. https://doi.org/10.3724/SP.J.1259.2013.00072

Magoc, T., & Salzberg, S. L. (2011). Flash: Fast length adjustment of short reads to improve genome assemblies. *Bioinformatics*, *27*, 2957–2963. https://doi.org/10.1093/bioinformatics/btr507

Marine, R. L., Magana, L. C., Castro, C. J., Zhao, K., Montmayeur, A. M., Schmidt, A., Diez-Valcarce, M., Ng, T. F. F., Vinje, J., Burns, C. C., Nix, W. A., Rota, P. A., & Oberste, M. S. (2020). Comparison of Illumina MiSeq and the Ion Torrent PGM and S5 platforms for whole-genome sequencing of picornaviruses and caliciviruses. *Journal of Virol Methods*, *280*, 113865. https://doi.org/10.1016/j.jviromet.2020.113865

Mora, C., Tittensor, D. P., Adl, S., Simpson, A. G., & Worm, B. (2011). How many species are there on earth and in the ocean? *PLoS Biology*, *9*, e1001127. https://doi.org/10.1371/journal.pbio.1001127

Nakamura, T., Yamada, K. D., Tomii, K., & Katoh, K. (2018). Parallelization of Mafft for large-scale multiple sequence alignments. *Bioinformatics*, *34*, 2490–2492. https://doi.org/10.1093/bioinformatics/bty121

Nithaniyal, S., Vassou, S. L., Poovitha, S., Raju, B., & Parani, M. (2017). Identification of species adulteration in traded medicinal plant raw drugs using DNA barcoding. *Genome*, *60*, 139–146. https://doi.org/10.1139/gen-2015-0225

Patel, R. K., & Jain, M. (2012). NGS QC toolkit: A toolkit for quality control of next generation sequencing data. *PLoS One*, *7*(2), e30619. https://doi.org/10.1371/journal.pone.0030619

Piry, S., Guivier, E., Realini, A., & Martin, J. F. (2012). |SE|S|AM|E| barcode: NGS-oriented software for amplicon characterization - application to species and environmental barcoding. *Molecular Ecology Resources*, *12*, 1151–1157. https://doi.org/10.1111/j.1755-0998.2012.03171.x

Quail, M. A., Smith, M., Coupland, P., Otto, T. D., Harris, S. R., Connor, T. R., Bertoni, A., Swerdlow, H. P., & Gu, Y. (2012). A tale of three next generation sequencing platforms: Comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics*, *13*, 341. https://doi.org/10.1186/1471-2164-13-341

Richardson, R. T., Bengtsson-Palme, J., & Johnson, R. M. (2017). Evaluating and optimizing the performance of software commonly used for the taxonomic classification of DNA metabarcoding sequence data. *Molecular Ecology Resources*, *17*, 760–769. https://doi.org/10.1111/1755-0998.12628

Rognes, T., Flouri, T., Nichols, B., Quince, C., & Mahe, F. (2016). Vsearch: A versatile open source tool for metagenomics. *Peerj*, *4*(10), 2584. https://doi.org/10.7717/peerj.2584

Salipante, S. J., Kawashima, T., Rosenthal, C., Hoogestraat, D. R., Cummings, L. A., Sengupta, D. J., Harkins, T. T., Cookson, B. T., & Hoffman, N. G. (2014). Performance comparison of Illumina and Ion Torrent next-generation sequencing platforms for 16S rRNA-based bacterial community profiling. *Applied and Environmental Microbiology*, *80*, 7583–7591. https://doi.org/10.1128/AEM.02206-14

Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., Lesniewski, R. A., Oakley, B. B., Parks, D. H., Robinson, C. J., Sahl, J. W., Stres, B., Thallinger, G. G., Van, D. J., & Weber, C. F. (2009). Introducing MOTHUR: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and Environmental Microbiology*, *75*, 7537–7541. https://doi.org/10.1128/AEM.01541-09

Schmidt, P. A., Bálint, M., Greshake, B., Bandow, C., Römbke, J., & Schmitt, I. (2013). Illumina metabarcoding of a soil fungal community. *Soil Biology and Biochemistry*, *65*, 128–132. https://doi.org/10.1016/j.soilbio.2013.05.014

Shi, Z. Y., Yang, C. Q., Hao, M. D., Wang, X. Y., Ward, R. D., & Zhang, A. B. (2018). FUZZYID 2: A software package for large data set species identification via barcoding and metabarcoding using hidden Markov models and fuzzy set methods. *Molecular Ecology Resources*, *18*, 666–675. https://doi.org/10.1111/1755-0998.12738

Shokralla, S., Gibson, J. F., Nikbakht, H., Janzen, D. H., Hallwachs, W., & Hajibabaei, M. (2014). Next-generation DNA barcoding: Using next-generation sequencing to enhance and accelerate DNA barcode capture from single specimens. *Molecular Ecology Resources*, *14*, 892–901. https://doi.org/10.1111/1755-0998.12236

Shokralla, S., Spall, J. L., Gibson, J. F., & Hajibabaei, M. (2012). Next-generation sequencing technologies for environmental DNA research. *Molecular Ecology*, *21*, 1794–1805. https://doi.org/10.1111/j.1365-294X.2012.05538.x

Speranskaya, A. S., Khafizov, K., Ayginin, A. A., Krinitsina, A. A., Omelchenko, D. O., Nilova, M. V., Severova, E. E., Samokhina, E. N., Shipulin, G. A., & Logacheva, M. D. (2018). Comparative analysis of Illumina and Ion Torrent high-throughput sequencing platforms for identification of plant components in herbal teas. *Food Control*, *93*, 315–324. https://doi.org/10.1016/j.foodcont.2018.04.040

Srivathsan, A., Baloglu, B., Wang, W., Tan, W. X., Bertrand, D., Ng, A. H. Q., Boey, E. J. H., Koh, J. J. Y., Nagarajan, N., & Meier, R. (2018). A minion-based pipeline for fast and cost-effective DNA barcoding. *Molecular Ecology Resources*, *18*, 1035–1049. https://doi.org/10.1111/1755-0998.12890

Thorne, R. F. (2002). How many species of seed plants are there? *Taxon*, *51*, 511–512. https://doi.org/10.2307/1554864

Toju, H. (2015). High-throughput DNA barcoding for ecological network studies. *Population Ecology*, *57*, 37–51. https://doi.org/10.1007/s10144-014-0472-z

Tyagi, K., Kumar, V., Kundu, S., Pakrashi, A., Prasad, P., Caleb, J. T. D., & Chandra, K. (2019). Identification of Indian spiders through DNA barcoding: Cryptic species and species complex. *Scientific Reports*, *9*, 14033. https://doi.org/10.1038/s41598-019-50510-8

Xu, C., Dong, W. P., Shi, S., Cheng, T., Li, C. H., Liu, Y. L., Wu, P., Wu, H. K., Gao, P., & Zhou, S. L. (2015). Accelerating plant DNA barcode reference library construction using herbarium specimens: Improved experimental techniques. *Molecular Ecology Resources*, *15*, 1366–1374. https://doi.org/10.1111/1755-0998.12413

Xu, S. Z., Li, Z. Y., & Jin, X. H. (2018). DNA barcoding of invasive plants in china: A resource for identifying invasive plants. *Molecular Ecology Resources*, *18*, 128–136. https://doi.org/10.1111/1755-0998.12715

Yamamoto, S., Masuda, R., Sato, Y., Sado, T., Araki, H., Kondoh, M., Minamoto, T., & Miya, M. (2017). Environmental DNA metabarcoding reveals local fish communities in a species-rich coastal sea. *Scientific Reports*, *7*, 40368. https://doi.org/10.1038/srep40368

Yu, J., Xue, J. H., & Zhou, S. L. (2011). New universal *matK* primers for DNA barcoding angiosperms. *Journal of Systematics and Evolution*, *49*, 176–181. https://doi.org/10.1111/j.1759-6831.2011.00134.x

Zhang, W., Sun, Y. Z., Liu, J., Xu, C., Zou, X. H., Chen, X., Liu, Y. L., Wu, P., Yang, X. Y., & Zhou, S. L. (2020). DNA barcoding of Oryza: Conventional, specific, and super barcodes. *Plant Molecular Biology*, *1*, 14. https://doi.org/10.1007/s11103-020-01054-3

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Liu, Y., Xu C., Sun Y., Chen X., Dong W., Yang X., & Zhou S. (2021). Method for quick DNA barcode reference library construction. *Ecology and Evolution*, 11, 11627–11638. https://doi.org/10.1002/ece3.7788