An rbcL Reference Library to Aid in the Identification of Plant Species Mixtures by DNA Metabarcoding

Karen L. Bell, Emory University
Virginia M. Loeffler, Emory University
Berry Brosi, Emory University

Journal Title: Applications in Plant Sciences
Volume: Volume 5, Number 3
Publisher: Botanical Society of America: APPS | 2017-03-10, Pages 1600110-1600110
Type of Work: Article | Final Publisher PDF
Publisher DOI: 10.3732/apps.1600110
Permanent URL: https://pid.emory.edu/ark:/25593/s0x2v

Final published version: http://dx.doi.org/10.3732/apps.1600110

Copyright information:
© 2017 Bell et al. Published by the Botanical Society of America
This is an Open Access work distributed under the terms of the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License (http://creativecommons.org/licenses/by-nc-sa/4.0/).

Accessed July 19, 2019 8:26 PM EDT
An *rbcL* Reference Library to Aid in the Identification of Plant Species Mixtures by DNA Metabarcoding

Author(s): Karen L. Bell, Virginia M. Loeffler, and Berry J. Brosi

Source: Applications in Plant Sciences, 5(3)

Published By: Botanical Society of America

DOI: [http://dx.doi.org/10.3732/apps.1600110](http://dx.doi.org/10.3732/apps.1600110)

AN *rbcL* REFERENCE LIBRARY TO AID IN THE IDENTIFICATION OF PLANT SPECIES MIXTURES BY DNA METABARCODING

KAREN L. BELL2,3,4, VIRGINIA M. LOEFFLER2, AND BERRY J. BROS2

2 Department of Environmental Science, Emory University, 400 Dowman Drive, Atlanta, Georgia 30322 USA

DNA metabarcoding, the use of high-throughput sequencing methods to simultaneously DNA barcode all species in a mixed sample, can be used to address a variety of questions in ecology, biosecurity, and forensics, among other fields (Cristescu, 2014; Bell et al., 2016a, 2016b). These methods enable taxonomic identifications of plant samples in which diagnostic characters are largely absent (e.g., roots in a soil sample, seeds from seed traps, or pollen), or for which few experts know the diagnostic characters. DNA metabarcoding has streamlined identifications of pollen for allergen monitoring (Kraaijeveld et al., 2015), plant-pollinator interactions (Keller et al., 2015; Richardson et al., 2015a, 2015b; Sickel et al., 2015; Pornon et al., 2016), and honey composition (Hawkins et al., 2015). The analysis of fragments of plant material in herbivore guts enables diet analysis (Valentini et al., 2009; Pomponio et al., 2012). DNA metabarcoding can also be used to determine the species composition of dietary supplements (Cheng et al., 2014) and foods (Hawkins et al., 2015) leading to improved health and safety. However, accurate determination of species by DNA metabarcoding is dependent on a comprehensive and accurate reference library of DNA sequences of a standard genetic marker, as well as an appropriate bioinformatics pipeline.

For plants, the Consortium for the Barcode of Life (CBOL) Plant Working group recommended the plastid DNA (ptDNA) genes *rbcL* and *matK* as standard DNA barcode markers, based on the availability of universal primers and the high level of taxonomic resolution (CBOL Plant Working Group, 2009; Hollingsworth et al., 2009). Recent studies using the standard *rbcL*+*matK* barcode for florals of moderate phylogenetic dispersion have shown that up to 92% of the species can be distinguished (Kress et al., 2009; Burgess et al., 2011). Amplicon fragment size is important in high-throughput sequencing (HTS) DNA metabarcoding, as the read lengths in many platforms are currently limited (e.g., ~600-bp paired-end reads on the Illumina MiSeq platform; Illumina, San Diego, California, USA). The long amplicon length generated by standard *matK* primers poses a technical limitation for these sequencing methods. In addition to ptDNA markers, the nuclear ribosomal spacer ITS2 has been shown to be an effective DNA barcoding marker, with 92.7% successful identifications in 6600 samples (Chen et al., 2010).

A bioinformatics pipeline has been published for the identification of species mixtures using ITS2 (Sickel et al., 2015). This bioinformatics pipeline was developed using dual-indexed mixed-amplicon sequence data from the Illumina MiSeq platform (adapting the method of Kozich et al., 2013), but can be used for any high-throughput sequencing platform that outputs data in FASTQ format. Sequencing reads are compared to...
a reference library of ITS2 sequences using two different programs—RDP classifier (Wang et al., 2007) and UTAX (Edgar, 2010). This pipeline currently cannot be used with either of the standard DNA barcode markers (rbcL or matK), due to the lack of reference libraries.

We report here on our development of reference libraries for taxonomic classification with RDP classifier and UTAX using rbcL sequences obtained from the National Center for Biotechnology Information (NCBI) database. We also made modifications to the bioinformatics pipeline of Sickel et al. (2015) to enable identifications from multiple markers amplified from the same sample, and these modifications could be applied to further additional barcodes, such as matK or trnL. We have used our rbcL reference libraries, along with existing ITS2 libraries and our modified bioinformatics pipeline, to classify sequences obtained from a constructed pollen sample composed of nine known species. We were able to successfully identify eight of the nine species using a combination of rbcL and ITS2, where neither marker was sufficient for identification to the species level for more than six species on its own.

MATERIALS AND METHODS

Developing the rbcL database—We selected rbcL, rather than matK, as our second DNA barcode, because the shorter PCR product allows for overlap of forward and reverse reads on the Illumina MiSeq. We downloaded all available seed plant rbcL sequences (as of 27 January 2016) from NCBI, using the following search: (rbcL[Gene Name] AND 50:400000000[Sequence Length]) AND “seed plants”[organism]. This included sequences that were predominantly rbcL, sequences with a small fragment of rbcL sequence along with a longer sequence of intergenic spacer, and complete tDNA genomes. We extracted the rbcL sequences from these records, using the “extract annotations” tool in Geneious version 7.1.9 (Biomatters, Auckland, New Zealand). We then filtered these extractions to remove any sequences of less than 100 bp. Previous studies have found interspecific sequence divergences of around 1–2% for rbcL (Hollingsworth et al., 2009; Kress and Erickson, 2007), so sequences of less than 100 bp would be unlikely to be informative. All sequences were exported as a single FASTA file from Geneious, with the “LOCUS” and “DEFINITION” fields from NCBI as the sequence header. We downloaded full details of each GenBank record, including the accession number and GenBank Identification (GI) number, from Geneious as a comma-separated value (.csv) file, prior to extraction of the rbcL annotated regions, because extraction with Geneious often causes this information to be lost, leaving only the “LOCUS” field to determine the origin of the sequence. We used R (R Core Team, 2016) scripts to replace the “LOCUS” field information from the sequence headers of the FASTA file with the GI number, accession number, and species for each sequence. We included both identifiers (rather than GI only) to be compatible with future bioinformatics pipelines, following the phasing out of GI numbers by NCBI in September 2016. Reference databases for RDP classifier and UTAX were built from the rbcL sequences using the method described in Sickel et al. (2015) for training of the ITS2 databases; these were then deposited on Figshare, along with the FASTA file described above (https://dx.doi.org/10.6084/m9.figshare.c.3463311). The FASTA file can be used to develop reference libraries for other classification programs, if desired.

Adapting existing bioinformatics pipeline—We accessed the bioinformatics pipeline of Sickel et al. (2015) through their website at https://github.com/imoj/mega-barcoding-dual-indexing, along with instructions for installing all of the required programs. Their bioinformatics pipeline includes commands for joining forward and reverse reads using QiIME version 1.8.0 (Caporaso et al., 2010), removing low-quality reads with USEARCH version 8.0.1477 (Edgar, 2010), and taxonomic classification of the remaining high-quality sequences using RDP classifier (Wang et al., 2007) and/or UTAX (Edgar, 2010). Although there are other algorithms for matching sequences (e.g., BLAST), algorithms like RDP and UTAX are faster because they cluster sequences before searching them against the databases, and they give more robust results as they consider the taxonomic hierarchy, rather than just returning the best hit.

The pipeline of Sickel et al. (2015) has been designed to be run with their database of ITS2 sequences (Ankenbrand et al., 2015), but can easily be run with other databases, provided the sequences are available. Once the pipeline has been run in its entirety, it is not necessary to run the entire pipeline to compare the joined and filtered FASTQ files against a second, or further, database of sequences. We removed the lines associated with joining and filtering FASTQ files from the script ‘classify_reads.pl’ of Sickel et al. (2015) to allow the FASTQ files that were joined and filtered before searching against the first database (in this case, ITS2) using the full bioinformatics pipeline, to be compared to other databases (in this case, rbcL sequences). We have called this script ‘classify_reads_subsequent.pl’ and made it available from https://github.com/KarenBell/ rbcL-dual-index-metabarcoding. The basic workflow of the modified bioinformatics pipeline, assuming analysis with ITS2 followed by analysis with rbcL, is depicted in Fig. 1.

Testing the database—We tested the newly developed rbcL database, in combination with the existing ITS2 database, using sequencing data from an artificial pollen sample composed of nine taxonomically diverse angiosperm species: *Populus tremuloides* Michx. (Salicaceae) (Sigma-Aldrich Co., St. Louis, Missouri, USA), *Populus deltoides* W. Bartram ex Marshall (Sigma-Aldrich Co.), *Broussonetia papyrifera* (L.) Vent. (Moraceae) (Polysciences, Warrington, Pennsylvania, USA), *Carya illinoinensis* (Wangenh.) K. Koch (Juglandaceae) (Polysciences), *Bassia scoparia* (L.) J. A. Scott (Amaranthaceae) (Sigma-Aldrich Co.), *Ambrosia artemisiifolia* L. (Asteraceae) (Polysciences), *Artemisia tridentata* Nutt. (Asteraceae) (Sigma-Aldrich Co.), *Poa pratensis* L. (Poaaceae) (Sigma-Aldrich Co.), and *Zea mays* L. (Poaaceae) (Carolina Biological Supply, Burlington, North Carolina, USA). These species were selected because they were available commercially and cover a broad range of plant taxonomic diversity. We suspended pollen of each species in a mixture of 1:3 glycerol:ethanol and estimated the concentration of pollen grains in these suspensions through microscopic examination. We added an appropriate volume of each suspension to make a mixture containing similar numbers of pollen grains from each species and a total of 1,000,000 pollen grains. We determined the actual composition of this mixture by examining four replicate microscope slides of approximately 200 pollen grains from each mixture using Classifyandyn, an automated pollen identification and counting system (Holt et al., 2011). We tested for any bias in the Classifyandyn identification and counting through visual examination of one slice by three different individuals, comparing these counts to the automated counts.

We isolated DNA from the constructed pollen mixture, and a negative control, using the FastDNA SPIN Kit for Soil (MP Biomedicals, Solon, Ohio, USA), with the following modifications: 150 μL 10% sodium dodecyl sulfate (SDS) (Promega Corporation, Madison, Wisconsin, USA) and 20 μL 10 mg/mL Proteinase K (QIAGEN, Valencia, California, USA) were added to the sample along with the MT Buffer; homogenization was conducted using a mini-bead beater (Biospec Products, Bartlesville, Oklahoma, USA) for 3 min, followed by incubation for 30 min at 55°C, followed by a further 3 min in the mini-bead beater (preliminary analysis of three different species showed that this was sufficient to rupture the majority of pollen grains); and DNA was eluted in 50 μL DES (DNasefree water), with a 5-min incubation at 55°C.

We amplified the rbcL and ITS2 markers in separate reactions using a modification of methods developed for mixed-amplicon sequencing of bacterial 16S (Kozich et al., 2013). Primers for the first round of PCR were standard DNA barcoding primers with Illumina overhang sequences appended (Kozich et al., 2013). For rbcL, we used the primers rbcL2, which binds near the 5’ end of the *rbcL* gene (Palmieri et al., 2009), and rbcLa-R, which binds near the middle of the *rbcL* gene (Kress and Erickson, 2007), which yield a PCR product of ~500 bp. For ITS2 we used the primers ITS2 S2F and ITS2 S3R, which bind to 5.8S and 3’ end of ITS2, respectively, and yield a PCR product of 350–400 bp (Chen et al., 2010). PCR reactions contained KAPA HiFi HotStart ReadyMix (KAPA Biosystems, Boston, Massachusetts, USA), each primer at a final concentration of 200 nM, and 12 μL of a 1/5 dilution of DNA isolate (43 ng total DNA) in a total volume of 25 μL. PCR cycles included an initial heat activation for 3 min at 95°C; followed by 30 cycles of 30 s at 95°C, 30 s at 55°C, and 50 s at 72°C; followed by a final extension of 7 min at 72°C. The PCR was conducted in triplicate. In addition to the negative control from the DNA isolation step, we included a PCR negative control (water only). We prepared indexed Illumina MiSeq libraries from these PCR products using a limited cycle PCR, adding Illumina sequencing adapters and Nextera XT dual-index barcodes (Illumina). We used different index combinations for each triplicate and negative control, with the same index being used for *rbcL* and ITS2 for the same sample (five index combinations and 10 PCR reactions in total). We purified the PCR products.
using Agencourt AMPure XP magnetic beads (Beckman Coulter, Danvers, Massachusetts, USA), and the DNA was quantified at the Emory Integrated Genomics Core (EIGC) using the Bioanalyzer instrument (Agilent Technologies, Santa Clara, California, USA). The PCR reactions were multiplexed with other samples, allocating each PCR reaction 1/192 of the total DNA content on a single flow cell (i.e., allocating each sample 1/96 of the total DNA once rbcL and ITS2 PCR are combined), except for negative controls that were multiplexed at the same volume as the most dilute sample (i.e., the maximum volume added).
Sequencing was conducted by the EIGC, with a 2 × 300 bp run of the Illumina MiSeq instrument.

We used the bioinformatics pipeline of Sickel et al. (2015) to classify ITS2 reads. We then used our modified Perl script and our databases to classify rbcL reads. After classification, we removed any identifications occurring at a lower frequency than identifications obtained in either negative control. We compared classification success of ITS2 alone, rbcL alone, and both markers combined, using the RDP and UTAX classification methods. Success was defined in terms of correct identification to any taxonomic level, correct identification to species level, and incorrect identifications.

When we were unable to detect a species with either ITS2 or rbcL, we checked for sequences in the NCBI database that might be able to identify the species (e.g., if the sequences were not identified as ITS2 or rbcL in the NCBI database). We selected one of our three triplicate samples, compared the first 1000 sequences against the entire NCBI database using a BLAST search, and checked for matches to any species that had not been identified using the method described in the previous paragraphs.

RESULTS

The rbcL database—We successfully developed a database containing rbcL sequences from 38,409 seed plant species. This compares to 72,325 species in the ITS2 database (Sickel et al., 2015) and an estimated 450,000 flowering plant species on earth (Pimm and Joppa, 2015). For the species included in our artificial pollen mixture, most were represented by multiple sequences, many vouched. For most species, voucher congeneric sequences were also available, although sometimes only for a small proportion of the total diversity of the genus (Table 1). The majority of sequences were in the length range of 501–1000 bp (Fig. 2). This category would include typical DNA barcode sequences, which are 600 bp for the most commonly used primer pair (Kress and Erickson, 2007). Full-length rbcL sequences are around 2500 bp in length and only represent a small portion of the sequences in our database.

Testing the database—Our three technical replicates produced 85,378, 38,398, and 74,843 sequencing reads, giving a total of 198,619 sequencing reads. Analysis with RDP identified 65,639 of these reads as ITS2 and 130,320 as rbcL, leaving 2660 unclassified. Analysis with UTAX identified 67,149 ITS2 reads, 131,260 rbcL reads, with 210 unclassified reads. Our isolation negative control contained 34 sequencing reads, while our PCR negative control contained 30 sequencing reads.

In most cases, taxonomic assignments were able to identify all nine species present in the mixture, although certain species were missed in some analyses, either due to low numbers of reads or absence in reference libraries (Table 2). Similar results were obtained with the same analysis method for each triplicate PCR (Table 3). Based on ITS2 analysis alone, we were unable to detect Artemisia tridentata or Populus tremuloides. In both cases, a closely related species was detected, A. keiskeana Miq. and P. alba L., respectively. Populus tremuloides is not represented in the ITS2 reference library. Zea mays could be detected, but with fewer reads than our false-positive threshold (i.e., fewer than the number of reads obtained in negative controls). Based on rbcL analysis alone, we were unable to detect Carvyla illinoinsis. Instead, C. illinoinsis was identified as Juglans regia L. or Pterocarya stenoptera C. DC., or the reads were identified correctly to family level. Ambrosia artemisiifolia was detected from only a small number of reads (below the contamination threshold) with rbcL, while Artemisia tridentata was entirely undetected, due to its absence in the rbcL reference library. The only species that could not be detected above the contamination threshold with at least one marker was A. tridentata. A BLAST search of our first 1000 sequence reads against the NCBI database (accessed 28–29 November 2016), saving the 10 best matches for each search, included identifications of A. tridentata ITS2, but these identifications were never the single highest score. No sequencing reads were identified as A. tridentata rbcL. A BLAST search of the complete data set of 198,619 sequences against the entire NCBI database was not feasible.

Table 1. Sequence representation on NCBI of the nine species used for testing our rbcL reference library, and congeneric species.

<table>
<thead>
<tr>
<th>Species</th>
<th>No. of rbcL sequences</th>
<th>No. of vouched rbcL sequences</th>
<th>No. of species in genus</th>
<th>No. of congeneric species with rbcL sequences</th>
<th>No. of congeneric species with vouched rbcL sequences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Populus deltoides</td>
<td>10</td>
<td>6</td>
<td>174</td>
<td>46</td>
<td>28</td>
</tr>
<tr>
<td>Populus tremuloides</td>
<td>14</td>
<td>7</td>
<td>174</td>
<td>46</td>
<td>28</td>
</tr>
<tr>
<td>Broussonetia papyrifera</td>
<td>15</td>
<td>11</td>
<td>31</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Carya illinoinsis</td>
<td>2</td>
<td>1</td>
<td>47</td>
<td>10</td>
<td>6</td>
</tr>
<tr>
<td>Bassia scoparia</td>
<td>3</td>
<td>1</td>
<td>217</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>Ambrosia artemisiifolia</td>
<td>9</td>
<td>9</td>
<td>76</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>Artemisia tridentata</td>
<td>0</td>
<td>0</td>
<td>870</td>
<td>27</td>
<td>16</td>
</tr>
<tr>
<td>Poa pratensis</td>
<td>62</td>
<td>45</td>
<td>1613</td>
<td>94</td>
<td>93</td>
</tr>
<tr>
<td>Zea mays</td>
<td>13</td>
<td>0</td>
<td>32</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

*A sequence is considered to be vouched if the word “voucher” was found in the NCBI record. Further sequences may be associated with vouchers that are not included in the NCBI record, but are recorded in the associated publication.

*Information obtained from the Tropicos online database (Missouri Botanical Garden, 2016). Species names that were invalid or illegitimate were excluded.

http://www.bioone.org/loi/apps
(3 sequences), *Poa pratensis* *rbcL* (2 sequences), *Populus* *rbcL* (2 sequences), *Ambrosia* *rbcL* (1 sequence), *Artemisia annua* L. *rbcL* (1 sequence), rosid *rbcL* (1 sequence), *Populus deltoides* *rbcL* (1 sequence), Rosales *rbcL* (1 sequence), *Bassia scoparia* *rbcL* (1 sequence), and *B. scoparia* ITS2 (1 sequence), with five sequences unidentified. Taxonomic identifications from the UTAX analysis were *Ambrosia artemisiifolia* ITS2 (8 sequences), *Populus tremuloides* *rbcL* (3 sequences), *Hulcus lanatus* L. *rbcL* (2 sequences), *Artemisia frigida* *rbcL* (2 sequences), *Artemisia maritima* *rbcL* (2 sequences), *Festuca subverticillata* (Pers.) E. B. Alexeev *rbcL* (1 sequence), *Vulpia unilateralis* (L.) Stace *rbcL* (1 sequence), *Poa pratensis* *rbcL* (1 sequence), *Ambrosia artemisiifolia* *rbcL* (1 sequence), *Bassia scoparia* *rbcL* (1 sequence), and *B. scoparia* ITS2 (1 sequence), with six sequences unidentified.

**DISCUSSION**

In this study, we have developed an *rbcL* reference sequence database and incorporated this into an existing ITS2 bioinformatics pipeline. By formatting the reference libraries for use with an existing bioinformatics pipeline, we have allowed for a streamlined analysis of sequencing data containing both ITS2 and *rbcL* reads. We have also allowed for improved standardization across studies, by developing a database for one of the two standard DNA barcoding markers (CBOL Plant Working Group, 2009). By combining these two markers, we were able to achieve more accurate species-level identifications than using ITS2 alone. The combination of ITS2 and *rbcL* sequence data enabled us to identify to species level eight of nine plant species in a mixture, and we were able to detect all genera. This compares to only six species-level identifications based on ITS2 only. We obtained fewer high-quality sequence reads from *rbcL* than ITS2. This may be because the *rbcL* fragment is longer, with less overlap between forward and reverse reads, causing more reads to be removed during filtering. If it were strongly desirable to have equal numbers of high-quality sequences from both markers, this could be corrected by adding more *rbcL* PCR products at the pooling step before MiSeq analysis.

The most common explanations for misidentifications were the absence of the species in the reference library or the lack of species-level variation in sequences. While our reference library represents a significant advance in plant DNA metabarcoding, our *rbcL* reference library and the ITS2 reference library of Sickel et al. (2015) fall far short of being comprehensive. There are an estimated 450,000 flowering plant species on earth (Pimm

---

**Table 2.** Summary of identification success, based on analysis of ITS2 and *rbcL* sequence data, using RDP and UTAX, for a nine-species mixture.

<table>
<thead>
<tr>
<th>Species</th>
<th>RDP</th>
<th></th>
<th></th>
<th></th>
<th>UTAX</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ITS2</td>
<td>rbcL</td>
<td>Both</td>
<td></td>
<td>ITS2</td>
<td>rbcL</td>
<td>Both</td>
<td></td>
</tr>
<tr>
<td><em>Ambrosia artemisiifolia</em></td>
<td>Species</td>
<td>Family</td>
<td>Species</td>
<td></td>
<td>Species</td>
<td>Undetected</td>
<td>Species</td>
<td></td>
</tr>
<tr>
<td><em>Artemisia tridentata</em></td>
<td>Misidentified</td>
<td>Family</td>
<td>Family</td>
<td></td>
<td>Misidentified</td>
<td>Misidentified</td>
<td>Misidentified</td>
<td></td>
</tr>
<tr>
<td><em>Bassia scoparia</em></td>
<td>Species</td>
<td>Species</td>
<td>Species</td>
<td></td>
<td>Species</td>
<td>Species</td>
<td>Species</td>
<td></td>
</tr>
<tr>
<td><em>Broussonetia papyrifera</em></td>
<td>Species</td>
<td>Undetected</td>
<td>Species</td>
<td></td>
<td>Species</td>
<td>Family</td>
<td>Species</td>
<td></td>
</tr>
<tr>
<td><em>Carya illinoinensis</em></td>
<td>Species</td>
<td>Species</td>
<td>Species</td>
<td></td>
<td>Species</td>
<td>Species</td>
<td>Species</td>
<td></td>
</tr>
<tr>
<td><em>Poa pratensis</em></td>
<td>Species</td>
<td>Species</td>
<td>Species</td>
<td></td>
<td>Species</td>
<td>Species</td>
<td>Species</td>
<td></td>
</tr>
<tr>
<td><em>Populus deltoides</em></td>
<td>Species</td>
<td>Species</td>
<td>Species</td>
<td></td>
<td>Species</td>
<td>Species</td>
<td>Species</td>
<td></td>
</tr>
<tr>
<td><em>Populus tremuloides</em></td>
<td>Misidentified</td>
<td>Species</td>
<td>Species</td>
<td></td>
<td>Misidentified</td>
<td>Species</td>
<td>Species</td>
<td></td>
</tr>
<tr>
<td><em>Zea mays</em></td>
<td>Undetected</td>
<td>Misidentified</td>
<td>Misidentified</td>
<td></td>
<td>Undetected</td>
<td>Species</td>
<td>Species</td>
<td></td>
</tr>
</tbody>
</table>
and Joppa, 2015). Our databases represent 10–20% of this biodiversity. Furthermore, many online data repositories, such as the International Nucleotide Sequences Database Collaboration (http://www.ncbi.nlm.nih.gov/genbank/collab), which incorporates GenBank (Benson et al., 2015), the DNA Data Bank of Japan (DDBJ; Mashima et al., 2015), and the European Molecular Biology Laboratory (EMBL) Nucleotide Sequence Database (Squizzato et al., 2015), do not require researchers submitting sequences to provide links to voucher specimens or raw sequence data, and may include sequences with incorrect species identification or low-quality sequence data. Other databases, such as the Barcode of Life Database (BOLD; Ratnasingham and Hebert, 2007), have stringent quality control but fewer plant \( rbcL \) sequences than the NCBI database, and no plant ITS2 sequences. As online databases become populated with more species, particularly those such as BOLD that have stringent quality control, these sequences can be added to our \( rbcL \) database, providing more comprehensive and accurate reference libraries for DNA metabarcoding analyses, which will lead to improvements in species-level identifications.

A lack of species-level variation was also found to hinder species-level identifications, particularly when only considering data from ITS2 or \( rbcL \) individually. This was likely the case with Artemisia tridentata being identified as A. keiskeana. Using ITS2 and \( rbcL \) in combination, most species in the mixtures could be identified. By providing a reference library for \( rbcL \), we have improved identification accuracy by allowing identification from multiple markers. When DNA metabarcoding a sample of unknown species composition, conflicting results from different loci or different analysis methods (e.g., the different results for \( rbcL \) from ITS2 or \( rbcL \) of ITS2 sequences) may be difficult to resolve. In these situations, it may be necessary to resolve the taxonomic identification at a higher rank (e.g., a genus-level identification of \( Populus \)). Adding a third marker, such as \( matK \) or \( trnL \), may also help here.

Two of the species we examined, \( Zea mays \) and Ambrosia artemisiifolia, were undetected above the contamination threshold with one of the two barcodes (ITS2 and \( rbcL \), respectively), rather than misidentified. This is not due to poor DNA isolation for these species, as they were both detected with the other marker. This is probably best explained by amplification bias, where certain species are less readily amplified under the combination of primers and PCR conditions, or copy number bias, where the markers are present in lower copy numbers in the genomes of some species. This emphasizes the importance of using multiple unlinked markers in DNA metabarcoding studies.

Increasing the accuracy of DNA barcoding and metabarcoding will advance research across a range of disciplines. DNA metabarcoding methods enable taxonomic identifications of mixed-species plant samples in which diagnostic characters may be absent, including pollen (Bell et al., 2016b), herbivore gut contents (Valentini et al., 2009; Pompanon et al., 2012), and traditional medicines (Cheng et al., 2014). All of these identifications are dependent on comprehensive and accurate reference libraries of DNA sequences of standard genetic markers, such as the \( rbcL \) library of the current study. In the future, as similar databases become available for other molecular markers, such as \( matK \) or \( trnL \), this will allow for even greater accuracy of taxonomic classification by DNA metabarcoding, and subsequent advancement in a multitude of fields of research.

### LITERATURE CITED


Applications in Plant Sciences 2017 5(3): 1600110

Bell et al.—rbcL reference library for plant species mixtures

doi:10.3732/apps.1600110

Applications in Plant Sciences 2017 5(3): 1600110 Bell et al.—rbcL reference library for plant species mixtures
doi:10.3732/apps.1600110


