Using Herbarium-Derived DNAs to Assemble a Large-Scale DNA Barcode Library for the Vascular Plants of Canada

Authors: Kuzmina, Maria L., Braukmann, Thomas W. A., Fazekas, Aron J., Graham, Sean W., Dewaard, Stephanie L., et al.

Source: Applications in Plant Sciences, 5(12)

Published By: Botanical Society of America

URL: https://doi.org/10.3732/apps.1700079
APPLICATION ARTICLE

USING HERBARIUM-DERIVED DNAs TO ASSEMBLE A LARGE-SCALE DNA BARCODE LIBRARY FOR THE VASCULAR PLANTS OF CANADA

MARI A L. KUZMINA2,15, THOMAS W. A. BRAUKMANN2, ARON F. FAZEKAS3, SEAN W. GRAH AM4, STEPHANIE L. DEWAARD2, ANUAR RODRIGUE5S5, BRUCE A. BENNETT6, TIMOTHY A. DICKINSON7,8, JEFFERY M. SAARELA9, PAUL M. CATLING10, STEVEN G. NEWMASTER11, DIANA M. PERCY12, ERIN FENNEMAN4, AU RÉLIEN LAURON-MOREAU13, BRUCE FORD14, LYNN GIL LESPIE9, RAGUPATHY SUBRAMANYAM11, JEANNETTE WHITTON4, LINDA JENNINGS4, DEBORAH METSGER7, CON N OR P. WARNE2, ALLISON BROWN2, ELI ZABE RSEARS2, JEREMY R. DEWAARD2, EVGENY V. ZAKHAROV2, AND PAUL D. N. HEBERT2

2Centre for Biodiversity Genomics, University of Guelph, 50 Stone Road East, Guelph, Ontario N1G 2W1, Canada; 3The Arboretum, University of Guelph, 50 Stone Road East, Ontario N1G 2W1, Canada; 4Department of Botany, University of British Columbia, 3200-6270 University Boulevard, Vancouver, British Columbia V6T 1Z4, Canada; 5Office of the Vice-Principal Academic and Dean, University of Toronto, 3359 Mississaug a Road, Mississauga, Ontario L5L 1C6, Canada; 6Yukon Conservation Data Centre (CDC), Whitehorse, Yukon Territory Y1A 2C6, Canada; 7Green Plant Herbarium (TRT), Department of Natural History, Royal Ontario Museum (ROM), 100 Queens Park, Toronto, Ontario M5S2C6, Canada; 8Department of Ecology and Evolutionary Biology, University of Toronto, 27 King’s College Circle, Toronto, Ontario M5S, Canada; 9Beatty Centre for Species Discovery and National Herbarium of Canada (CAN), Botany Section, Research and Collections, National Heritage Campus of the Canadian Museum of Nature, Gatineau, Québec J9J 3N7, Canada;10The Agriculture and Agri-Food Canada Collection of Vascular Plants (DAO), 960 Carling Avenue, Ottawa, Ontario K1A 0C6, Canada; 11BIO Herbarium (OAC), University of Guelph, 50 Stone Road East, Guelph, Ontario N1G2W1, Canada; 12Natural History Museum, Cromwell Road, Kensington, London SW75BD, United Kingdom; 13Institut de recherché en biologie végétale, Université de Montréal, 2900 Edouard Montpetit Boulevard, Montréal, Québec H3T 1J4, Canada; and 14University of Manitoba Vascular Plant Herbarium (WIN), Department of Biological Sciences, University of Manitoba, 66 Chancellors Circle, Winnipeg, Manitoba R3T 2N2, Canada

- **Premise of the study:** Constructing complete, accurate plant DNA barcode reference libraries can be logistically challenging for large-scale floras. Here we demonstrate the promise and challenges of using herbarium collections for building a DNA barcode reference library for the vascular plant flora of Canada.
- **Methods:** Our study examined 20,816 specimens representing 5076 of 5190 vascular plant species in Canada (98%). For 98% of the specimens, at least one of the DNA barcode regions was recovered from the plastid loci *rbcL* and *matK* and from the nuclear ITS2 region. We used beta regression to quantify the effects of age, type of preservation, and taxonomic affiliation (family) on DNA sequence recovery.
- **Results:** Specimen age and method of preservation had significant effects on sequence recovery for all markers, but influenced some families more (e.g., Boraginaceae) than others (e.g., Asteraceae).
- **Discussion:** Our DNA barcode library represents an unparalleled resource for metagenomic and ecological genetic research working on temperate and arctic biomes. An observed decline in sequence recovery with specimen age may be associated with poor primer matches, intragenomic variation (for ITS2), or inhibitory secondary compounds in some taxa.

**Key words:** Canadian vascular flora; DNA barcode reference library; herbarium; ITS2; *matK; rbcL*.

There is a need for high-quality DNA barcode reference libraries to facilitate the routine identification of plants and to support rapidly emerging metagenomic studies in the regulation of plant-based foods and food supplements (Ivanova et al., 2016; Prosser and Hebert, 2017), ecological forensics (Kartzhinel et al., 2015; Richardson et al., 2015; Erickson et al., 2017), environmental DNA detection (Kraaijeveld et al., 2015; Scrivat et al., 2015; Bell et al., 2017), and ancient DNA analysis (Birks and Birks, 2016). The generation of reliable DNA-based identifications requires a comprehensive, accurate reference DNA barcode library based on associated voucher specimens (Hebert et al., 2017; Prosser and Hebert, 2017).

1 Manuscript received 31 July 2017; revision accepted 26 October 2017. Funding for this study was provided by the Ontario Ministry of Research and Innovation and by the government of Canada through Genome Canada and Ontario Genomics. This is a contribution to the Food from Thought program, which is supported by the Canada First Research Excellence Fund. The authors thank A. Shipunov for contributing his data from plants collected in North Dakota as well as D. Fabijan, M. Fatahi, G. Mitrow, E. Punter, P. Sokoloff, and A. Ward for aiding their work in the herbaria. The authors also thank all herbaria that contributed specimens for this analysis.

15 Author for correspondence: kuzminam@uoguelph.ca

doi:10.3732/apps.1700079


This is an open access article distributed under the terms of the Creative Commons Attribution License (CC-BY-NC-SA 4.0), which permits unrestricted noncommercial use and redistribution provided that the original author and source are credited and the new work is distributed under the same license as the original.
et al., 2003). A well-curated local reference DNA barcode library can increase the precision and accuracy of the species assignment for a query sequence (Landi et al., 2014). In some cases, results can be improved when the reference library reflects broad geographical sampling (Bergsten et al., 2012). A unique challenge for plants is that plastid and nuclear DNA barcodes yield lower species resolution compared to the mitochondrial and nuclear barcodes used for animals and fungi (Naciri and Linder, 2015; Hollingsworth et al., 2016). However, building a geographically circumscribed reference library can improve the effectiveness of plant DNA barcoding (Clerc-Blain et al., 2010; Burgess et al., 2011; de Vere et al., 2012; Parmentier et al., 2013; Elliott and Davies, 2014; Erickson et al., 2014; Braukmann et al., 2017a) recently showed that the discriminatory power of the most commonly used plant DNA barcodes (rbcL, matK, and ITS2) for the Canadian vascular plant species varies depending on the method of analysis (BLAST vs. Mothur) and biogeographic region (e.g., Canadian arctic vs. woodland). Considering individual markers, the highest resolution was provided by matK (~81%), followed by ITS2 (~72%) and rbcL (~44%). All three DNA barcodes performed strongly in assigning taxa to the correct genus (91–98%).

Obtaining a geographically representative sample of plant species from a large and diverse area such as Canada is a logistical challenge, but use of the rich collections housed in herbaria, which document plant diversity in time and space, can potentially overcome this barrier. Herbarium vouchers include detailed collection data and identifications that in many cases are annotated by experts. Importantly, voucher specimens are available for re-examination. The use of herbarium collections can significantly reduce costs and project time in comparison to making fresh collections for large-scale floras. Most North American herbaria include relatively few recently collected specimens (Deng, 2015), potentially limiting sequence recovery due to DNA degradation (Staats et al., 2011). Advances in high-throughput sequencing using approaches not based on PCR may reduce this problem, but costs are currently too high to sequence large numbers of specimens (Staats et al., 2013; Coissac et al., 2016; Bakker et al., 2016; Zhang et al., 2017a). Although several studies have demonstrated the utility of material sourced from herbaria for constructing DNA barcode libraries (de Vere et al., 2012; Kuzmina et al., 2012; Saarala et al., 2013), quantitative analysis of DNA degradation in old herbarium specimens has only been performed using a limited number of specimens and taxa (Staats et al., 2011).

Until recently, the Canadian flora lacked a standardized and comprehensive checklist across both its Nearctic endemics and taxa with Holarctic distributions (Takahajan, 1986; Thorne, 1993). The latter often possess conflicting taxonomic assignments in North American and Eurasian treatments (Flora of North America Editorial Committee, 1993; Cody, 2000; Aiken et al., 2007; Elven et al., 2011; Klinkenberg, 2013). A further complication is that approximately 22% of the modern Canadian flora comprises species introduced by human activity (Vitousek et al., 2007; Elven et al., 2011; Klinkenberg, 2013). The Database of Vascular Plants of Canada (VASCAN) was developed to standardize names for all vascular plant taxa (species, subspecies, and varieties) recorded in Canada, providing an up-to-date checklist of accepted names, synonyms, and distribution status (Brouillet et al., 2010; Desmet and Brouillet, 2013). The checklist is continuously updated based on new findings and recent taxonomic treatments. In addition to using this checklist as a framework for the species included here, we validated taxonomic information associated with all barcoded specimens in collaboration with plant taxonomists.

Here we demonstrate how herbarium material can be exploited as a large-scale resource for creating a reference DNA barcode library for a major vascular plant flora. We supplemented our primarily herbarium-derived DNAs with vouchered field- or garden-collected specimens preserved in silica gel. We assembled a reference library for three plant DNA barcodes: two based on the plastid genes rbcL and matK (CBOL Working Group, 2009), and a third, ITS2, comprising one of the two nuclear internal transcribed ribosomal spacer regions (China Plant BOL Group, 2011). We examined how sequence recovery for these markers depends on specimen age, method of preservation (i.e., herbarium specimens vs. tissue preserved in silica gel), and taxonomic affiliation (i.e., family). These parameters may represent important constraints in employing herbarium specimens in DNA barcoding studies or reference library development.

MATERIALS AND METHODS

We examined 20,816 specimens from Canada and adjacent regions of the United States (Fig. 1A). This included 13,170 specimens selected from 27 herbaria with priority to the most recently collected representatives for each species (Table 1). The remaining 7660 specimens represent freshly collected material obtained from field trips or botanical gardens in 2006–2013, which were immediately preserved in silica gel. Voucher specimens associated with these silica gel samples were deposited in associated herbarium (Table 1). Taxonomic assignments and geographic information were recorded during field collection or were obtained from herbarium voucher labels. To ensure that the record from each specimen was traceable on Barcode of Life Data Systems (BOLD), it was associated with the nomenclatural combination provided on the herbarium label (International Society for Biological and Environmental Repositories, 2012). This information, along with an image where possible, was uploaded to BOLD (dx.doi.org/10.5883/DS-VASCAN). We redacted geographic data from 252 records representing plant species assessed as endangered or threatened by the Committee on the Status of Endangered Wildlife in Canada (COSEWIC), following recommendations made by NatureServe Canada (Amie Enns and Patrick Henry, NatureServe Canada, personal communication).

We used VASCAN (Brouillet et al., 2010; Desmet and Brouillet, 2013) to provide standardized nomenclature, represented as a supplementary field in BOLD (“associated taxonomy”). The complete checklist of 5190 accepted species names of vascular plants reported from Canada includes species categorized as “native” (i.e., that are present as a result of natural processes), “introduced” (taxa established or naturalized as a result of human activity), and “ephemeral” (not established permanently, but recurring in the wild on a near-annual basis). Species of known hybrid origin, defined in literature as nothospecies (McNeill et al., 2012), were not included in the final checklist. The vouchers that were analyzed were associated with 4974 species on the VASCAN checklist. Additionally, our library includes 101 native and alien (including cultivated) species collected from Canada and the adjacent United States not listed in VASCAN (Appendix S1). The latter were included in our database following the nomenclature accepted in the Flora of North America for the relevant taxa (Flora of North America Editorial Committee, 1993), or otherwise following accepted names in The Plant List (2013). As a result, we analyzed representatives of 5076 species of vascular plants belonging to 146 of the 416 families and 43 of the 64 orders of angiosperms (Chase et al., 2016), with an additional 23 families and 13 orders of nonflowering land plants also represented (see Smith et al., 2006 for fern classification).

DNA extraction, PCR, and sequencing were performed with semiautomated protocols at the Canadian Centre for DNA Barcoding (Ivanova and Grainger, 2006; Ivanova et al., 2008, 2011; Kuzmina and Ivanova, 2011; Fazekas et al., 2012). In brief, 1–5 mg of herbarium or silica gel-dried plant tissue was ground into fine powder using a TissueLyser II (QIAGEN, Germantown, Maryland, USA) at 28 Hz for 60–90 s at room temperature, using the Axygen Mini Tube System (QIAGEN, Union City, California, USA) with one 3.17-mm stainless steel bead per tube. Following disruption, cells were lysed with 250–400 µL of 2% cetyltrimethylammonium bromide (CTAB) buffer incubated in 65°C for 60–90 min. After incubation, 50 µL of lysate from each sample was transferred into 96-well microplates (250 µL, semiskirted; Eppendorf, Hamburg, Germany) using a Liquidator 96 (200 µL; Mettler Toledo, Mississauga, Ontario, Canada). DNA was isolated and purified through binding to glass fiber filtration

http://www.bioone.org/loi/apps

Kuzmina et al.—DNA barcoding Canadian plants

Applications in Plant Sciences 2017 5(12): 1700079

doi:10.3732/apps.1700079

2 of 11

Downloaded From: https://bioone.org/journals/Applications-in-Plant-Sciences on 29 Jun 2021
Terms of Use: https://bioone.org/terms-of-use
columns (Ivanova et al., 2008) on a Biomek FX Workstation (Beckman Coulter, Mississauga, Ontario, Canada). This protocol generated 40-µL long-term storage extracts with DNA concentrations of 20–40 ng/µL, sufficient for PCR amplification of multiple DNA target regions (rbcL, matK, and ITS2) (Kuzmina and Ivanova, 2011; Fazekas et al., 2012). Details on the primers used for PCR and sequencing, as well as PCR conditions, are provided in Appendix S2. Initial PCR was performed with Phusion High-Fidelity DNA polymerase (Fisher Scientific, Hampton, New Hampshire, USA) using primers matK-xf and matK-MALP, and sequencing was done with the internal forward primer matK-1RKIM-f and matK-MALP. Owing to very low PCR success for species of Boraginaceae, a subset of specimens in this family was subjected to an additional round of DNA purification (Ivanova et al., 2008) with the goal of removing secondary compounds that might inhibit PCR. Amplification of rbcL was attempted on all 20,816 specimens, matK on 9412 specimens, and ITS2 on 13,233 specimens (Appendix S3) (differences in sample size reflect fluctuations in funding linked to specific projects, but we made substantial effort to represent all species with at least one sample per marker, where feasible). A subset of 2439 specimens was tested with all three markers. PCR products were diluted in water (Appendix S2) and sequenced on an ABI 3730xl DNA Analyzer (Applied Biosystems, Foster City, California, USA) following standard procedures. Chromatograms were edited with CodonCode Aligner versions 3.7.1–6.0.2 (CodonCode Co., Centerville, Massachusetts, USA). Sequence alignments generated in MUSCLE (Edgar, 2004) were used as a basis for removing primer sequences and to aid the identification of errors in sequence editing (e.g., nucleotide shifts in homopolymer tracts) within rbcL and matK in the partial sequences for the 5.8S and 26S ribosomal subunit genes that flank ITS2. BLASTN searches (Altschul et al., 1990) were used to identify and remove fungal, algal, and liverwort sequences that were occasionally recovered when the DNA of the target specimen was degraded and/or the match of primers to template DNA was poor. Both rbcL and matK sequences were aligned using the back-translation sequence alignment program, transAlign, to identify sequences with frameshift mutations (Bininda-Emonds, 2005). After filtering out contaminants and correcting editing errors, the sequences were uploaded to BOLD (www.boldsystems.org).

Additional quality control of the DNA barcode data was accomplished by inspecting the correspondence between phylogenies reconstructed with DNA barcodes (rbcL and matK) and the Angiosperm Phylogeny Group (APG) topologies (Kress et al., 2009; Chase et al., 2016). A preliminary neighbor-joining tree for each of these markers was constructed in BOLD (Ratnasingham and Hebert, 2007) to aid the identification of contaminants and errors in sampling, identification, and/or data entry. Cases of potential errors in identification led to reexamination of the specimens, typically in consultation with the curator of the source specimen, before taxonomic information on BOLD was updated. As a noncoding region, ITS2 required a different approach to remove contaminants and paralogous copies prior to data submission; MAFFT was used to generate an initial sequence alignment for all ITS2 sequences (Katoh et al., 2002). A maximum likelihood phylogenetic tree for the entire ITS2 data set was constructed in SATé (Liu et al., 2009) using default parameters (aligner: MAFFT; merger: MUSCLE; tree estimator: FASTTREE; model: GTR+G20) to identify and remove erroneous sequences, as described above for matK and rbcL. Our sequence length thresholds for accepting DNA barcodes were 450 bp for rbcL, 500 bp for matK, and 180 bp for ITS2.

The success in recovery of each barcode region was calculated separately (Fig. 1). For this comparison, the effects of uneven sampling were minimized by focusing on data from 25 species-rich families (23 angiosperms, one gymnosperm, one fern) with the most complete sample size for each marker (Appendix S4). The three markers (rbcL, matK, and ITS2) for these families are represented by corresponding data sets from 15,173, 7416, and 9404 specimens, respectively. The data sets analyzed for each marker have minor differences in taxonomic composition. The matK data set omits Lamiaceae (small sample size), Pinaceae, and Dryopteridaceae (no amplification), while the ITS2 data set omits Dryopteridaceae (no amplification).

To examine the effects of age and family (taxonomy affiliation) on sequence recovery from herbarium specimens, we used a beta regression analysis for modeling proportions, implemented in R version 3.2.0 (betareg package) (Ferrari and Cribari-Neto, 2004; R Development Core Team, 2008; Cribari-Neto and Zeileis, 2009). A second beta regression compared sequence recovery from specimens preserved in silica gel vs. standard herbarium material. Both tests were performed for each DNA barcode marker separately (Appendix S4). Data obtained from herbarium material were sorted by family into seven age groups (decades 1–7, Fig. 2). Sequencing success in specimens from the first (most recent) decade served as a reference to evaluate the decline in sequence recovery with time. For the second test, we compared success from silica gel-preserved material with herbarium specimens of equivalent age (1–10 yr) and performed a beta regression to avoid type I error associated with multiple pairwise tests (25 for each marker). For all tests, a high-performing reference family provided a basis for comparison with other families to identify families with low sequence recovery. (A) Specimen source location. (B) DNA barcode recovery for each marker as a percentage of specimens and species analyzed. (C) DNA barcode recovery as a percentage of samples analyzed. (D) DNA barcode recovery as a percentage of total species analyzed.
Table 1. Herbaria contributing specimens that were analyzed in this study.

<table>
<thead>
<tr>
<th>Institution</th>
<th>Index Herbariorum code</th>
<th>No. of specimens collected from tissue preserved in silica gel</th>
<th>No. of specimens collected from herbarium vouchers</th>
<th>Total no. of specimens</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acadia University</td>
<td>ACAD</td>
<td>—</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>University of Alaska Museum</td>
<td>ALA</td>
<td>—</td>
<td>46</td>
<td>46</td>
</tr>
<tr>
<td>University of Alberta</td>
<td>ALTA</td>
<td>7</td>
<td>5</td>
<td>12</td>
</tr>
<tr>
<td>B. A. Bennett Herbarium, Yukon Government</td>
<td>BABY</td>
<td>—</td>
<td>2608</td>
<td>2608</td>
</tr>
<tr>
<td>Canadian Museum of Nature</td>
<td>CAN</td>
<td>2057</td>
<td>1989</td>
<td>4046</td>
</tr>
<tr>
<td>University of Colorado, Museum of Natural History</td>
<td>COLO</td>
<td>—</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Agriculture and Agri-Food Canada</td>
<td>DAO</td>
<td>—</td>
<td>1072</td>
<td>1072</td>
</tr>
<tr>
<td>Delaware State University</td>
<td>DOV</td>
<td>—</td>
<td>210</td>
<td>210</td>
</tr>
<tr>
<td>Aberystwyth University</td>
<td>IBERS</td>
<td>—</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>University of Maine</td>
<td>MAINE</td>
<td>—</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>University of Michigan</td>
<td>MICH</td>
<td>—</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>University of Minnesota</td>
<td>MIN</td>
<td>—</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Minot State University</td>
<td>MUSU</td>
<td>281</td>
<td>—</td>
<td>281</td>
</tr>
<tr>
<td>The Manitoba Museum</td>
<td>MMMN</td>
<td>—</td>
<td>74</td>
<td>74</td>
</tr>
<tr>
<td>Université de Montréal, Herbart Marie-Victorin</td>
<td>MT</td>
<td>458</td>
<td>73</td>
<td>531</td>
</tr>
<tr>
<td>McGill University, Macdonald Campus</td>
<td>MMTG</td>
<td>149</td>
<td>—</td>
<td>149</td>
</tr>
<tr>
<td>University of Guelph</td>
<td>OAC</td>
<td>3029</td>
<td>2508</td>
<td>5537</td>
</tr>
<tr>
<td>Ontario Natural Heritage Information Centre</td>
<td>ONHIC</td>
<td>—</td>
<td>24</td>
<td>24</td>
</tr>
<tr>
<td>Université Laval, Herbart Louis-Marie</td>
<td>QFA</td>
<td>—</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>Royal Ontario Museum, Green Plant Herbarium</td>
<td>TRT</td>
<td>192</td>
<td>2641</td>
<td>2833</td>
</tr>
<tr>
<td>University of Toronto at Mississauga</td>
<td>TRITE</td>
<td>—</td>
<td>14</td>
<td>14</td>
</tr>
<tr>
<td>University of British Columbia</td>
<td>UBC</td>
<td>1436</td>
<td>1482</td>
<td>2918</td>
</tr>
<tr>
<td>University of California, Riverside</td>
<td>UCR</td>
<td>—</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>University of New Brunswick</td>
<td>UNB</td>
<td>—</td>
<td>11</td>
<td>11</td>
</tr>
<tr>
<td>University of Northern British Columbia</td>
<td>UNBC</td>
<td>33</td>
<td>—</td>
<td>33</td>
</tr>
<tr>
<td>Smithsonian Institution, United States National Herbarium</td>
<td>US</td>
<td>18</td>
<td>—</td>
<td>18</td>
</tr>
<tr>
<td>University of Waterloo</td>
<td>WAT</td>
<td>—</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>University of Manitoba</td>
<td>WIN</td>
<td>—</td>
<td>316</td>
<td>316</td>
</tr>
<tr>
<td>University of Wisconsin—Madison</td>
<td>WIS</td>
<td>—</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>University of Washington</td>
<td>WTU</td>
<td>—</td>
<td>50</td>
<td>50</td>
</tr>
</tbody>
</table>

*Acronyms for the herbaria are used in accordance with Index Herbariorum (Thiers, 2017).

recovery. The reference families were selected based on consistently high sequence recovery across all age groups and a large sample size (Fabaceae for rbcL and matK, Brassicaceae for ITS2). Families identified as having low or problematic sequence recovery were analyzed for possible primer–template incompatibility by comparing primer sequences to corresponding full-length sequences available on GenBank for each taxon.

RESULTS

Specimen, species, and family coverage—The sequencing success for each marker was evaluated based on the number of specimens that successfully generated a sequence for a particular marker (Fig. 1B, 1C, and 1D; Appendix S3). The values of sequence recovery for rbcL, matK, and ITS2 for all 169 families of vascular plants in Canada are reported in Appendix S5, together with the sample size (number of specimens attempted) for each family.

To identify trends in sequencing success, we grouped families with more than nine specimens into four categories: those with no sequence recovery (0%), low recovery (1–50%), moderate recovery (51–75%), and high recovery (76–100%) (Table 2, Appendix S5). The proportion of families distributed across the four groups for rbcL was 0.00/0.03/0.20/0.77. For matK and ITS2, these proportions were 0.10/0.88/0.60 and 0.11/0.21/0.33/0.35, respectively. No family completely failed to generate sequences for rbcL, and low recovery was restricted to two nonangiosperm families (Ophioglossaceae, Selaginellaceae) and three angiosperm families (Boraginaceae, Cistaceae, Pontederiaceae). A complete failure to recover matK was observed in all eight nonangiosperm families (Appendix S5), and seven angiosperm families (Boraginaceae, Crassulaceae, Drosaceae, Geraniaceae, Haloragaceae, Hypericaceae, Juncaceae) showed low recovery. ITS2 failed for one angiosperm family (Pontederiaceae) and for 11 of 16 nonangiosperm families. The families with low ITS2 recovery were mainly monocots (7), rosids (6), and asterids (6). Among poorly performing families, five were characterized by extensive sampling (>100 specimens): Boraginaceae, Caprifoliaceae, Juncaceae, Lamiaceae, and Polygonaceae (see Appendix S5 for details).

Recovery from herbarium specimens over time—Beta regression analysis demonstrated that sequence recovery for rbcL and matK was significantly lower for herbarium specimens ranging from 10–30 yr of age than for those collected in the past decade (Fig. 2, Appendix S4). Success was further reduced among older specimens. By contrast, a significant decline in sequence recovery for ITS2 was only noted in specimens older than 50 yr, although the average success for this marker at this age was comparable to that seen for the plastid markers.

Beta regression identified four groupings with similar patterns in sequence recovery over time among the 25 families analyzed with this method. These groupings were clearest for the rbcL data set because of its comprehensive sampling across families and age groups (Fig. 3). The first group of families (including Apiaceae, Asteraceae, Brassicaceae, Poaceae) had consistently high sequence recovery for all age groups with no difference from the reference family (Fabaceae) (P > 0.05; Fig. 3A). The remaining three groups showed declining sequence...
recovery with age, but did so differently. The second group of families (Cyperaceae, Juncaceae, Plantaginaceae, Ranunculaceae) had high sequence recovery for herbarium material less than 50 yr old, but a noticeable decline in older specimens (0.01 < P < 0.05; Fig. 3B). The third group (Dryopteridaceae, Onagraceae, Polygonaceae, Saxifragaceae) delivered high sequencing success for material less than 10 yr old followed by a gradual decline with each decade (0.001 < P < 0.01; Fig. 3C). The final group showed a rapid decline in sequence recovery with time (P < 0.001; Fig. 3D). In particular, three families (Ericaceae, Pinaceae, Rosaceae) had initially high success (decade 1: 88–100%) that declined rapidly with age, while two families (Boraginaceae, Orchidaceae) had poor sequence recovery even for the youngest material (<10 yr) that declined further in older specimens (<40%).

The results of the beta regression analyses for matK and ITS2 sequencing success were consistent with those for rbcL (Fig. 2).

For example, sequencing success was high for all specimens and markers for four families (Apiaceae, Asteraceae, Caryophyllaceae, Salicaceae). Five families had high recovery for all material for at least two markers (Amaranthaceae, Betulaceae, Caprifoliaceae, Orobanchaceae, Saxifragaceae). Conversely, six families (Boraginaceae, Cyperaceae, Juncaceae, Orchidaceae, Plantaginaceae, Ranunculaceae) showed a significant decline in recovery with age for all three barcode regions. Among those,
Boraginaceae and Orchidaceae showed strikingly low average sequencing success for \textit{rbcL} (Fig. 4).

**Recovery from silica gel vs. herbarium material**—Beta regression analysis demonstrated that the three markers responded differently to preservation method with respect to sequence recovery (Fig. 2, Appendix S4). For ITS2, seven of 23 families showed recovery similar to the reference family (Brassicaceae), while 11 families had lower recovery and five had higher recovery for material preserved in silica gel compared to herbarium. For \textit{rbcL} and \textit{matK}, preservation method had no effect on sequence recovery for 16 families, while the other eight families showed a significant difference in sequence recovery but with no consistent pattern.

**Additional DNA purification of the specimens belonging to Boraginaceae**—All specimens of Boraginaceae, regardless of age or preservation method, delivered low average sequence recovery (\textit{rbcL} 31\%, \textit{matK} 46\%, ITS2 17\%), and a secondary DNA purification step (Ivanova et al., 2008) failed to improve recovery. The potential for mismatch of the standard \textit{rbcL} primers with this family was checked using GenBank reference \textit{rbcL} sequences for \textit{Myosotis} L. (KJ841424) and \textit{Hydrophyllum} L. (HQ590137), but neither sequence had mismatches.

**Identification with DNA barcodes and new alien species for Canada**—Sequence data from the three markers led to the re-identification of 192 herbarium specimens (dx.doi.org/10.5883/DS-VASCAN; the process IDs below indicate the records in this database). These cases of misidentification often involved morphologically similar species (e.g., \textit{Ribes americanum} Mill. vs. \textit{R. nigrum} L.) or species that require microscopic examination to evaluate characters (e.g., members of Chenopodiaceae). Some of these updates resulted in corrections to distributional data. For example, \textit{R. laxiflorum} Pursh was reported from Yukon based on a single record (BBYUK2558-16), but DNA barcode analysis led to the reassignment of this specimen to \textit{R. glandulosum} Grauer and excluded \textit{R. laxiflorum} from the flora of Yukon. Conversely, DNA barcodes confirmed the presence of \textit{Ranunculus occidentalis} Nutt. (BBYUK1525-12) in the Northwest Territories, which was previously reported only from Yukon, British Columbia, and Alberta.

Our DNA barcode database includes representatives of 101 species that were collected in Canada and the adjacent United States, but not listed in VASCAN (Appendix S1; dx.doi.org/10.5883/DS-VASCAN). These species, some with medicinal properties, are often cultivated in botanical gardens (e.g., \textit{Trigonella foenum-graecum} L., \textit{Atropa belladonna} L., \textit{Crataegus viridis} L.) or are widely cultivated as ornamentals (e.g., \textit{Ginkgo biloba} L., \textit{Wisteria sinensis} (Sims) DC.). Some were previously reported from the United States (Cullina et al., 2011; Glenn, 2013) as native (\textit{Coreopsis pubescens} Elliott: KSR336-07) or as being introduced (\textit{Euphrasia micrantha} Rchb.: BBYUK2400-13, VASCAS92-15; \textit{Verbascum phoeniceum} L.: PLWE109-10). Other species have been reported from Canada (Zika, 2013) but are not included in VASCAN yet (\textit{Juncus hesperius} (Piper) Lint: VPSBC1118-13;
J. laccatus Zika: VPSBC1121-13). Finally, our database provides coverage for 42 native North American species recorded from adjacent regions in the United States (e.g., *Claytonia acutifolia* Pall. ex Willd.: BBYUK2574-16; *Papaver alaskanum* Hultén: BBYUK1286-12, BBYUK1287-12) that may occur in Canada.

**DISCUSSION**

Factors contributing to sequence recovery—The comprehensive nature of our DNA barcode library for the Canadian flora made it possible to examine factors influencing sequence recovery at a large scale. Two-thirds of the material we sampled was derived from herbarium specimens. Given our extensive use of older material (84% >10 yr), we expected that DNA degradation would hinder sequence recovery (de Vere et al., 2012). Our results confirmed that sequence recovery was problematic for the longer plastid barcodes *rbcL* (552 bp) and *matK* (~800 bp) from material older than 10 yr (Fig. 2, Appendix S4). This effect was less noticeable for the shorter nuclear marker ITS2 (~350 bp), as its sequence recovery significantly declined only in material older than 50 yr. This difference in recovery rate between plastid and nuclear markers is likely caused by the different lengths of the target regions, rather than any inherent difference between organellar compartments (Lister et al., 2008), as all three markers are expected to be present in high copy number in each cell.

Species-rich families with consistently high sequence recovery for all material (e.g., Asteraceae; Fig. 2, Test 1; Appendix S4) represent three major phylogenetic clades (rosids, caryophyllids, and asterids) that have undergone recent, explosive radiations (e.g., Liu et al., 2006; Valente et al., 2010; Zhang et al., 2017b). These groups were the main consideration during design of the standard primer sets because they represent so much plant diversity at the species level. The strong affiliation of the primers to conserved sites flanking the DNA barcode regions in these taxonomic groups allowed consistent recovery, even from older degraded material. By contrast, another group of families (including Cyperaceae, Poaceae, Pinaceae) demonstrated clear evidence of a decline in sequence recovery with specimen age. Using sequence data available in GenBank, we found that many had primer mismatches for at least one of the DNA barcode regions. We hypothesize that the long history of Pinaceae, which has a crown age of 198 Ma (Lu et al., 2014), has allowed the accumulation of substitutions in the primer-binding regions. Although Cyperaceae and Poaceae form much younger crown groups (82 Ma and 65 Ma, respectively; Bouchenak-Khelladi et al., 2014), their elevated rates of substitutions (e.g., Hilu and Liang, 1997) likely created the same effect. Primer mismatches for some groups of monocots may similarly reflect the well-documented acceleration of plastome evolution in the order Poales (grasses, sedges, rushes, and relatives) vs. most other monocots and eudicots (e.g., Saarela and Graham, 2010). Only two cases of primer mismatch were recorded for *rbcL* (Ophioglossaceae, Selaginellaceae) (Appendix S5). Such cases were more frequent for *matK* in three species-rich families (Juncaceae, Plantaginaceae, Ranunculaceae) and in some smaller families (e.g., Crassulaceae, Geraniaceae, Haloragaceae, Hypericaceae). Although partial primer incompatibility may not prevent amplification, it undoubtedly contributes to the reduced sequencing success in older herbarium material owing to DNA degradation and the resulting substantial decline in copy number of nondegraded molecules (Staats et al., 2011, 2013). The lack of universal priming sites as a cause of reduced sequence recovery is well-recognized for *matK* (Dunning and Savolainen, 2010). This issue is particularly challenging for most ferns, in which the plastid genome has lost the conserved *trnK* exons that are usually used as flanking conservative sites for amplification of the entire *matK* gene (Kuo et al., 2011).

The priming sites for ITS2 were previously proposed as mostly “universal” (Chen et al., 2010; China Plant BOL Group, 2011), but our results suggest this is not the case. Primer sites in

![Fig. 4. Average sequence recovery of each DNA barcode from herbarium specimens of six families with the lowest success.](https://bioone.org/journals/Applications-in-Plant-Sciences/2017/5/12/1700079)
template sequences are not completely conserved within certain species-rich (Cyperaceae, Juncaceae, Poaceae) and other less diverse families (Araceae, Commelinaceae, Ginkgoaceae, Isoetae- ceae, Orchidaceae, Pinaceae, Pontederiaceae, Typhaceae) (Appendix S6). The limited availability of sequence records for ITS2 for ferns makes it difficult to ascertain if amplification of this DNA region fails for the same reason. We recovered ITS2 data for eight genera (Asplenium L., Azolla Lam., Cyathea Sm., Ceratopteris Brong., Dryopteris Adans., Equisetum L., Lygodium Sw., and Psilotum Sw.) of monophytes (ferns sensu lato) in GenBank. Other ITS2 sequences on GenBank, annotated as being from ferns, actually originate from fungal or angiosperm contaminants based on BLAST.

ITS2 can also possess divergent copies within the nuclear genome, which can lead to the amplification of paralogous sequences from multiple templates (e.g., Xu et al., 2017). Song et al. (2012) suggested that intragenomic variation within this DNA region occurs more frequently than previously reported (Chen et al., 2010; China Plant BOL Group, 2011). Our data showed the recovery of ITS2 sequences was <75% in almost half of the species-rich families (e.g., Apiaceae, Caprifoliaceae, Orobanchaceae, Poaceae, Polygonaceae). In most instances, PCR products were obtained, but chromatograms acquired with Sanger sequencing were not interpretable. This result is most easily explained by the presence of multiple paralogous copies of ITS2. Whether as a result of gene duplication, hybridization, or polyploidy, the presence of multiple variants in high proportions negatively affects the successful recovery and application of ITS2 as a DNA barcoding marker in a substantial fraction of taxonomic groups (e.g., Zarrei et al., 2015).

Rapid DNA degradation after sample collection is likely a significant cause of amplification failure in some groups. We observed very low success in the recovery of all three DNA barcode regions, regardless of the method of preservation, in two families (Boraginaceae and Orchidaceae) in which we had little reason to suspect primer incompatibility (Fig. 3, Appendix S4). We hypothesize that DNA degradation occurs soon after collection in these families owing to the presence of compounds that degrade DNA or irreversibly bind to it. For example, most genera of Boraginaceae synthesize and store pyrrolizidine alkaloids, compounds that cause rapid and permanent DNA damage (El-Shazly and Wink, 2014). We suggest this type of metabolite may prevent DNA preservation during the early phases of drying, immediately following specimen collection. The polyphenols found in some Orchidaceae species are also well known for their DNA-binding activity, particularly in the presence of polyphenol oxidase, which is liberated during plant tissue damage (Ho, 1999; Mazo et al., 2012). Irreversible DNA damage in specimens from these families may only be prevented by neutralizing enzymatic activity at the earliest stages of plant tissue preservation. To achieve consistent sequence recovery, immediate DNA extraction with modified DNA extraction methods (e.g., Ivanova et al., 2011) after specimen collection may be essential.

Different stages of degradation of the plastid genome in holoparasitic or mycoheterotrophic plants can lead to the pseudo-digenization or complete loss of plastid-encoded loci (Graham et al., 2017) including rbcL (e.g., Corallorhiza Gagnepin [Barrett and Davis, 2012], mycoheterotrophic Ericaceae [Braukmann et al., 2017b], holoparasitic Orobanchaceae [Wicke et al., 2013], Cuscuta L. [McNeal et al., 2007; Braukmann et al., 2013]) and matK (e.g., Cuscuta [McNeal et al., 2007; Braukmann et al., 2013], Monotropis Schwein. ex Elliott [Braukmann et al., 2017b]). Therefore, the amplification of plastid loci (especially rbcL) failed for these plants, or in some cases led to the recovery of pseudogenes. Barcoding efforts should focus on the nuclear-encoded loci (i.e., ITS2) given plastome degradation in these plants, or on alternate plastid loci that are commonly retained in nonphotosynthetic plants, such as accD (Lam et al., 2016).

Contrary to initial expectations, we encountered cases where sequence recovery from herbarium material was better than that from material preserved in silica gel. Most families with low recovery from samples stored in silica gel also have low rates of recovery from older herbarium material. It is likely that PCR amplification in these families was strongly affected by extrinsic factors (poor primer matches or presence of DNA-degrading metabolites) that are exacerbated over time, causing a substantial reduction in template copy number for the targeted DNA regions (Staats et al., 2011). The better performance of herbarium specimens vs. material preserved in silica gel may also reflect deviation from proper handling of samples prior to their storage in silica gel. Any delay between collection and storage on silica increases the opportunity for metabolites to damage DNA before the sample is fully dried. Long-term storage of specimens in silica gel requires additional control of humidity and isolation from light. Improper storage conditions may result in greater DNA damage than for herbarium specimens preserved in constantly dark, dry conditions.

The effects of DNA degradation seem to be exacerbated in families where there is mismatch between the standard primers and the priming regions, potential intragenomic variation (paralogy in ITS), and/or presence of certain metabolites that affect DNA integrity. All these factors may affect the quality and completeness of a DNA barcode reference library and thus have a direct influence on the interpretation of results from any applications using it. Further customization of protocols to accommodate different primer sets for phylogenetically diverse clades and to neutralize secondary metabolites will improve overall accuracy and completeness of the sequence data. Our results demonstrate that herbarium specimens are suitable for most plant families as the main source of material and that several large families (e.g., Asteraceae, Fabaceae, Brassicaceae) can be successfully sampled from older herbarium material without significantly affecting the quality of the results.

**Taxonomic representation of the Canadian flora in the DNA barcode library**—The use of a substantial fraction of herbarium specimens in our study ensured nearly complete representation of the vascular plant species in the flora of Canada. In addition to the names used in the standard checklist, our database retained all primary taxonomic annotations that accompanied the herbarium vouchers, which are mostly also available as digitized images. This information remains critical for applying the most recent and accurate taxonomic updates to a herbarium voucher and the associated record in BOLD (dx.doi.org/10.5883/DS-VASCAN).

The existing VASCAN checklist is subject to periodic revision based on new information from local checklists, nomenclatural changes, species discovery, and inventories of alien species. For example, the British Columbia Conservation Data Centre (2016) reported about 100 taxonomic name changes, 39 taxonomic rank changes, and 53 new taxon records for the vascular plants for British Columbia in 2016. Although the discovery of new vascular plant species in Canada is less common, one was recently described from the Yukon (Draba bruce-bennettii...
Al-Shehbaz, Al-Shehbaz, 2016). An inventory of nonnative species is particularly challenging because their distribution constantly changes (Cox, 1999; Davis et al., 2011). However, DNA barcode libraries must include these species because of their utility in forensic cases (Ivanova et al., 2016), ecological surveys (Bell et al., 2017), and in the detection of invasive species in local communities (van de Wiel et al., 2009; Gahramanazadeh et al., 2013). The inclusion of both native and alien species occurring in the adjacent United States complements the generic diversity of these taxa in our DNA barcode library with their closest counterparts outside Canada and provides more complete information with respect to their phylogeographic status. Future additions to the DNA barcode library of the vascular plants of Canada should focus on the inclusion of such species.

In addition to the advantages for obtaining DNA barcodes (easy accessibility, completeness of documentation, reduced costs and time), herbarium vouchers also provide an opportunity to cross-check voucher annotation with a molecular data set, which improves the quality of both collections. The robust reference library presented here has facilitated the improvement of local floristic checklists and the tracking of alien or invasive species. It also contributes to updates on the distribution of species. As a publicly available, actively managed database, the Vascular Plants of Canada library in BOLD is a comprehensive and effective system that facilitates plant diversity information sharing and creates an unparalleled genetic resource for the study of temperate and arctic biomes.

DATA ACCESSIBILITY

BOLD (dx.doi.org/10.5883/DS-VASCAN); GenBank (accession numbers associated with the process identification numbers in BOLD).

LITERATURE CITED


Applications in Plant Sciences 2017 5(12): 1700079 Kuzmina et al.—DNA barcoding Canadian plants

doi:10.3732/apps.1700079

Applications in Plant Sciences 2017 5(12): 1700079 Kuzmina et al.—DNA barcoding Canadian plants

Applications in Plant Sciences 2017 5(12): 1700079 Kuzmina et al.—DNA barcoding Canadian plants


