Methods in Ecology and Evolution

British Ecological Society

Methods in Ecology and Evolution 2011, **2**, 333–340 doi: 10.1111/j.2041-210X.2011.00092.x

Discriminating plant species in a local temperate flora using the *rbcL* + *matK* DNA barcode

Kevin S. Burgess[1]*, Aron J. Fazekas[2], Prasad R. Kesanakurti[2], Sean W. Graham[3], Brian C. Husband[2], Steven G. Newmaster[2], Diana M. Percy[3], Mehrdad Hajibabaei[4] and Spencer C. H. Barrett[5]

[1]*Department of Biology, Columbus State University, Columbus, GA 31907-5645, USA;* [2]*Department of Integrative Biology, University of Guelph, Guelph ON N1G 2W1, Canada;* [3]*UBC Botanical Garden & Centre for Plant Research, and Department of Botany, University of British Columbia, Vancouver, BC V6T 1Z4, Canada;* [4]*Department of Integrative Biology, Biodiversity Institute of Ontario, University of Guelph, Guelph, ON N1G 2W1, Canada and* [5]*Department of Ecology & Evolutionary Biology, University of Toronto, Toronto, ON M5S 3B2, Canada*

Summary

1. A major goal of DNA barcoding is to identify species in local floras and ecological communities. With the consensus of a two-locus DNA barcode (*rbcL* + *matK*) by the Consortium for the Barcode of Life (CBOL) Plant Working Group (2009), barcoding efforts have begun to focus on building the barcode library for land plants.

2. Here, we establish a barcoding database for a temperate flora of moderate taxonomic breadth at the Koffler Scientific Reserve, Ontario, Canada based on the *rbcL* + *matK* barcode. We evaluated the performance of this combination in comparison with three other potential supplementary regions (the coding region *rpoC1* and two non-coding intergenic spacers *trnH-psbA* and *atpF-atpH*). We examined these markers singly and in combination to evaluate their discriminatory power among 436 species in 269 genera of land plants.

3. Using high-throughput techniques, we recovered a high-quality sequence from at least one region for 98.2% of the 513 samples screened; 55% had complete coverage across all five gene regions. Sequencing success was highest for *rbcL* (91.4% of samples collected) and lowest for *rpoC1* (74.5%). The two coding regions *rbcL* and *matK* provided a relatively high number of high-quality bi-directional sequences compared with the non-coding intergenic spacers, and in combination were able to correctly identify 93.1% of the species sampled. Marginal increases in species resolution were obtained with the inclusion of the *trnH-psbA* intergenic spacer (95.3%), or by using all five gene regions combined (97.3%).

4. There was a weak relation between the number of species per genus and identification success rate using *rbcL* + *matK*; 100% for monotypic genera (70.5% of the flora) and 83.6% for polytypic genera. Identification success using the *rbcL* + *matK* barcode was higher (100%) for gymnosperms, bryophytes, lycophytes and monilophytes (collectively representing 5% of the flora), compared with angiosperms (92.7%).

5. Our results indicate that the *rbcL* + *matK* barcode can provide an acceptably high rate of species resolution in the context of this and other local northern temperate floras. It does so in a cost-effective manner, with relatively modest laboratory effort, and despite the presence of missing data from individual plastid regions in a subset of samples.

Key-words: *atpF-atpH*, barcoding, ecological applications, floristic surveys, plants, *rpoC1*, species identification, *trnH-psbA*

*Correspondence author. E-mail: burgess_kevin@colstate.edu
Correspondence site: http://www.respond2articles.com/MEE/

Introduction

DNA barcoding is an effective method for species identification and for documenting animal biodiversity in certain taxonomic groups (Hebert, Ratnasingham, & deWaard 2003; Hebert *et al.* 2003; Borisenko, Sones, & Hebert 2009; Janzen *et al.* 2009). By assessing DNA sequence variation in a short standardized region of the mitochondrial cytochrome c oxidase 1 (COI or *cox1*) gene, numerous initiatives are underway barcoding the earth's animal biota at an unprecedented rate [Barcode of Life Data Systems (http://www.boldsystems.org/views/login.php); Consortium for the Barcode of Life (CBOL: http://www.barcoding.si.edu); International Barcode of Life (iBOL: http://ibol.org/); Ratnasingham & Hebert 2007]. However, this region is ineffective for barcoding plants due to generally low nucleotide substitution rates in plant mitochondria (Chase *et al.* 2005; Cowan *et al.* 2006; Taberlet *et al.* 2006; Mower *et al.* 2007; Fazekas *et al.* 2009). Although the nuclear ribosomal internal transcribed spacer region (ITS) has been suggested as a possible plant barcode (e.g. Kress *et al.* 2005; Chen *et al.* 2010), individuals can host numerous paralogous copies (Buckler, Ippolito, & Holtsford 1997; Álvarez & Wendel 2003; Bailey *et al.* 2003; King & Roalson 2008) resulting in multiple barcode haplotypes per individual or inappropriate homology assignments. Furthermore, while the nuclear genome has been posited as a source of barcoding markers (Chase *et al.* 2005), the considerable fluidity of this genome (e.g. Kellogg & Bennetzen 2004; Chase *et al.* 2005) means that there are significant technical hurdles to identifying appropriate regions.

As a consequence, numerous multigene approaches that use combinations of variable non-coding and relatively conserved coding regions of the plastid genome have been proposed (e.g. Kress *et al.* 2005; Newmaster, Fazekas, & Ragupathy 2006; Chase *et al.* 2007; Fazekas *et al.* 2008; Kress *et al.* 2009). Recently, CBOL has provisionally adopted a two-locus barcode for land plants comprising portions of the *rbcL* and *matK* coding regions (CBOL Plant Working Group: Hollingsworth *et al.* 2009). These regions were chosen based on two main criteria: a high level of recoverability of high-quality sequences and high levels of species discrimination. The combination *rbcL* + *matK* was adopted as a core barcode, with the recognition that in some circumstances other supplementary regions will be necessary to provide the desired level of species resolution. With the adoption of this barcode, researchers have begun to investigate a variety of plant barcoding applications.

DNA barcodes allow the identification of species within a community associated by geography or ecology (Chase *et al.* 2005; Kress & Erickson 2008). However, most studies to date have involved an assessment of species resolution within defined taxonomic groups, highlighting areas of concordance and discordance with traditional, morphological-based taxonomic approaches (e.g. Sass *et al.* 2007; Newmaster & Ragupathy 2009; Seberg & Petersen 2009; Starr, Naczi, & Chouinard 2009). In addition, considerable effort has been focused on the performance of barcoding regions (and their combinations) across diverse collections of land plants (e.g., Newmaster, Fazekas, & Ragupathy 2006; Kress & Erickson 2007; Fazekas *et al.* 2008; CBOL Plant Working Group: Hollingsworth *et al.* 2009), medicinal floras (Chen *et al.* 2010) and mesoamerican and South African biodiversity hotspots (Lahaye *et al.* 2008). Species resolution for the *rbcL* + *matK* barcode, in particular, has been shown to plateau around 70% across a broad sampling of land plants (CBOL Plant Working Group: Hollingsworth *et al.* 2009; see also Fazekas *et al.* 2008).

In a recent review of the potential ecological applications of DNA barcoding, Valentini, Pompanon, & Taberlet (2008) highlighted how this approach can be used for biodiversity assessments of both contemporary and past communities. Such applications will be greatly enhanced by the development of localized barcoding libraries for determining the identity of unknown samples, using algorithms such as BLAST (Chase *et al.* 2005; Kress & Erickson 2008). For ecological barcoding applications, the plateau of ~70% species resolution found in several studies (e.g. Fazekas *et al.* 2008; CBOL Plant Working Group: Hollingsworth *et al.* 2009) may often be a conservative estimate; species resolution may be higher in a geographically restricted context because of the reduced number of closely related species occurring within a given locale. Indeed, this was demonstrated by Kress *et al.* (2009) who reported that the *rbcL* + *matK* barcode correctly identified 92% of all woody species in a 50 ha tropical forest plot in Panama. This value increased to 98% with the inclusion of the supplementary *trnH-psbA* region. Although they did not test this three-region combination, Gonzalez *et al.* (2009) reported species resolution to be much lower (< 50%) for the *rbcL* + *matK* combination in a study of tropical tree species from a 2 ha plot in French Guiana. This finding was largely due to the presence in their sample of species-rich genera with low sequence variation for the plastid genome. The expected rate of species resolution with the *rbcL* + *matK* barcode in temperate floras of similar spatial scales is not known.

Here, we establish a plant DNA barcode database for the Koffler Scientific Reserve (KSR), a 350 ha field station administered by the University of Toronto, Ontario, Canada, for use in ecological applications at the site and more generally for south-central Ontario, Canada. We also evaluate the efficacy of the *rbcL* + *matK* barcode in comparison with other proposed coding (*rpoC1*) and non-coding regions (*trnH-psbA* and *atpF-atpH*) for species discrimination in this local temperate flora. We use these markers singly and in combination to evaluate sequence recoverability, species identification success and cost-effectiveness when applied to a sample of 436 species distributed among 269 genera of land plants. Specifically, we focus our analyses on: (i) individual plastid regions; (ii) *rbcL* + *matK*; (iii) this two-locus barcode combination plus a non-coding region (*trnH-psbA*) previously identified as a potential supplementary region (CBOL Plant Working Group: Hollingsworth *et al.* 2009); and (iv) the combination of this three-locus barcode plus two other previously identified candidate regions: *rpoC1* and *atpF-atpH* (e.g. Fazekas *et al.* 2008).

Materials and methods

SAMPLING

During the summers of 2006–2008, we sampled 436 plant species comprising 513 samples from the KSR at Jokers Hill near Newmarket, southern Ontario, Canada (44° 03′ N, 79° 29′ E; Table S1). Our sampling represents ~70% of the approximately 628 taxa on the vascular plant species list created for this site (http://ksr.utoronto.ca/Plants). This list includes species for which only a single record is known and in many cases these species have not been recollected. The taxa sampled here include a number of taxonomically complex groups that present considerable challenges for routine morphological identification because of hybridization, polyploidy and/or agamospermy (see Fazekas *et al.* 2008, 2009). Sixty-two (~12%) of the samples were from a previously published barcoding study collected during the same time period (Fazekas *et al.* 2008). Seventy-five replicate samples were included in our analysis. All specimens were mounted on herbarium sheets, photographed and stored at the University of Guelph Herbarium as barcode vouchers with duplicates held at the Royal Ontario Museum, Toronto, Canada. For each specimen, we also sampled 3–5 cm^2 of leaf tissue and stored this in silica gel for subsequent DNA isolation.

DNA ISOLATION, AMPLIFICATION AND SEQUENCING

We isolated DNA, amplified specific regions and sequenced amplicons for each of 513 samples using modified protocols from the Canadian Centre for DNA Barcoding (CCDB) (http://www.ccdb.ca/pa/ge/research/protocols/). Specifically, we isolated total genomic DNA from approximately 10 mg of dried leaf material using a semi-automated, membrane-based protocol (Ivanova, Fazekas, & Hebert 2008). We then amplified DNA for three coding (*rbcL, matK* and *rpoC1*) and two non-coding plastid regions (*trnH-psbA* and *atpF-atpH*) using primers with broad taxonomic versatility (Table S2). We amplified DNA in 12.5 μL reaction mixtures containing 0.06 μL of *Taq* polymerase (5 U μL^{-1}), 1.25 μL of 10× buffer, 0.625 μL of 50 mM $MgCl_2$, 0.0625 μL of 10 mM dNTPs, 0.125 μL of 10 μM of each primer, 6.25 μL of 10% trehalose, 2 μL of ddH_2O and 2 μL template DNA. Amplification of each gene region was performed using the following protocol: initial denaturation at 95 °C for 1 min, 35 cycles of 95 °C for 30 s, annealing at 55 °C for 40 s and extension at 72 °C for 1 min, followed by a final extension at 72 °C for 10 min and final hold at 4 °C. We sequenced amplification products directly on both strands with the primers used for amplification in sequencing reactions containing 0.25 μL of BigDye terminator mix v3.1, 1.875 μL of 5× sequencing buffer (400 mM Tris-HCl pH 9.0 + 10 mM $MgCl_2$), 5 μL of 10% trehalose, 1 μL of 10 μM primer, 0.875 μL of ddH_2O and 0.5–1.2 μL of PCR product. Sequencing reactions were performed using the following conditions: initial denaturation at 96 °C for 2 min, 30 cycles of 96 °C for 30 s, annealing at 55 °C for 15 s and extension at 60 °C for 4 min followed by a final hold at 4 °C. We obtained bidirectional sequence reads from most PCR products, but in some cases either the forward or reverse sequencing reaction consistently failed, or only a partial sequence was recovered, frequently due to homopolymer runs in non-coding gene regions. For these samples, a minimum of two-fold coverage was obtained by repeating the sequencing reactions in the direction that was initially successful (see methods in Fazekas *et al.* 2008). For *matK* in particular, PCR and sequencing success were low (~60%) using the single primer set 3F_KIM, 1R_KIM (Table S2). To increase the taxonomic coverage for *matK*, we employed a second set of primers, 390F, 1326R, for use on samples that failed with the first set (Table S2). Using this primer set, we were able to obtain *matK* sequences for approximately 20% more samples.

ANALYSIS

We assembled and base-called sequences using Sequencher 4.5 (Gene Codes Corp, Ann Arbor, MI http://www.genecodes.com/). To determine percentage species resolution/gene region, we first created a local sequence database in Geneious Pro 4.8.5 (Drummond *et al.* 2009) based on the 504 samples for which we obtained a sequence for at least one of the five gene regions. To determine species assignment for each sample, we then conducted an all-to-all BLAST analysis for short nearly exact searches using BLASTN (v. 2.2.22 as a plugin in Geneious Pro 4.8.5): all sequences served as both database and query but were queried individually to the database. Query sequences having 100% identical sites to sequences in the database were counted as correct assignments compared with all other sequences in the database. This analysis was also conducted for a subset of taxa (282 samples, representing 256 species) in which we had complete sequence coverage for all five gene regions. For both analyses, if no identical match was found for a species other than itself, we considered it as having a unique sequence; the species was then scored as 100% resolved using that particular gene region. We then calculated percentage species resolution for a given gene region as the percent of species that had unique sequences for each respective gene region in the database. To assess the effect of multiple gene-region combinations on species resolution, we calculated species resolution as a cumulative percentage for each combination of gene regions. This enabled us to address the extent to which species resolution (identification) in our local context may increase with the addition of one or more gene regions.

We used contingency analyses to further evaluate the performance of the *rbcL* + *matK* barcode among major taxonomic groups (specifically, non-angiosperms vs. angiosperms) and mono- vs. polytypic-genera. Angiosperms are by far the most species-rich group of land plants and the clade as a whole originated relatively recently, which may lead to differences in ability to discriminate among close relatives compared with other land plants. In addition, a species of a monotypic genus may be expected to be more distinctive than species belonging to a polytypic genus if their morphological distinctiveness parallels their genetic divergence. A nominal logistic regression was used to evaluate the relation between the number of correctly assigned species and the number of species/genus locally as a possible source of variation for identification success within genera. All statistical analyses were performed using JMP® statistical software, version 5.0 (SAS Institute, 2002).

Results

RECOVERY OF SEQUENCES

In total, we obtained 2130 sequences from 436 plant species, representing 269 genera and 90 families (Fig. 1). We recovered one sequence for at least one of the five gene regions in 504 (98.2%) samples; only nine samples failed across all five gene regions (Tables S1 and S3). Sequence recovery was highest for *rbcL* (91.4% samples) and lowest for *rpoC1* (74.5% samples) (Fig. 1). Of the 513 samples that were initially screened, we were unable to obtain high-quality sequences for 22 samples

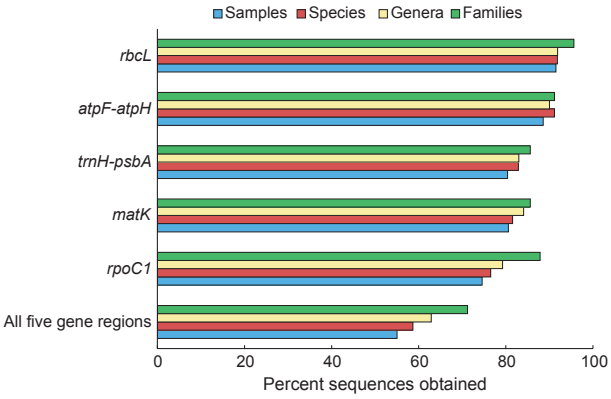

Fig. 1. Sequence recovery for five plastid regions from 436 species (513 samples) representing 269 genera and 90 families collected from a local flora in southern Ontario, Canada. In total, 2130 sequences were recovered (504 samples) but only a subset [282 samples (1410 sequences)] provided sequences for all five plastid gene regions.

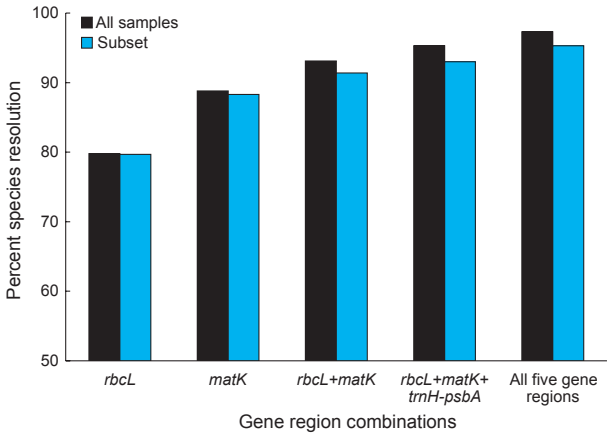

Fig. 2. Percent species resolution for single and multigene regions based on an all-to-all BLAST analysis of all 504 samples, as well as a subset of samples (282) having complete sequence coverage for all five gene regions, collected from a local flora in southern Ontario, Canada. Because of differential recovery of sequences across regions, the taxonomic complement for each gene region differed slightly among regions and combinations of the subsets.

(4.3%) for both regions of the *rbcL*+*matK* combination. Fourteen samples (2.7%) were missing all regions of the three-locus combination, *rbcL*+*matK*+*trnH-psbA*. We obtained complete coverage for all five gene regions for 55% of the samples collected, representing 58.6% of the species and 62.8% of the genera within the flora (Fig. 1; Tables S1 and S3).

SPECIES RESOLUTION: SINGLE-REGION ANALYSIS

For the 504 samples for which we obtained at least one sequence, percent species resolution (expressed relative to the number of species that amplified for each plastid region) ranged from 88.8% (*matK*) to 73.1% (*rpoC1*) with *atpF-atpH*, *rbcL* and *trnH-psbA* having intermediate values of 82.4%, 79.8% and 79.3%, respectively. Because each of the values for the respective gene regions are based on different subsamples, we reduced the data set to the 282 samples that had complete coverage for all five gene regions so that direct comparisons of species resolution among gene regions could be made. For this reduced data set, similar patterns were observed: *matK* provided the highest species resolution (88.3%) and *rpoC1* the lowest (72.7%); *atpF-atpH* (87.9%), *rbcL* (79.7%) and *trnH-psbA* (81.3%) gave intermediate values for species resolution. Thus, percentage species resolution was marginally higher for two of the five gene regions for the subset of taxa (282 samples) compared with the data set of 504 samples in which gene region coverage was incomplete; the increase in species resolution ranged from 2% (*trnH-psbA*) to 5.5% (*atpF-atpH*).

SPECIES RESOLUTION: MULTI-REGION ANALYSIS

The *rbcL*+*matK* barcode identified 93.1% of the taxa we sampled from the local flora (all samples, Fig. 2). The addition of a non-coding region to this barcode increased resolution by 2.2% (to 95.3% for *rbcL*+*matK*+*trnH-psbA*; Fig. 2) and the addition of all five gene regions resulted in 97.3% species resolution. When we reduced the data set to include only those samples that had sequences for all five gene regions, similar levels of species resolution were found, with results ranging from 91.4% for the *rbcL*+*matK* barcode to 95.3% for the five-locus combination (Fig. 2). Overall percentage species resolution was slightly lower for the reduced data set compared with that for the complete sample; *rbcL*+*matK* species resolution decreased by 1.7% whereas it decreased by 2.3% and 2.0% for the three- and five-gene region combinations, respectively (Fig. 2). This was a function of the particular suite of species included in the reduced set; in particular, the proportion of monotypic genera was slightly lower than for the complete set of samples.

The number of species that were resolved by the *rbcL*+*matK* barcode in our entire sample of the KSR flora was significantly lower for angiosperms (92.7%) compared with other land plants (100%; $\chi^2 = 4.53$, $P < 0.05$), which only comprised 5% of our sample (Tables S1 and S3). Identification success was significantly lower in genera that had more than one species (e.g. 83.6% using the *rbcL*+*matK* barcode) compared with monotypic genera (100%; $\chi^2 = 53.7$, $P < 0.001$). There was only a weak overall relation between the number of species per genus and the number of correctly assigned species ($\chi^2 = 77.7$, d.f. = 8, $P < 0.001$; $R^2 = 0.27$). Using the *rbcL*+*matK* barcode, percent species resolution for polytypic genera with at least some species unresolved ranged from 0% to 92% (Fig. 3).

Discussion

Our study represents one of the first attempts to barcode a local flora in the temperate zone. Our results, based on five barcoding regions used singly and in combination, provide answers to several practical questions associated with the establishment of a barcode database for local ecological applications: (i) what proportion of samples provide high-quality barcodes in a typical survey of a local flora (*recoverability*);

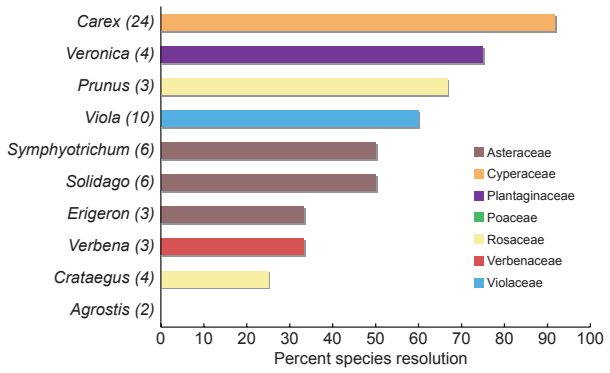

Fig. 3. Percent species resolution for polytypic genera (number of species in brackets) that were not fully resolved using the *rbcL* + *matK* barcode based on a survey of 269 genera at the Koffler Scientific Reserve, southern Ontario, Canada.

(ii) what is the level of error in species identification when an unknown sample is compared with a database in which at least some samples lack complete gene region coverage (*percent species resolution*), and (iii) what is the most cost-effective way to use barcoding for local floristic surveys: one, two or several regions? (*cost-effectiveness*). We address each of these points below.

SEQUENCE RECOVERY

We were able to recover 2130 sequences across the five gene regions screened in this study. Sequence failures across samples and regions were quite low (~2%) and were apparently randomly distributed among taxa (Fig 1; Table S1). We therefore were able to obtain at least one sequence for at least one of five regions in ~98% of the samples in our study. Only 22 (4.3%) of the samples failed for both *rbcL* + *matK* and only 14 samples (2.7%) failed for the three-locus combination, *rbcL* + *matK* + *trnH-psbA*. Our recovery rate of 97.3% for at least one of the regions for this combination is very similar to the 98% value obtained by Kress *et al.* (2009). Furthermore, in a much broader taxonomic survey, the CBOL Plant Working Group: Hollingsworth *et al.* (2009) also found similar recovery rates for angiosperms using the same three gene regions (90-98%), although they obtained reduced values for non-angiosperm plant taxa (< 50%), largely due to a lack of suitable primers. Although we did not identify the source of sequencing failures in our high-throughput workflow, the relatively high recovery rates we obtained across a broad range of taxa demonstrate that when barcoding a local flora only a small number of samples may fail to produce high-quality contigs. These failures appear to arise because some specimens produce poor quality DNA using high-throughput isolation techniques, although we cannot rule out the possibility that primer matching may also have been incomplete, particularly for *matK*.

Our recovery rate of 95.7% for samples having either component of the *rbcL* + *matK* barcode is comparable with previous results for a tropical angiosperm woody flora sampled on a similar spatial scale (94%, Kress *et al.* 2009). However, when compared with the rate of the least successfully recovered marker (*matK* in both cases), our recovery rate is significantly higher (80.5%) compared with the tropical study (69%). Overall, our higher recovery rates across a modest taxonomic assemblage of taxa are likely due to the specific subsamples examined, our use of multiple primers for *matK* (two sets, Table S2), and increased effort (multiple runs) to obtain two-fold sequence coverage, compared with previous studies. Given that the *matK* primers used in our study have relatively broad taxonomic coverage, it is reasonable to suggest that the *rbcL* + *matK* barcode can provide recovery rates approaching 95% for a local flora with relatively modest effort using two-pass runs and alternative primer pairs for *matK*, although there is still substantial room for improvement in obtaining high-quality sequences for this particular gene region (see Dunning & Savolainen 2010).

RESOLUTION

We compared identification success (species resolution) with two data sets: one comprising samples with a high-quality sequence for at least one of the five regions screened (504 samples), and a reduced data set with sequences for all five regions (N = 282 samples) (Tables S1 and S3). We made the latter comparison because our ability to recover sequences for all five gene regions for every species in our total sample was quite low (55%: Fig. 1), and this permitted direct comparisons of genomic regions using an identical taxonomic complement. For example, based on the entire sample the 79.8% of species that were resolved using *rbcL* is not the same subset of species that were resolved using *matK*. However, the percent species resolution in the reduced data set (Subset, Fig. 2), where there are no missing cells, is only marginally different from that of the entire flora for both *rbcL* and *matK* (79.7% and 88.3%, respectively). This indicates that our estimate for the complete data set (504 samples) is likely an accurate estimate of species identification success for each of these gene regions. Indeed, the results we obtained for the complete data set are similar to previous findings for a local flora in both scope and magnitude [*rbcL*: 75%, *matK*: 99% (Kress *et al.* 2009)], although substantially lower values have been reported for Amazonian trees sampled at a much smaller spatial scale [*rbcL*: 65%, *matK*: 63% (Gonzalez *et al.* 2009)], or in studies with broader taxonomic and geographic coverage [*rbcL*: 61%, *matK*: 66% (CBOL Plant Working Group: Hollingsworth *et al.* 2009)].

The ability of the combined *rbcL* + *matK* barcode to identify species in our local flora (~93%) matched that found for a local tropical angiosperm flora (92%, Kress *et al.* 2009) and exceeded that of a broader taxonomic survey where only 72% of species surveyed were successfully discriminated (CBOL Plant Working Group: Hollingsworth *et al.* 2009). Although we found that identification success was higher in non-angiosperms (100%) compared with that for angiosperms (92.7%), probably because non-angiosperm taxa are more phylogenetically isolated, angiosperms represented 95% of our sample. These results are encouraging and suggest that high rates of species resolution may be expected in a localized flora typified

by a high proportion of diverse monotypic genera (Table S3; Kress *et al*. 2009; although see Gonzalez *et al*. 2009). For our study, species identification was 100% for monotypic genera, which comprise 70.5% of the genera sampled. Furthermore, species resolution for the *rbcL* + *matK* barcode was higher than that obtained for each region separately for both the entire sample, where there are missing cells in the data set (93.1%), and the reduced data set where we had complete sequence coverage for all samples (91.4%). Collectively, our results indicate that substantial effort (and resources) to obtain complete coverage for every sample may not be necessary when establishing and querying a barcoding database for many ecological and floristic applications on a local scale.

Identification success of species in monotypic genera (100%) was significantly higher than that within polytypic genera (83.6%) using the *rbcL* + *matK* barcode. This result can be explained because species of monotypic genera would be expected to be more distinctive at both morphological and phylogenetic levels. However, this relation was quite weak; only 13% of polytypic genera contained species that were not fully resolved using the *rbcL* + *matK* barcode (Fig. 3). A closer look at the unresolved genera in our study revealed that they were taxa that are likely to have had a recent history of hybridization and/or polyploidy, which may be contributing an important source of variation to the resolution rates among these taxa (e.g. *Symphyotrichum*, *Solidago*; see Fazekas *et al*. 2008, 2009; Fig. 3). The relatively high incidence of paraphyly for these groups may also be due to coalescence failure of the plastid genome, or a simple lack of variation associated with rapid or recent radiations (Fazekas *et al*. 2008, 2009). Similar results may be expected in other barcoding applications for local floras typified by a high proportion of monotypic genera.

COST-EFFECTIVENESS

The inclusion of a third gene region – the *trnH-psbA* intergenic spacer – as a candidate plant DNA barcode has been thoroughly discussed elsewhere (see Kress *et al*. 2005, 2009; Kress & Erickson 2007; CBOL Plant Working Group: Hollingsworth *et al*. 2009). Interest in this region is due, in large part, to historical precedence, but also the greater resolution provided by this region over *rbcL* and *matK* in some groups, and the availability of primers that have wide taxonomic applicability (although the region has several potential disadvantages, see CBOL Plant Working Group: Hollingsworth *et al*. 2009). In our study, the *rbcL* + *matK* + *trnH-psbA* barcode increased species resolution from 93.1% (*rbcL* + *matK*) to 95.3% (Fig 2). Kress *et al*. (2009) reported a somewhat larger improvement (6% increase) in species discrimination using this three-locus barcode compared with *rbcL* + *matK* alone for a tropical plot. By contrast, in a broader taxonomic survey the increase in species resolution was only marginal (~1%; CBOL Plant Working Group: Hollingsworth *et al*. 2009). Several practical issues, for example, failure to obtain full length bi-directional reads due to homopolymer runs in many samples, limited our ability to obtain high-quality *trnH-psbA* sequences (also found by CBOL Plant Working Group: Hollingsworth *et al*. 2009; although see Fazekas *et al*. 2010 for a possible solution). In addition, because non-coding regions do not align well across major taxonomic groups, and thus may be difficult to apply to additional phylogenetic or taxonomic analyses (CBOL Plant Working Group: Hollingsworth *et al*. 2009; although see Kress & Erickson 2007, 2008; Kress *et al*. 2009), the addition of non-coding regions to the *rbcL* + *matK* barcode may not be worth the cost/effort involved to obtain sequences, at least for temperate local floras with modest diversity.

The inclusion of all five gene regions in our local barcoding database provided only marginal returns for increased percentage species resolution compared with that obtained with the two-locus, *rbcL* + *matK* barcode (an increase of 2%, Fig. 2). In empirical tests of the utility of five coding and three non-coding plastid gene regions (and their combinations) in a larger regional context (flora of Ontario), Fazekas *et al*. (2008) reported that species resolution reached a plateau at ~70% with little difference in cost/effort among the various combinations. Although these authors had a targeted over-representation of taxonomically complex groups in their sample, which likely contributed to lower values for species resolution within the same geographic region compared with those reported here, the plateau in performance of their different multilocus barcode combinations suggests that plant species resolution may have a fundamental upper limit in local floras (Fazekas *et al*. 2008, 2009; CBOL Plant Working Group: Hollingsworth *et al*. 2009). Our results also highlight the disparity between increasing resolution for a data set with missing cells (where we obtained 97.3% species resolution) vs. only marginal increases in actual sample resolution (to 95.3%) where gene coverage was 100% in our local flora (Fig. 2). Given such diminishing returns for increases in species resolution, the recovery and discriminatory success of the *rbcL* + *matK* barcode is encouraging and highlights the importance of expending a modest amount of effort in trying out alternative primer sets for *matK*.

In conclusion, our results provide answers to several key questions on the expected rates of sequence recovery and species resolution for recently proposed multigene barcodes, and their cost-effectiveness in local barcoding applications in which complete coverage may be difficult to achieve. In particular, the generic methods and results described in this study directly inform researchers interested in conducting biodiversity surveys of local temperate floras. Our study demonstrates that for building barcoding databases for local floras in-depth sampling of multiple collections per species may not be as necessary, compared with that required for larger scale floras in which geographical variation and a higher proportion of taxonomically complex groups are likely to occur. Furthermore, high rates of species resolution for the *rbcL* + *matK* barcode may be expected for local temperate floras of modest taxonomic breadth, thus assisting in the rapid assessment of biodiversity.

Establishing a local barcode data base will be valuable for a broad range of potential ecological applications, including the building of community phylogenies (Kress *et al*. 2009), palaeoecological investigations of past ecosystems (Valentini,

Pompanon, & Taberlet 2008), describing insect host-plant associations (Matheson *et al*. 2008; Jurado-Rivera *et al*. 2009; Navarro *et al*. 2010) or analysing the diets of mammals and birds (Bradley *et al*. 2007; Valentini *et al*. 2009). In our case, we have used the barcode data base developed in this study to identify root fragments in old-field plots to compare patterns of below- and above-ground plant diversity (Kesanakurti *et al*. 2011). These types of applications of barcoding data bases are likely to provide novel insights into plant and animal community structure and facilitate future hypothesis testing in ecology and evolution.

Acknowledgements

We thank Peter M Kotanen, Ann Zimmerman and Art Weis and staff at the Koffler Scientific Reserve at Jokers Hill for allowing us to collect samples and the Canadian Centre for DNA Barcoding at the Biodiversity Institute of Ontario and the University of Guelph Genomics Facility for DNA extraction, PCR and sequencing support. We also thank Annabel Por, John Gerrath and Carole Ann Lacroix who helped with collections and identifications, and the staff of the University of Guelph Herbarium (OAC) who provided help with collection and specimen preparation. This research was funded by a grant from Genome Canada through the Ontario Genomics Institute to the Canadian Barcode of Life Network.

References

Álvarez, I. & Wendel, J.F. (2003) Ribosomal ITS sequences and plant phylogenetic inference. *Molecular Phylogenetics & Evolution*, **29**, 417–434.

Bailey, D.C., Carrb, T.G., Harris, S.A. & Hughes, C.E. (2003) Characterization of angiosperm nrDNA polymorphism, paralogy, and pseudogenes. *Molecular Phylogenetics & Evolution*, **29**, 435–455.

Borisenko, A.V., Sones, J.E. & Hebert, P.D.N. (2009) The front-end logistics of DNA barcoding: challenges and prospects. *Molecular Ecology Resources*, **9**, 27–34.

Bradley, B.J., Stiller, M., Doran-Sheehy, D.M., Harris, T., Chapman, C.A., Vigilant, L. & Poinar, H. (2007) Plant DNA sequences from feces: potential means for assessing diets of wild primates. *American Journal of Primatology*, **69**, 699–705.

Buckler, E.S.I., Ippolito, A. & Holtsford, T.P. (1997) The evolution of ribosomal DNA: divergent paralogues and phylogenetic implications. *Genetics*, **145**, 821–832.

CBOL Plant Working Group: Hollingsworth, P.M., Forrest, L.L., Spouge, J.L., Hajibabaei, M., Ratnasingham, S., van der Bank, M., Chase, M.W., Cowan, R.S., Erickson, D.L., Fazekas, A.J., Graham, S.W., James, K.E., Kim, K-J., Kress, W.J., Schneider, H., van AlphenStahl, J., Barrett, S.C.H., van den Berg, C., Bogarin, D., Burgess, K.S., Cameron, K.M., Carine, M., Chacón, J., Clark, A., Clarkson, J.J., Conrad, F., Devey, D.S., Ford, C.S., Hedderson, T.A.J., Hollingsworth, M.L., Husband, B.C., Kelly, L.J., Kesanakurti, P.R., Kim, J.S., Kim, Y-D., Lahaye, R., Lee, H-L., Long, D.G., Madriñán, S., Maurin, O., Meusnier, I., Newmaster, S.G., Park, C-W., Percy, D.M., Petersen, G., Richardson, J.E., Salazar, G.A., Savolainen, V., Seberg, O., Wilkinson, M.J., Yi, D-K. & Little, D.P. (2009) A DNA barcode for land plants. *Proceedings of the National Academy of Sciences of the USA*, **106**, 12794–12797.

Chase, M.W., Cowan, R.S., Hollingsworth, P.M., van den Berg, C., Madriñán, S., Petersen, P. *et al.* (2007) A proposal for a standardised protocol to barcode all land plants. *Taxon*, **56**, 295–299.

Chase, M.W., Salamin, N., Wilkinson, M., Dunwell, J.M., Kesanakurthi, R.O., Haidar, N. & Savolainen, V. (2005) Land plants and DNA barcodes: short-term and long-term goals. *Philosophical Transactions of the Royal Society of London B*, **360**, 1889–1895.

Chen, S., Yao, H., Han, J., Liu, C., Song, J., Shi, L., Zhu, Y., Ma, X., Gao, T., Pang, X., Luo, K., Li, Y., Li, X., Jia, X., Lin, Y. & Leon, C. (2010) Validation of the ITS2 region as a novel DNA barcode for identifying medicinal plant species. *PLoS ONE*, **5**, e8613.

Cowan, R.S., Chase, M.W., Kress, W.J. & Savolainen, V. (2006) 300,000 species to identify: problems, progress, and prospects in DNA barcoding of land plants. *Taxon*, **55**, 611–616.

Drummond, A.J., Ashton, B., Cheung, M., Heled, J., Kearse, M., Moir, R., Stones-Havas, S., Thierer, T. & Wilson, A. (2009) Geneious v4.8. Available at: http://www.geneious.com/.

Dunning, L.T. & Savolainen, V. (2010) Broad-scale amplification of *matK* for DNA barcoding plants, a technical note. *Botanical Journal of the Linnean Society*, **164**, 1–9.

Fazekas, A.J., Burgess, K.S., Kesanakurti, P.R., Graham, S.W., Newmaster, S.G., Husband, B.C., Percy, D.M., Hajibabaei, M. & Barrett, S.C.H. (2008) Multiple multilocus DNA barcodes from the plastid genome discriminate plant species equally well. *PLoS ONE*, **3**, e2802. doi:10.1371/journal.pone.0002802.

Fazekas, A.J., Kesanakurti, P.R., Burgess, K.S., Percy, D.M., Graham, S.W., Barrett, S.C.H., Newmaster, S.G., Hajibabaei, M. & Husband, B.C. (2009) Are plant species inherently harder to discriminate than animal species using DNA barcoding markers? *Molecular Ecology Resources*, **9**, 130–139.

Fazekas, A.J., Steeves, R., Newmaster, S.G. & Hollingsworth, P.M. (2010) Stopping the stutter: improvements in sequence quality from regions with mononucleotide repeats can increase the usefulness of non-coding regions for DNA barcoding. *Taxon*, **59**, 694–697.

Gonzalez, M.A., Baraloto, C., Engel, J., Mori, S.A., Pétronelli, P., Riéra, B., Roger, A., Thébaud, C. & Chave, J. (2009) Identification of amazonian trees with DNA barcodes. *PLoS ONE*, **4**, e7483.

Hebert, P.D.N., Ratnasingham, S. & deWaard, J.R. (2003) Barcoding animal life: cytochrome *c* oxidase subunit 1 divergences among closely related species. *Proceedings of the Royal Society of London B*, **270**, S96–S99.

Hebert, P.D.N., Cywinska, A., Ball, S.L. & DeWaard, J.R. (2003) Biological identification through DNA barcodes. *Proceedings of the Royal Society of London B*, **270**, 313–321.

Ivanova, N.V., Fazekas, A.J. & Hebert, P.D.N. (2008) Semi-automated, membrane-based protocol for DNA isolation from plants. *Plant Molecular Biology Reporter*, **26**, 186–198. doi:10.1007/s11105-008-0029-4.

Janzen, D.H., Hallwachs, W., Blandin, P., Burns, J.M., Cadiou, J., Chacon, I., Dapkey, T., Deans, A.R., Epstein, M.E., Espinoza, B., Franclemont, J. G., Haber, W.A., Hajibabaei, M., Hall, J.P.W., Hebert, P.D.N., Gauld, I.D., Harvey, D.J., Hausmann, A., Kitching, I., Lafontaine, D., Landry, J., Lemaire, C., Miller, J.Y., Miller, J.S., Miller, L., Miller, S.E., Montero, J., Munroe, E., Green, R.S., Ratnasingham, S., Rawlins, J.E., Robbins, R.K., Rodriguez, J.J., Rougerie, R., Sharkey, M.J., Smith, M.A., Solis, M.A., Sullivan, J.B., Thiaucourt, P., Wahl, D.B., Weller, S.J., Whitfield, J.B., Willmott, K.R., Wood, D.M., Woodley, N.E. & Wilson, J.J. (2009) Integration of DNA barcoding into an ongoing inventory of complex tropical biodiversity. *Molecular Ecology Resources*, **9**, 1–26.

Jurado-Rivera, J.A., Vogler, A.P., Reid, C.A.M., Petitpierre, E. & Gómez-Zurita, J. (2009) DNA barcoding insect–host plant associations. *Proceedings of the Royal Society of London B*, **276**, 639–648.

Kellogg, E.A. & Bennetzen, J.L. (2004) The evolution of nuclear genome structure in seed plants. *American Journal of Botany*, **91**, 1709–1725.

Kesanakurti, P.R., Fazekas, A.J., Burgess, K.S., Percy, D.M., Newmaster, S.G., Graham, S.W., Barrett, S.C.H., Hajibabaei, M. & Husband, B.C. (2011) Spatial patterns of plant diversity below ground as revealed by DNA barcoding. *Molecular Ecology* DOI: 10.1111/j.1365-294X.2010.04989.x.

King, M.G. & Roalson, E.H. (2008) Exploring evolutionary dynamics of nrDNA in *Carex* subgenus Vignea (Cyperaceae). *Systematic Botany*, **33**, 514–524.

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**, e508.

Kress, W.J. & Erickson, D.L. (2008) DNA barcodes: genes, genomics, and bioinformatics. *Proceedings of the National Academy of Sciences of the USA*, **105**, 2761–2762.

Kress, W.J., Wurdack, K.J., Zimmer, E.A., Weigt, L.A. & Janzen, D.H. (2005) Use of DNA barcodes to identify flowering plants. *Proceedings of the National Academy of Sciences of the USA*, **102**, 8369–8374.

Kress, W.J., Erickson, D.L., Jones, F.A., Swenson, N.G., Perez, R., Sanjur, O. & Bermingham, E. (2009) Plant DNA barcodes and a community phylogeny of a tropical forest dynamics plot in Panama. *Proceedings of the National Academy of Sciences of the USA*, **106**, 18621–18626.

Lahaye, R., van der Bank, M., Bogarin, D., Warner, J., Pupulin, F., Gigot, G., Maurin, O., Duthoit, S., Barraclough, T.G. & Savolainen, V. (2008) DNA barcoding the floras of biodiversity hotspots. *Proceedings of the National Academy of Sciences of the USA*, **105**, 2923–2928. doi: 10.1073/pnas.0709936105.

Matheson, C.D., Muller, G.C., Junnila, A., Vernon, K., Hausmann, A., Miller, M.A., Greenblatt, C. & Schlein, Y. (2008) A PCR method for detection of

plant meals from the guts of insects. *Organisms Diversity & Evolution*, **7**, 294–303.
Mower, J.P., Touzet, P., Gummow, J.S., Delph, L.F. & Palmer, J.D. (2007) Extensive variation in synonymous substitution rates in mitochondrial gene of seed plants. *BMC Evolutionary Biology*, **7**, 135. doi:10.1186/1471-2148-7-135.
Navarro, S.P., Jurado-Rivera, J.A., Gómez-Zurita, J., Lyal, C.H.C. & Vogler, A.P. (2010) DNA profiling of host–herbivore interactions in tropical forests. *Ecological Entomology*, **35**, 18–32.
Newmaster, S.G., Fazekas, A.J. & Ragupathy, S. (2006) DNA barcoding in the land plants: evaluation of *rbcL* in a multigene tiered approach. *Canadian Journal of Botany*, **84**, 335–341.
Newmaster, S.G. & Ragupathy, S. (2009) Testing plant barcoding in a sister species complex of pantropical *Acacia* (Mimosoideae, Fabaceae). *Molecular Ecology Resources*, **9**, 172–180.
Ratnasingham, S. & Hebert, P.D.N. (2007) BOLD: The Barcode of Life Data System (http://www.barcodinglife.org). *Molecular Ecology Notes*, **7**, 355–364. doi: 10.1111/j.1471-8286.2006.01678.x.
SAS Institute (2002) *JMP – The Statistical Discovery Software, Version 5.0*. SAS Institute, Cary, NC, USA.
Sass, C., Little, D.P., Stevenson, D.W. & Specht, C.D. (2007) DNA barcoding in the cycadales: testing the potential of proposed barcoding markers for species identification of cycads. *PLoS ONE*, **2**, e1154.
Seberg, O. & Petersen, G. (2009) How many loci does it take to DNA barcode a Crocus? *PLoS ONE*, **4**, e4598. doi: 10.1371/journal.pone.0004598.
Starr, J., Naczi, R. & Chouinard, B. (2009) Plant DNA barcodes and species resolution in sedges (*Carex*, Cyperaceae). *Molecular Ecology Resources*, **9**, 151–163.
Taberlet, P., Coissac, E., Pompanon, F., Gielly, L., Miquel, C., Valentini, A., Vermat, T., Corthier, G., Brochmann, C. & Willerslev, E. (2006) Power and limitations of the chloroplast *trn*L (UAA) intron for plant DNA barcoding. *Nucleic Acids Research*, **35**, e1–e8.
Valentini, A., Pompanon, F. & Taberlet, P. (2008) DNA barcoding for ecologists. *Trends in Ecology & Evolution*, **24**, 110–117.
Valentini, A., Miquel, C., Nawaz, M., Bellemain, E., Coissac, E., Pompanon, F., Gielly, L., Cruaud, C., Nascetti, G., Wincker, P., Swenson, J. & Taberlet, P. (2009) New perspectives in diet analyses based on DNA barcoding and parallel pyrosequencing: the *trnL* approach. *Molecular Ecology Resources*, **9**, 51–60.

Received 28 September 2010; accepted 11 January 2011
Handling Editor: Robert P. Freckleton

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Table S1. GenBank and collection accession numbers (OAC Herbarium, University of Guelph, Ontario, Canada) for each sample collected at the Koffler Scientific Reserve, Ontario, Canada.

Table S2. Primer sequences for five plastid genomic regions examined in this study.

Table S3. The number of plastid genomic regions recovered across all 269 genera.

As a service to our authors and readers, this journal provides supporting information supplied by the authors. Such materials may be re-organized for online delivery, but are not copy-edited or typeset. Technical support issues arising from supporting information (other than missing files) should be addressed to the authors.