Testing the Efficacy of DNA Barcodes for Identifying the Vascular Plants of Canada

Thomas W. A. Braukmann*, Maria L. Kuzmina, Jesse Sills, Evgeny V. Zakharov, Paul D. N. Hebert

Centre for Biodiversity Genomics, Biodiversity Institute of Ontario, University of Guelph, Guelph, Ontario, Canada

* tbraukma@uoguelph.ca

Abstract

Their relatively slow rates of molecular evolution, as well as frequent exposure to hybridization and introgression, often make it difficult to discriminate species of vascular plants with the standard barcode markers (\textit{rbcL}, \textit{matK}, ITS2). Previous studies have examined these constraints in narrow geographic or taxonomic contexts, but the present investigation expands analysis to consider the performance of these gene regions in discriminating the species in local floras at sites across Canada. To test identification success, we employed a DNA barcode reference library with sequence records for 96\% of the 5108 vascular plant species known from Canada, but coverage varied from 94\% for \textit{rbcL} to 60\% for ITS2 and 39\% for \textit{matK}. Using plant lists from 27 national parks and one scientific reserve, we tested the efficacy of DNA barcodes in identifying the plants in simulated species assemblages from six biogeographic regions of Canada using BLAST and mothur. Mean pairwise distance (MPD) and mean nearest taxon distance (MNTD) were strong predictors of barcode performance for different plant families and genera, and both metrics supported ITS2 as possessing the highest genetic diversity. All three genes performed strongly in assigning the taxa present in local floras to the correct genus with values ranging from 91\% for \textit{rbcL} to 97\% for ITS2 and 98\% for \textit{matK}. However, \textit{matK} delivered the highest species discrimination (~81\%) followed by ITS2 (~72\%) and \textit{rbcL} (~44\%). Despite the low number of plant taxa in the Canadian Arctic, DNA barcodes had the least success in discriminating species from this biogeographic region with resolution ranging from 36\% with \textit{rbcL} to 69\% with \textit{matK}. Species resolution was higher in the other settings, peaking in the Woodland region at 52\% for \textit{rbcL} and 87\% for \textit{matK}. Our results indicate that DNA barcoding is very effective in identifying Canadian plants to a genus, and that it performs well in discriminating species in regions where floristic diversity is highest.

Introduction

DNA barcoding employs sequence variation in short, standardized gene regions as a tool to discriminate species [1]. The ideal DNA barcode region is reliably amplified and sequenced across large assemblages of taxa and provides a high level of species discrimination [2]. The
success of the 5’ region of the mitochondrial cytochrome c oxidase I (COI) gene in discriminating animal species motivated efforts to identify gene regions that might deliver similar resolution for plants. Due to the extremely low rates of nucleotide substitution in mitochondrial genes in most plant lineages [3], COI was not a candidate. However, building on their intense use for phylogenetics and molecular systematics, two plastid gene regions were considered as DNA barcodes for vascular plants and the large subunit of RuBisCo (rbcL) in combination with an intron maturase (matK) were adopted as standards [4; 5]. Because these regions often fail to resolve congeners [6; 7; 8; 9; 10; 11], there has been a subsequent trend, building on earlier suggestions [12; 13; 14], to couple them with the nuclear-encoded ribosomal internal transcribed spacer, ITS2 [2; 15; 16].

A considerable number of studies have now examined the performance of different markers with respect to both their ease of amplification and their capacity to resolve plant species [9; 10; 15; 17; 18; 19; 7; 20; 21; 22; 23; 24; 25; 26]. This work has indicated that rbcL has the highest level of sequence recovery (90–100%), followed by ITS2 (~90%), while matK is more difficult (56–90%). The efficacy of these gene regions in discriminating species has been determined by tree-based (phylogenetic) or basic local alignment (BLAST) algorithms. ITS2 has been reported to deliver the highest species resolution (79–93%) followed by matK (45–80%), and rbcL (17%–92%). It was suggested that the efficacy of DNA barcodes in delivering species-level identifications could be improved by developing local libraries [7; 27], and it was later demonstrated that this approach did indeed improve resolution [9; 23]. The effectiveness of such libraries depends upon complete sampling of local floras, accurate identification of the specimens that are analyzed, and quality of the resultant sequences [28].

Comparisons among past studies are difficult due to high variance in taxonomic scope (30–4800 species), biogeographic focus (e.g. Arctic and temperate floras, tropical trees), the number of DNA barcode markers employed (2–8 chloroplast and nuclear), and the methodologies used for making taxonomic assignments. In fact, no prior study has involved a large-scale comparative analysis of the capacity of the standard barcode markers (rbcL, matK, ITS2) to deliver a species-level identification for different biogeographic communities using a standard barcode library with the same methods. This study addresses this gap by employing a DNA barcode library for the vascular plants of Canada to determine the method that yields the best species resolution and the marker (rbcL, matK, ITS2) with the highest performance. As well, this study examines the efficacy of custom DNA barcode libraries for identification success, and compares phylogenetic diversity measures between sites and among species–rich families to determine factors affecting species resolution.

Materials and Methods

Taxonomic sampling

Sequences for three DNA barcode regions (rbcL, matK and ITS2) were generated for the vascular plants of Canada at the Canadian Center for DNA Barcoding [29]. Complete taxonomic information, collection records, voucher images and sequences for 17,995 specimens are publicly available through BOLD [30] in the plants of Canada project (Available as of January 4, 2016; doi: dx.doi.org/10.5883/DS-VASCAN). This sequence library includes records for 4923 of the 5108 species of non-hybrid origin (~96%) with coverage for all 1153 genera and 171 families in the Database of Vascular Plants of Canada (VASCAN; [31]). Coverage varies among the three gene regions; the rbcL dataset is most complete with 16,008 sequences spanning 4790 species (~93.8%) in 168 families (Table 1). The ITS2 library includes 6630 sequences representing 3044 species (~59.6%) in 125 families while the matK dataset includes 6599 sequences covering 2000 species (39%) across 118 families. Overall, 78% of the species (3839)
possess records for some combination of two markers, but only 1074 species (22%) have data for all three.

To test the taxonomic resolution of the DNA library we created ‘synthetic’ floras based on the checklist of vascular plants for each of 27 Canadian National Parks and the Koffler Scientific Reserve (KSR). Initial checklists were generated using the Parks Canada Biotics Web explorer at http://www.pc.gc.ca/apps/bos/bosfieldselection_e.asp, with more recent updates for Ellesmere, Ivivavik, Nahannii, Point Pelee, Torngat, Ukkusiksalik, and Wapusk National Parks (Bruce Bennett and Sergei Ponomarenko, personal communication). The species list for KSR was obtained from http://ksr.utoronto.ca/research/species-list/ksr-plant-list/. Plant species on the checklists were best represented by rbcL (> 95% coverage), followed by ITS2 and matK with comparable coverage (54–83% depending on the community; see Fig 1 for details). For the purpose of further analyses, the 28 checklists were clustered into six biogeographic regions: Arctic, Atlantic, Boreal, Pacific, Prairies, and Woodland (Table 1) representing 12 of the 15 terrestrial Canadian ecozones [32]. To ensure standardization of naming, all specimens and checklists used in this study followed the nomenclature accepted by VASCAN [31].

<table>
<thead>
<tr>
<th>Park (species)</th>
<th>Terrestrial Ecozones</th>
<th>Region</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ellesmere Island NP (135)</td>
<td>Northern Arctic</td>
<td>Arctic</td>
</tr>
<tr>
<td>Ivivavik NP (402)</td>
<td>Southern Arctic</td>
<td>Arctic</td>
</tr>
<tr>
<td>Nahannii NP (652)</td>
<td>Taiga Plains</td>
<td>Arctic</td>
</tr>
<tr>
<td>Torngat Mountains NP (368)</td>
<td>Arctic Cordillera</td>
<td>Arctic</td>
</tr>
<tr>
<td>Ukkusiksalik NP (150)</td>
<td>Northern Arctic</td>
<td>Arctic</td>
</tr>
<tr>
<td>Wapusk NP (272)</td>
<td>Hudson Plains</td>
<td>Arctic</td>
</tr>
<tr>
<td>Forillon NP (622)</td>
<td>Atlantic Maritime</td>
<td>Atlantic</td>
</tr>
<tr>
<td>Fundy NP (732)</td>
<td>Atlantic Maritime</td>
<td>Atlantic</td>
</tr>
<tr>
<td>Kejimkujik NP (578)</td>
<td>Atlantic Maritime</td>
<td>Atlantic</td>
</tr>
<tr>
<td>Prince Edward Island NP (648)</td>
<td>Atlantic Maritime</td>
<td>Atlantic</td>
</tr>
<tr>
<td>La Mauricie NP (461)</td>
<td>Boreal Shield</td>
<td>Boreal</td>
</tr>
<tr>
<td>Mingan Archipelago NP (441)</td>
<td>Boreal Shield</td>
<td>Boreal</td>
</tr>
<tr>
<td>Pukaskwa NP (545)</td>
<td>Boreal Shield</td>
<td>Boreal</td>
</tr>
<tr>
<td>Terra Nova NP (506)</td>
<td>Boreal Shield</td>
<td>Boreal</td>
</tr>
<tr>
<td>Banff NP (910)</td>
<td>Montane Cordillera</td>
<td>Pacific</td>
</tr>
<tr>
<td>Glacier NP (634)</td>
<td>Montane Cordillera</td>
<td>Pacific</td>
</tr>
<tr>
<td>Mount Revelstoke NP (435)</td>
<td>Montane Cordillera</td>
<td>Pacific</td>
</tr>
<tr>
<td>Pacific Rim NP (436)</td>
<td>Pacific Maritime</td>
<td>Pacific</td>
</tr>
<tr>
<td>Yoho NP (658)</td>
<td>Montane Cordillera</td>
<td>Pacific</td>
</tr>
<tr>
<td>Elk Island NP (482)</td>
<td>Boreal Plains</td>
<td>Prairies</td>
</tr>
<tr>
<td>Grasslands NP (427)</td>
<td>Prairies</td>
<td>Prairies</td>
</tr>
<tr>
<td>Prince Albert NP (643)</td>
<td>Boreal Plains</td>
<td>Prairies</td>
</tr>
<tr>
<td>Riding Mountain NP (714)</td>
<td>Prairies</td>
<td>Prairies</td>
</tr>
<tr>
<td>Waterton Lakes NP (976)</td>
<td>Prairies</td>
<td>Prairies</td>
</tr>
<tr>
<td>1000 Islands NP (1631)</td>
<td>Mixed wood Plains</td>
<td>Woodland</td>
</tr>
<tr>
<td>Bruce Peninsula NP (877)</td>
<td>Mixed wood Plains</td>
<td>Woodland</td>
</tr>
<tr>
<td>Koffler Scientific Reserve (621)</td>
<td>Mixed wood Plains</td>
<td>Woodland</td>
</tr>
<tr>
<td>Point Pelee NP (858)</td>
<td>Mixed wood Plains</td>
<td>Woodland</td>
</tr>
</tbody>
</table>

doi:10.1371/journal.pone.0169515.t001
Sequencing and analysis of libraries

Data validation. To reduce redundancy, identical sequences were clustered in UCLUST [33] and each cluster was parsed to its respective species (one species could be represented by more than one cluster). Sequences were then aligned using transAlign [34] for rbcL and matK (universal codon table), and MAFFT ver 7.221 for ITS2 under default parameters (FFT-NS-2 strategy) [35]. Maximum likelihood phylogenies were inferred for each alignment using RAxML Black box [36] on XCEDE via the CIPRES portal [37]. A dataset of 1074 species with records for all three gene regions was used to evaluate variation in taxonomic resolution (via BLAST and mothur) and phylogenetic metrics (MPD and MNTD). To estimate the number of unique sequences as a proxy for sequence variation, we clustered each marker at 100% using UCLUST [33].

Phylogenetic matrices. We calculated two metrics for each barcode region, mean phylogenetic distance (MPD) and mean nearest taxon distance (MNTD) [38] to examine their potential as predictors of the capacity of each region to resolve species. MPD is the average of the branch lengths (or distances) across all pairs of taxa in a phylogeny. It summarizes the overall phylogenetic diversity of a community and is influenced by the number of taxa in a tree [39]. By comparison, MNTD is an average of the distance between nearest neighbours so it describes the terminal phylogenetic structure. MNTD is the more appropriate measure of species resolution because it excludes internal nodes and instead calculates the mean distance between closely related species. Because both measures are influenced by polytomies in a phylogeny [40], we only included one representative per species to avoid bias created by an unequal number of sequences per species.
MPD and MNTD were estimated using the picante package [41] in R ver 3.2.0 [42]. The phylogenetic matrices for each barcode were calculated using the maximum likelihood tree, and were partitioned by family and genus. Regression analysis was used to determine if there was a correlation between each phylogenetic diversity metric and the number of sequences for a family. We also compared MNTD values for the three markers using a common set of genera to determine the strength of the correlation in their divergence values. To determine if significant differences existed between markers, Kruskal-Wallis (KW) tests followed by a Dunn’s posthoc were carried out in R ver 3.2.0 [42].

A similar analysis was conducted for each park community using RAxML-based trees to calculate MPD and MNTD partitioned by family. The percentage of congenerers in the six large families with low MNTD (Asteraceae, Brassicaceae, Cyperaceae, Poaceae, Rosaceae, and Salicaceae) was evaluated for the datasets representing the three barcodes for the six biogeographic regions.

**Taxonomic resolution.** Our custom sequence library for Canadian plants was parsed based on the species present at each locality and the taxonomic resolution provided by each barcode was then evaluated using BLAST searches and by mothur in Qiime [43]. For both methods, the species known for each park were compared with the parent library to ascertain if barcode records allowed their identification to a family, genus, or species level. The resolution for species with multiple sequences was recorded as that where the taxonomic assignment for all individuals was consistent (e.g. if there were four sequences for species A and three were unambiguously identified to a species and one was to a genus, the recorded level of resolution would be to a genus). This approach generates a ‘worst case’ outcome for the capacity to identify a particular species. Mothur employs a distance matrix to assign a sequence (or cluster) to a species based on a parent library. For its use, identifications were predicted using a posterior probability cut-off of 0.95. We also report the true level of success of mothur by comparing the taxonomic identification assigned to a given sequence by mothur with its correct assignment. The data for each park was then used to generate a mean level of taxonomic resolution for each family, genus, and species. Data was checked for normality prior to conducting a Kruskal-Wallis (KW) test or one-way ANOVA to test for significant differences in species resolution among the three markers. Any significant test was followed up with the appropriate posthoc tests (Tukey’s HSD for ANOVA or Dunn for KW). The parks were then subdivided into six biogeographic regions (Arctic, Atlantic, Boreal, Pacific, Prairies, Woodland) and the data was pooled for each region to estimate the mean level of taxonomic resolution for the floras that were examined. After checking for normality, KW or one-way ANOVA was used to test for a significant difference in species resolution among the regions for a particular barcode marker. We also evaluated taxonomic resolution for the 1074 species with data for all three barcode genes to compare the mean of the parks and the performance of different markers using an identical set of taxa. The performance of the barcodes for 25 families with the most species was then compared based on the BLAST results to identify groups where barcodes delivered low taxonomic resolution. All statistical tests were performed in R ver 3.2.0 [42] with Bonferroni error corrections for multiple tests (adjusted p = 0.005).

**Results**

**Clustering and phylogenetic matrices**

After the removal of identical sequences within any one species, the read library was reduced to 5919 sequences for *rbcL*, 2891 sequences for *matK*, and 4423 sequences for *ITS2*. The plastid markers were much less variable than ITS2 as evidenced when the read libraries were clustered at 100% identity which collapsed the sequence count when different species shared a particular
sequence. This analysis showed that \textit{rbcl} had considerably less sequence variation (2895 clusters; 5919 sequences) than \textit{matK} (2145 clusters; 2891 sequences) while \textit{ITS2} was most diverse (4418 clusters; 4423 sequences). This pattern was reinforced by the global estimates for MPD and MNTD that rated \textit{ITS2} as the most variable marker followed by \textit{matK} and \textit{rbcl} (Fig 2, Table A and Table B in S1 File). For each measure, markers were significantly different from one another (KW and Dunn’s posthoc p < 0.0005).

The Asteraceae had low values for both metrics across all three barcodes (Table 2), but those for the family Salicaceae were exceptionally so for MPD (\textit{rbcl} = 0.017; \textit{matK} = 0.009; \textit{ITS2} = 0.036) and MNTD (\textit{rbcl} = 0.0005; \textit{matK} = 0.0007; \textit{ITS2} = 0.013). The latter result reflected the low MNTD values within \textit{Salix} (41–90 species per region; \textit{matK} and \textit{rbcl} = 0.0005; \textit{ITS2} = 0.009; Table B in S1 File). The Asteraceae also had low MNTD (\textit{rbcl} = 0.002; \textit{matK} = 0.006; \textit{ITS2} = 0.021), strongly influenced by four genera with MNTD < 2 e−06. Interestingly, MPD did not predict low species resolution for Asteraceae (\textit{rbcl} = 0.073; \textit{matK} = 0.07; \textit{ITS2} = 0.373), because some long internal branches raised this measure (Table B in S1 File).

High phylogenetic diversity for families lacking genera with a low MNTD or MPD is a strong predictor of strong species resolution. For example, the Caryophyllaceae and Fabaceae have high MPD and MNTD for \textit{rbcl} (MPD = 0.081, 0.160; MNTD = 0.008, 0.009 respectively; Table 2), but several of their genera have near zero values for both metrics (< 0.001; Table B in S1 File) suggesting that these lineages will have much lower species resolution than highly variable genera. By contrast, nearly all genera of the Orchidaceae and Primulaceae have high MPD...
and MNTD, ensuring high species resolution (see Table B in S1 File). Species resolution is also strong for the Lamiaceae (ITS2), Onagraceae (matK), and Polygonaceae (ITS2 and matK) (Table 2) due to their high genetic diversity. There was no correlation between the number of species in a family or genus and either MPD or MNTD ($r^2 < 0.05$ for all comparisons).

There was also no correlation between markers for MNTD values ($r^2 < 0.007$ for all comparisons; S1 Fig).

MPD and MNTD were used to predict those parks and biogeographic regions where DNA barcodes would deliver poor taxonomic resolution. Both values were generally lower in the Arctic than in the other biogeographic regions for all three markers (see Table C in S1 File for details), suggesting that species resolution should be most challenging in the north (Table C in S1 File). These estimates of genetic diversity further predict that ITS2 will deliver the best taxonomic resolution followed by matK and rbcL.

### Taxonomic resolution

**Overall.** Performance comparison of BLAST and mothur in identifying plants from the 28 localities (Table 3; Fig 1) indicated that BLAST delivered higher species resolution for all three barcodes (Fig 3a–3c). When employing a posterior probability cut-off of 0.95, mothur
underestimated the capacity to make species-level identifications, but overestimated it at a genus level (Table D in S1 File). Both BLAST and mothur indicated that *rbcl* has the lowest species (45% and 31% with BLAST and mothur respectively) and generic (91% and 84% with BLAST and mothur respectively) resolution (Table D in S1 File), but diverged on which marker provides the highest species resolution. BLAST generates the highest species resolution with *matK* (80%) followed by ITS2 (73%). By comparison, mothur ranks ITS2 as the best barcode when resolving taxa with both posterior probability (ITS2 mean = 64% vs. *matK* mean = 58%) and true species resolution (69% versus 62%). Generic resolution was high for both *matK* (~96–98%) and ITS2 (96–99%) using either approach (Table 3). The difference in species resolution was significant between markers for both algorithms (p < 0.005; Fig 3a–3c).

Analysis of the dataset consisting of 1074 species represented by all three barcodes generated similar results to the park data (Table D in S1 File). Since BLAST yielded the highest species resolution for each marker, these results were employed for the further analyses.

**Species resolution by family.** For most families, *matK* delivered the highest resolution followed by ITS2, but the two gene regions were complementary, jointly delivering 85% species
resolution if two families were excluded (Salicaceae, Asteraceae) (Table 4; Fig 4). In fact, matK delivered perfect resolution for four families (Onagraceae, Polemoniaceae, Boraginaceae, Caprifoliaceae), while ITS2 did well for Lamiaceae (98%) and Orchidaceae (92%). By comparison, rbcL had low species resolution (<60%) for all families except Orchidaceae (78%), Ericaceae (65%), Plantaginaceae (64%), Primulaceae (71%), and Saxifragaceae (69%). Generic resolution was high for matK (98%) and ITS2 (97%) but slightly lower for rbcL (91%). Families with compromised generic resolution included the Asteraceae (rbcL = 78%; matK = 97%; ITS2 = 92%), Fabaceae (rbcL = 81%; matK = 93%; ITS2 = 89%) and Poaceae (rbcL = 82%; matK = 95%; ITS2 = 95%) (Table 4; Fig 4).

Consistent with their low values for MNTD and MPD, species resolution was poor for the Salicaceae (<31%) and Asteraceae (<68%). Their low MPD and MNTD values also predicted that certain genes would fail to distinguish species of Violaceae (rbcL), Rosaceae (matK), and Onagraceae (ITS2). When accounting for genera with low MPD and MNTD within families,
low species resolution was apparent for Fabaceae (rbcL and matK) and Caryophyllaceae (rbcL), while resolution was high for Orchidaceae (rbcL) and Primulaceae (rbcL). The lack of low resolution genera in the Polygonaceae (matK and ITS2), Caprifoliaceae (ITS2 and matK), Polemoniaceae (matK), and Lamiaceae (ITS2) accounts for the relatively high success of barcoding in these taxa (Table B in S1 File; Fig 4).

<table>
<thead>
<tr>
<th>Blast Family</th>
<th>rbcL Family</th>
<th>Genus</th>
<th>species</th>
<th>matK Family</th>
<th>Genus</th>
<th>species</th>
<th>ITS2 Family</th>
<th>Genus</th>
<th>species</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amaranthaceae</td>
<td>100.00%</td>
<td>96.48%</td>
<td>50.70%</td>
<td>100.00%</td>
<td>100.00%</td>
<td>91.09%</td>
<td>100.00%</td>
<td>95.85%</td>
<td>54.35%</td>
</tr>
<tr>
<td>Apiaceae</td>
<td>100.00%</td>
<td>87.34%</td>
<td>56.54%</td>
<td>100.00%</td>
<td>97.86%</td>
<td>94.12%</td>
<td>100.00%</td>
<td>97.62%</td>
<td>86.19%</td>
</tr>
<tr>
<td>Asteraceae</td>
<td>100.00%</td>
<td>77.97%</td>
<td>26.39%</td>
<td>99.76%</td>
<td>96.76%</td>
<td>67.17%</td>
<td>100.00%</td>
<td>91.50%</td>
<td>58.37%</td>
</tr>
<tr>
<td>Boraginaceae</td>
<td>100.00%</td>
<td>88.89%</td>
<td>52.59%</td>
<td>100.00%</td>
<td>100.00%</td>
<td>100.00%</td>
<td>100.00%</td>
<td>100.00%</td>
<td>83.56%</td>
</tr>
<tr>
<td>Brassicaceae</td>
<td>100.00%</td>
<td>90.29%</td>
<td>34.44%</td>
<td>100.00%</td>
<td>100.00%</td>
<td>100.00%</td>
<td>100.00%</td>
<td>100.00%</td>
<td>49.67%</td>
</tr>
<tr>
<td>Caprifoliaceae</td>
<td>100.00%</td>
<td>85.23%</td>
<td>53.69%</td>
<td>100.00%</td>
<td>100.00%</td>
<td>100.00%</td>
<td>100.00%</td>
<td>100.00%</td>
<td>89.66%</td>
</tr>
<tr>
<td>Caryophyllaceae</td>
<td>100.00%</td>
<td>81.49%</td>
<td>33.63%</td>
<td>100.00%</td>
<td>100.00%</td>
<td>92.83%</td>
<td>100.00%</td>
<td>95.61%</td>
<td>67.80%</td>
</tr>
<tr>
<td>Cyperaceae</td>
<td>99.94%</td>
<td>95.15%</td>
<td>30.08%</td>
<td>100.00%</td>
<td>97.29%</td>
<td>85.67%</td>
<td>100.00%</td>
<td>100.00%</td>
<td>89.19%</td>
</tr>
<tr>
<td>Ericaceae</td>
<td>100.00%</td>
<td>99.62%</td>
<td>64.82%</td>
<td>100.00%</td>
<td>100.00%</td>
<td>95.69%</td>
<td>100.00%</td>
<td>100.00%</td>
<td>72.22%</td>
</tr>
<tr>
<td>Fabaceae</td>
<td>95.66%</td>
<td>81.07%</td>
<td>41.42%</td>
<td>100.00%</td>
<td>92.48%</td>
<td>72.09%</td>
<td>100.00%</td>
<td>89.39%</td>
<td>70.88%</td>
</tr>
<tr>
<td>Juncaceae</td>
<td>100.00%</td>
<td>99.08%</td>
<td>54.29%</td>
<td>100.00%</td>
<td>100.00%</td>
<td>86.05%</td>
<td>100.00%</td>
<td>100.00%</td>
<td>88.81%</td>
</tr>
<tr>
<td>Lamiaceae</td>
<td>100.00%</td>
<td>88.93%</td>
<td>56.15%</td>
<td>100.00%</td>
<td>100.00%</td>
<td>88.04%</td>
<td>100.00%</td>
<td>99.04%</td>
<td>98.08%</td>
</tr>
<tr>
<td>Onagraceae</td>
<td>100.00%</td>
<td>99.55%</td>
<td>49.77%</td>
<td>100.00%</td>
<td>100.00%</td>
<td>100.00%</td>
<td>100.00%</td>
<td>100.00%</td>
<td>61.75%</td>
</tr>
<tr>
<td>Orchidaceae</td>
<td>100.00%</td>
<td>89.08%</td>
<td>77.59%</td>
<td>100.00%</td>
<td>100.00%</td>
<td>83.70%</td>
<td>100.00%</td>
<td>96.12%</td>
<td>92.24%</td>
</tr>
<tr>
<td>Orobanchaceae</td>
<td>100.00%</td>
<td>99.44%</td>
<td>52.54%</td>
<td>100.00%</td>
<td>100.00%</td>
<td>76.40%</td>
<td>100.00%</td>
<td>100.00%</td>
<td>79.35%</td>
</tr>
<tr>
<td>Plantaginaceae</td>
<td>100.00%</td>
<td>98.64%</td>
<td>64.07%</td>
<td>100.00%</td>
<td>100.00%</td>
<td>91.88%</td>
<td>100.00%</td>
<td>100.00%</td>
<td>83.54%</td>
</tr>
<tr>
<td>Poaceae</td>
<td>100.00%</td>
<td>81.48%</td>
<td>39.89%</td>
<td>100.00%</td>
<td>95.00%</td>
<td>75.25%</td>
<td>100.00%</td>
<td>95.23%</td>
<td>66.84%</td>
</tr>
<tr>
<td>Polemoniaceae</td>
<td>100.00%</td>
<td>96.77%</td>
<td>29.03%</td>
<td>100.00%</td>
<td>100.00%</td>
<td>100.00%</td>
<td>100.00%</td>
<td>100.00%</td>
<td>75.86%</td>
</tr>
<tr>
<td>Polygonaceae</td>
<td>92.61%</td>
<td>89.79%</td>
<td>48.59%</td>
<td>100.00%</td>
<td>100.00%</td>
<td>94.71%</td>
<td>100.00%</td>
<td>100.00%</td>
<td>88.67%</td>
</tr>
<tr>
<td>Primulaceae</td>
<td>100.00%</td>
<td>99.17%</td>
<td>71.07%</td>
<td>100.00%</td>
<td>100.00%</td>
<td>88.16%</td>
<td>100.00%</td>
<td>100.00%</td>
<td>63.11%</td>
</tr>
<tr>
<td>Ranunculaceae</td>
<td>100.00%</td>
<td>97.97%</td>
<td>54.07%</td>
<td>100.00%</td>
<td>100.00%</td>
<td>85.37%</td>
<td>100.00%</td>
<td>100.00%</td>
<td>81.75%</td>
</tr>
<tr>
<td>Rosaceae</td>
<td>100.00%</td>
<td>88.78%</td>
<td>39.86%</td>
<td>100.00%</td>
<td>100.00%</td>
<td>80.37%</td>
<td>100.00%</td>
<td>99.84%</td>
<td>85.02%</td>
</tr>
<tr>
<td>Salicaceae</td>
<td>100.00%</td>
<td>99.76%</td>
<td>12.24%</td>
<td>100.00%</td>
<td>100.00%</td>
<td>26.47%</td>
<td>100.00%</td>
<td>100.00%</td>
<td>30.96%</td>
</tr>
<tr>
<td>Saxifragaceae</td>
<td>100.00%</td>
<td>95.06%</td>
<td>68.72%</td>
<td>100.00%</td>
<td>97.69%</td>
<td>91.33%</td>
<td>100.00%</td>
<td>100.00%</td>
<td>89.24%</td>
</tr>
<tr>
<td>Violaceae</td>
<td>100.00%</td>
<td>100.00%</td>
<td>30.41%</td>
<td>100.00%</td>
<td>100.00%</td>
<td>90.00%</td>
<td>100.00%</td>
<td>100.00%</td>
<td>69.78%</td>
</tr>
</tbody>
</table>

doi:10.1371/journal.pone.0169515.t004

Fig 4. Level of taxonomic resolution provided by rbcL, matK or ITS2 for 25 families. Level of taxonomic resolution provided by rbcL, matK or ITS2 for 25 families of vascular plant that are species-rich in Canada. The three colours show the proportion of species identified to a family (blue), genus (orange) or species (green) level.

doi:10.1371/journal.pone.0169515.g004
**Species resolution by region.** When the 28 localities were organized into six biogeographic regions (Table 1), Arctic sites had significantly lower levels of species resolution than those in the other five regions (p < 0.05 for all markers; Fig 3d–3f) although results varied by marker. For *rbcl*, the Atlantic and Woodland regions have significantly higher taxonomic resolution than all others (p < 0.005) while Boreal and Pacific communities have significantly higher taxonomic resolution than Arctic and Prairie assemblages (p < 0.005). For *matK*, only Arctic communities have significantly lower species resolution (p < 0.005) than the other localities. Atlantic and Woodland have significantly higher resolution with ITS2 than the other regions (p < 0.005; Fig 3d–3f), as predicted by MPD and MNTD.

**Discussion**

This study examined the effectiveness of DNA barcoding in the identification of plants from six biogeographic regions of Canada using both local alignment (BLAST) and phylogeny-based (MNTD and MPD) approaches. MPD and MNTD were first proposed as measures of phylogenetic diversity within a community [38], and have commonly been used to study community assembly [44; 45; 46]. MPD was previously used to compare substitution rates among plant families for three barcode regions (*rbcl*, *matK*, ITS2), and a positive correlation was reported between these rates and their capacity to discriminate species [26]. The present study extended this work by examining the utility of MPD and MNTD as predictors of species resolution for the same three gene regions.

Both MPD and MNTD indicated that ITS2 should deliver the best species resolution, an expected result given the higher rates of nucleotide substitution in nuclear than organellar genomes of plants [47; 48]. The prediction was supported when mothur was used to generate taxonomic assignments, but *matK* delivered the best species resolution with BLAST. Interestingly, BLAST yielded higher species resolution than mothur for all three markers, a result which was maintained even when analysis was restricted to the 1074 species with sequence data for all three regions. BLAST’s higher resolution is explained by its greater sensitivity to sequence length [49], as well its inclusion of indel variation, which phylogenetic approaches typically overlook. Although *matK* has less sequence variation than ITS2, it contains more indels which helped it to achieve higher species resolution with BLAST. Our results support the need for DNA barcoding to utilize phylogenetic methods that incorporate indels to maximize the resolving power of a given marker.

The genome compartment exposed to the highest intraspecific gene flow is generally the best suited for making species assignments because it reduces the likelihood that introgressed alleles will gain establishment and blur species diagnosis. Gene flow raises effective population size, reducing exposure to genetic drift, diminishing the chance of introgressed alleles gaining fixation in the gene pool, and increasing the probability that a particular gene will track species relationships [50]. Since the nuclear genome tends to experience greater dispersal and gene flow than the plastid genome, nuclear markers are generally more effective in species diagnosis than their plastid counterparts [50; 51]. Hence, the incorporation of a nuclear marker with the core (plastid) barcodes offers the advantage of compensating for situations where plastid markers fail to provide resolution. ITS2 did outperform its plastid counterparts in several species-rich families (i.e. Lamiaceae, Poaceae, Cyperaceae, and Saxifragaceae; Table 4) examined in this study. However, it was less effective in other families, likely reflecting incomplete lineage sorting stemming from its larger effective population sizes or, in rare cases, when plastid dispersal exceeds that of the nucleus [50; 51]. Despite these situations, the incorporation of ITS2 is preferable over additional plastid markers such as *psbA-trnH* because it occurs in all plant species which is not true for any plastid marker (including *rbcl* and *psbA-trnH*) and
existing primers are nearly universal (as opposed to those for *matK*). Moreover, it delivers high resolution despite its short length (~350 bp), making it an ideal marker for studies using high-throughput sequencing platforms which cannot recover full-length sequences for longer barcodes such as *matK* (~800 bp) or those variable in length (*psbA-trnH*: 50–1000 bp).

The observed differences in taxonomic resolution for the three barcodes are undoubtedly influenced by selection, species demography, hybridization, lineage sorting, and phylogeographic structure (reviewed by [52]). The higher resolution of *matK* compared to *rbcL* reflects the different selective pressures acting on these genes. Because it encodes the large subunit of RuBisCo which has an essential role in photosynthesis, *rbcL* is under strong purifying selection in autotrophic plants [53] reducing its rate of evolution and constraining its utility for distinguishing closely related species. By contrast *matK*, an intron maturase involved in the splicing of group IIa introns, appears to be under relaxed purifying selection as evidenced by nearly equal substitution rates for all three coding positions [53; 54; 55; 56]. The relatively high rates of nucleotide substitution in *matK* compared to other plastid genes is useful for species delimitation, but a lack of conserved priming sites often undermines sequence recovery. Nuclear markers have a larger effective population size than plastid markers and tend to evolve more rapidly [47; 48]. The higher rates of nucleotide substitution and dispersal in plant nuclear genomes support their inclusion for plant DNA barcoding [51]. Additionally, the presence of multiple alleles for nuclear genes makes it possible to identify hybrids. Currently the only plant nuclear locus that meets barcoding criteria is ITS2 (see above) and its inclusion adds depth to barcode reference libraries by tracking a different genomic compartment.

The present analysis shows that MPD and MNTD are strong predictors of barcode resolution, identifying families and genera where taxonomic resolution is low. They were particularly useful in revealing genera with low resolution in families where divergences are high. For example, the Fabaceae has a relatively high MPD for both *rbcL* and *matK* but low species resolution, reflecting its inclusion of several genera (*e.g.* *Lupinus*, *Oxytropis*) with low MNTD. The latter genera explain the lower than average species resolution in this family for all markers, but this outcome was especially surprising for *Lupinus* because it was previously observed to have high genetic diversity in North America and low genetic diversity in the Andes due to a recent adaptive radiation [57]. Further research is needed to determine if a similar radiation occurred in North America. MNTD is a better predictor of species resolution than MPD because it quantifies the distance between pairs of closely related species and it is also less influenced by polytomies than MPD or PD [40]. As such, it is a better estimator of the efficacy of a given DNA barcode. The low correlation between MNTD values for the three barcode regions in different genera implies they are evolving independently (S1 Fig). As a consequence, the use of multiple barcode markers consistently improves taxonomic resolution because a particular marker can compensate for the deficits in resolution of its counterparts [7; 9; 10; 15; 17; 18; 19; 20; 21; 22; 23; 24; 25; 26]. This complementarity supports the use of specific barcodes that optimize species resolution for different groups [58; 59].

The patterns of variation in the phylogenetic matrices (MPD, MNTD) agree with the earlier conclusion [15; 16] that ITS2 has higher discriminatory power than *matK* or *rbcL* when specimens are analyzed against a local reference library. However, this conclusion may not extend to other situations, such as the present study, where taxonomic resolution is compared against a more comprehensive parent library, an approach which provides a ‘real world’ outcome of DNA barcoding. For example, Burgess et al. [23] reported 88% species resolution with *matK* and 80% with *rbcL* when identifications were driven by a barcode library comprised solely of plant species known from the site. Analysis of the same community with the barcode library for all Canadian plants lowered species resolution (i.e. 86% for *matK*, 54% for *rbcL*) but with the advantage that newly encountered plants would potentially be identified.
The low levels of sequence variation in several plant families likely reflects the joint impacts of polyploidization, hybridization, phylogeographic effects such as allele surfing during range expansions, and demographic effects including bottlenecks which reduce intra- and inter-specific variation [52; 60; 61]. These effects are more prominent in Arctic communities that might explain the frequent failure of both nuclear and plastid markers in discriminating species in this region [52; 62; 63; 64; 65]. Although these processes (singly or in combination) compromise the effectiveness of DNA barcoding in discriminating plant species, they do provide an opportunity to understand the factors that shape plant populations and genomes. While our dataset lacks the extensive sampling needed to differentiate between these processes, it does reveal taxa that require further study.

Among the 171 plant families in Canada, Salicaceae has the lowest species discrimination, largely due to the very limited genetic diversity among the 90 species of Salix. Its lack of variation in seven regions of the plastid genome was linked to frequent hybridization, incomplete lineage sorting, or repeated plastid capture events [66]. However, the same lack of resolution was observed with our ITS2 data and in a more detailed analysis of 22 species [26], indicating that the lack of divergence extends into the nuclear genome. This difficulty in differentiating Salix species using molecular markers may reflect hybridization, introgression, recent speciation, allele surfing via range expansion, and low rates of molecular evolution [52; 59; 67; 68; 69]. More extensive phylogenetic studies targeting nuclear markers or whole plastid genomes are necessary to clarify the processes driving the unusually low divergences in Salix.

The Asteraceae is another family where DNA barcodes deliver poor species resolution, but the underlying factors differ from those in the Salicaceae. The Asteraceae is a species-rich group that lacks reciprocal monophyly between some closely related species and genera [70], taxa that are difficult to differentiate with molecular data [71]. The Fabaceae also showed poor species resolution with all three loci (41–72%), an expected result given the number of poorly resolved genera in this family [72]. Although species resolution in certain groups may never be resolved by the targeted analysis of a few barcode loci, they do represent interesting models for testing the effectiveness of whole plastid genomes as a tool for species discrimination (reviewed by [58]).

When comparisons were extended across the six biogeographic regions of Canada, DNA barcoding delivered the poorest species discrimination in the Arctic, perhaps reflecting the higher incidence of congeners in Arctic communities (47–53%) than in other regions (39–48%; S2 Fig). As well, the arctic flora is rich in recently radiated congeners that have not achieved reciprocal monophyly [63; 65]. As a consequence, DNA barcoding delivered higher species resolution in the more floristically diverse regions. In fact, the most floristically diverse regions (Atlantic and Woodland) had the highest species resolution for all markers, suggesting that increased species diversity is correlated with genetic diversity. This difference likely reflects both demographic effects and shifts in community composition. For example, species of Salix comprise 8% of the flora near Churchill, reducing the overall success of barcoding at this arctic site [26]. We do not observe similar compositional biases in more floristically diverse communities that would influence overall barcoding success.

The present study has established that DNA barcoding delivers approximately 80% species resolution for plant communities in the temperate regions of Canada when either matK or ITS2 are employed, meaning that DNA barcoding can provide a standardized, rapid approach for ecological surveys in these settings. The same gene regions deliver near-perfect resolution to a generic level, a level of taxonomic placement useful for characterizing both past [73; 74] and present [11; 75] plant communities, for forensic applications [76; 77; 78], for validating the accuracy of specimen identifications in herbaria [28; 79], and for assessing herbivore diets [11; 80; 81; 82; 83]. The present study also demonstrates the ability of DNA barcoding to
deliver particularly high levels of taxonomic resolution when comprehensive reference libraries are available for matK and ITS2, providing motivation for efforts to extend coverage for these genes.

Conclusions
Comprehensive sampling (~96% taxonomic coverage) of the Canadian flora provided a unique opportunity to test the efficacy of DNA barcoding across a diverse set of communities. Analyses based on this library indicate that any one of the three barcode regions is very effective (>90%) in delivering a generic assignment while species resolution is often possible with ITS2 (72%) and matK (80%). BLAST demonstrated higher performance than mothur in assigning specimens to a species in all datasets, including those at a community level and for 1074 species with data for all three barcode regions. The higher performance of BLAST reflects its consideration of indel variation and absolute length of the marker, leading matK to deliver the highest resolution. Although ITS2 showed slightly lower performance, it has two important advantages; its short length makes it suitable for HTS-based applications, and it is readily recovered from diverse taxa, including vascular plants and fungi.

Supporting Information
S1 Fig. MNTD values for the three barcodes for genera. Comparison of MNTD values for the three barcode regions for genera of Canadian vascular plants. A) Three-dimensional scatter plot of 243 genera; B) Three-dimensional scatter plot of a subset of 171 genera with low MNTD values. The r² is less than 0.007 for all comparisons. (PDF)

S2 Fig. The percentage of congeners for the six most species-rich families with low MNTD. The percentage of congeners for the six most species-rich families with low MNTD by barcode and region. (PDF)

S1 File. A supplemental file containing four tables. Raw MPD and MNTD values for the vascular plant families of Canada (Table A). Raw MPD and MNTD values for the vascular plant genera of Canada (Table B). The biogeographic region, number of families, MPD, and MNTD for the 28 Canadian localities employed as a basis to test barcode resolution (Table C). The mean taxonomic resolution to family, genus, and species for all 28 localities employed as a basis to test barcode resolution and for the subset of 1074 species with sequence data for all three barcode regions (Table D). (XLSX)

Acknowledgments
We thank two reviewers for helpful comments on an earlier version of this paper. We extend special gratitude to David Erikson and Ina Anreiter for friendly suggestions and comments during the preparation of our manuscript. We also thank Bruce Bennet and Sergei Ponomarenko for providing updated species lists for several national parks. This research was funded through the International Barcode of Life project supported by the Genome Canada through Ontario Genomics, Canada Foundation for Innovation, and by the Ontario Ministry for Research and Innovation. It is also a contribution to the Food From Thought research program supported by the Canada First Research Excellence Fund. The funders
had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author Contributions

Conceptualization: TWAB MLK PDNH EVZ.

Data curation: TWAB MLK JS.

Formal analysis: TWAB.

Funding acquisition: PDNH.

Investigation: TWAB MLK.

Methodology: TWAB MLK.

Supervision: EVZ PDNH.

Validation: TWAB MLK JS.

Visualization: TWAB MLK.

Writing – original draft: TWAB MLK.

Writing – review & editing: TWAB MLK PDNH EVZ.

References


42. R Development Core Team. R: A language and environment for statistical computing. R Foundation, Vienna, Austria. 2008


52. Naciri Y, Linder P. Species identification and delimitation: the dance of the seven veils. Taxon. 2015; 64: 3–16


