Plant Systematics and Evolution (2019) 305:319–324
https://doi.org/10.1007/s00606-019-01568-4

SHORT COMMUNICATION

The effects of sample age and taxonomic origin on the success rate of DNA barcoding when using herbarium material

Helena Korpelainen[1] · Maria Pietiläinen[1]

Received: 26 July 2018 / Accepted: 5 February 2019 / Published online: 26 February 2019

Abstract

We have produced DNA barcodes for Finnish plant taxa. In this study, we specifically report the barcoding success for herbarium materials varying widely in age, also paying attention on success rate variation and genetic distances among different plant families. Additionally, we investigated whether the level of intraspecific variation differs between native and introduced species. The specimens had been collected between years 1867 and 2013. Among all studied specimens, the average success rates for any barcode (*mat*K or *rbc*L), *rbc*L, *mat*K and both barcodes equaled 81, 79, 55 and 53%, respectively, and among species (at least one specimen per species barcoded successfully) 95, 95, 74 and 73%, respectively. We found significant age effects on the barcoding success, the greatest decline being visible in over 100-year-old samples. Plant families showed differences in overall success rates and sample age effects, as well as in intraspecific and interspecific variation levels, while the average level of intraspecific variation appeared similar among native and introduced species. Besides being valuable for the identification of species, DNA barcoding with sufficient sampling is also a tool to investigate specific evolutionary questions, such as biogeographic patterns or the adaptive capacity of invasive and other alien plant species.

Keywords Angiosperms · DNA barcoding · Genetic distances · Herbarium collections · Native and introduced species · Sample age effect

Introduction

The correct identification of any biological species is important in many situations, e.g., biodiversity analyses, monitoring of endangered species, control actions of weeds, pests and invasive species, and studies on the impact of climate change on species distribution. However, this is not always simple or even possible based on morphology alone. Therefore, precise DNA sequencing-based methods of identification have become increasingly common, especially after the development of the idea of DNA barcoding (Hebert et al. 2003), i.e., the use of short DNA sequences from a standardized position in the genome. The practical use of DNA barcodes requires reference databases, namely compiled public libraries of sequences linked to named specimens. There has been considerable international effort to build databases, many of them being region- or taxonomic group-based. The global International Barcode of Life Project (iBOL; www.ibol.org) is a large biodiversity genomics initiative that paves the way for a global digital identification system for life.

The biological materials used for compiling reference libraries for DNA barcodes are either fresh or specimens available in museums, herbaria, botanical gardens and zoos, or in other ex situ conservation collections (von Cräutlein et al. 2011). For plant barcoding, herbarium collections are a major source of materials. It has been well documented that DNA is often well preserved in herbarium samples and, therefore, can be used for DNA analyses, such as sequencing in molecular systematic and phylogenetic studies, and DNA barcoding (e.g., Lehtonen and Christenhusz 2010; von Cräutlein et al. 2011; Xu et a. 2015; Kuzmina et al. 2017).

In this study, we developed *rbc*L and *mat*K barcodes for angiosperm species collected in Finland during a period of

Handling Editor: Alexandra Nora Muellner-Riehl.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00606-019-01568-4) contains supplementary material, which is available to authorized users.

✉ Helena Korpelainen
helena.korpelainen@helsinki.fi

[1] Department of Agricultural Sciences, Viikki Plant Science Centre, University of Helsinki, PO Box 27, 00014 Helsinki, Finland

well over 100 years and maintained in herbarium collections. This work is a part of the Finnish Barcode of Life (FinBOL; www.finbol.org) initiative. Besides producing barcodes for Finnish taxa, we specifically report the barcoding success for herbarium materials varying widely in age, also paying attention on success rate variation and genetic distances among different plant families. An additional attempt was to reveal, whether the level of intraspecific variation differs between native and introduced species, the hypothesis being that the non-native species contain less variation due to their possibly sporadic introductions and short evolutionary history in the area.

Materials and methods

Sampling and barcoding

Angiosperm plant materials used for DNA barcoding and included in the present study consisted of herbarium samples originating from the collections of the Botanical Museum (H), Finnish Museum of Natural History, University of Helsinki. Sampling was conducted during years 2012–2016, including 3176 specimens, 1068 species, 456 genera and 77 families. All included specimens had been collected in Finland between years 1867 and 2013, the mean age equaling 39 years at the time of barcoding (conducted in 2012–2016). The number of very old samples (collected between 1867 and 1910) was only 59 (1.8%). Taxonomic assignments and geographic information were obtained from herbarium voucher labels. All specimens were photographed. It was also recorded whether a species was native (including archaeophytes) or introduced (definitions following Hämet-Ahti et al. 1998). All plant samples were shipped to the Canadian Centre for DNA Barcoding for DNA extraction, PCR and sequencing of chloroplast *mat*K and *rbc*L barcodes using standard plant barcoding protocols (Ivanova et al. 2008; Kuzmina and Ivanova 2011; Kuzmina et al. 2017). All barcodes, images and other sample informations are available in BOLD (http://www.boldsystems.org) under the FinBOL-Plants (FBPL) project (https://doi.org/10.5883/ds-fbpl1).

Data analysis

First, all sequences were subjected to genus-level similarity search against the National Center for Biotechnology Information (NCBI) database using BLASTn to confirm the assignment of taxonomic identities. Matches were found for all sequences. Then, the barcoding success of samples, i.e., barcode compliance (sequences > 500 bp long, *mat*K or *rbc*L or both), was calculated for all specimens and species, and for each family and sample age group. Final age comparisons were conducted for the whole data set and for the families with greatest sample sizes, including Asteraceae, Caryophyllaceae, Cyperaceae, Lamiaceae, Plantaginaceae, Poaceae, Ranunculaceae and Rosaceae. Age-specific barcoding success rates among age groups and among the eight above-mentioned families were compared with Chi-square tests (Fisher 1922).

Based on barcode-compliant *mat*K and *rbc*L sequences, intra- and interspecific genetic distances were calculated. The matrices of pairwise distances were generated using MEGA6 (Tamura et al. 2013). The MEGA output files were used as input files for the ExcaliBAR software to produce intra- and interspecific pairwise genetic distances (Aliabadian et al. 2014). Average intraspecific and interspecific genetic distances based on *mat*K and *rbc*L, separately for each marker, were produced for all samples and separately for families with numbers of samples over 12. Intraspecific genetic distances were generated separately also for native (including archaeophytes) and introduced species, but only for those genera that contained both native and introduced species. Diversity values were compared using *t* tests (Student 1908).

Results and discussion

The barcoding success rates are shown in Fig. 1 for all data and in Online Resource 1 for those families that contained at least 12 samples (41 out of 77 families). The average success rates for any barcode (*mat*K or *rbc*L), *rbc*L, *mat*K and both barcodes among all specimens equaled 81, 79, 55 and 53%, respectively, and among species (at least one specimen per species barcoded successfully) 95, 95, 74 and 73%, respectively. Success rates for *rbc*L varied from 33% in Nymphaeaceae specimens to 100% in Alismataceae, Betulaceae, Papaveraceae and Rubiaceae, while for *mat*K success rates ranged from 0% in Boraginaceae, Crassulaceae, Geraniaceae, Liliaceae, Onagraceae, Orobanchaceae, Pinaceae

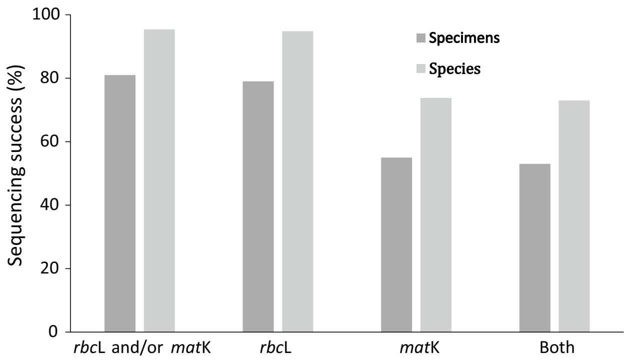

Fig. 1 Barcoding success rates for any barcode (*rbc*L and/or *mat*K), *rbc*L, *mat*K and both of them among all specimens and species

and Saxifragaceae specimens to 100% in Gentianaceae and Papaveraceae. Only in Papaveraceae, the success rate was 100% for both barcoding regions. Previously, Kuzmina et al. (2017) have observed great differences in barcoding success among plant families, especially concerning the *mat*K barcode, where the used primers are not fully universal.

Among very old samples, 17 were collected between years 1867 and 1899. Only six of them (35%) provided barcode-compliant results, the oldest one being from the year 1880 (*Fumaria officinalis*, Papaveraceae). Overall, 37 samples were collected between years 1900–1909 and 36 samples between years 1910 and 1919. Among them, 15 (41%) and 24 samples (67%), respectively, gave barcode-compliant results. Figure 2 shows age-specific success rates for *rbc*L, *mat*K and both in the whole data set and separately for eight families. Since the oldest age groups had only small numbers of samples, old materials formed one group, i.e., collecting before the year 1931. The success rates were tested by a Chi-square test only for *rbc*L results, because the amplification of *mat*K was poorer even in young material and strongly dependent on factors other than age, largely due to the low performance of standard primers in many plant groups. Among the nine age groups (collecting < 1931, 1931–1940, 1941–1950, 1951–1960, 1961–1970, 1971–1980, 1981–1990, 1991–2000, 2001), the results differed significantly ($P < 0.001$). The *rbc*L barcoding success increased steadily from 51.4 to 88.0%, the biggest change being between the age groups 1931–1940 and 1941–1950 (success rates 56.6% and 68.3%, respectively).

When comparing *rbc*L barcoding success statistically in eight individual families, samples were combined into three age groups (< 1950, 1951–1990 and 1991) due to a limited amount of data per family. The Chi-square testing was not possible for two families, Lamiaceae and Poaceae. Among other six tested families, five families showed significant age differences (i.e., a negative age effect), but Plantaginaceae was an exception without any significant age effect. When testing barcoding success across families in each of the three age groups, significant differences among families were found in the age group < 1950 ($P > 0.001$, Lamiaceae and Poaceae excluded) and 1951–90 ($P < 0.001$, all eight families), while in the youngest age group 1991- (Lamiaceae and Poaceae excluded), no significant differences among families were detected. Ranunculaceae and Caryophyllaceae were the families with the poorest barcoding performance among old samples. Previously, Kuzmina et al. (2017) have also observed difference in specimen age effects among plant families.

Herbarium samples are commonly used for studies on molecular systematics or for the generation of DNA barcoding libraries. For instance, Lehtonen and Christenhusz (2010) have used historical herbarium specimens to study the molecular systematics of the fern genus *Lindsaea* using two chloroplast loci (*trnL-trnF* and *trnS*^{GGA}-*rps4*). The age of the samples ranged from 4 to 172 years, and the total success rate was 57%. For *Lindsaea*, the specimens age was found to be of little importance for sequencing success when less than 75 years, while among older samples the sequencing success reduced considerably. In our study, the greatest decline in the sequencing success occurred for over 100-year-old specimens. Although fresh specimens generally provide a uniformly good sequencing results (H. Korpelainen, pers. obs.) and herbarium samples quite good results (e.g., the present study; Kuzmina et al. 2017), some investigations have reported low success rates for herbarium samples. For instance, in a small-scale investigation, Enan et al. (2017) have reported that the success rates for *mat*K and *rbc*L barcodes were 90% and 90% for fresh samples, but 40% and 50% for herbarium samples, respectively.

Aging or inappropriately preserved herbarium specimens lead to complications in DNA analyses, such as DNA fragmentation and other quality issues (see, Xu et al. 2015). In addition, the taxonomic origin affects barcoding success (e.g., the present study; Kuzmina et al. 2017). Kuzmina et al. (2017) have hypothesized that in certain plant families, DNA degradation occurs soon after sample collection owing to the presence of compounds that degrade DNA or irreversibly bind to it. One way to solve obstacles caused by poor DNA qualities would be to use improved techniques, for instance, DNA reconstruction (Xu et al. 2015) that enables the amplification of sufficiently long fragments for DNA barcoding and other purposes even for otherwise failing specimens.

Table 1 and Online Resource 2 give the values for intraspecific and interspecific genetic distances across all data and in different plant families. Intraspecific distances per family ranged from 0 to 0.0062 (average 0.0005) and from 0 to 0.0048 (average 0.0007) based on *rbc*L and *mat*K barcodes, respectively, while interspecific distances ranged from 0 to 0.0216 (average 0.0064) and from 0.0019 to 0.0497 (average 0.0150) based on *rbc*L and *mat*K, respectively. Mean intraspecific distances were significantly lower than interspecific distances for both *rbc*L and *mat*K barcodes ($P < 0.001$). The presence of intraspecific variation means that DNA barcodes can potentially be used as a tool for inferring biogeographic patterns, as also suggested by Costion et al. (2016). However, sampling should be sufficiently comprehensive to reveal geographic patterns in molecular variation. Increasingly, common and cost-efficient whole-chloroplast genome sequencing for plant DNA barcoding purposes will not only provide more accurate identification (Li et al. 2015; Coissac et al. 2016; Zhang et al. 2017) but also improve insights into biogeographic variation patterns.

Both native and introduced species were present in 50 genera, including 276 native species and 118 introduced species (2–3 successfully barcoded specimens per species). Intraspecific genetic distances based on *rbc*L equaled

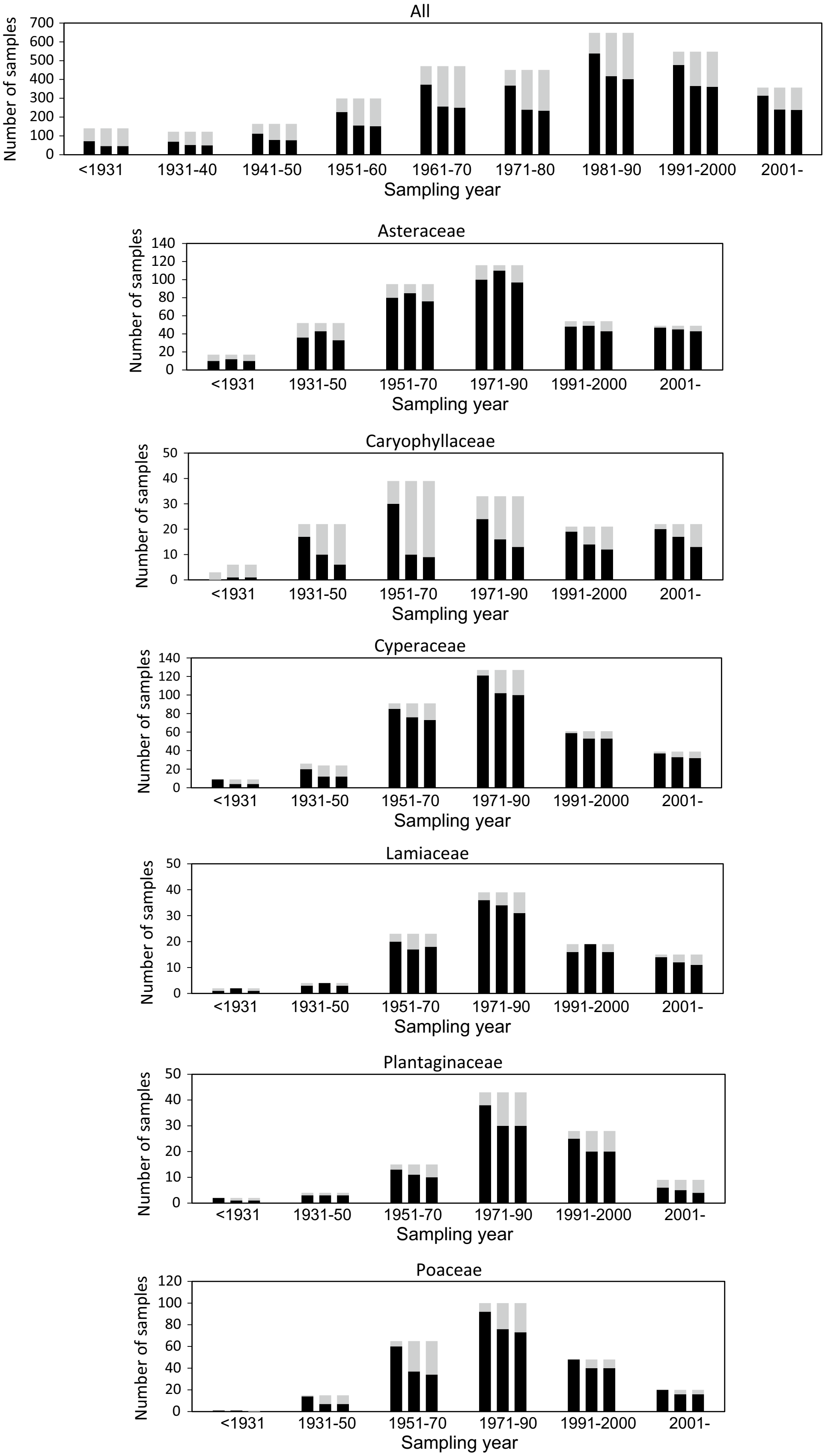

Fig. 2 Barcoding success rates for *rbc*L, *mat*K and both of them in each age group for the whole data set and separately for eight plant families. Barcoding was conducted in 2012–2016. Black columns, the proportion of successful barcodes; gray column, the proportion of unsuccessful barcodes

Fig. 2 (continued)

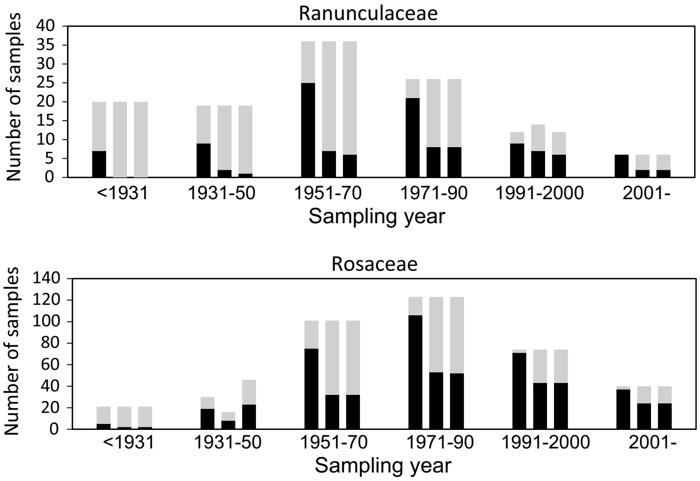

Table 1 Minimum, maximum and average pairwise intraspecific and interspecific (intra-generic) genetic distances based on DNA barcodes (*rbc*L and *mat*K)

Family	Average genetic distance			
	Intraspecific (*rbc*L)	Interspecific (*rbc*L)	Intraspecific (*mat*K)	Interspecific (*mat*K)
(a) Across families with numbers of samples > 12				
Minimum	0	0	0	0.0019
Maximum	0.0062	0.0216	0.0048	0.0497
Average	0.0005 ± 0.0011	0.0064 ± 0.0053	0.0007 ± 0.0010	0.0150 ± 0.0133
(b) Across all data				
Minimum	0	0	0	0
Maximum	0.0109	0.0327	0.0073	0.0576
Average	0.0013 ± 0.0089	0.0081 ± 0.0118	0.0015 ± 0.0051	0.0128 ± 0.0161

The analysis was based on (a) families with numbers of samples > 12 and (b) all data

0.0021 ± 0.0018 and 0.0016 ± 0.0042 in native and non-native species, respectively. Comparable values based on *mat*K equaled 0.0007 ± 0.0013 and 0.0018 ± 0.0038, respectively. The differences were nonsignificant. Previously, Dlugosch and Parker (2008) have provided evidence that introduced populations possess lower levels of neutral genetic diversity than their conspecific native-range populations. On the other hand, Oduor et al. (2016) have discovered that invasive and native plant species do not differ consistently in the extent and frequency of local adaptation, which supports the view that rapid post-introduction adaptive evolution may enable invasive plant species to persist and expand their ecological niche in introduced ranges. Then again, Bock et al. (2015) have emphasized that numerous unknowns remain in invasion genetics, such as the sources of genetic variation, the role of so-called expansion load, and the relative importance of propagule pressure versus genetic diversity for successful establishment. Our comparison of intraspecific variation on partial chloroplast genes (DNA barcodes *rbc*L and *mat*K) conducted for native and alien plants (most not considered invasive) showed that there is no significant difference in the genetic variation between these two groups. It remains unresolved, whether this may be due to multiple introductions and recombination or rapid adaptive evolution among alien plants. However, our sampling is not sufficient to allow definite conclusions on the patterns of intraspecific variation in native and alien plants. Yet, DNA barcoding with sufficient sampling is a tool to investigate specific evolutionary questions, such as the adaptive capacity of invasive and other alien plants or biogeographic patterns.

Acknowledgements Open access funding provided by University of Helsinki. We thank the iBOL and FinBOL organizations for funding, and Marko Mutanen and the staff of the Canadian Centre for DNA Barcoding for many kinds of help during the project.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Information on electronic supplementary material

All collecting and sequence data are available on the Barcode of Life Data System (BOLD) under the project "FinBOL-Plants" (https://doi.org/10.5883/ds-fbpl1).

Online resource 1. Success rates (%) for DNA barcodes (*rbc*L and *mat*K) in different plant families.
Online resource 2. Average pairwise intra-specific and inter-specific (intra-generic) genetic distances based on DNA barcodes (*rbc*L and *mat*K) in different plant families.

References

Aliabadian M, Nijman V, Mahmoudi A, Naderi M, Vonk R, Vences M (2014) ExcaliBAR: a simple and fast software utility to calculate intra- and interspecific distances from DNA barcodes. Contr Zool 83:79–83

Bock DG, Caseys C, Cousens RD, Hahn MA, Heredia SM, Hübner S, Turner KG, Whitney KD, Rieseberg LH (2015) What we still don't know about invasion genetics. Molec Ecol 24:2277–2297. https://doi.org/10.1111/mec.13032

Coissac E, Hollingsworth PM, Lavergne S, Taberlet P (2016) From barcodes to genomes: extending the concept of DNA barcoding. Molec Ecol 25:1423–1428. https://doi.org/10.1111/mec.13549

Costion CM, Lowe AJ, Rossetto M, Kooyman RM, Breed MF, Ford A, Crayn DM (2016) Building a plant DNA barcode reference library for a diverse tropical flora: an example from Queensland, Australia. Diversity 8:5. https://doi.org/10.3390/d8010005

Dlugosch KM, Parker IM (2008) Founding events in species invasions: genetic variation, adaptive evolution, and the role of multiple introductions. Molec Ecol 17:431–449. https://doi.org/10.1111/j.1365-294X.2007.03538.x

Enan MR, Palakkott AR, Ksiksi TS (2017) DNA barcoding of selected UAE medicinal plant species: a comparative assessment of herbarium and fresh samples. Physiol Molec Biol Pl 23:221–227

Fisher RA (1922) On the interpretation of Chi squared from contingency tables, and the calculation of P. J Roy Statist Soc 85:87–94. https://doi.org/10.2307/2340521.JSTOR2340521

Hämet-Ahti L, Suominen J, Ulvinen T, Uotila P (eds) (1998) Retkeilykasvio (Field Flora of Finland), 4th edn. Finnish Museum of Natural History, Botanical Museum, Helsinki

Hebert PDN, Cywinska A, Ball SL, de Waard JR (2003) Biological identification through DNA barcodes. Proc Roy Soc B Biol Sci 270:313–321. https://doi.org/10.1098/rspb.2002.2218

Ivanova NV, Fazekas AJ, Hebert PDN (2008) Semiautomated, membrane-based protocol for DNA isolation from plants. Pl Molec Biol Rep 26:186–198. https://doi.org/10.1007/s11105-008-0029-4

Kuzmina M, Ivanova N (2011) CCDB protocols: PCR amplification for plants and fungi. http://ccdb.ca/site/wp-content/uploads/2016/09/CCDB_Amplification-Plants.pdf

Kuzmina ML, Braukmann TWA, Fazekas AJ, Graham SW, Dewaard SL, Rodrigues A, Bennett BA, Dickinson TA, Saarela JM, Catling PM, Newmaster SG, Percy DM, Fenneman E, Lauron-Moreau A, Ford B, Gillespie L, Subramanyam R, Whitton J, Jennings L, Metsger D, Warne CP, Brown A, Sears E, Dewaard JR, Zakharov EV, Hebert PDN (2017) Using herbarium-derived DNAs to assemble a large-scale DNA barcode library for the vascular-plants of Canada. Appl Pl Sci 5:1700079. https://doi.org/10.3732/apps.1700079

Lehtonen S, Christenhusz MJM (2010) Historical herbarium specimens in plant molecular systematics: an example from the fern genus *Lindsaea* (Lindsaeaceae). Biologia (Bratislava) 65:204–208. https://doi.org/10.2478/s11756-010-0008-8

Li X, Yang Y, Henry RJ, Rossetto M, Wang Y, Chen S (2015) Plant DNA barcoding: from gene to genome. Biol Rev 90:157–166. https://doi.org/10.1111/brv.12104

Oduor AMO, Leimu R, van Kleunen M (2016) Invasive plant species are locally adapted just as frequently and at least as strongly as native plant species. J Ecol 104:957–968. https://doi.org/10.1111/1365-2745.12578

Student (1908) The probable error of a mean. Biometrika 6:1–25. https://doi.org/10.2307/2331554

Tamura K, Stecher G, Peterson D, Filipski A, Kumar S (2013) MEGA6: molecular evolutionary genetics analysis version 6.0. Molec Biol Evol 30:2725–2729. https://doi.org/10.1093/molbev/mst197

von Cräutlein M, Korpelainen H, Pietiläinen M, Rikkinen J (2011) DNA barcoding: a tool for improved taxon identification and management of species diversity. Biodivers Conservation 20:373–389. https://doi.org/10.1007/s10531-010-9964-0

Xu C, Dong W, Shi S, Cheng T, Li C, Liu Y, Wu P, Wu H, Gao P, Zhou S (2015) Accelerating plant DNA barcode reference library construction using herbarium specimens: improved experimental techniques. Molec Ecol Res 15:1366–1374. https://doi.org/10.1111/1755-0998.12413

Zhang N, Ramachandran P, Wen J, Duke JA, Metzman H, McLaughlin W, Ottesen AR, Timme RE, Handy SM (2017) Development of a reference standard library of chloroplast genome sequences, GenomeTrakrCP. Pl Med (Stuttgart) 83:1420–1430. https://doi.org/10.1055/s-0043-113449