Journal of Plant Research
https://doi.org/10.1007/s10265-021-01265-9

REGULAR PAPER – TAXONOMY/PHYLOGENETICS/EVOLUTIONARY BIOLOGY

Using near-infrared spectroscopy to discriminate closely related species: a case study of neotropical ferns

Darlem Nikerlly Amaral Paiva[1] · **Ricardo de Oliveira Perdiz**[2] · **Thaís Elias Almeida**[1]

Received: 23 October 2020 / Accepted: 7 February 2021

Abstract
Identifying plant species requires considerable knowledge and can be difficult without complete specimens. Fourier-transform near-infrared spectroscopy (FT-NIR) is an effective technique for discriminating plant species, especially angiosperms. However, its efficacy has never been tested on ferns. Here we tested the accuracy of FT-NIR at discriminating species of the genus *Microgramma*. We obtained 16 spectral readings per individual from the adaxial and abaxial surfaces of 100 specimens belonging to 13 species. The analyses included all 1557 spectral variables. We tested different datasets (adaxial + abaxial, adaxial, and abaxial) to compare the correct identification of species through the construction of discriminant models (Linear discriminant analysis and partial least squares discriminant analysis) and cross-validation techniques (leave-one-out, K-fold). All analyses recovered an overall high percentage (> 90%) of correct predictions of specimen identifications for all datasets, regardless of the model or cross-validation used. On average, there was > 95% accuracy when using partial least squares discriminant analysis and both cross-validations. Our results show the high predictive power of FT-NIR at correctly discriminating fern species when using leaves of dried herbarium specimens. The technique is sensitive enough to reflect species delimitation problems and possible hybridization, and it has the potential of helping better delimit and identify fern species.

Keywords Barcoding · Discrimination of plant species · FT-NIR · Integrative taxonomy · Metabolomics · *Microgramma*

Introduction

Defining and identifying species using qualitative morphological traits can be challenging even though species identification is fundamental to some areas of science and sustainable dynamics (Galtier 2018; Pinheiro et al. 2018). Correct identifications also contribute significantly to understanding the evolutionary history of many species and the diversity of biological groups in rich and threatened areas, such as tropical forests (Costello 2015). Considering the biological and historical diversity of polymorphisms in plants, allied with centuries of describing species using alpha taxonomy tools, the correct identification of a specimen requires experts with considerable knowledge (Ahrends et al. 2011; Lacerda and Nimmo 2010; Richard and Evans 2006).

A problem when identifying plant species is the absence of complete specimens, including both sterile and fertile material, such as flowers or fruits of seed plants (Gomes et al. 2013). Difficult to access and insufficient or unrepresentative collections of species widely distributed in highly diverse areas can also pose a problem when identifying specimens (Lacerda and Nimmo 2010). Among the traditional identification methods used for plants, keys stand out and are widely employed (Smith 2017). However, polymorphisms and the complexity of shapes, associated with homoplasies and cryptic taxa, for example, create the need for more elaborate tools aimed primarily at the identification, conservation, and elucidation of unclear relationships of plants (Durgante et al. 2013; Pinheiro et al. 2018).

In addition to the use of macromorphology, DNA barcoding is an internationally recognized tool and widely used

✉ Darlem Nikerlly Amaral Paiva
nikerllyjc@hotmail.com; paivadarlem@gmail.com

Ricardo de Oliveira Perdiz
ricoperdiz@gmail.com

Thaís Elias Almeida
blotiella@gmail.com

[1] Universidade Federal do Oeste do Pará, Programa de Pós-graduação em Biodiversidade, Rua Vera Paz, s/n (Unidade Tapajós) Bairro Salé, Santarém, PA 68040-255, Brazil

[2] Instituto Nacional de Pesquisas da Amazônia, Programa de Pós-graduação em Ciências Biológicas, Avenida André Araújo, Manaus, AM 293669060-001, Brazil

Published online: 07 April 2021

in species identification, ecological studies, and forensic analyses (Li et al. 2015; Shokralla et al. 2014). In studies of animal groups that used this molecular approach, the technique proved to be highly efficient (e.g., Ohira et al. 2018; Pérez-Losada et al. 2012; Porco et al. 2012). Using DNA barcoding has been less successful at identifying plants compared to animals (Li et al. 2015). According to Fazekas et al. (2012), this is partially due to hybridization, polyploidy, and speciation related to reproductive systems. However, these are not problems common to all plant groups; the success in using DNA barcoding is lineage-dependent (Li et al. 2015). Identifying herbarium specimens using this method is also more difficult compared to fresh material, since DNA extracted from dried specimens is usually fragmented, requiring a greater combination of primers, which increases the chances of incorrect sequencing (Li et al. 2015; Vere et al. 2012). Furthermore, the widespread use of this technique is still limited because of the high cost (Stein et al. 2014).

One of the most promising alternative tools currently used in botanical identification is Fourier transformed near-infrared spectroscopy (FT-NIR) (e.g., Lang et al. 2017; Rodríguez- Fernández et al. 2011). The principle of the technique is to irradiate fractions of biological material (e.g., a dry leaf) in the infrared region. As a result, a set of absorbance values at different wavelengths (the spectra) is defined for the material (Workman and Weyer 2007). The spectra reflect molecular bonds, such as C–H, N–H, S–H or O–H, and are therefore related to biological molecules and the metabolome of the irradiated tissue (Stuart 2005; Asner et al. 2014).

Research using near-infrared spectroscopy to discriminate plant species is gaining more and more attention in plant taxonomy, especially for angiosperms (Durgante et al. 2013; Kim et al. 2004; Krajšek et al. 2008; Lang et al. 2017). The tool has been shown to be more practical and accurate than genetic or morphological methods (Castillo et al. 2008), is capable of consistently discriminating phylogenetic relationships of flowering plant species (Kim et al. 2004) and has been used in different works to aid in species circumscription and identification of several plant groups (Damasco et al. 2019; Durgante et al. 2013; Lang et al. 2017; Meireles et al. 2020; Prata et al. 2018; Shen et al. 2020). However, FT-NIR has not been tested to identify other groups of embryophytes, such as ferns and lycophytes, or bryophytes (Guzmán and Antonio 2020).

Ferns are the second most diverse group of vascular plants, occur from the tundra to tropical forests, and occupy niches from the ground to the canopy (Moran 2008). Due to the absence of flowers, fruits, and seeds, fern identification relies mainly on rhizome, frond, and sorus morphology (Tryon and Tryon 1982). Sporophyte characters such as indument (trichomes and scales), leaf shape, and the structure and arrangement of sori are fundamental elements in the differentiation between species (Christenhusz and Chase 2014; Tryon and Tryon 1982).

Microgramma (Polypodiaceae) comprises ca. 30 species, occurs in the Neotropics and tropical Africa (Almeida 2014), and is monophyletic according to the most recent circumscriptions (Almeida et al. 2021; Salino et al. 2008). The genus exhibits wide morphological variation, especially in the leaves (e.g., it has both monomorphic and dimorphic species), leaf indument, and sorus arrangement (Almeida 2014). Additionally, intraspecific phenotypic variation and interspecific morphological overlap are found in closely related species, and there are species complexes, which may result in misidentifications in the genus (Almeida et al. 2021). Using *Microgramma* as a model, our goal was to test the effectiveness of Fourier-transform near-infrared spectroscopy (FT-NIR) at discriminating and identifying closely related species in a fern lineage.

Methods

Sampling

Dried leaves were selected from specimens at the BHCB, HSTM, and INPA herbaria (acronyms according to Thiers 2020 onwards: http://sweetgum.nybg.org/science/ih/). One hundred specimens belonging to thirteen species of *Microgramma* were analyzed (Table S1): *M. baldwinii* Brade, *M. crispata* (Fée) R.M.Tryon & A.F.Tryon, *M. dictyophylla* (Kunze ex Mett.) de la Sota, *M. geminata* (Schrad.) R.M.Tryon & A.F.Tryon, *M. lindbergii* (Mett. ex Kuhn) de la Sota, *M. lycopodioides* (L.) Copel., *M. megalophylla* (Desv.) de la Sota, *M. percussa* (Cav.) de la Sota, *M. persicariifolia* (Schrad.) C.Presl, *M. reptans* (Cav.) A.R.Sm., *M. squamulosa* (Kaulf.) de la Sota, *M. thurnii* (Baker) R.M.Tryon & Stolze, and *M. vacciniifolia* (Langsd. & Fisch.) Copel. All specimens had their identification confirmed by an expert (senior author). Only species with a minimum of five available specimens, with both fertile and sterile leaves, were selected. When possible, samples with fronds that were severly damaged by insects or with signs of fungi or other epiphylic organisms were avoided. Sixteen spectral readings were obtained for each specimen (when possible), which included four readings, two on the adaxial surface and two on the abaxial surface, of four different leaves. No distinction between fertile and sterile leaves was made.

The acquisition of the spectra lasted 30 s per reading and was taken using a Thermo Nicolet spectrophotometer, FT-NIR Antaris II Method Development System (MDS). The spectral readings consisted of 1,557 leaf absorbance values in the region of 4,000 to 10,000 cm^{-1} (1,000 to 2,500 nm). Each measurement produced by the equipment was an average of 16 readings with a wavelength resolution of 8 cm^{-1}. The equipment was calibrated every 4 h of use. A black body was placed over the frond to prevent light scattering.

Analyses

All analyses were implemented in the statistical program R version 4.0.2 (R Core Team 2020), under the framework provided by R package caret (Kuhn 2008), using the R packages MASS (Venables and Ripley 2002) and pls (Mevik et al. 2019). Three datasets using all FT-NIR spectrum wavelengths were tested to construct the spectral models: data of (i) adaxial + abaxial surfaces, (ii) adaxial surface only, and (iii) abaxial surface only. The datasets were explored using a principal component analysis (PCA). This technique allows the visualization of data of a smaller set of variables but still preserves the maximum information from the original variable set (Hongyu et al. 2016), thus allowing an exploratory analysis of the behavior of the spectra. The results of the PCA were represented in two-dimensional graphs using the first two main components.

To predict species based on spectral data, we used two supervised pattern recognition techniques: linear discriminant analysis (LDA) and partial least squares discriminant analysis (PLS-DA) (Berrueta et al. 2007). The LDA is a technique that discriminates and classifies objects based on previously defined groups (Sharma and Paliwal 2015), where the dependent variables corresponded to the species (categories) and the independent variables represent the absorbance values in the near-infrared. The PLS-DA, which also classifies the samples according to defined categories, is based on finding components that better explain the variations of the variables between classes, giving less weight to the noise and uncorrelated variations (Mevik and Cederkvist 2004). Both models were tested using the three different datasets.

Cross-validation techniques were used to assess model performance and species discrimination. The K-fold validation technique (Burman 1989) is where the set of calibration samples is divided into K subsets, with a subset taken out for validation and the remaining K-1 subsets used to build the model. Thus, at the end of K steps, the data is used in both test subsets and validation (Yadav and Shukla 2016). Here we use K = 10, described as the value that presents the best performance in the sampling, with the least bias in the error rate estimates (Kohavi 1995). The leave-one-out (LOO) technique uses k-1 samples to generate the discriminant function and the sample not included in the model serves to validate it, obtaining the percentage of the model's prediction (Kohavi 1995). Thus, we compared the predictions of individual identities for each species in each of the datasets.

Results

We found considerable variation in the near-infrared spectral data among the sampled species (Fig. 1). Among the three datasets tested, the adaxial + abaxial (i) dataset showed 97.8% of the spectral variation, the adaxial (ii) dataset showed 97.6% of the spectral variation, and the abaxial (iii) dataset was the most representative with a spectral variation of 98.1% (Fig. 2). For the abaxial (iii) dataset, individuals belonging to the same species tended to group more cohesively and consequently less mixed compared to the remaining two datasets (adaxial + abaxial [i] and adaxial [ii]) (Fig. 2).

All datasets had high predictive results in the identification of species (correct predictions higher than 90%) for both models (PLS and LDA) and validation techniques (K-fold and LOO) (Table 1). Among the best percentages for plant discrimination (over 96%) were the LDA model with the (iii) abaxial dataset for both the K-fold

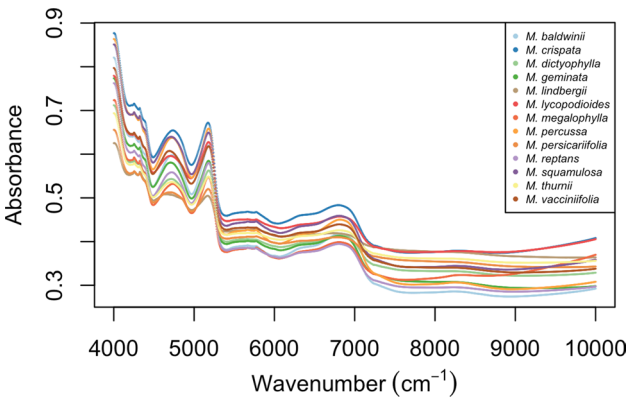
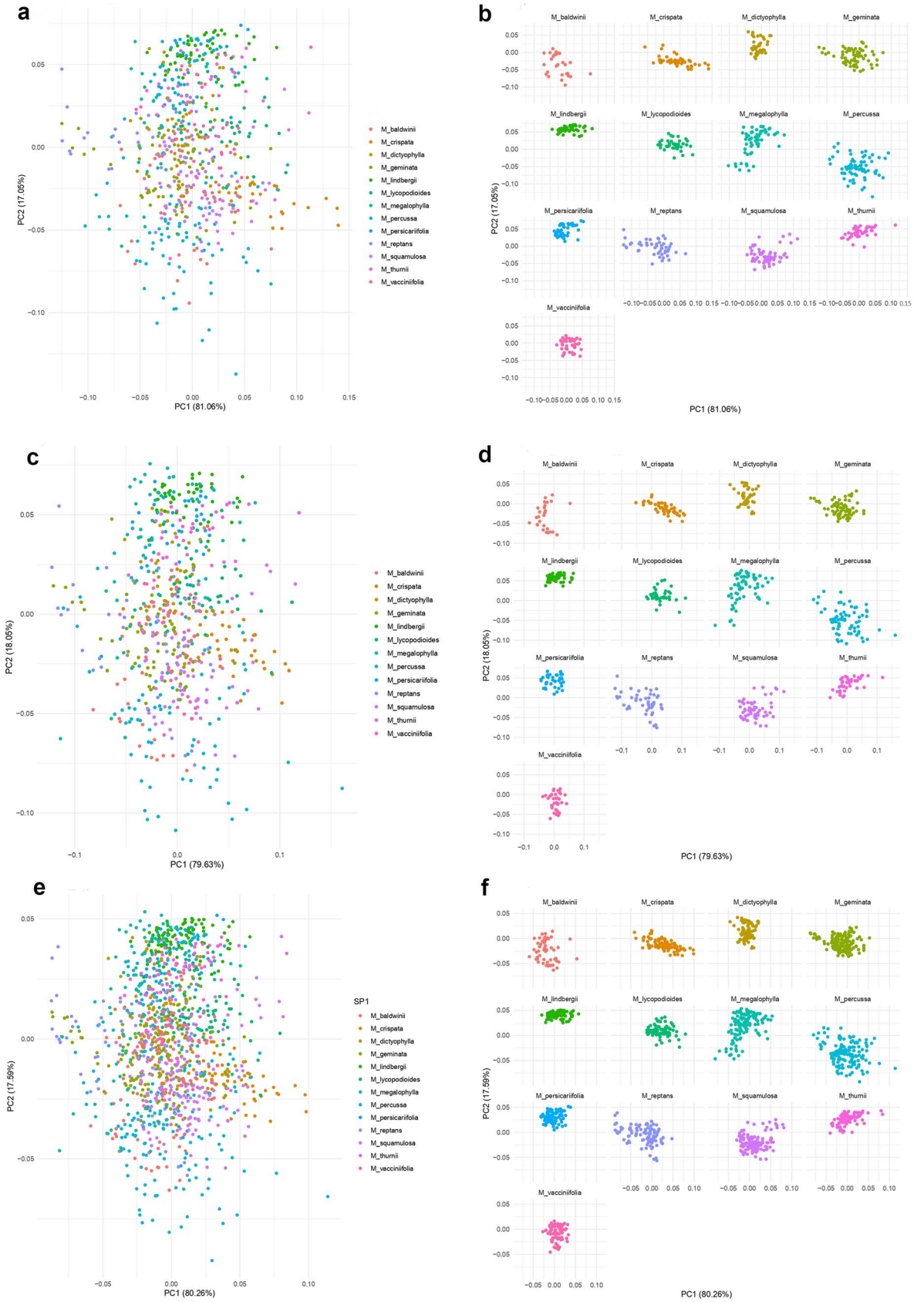


Fig. 1 Average near-infrared spectral data for the thirteen sampled *Microgramma* species


a
PC2 (17.05%)
PC1 (81.06%)
M_baldwinii
M_crispata
M_dictyophylla
M_geminata
M_lindbergii
M_lycopodioides
M_megalophylla
M_percussa
M_persicariifolia
M_reptans
M_squamulosa
M_thurnii
M_vacciniifolia
b
PC2 (17.05%)
PC1 (81.06%)
c
PC2 (18.05%)
PC1 (79.63%)
d
PC2 (18.05%)
PC1 (79.63%)
e
SP1
PC2 (17.59%)
PC1 (80.26%)
f
PC2 (17.59%)
PC1 (80.26%)

◂**Fig. 2** Principal component analysis (PCA) plot of the first two principal component axes for spectral data. **a** Abaxial surface dataset (iii), all species. **b** Abaxial surface dataset (iii), individually represented species. **c** Adaxial surface dataset (ii), all species. **d** Adaxial surface dataset (ii), individually represented species. **e** Adaxial + abaxial surface dataset (i), all species. **f** Adaxial + abaxial surface dataset (i), individually represented species

and leave-one-out validation techniques, and the PLS-DA model with the (i) adaxial + abaxial dataset and leave-one-out validation.

The adaxial + abaxial (i) dataset alone had the best percentage only for the PLS model and leave-one-out validation (96.7%), and both validations in a similar way resulted in elevated correct identifications for the three datasets tested (adaxial + abaxial, adaxial, abaxial).

All individuals of *M. crispata* and *M. megalophylla* were 100% correctly predicted in both models and validations tests, with no confusion of readings with any sample related to any other species (Figs. 3, 4, Table 2). For six species, *M. dictyophylla*, *M. geminata*, *M. lindbergii, M. lycopodioides*, *M. percussa*, and *M. reptans*, the correct prediction of the identities of individuals in all models and validations ranged from 90 to 100% (Figs. 3, 4, Table 2).

Two species (15.3%), *M. persicariifolia* and *M. squamulosa*, had correct predictions between 80 and 100% among the models and validations tests (Figs. 3, 4). For *M. baldwinii*, the abaxial (iii) dataset underperformed in both the LDA and PLS-DA models and validations (Figs. 3, 4, Table 2). Additionally, in the PLS-DA model for this species, the adaxial + abaxial (i) dataset had 88% and 81% correct predictions for the K-fold and LOO validations, respectively. The remaining models and validations recovered 100% correct predictions. For *M. persicariifolia*, the lowest prediction value (83%) was found in the PLS model for the abaxial dataset, for both validations; the remaining models and validations recovered 100% correct predictions. Regarding *M. squamulosa*, only the abaxial (iii) dataset had 100% correct predictions in both models and validations (Figs. 3, 4).

The two species with the lowest percentages of correct predictions were *M. thurnii* and *M. vacciniifolia*. For *M. thurnii*, of the 12 different combinations of the datasets and tests, only two recovered one of the lowest percentages of correct predictions (75%): the PLS model, with the adaxial (ii) dataset, for both validations. Four tests recovered more than 90% correct predictions, and six had 100% correct identifications (Figs. 3, 4, Table 2). For this species, the PLS-DA models underperformed compared to the LDA models.

Microgramma vacciniifolia was the species with the lowest percentage of correct predictions (73%), which was found by the PLS model with the adaxial + abaxial and adaxial datasets; although, for the abaxial dataset there was 100% accuracy in the identifications (data not shown). Prediction errors for *M. vacciniifolia* individuals occurred mainly with spectra associated with *M. geminata* and *M. squamulosa* in both models. Even so, the lowest percentage observed was 73% in the PLS model (Fig. 4).

Discussion

This is the first time that Fourier-transform near-infrared spectroscopy (FT-NIR) was tested for discriminating ferns species. Our results show that FT-NIR is a powerful tool that can be easily applied to species identification using spectral data of leaves. For all different scenarios tested (species, datasets, models, and validations), more than 85% had an accuracy equal or greater than 90% (Figs. 3, 4, Table 2), with an average above 93% (Table 1).

Regarding the accuracy of FT-NIR at species identification, we recognize the importance of using well-defined species circumscriptions as premises and well-identified samples for constructing spectral models. In this work, when incorrectly identified specimens were used (an individual of *M. reptans* incorrectly determined as *M. baldwinii*), the accuracy decreased to 85.5%, and after redoing the analysis with the correct identification, the correct prediction of the individuals of *M. baldwinii* reached 93.4%.

Ferns are characterized by the presence of sori, which are usually on the abaxial surface of the fronds (Christenhusz and Chase 2014). Some lineages exhibit leaf dimorphism, with leaves morphologically and physiologically specialized for photosynthesis or reproduction, and in some cases, there are extreme differences between both types (Wagner and Wagner 1977). In our study, we used species that are both monomorphic (*M. baldwinii, M. dictyophylla*, *M. geminata*, *M. lindbergii*, *M. lycopodioides*, *M. megalophylla*, *M. percussa*, *M. persicariifolia*, and *M. thurnii*) and dimorphic (*M. crispata*, *M. reptans*, *M. squamulosa*, and *M. vacciniifolia*). The dataset for the abaxial leaf surface, which can be more affected by the presence of sori, had (on average) higher percentages of discriminating samples than the other tested datasets (Table 1). Given our results, we believe the presence of sori has minimal influence on the spectral readings and subsequent discrimination power among species. However, this can vary among different lineages, and further tests controlling for fertile and sterile frond spectral readings are recommended.

Table 1 Average percentage of correct identifications using a discriminant analysis and all of the FT-NIR spectrum wavelength data (1000–2500 nm) for the three datasets, (i) adaxial + abaxial leaf surfaces, (ii) adaxial surface only and (iii) abaxial surface only, for both models (LDA = linear discriminant analysis, PLS-DA = partial least squares discriminant analysis) and validation tests (K-fold and leave-one-out)

	LDA		PLS-DA	
Dataset	K-fold	LOO	K-fold	LOO
Adaxial + abaxial	95.3	93.0	96.2	96.7
Adaxial	95.4	95.8	93.9	94.3
Abaxial	96.2	96.2	94.6	95.1

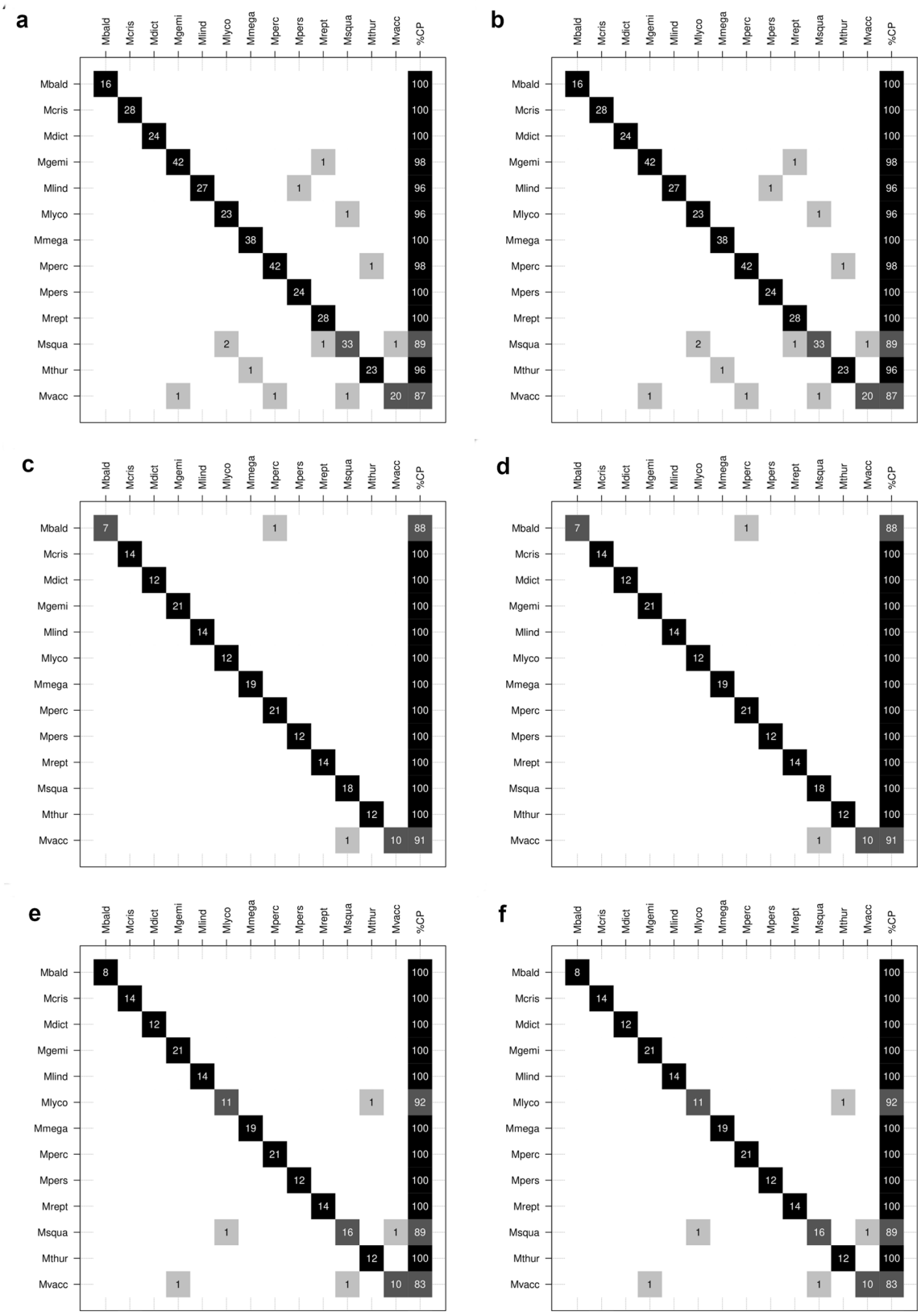
Fig. 3 Confusion matrices resulting from the linear discriminant analysis (LDA) for the LOO and K-fold validations. **a** LOO validation, adaxial + abaxial surface data. **b** K-fold validation, adaxial + abaxial surface data. **c** LOO validation, abaxial surface data only. **d** K-fold validation, abaxial surface data only. **e** LOO validation, adaxial surface data only. **f** K-fold validation, adaxial surface data only. The names of the species observed are in rows and columns. The values on the diagonal correspond to correct predictions and those outside the diagonal correspond to incorrect predictions. *Mbald M. baldwinii*, *Mcris M. crispata*, *Mdict M. dictyophylla*, *Mgemi M. geminata*, *Mlind M. lindbergii*, *Mlyco M. lycopodioides*, *Mmega M. megalophylla*, *Mperc M. percussa*, *Mpers M. persicariifolia*, *Mrept M. reptans*, *Msqua M. squamulosa*, *Mthur M. thurnii*, *Mvacc M. vacciniifolia*

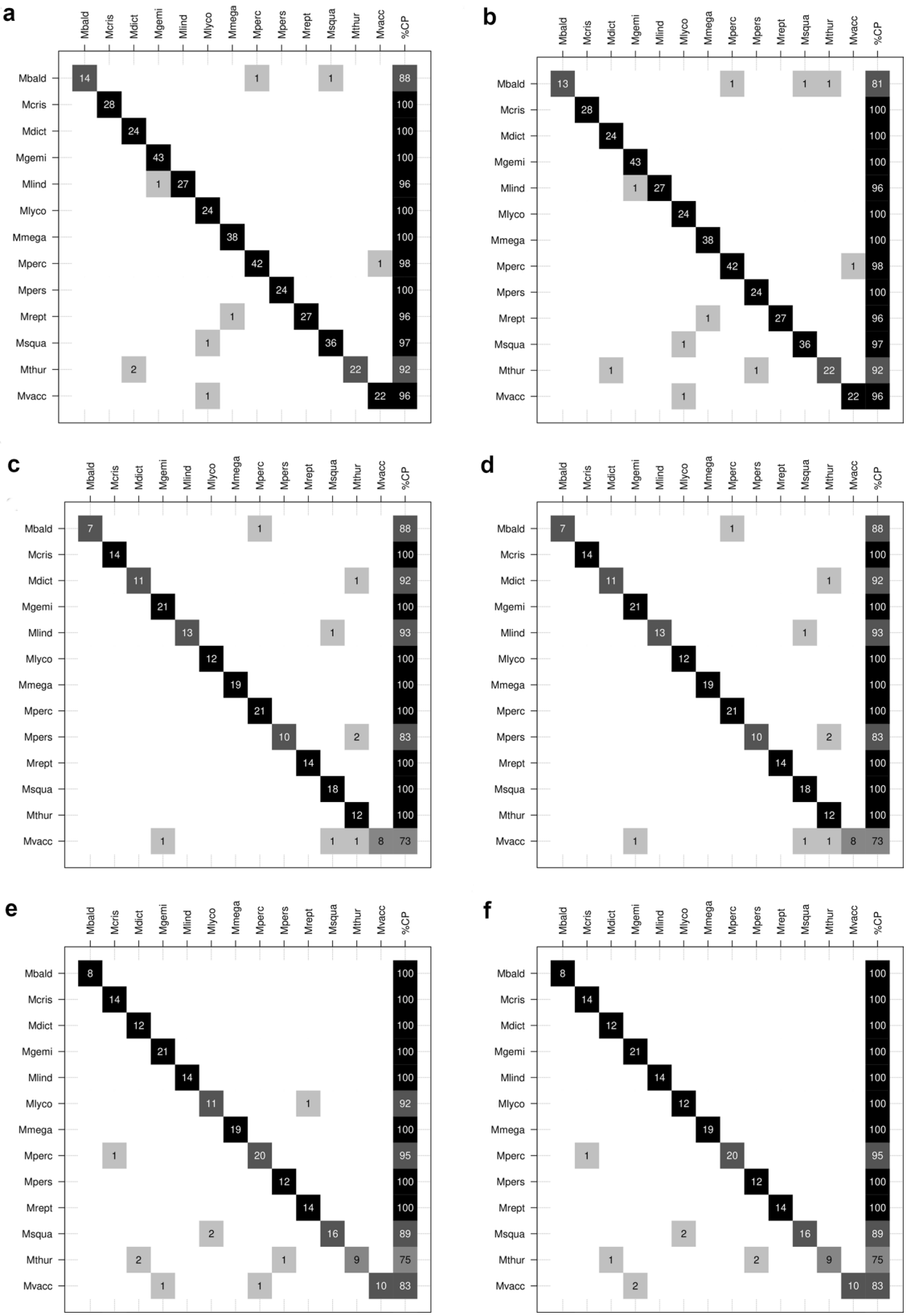
One of the species that was more difficult to discriminate was *M. vacciniifolia*, where ca. 40% of the samples were incorrectly predicted as *M. squamulosa* (Figs. 3, 4, Table 2). These species are sympatric in eastern and central Brazil and exhibit wide morphological variation (Almeida 2014). Our results using near-infrared spectroscopy (NIR) could be revealing inconsistencies in their current taxonomic circumscriptions. Also, the existence of hybrids between these species (Sota 1973) might explain the related spectral readings and the lower percentage of correct predictions. The technique has been shown to detect differences in the physical and biochemical compositions expressed in plant samples, even between closely related species, populations, and hybrids (Atkinson et al. 1997; Cui et al. 2012; Espinoza et al. 2012; Humphreys et al. 2008). Spectral properties have been shown to directly reflect phylogenetic patterns (Asner et al. 2014; Meireles et al. 2020), so it is expected that existing intra-specific variation between individuals and populations can be captured in the spectral model with a broad sampling of phenotypic and geographic variation within species.

Our results show that the best models and validations can, on average, correctly predict the identification of species 96.7% of the time when using all wavelengths to construct the models, which is comparable to previously published taxonomic works. Fan et al. (2010) tested and indicated the reliability of the technique at discriminating *Ephedra* plants from different habitats and collection seasons, while Lang et al. (2017) showed the effectiveness of the technique at discriminating species, genera, and families of tree species from eighteen different angiosperm families. Prata et al. (2018) demonstrated for the first time that near-infrared spectroscopy on leaves of subspecies of the *Pagamea guianensis* complex can discriminate taxa with high precision. For groups of closely related plants, the technique has shown excellent results for species of *Protium* (Burseraceae), confirming the differences in spectral signatures among species (Damasco et al. 2019).

There are many sources of potential variation that could impact the spectral variation in leaves (Asner et al. 2014). The chemical composition and other structural characteristics of leaves vary within and between species, as a result of the developmental stage and a combination of environmental factors, ontogeny, and composition of the plant epidermis (Mediavilla and Escudero 2009; Mediavilla et al. 2014). Young and mature leaves generally exhibit differences in morphology, structure, and biochemical composition. The development process in plants causes changes in leaf structure, and these changes can be expressed in the spectra of the leaf through changes of certain compounds such as pigments, water and an increase in the mass of the secondary cell wall components (Ustin and Jacquemoud 2020). We controlled phenological sources of variation using only fully developed leaves in individuals that already presented fertile leaves. Moreover, to avoid variation coming from the presence of epiphylls, such as bryophytes, lichens, or fungi, or response to herbivory and infections that can also affect chemical composition, we made a rigorous selection of specimens with clean, undamaged leaves. Chemical composition in leaves can vary with soil properties, as already shown for canopy trees (Asner et al. 2014). Despite this variation, FT-NIR was still able to discriminate between species.

The high predictive power of FT-NIR at discriminating fern species, presented here, is superior to that observed for that group using a single region of two combined DNA barcodes, for which the best performance was 75% correct predictions (Li et al. 2011; Wang et al. 2016). Identifying ferns and other plant groups using DNA barcoding is an expensive and time demanding technique and different lineages require specific combinations of molecular markers, which can make this


a
b
c
d
e
f
Mbald
Mcris
Mdict
Mgemi
Mlind
Mlyco
Mmega
Mperc
Mpers
Mrept
Msqua
Mthur
Mvacc
%CP


a
b
c
d
e
f
Mbald
Mcris
Mdict
Mgemi
Mlind
Mlyco
Mmega
Mperc
Mpers
Mrept
Msqua
Mthur
Mvacc
%CP

◂**Fig. 4** Confusion matrices resulting from the Partial Least Squares Discriminant Analysis (PLS) in the LOO and K-fold validations. **a** K-fold validation, adaxial + abaxial surface data. **b** LOO validation, adaxial + abaxial surface data. **c** LOO validation, abaxial surface data only. **d** K-fold validation, abaxial surface data only. **e** K-fold validation, adaxial surface data only. **f** LOO validation, adaxial surface data only. The names of the species observed are in rows and columns. The values on the diagonal correspond to correct predictions and those outside the diagonal correspond to incorrect predictions. *Mbald M. baldwinii*, *Mcris M. crispata*, *Mdict M. dictyophylla*, *Mgemi M. geminata*, *Mlind M. lindbergii*, *Mlyco M. lycopodioides*, *Mmega M. megalophylla*, *Mperc M. percussa*, *Mpers M. persicariifolia*, *Mrept M. reptans*, *Msqua M. squamulosa*, *Mthur M. thurnii*, *Mvacc M. vacciniifolia*

technique complicated (Li et al. 2011; Lima et al. 2018). However, barcoding gametophytes has shown promising results for identifying species of ferns, which represents a great contribution to what is known about the evolution of this group (Schneider and Schuettpelz 2006). Our work does not intend to minimize the importance of other techniques used in plant systematics, but rather tested the reliability and effectiveness of FT-NIR at discriminating species in a group known to be problematic. Further, it highlights the potential of using this method in studies about plant systematics.

Conclusion

Our results show that near-infrared spectroscopy (NIR) is a highly effective, cost-effective, and non-destructive technique that can be used to discriminate closely related species. In addition to the possibility of obtaining spectral data quickly with minimal damage to samples, the technique provides greater reliability at discriminating morphologically similar fern species, as previously found for some angiosperms. The accuracy of the identifications is comparable to and surpasses that of DNA barcoding, even for species from highly diverse and heterogeneous areas, such as tropical forests. We believe that NIR has great potential to be used in integrative taxonomic studies that aim to better understand species circumscriptions in the fern lineage.

Table 2 Summary of the percentage of incorrect predictions per species analyzed among the different datasets, analyses, and validations tested

Dataset	Analysis	Validation	Mbald	Mcris	Mdict	Mgemi	Mlind	Mlyco	Mmega	Mperc	Mpers	Mrept	Msqua	Mthurn	Mvacc
	LDA	LOO	12%	0	0	0	0	0	0	0	0	0	0	0	9%
Abaxial		K-fold	12%	0	0	0	0	0	0	0	0	0	0	0	9%
	PLS	LOO	12%	0	8%	0	7%	0	0	0	17%	0	0	0	27%
		K-fold	12%	0	8%	0	7%	0	0	0	17%	0	0	0	27%
	LDA	LOO	0	0	0	0	0	8%	0	0	0	0	11%	0	17%
Adaxial		K-fold	0	0	0	0	0	8%	0	0	0	0	11%	0	17%
	PLS	LOO	0	0	0	0	0	0	0	5%	0	0	11%	25%	17%
		K-fold	0	0	0	0	0	8%	0	5%	0	0	11%	25%	17%
	LDA	LOO	0	0	0	2%	4%	4%	0	2%	0	0	11%	4%	13%
Both		K-fold	0	0	0	2%	4%	4%	0	2%	0	0	11%	4%	13%
	PLS	LOO	19%	0	0	0	4%	0	0	2%	0	4%	3%	8%	4%
		K-fold	12%	0	0	0	4%	0	0	2%	0	4%	3%	8%	4%

LDA linear discriminant analysis, *PLS* partial least squares discriminant analysis, *Mbald M. baldwinii*, *Mcris M. crispata*, *Mdict M. dictyophylla*, *Mgemi M. geminata*, *Mlind M. lindbergii*, *Mlyco M. lycopodioides*, *Mmega M. megalophylla*, *Mperc M. percussa*, *Mpers M. persicariifolia*, *Mrept M. reptans*, *Msqua M. squamulosa*, *Mthur M. thurnii*, *Mvacc M. vacciniifolia*

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10265-021-01265-9.

Acknowledgements This study was partly financed by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior—Brasil (CAPES)—financing code 001, and Programa Nacional de Cooperação Acadêmica na Amazônia (PROCAD-AM/CAPES 21/2018, no 88887.200472/2018-00). RO Perdiz received a scholarship from Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) (process no. 142243/2015-9). The authors thank the following: B. Leal, L.L. Giacomin, M.A. Buitrago, and T. André for contributing to the manuscript; Mike Hopkins for all the support given to the first author while working at INPA; the herbaria that kindly gave us access to the specimens used in this research; and our colleagues from HSTM and INPA for the support and willingness to help.

Author contributions Conceptualization: DNAP and TEA; methodology: DNAP, ROP, and TEA; formal analysis and investigation: DNAP, ROP, and TEA; writing—original draft preparation: DNAP and TEA; writing—review and editing: DNAP and TEA; funding acquisition: TEA; supervision: TEA.

References

Ahrends A, Rahbek C, Bulling MT, Burgess ND, Platts PJ, Lovett JC, Kindemba VW, Owen N, Sallu AN, Marshall AR, Mhoro BE, Fanning E, Marchant R (2011) Conservation and the botanist effect. Biol Conserv 144:131–140. https://doi.org/10.1016/j.biocon.2010.08.008

Almeida TE (2014) Systematic studies in the genus *Micrograma* C.Presl (Polypodiaceae-Polypodiopsida). Dissertation, Universidade Federal de Minas Gerais

Almeida TE, Salino A, Dubuisson J-Y, Hennequin S (2021) Insights on long-distance dispersal, ecological and morphological evolution in the fern genus *Micrograma* from phylogenetic inferences. Bot J Linn Soc. https://doi.org/10.1093/botlinnean/boaa107 (**in press**)

Asner GP, Martin RE, Carranza-Jimenez L, Sinca F, Tupayachi R, Anderson CB, Martinez P (2014) Functional and biological diversity of foliar spectra in tree canopies throughout the Andes to Amazon region. New Phytol 204:127–139. https://doi.org/10.1111/nph.12895

Atkinson MD, Jervis PA, Sangha SR (1997) Discrimination between *Betula pendula*, *Betula pubescens*, and their hybrids using near-infrared reflectance spectroscopy. Can J Forest Res 26:1896–1900. https://doi.org/10.1139/x97-141

Berrueta LA, Alonso-Salces MR, Héberger K (2007) Supervised pattern recognition in food analysis. J Chromatogr A 1158:196–214. https://doi.org/10.1016/j.chroma.2007.05.024

Burman P (1989) A comparative study of ordinary cross-validation, r-fold cross-validation and the repeated learning-testing methods. Biometrika 76:503–514. https://doi.org/10.2307/2336116

Castillo R, Contreras D, Freer J, Ruiz J, Valenzuela S (2008) Supervised pattern recognition techniques for classification of *Eucalyptus* species from leaves NIR spectra. J Chil Chem Soc 53:1709–1713. https://doi.org/10.4067/S0717-97072008000400016

Christenhusz MJM, Chase MW (2014) Trends and concepts in fern classification. Ann Bot 113:571–594. https://doi.org/10.1093/aob/mct299

Costello MJ (2015) Biodiversity: the known, unknown, and rates of extinction. Curr Biol 25:R368–R371. https://doi.org/10.1016/j.cub.2015.03.051

Cui H-F, Ye Z-H, Xu L, Fu X-S, Fan C-W, Yu X-P (2012) Automatic and rapid discrimination of cotton genotypes by near-infrared spectroscopy and chemometrics. J Anal Methods Chem 2012:793468. https://doi.org/10.1155/2012/793468

Damasco G, Daly DC, Vicentini A, Fine AVP (2019) Reestablishment of *Protium cordatum* (Burseraceae) based on integrative taxonomy. Taxon 68:34–46. https://doi.org/10.1002/tax.12022

de la Sota ER (1973) A new species of *Micrograma* from Argentina. Am Fern J 63:61–64. https://www.jstor.org/stable/1546180

de Lima RAF, Oliveira AA, Colletta GD, Flores TB, Coelho RLG, Dias P, Frey GP, Iribar A, Rodrigues RR, Souza VC, Chave J (2018) Can plant DNA barcoding be implemented in species-rich tropical regions? A perspective from São Paulo State, Brazil. Genet Mol Biol 41:661–670. https://doi.org/10.1590/1678-4685-gmb-2017-0282

Durgante FM, Higuchi N, Almeida A, Vicentini A (2013) Species spectral signature: discriminating closely related plant species in the Amazon with near-infrared leaf-spectroscopy. Forest Ecol Manag 291:240–248. https://doi.org/10.1016/j.foreco.2012.10.045

Espinoza JA, Hodge GR, Dvorak WS (2012) The potential use of near infrared spectroscopy to discriminate between different pine species and their hybrids. J Near Infrared Spectrosc 20:437–447. https://doi.org/10.1255/jnirs.1006

Fan Q, Wang Y, Sun P, Liu S, Li Y (2010) Discrimination of *Ephedra* plants with diffuse reflectance FT-NIRS and multivariate analysis. Talanta 80:1245–1250. https://doi.org/10.1016/j.talanta.2009.09.018

Fazekas JA, Kuzmina LM, Newmaster GS, Hollingsworth MP (2012) DNA barcoding methods for land plants. In: Kress WJ, Erickson, DL (eds) DNA barcodes. Humana Press, Totowa, pp 232–252. https://doi.org/10.1007/978-1-61779-591-6_11

Galtier N (2018) Delineating species in the speciation continuum: a proposal. Evol Appl 12:12657–12663. https://doi.org/10.1111/eva.12748

Gomes ACS, Andrade A, Barreto-Silva JS, Brenes-Arguedas T, Lopez DC, de Freitas CC, Lang C, Oliveira AA, Pérez AJ, Perez R, Silva JB, Silveira AMF, Vaz MC, Vendrami J, Vincentini A (2013) Local plant species delimitation in a highly diverse Amazonian forest: Do we all see the same species? J Veg Sci. https://doi.org/10.1111/j.1654-1103.2012.01441.x

Guzmán Q, Antonio J (2020) Using visible-near-Infrared spectroscopy to classify Lichens at a neotropical dry forest. Ecol Indic 111:105999. https://doi.org/10.1016/j.ecolind.2019.105999

Hongyu G, Sandanielo MLV, Junior OJG (2016) Principal component analysis: theory, interpretations and applications. Eng Sci 1:83–90. https://doi.org/10.18607/ES20165053

Humphreys RJ, Reilly-Wapstra JM, Harbard JL, Davies NW, Griffin AR, Jordan GJ, Potts BM (2008) Discrimination between seedlings of *Eucalyptus globulus*, *E. nitens* and their F1 hybrid using near-infrared reflectance spectroscopy and foliar oil content. Silvae Genet 57:262–269. https://doi.org/10.1515/sg-2008-0040

Kim SW, Ban SH, Chung H, Cho S, Chung HJ, Choi OS, Yoo OJ, Liu JR (2004) Taxonomic discrimination of flowering plants by multivariate analysis of Fourier Transform Infrared Spectroscopy data. Plant Cell Rep 23:246–250. https://doi.org/10.1007/s00299-004-0811-1

Kohavi R (1995) A study of crossvalidation and bootstrap for accuracy estimation and model selection. In: International joint conference on artificial intelligence, pp 456–461

Krajšek SS, Buh P, Zega A, Kreft S (2008) Identification of herbarium whole-leaf samples of *Epilobium* species by ATR-IR

spectroscopy. Chem Biodivers 5:310–317. https://doi.org/10.1002/cbdv.200890028
Kuhn M (2008) Building predictive models in R using the caret package. J Stat Softw 28(5):1–26. https://doi.org/10.18637/jss.v028.i05
Lacerda AEB, Nimmo ER (2010) Can we really manage tropical forests without knowing the species within? Getting back to the basics of forest management through taxonomy. Forest Ecol Manag 259:995–1002. https://doi.org/10.1016/j.foreco.2009.12.005
Lang C, Almeida DRA, Costa FRC (2017) Discrimination of taxonomic identity at species, genus and family levels using Fourier Transformed Near-Infrared Spectroscopy (FT-NIR). Forest Ecol Manag 406:219–227. https://doi.org/10.1016/j.foreco.2017.09.003
Li F-W, Kuo L-Y, Rothfels CJ, Ebihara A, Chiou W-L, Windham MD, Pryer KM (2011) rbcL and matK earn two thumbs up as the core DNA barcode for ferns. PLoS ONE 6:e26597. https://doi.org/10.1371/journal.pone.0026597
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
Mediavilla S, Escudero A (2009) Ontogenetic changes in leaf phenology of two co-occurring Mediterranean oaks differing in leaf life span. Ecol Res 24:1083–1090. https://doi.org/10.1007/s11284-009-0587-4
Mediavilla S, Herranz M, González-Zurdo P, Escudero A (2014) Ontogenetic transition in leaf traits: a new cost associated with the increase in leaf longevity. J Plant Ecol 7:567–575. https://doi.org/10.1093/jpe/rtt059
Meireles JE, Cavender-Bares J, Townsend PA, Ustin S, Gamon JA, Schweiger AK, Schaepman ME, Asner GP, Martin RE, Singh A, Schrodt F, Chlus A, O'Meara BC (2020) Leaf reflectance spectra capture the evolutionary history of seed plants. New Phytol 228:485–493. https://doi.org/10.1111/nph.16771
Mevik BH, Cederkvist RH (2004) Mean squared error of prediction (MSEP) estimates for principal component regression (PCR) and partial least squares regression (PLSR). J Chemometr 18:422–429. https://doi.org/10.1002/cem.887
Mevik BH, Wehrens R, Liland KH (2019) pls: partial least squares and principal component regression. https://CRAN.R-project.org/package=pls. Accessed 5 Apr 2020
Moran RC (2008) Diversity, biogeography, and floristics. In: Ranker TA, Haufler CH (eds) Biology and evolution of ferns and lycophytes. Cambridge University Press, New York, pp 367–394. https://doi.org/10.1017/CBO9780511541827.015
Ohira H, Koichi S, Tadaaki T, Shingo K, Hyeok-Jae C (2018) DNA barcoding suggested the existence of cryptic species and high biodiversity of South Korean Pseudoscorpions (*Arachnida*, Pseudoscorpiones). J Asia-Pac Biod 11:399–407. https://doi.org/10.1016/j.japb.2018.04.005
Pérez-Losada M, Bloch R, Breinholt JW, Pfenninger M, Domínguez J (2012) Taxonomic assessment of Lumbricidae (Oligochaeta) earthworm genera using DNA barcodes. Eur J Soil Biol 48:41–47. https://doi.org/10.1016/j.ejsobi.2011.10.003
Pinheiro F, Dantas-Queiroz MV, Palma-Silva C (2018) Plant species complexes as models to understand speciation and evolution: a review of South American studies. CRC Cr Rev Plant Sci 37:54–80. https://doi.org/10.1080/07352689.2018.1471565
Porco D, Bedos A, Greenslade P, Janion C, Skarzynski D, Stevens ID, van Vuuren J, Deharveng L (2012) Challenging species delimitation in Collembola: cryptic diversity among common springtails unveiled by DNA barcoding. Invertebr Syst 26:470–477. https://doi.org/10.1071/IS12026
Prata EM, Sass C, Rodrigues DP, Domingos FM, Specht CD, Damasco G, Ribas CC, Fine PVA, Vincentini A (2018) Towards integrative taxonomy in Neotropical botany: disentangling the *Pagamea guianensis* species complex (Rubiaceae). Bot J Linn Soc 188:213–231. https://doi.org/10.1093/botlinnean/boy051
R Core Team (2020) R: A language and environment for statistical computing http://www.R-project.org/. R Foundation for Statistical Computing, Vienna
Richard D, Evans D (2006) The need for plant taxonomy in setting priorities for designated areas and conservation management plans: a European perspective. In: Leadlay E, Jury S (eds) Taxonomy and plant conservation. Cambridge University Press, Cambridge, pp 162–176
Rodríguez-Fernández JI, de Carvalho CJB, Pasquini C, de Lima KMG, Moura MO, Arízaga GGC (2011) Barcoding without DNA? Species identification using near infrared spectroscopy. Zootaxa 2933:46–54. https://doi.org/10.11646/zootaxa.2933.1.3
Salino A, Almeida TE, Smith AR, Gómez AN, Kreier H-P, Schneider H (2008) A new species of *Microgramma* (Polypodiaceae) from Brazil and recircumscription of the genus based on phylogenetic evidence. Syst Bot 33:630–635. https://doi.org/10.1600/036364408786500208
Schneider H, Schuettpelz E (2006) Identifying fern gametophytes using DNA Sequences. Mol Ecol Notes 6:989–991. https://doi.org/10.1111/j.1471-8286.2006.01424.x
Sharma A, Paliwal KK (2015) Linear discriminant analysis for the small sample size problem: an overview. Int J Mach Learn Cyb 6:443–454. https://doi.org/10.1007/s13042-013-0226-9
Shen T, Hong Y, Wang ZY (2020) Discrimination of *Gentiana* and its related species using IR spectroscopy combined with feature selection and stacked generalization. Molecules 25:1442. https://doi.org/10.3390/molecules25061442
Shokralla S, Gibson JF, Nikbakht H, Janzen DH, Hallwachs W, Hajibabaei M (2014) Next-generation DNA barcoding: using next-generation sequencing to enhance and accelerate DNA barcode capture from single specimens. Mol Ecol Resour. https://doi.org/10.1111/1755-0998.12236
Smith JP Jr (2017) Dichotomous keys—their structure and use. Botanical Studies 58, Humboldt State University, Arcata. http://digitalcommons.humboldt.edu/botany_jps/58. Accessed 16 Oct 2020
Stein ED, Martinez MC, Stiles S (2014) Is DNA barcoding actually cheaper and faster than traditional morphological methods: results from a survey of freshwater bioassessment efforts in the United States? PLoS ONE 9:e95525. https://doi.org/10.1371/journal.pone.0095525
Stuart BH (2005) Infrared Spectroscopy: fundamentals and applications. Wiley, Chichester
Tryon RM, Tryon AF (1982) Ferns and allied plants, with special reference to Tropical America. Springer-Verlag, New York
Ustin SL, Jacquemoud S (2020) How the optical properties of fronds modify the absorption and scattering of energy and enhance frond functionality. In: Cavender-Bares J, Gamon J, Townsend P (eds) Remote sensing of plant biodiversity. Springer International Publishing, Cham, pp 349–384
Venables WN, Ripley BD (2002) Modern applied statistics with S, 4th edn. Springer, New York. http://www.stats.ox.ac.uk/pub/MASS4/
Vere N, Rich TCG, Ford CR, Trinder SA, Long C, Moore CW, Satterthwaite D, Davies H, Allainguillaume J, Ronca S, Tatarinova T, Garbett H, Walker K, Wilkinson MJ (2012) DNA barcoding the native flowering plants and conifers of Wales. PLoS ONE 7:e37945. https://doi.org/10.1371/journal.pone.0037945

Wagner WH, Wagner FS (1977) Fertile-sterile leaf dimorphy in ferns. Gard Bull (Singapore) 30:251–267. https://www.biodiversitylibrary.org/part/163903

Wang HH, Lu MJ, Wen J, Ebihara A, Li D (2016) Applying DNA Barcodes to identify closely related species of ferns: a case study of the chinese *Adiantum* (Pteridaceae). PLoS ONE 11:e0160611. https://doi.org/10.1371/journal.pone.0160611

Workman J, Weyer L (2007) Practical guide to interpretive near-infrared spectroscopy. CRC Press, Boca Raton

Yadav S, Shukla S (2016) Analysis of k-fold cross-validation over hold-out validation on colossal datasets for quality classification. In: Advanced computing (IACC), IEEE 6th international conference on IEEE, pp 78–83. https://doi.org/10.1109/IACC.2016.25