Received: 8 June 2022 | Accepted: 20 December 2022

DOI: 10.1002/aps3.11516

APPLICATION ARTICLE

Applications in Plant Sciences

Infrared spectroscopy for ploidy estimation: An example in two species of *Veronica* using fresh and herbarium specimens

Daniele Buono[1,2] | **Dirk C. Albach**[1]

[1]AG Plant Biodiversity and Evolution, Carl von Ossietzky University, Ammerlaender Heerstrasse 114-118, 26129 Oldenburg, Germany

[2]Institute of Botany, Technical University of Dresden, Obergraben 6, 01097 Dresden, Germany

Correspondence

Dirk C. Albach, AG Plant Biodiversity and Evolution, Carl von Ossietzky University, Ammerlaender Heerstrasse 114-118, 26129 Oldenburg, Germany.
Email: dirk.albach@uol.de

Present address

Daniele Buono, Systematik, Biodiversität und Evolution der Pflanzen, Ludwig-Maximilians-University, Menzinger Str. 67, 80638 Munich, Germany.

Abstract

Premise: Polyploidy has become a central factor in plant evolutionary biological research in recent decades. Methods such as flow cytometry have revealed the widespread occurrence of polyploidy; however, its inference relies on expensive lab equipment and is largely restricted to fresh or recently dried material.

Methods: Here, we assess the applicability of infrared spectroscopy to infer ploidy in two related species of *Veronica* (Plantaginaceae). Infrared spectroscopy relies on differences in the absorbance of tissues, which could be affected by primary and secondary metabolites related to polyploidy. We sampled 33 living plants from the greenhouse and 74 herbarium specimens with ploidy known through flow cytometrical measurements and analyzed the resulting spectra using discriminant analysis of principal components (DAPC) and neural network (NNET) classifiers.

Results: Living material of both species combined was classified with 70% (DAPC) to 75% (NNET) accuracy, whereas herbarium material was classified with 84% (DAPC) to 85% (NNET) accuracy. Analyzing both species separately resulted in less clear results.

Discussion: Infrared spectroscopy is quite reliable but is not a certain method for assessing intraspecific ploidy level differences in two species of *Veronica*. More accurate inferences rely on large training data sets and herbarium material. This study demonstrates an important way to expand the field of polyploid research to herbaria.

KEYWORDS

discriminant analysis of principal components (DAPC), herbaria, infrared spectroscopy, neural network, polyploidy, *Veronica*

Whole-genome duplication is an important process in plant evolution and breeding. It generates polyploid individuals with new combinations of traits and possibly new characters, potentially leading to new species (Soltis et al., 2014). Large research efforts have therefore been devoted to understanding the processes leading to polyploidy and to differentiate between the general and lineage-specific characteristics of polyploids (Doyle and Coate, 2019). Given the importance of polyploidy in plant breeding (Salman-Minkov et al., 2016), most of our knowledge of this phenomenon comes from studies of crops. Despite the possibility of generating polyploids in the lab (Tate et al., 2009), the investigation of natural polyploids remains the preeminent method for understanding their evolution for the majority of species. The microscale evolutionary and ecological advantages of polyploidy are often not studied because polyploid individuals are difficult to recognize. Furthermore, temporal trends in the evolution of polyploids are not recognized because the ploidy levels of historical specimens often cannot be inferred. The early recognition of the evolution of tetraploid *Tragopogon* L. species is a rare exception (Ownbey and McCollum, 1954).

A major problem is that many young polyploid lineages are morphologically difficult to distinguish from their parents or homoploid hybrids (e.g., López-González et al., 2018). Size is often an indicator, but not reliably so

(e.g., Lavania et al., 2012); therefore, the frequency of intraspecific variation and mixed ploidy populations is likely underestimated. Traditional chromosome counts are tedious and rely on the availability of fresh material (Windham et al., 2020). Flow cytometry has facilitated a rapid increase in ploidy measurements due to the faster processing of materials and the possibility of using field-collected silica-dried material (Suda and Trávníček, 2006; Meudt et al., 2015). Nevertheless, flow cytometry is still limited by the quality and age of the material and requires bringing material to the lab. Similar restrictions apply to other methods that generate ploidy information parallel to genotyping, such as microsatellite analyses (Huang et al., 2019) or genotyping-by-sequencing (Gompert and Mock, 2017; Siadjeu et al., 2018).

There is also a long tradition of recognizing polyploids based on morphological characteristics such as stomatal guard cell length, epidermal cell area, and pollen size (Tan and Dunn, 1973; Xue et al., 2017), which may even be extended across all angiosperms (Beaulieu et al., 2008) and to fossil plants (Lomax et al., 2014). Aside from these general characters, other more species-specific morphological, anatomical, and physiological ploidy indicators have been identified (e.g., Qiao et al., 2017; Mtileni et al., 2021). An especially interesting aspect is the influence of polyploidy on photosynthesis and chloroplast number (Coate et al., 2012; Qiao et al., 2017). Changes in metabolism and phytochemistry have been used to differentiate polyploids from diploids, but sophisticated laboratory experiments are needed to measure these (Hull-Sanders et al., 2009; Lavania et al., 2012). All these latter methods, however, require living plants and a robust comparison of closely related diploid and polyploid plants.

Here, we explore the use of infrared spectroscopy (IRS) to detect differences between ploidy levels. The method has potential far-reaching applicability and is fast, theoretically allowing the measurement of dozens of plants in a single day. It can be used in the field or with herbarium specimens, is non-destructive, and has no associated costs apart from the equipment. Infrared spectroscopy measures light reflectance in the visible and mid-infrared spectrum, in which the hydrogen bonds with nitrogen, carbon, oxygen, and sulfur generate the typical spectra (Chen et al., 2014). Water, proteins, carbohydrates, and other metabolites therefore leave a specific imprint on tissues, which can be identified by applying this method. The specific array of compounds, each reflecting light at a distinct wavelength, produces a specific pattern that can be measured and interpreted. Infrared spectroscopy has been used since the early 1970s to determine the quality of forage and crops (e.g., Norris et al., 1976). Its subsequent uses include phenotyping in breeding programs (Cabrera-Bosquet et al., 2012) and vegetation analyses focusing on diversity (Hill et al., 1988) and vegetation health (Pontius et al., 2005). A number of taxonomic studies have used IRS to differentiate species (Lu et al., 2008a; Lang et al., 2015; Prata et al., 2018), hybrids (Stasinski et al., 2021), or even intraspecific units (Gao et al., 2012), sometimes demonstrating superior results to DNA barcoding (Durgante et al., 2013). Meireles et al. (2020) even studied spectra across seed plants and found a phylogenetic signal. Whereas most of these studies used extracts (Werner et al., 2006), powders (Fan et al., 2010), airborne remote sensing (Feret and Asner, 2013), or fresh plants (Asner et al., 2014b), the use of IRS with herbarium material has been previously demonstrated (Strgulc Krajšek et al., 2008; Lang et al., 2015) but has not gained wide attention.

Few researchers have used IRS to differentiate and measure polyploids. Demir et al. (2015) used IRS to differentiate diploid and allopolyploid *Triticum* L. and *Aegilops* L. species, while Atkinson et al. (1997) differentiated diploid and tetraploid birch (*Betula* L.) species and their hybrids. Blonder et al. (2020, 2021) used the method to differentiate diploid and triploid individuals of quaking aspen (*Populus tremuloides* Michx.) in a remote-sensing study to infer microscale niche differences between two morphologically indistinguishable cytotypes. Closely related individuals of different ploidy may differ in a number of ways, which influence reflectance patterns; for example, Blonder et al. (2020) suggested that diploid and triploid aspens differ in their chlorophyll content. Furthermore, different cytotypes can accumulate different levels of secondary metabolites, especially lignin, cellulose, and other cell wall compounds (Ferreira de Carvalho et al., 2017), which are detectable using IRS (Robinson and Mansfield, 2009; Tucker et al., 2017). However, a number of potentially compounding factors need to be considered; for example, Lang et al. (2015) noticed leaf age can alter spectral reflectance, while Asner et al. (2014a) reported differences between sunlit or shaded leaves of Amazonian trees, but also considered possible variation due to herbivory.

Here, we explore the ability to differentiate intraspecific ploidy variation in two species of *Veronica* L. (*V. spicata* L. and *V. longifolia* L.) that have been shown to have multiple origins of tetraploids with no clear geographic pattern in their distribution (Buono et al., 2021). The analysis of ploidy by Buono et al. (2021) was limited to flow cytometry and, consequently, utilized recently collected material. Expanding our knowledge of ploidy distribution to samples present in herbaria would, therefore, significantly increase our capacity to infer distribution patterns and the history of polyploid formation, as well as potential climatic correlates of polyploidy. We were not able to morphologically differentiate diploids and tetraploids in either species using either fresh or herbarium material; thus, our study provides the first possible method for determining the ploidy of herbarium specimens in species with morphologically indistinguishable cytotypes.

METHODS

Plant material

We sampled 33 living plants and 74 herbarium specimens with ploidy previously estimated using flow cytometry

(Buono et al., 2021). All herbarium specimens were less than 10 years old and in excellent condition. Herbarium specimens were made from the same individuals analyzed as fresh material, and all specimens were prepared in the same way. The sampling of the living plants included two diploid and 13 tetraploid *V. spicata* and five diploid and 13 tetraploid *V. longifolia*. Living individuals were cultivated in the greenhouses of the Botanical Garden of the Carl von Ossietzky University in Oldenburg, Germany. All plants were well hydrated at the moment of sampling. The sampling of the herbarium material included 31 diploid and nine tetraploid *V. spicata* and 22 diploid and 23 tetraploid *V. longifolia* from the herbarium of Carl von Ossietzky University (OLD), the ploidy of which were measured in a previous study (Buono et al., 2021). Voucher specimens and information on the origin of the material are provided in Appendix S1 (see Supporting Information with this article).

Acquisition of reflectance spectra

The reflectance spectra of the leaves of fresh and dry plant material were sampled using a SR-3500 spectroradiometer (Spectral Evolution, Haverhill, Massachusetts, USA). The instrument was calibrated with the ceramic white background provided by the manufacturer. Each spectrum obtained consists of 2151 reflectance values (expressed in percentages) for each electromagnetic wavelength, ranging from 350 nm (ultraviolet A) to 2500 nm (shortwave infrared). On each specimen, when available, a total of 24 spectra were taken at four sampling points: (i) six spectra on the adaxial side of new leaves (leaves close to the shoot apical meristem), (ii) six spectra on the abaxial side of new leaves, (iii) six spectra on the adaxial side of old leaves (leaves close to the basal end of the stem), (iv) six spectra on the abaxial side of old leaves (Figure 1). Each spectrum was taken on a different leaf when available. For some specimens, it was not possible to sample specific points (e.g., in herbarium material, often the abaxial or adaxial leaf side was not exposed), so for those individuals fewer than 24 spectra were obtained. As a result, 22 herbarium specimens were represented by 24 spectra, with 57 herbarium specimens represented by at least 12 spectra. For the fresh material, 26 specimens were represented by 24 spectra, and 32 by more than 20 spectra.

Data analysis

All analyses were performed using a self-written R script using R 3.6.3 (R Core Team, 2014) (see Appendices S2 and S3 for code). The code was executed on a Linux laptop, except the neural network (NNET) parameter optimization, which was executed on the high-performance computing cluster facility at Carl von Ossietzky University.

The raw spectra that showed reflectance values higher than 100% were discarded. A Savitzky–Golay filter was applied to reduce sampling noise using the R package SIGNAL (Ligges et al., 2021). After a visual inspection of the filtering effect, a filter order of 3 was chosen. Spectra too similar to that of the background (i.e., paper for the herbarium specimens, probe-white for the living individuals) were removed. To alleviate the computational burden, we used one reflectance measurement every 5 nm instead of one every 1 nm, calculated as the average over the adjacent reflectance values. In this way, the data set size was reduced by about 80%, resulting in 430 reflectance values on each spectrum.

To test whether the freshness of the sample, the sampling point on the leaves, or averaging the reflectance spectra altered the ploidy estimation capabilities, the analyses were performed on different data subsets. We used three data sets: (1) both fresh and dry material, (2) only dry

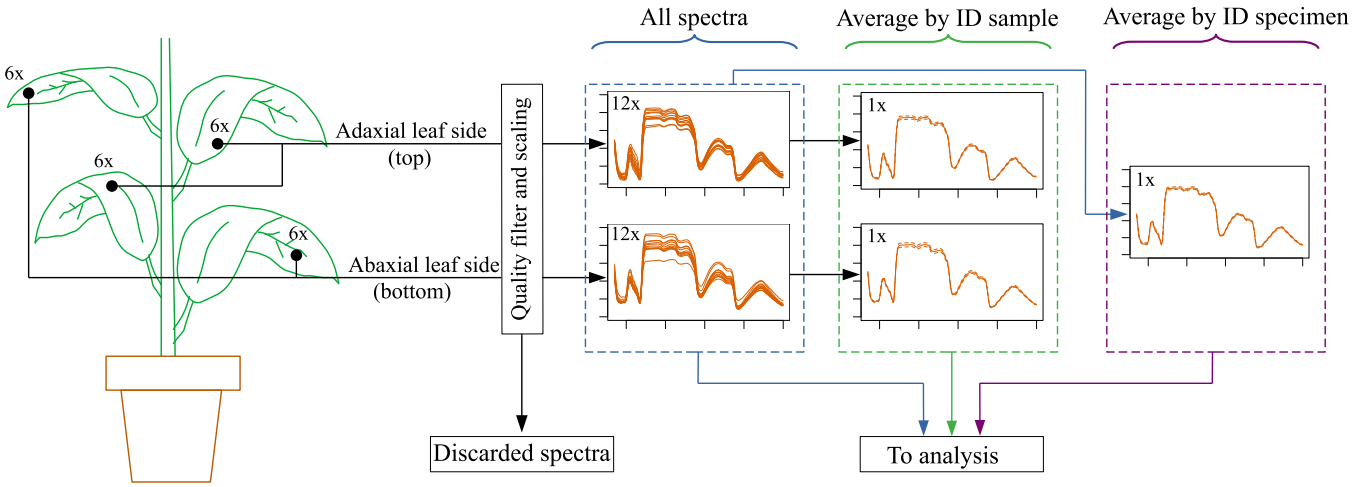

FIGURE 1 Scheme of reflectance spectra sampling, quality filtering and scaling, subsetting, and averaging. As an example, real spectra sampled from individual OLD01319 are shown.

material, and (3) only fresh material. Each of those data sets was also divided into three subsets: (1) all spectra, (2) spectra averaged by specimen, and (3) spectra averaged by sampling point. Each of these subsets was then divided into (1) both leaf sides analyzed, (2) abaxial leaf side only, or (3) adaxial leaf side only. The data subsets were analyzed separately for each species and for the two species combined.

We performed three types of data analyses: a hierarchical cluster analysis, a discriminant analysis of principal components (DAPC), and a NNET analysis. The hierarchical cluster analysis was performed on the data sets to assess a possible division between samples without providing any information on ploidy level, using the R package DENDEXTEND (Galili, 2015). The DAPC analysis was implemented with the package ADEGENET (Jombart and Ahmed, 2011). The DAPC method was developed to discriminate genetic clusters, and as it is based on a principal component (PC) analysis, it can be used to analyze big data sets relatively quickly. To reach a balance between the discriminant power and the capability to handle samples that were not included in the construction of the discriminant model (i.e., to be able to predict the ploidy of the additional samples), a proper number of PCs must be determined. To do that, we assessed the accuracy in ploidy discrimination for each PC value from 10 to 200 with a leave-one-out method implemented with a self-written R script (Appendix S3). This method requires the removal of all spectra belonging to one individual from the data set and using them as a test set. The DAPC model that discriminates the ploidy level is constructed based on spectra of the remaining individuals (training set). The capability of the model to infer the ploidy of the test set is then evaluated. This procedure was repeated for all individuals, so that each individual was a test individual one time. The accuracy of the prediction was calculated as a percentage of the correct ploidy classification.

A NNET analysis with a resilient back-propagation was implemented using the R package NEURALNET (Fritsch et al., 2019). Neural networks are commonly used to detect patterns in data and for classification, and can produce easy-to-use models with large inputs (Abiodun et al., 2018). To determine the best parameters for the NNET (number of neurons [i.e., nodes] in each hidden layer and the number of hidden layers), a *k*-fold test was performed for each set of parameters using a self-written R script (Appendix S3). This test consists of dividing the data set into *k* random groups, so that a group contains all the spectra of a certain number of individuals. One group (test group) is taken out, and the NNET is then constructed and trained with the data in the remaining groups (training groups). The performances of the NNET are then tested on each individual of the test group. The procedure is repeated for each group, so that each group is a test group one time. Note that when $k = n$, each group consists of a single individual, which is equivalent to the leave-one-out approach described above. We decided to implement this method instead of leave-one-out for the NNET to reduce the computational burden, using a value for *k* that led to about three individuals in each group. The *k*-fold test was repeated for each combination of the parameters: number of layers from 1 to 3, number of neurons in each layer from 20 to 120, and the ratio between the neurons in each layer from 0.3 to 0.8. The seed parameter, which initialized the pseudo-random number generator in the functions, is typically subject to a high degree of stochasticity, and thus was always set to the value 222 to guarantee the repeatability of the results.

To assess whether the differences in prediction accuracy between the two classifiers (NNET and DAPC) were statistically significant, we computed two-proportion *Z*-tests using the function prop.test in the base R package stats. Before the analyses, the data sets were balanced using the function ROSE in the R package ROSE (Lunardon et al., 2014). A *Z*-test was executed for the four best results (Table 1).

RESULTS

A total of 2522 and 782 spectra were sampled from herbarium material and living plants, respectively. After all filtering steps described above and the removal of spectra from individuals of unknown ploidy, 1309 and 778 spectra were subsequently analyzed for the herbarium and living materials, respectively. Those spectra represented 74 herbarium specimens (37 *V. longifolia*, nine diploid and 28 tetraploid; 37 *V. spicata*, 23 diploid and 14 tetraploid) and 33 living individuals (15 *V. longifolia*, seven diploid and eight tetraploid; 18 *V. spicata*, 10 diploid and eight tetraploid) (Table 1). Sample spectra are shown in Figure 2.

Hierarchical clustering did not reveal a clear pattern of clustering in terms of either ploidy or species (Appendix S4). The herbarium material showed a division into five groups, which are more or less homogenous in terms of species. The living material showed a division into a pure *V. spicata* cluster and a second cluster composed of two-thirds *V. longifolia* and one-third *V. spicata*. This demonstrates that the spectra of the two species are sufficiently similar to combine them, rather than separating both species, which leads to smaller sample size and reduced statistical power (Appendix S5).

Optimization procedures for DAPC showed only a slight variation in ploidy prediction accuracy. The best results were obtained with different values of PC for each subset (herbarium, fresh, combined; Table 1). The use of DAPC for herbarium material alone with one average spectrum per specimen obtained from the adaxial surface of leaves only led to the best performances, with 84.1% accuracy (78.3% balanced accuracy taking sample numbers into account) in the predictions (50 PC retained) compared with 67.6% and 64.9% for *V. spicata* and *V. longifolia*, respectively, if analyzed individually. For the living samples, the highest accuracy reached was 70.8% (83.7% balanced accuracy; 20 PC), using one average spectrum per sampling point on the bottom side of the leaf only (50.0% and 86.7%

TABLE 1 Percentages of correct predictions obtained using the different data subsets and methods, with optimized parameters. The values in parentheses after the percentages indicate the number of principal components (PCs) for discriminant analysis of principal components (DAPC) and the number of neurons on each layer for the neural network (NNET) analysis. For the NNET, the *k*-fold values of 25, 11, and 35 were used for the herbarium, fresh, and combined data sets, respectively. The DAPC was optimized with the leave-one-out method. Bold numbers indicate the best NNET/DAPC settings used to predict ploidy levels in Appendix S1.

		Leaf side					
		Both		Adaxial		Abaxial	
Data set	Grouping[a]	DAPC	NNET	DAPC	NNET	DAPC	NNET
Herbarium	All	78.6% (110)	75.4% (33, 14, 6)	79.0% (80)	75.1% (76, 39, 20)	79.4% (80)	77.8% (103, 62, 38)
	ID specimen	81.1% (60)	**85.1% (36, 11, 4)**	**84.1% (50)**	78.3% (24, 8, 3)	79.2% (30)	81.9% (77, 24, 8)
	ID sample	79.7% (40)	81.9% (93, 28, 9)	81.6% (60)	79.8% (29, 9, 3)	80.5% (30)	83.7% (103, 62, 38)
Fresh	All	56.0% (40)	63.0% (120, 97, 78)	56.6% (30)	63.0% (120, 97, 78)	67.9% (30)	62.2% (89, 63, 45)
	ID specimen	66.7% (30)	63.6% (96, 68, 48)	57.6% (40)	57.56% (32, 13, 6)	69.7% (20)	66.7% (41, 21, 11)
	ID sample	62.3% (40)	63.1% (96, 68, 48)	55.4% (20)	63.1% (96, 68, 48)	**70.8% (20)**	63.1% (96, 68, 48)
Combined	All	62.6% (30)	68.0% (27, 11)	63.6% (30)	68.0% (34, 18)	68.6% (80)	67.6% (24, 10)
	ID specimen	65.4% (50)	**74.8% (41, 17, 7)**	68.6% (60)	68.6% (25, 8)	70.5% (40)	74.3% (44, 27, 17)
	ID sample	62.9% (50)	70.6% (46, 28, 17)	65.9% (30)	70.4% (49, 40, 33)	70.7% (50)	72.3% (170, 52, 16)

[a]"All" indicates all spectra sampled, "ID specimen" indicates one average spectrum for each specimen, and "ID sample" indicates one average spectrum for each sampling point (four spectra per specimen).

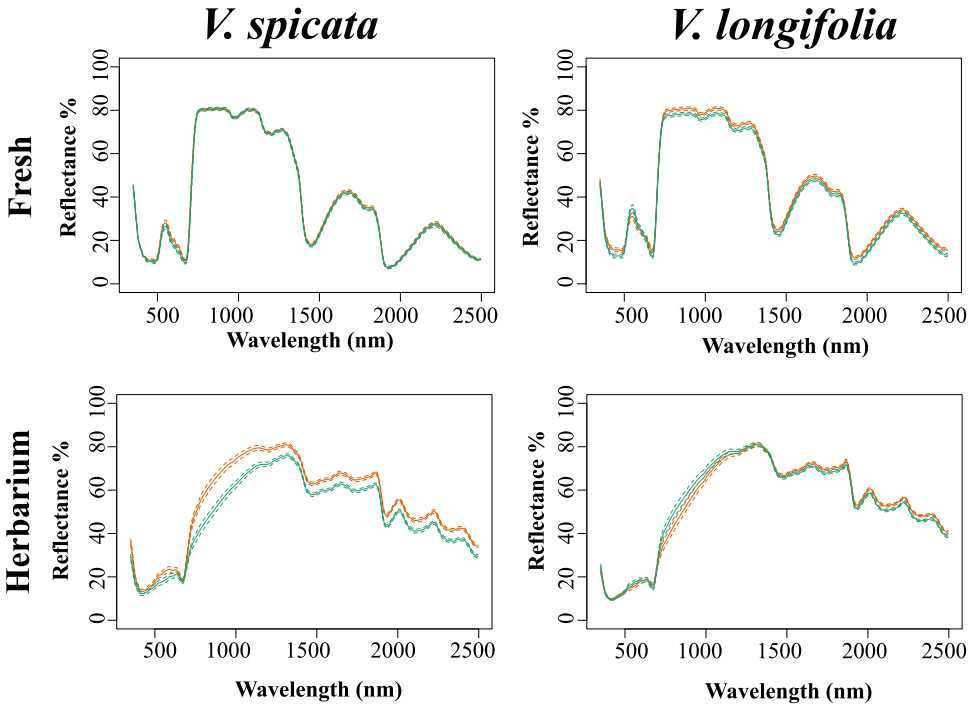

FIGURE 2 Average reflectance spectra in the four data subsets. Dashed lines indicate the standard error. Different colors indicate diploids (orange) and tetraploids (green).

for *V. spicata* and *V. longifolia*, respectively, if analyzed individually). Combining both dry and fresh individuals in a unique data set led to 70.7% accuracy (50 PC; 73.1% and 76.9% for *V. spicata* and *V. longifolia*, respectively; Appendix S6).

The NNET optimization resulted in a different number of neurons and layers for each data subset (Table 1). The highest precision in the prediction, 85.1% (84.7% balanced accuracy), was achieved using herbarium spectra, with one average spectrum per specimen from both leaf sides, using 36, 11, and four neurons in three layers (81% and 75% accuracy was achieved for *V. spicata* and *V. longifolia*, respectively, if analyzed individually; Appendix S6). On the other hand, the prediction of ploidy was more accurate for

the living specimens (although still not reliably so) when the herbarium specimens were also included in the training set, leading to a 74.8% accuracy (56.3% balanced accuracy) using 41, 17, and seven neurons in the three layers and also using spectra from both leaf surfaces (33% and 46% for *V. spicata* and *V. longifolia*, respectively, if analyzed individually; Appendix S6). The differences in the prediction accuracies observed are statistically insignificant ($\chi^2 = 0.136–1.806$, $df = 1$, $P > 0.05$; Appendix S7).

DISCUSSION

Here, we demonstrate that IRS is a suitable method for discriminating intraspecific variation in ploidy in both living and herbarium materials. The identification of ploidy levels is often difficult because differences in ploidy do not necessarily translate into obvious phenotypic differences (Soltis et al., 2007), meaning the extent of ploidy variation within species is consequently often underestimated. This also means that geographical and/or ecological patterns in the distribution of certain cytotypes have been overlooked, hindering the progress in our understanding of the consequences of polyploidization (e.g., Bardy et al., 2010). Infrared spectroscopy was recently shown to be efficient in differentiating diploid and triploid quaking aspen in the field on a large scale (Blonder et al., 2020, 2021); however, neither the generality of the method to differentiate ploidy levels across plants, nor the ability to differentiate herbarium specimens and thus historical collections, has previously been demonstrated. Here, we demonstrate the general applicability of the method and its limitations.

Intraspecific variation

Veronica spicata and *V. longifolia* are both known to have diploid and autotetraploid individuals across much of their total ranges (Buono et al., 2021). These ploidy levels have previously been determined using pollen size, stomata size (Trávníček and Vinter, 1999), chromosome counts (summarized by Albach et al., 2008), and flow cytometry (Buono et al., 2021); however, the latter two methods require living plants or at least recently well-dried material. In addition, pollen and stomata size were shown to have significant differences in means but large overlaps, making the determination of ploidy based on these characters unreliable (Trávníček and Vinter, 1999). The possibility of using IRS to accurately differentiate ploidy based on dried material such as herbarium specimens therefore has great promise for a variety of studies (Figure 3). Whereas Blonder et al. (2020) only sampled in one population of aspen using IRS, we were able to differentiate the ploidy levels of *V. spicata* and *V. longifolia* across the Eurasian distribution range of both species. This allowed us to evaluate the use of IRS across the intraspecific variation of the species irrespective of ploidy variation, consequently integrating across different climates, water availability, seasons, light, and nutrient availability. In the field, factors such as leaf age and water stress have been shown to cause variation in the spectra (Blonder et al., 2020), as well as variation in phosphorus content and phenolic

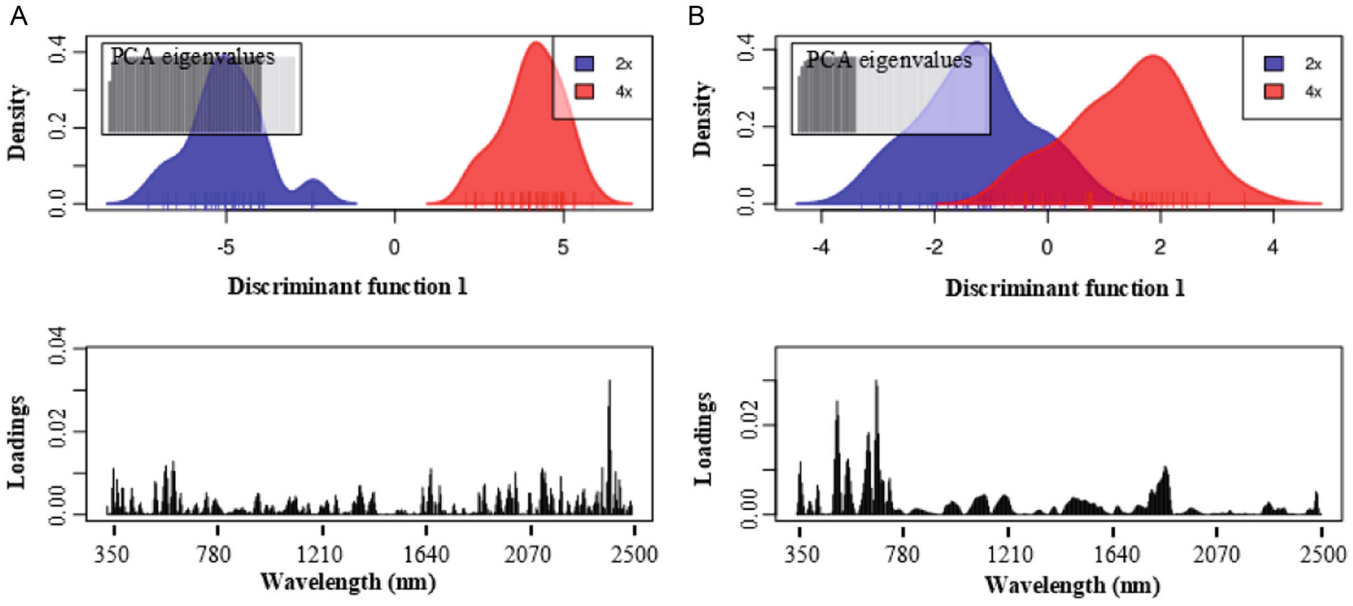

FIGURE 3 Discriminant analysis of principal components (DAPC) scatterplots for the (A) herbarium material and (B) fresh material. The best-discriminating results are shown according to the results summarized in Table 1. In the loading plots (bottom), it is possible to observe which wavelengths influence the discrimination capability between 2*x* and 4*x* most strongly. For the herbarium material, only the adaxial surface of the leaf was used, and each specimen was represented by one average spectrum. For the fresh material, only the abaxial surface of the leaf was used, and each specimen was represented by several spectra averaged by sampling points.

compounds (Asner et al., 2014a). Across the range of a species this may be further amplified by intraspecific genotypic variation and different phenologies. In the living material, we avoided most of these effects by using plants grown in the greenhouse under the same conditions; however, our sample size was not large enough to evaluate the effects of known intraspecific genotypic variation (Buono et al., 2021). We estimate that this would require at least 30–50 specimens per phylogroup.

It is important to note that the ability to discriminate ploidies was markedly different among the two species when using the fresh material. The ploidy levels determined using flow cytometry and IRS were more commonly consistent for *V. longifolia* (11/15 specimens in fresh material) than for *V. spicata* (7/18 specimens); however, the two ploidies were equally likely to be misidentified (flow cytometry diploids identified as tetraploid using IRS or vice versa; Appendix S1). The difference between species was much lower when using the dried material, and *V. spicata* was better classified. This suggests that the differences in ploidy discrimination in fresh material were caused either by different water contents or dilutions of discriminative compounds by water, or that the lower sample size in the living material reduced the success of identifying ploidy levels. However, it should also be noted that the two types of classifiers (DAPC and NNET) differ, and the two methods agree on a different ploidy than the one obtained by flow cytometry only for two fresh and three dried specimens. In the other cases, one classifier disagreed with flow cytometry and the other classifier agreed or did not unanimously identify a ploidy level.

Analytical considerations

Previous studies using IRS used different types of analyses to delimit the groups. Here, we were not interested in the similarities but rather the discrimination of groups, and so did not use phenetic clustering algorithms (Lu et al., 2008b). Most studies used some kind of PC and/or discriminant analyses similar to our DAPC method (Durgante et al., 2013; Feret and Asner, 2013; Lang et al., 2015); however, these methods assume a normal distribution of the data and require more training data (Castro-Esau et al., 2004; Feret and Asner, 2013). We therefore employed non-parametric NNET analyses that were shown to have higher prediction accuracies in those few studies comparing different methods (Castro-Esau et al., 2004; Fan et al., 2010), although this seems to depend on specific method and amount of training data (Feret and Asner, 2013). In contrast to our expectations, the NNET analysis disagreed with flow cytometry results more often than the DAPC analysis, which was especially pronounced in the living material. Given the large difference between directly measured accuracy and balanced accuracy, this is likely a sampling artifact. Thus, we conclude that our sample size may be too low for the reliable estimation of ploidy levels in our group, but this is not a problem of the method per se.

Dried material

Only a few studies have compared the differences between dry and fresh plant material, most notably Elvidge (1990). These studies reported that the use of dried material is more efficient for estimating leaf nutrients (Prananto et al., 2021) or secondary metabolites (Couture et al., 2016), which is attributed to the direct and indirect effects of water on the reflectance spectra (Ollinger, 2011). In line with this reasoning, we not only found large radiation-reflectance differences between the dried and fresh material (Figure 4), but also found that using dried material led to a more efficient differentiation of ploidy levels (Table 1). Given the

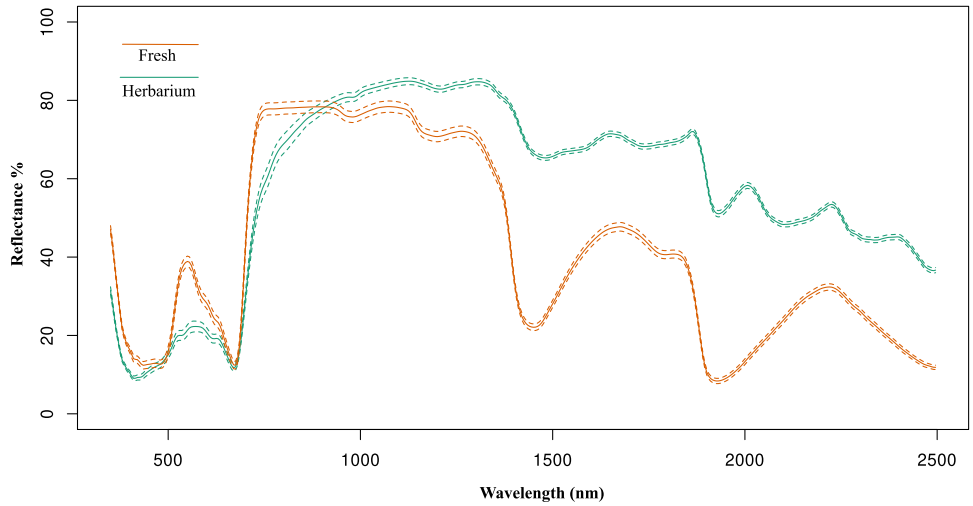

FIGURE 4 Example of reflectance spectra from the same individual of *Veronica longifolia* (Boiko 20, OLD01319). The green line represents the herbarium sample, while orange represents the fresh sample. The spectra shown were obtained as an average of several spectra sampled from different points on the plant. The dashed line indicates the standard error.

difference in sampling size, we cannot clearly state whether this enhanced discrimination was due to dried material performing better, more training data in the dried material, or both. Nevertheless, the correspondence to earlier results suggests both may have at least some influence. The spectra of dried material differs not only in the characteristic decrease in reflectance at about 1450 nm, 1950 nm, and above 2300 nm caused by water, but also in an overall increase of reflectance, as noted earlier (Figure 4) (Ollinger, 2011), and in the visible light wavelengths. The latter difference, which presumably corresponds to chlorophyll, is much lower in the herbarium material, suggesting degradation during the drying procedure. Another characteristic difference is the curved rise in reflectance between 800 and 900 nm, which Castro and Sanchez-Azofeifa (2008) associated with a reduction in air spaces and the width of spongy mesophyll.

The variation among the spectra of dried leaves irrespective of ploidy should also be recognized. This variation is partly due to differences in the living material, as discussed above, but also caused by differences in the drying procedure. It is well known that different herbarium specimens were dried with varying degrees of effectiveness, which influences DNA preservation (Forrest et al., 2019), but this may also alter the IRS reading and potentially the ability to differentiate ploidy levels as well. This intraspecific variation may in fact be much larger in more extensive studies than ours, for example, in projects including specimens from different herbaria that were collected using different collection and drying methods. Furthermore, age effects may influence the ability to differentiate ploidy levels, although DNA extraction studies have shown that the quality of herbarium specimen preparation may be at least as important as the specimen's age (Choi et al., 2015; Höpke et al., 2018).

Differences between ploidies

Given all these potential causes for intraspecific and analytical variation, the fact that we were able to differentiate ploidy levels with some success suggests that there are some underlying general differences between diploid and tetraploid *Veronica*. Blonder et al. (2020, 2021) used spectroscopy to differentiate diploid and triploid aspen, but only used a smaller part of the spectrum. In their analysis, the largest difference was considered to be related to the amount of chlorophyll, with some additional variation caused by the water/dry matter content. The high correlation between chlorophyll content and reflectance in the range between 510 to 650 nm suggests that our highly discriminating wavelengths at 590 and 620 nm (Figure 3) are related to chlorophyll content, similar to the findings of Blonder et al. (2020). This importance of chlorophyll for the distinction of ploidy levels was anticipated by studies demonstrating that polyploids have increased rates of photosynthesis and numbers of chloroplasts per cell, which can only be compensated at the per-area level by a proportionally higher decrease of cells per leaf area (Warner and Edwards, 1993). Although we did not measure cell size, such a decrease is unlikely given the lack of correlation between guard cell size and ploidy in our study species (Trávníček and Vinter, 1999).

In addition to the differences in the visible light range, the spectra also reveal other differences between the ploidy levels. Atkinson et al. (1997) used a similar spectral range to ours, and also found most variation between ploidies in the range of 2300–2500 nm. Unfortunately, a number of compounds could be responsible for these differences, such as lignin, cellulose, and other carbohydrates (Elvidge, 1990), which are also likely to be responsible for the peak around 1675–1680 nm (Flinn et al., 1996). Both lignin- and cellulose-related genes have been demonstrated to be overexpressed in polyploids (Ferreira de Carvalho et al., 2017; Zhang et al., 2022), although studies in *Salix* L. and *Arabidopsis* Heynh. suggest that these changes may be compound-specific, with a decrease of up to 50% in lignin content and increases in hemicellulose and pectins (Serapiglia et al., 2015; Corneillie et al., 2018). Thus, our study supports the hypothesis of a general restructuring of the cell walls in polyploids compared with their diploid relatives (Corneillie et al., 2018).

A final region of differentiation between the ploidy levels was detected in the range of 2130–2135 nm, which corresponds to proteins (Shenk and Westerhaus, 1993). This suggests a more general change in the metabolic activity and the proteome size (Doyle and Coate, 2019); however, IRS is not suitable for investigating changes in the proteome, not even its size. Finally, the region of least differentiation (1400–1650 nm, 1900–2000 nm) was similar to that reported by Atkinson et al. (1997), suggesting more general similarities between closely related plants of different ploidies in this range. Interestingly though, the latter of these two regions is the one in which the most phylogenetic signal was found across the seed plants (Meireles et al., 2020).

We here demonstrate that IRS is a reliable but not certain way to assess ploidy level in *V. spicata* and *V. longifolia*. Flow cytometry will consequently remain the gold standard for ploidy estimation. The success of IRS largely depends on a large and high-quality training set of plants with known ploidy. The need for large training data sets will likely be a limiting factor for small institutions or for studies on rare species. One solution to this problem may be to incorporate data from closely related species, as the differences in spectra induced by polyploidy may be larger than between closely related species, as was the case here. This may not be generalizable to all genera, but studies from a wide range of species have shown that polyploids have a larger chlorophyll content and different amounts of cell wall compounds (Warner and Edwards, 1993; Podwyszyńska et al., 2015, 2021), arguing against species-specific patterns of discrimination between cytotypes.

Further refinements of the method are necessary in terms of the sampling strategy and statistical analysis of

spectra. Nevertheless, based on our approach, we have been able to infer ploidy levels with an error margin of approximately 15%. Even though it is not clear which components are responsible for the detection of ploidy levels using IRS, our results provide initial hypotheses about the wavelengths best able to discriminate ploidy levels (Figure 3). We further provide initial guidance for the field application of this method and the investigation of herbarium material. Finally, a number of different types of statistical analyses are possible, and the properties of these analyses need to be further explored in their application to IRS data.

AUTHOR CONTRIBUTIONS

D.C.A. conceived the research. D.B. and D.C.A. designed the experiments. D.B. performed all experiments and analyzed the data. D.B. and D.C.A. wrote the manuscript. D.C.A. acquired the funding. All authors approved the final version of the manuscript.

ACKNOWLEDGMENTS

This research was funded by the Deutsche Forschungsgemeinschaft (grant number AL632/19-1) within the priority program "Taxon-Omics: New Approaches for Discovering and Naming Biodiversity" (SPP 1991). The authors are grateful to Michael Kleyer, Cord Peppler-Lisbach, and Kertu Lohmus (Carl von Ossietzky University) for providing access and an introduction to the spectroradiometer. Open Access funding enabled and organized by Projekt DEAL.

DATA AVAILABILITY STATEMENT

The original spectral data are available at https://doi.org/10.5281/zenodo.7729563 (Buono and Albach, 2023).

ORCID

Dirk C. Albach http://orcid.org/0000-0001-9056-7382

REFERENCES

Abiodun, O. I., J. Aman, E. O. Abiodun, V. D. Kemi, A. M. Nachaat, and A. Humaira. 2018. State-of-the-art in artificial neural network applications: A survey. *Heliyon* 4: e00938.

Albach, D. C., M. M. Martinez-Ortega, L. Delgado Sanchez, H. Weiss-Schneeweiss, F. Özgökce, and M. A. Fischer. 2008. Chromosome numbers in Veroniceae: Review and several new counts. *Annals of the Missouri Botanical Garden* 95: 543–566.

Asner, G. P., R. E. Martin, R. Tupayachi, C. B. Anderson, F. Sinca, L. Carranza-Jiménez, and P. Martinez. 2014a. Amazonian functional diversity from forest canopy chemical assembly. *Proceedings of the National Academy of Sciences, USA* 111: 5604–5609.

Asner, G. P., R. E. Martin, L. Carranza-Jiménez, F. Sinca, R. Tupayachi, C. B. Anderson, and P. Martinez. 2014b. Functional and biological diversity of foliar spectra in tree canopies throughout the Andes to Amazon region. *New Phytologist* 204: 127–139.

Atkinson, M. D., A. P. Jervis, and R. S. Sangha. 1997. Discrimination between *Betula pendula*, *Betula pubescens*, and their hybrids using near-infrared reflectance spectroscopy. *Canadian Journal of Forest Research* 27: 1896–1900.

Bardy, K. E., D. C. Albach, G. M. Schneeweiss, M. A. Fischer, and P. Schönswetter. 2010. Disentangling phylogeography, polyploid evolution and taxonomy of a woodland herb (*Veronica chamaedrys* group, Plantaginaceae s.l.) in southeastern Europe. *Molecular Phylogenetics and Evolution* 57: 771–786.

Beaulieu, J. M., I. J. Leitch, S. Patel, A. Pendharkar, and C. A. Knight. 2008. Genome size is a strong predictor of cell size and stomatal density in angiosperms. *New Phytologist* 179: 975–986.

Blonder, B., B. J. Graae, B. Greer, M. Haagsma, K. Helsen, R. E. Kapás, H. Pai, et al. 2020. Remote sensing of ploidy level in quaking aspen (*Populus tremuloides* Michx.). *Journal of Ecology* 108: 175–188.

Blonder, B., P. G. Brodrick, J. A. Walton, K. D. Chadwick, I. K. Breckheimer, S. Marchetti, C. A. Ray, and K. Mock. 2021. Remote sensing of cytotype and its consequences for canopy damage in quaking aspen. *Global Change Biology* 28: 2491–2504.

Buono, D., G. Khan, K. B. von Hagen, P. A. Kosachev, E. Mayland-Quellhorst, S. L. Mosyakin, and D. C. Albach. 2021. Comparative phylogeography of *Veronica spicata* and *V. longifolia* (Plantaginaceae) across Europe: Integrating hybridization and polyploidy in phylogeography. *Frontiers in Plant Science* 11: 588354.

Buono, D., and D. C. Albach. 2023. Data from: Infrared spectroscopy for ploidy estimation: An example in two species of *Veronica* using fresh and herbarium specimens. Available at Zenodo repository https://doi.org/10.5281/zenodo.7729563 [accessed 14 March 2023].

Cabrera-Bosquet, L., J. Crossa, J. von Zitzewitz, M. D. Serret, and J. Luis Araus. 2012. High-throughput phenotyping and genomic selection: The frontiers of crop breeding converge. *Journal of Integrative Plant Biology* 54: 312–320.

Castro, K. L., and G. A. Sanchez-Azofeifa. 2008. Changes in spectral properties, chlorophyll content and internal mesophyll structure of senescing *Populus balsamifera* and *Populus tremuloides* leaves. *Sensors* 8: 51–69.

Castro-Esau, K. L., G. Sánchez-Azofeifa, and T. Caelli. 2004. Discrimination of lianas and trees with leaf-level hyperspectral data. *Remote Sensing of Environment* 90: 353–372.

Chen, C.-W., H. Yan, and B.-X. Han. 2014. Rapid identification of three varieties of *Chrysanthemum* with near infrared spectroscopy. *Revista Brasileira de Farmacognosia* 24: 33–37.

Choi, J., H. Lee, and A. Shipunov. 2015. All that is gold does not glitter? Age, taxonomy, and ancient plant DNA quality. *PeerJ* 3: e1087.

Coate, J. E., A. K. Luciano, V. Seralathan, K. J. Minchew, T. G. Owens, and J. J. Doyle. 2012. Anatomical, biochemical, and photosynthetic responses to recent allopolyploidy in *Glycine dolichocarpa* (Fabaceae). *American Journal of Botany* 99: 55–67.

Corneillie, S., N. De Storme, R. Van Acker, J. U. Fangel, M. De Bruyne, R. De Rycke, D. Geelen, et al. 2018. Polyploidy affects plant growth and alters cell wall composition. *Plant Physiology* 179: 74–87.

Couture, J. J., A. Singh, K. F. Rubert-Nason, S. P. Serbin, R. L. Lindroth, and P. A. Townsend. 2016. Spectroscopic determination of ecologically relevant plant secondary metabolites. *Methods in Ecology and Evolution* 7: 1402–1412.

Demir, P., S. Onde, and F. Severcan. 2015. Phylogeny of cultivated and wild wheat species using ATR–FTIR spectroscopy. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy* 135: 757–763.

Doyle, J. J., and J. E. Coate. 2019. Polyploidy, the nucleotype, and novelty: The impact of genome doubling on the biology of the cell. *International Journal of Plant Sciences* 180: 1–52.

Durgante, F. M., N. Higuchi, A. Almeida, and A. Vicentini. 2013. Species spectral signature: Discriminating closely related plant species in the Amazon with near-infrared leaf-spectroscopy. *Forest Ecology and Management* 291: 240–248.

Elvidge, C. D. 1990. Visible and near infrared reflectance characteristics of dry plant materials. *International Journal of Remote Sensing* 11: 1775–1795.

Fan, Q., Y. Wang, P. Sun, S. Liu, and Y. Li. 2010. Discrimination of *Ephedra* plants with diffuse reflectance FT-NIRS and multivariate analysis. *Talanta* 80: 1245–1250.

Feret, J., and G. P. Asner. 2013. Tree species discrimination in tropical forests using airborne imaging spectroscopy. *IEEE Transactions on Geoscience and Remote Sensing* 51: 73–84.

Ferreira de Carvalho, J., J. Boutte, P. Bourdaud, H. Chelaifa, K. Ainouche, A. Salmon, and M. Ainouche. 2017. Gene expression variation in natural populations of hexaploid and allododecaploid *Spartina* species (Poaceae). *Plant Systematics and Evolution* 303: 1061–1079.

Flinn, P., N. Edwards, C. Oldham, and M. McNeil. 1996. Near infrared analysis of the fodder shrub tagasaste (*Chamaecytisus proliferus*) for nutritive value and anti-nutritive factors. *In* M. C. Davies and P. C. Williams [eds.], Near infrared spectroscopy: the future waves, 576–580. NIR Publications, Chichester, United Kingdom.

Forrest, L. L., M. L. Hart, M. Hughes, H. P. Wilson, K.-F. Chung, Y.-H. Tseng, and C. A. Kidner. 2019. The limits of Hyb-Seq for herbarium specimens: impact of preservation techniques. *Frontiers in Ecology and Evolution* 7: 10.3389.

Fritsch, S., F. Guenther, M. N. Wright, M. Suling, and S. M. Mueller. 2019. Package 'neuralnet': Training of Neural Networks. Website: https://github.com/bips-hb/neuralnet [accessed 2 March 2023].

Galili, T. 2015. dendextend: an R package for visualizing, adjusting and comparing trees of hierarchical clustering. *Bioinformatics* 31: 3718–3720.

Gao, J.-G., Y.-H. Wu, G.-D. Xu, W.-Q. Li, G.-H. Yao, J. Ma, and P. Liu. 2012. Phylogeography of *Ulmus elongata* based on Fourier transform-infrared spectroscopy (FTIR), thermal gravimetric and differential thermal analyses. *Biochemical Systematics and Ecology* 40: 184–191.

Gompert, Z., and K. E. Mock. 2017. Detection of individual ploidy levels with genotyping-by-sequencing (GBS) analysis. *Molecular Ecology Resources* 17: 1156–1167.

Hill, N., J. Petersen, J. Stuedmann, and F. Barton II. 1988. Prediction of percentage leaf in stratified canopies of alfalfa with near infrared reflectance spectroscopy. *Crop Science* 28: 354–358.

Höpke, J., G. Brewer, S. Dodsworth, E. M. Ortiz, and D. C. Albach. 2018. DNA extraction from old herbarium material of *Veronica* subgen. *Pseudolysimachium* (Plantaginaceae). *Ukrainian Botanical Journal* 75: 564–575.

Huang, K., D. W. Dunn, Z. Li, P. Zhang, Y. Dai, and B. Li. 2019. Inference of individual ploidy level using codominant markers. *Molecular Ecology Resources* 19: 1218–1229.

Hull-Sanders, H. M., R. H. Johnson, H. A. Owen, and G. A. Meyer. 2009. Effects of polyploidy on secondary chemistry, physiology, and performance of native and invasive genotypes of *Solidago gigantea* (Asteraceae). *American Journal of Botany* 96: 762–770.

Jombart, T., and I. Ahmed. 2011. adegenet 1.3-1: New tools for the analysis of genome-wide SNP data. *Bioinformatics* 27: 3070–3071.

Lang, C., F. R. C. Costa, J. L. C. Camargo, F. M. Durgante, and A. Vicentini. 2015. Near infrared spectroscopy facilitates rapid identification of both young and mature Amazonian tree species. *PLoS ONE* 10: e0134521.

Lavania, U. C., S. Srivastava, S. Lavania, S. Basu, N. K. Misra, and Y. Mukai. 2012. Autopolyploidy differentially influences body size in plants, but facilitates enhanced accumulation of secondary metabolites, causing increased cytosine methylation. *The Plant Journal* 71: 539–549.

Ligges, U., T. Short, P. Kienzle, S. Schnackenberg, D. Billinghurst, H.-W. Borchers, A. Carezia, et al. 2021. Signal: signal processing. Website: https://cran.r-project.org/web/packages/signal/index.html [accessed 2 March 2023].

Lomax, B. H., J. Hilton, R. M. Bateman, G. R. Upchurch, J. A. Lake, I. J. Leitch, A. Cromwell, and C. A. Knight. 2014. Reconstructing relative genome size of vascular plants through geological time. *New Phytologist* 201: 636–644.

López-González, N., S. Andrés-Sánchez, B. M. Rojas-Andrés, and M. M. Martínez-Ortega. 2018. Divide and conquer! Data-mining tools and sequential multivariate analysis to search for diagnostic morphological characters within a plant polyploid complex (*Veronica* subsect. *Pentasepalae*, Plantaginaceae). *PLoS ONE* 13: e0199818.

Lu, H.-F., J.-B. Shen, X.-Y. Lin, and J.-L. Fu. 2008a. Relevance of Fourier transform infrared spectroscopy and leaf anatomy for species classification in *Camellia* (Theaceae). *Taxon* 57: 1274–1278E.

Lu, H. F., B. Jiang, Z. G. Shen, J. B. Shen, Q. F. Peng, and C. G. Cheng. 2008b. Comparative leaf anatomy, FTIR discrimination and biogeographical analysis of *Camellia* section *Tuberculata* (Theaceae) with a discussion of its taxonomic treatments. *Plant Systematics and Evolution* 274: 223.

Lunardon, N., G. Menardi, and N. Torelli. 2014. ROSE: A package for binary imbalanced learning. *R Journal* 6: 79–89.

Meireles, J. E., J. Cavender-Bares, P. A. Townsend, S. Ustin, J. A. Gamon, A. K. Schweiger, M. E. Schaepman, et al. 2020. Leaf reflectance spectra capture the evolutionary history of seed plants. *New Phytologist* 228: 485–493.

Meudt, H. M., B. M. Rojas-Andrés, J. M. Prebble, E. Low, P. J. Garnock-Jones, and D. C. Albach. 2015. Is genome downsizing associated with diversification in polyploid lineages of *Veronica*? *Botanical Journal of the Linnean Society* 178: 243–266.

Mtileni, M. P., N. Venter, and K. L. Glennon. 2021. Ploidy differences affect leaf functional traits, but not water stress responses in a mountain endemic plant population. *South African Journal of Botany* 138: 76–83.

Norris, K., R. Barnes, J. Moore, and J. Shenk. 1976. Predicting forage quality by near infrared reflectance spectroscopy. *Journal of Animal Science* 43: 889–897.

Ollinger, S. V. 2011. Sources of variability in canopy reflectance and the convergent properties of plants. *New Phytologist* 189: 375–394.

Ownbey, M., and G. D. McCollum. 1954. The chromosomes of *Tragopogon*. *Rhodora* 56: 7–21.

Podwyszyńska, M., E. Gabryszewska, B. Dyki, A. A. Steptowska, A. Kowalski, and A. Jasiński. 2015. Phenotypic and genome size changes (variation) in synthetic tetraploids of daylily (*Hemerocallis*) in relation to their diploid counterparts. *Euphytica* 203: 1–16.

Podwyszyńska, M., M. Markiewicz, A. Broniarek-Niemiec, B. Matysiak, and A. Marasek-Ciolakowska. 2021. Apple autotetraploids with enhanced resistance to apple scab (*Venturia inaequalis*) due to genome duplication-phenotypic and genetic evaluation. *International Journal of Molecular Science* 22: 527.

Pontius, J., R. Hallett, and M. Martin. 2005. Assessing hemlock decline using visible and near-infrared spectroscopy: indices comparison and algorithm development. *Applied Spectroscopy* 59: 836–843.

Prananto, J. A., B. Minasny, and T. Weaver. 2021. Rapid and cost-effective nutrient content analysis of cotton leaves using near-infrared spectroscopy (NIRS). *PeerJ* 9: e11042.

Prata, E. M. B., C. Sass, D. P. Rodrigues, F. M. C. B. Domingos, C. D. Specht, G. Damasco, P. V. A. Fine, and C. C. Ribas. 2018. Towards integrative taxonomy in Neotropical botany: disentangling the *Pagamea guianensis* species complex (Rubiaceae). *Botanical Journal of the Linnean Society* 188: 213–231.

Qiao, G., M. Liu, K. Song, H. Li, H. Yang, Y. Yin, and R. Zhuo. 2017. Phenotypic and comparative transcriptome analysis of different ploidy plants in *Dendrocalamus latiflorus* Munro. *Frontiers in Plant Science* 8: 1371.

R Core Team. 2014. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Website: http://www.R-project.org/ [accessed 2 March 2023].

Robinson, A. R., and S. D. Mansfield. 2009. Rapid analysis of poplar lignin monomer composition by a streamlined thioacidolysis procedure and near-infrared reflectance-based prediction modeling. *The Plant Journal* 58: 706–714.

Salman-Minkov, A., N. Sabath, and I. Mayrose. 2016. Whole-genome duplication as a key factor in crop domestication. *Nature Plants* 2: 16115.

Serapiglia, M. J., F. E. Gouker, J. F. Hart, F. Unda, S. D. Mansfield, A. J. Stipanovic, and L. B. Smart. 2015. Ploidy level affects important biomass traits of novel shrub willow (*Salix*) hybrids. *BioEnergy Research* 8: 259–269.

Shenk, J. S., and M. O. Westerhaus. 1993. Analysis of agriculture and food products by near infrared reflectance spectroscopy. Infrasoft International, Luxembourg.

Siadjeu, C., E. Mayland-Quellhorst, and D. C. Albach. 2018. Genetic diversity and population structure of trifoliate yam (*Dioscorea dumetorum* Kunth) in Cameroon revealed by genotyping-by-sequencing (GBS). *BMC Plant Biology* 18: 359.

Soltis, D. E., P. S. Soltis, D. W. Schemske, J. F. Hancock, J. N. Thompson, B. C. Husband, and W. S. Judd. 2007. Autopolyploidy in angiosperms: have we grossly underestimated the number of species? *Taxon* 56: 13–30.

Soltis, P. S., X. Liu, D. B. Marchant, C. J. Visger, and D. E. Soltis. 2014. Polyploidy and novelty: Gottlieb's legacy. *Philosophical Transactions of the Royal Society B: Biological Sciences* 369: 1648.

Stasinski, L., D. M. White, P. R. Nelson, R. H. Ree, and J. E. Meireles. 2021. Reading light: leaf spectra capture fine-scale diversity of closely related, hybridizing arctic shrubs. *New Phytologist* 232: 2283–2294.

Strgulc Krajšek, S., P. Buh, A. Zega, and S. Kreft. 2008. Identification of herbarium whole-leaf samples of *Epilobium* species by ATR-IR spectroscopy. *Chemistry and Biodiversity* 5: 310–317.

Suda, J., and P. Trávníček. 2006. Reliable DNA ploidy determination in dehydrated tissues of vascular plants by DAPI flow cytometry: new prospects for plant research. *Cytometry Part A* 69A: 273–280.

Tan, G.-Y., and G. M. Dunn. 1973. Relationship of stomatal length and frequency and pollen-grain diameter to ploidy level in *Bromus inermis* Leyss. *Crop Science* 13: 332–334.

Tate, J. A., V. V. Symonds, A. N. Doust, R. J. A. Buggs, E. Mavrodiev, L. C. Majure, P. S. Soltis, and D. E. Soltis. 2009. Synthetic polyploids of *Tragopogon miscellus* and *T. mirus* (Asteraceae): 60 years after Ownbey's discovery. *American Journal of Botany* 96: 979–988.

Trávníček, B., and V. Vinter. 1999. Studium der Beziehung der Ploidie zur Pollengrösse und Stomatalänge bei der Art *Pseudolysimachion maritimum* (*Veronica longifolia* auct., Scrophulariaceae). *Acta Universitatis Palackianae Olomucensis Facultas Rerum Naturalium* 37: 35–45.

Tucker, M. R., C. Ma, J. Phan, K. Neumann, N. J. Shirley, M. G. Hahn, D. Cozzolino, and R. A. Burton. 2017. Dissecting the genetic basis for seed coat mucilage heteroxylan biosynthesis in *Plantago ovata* using gamma irradiation and infrared spectroscopy. *Frontiers in Plant Science* 8: 326.

Warner, D. A., and G. E. Edwards. 1993. Effects of polyploidy on photosynthesis. *Photosynthesis Research* 35: 135–147.

Werner, I., S. Glasl, and G. Reznicek. 2006. Infrared spectroscopy as a tool for chemotaxonomic investigations within the *Achillea millefolium* group. *Chemistry and Biodiversity* 3: 27–33.

Windham, M. D., K. M. Pryer, D. B. Poindexter, F. W. Li, C. J. Rothfels, and J. B. Beck. 2020. A step-by-step protocol for meiotic chromosome counts in flowering plants: A powerful and economical technique revisited. *Applications in Plant Sciences* 8: e11342.

Xue, H., B. Zhang, J.-R. Tian, M.-M. Chen, Y.-Y. Zhang, Z.-H. Zhang and Y. Ma 2017. Comparison of the morphology, growth and development of diploid and autotetraploid 'Hanfu' apple trees. *Scientia Horticulturae* 225: 277–285.

Zhang, S., Z. Xia, C. Li, X. Wang, X. Lu, W. Zhang, H. Ma, et al. 2022. Chromosome-scale genome assembly provides insights into speciation of allotetraploid and massive biomass accumulation of elephant grass (*Pennisetum purpureum* Schum.). *Molecular Ecology Resources* 22: 2363–2378.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

Appendix S1. Location, voucher code, and herbarium ID for each sample of *Veronica longifolia* and *V. spicata* included in both the living and herbarium data sets. The ploidy estimated using flow cytometry, neural network (NNET), and discriminant analysis of principal components (DAPC) is indicated. Training the NNET was based on a data set containing only herbarium specimens for the herbarium predictions (one average spectrum per specimen, using both leaf surfaces, with three layers comprising 36, 11, and four neurons), while for fresh material predictions NNET was trained with a combined data set (one average spectrum per specimen, using both leaf surfaces, with three layers comprising 41, 17, and seven neurons). The DAPC predictions for the herbarium materials were obtained with a model trained on only herbarium spectra (one average spectrum per specimen, including only adaxial leaf reflectance spectra; 50 principal components [PCs]), while for fresh material predictions, the training set consisted of only fresh specimen spectra (one averaged among each sampling point, typically resulting in two averaged spectra per specimen; 20 PCs).

Appendix S2. R script for the reading and initial analysis of the IRS spectra.

Appendix S3. R script for the data analysis (hierarchical clustering, discriminant analysis of principal components [DAPC], and neural network [NNET]).

Appendix S4. Hierarchical clustering analysis tree obtained for the (A) herbarium material and (B) living individuals. Specimen labels that start with "sp" and "lo" indicate *Veronica spicata* and *V. longifolia*, respectively.

Appendix S5. Discriminant analysis of principal components (DAPC) scatterplots for the (A) herbarium material and (B) fresh material. The best-discriminating results are shown according to the results summarized in Table 1.

Appendix S6. Separate analyses of *Veronica spicata* and *V. longifolia*. The percentages of correct predictions for the analyses separated by species obtained with different data subsets and methods, using optimized parameters, are presented. The values in parentheses after the percentages indicate the number of principal components (PCs) for the discriminant analysis of principal components (DAPC) and the number of neurons on each layer for the neural network (NNET). For the herbarium specimens, $n = 37$ *V. spicata* and $n = 37$ *V. longifolia*; for the fresh material, $n = 18$ *V. spicata* and $n = 15$ *V. longifolia*.

Appendix S7. Two-proportion Z-test results on higher accuracies obtained in the optimization test. The k-fold values of 25, 11, and 35 were used for the herbarium, fresh, and combined data sets, respectively.

How to cite this article: Buono, D., and D. C. Albach. 2023. Infrared spectroscopy for ploidy estimation: An example in two species of *Veronica* using fresh and herbarium specimens. *Applications in Plant Sciences* 11(2): e11516. https://doi.org/10.1002/aps3.11516