ISSN 1479-2621

Plant Genetic Resources: Characterization and Utilization (2017) 15(4); 286–295
doi:10.1017/S1479262115000623

Comparative efficacy of four candidate DNA barcode regions for identification of *Vicia* species

Sebastin Raveendar[1], Jung-Ro Lee[1], Donghwan Shim[2], Gi-An Lee[1], Young-Ah Jeon[1], Gyu-Taek Cho[1], Kyung-Ho Ma[1], Sok-Young Lee[1], Gi-Ho Sung[3,4]* and Jong-Wook Chung[1]*

[1]*National Agrobiodiversity Centre, National Academy of Agricultural Science, Rural Development Administration, Jeonju 560-500, Republic of Korea,* [2]*Department of Forest Genetic Resources, Korea Forest Research Institute, Suwon 441-350, Republic of Korea,* [3]*Institute for Bio-medical Convergence, College of Medicine, Catholic Kwandong University, Gangneung 210-701, Republic of Korea and* [4]*Catholic Kwandong University, International St. Mary's Hospital, Incheon Metropolitan City 404-834, Republic of Korea*

Received 15 July 2015; Revised 15 October 2015; Accepted 6 November 2015

First published online 11 December 2015

Abstract

The genus *Vicia* L., one of the earliest domesticated plant genera, is a member of the legume tribe *Fabeae* of the subfamily *Papilionoideae* (*Fabaceae*). The taxonomic history of this genus is extensive and controversial, which has hindered the development of taxonomic procedures and made it difficult to identify and share these economically important crop resources. Species identification through DNA barcoding is a valuable taxonomic classification tool. In this study, four DNA barcodes (ITS2, *matK*, *rbcL* and *psbA-trnH*) were evaluated on 110 samples that represented 34 taxonomically best-known species in the *Vicia* genus. Topologies of the phylogenetic trees based on an individual locus were similar. Individual locus-based analyses could not discriminate closely related *Vicia* species. We proposed a concatenated data approach to increase the resolving power of ITS2. The DNA barcodes *matK*, *psbA-trnH* and *rbcL* were used as an additional tool for phylogenetic analysis. Among the four barcodes, three-barcode combinations that included *psbA-trnH* with any two of the other barcodes (ITS2, *matK* or *rbcL*) provided the best discrimination among *Vicia* species. Species discrimination was assessed with bootstrap values and considered successful only when all the conspecific individuals formed a single clade. Through sequencing of these barcodes from additional *Vicia* accessions, 17 of the 34 known *Vicia* species could be identified with varying levels of confidence. From our analyses, the combined barcoding markers are useful in the early diagnosis of targeted *Vicia* species and can provide essential baseline data for conservation strategies, as well as guidance in assembling germplasm collections.

Keywords: DNA barcoding; phylogenetic analysis; species differentiation; *Vicia*

Introduction

The genus *Vicia* L. is a member of the legume tribe *Fabeae* of the subfamily *Papilionoideae* (*Fabaceae*) (Frediani *et al*., 2004). This temperate herbaceous genus comprises 210 annual or perennial species that are

* Corresponding authors.
E-mail: sung97330@gmail.com; jwchung73@korea.kr

widely distributed throughout the temperate regions (Maxted, 1993; Jaaska, 2005). Archaeological evidence suggests that the Mediterranean region is the principal centre of diversification (Naranjo *et al.*, 1998), although secondary centres with high genetic variability have been found in southern Siberia, Europe, and North and South America, including Argentina (Maxted, 1995). Knowledge of the natural distribution, taxonomy and production potential of the *Vicia* genus has not been fully exploited. Knowledge of genetic diversity is a useful tool in genebank management and breeding programmes to tag germplasm, identify and/or eliminate duplicates in the genebank stock, and establish core collections (Ma *et al.*, 2009).

Identification of *Vicia* species using a method based on morphological characteristics has its limitations, as it is difficult to describe the genetic variation present in the *Vicia* species (Haider *et al.*, 2012). Hosseinzadeh *et al.* (2008) identified several *Vicia* species with intermediate morphological characters, which make it difficult to distinguish species. Large-scale structural changes have been observed in *Vicia* chromosomes, and cytological and karyological studies have been performed for taxonomical discrimination (Maxted *et al.*, 1991; Navratilova *et al.*, 2003). However, earlier studies had limited success in discriminating *Vicia* species. Therefore, a robust and reliable method is needed to discriminate *Vicia* species in order to secure their diversity.

DNA barcoding has recently been proposed as a taxonomic tool that allows taxonomists to revise, describe, reorder and even identify species (Gregory, 2005). The barcoding method has been used to identify a specific region in the plant genome that can be sequenced routinely in a diverse set of samples, resulting in easily comparable data that enable species to be distinguished (Chen *et al.*, 2010). Many recent papers have reviewed the applications of DNA barcoding (Vijayan and Tsou, 2010; Hollingsworth *et al.*, 2011). Numerous single and combined loci have been proposed as barcode sequences (Chase *et al.*, 2007; Kress and Erickson, 2007), but no consensus has emerged yet on the use of a standard genomic region.

An ideal barcode sequence must allow efficient discrimination of closely related species in a set of diverse samples. Several researchers have already demonstrated the potential of internal transcribed spacer (ITS) for taxonomic classification and phylogenetic reconstruction of *Vicia* L. species (Endo *et al.*, 2008; Ruffini Castiglione *et al.*, 2011; Haider *et al.*, 2012; Ruffini Castiglione *et al.*, 2012; Schaefer *et al.*, 2012; Caputo *et al.*, 2013; Shiran *et al.*, 2014). In a previous study, we tested the use of ITS2 and *matK* for taxonomic classification of *Vicia* L. species (Raveendar *et al.*, 2015). Here, we evaluated the most widely used DNA barcodes (ITS2, *matK*, *psbA-trnH* and *rbcL*) in discriminating *Vicia* species, the earliest domesticated plant genus.

Materials and methods

Plant material and DNA extraction

A total of 110 genebank accessions representing 34 *Vicia* species were provided by the Genetic Resource Center of the National Academy of Agricultural Science, Rural Development Administration, Republic of Korea (Supplementary Table S1, available online). To identify the subset of the proposed barcoding loci required to distinguish *Vicia* species, we sampled one to five individual accessions. Seeds were germinated and leaf tissues were harvested from 3-week-old seedlings grown under controlled conditions in the greenhouse. Total DNA was extracted using the DNeasy® Plant Mini-kit (Qiagen, Valencia, CA, USA), according to the manufacturer's instructions. Fresh leaf tissue from each accession was used for each extraction and pulverized in liquid nitrogen. DNA was resuspended in 100 μl of water, and dilutions were made to 10 ng/μl followed by storage at −20 or −80°C. Genomic DNA was quantified using a Nanodrop/UVS-99 instrument (ACTGene, Piscataway, NJ, USA), and the A260/A280 nm ratio was established. DNA quality was confirmed on a 0.8% agarose gel.

PCR amplification and sequencing

Sequences of the universal primers for four barcoding regions (ITS2, *matK*, *psbA-trnH* and *rbcL*) and thermocycling reaction conditions were obtained from Chen *et al.* (2010). Amplification reactions were performed in a total volume of 20 μl containing 1× PCR buffer, 0.1 mM primers, 0.2 mM each dNTP, 1 U *Taq* DNA polymerase and 200 ng of template DNA. Sequencing of PCR amplicons was performed by Macrogen Company, South Korea. Forward and reverse sequences were assembled and aligned for consensus sequences using the Sequence Scanner (Version 1.0). All sequences were submitted to the NCBI GenBank (Supplementary Table S1, available online).

Phylogenetic analysis

Consensus sequences of each region (ITS2, *matK*, *psbA-trnH* and *rbcL*) were manually edited with MEGA6 (Tamura *et al.*, 2013) and aligned using the ClustalW program. Manual adjustments in the alignment of the nucleotide sequences for the barcoding region were made to improve alignment. All variable sites were rechecked with the original trace files. To assess the barcoding gap, the relative distribution of pairwise genetic distances was calculated using TAXONDNA (Meier *et al.*, 2006) based on the Kimura-2 parameter (K2P)-corrected pairwise distance model. For efficient species discrimination, aligned barcodes (ITS2, *matK*, *psbA-trnH* and *rbcL*) were evaluated by bootstrap analysis (5000 replicates) with pairwise deletion (Felsenstein, 1985). K2P distances (Kimura, 1980) were calculated to construct an unrooted neighbour-joining (NJ) dendrogram using MEGA6 (Tamura *et al.*, 2013). Species discrimination was considered successful only when all the conspecific individuals formed a single clade. Species identification was assessed with bootstrap values, as described by Liu *et al.* (2011). Additionally, the functions of the 'best match' and the 'best close match' based on the presence or absence of a 'barcode gap' were used to test the individual-level discrimination rates for each single marker and all possible combinations using TAXONDNA based on the K2P-corrected distance model (Meyer and Paulay, 2005). Using the barcode gap criterion, a species was distinct from its nearest neighbour (NN) if its maximum intraspecific distance was less than the distance to its NN sequence. Finally, the publicly available National Center for Biotechnology Information (NCBI) database for reported DNA sequences, the nucleotide BLAST (blastn) application, was searched for barcode sequences.

Results

Sequence character of the four loci

When the universal primers were evaluated separately, all the four loci showed very high success rates (93–100%) for PCR amplification and sequencing (Table 1). All individual loci had length variation, with ranges of 319–335 bp for ITS2, 685–687 bp for *matK*, 307–567 bp for *psbA-trnH* and 545–549 bp for *rbcL*. The GC content ranged from 27 to 49% (Table 1). Efforts to align all barcode sequences using ClustalW were successful across all species. However, the ClustalW-aligned sequences showed considerable size variation among the four targeted loci. The aligned length was 342 bp for ITS2, 687 bp for *matK*, 567 bp for *psbA-trnH* and 549 bp for *rbcL* (Table 2). A total of 426 sequences for 110 individuals were available from the 34 *Vicia* species. Sequence analysis of the multiple alignment revealed that nucleotide diversity (π) was similar among the four loci, with *psbA-trnH* exhibiting the highest π value (0.11). Sequence characteristics of the four barcoding regions are summarized in Table 2.

Table 1. Primer sequences, PCR conditions and sequencing successes for the samples used in this study

Regions	Primer sequences (5′–3′)[a]	PCR conditions[a]	Sequencing sample tried	Sequencing reaction failure, *N* (%)	Sequence length	GC content (%)
ITS2	Forward: ATGCGATACTTGGTGTGAAT Reverse: GACGCTTCTCCAGACTACAAT	94°C 3 min; 95°C 30 s, 56°C 30 s, 72°C 30 s, 35 cycles; 72°C 7 min	110	3 (2.72)	319–335	49.0
matK	Forward: CGTACAGTACTTTTGTGTTTACGAG Reverse: ACCCAGTCCATCTGGAAATCTTGGTTC	94°C 2 min 30 s; 94°C 30 s, 54°C 30 s, 72°C 30 s, 10 cycles and 88°C 30 s, 54°C 30 s, 72°C 30 s, 25 cycles; 72°C 10 min	110	3 (2.72)	685–687	30.5
psbA-trnH	Forward: GTTATGCATGAACGTAATGCTC Reverse: CGCGCATGGTGGATTCACAATCC	95°C 2 min 30 s; 95°C 30 s, 58°C 30 s, 64°C 1 min, 35 cycles; 72°C 7 min	110	8 (7.27)	307–567	26.9
rbcL	Forward: ATGTCACCACAAACAGAGACTAAAGC	94°C 2 min 30 s; 94°C 30 s, 54°C 30 s, 72°C 30 s, 10 cycles and 88°C 30 s, 54°C 30 s, 72°C 30 s, 25 cycles; 72°C 10 min	110	0 (0)	545–549	43.2

[a] Chen *et al.* (2010).

https://doi.org/10.1017/S1479262115000623 Published online by Cambridge University Press

Table 2. Genetic diversity of marker combinations among the four markers used in this study

	Individuals (n)	No. of species	Aligned length	Variable characters	No. of segregating sites	Nucleotide diversity (π)	Species resolved (%)	Discrimination (%)
ITS2	107	35	342	8	63	0.021440	78.5	66
matK	107	35	687	13	150	0.031559	80.7	71
psbA-trnH	102	34	576	10	154	0.117768	83.3	68
rbcL	110	35	549	9	42	0.016032	64.4	71
ITS2 + *matK*	98	34	1029	15	208	0.029145	92.1	68
ITS2 + *psbA-trnH*	98	34	918	12	222	0.059444	86.6	82
ITS2 + *rbcL*	98	34	891	12	103	0.019670	89.3	73
matK + *psbA-trnH*	98	34	1263	14	306	0.046866	97.7	76
matK + *rbcL*	98	34	1236	15	187	0.024564	87.2	82
psbA-trnH + *rbcL*	93	34	1125	11	173	0.038938	88.2	82
ITS2 + *matK* + *psbA-trnH*	93	34	1605	14	360	0.041923	87.9	85
ITS2 + *matK* + *rbcL*	93	34	1578	15	248	0.024628	91.6	82
ITS2 + *psbA-trnH* + *rbcL*	93	34	1467	12	243	0.034454	91.7	85
matK + *psbA-trnH* + *rbcL*	93	34	1812	14	334	0.035160	93.0	85
ITS2 + *matK* + *psbA-trnH* + *rbcL*	86	34	2154	15	415	0.034783	91.7	88

Genetic distance and barcoding gap assessment

The relative distribution of K2P distances based on single barcodes and combinations ITS2 + *matK* + *rbcL* and ITS2 + *matK* + *psbA-trnH* + *rbcL* demonstrated significant overlap and no barcoding gap (Fig. 1). Among the single barcodes, *psbA-trnH* had the highest variation in inter-specific divergence, followed by ITS2, when compared with the range of intra-specific distances (Fig. 1). Similarly, among the barcode combinations, ITS2 + *matK* + *psbA-trnH* + *rbcL* showed the highest variation in inter-specific divergence compared with the range of intra-specific distances (Fig. 1).

Utility of barcodes for resolving species

The utility of loci and their sequences for barcoding alone and in multigene combinations is presented in Table 2. The results indicated that *psbA-trnH* resolved a much higher percentage of species (83%) than did other loci. However, multigene combinations marginally resolved a greater percentage of taxa and provided greater support compared with *psbA-trnH* alone (Table 2); *matK* + *psbA-trnH* provided considerably higher resolution (98%). When comparing the results of the 'best match' and 'best close match' analyses, the former always showed higher or equal individual identification rates compared with the latter (Table 3). In the 'best match' analysis, each query found the closest barcode match. If both sequences were from the same species, the identification was considered a success, whereas mismatched names were counted as failures. Several equally good best matches from different species were considered ambiguous. The 'best close match' analysis determined that all queries without barcode matches below the threshold value were unidentified, and their identity was compared with the species identity of their closest barcode. If the names were identical, the query was considered an identification success. The identification was considered a failure if the names were mismatched, and it was considered ambiguous when several equally good best matches were found that belonged to a minimum of two species. Identification efficiency at the individual level was higher than at the species level for each barcode and all combinations (Fig. 1; Table 3).

Discrimination efficiency in Vicia

The NJ tree was used to assess identification efficiency within the genus *Vicia*. Phylogenetic trees based on unrooted NJ analysis and K2P (Kimura, 1980) distances of the nucleotide sequences of the four loci were topologically similar (Fig. 2). However, trees of each locus and

https://doi.org/10.1017/S1479262115000623 Published online by Cambridge University Press

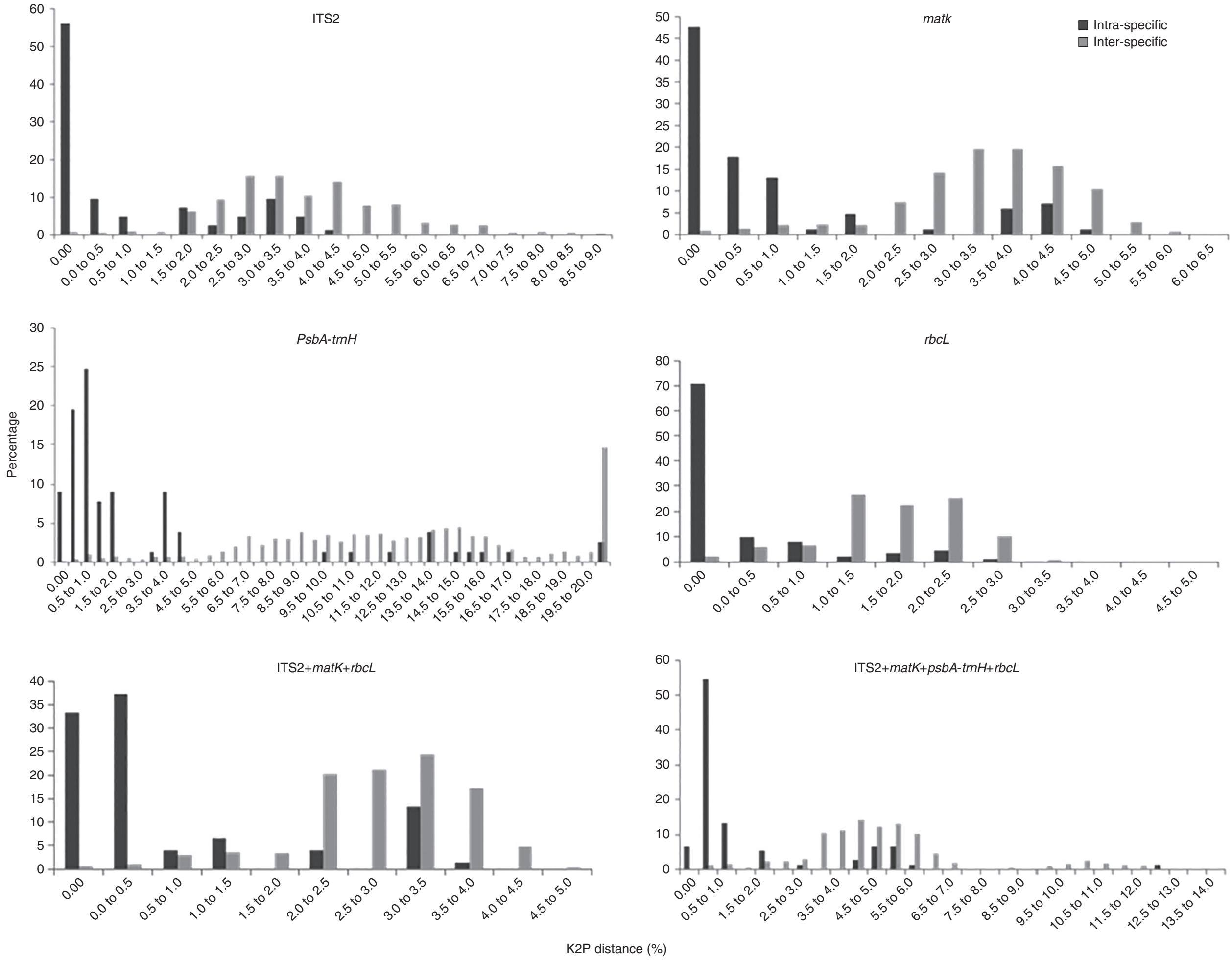

Fig. 1. Relative distribution of K2P distances across all the sequence pairs of *Vicia* datasets for different markers.

multigene combinations generated by NJ tree analyses differed in their branch length values (Supplementary Figs S1–S14, available online). Phylogenetic analyses based on the nucleotide sequences of the four loci were generally successful in discriminating the *Vicia* species (Fig. 2). If individuals within a species or subspecies formed a distinct clade, it was considered a successful identification at the species level. Also, species with a single or fewer than three samples were only considered successful when they formed a distinct clade themselves. The species and subspecies of *V. articulata*, *V. benghalensis*, *V. cassubica*, *V. costata*, *V. cracca*, *V. ervilia*, *V. faba*, *V. hyrcanica*, *V. michauxii*, *V. montbretii*, *V. narbonensis*, *V. sativa* and *V. villosa* were identified successfully using the four DNA barcodes, singly or in multigene combinations (Supplementary Figs S1–S14, available online). Individual and multigene combinations, through NJ tree analyses, separated most of the species and subspecies in the genus *Vicia*, although some species could not be discriminated.

Phylogenetic analysis of the four loci also demonstrated that, for several species, accessions were located in distant clades. For example, an accession of *V. villosa* subsp. *villosa* was located in a clade with the accessions of *V. sativa* (Supplementary Figs S1–S4, available online). Similarly, accessions of *V. anatolica*, *V. amoena*, *V. articulata*, *V. benghalensis* and *V. villosa* were placed in the same clade, with no clear differences. The species *V. amurensis*, *V. anatolica* and *V. hirsuta* were not identified by the loci used, neither singly nor in combination (Supplementary Figs S1–S14, available online).

The NJ tree of the 34 species based on the combinations of the four DNA barcoding regions is shown in Fig. 2. All 86 individuals fell into distinct clades, with high bootstrap support values corresponding to the 34 species. The clustering relationships among the species were similar to those found by Liu *et al.* (2011), who used smaller sample sizes for each species. Topologies of the phylogenetic trees based

https://doi.org/10.1017/S1479262115000623 Published online by Cambridge University Press

Table 3. Number (rates) of sample identification based on the 'best match' and 'best close match' functions using TAXONDNA software for each DNA barcoding marker and combinations from 110 individuals

	Best match, *N* (%)			Best close match, *N* (%)				
Barcoding region[a]	Correct	Ambiguous	Incorrect	Correct	Ambiguous	Incorrect	No match	Threshold (%)
I	44 (41.12)	45 (42.05)	18 (16.82)	44 (41.12)	45 (42.05)	17 (15.88)	1 (0.93)	0.68
M	52 (48.59)	38 (35.51)	17 (15.88)	52 (48.59)	38 (35.51)	17 (15.88)	0 (0.0)	0.61
P	62 (60.78)	10 (9.8)	30 (29.41)	59 (57.84)	10 (9.8)	25 (24.5)	8 (7.84)	3.21
R	39 (35.45)	61 (55.45)	10 (9.09)	39 (35.45)	61 (55.45)	10 (9.09)	0 (0.0)	0.23
I + M	48 (48.97)	32 (32.65)	18 (18.36)	48 (48.97)	32 (32.65)	18 (18.36)	0 (0.0)	0.64
I + P	59 (60.2)	6 (6.12)	33 (33.67)	58 (59.18)	6 (6.12)	28 (28.57)	6 (6.12)	1.77
I + R	38 (38.77)	43 (43.87)	17 (17.34)	38 (38.77)	43 (43.87)	17 (17.34)	0 (0.0)	0.41
M + P	64 (65.3)	8 (8.16)	26 (26.52)	64 (65.3)	8 (8.16)	20 (20.4)	6 (6.12)	1.31
M + R	49 (50.0)	37 (37.75)	12 (12.24)	49 (50.0)	37 (37.75)	12 (12.24)	0 (0.0)	0.47
P + R	57 (58.16)	13 (13.26)	28 (28.57)	56 (57.14)	13 (13.26)	27 (27.55)	2 (2.04)	1.18
I + M + P	62 (63.26)	9 (9.18)	27 (27.55)	61 (62.24)	9 (9.18)	22 (22.44)	6 (6.12)	1.21
I + M + R	50 (51.02)	28 (28.57)	20 (20.4)	50 (51.02)	28 (28.57)	20 (20.4)	0 (0.0)	0.53
I + P + R	57 (58.16)	8 (8.16)	33 (33.67)	56 (57.14)	8 (8.16)	32 (32.65)	2 (2.04)	1.09
M + P + R	63 (64.28)	9 (9.18)	26 (26.52)	63 (64.28)	9 (9.18)	24 (24.48)	2 (2.04)	0.94
I + M + P + R	62 (63.26)	8 (8.16)	28 (28.57)	61 (62.24)	8 (8.16)	27 (27.55)	2 (2.04)	0.95

[a] For the abbreviations, see Table 2.

on sequence concatenation were similar, but some *Vicia* species were placed in different clades when analysed individually. Moreover, phylogenetic trees based on concatenated sequences of the four loci improved the topologies of the phylogenetic tree and successfully discriminated all the 34 *Vicia* species (Fig. 2).

Of the four single barcodes, *matK* and *rbcL* showed the highest discriminatory power, with 71% of all the species discriminated, followed by *psbA-trnH* (68%) and ITS2 (66%). The combination of the four barcodes led to higher discrimination rates (Table 2). Three-marker combinations significantly increased the species discrimination ability when they included *psbA-trnH*. Any combination of three barcodes that included *psbA-trnH* showed the best (85%) species discrimination. The four-way combination, ITS2 + *matK* + *psbA-trnH* + *rbcL*, had higher species discrimination (88%), but some closely related species could not be discriminated with bootstrap support values. Accessions of *V. villosa*, *V. benghalensis*, *V. cracca* and *V. montbretii* were placed with closely related species in the same monophyletic clade. Across all locus combinations and the single loci, 17 of the 34 species received >90% branch support on average, with the highest average for *V. faba*, followed by *V. sativa* (Supplementary Figs S1–S14, available online), indicating a clear genetic distinction of the species.

Discussion

Taxonomy of the *Vicia* L. genus has been problematic, as the taxonomic history of the genus is extensive and contentious (Maxted, 1993). The high economic importance of this genus has led to numerous studies on the molecular characterization and investigation of phylogenetic relationships among *Vicia* species. Various studies have investigated phylogenetic relationships among species based on rDNA (Raina and Ogihara, 1995), *in situ* hybridization with repetitive sequences (Navratilova *et al.*, 2003), RAPD analysis (Haider *et al.*, 2000; Sakowicz and Cieslikowski, 2006), repetitive DNA sequences as probes (Frediani *et al.*, 2004), capillary electrophoresis (Piergiovanni and Taranto, 2005) and SDS–PAGE on seed storage proteins (MirAli *et al.*, 2007). However, none of these studies could resolve the taxonomic problems of the genus *Vicia*. Our research study aimed to validate DNA barcodes that could distinguish *Vicia* species.

DNA barcode evaluation

The universality of the PCR primers and sequencing success are important criteria for DNA barcoding (Chase *et al.*, 2007; Kress and Erickson, 2007). If the loci fail to

https://doi.org/10.1017/S1479262115000623 Published online by Cambridge University Press

Fig. 2. Phylogenetic analysis of *Vicia* species based on the combinations of barcode loci. The NJ tree was developed by applying the K2P method to the nucleotide sequences of the ITS2 + *matK* + *psbA-trnH* + *rbcL* region. Numbers next to the branches are the bootstrap test values.

amplify well, no sequencing data can be generated. Therefore, a valid DNA barcode must be evaluated for PCR amplification success (CBOL Plant Working Group: Hollingsworth *et al.*, 2009). In previous studies, the ITS2 region was suggested for phylogenetic analysis of plant species (Chen *et al.*, 2010; Gao *et al.*, 2010). Recently, nucleotide sequences of some regions of the chloroplast DNA (*matK*, *rpoC1 rpoB*, *trnH-PsbA rbcL*, *atpF-atpH* and *psbK-psbI*) and their combinations were tested for barcoding of plant species (e.g. Starr *et al.*, 2009). Among these DNA regions, *matK* and *rbcL* were accepted as a two-locus DNA barcode by the CBOL (CBOL Plant Working Group: Hollingsworth *et al.*, 2009). In the present study, three chloroplast (*matK*, *psbA-trnH* and *rbcL*) and one nuclear-specific (ITS2) barcoding marker were tested. All regions were successfully amplified, except for the *psbA-trnH*, ITS2 and *matK* regions for a small percentage of individuals (Table 1).

Sequence quality and coverage are important criteria for DNA barcoding. High-quality sequences were routinely obtained for most of the four loci evaluated in this study (Table 1). However, a few ambiguous bases occurred with *psbA-trnH* sequences, which were previously considered a limitation for a barcode due to the potential for alignment ambiguities (CBOL Plant Working Group: Hollingsworth *et al.*, 2009). In this

https://doi.org/10.1017/S1479262115000623 Published online by Cambridge University Press

study, we discovered that the *psbA-trnH* region had the highest sequence length variation due to their highest sequence divergence (Table 2). In this study, PCR amplification success using universal primers showed that they are vital for screening DNA barcodes in the *Vicia* genus.

DNA barcode species resolution

As a single-region barcode, *psbA-trnH* resolved the greatest number of species (Table 2). Multigene combinations improved species resolution when other barcodes were combined with *psbA-trnH* (Table 2). Only marginal gains in taxon resolution (83.3% vs. 97.7%) could be achieved when *psbA-trnH* was included in a two- (*matK* + *psbA-trnH*) or three-locus barcode (*matK* + *psbA-trnH* + *rbcL*). Even though we found few ambiguous bases when sequencing the *psbA-trnH* region, it is unlikely that *psbA-trnH* would significantly increase species resolution. In other words, it would most probably provide the same level of species resolution as observed in *matK*, but it would require more sequencing effort (Chase *et al.*, 2005; Kress and Erickson, 2007).

Among the four barcodes, *psbA-trnH* provided the highest species resolution. Due to the high level of sequence divergence and species discrimination, *psbA-trnH* has been considered the best candidate plant barcode in many studies (Hollingsworth *et al.*, 2011). Similarly, in the present study, the *psbA-trnH* region had the highest sequence length variation and genetic divergence. Therefore, as a single barcode, *psbA-trnH* is the best candidate to distinguish *Vicia* species.

Barcoding discriminates Vicia *species*

We barcoded *Vicia* species that were difficult for taxonomists to differentiate using morphological characters. One of the most difficult tasks in reviving genebanks is the proper maintenance of genetic variation in the form of accessions. There are two risks during revival of accessions: loss of diversity, which requires critical attention to minimum population size, and loss of identity due to migration among accessions (Vencovsky and Crossa, 1999). We found that the misidentified genebank specimens were those that only had vegetative characters, underscoring the difficulty of identifying species. Given the important economic value of *Vicia*, it would be very useful to have a reliable identification tool that can differentiate *Vicia* species by sampling only the leaves for DNA barcoding, which are easily accessible.

The success of molecular identification using DNA barcoding lies in the cohesiveness and distinctness of the clusters in the analysis (Steinke *et al.*, 2009). In the present study, conspecific samples formed monophyletic clusters, supported by a high bootstrap value, which provided reliability of barcoding sequences to identify *Vicia* species. Species discrimination with single DNA barcoding regions was similar, as the topologies of trees were similar (Supplementary Figs S1–S4, available online). DNA barcoding regions generally created separated clusters in the *Vicia* genus. However, several species could not be discriminated and are incorrectly grouped in the NJ tree of Fig. 2, suggesting that there might be misidentification of these accessions. Sample sizes of five to ten specimens per species have been suggested in the DNA barcoding database (www.barcodinglife.org), but optimal representation of intra-specific variation remains unclear (Zhang *et al.*, 2010). It has been reported that the nuclear ITS2 region requires cloning before sequencing because of the allelic polymorphisms, pseudogenes and paralogous copies of the ITS2 region in a plant species (Bailey *et al.*, 2003; King and Roalson, 2008). In contrast, there are no allelic polymorphisms or insertions/deletions in the chloroplast region within the plastid genome. Therefore, we were able to efficiently amplify and sequence-characterize the chloroplast region without cloning.

Sequence data for ITS2, *matK*, *psbA-trnH* and *rbcL* were used to discriminate the *Vicia* species. The ITS region was proposed by others as a suitable barcode (Kress *et al.*, 2005). Genetic relationships among 49 *Vicia* species have recently been analysed using polymorphisms within the region of nrDNA, which includes the internally transcribed spacers ITS1 and ITS2 (Shiran *et al.*, 2014). When only ITS2 was used, it discriminated only 66% of the species, which might be explained by allelic polymorphisms, pseudogenes and paralogous copies of the ITS2 region (Bailey *et al.*, 2003; King and Roalson, 2008).

Gao and Chen (2009) tested the potential of four coding chloroplast regions (*rpoB*, *rpoC1*, *rbcL* and *matK*) and two non-coding nuclear regions (ITS, ITS2) as barcodes for medicinal plants. Similarly, in the present study, species discrimination with barcode combinations (88%) was significantly higher than that with a single barcode (71%). When we searched the NCBI database for barcoding sequences generated in this study for ITS2, *matK*, *psbA-trnH* and *rbcL*, we retrieved accessions containing the highest hits (Supplementary Table S2, available online). We failed to obtain identical sequences for some species (e.g. *V. faba* var. *faba*, *V. faba* var. *minuta* and *V. costata*) as the sequences in the database had not been annotated or the sequences of the specific species were absent from the database.

In conclusion, a total of 110 individuals of 34 species in the *Vicia* genus were used to evaluate the discriminatory power of barcoding with four DNA barcodes. Based on the results from our study, *psbA-trnH* and *matK* are recommended as single DNA barcodes for *Vicia*. The combination of ITS2 + *matK* + *psbA-trnH* + *rbcL* provided the most accurate (100% species ID) and efficient multi-locus DNA barcoding tool to identify *Vicia* species. Single-locus barcoding did not differentiate the 34 *Vicia* species, which was also the conclusion from multiple studies focusing on species other than the *Vicia*. Although the combination of *psbA-trnH* and any two of the other tested markers increased the percentage of species discrimination, further confirmation is required after a more complete sampling of the genus. The K2P-corrected pairwise distance analysis revealed considerable sequence variation that might easily differentiate all *Vicia* species. Giving consideration to universal amplification and divergence as needed, *psbA-trnH* and *matK* could serve as potential markers to discriminate *Vicia* species. It is unlikely that more than two samples, but a minimum of two, would be needed for DNA barcoding of any specific group of plant species.

Supplementary material

To view supplementary material for this article, please visit http://dx.doi.org/10.1017/S1479262115000623

Acknowledgements

This study was carried out with the support of the 'Research Program for Agricultural Science & Technology Development (Project No. PJ008623)' and was supported by the 2014 Postdoctoral Fellowship Program of National Academy of Agricultural Science, Rural Development Administration, Korea.

References

Bailey CD, Carr TG, Harris SA and Hughes CE (2003) Characterization of angiosperm nrDNA polymorphism, paralogy, and pseudogenes. *Molecular Phylogenetics and Evolution* 29: 435–455.

Caputo P, Frediani M, Gelati MT, Venora G, Cremonini R and Ruffini Castiglione M (2013) Karyological and molecular characterisation of subgenus *Vicia* (*Fabaceae*). *Plant Biosystems* 147: 1242–1252.

CBOL Plant Working Group: Hollingsworth PM, Forrest LL, Spouge JL, Hajibabaei M, Ratnasingham S, van der Bank M, Chase MW, Cowan RS, Erickson DL, Fazekas AJ, Graham SW, James KE, Kim KJ, Kress WJ, Schneider H, van AlphenStahl J, Barrett SCH, van den Berg C, Bogarin D, Burgess KS, Cameron KM, Carine M, Chacón J, Clark A, Clarkson JJ, Conrad F, Devey DS, Ford CS, Hedderson TAJ, Hollingsworth ML, Husband BC, Kelly LJ, Kesanakurti PR, Kim JS, Kim YD, Lahaye R, Lee HL, Long DG, Madriñn S, Maurin O, Meusnier I, Newmaster SG, Park CW, Percy DM, Petersen G, Richardson JE, Salazar GA, Savolainen V, Seberg O, Wilkinson MJ, Yi DK and Little DP (2009) A DNA barcode for land plants. *Proceedings of the National Academy of Sciences* 106: 12794–12797.

Chase MW, Salamin N, Wilkinson M, Dunwell JM, Kesanakurthi RP, Haidar N and Savolainen V (2005) Land plants and DNA barcodes: short-term and long-term goals. *Philosophical Transactions of the Royal Society B: Biological Sciences* 360: 1889–1895.

Chase MW, Cowan RS, Hollingsworth PM, van den Berg C, Madrinan S, Petersen G, Seberg O, Jorgsensen T, Cameron KM, Carine M, Pedersen N, Hedderson TAJ, Conrad F, Salazar GA, Richardson JE, Hollingsworth ML, Barraclough TG, Kelly L and Wilkinson M (2007) A proposal for a standardised protocol to barcode all land plants. *Taxon* 56: 295–299.

Chen SL, Yao H, Han JP, Liu C, Song JY, Shi LC, Zhu YJ, Ma XY, Gao T, Pang XH, Luo K, Li Y, Li XW, Jia XC, Lin YL and Leon C (2010) Validation of the ITS2 region as a novel DNA barcode for identifying medicinal plant species. *PLoS ONE* 5: e8613.

Endo Y, Choi BH, Ohashi H and Delgado-Salinas A (2008) Phylogenetic relationships of new world *Vicia* (*Leguminosae*) inferred from nrDNA internal transcribed spacer sequences and floral characters. *Systematic Botany* 33: 356–363.

Felsenstein J (1985) Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* 39: 783–791.

Frediani M, Maggini F, Gelati MT and Cremonini R (2004) Repetitive DNA sequences as probes for phylogenetic analysis in *Vicia* genus. *Caryologia* 57: 379–386.

Gao T and Chen SL (2009) Authentication of the medicinal plants in Fabaceae by DNA barcoding technique. *Planta Medica* 75: 417.

Gao T, Yao H, Song J, Liu C, Zhu Y, Ma X, Pang X, Xu H and Chen S (2010) Identification of medicinal plants in the family Fabaceae using a potential DNA barcode ITS2. *Journal of Ethnopharmacology* 130: 116–121.

Gregory TR (2005) DNA barcoding does not compete with taxonomy. *Nature* 434: 1067.

Haider A, Hassanin BR, Mahmoud N and Madkour M (2000) *Molecular Characterization of Some Species of the Genus* Vicia. Cairo: Arab Council for Graduate Studies and Scientific Research.

Haider N, Nabulsi I and MirAli N (2012) Identification of species of *Vicia* subgenus *Vicia* (Fabaceae) using chloroplast DNA data. *Turkish Journal of Agriculture and Forestry* 36: 297–308.

Hollingsworth PM, Graham SW and Little DP (2011) Choosing and using a plant DNA barcode. *PLoS ONE* 6: e19254.

Hosseinzadeh Z, Pakravan M and Tavassoli A (2008) Micromorphology of seed in some *Vicia* species from Iran. *Rostaniha* 9: 96–107.

Jaaska V (2005) Isozyme variation and phylogenetic relationships in *Vicia* subgenus *Cracca* (Fabaceae). *Annals of Botany* 96: 1085–1096.

Kimura M (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution* 16: 111–120.

King MG and Roalson EH (2008) Exploring evolutionary dynamics of nrDNA in *Carex* subgenus *Vignea* (Cyperaceae). *Systematic Botany* 33: 514–524.

Kress WJ and Erickson DL (2007) A two-locus global DNA barcode for land plants: the coding rbcL gene complements the non-coding trnH-psbA spacer region. *PLoS ONE* 2: e508.

Kress WJ, Wurdack KJ, Zimmer EA, Weigt LA and Janzen DH (2005) Use of DNA barcodes to identify flowering plants. *Proceedings of the National Academy of Sciences* 102: 8369–8374.

Liu J, Moller M, Gao LM, Zhang DQ and Li DZ (2011) DNA barcoding for the discrimination of Eurasian yews (*Taxus* L., Taxaceae) and the discovery of cryptic species. *Molecular Ecology Resources* 11: 89–100.

Ma KH, Kim NS, Lee GA, Lee SY, Lee JK, Yi JY, Park YJ, Kim TS, Gwag JG and Kwon SJ (2009) Development of SSR markers for studies of diversity in the genus *Fagopyrum*. *Theoretical and Applied Genetics*. 119: 1247–1254.

Maxted N (1993) A phenetic investigation of *Vicia* L. subgenus *Vicia* (Leguminosae, Vicieae). *Botanical Journal of the Linnean Society* 111: 155–182.

Maxted N (1995) *An Ecogeographic Study of* Vicia *Subgenus* Vicia. Systematic and Ecogeographic Studies in Crop Genepools 8. Rome, Italy: IBPGR.

Maxted N, Callimassia MA and Bennett MD (1991) Cytotaxonomic studies of eastern Mediterranean *Vicia* species (Leguminosae). *Plant Systematics and Evolution* 177: 221–234.

Meier R, Shiyang K, Vaidya G and Ng PKL (2006) DNA barcoding and taxonomy in Diptera: a tale of high intraspecific variability and low identification success. *Systematic Biology* 55: 715–728.

Meyer CP and Paulay G (2005) DNA barcoding: error rates based on comprehensive sampling. *PLoS Biology* 3: e422.

MirAli N, El-Khouri S and Rizq F (2007) Genetic diversity and relationships in some *Vicia* species as determined by SDS-PAGE of seed proteins. *Biologia Plantarum* 51: 660–666.

Naranjo CA, Ferrari MR, Palermo AM and Poggio L (1998) Karyotype, DNA content and meiotic behaviour in five South American species of *Vicia* (Fabaceae). *Annals of Botany* 82: 757–764.

Navratilova A, Neumann P and Macas J (2003) Karyotype analysis of four *Vicia* species using *in situ* hybridization with repetitive sequences. *Annals of Botany* 91: 921–926.

Piergiovanni AR and Taranto G (2005) Specific differentiation in *Vicia* genus by means of capillary electrophoresis. *Journal of Chromatography A* 1069: 253–260.

Raina SN and Ogihara Y (1995) Ribosomal DNA repeat unit polymorphism in 49 *Vicia* species. *Theoretical and Applied Genetics* 90: 477–486.

Raveendar S, Lee JR, Park JW, Lee GA, Jeon YA, Lee YJ, Cho GT, Ma KH, Lee SY and Chung JW (2015) Potential use of ITS2 and *matK* as a two-locus DNA barcode for identification of *Vicia* species. *Plant Breeding and Biotechnology* 3: 58–66.

Ruffini Castiglione MR, Frediani M, Gelati MT, Ravalli C, Venora G, Caputo P and Cremonini R (2011) Cytology of *Vicia* species. X. Karyotype evolution and phylogenetic implication in *Vicia* species of the sections *Atossa*, *Microcarinae*, *Wiggersia* and *Vicia*. *Protoplasma* 248: 707–716.

Ruffini Castiglione MR, Frediani M, Gelati MT, Venora G, Giorgetti L, Caputo P and Cremonini R (2012) Cytological and molecular characterization of *Vicia barbazitae* Ten. & Guss. *Protoplasma* 249: 779–788.

Sakowicz T and Cieslikowski T (2006) Phylogenetic analyses within three sections of the genus *Vicia*. *Cellular & Molecular Biology Letters* 11: 594–615.

Schaefer H, Hechenleitner P, Santos-Guerra A, de Sequeira MM, Pennington RT, Kenicer G and Carine MA (2012) Systematics, biogeography, and character evolution of the legume tribe *Fabeae* with special focus on the middle-Atlantic island lineages. *BMC Evolutionary Biology* 12: 250.

Shiran B, Kiani S, Sehgal D, Hafizi A, ul-Hassan T, Chaudhary M and Raina SN (2014) Internal transcribed spacer sequences of nuclear ribosomal DNA resolving complex taxonomic history in the genus *Vicia* L. *Genetic Resources and Crop Evolution* 61: 909–925.

Starr JR, Naczi RF and Chouinard BN (2009) Plant DNA barcodes and species resolution in sedges (*Carex*, Cyperaceae). *Molecular Ecology Resources* 9: 151–163.

Steinke D, Zemlak TS, Boutillier JA and Hebert PDN (2009) DNA barcoding of Pacific Canada's fishes. *Marine Biology* 156: 2641–2647.

Tamura K, Stecher G, Peterson D, Filipski A and Kumar S (2013) MEGA6: Molecular Evolutionary Genetics Analysis Version 6.0. *Molecular Biology and Evolution* 30: 2725–2729.

Vencovsky R and Crossa J (1999) Variance effective population size under mixed self and random mating with applications to genetic conservation of species. *Crop Science* 39: 1282–1294.

Vijayan K and Tsou CH (2010) DNA barcoding in plants: taxonomy in a new perspective. *Current Science* 99: 1530–1541.

Zhang AB, He LJ, Crozier RH, Muster C and Zhu CD (2010) Estimating sample sizes for DNA barcoding. *Molecular Phylogenetics and Evolution* 54: 1035–1039.