MOLECULAR ECOLOGY
RESOURCES
Molecular Ecology Resources (2011) doi: 10.1111/j.1755-0998.2011.03009.x

DNA barcodes for Mexican Cactaceae, plants under pressure from wild collecting

CHRIS YESSON,*§ ROLANDO T. BÁRCENAS,† HÉCTOR M. HERNÁNDEZ,‡ MARÍA DE LA LUZ RUIZ-MAQUEDA,† ALBERTO PRADO,†¶ VÍCTOR M. RODRÍGUEZ* and JULIE A. HAWKINS*
School of Biological Sciences, Lyle Tower, The University of Reading, Reading, Berkshire RG6 6AS, UK, †Laboratorio Darwin de Sistemática Molecular y Evolución, Facultad de Ciencias Naturales, Universidad Autónoma de Querétaro, Av. de las Ciencias s/n, Juriquilla, Querétaro CP 76230, México, ‡Departamento de Botánica, Instituto de Biología, Universidad Nacional Autónoma de México, Apartado Postal 70-233, Ciudad Universitaria, 04510 Mexico City, México

Abstract

DNA barcodes could be a useful tool for plant conservation. Of particular importance is the ability to identify unknown plant material, such as from customs seizures of illegally collected specimens. Mexican cacti are an example of a threatened group, under pressure because of wild collection for the xeriscaping trade and private collectors. Mexican cacti also provide a taxonomically and geographically coherent group with which to test DNA barcodes. Here, we sample the *matK* barcode for 528 species of Cactaceae including approximately 75% of Mexican species and test the utility of the *matK* region for species-level identification. We find that the *matK* DNA barcode can be used to identify uniquely 77% of species sampled, and 79–87% of species of particular conservation importance. However, this is far below the desired rate of 95% and there are significant issues for PCR amplification because of the variability of primer sites. Additionally, we test the nuclear *ITS* regions for the cactus subfamily Opuntioideae and for the genus *Ariocarpus* (subfamily Cactoideae). We observed higher rates of variation for *ITS* (86% unique for Opuntioideae sampled) but a much lower PCR success, encountering significant intra-individual polymorphism in *Ariocarpus* precluding the use of this marker in this taxon. We conclude that the *matK* region should provide useful information as a DNA barcode for Cactaceae if the problems with primers can be addressed, but matK alone is not sufficiently variable to achieve species-level identification. Additional complementary regions should be investigated as *ITS* is shown to be unsuitable.

Keywords: conservation, DNA barcodes, internal transcribed spacer, *matK*, Opuntioideae, species identification

Received 16 November 2010; revision received 3 February 2011; accepted 17 February 2011

Introduction

DNA barcodes have tremendous potential to provide benefit to conservation efforts. This includes inventories of biodiversity, biosecurity, discovery of new species and the prevention or detection of illegal trade (Armstrong & Ball 2005; Hajibabaei *et al.* 2006; Lahaye *et al.* 2008; Newmaster *et al.* 2006). DNA barcodes can be used to identify unknown material rapidly. This material might be impossible to identify morphologically without examination of ephemeral characteristics, such as flowers or fruits, by an expert taxonomist. Such a utility could be useful for the identification of customs seizures of illegally traded specimens.

To date, barcoding efforts have met with greatest success for the animal kingdom, whilst progress with plants has been slower, particularly in the selection of a sequence region sufficiently variable to provide species-specific detail, and sufficiently conserved to provide comparison between organisms (Chase *et al.* 2007; Cowan *et al.* 2006; Newmaster *et al.* 2008). Significant effort and resource have been directed towards the selection of plant barcodes, most notably by the Plant Working Group of the Consortium for the Barcoding of Life (CBOL). The recently agreed-upon standard regions for DNA barcoding land plants, the maturase K (*matK*) gene region and the ribulose-bisphosphate carboxylase gene (*rbcL*), were selected to optimize three criteria:

Correspondence: Chris Yesson, Tel.: 020 7449 6267;
E-mail: chris.yesson@ioz.ac.uk
§Present address: Institute of Zoology, Zoological Society of London, Regent's Park, London, NW1 4RY, UK.
¶Present address: Department of Plant Science, McGill University, 21, 111 Lakeshore Road, Ste. Anne de Bellevue, QC H9X 3V9, Canada.

universality, sequence quality and discriminatory power (CBOL Plant Working Group, 2009). Both of the selected regions are widely used for phylogenetic study and so have a significant 'back catalogue' of sequences available on publicly accessible databases, and in previous studies of both floras and taxa, *matK* has performed well as a barcode (Lahaye *et al.* 2008; Starr *et al.* 2009). The *matK* region was shown to be more variable than *rbcL* by the CBOL Plant Working Group (2009). The variability of this region is both an advantage for differentiating closely related species (Shaw *et al.* 2005), and a disadvantage because of the variability of the primer binding sites (Kress & Erickson 2007; CBOL Plant Working Group, 2009). Indeed, the CBOL working group noted that *matK* primer design would likely need optimization in the light of difficulties working with some clades.

The CBOL Plant Working Group (2009) reported that the two-region barcode identified 72% of species tested and proposed that further discrimination could be facilitated by selecting additional markers on a taxon-by-taxon basis. They presented a short list of supplementary noncoding chloroplast loci which might be suitable for individual studies. They also highlight the nuclear internal transcribed spacer region (*ITS*) as an appropriate supplementary region, in cases where its use is not precluded by paralogy, incomplete concerted evolution or pseudogene copies. Markers such as *ITS*, which are very variable and therefore powerful in some groups but present technical problems in others, are likely candidates for taxon-specific second-level barcode markers. The *ITS* region is amongst the most variable regions used for species-level phylogeny reconstruction, suggesting discriminatory power, and it has been shown to be useful in studies by authors including Kress *et al.* (2005) and Edwards *et al.* (2008). Furthermore, conflict between nuclear and chloroplast-based barcode identifications, or the presence of multiple peaks in direct *ITS* sequences, could highlight putative hybrids, minimizing misidentifications of introgressed individuals.

The Cactaceae provide a good test of the utility of barcodes for the identification of plants. Cacti are popular with collectors, often illegally traded, difficult to identify and likely to be a group where investment in molecular identification tools will pay dividends in terms of conservation outcomes (Hardesty *et al.* 2008; Hughes *et al.* 2008). Although the family as a whole show a wide range of morphology, closely related species are morphologically similar and are difficult to identify when not in flower (Anderson 2001; Hunt *et al.* 2006). Mexico is a major centre of cactus diversity, containing approximately one-third of the 1438 species currently recognized (Hunt 1999; Hunt *et al.* 2006). Almost all Cactaceae, excluding those of economic importance, are listed in the Convention on International Trade in Endangered Species of Wild Fauna and Flora (CITES) Appendices I & II (Hunt 1999). Furthermore, there are 123 species listed as vulnerable or more severely threatened on the IUCN Red List, of which more than half (62) are found in Mexico (IUCN 2008). Cactaceae face severe collection pressure, with demand for xeriscaping and from private collectors. In terms of private collectors, pressure can be particularly intense for newly discovered populations, particularly if they are for rare species (Bárcenas-Luna 2003, 2006; Ortega-Baes & Godínez-Alvarez 2006). However, identifications of trafficked plants are often difficult, limiting the utility of seized plants for conservation *ex situ*, or re-introduction. For example, 318 cactus seizures were made of material en route to the USA in 1998. These seizures comprised 1751 plants. In 132 of the 318 cases, initial identifications were made to family level only (42%) and in 137 cases to genus only (43%) (unpublished data made available by U.S. Fish and Wildlife Service, requested under Freedom of Information Act). Many of these plant collections could be of extremely rare species and represent a valuable genetic resource. Better insights into commercial collecting patterns could also be used by ecologists and conservationists to better understand the extent and impact of cactus harvesting in Mexico.

The combination of taxonomic and regional focus employed here provides a realistic test of the *matK* region as a potential DNA barcode. In developing *matK* barcodes for the majority of Mexican cacti, we provide a useful resource, although we highlight the limitations of the marker, both in terms of primer design and in its ability to discriminate species. We also test a potentially complementary barcoding markers, *ITS*, for a subset of the Cacti. The *ITS* marker is tested in a survey of the subfamily Opuntioideae and for all seven species of *Ariocarpus*, a genus of subfamily Cactoideae, tribe Cacteae that is subject to extreme collecting pressure.

Materials and methods

Taxonomic sampling

For the *matK* sequencing, we sampled 655 specimens representing 528 species of Cactaceae including 426 Mexican species, and representing c.75% of all Mexican species (Hunt 1999; Hunt *et al.* 2006). Appendix I contains a complete list of specimens and accession numbers. Appendix I also indicates the subset of specimens which were included in the *ITS* study. Our Opuntioideae data set for *ITS* includes 110 samples, representing 86 of the 220–350 species in the subfamily. We sampled all seven species of *Ariocarpus*, a genus endemic to Mexico.

DNA isolation and amplification

Total genomic DNA was isolated from either fresh material sourced from living collections, field-collected silica-dried cortex peels or herbarium specimens, using Qiagen kits following the manufacturer's instructions (Qiagen, Crawley, West Sussex). For some specimens, this extraction procedure failed to produce adequate DNA; in these cases, a modified CTAB method (Drabkova *et al.* 2002) was used for extraction. The nuclear *ITS* region was amplified using the primers 5F and 4R (Baldwin 1993), using the PCR protocol of Columbus *et al.* (1998). In the case of *Ariocarpus*, the 5F and 4R primers only successfully amplified three species, and so two specific primers were used, *ITS*-AF: GTCGTAACAAGGTTTCCATA and *ITS*-AR: CTTAAACTCAGCGGGCAGCC (V. M. Rodríguez, unpublished). As PCR for *ITS* from all *Ariocarpus* species resulted in multiple bands, bands were excised from agarose gels and purified using a Qiagen gel extraction kit following the manufacturer's protocols prior to sequencing. The *matK* sequences tested here were generated using a suite of primers and methods published elsewhere (Bárcenas *et al.* 2011). For this study, the entire *trnK/matK* region (2916 bp) bounded by the primers 3914F and 2R (Johnson & Soltis 1994) was amplified using the internal primers 23F, 31R, 41R, 44F and 52F (Nyffeler 2002) to reduce each sequence read to a manageable size (<1000 bp). Note, because of amplification failures of some internal primers for some specimens, the reverse complement of internal primers 41R and 52F were used to sequence in the opposite directions; we refer to these as 41F and 52R. The use of these internal primers resulted in some sequences with central sections of the barcode missing because of short or missing fragments.

Sequencing, assembly and alignment

Successful PCR products (as assessed by visualizing with ethidium bromide staining on a 1–1.5% agarose gel) were sent to Macrogen Inc (Seoul, South Korea—http://dna.macrogen.com) for purification and sequencing using the Johnson & Soltis (1994) and Nyffeler (2002) primers. *ITS* products were sequenced using the same primer pairs as used for the initial amplification. Raw sequences were assembled using SeqMan™ II, in the Lasergene® software package (DNASTAR, Inc.). Sequences were aligned automatically with ClustalW 1.83 (Thompson *et al.* 1994) and adjusted by hand. The *trnK/matK* sequences were trimmed at the barcoding primer binding sites 390F (Cuenoud *et al.* 2002) and 3.2R (http://www.kew.org/barcoding/protocols.html). These primers represent the widest implementation of the *matK* barcoding region encompassing other widely used barcoding primers, including the other primers specified on the Kew barcoding website (2.1F, 2.1aF, 5R; http://www.kew.org/barcoding/protocols.html), the reverse primer sequence published by Cuenoud *et al.* (2002) (1326R) and primers developed by Ki-Joong Kim and tested by Fazekas *et al.* (2008) (1R and 3F). Alignments were screened to determine the extent of conservation of these barcoding primer sites.

Species identification

Blast searches have been shown to be one of the quickest and best methods for identifications using DNA barcodes (Little & Stevenson 2007). A blast method does not rely on the subjectivity of sequence alignment and have been used to assess DNA barcodes for plants (Kress & Erickson 2007; Sass *et al.* 2007). All DNA sequences were deposited in a database using the formatdb utility in the BLAST software package (http://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastDocs&DOC_TYPE=Download). Three databases were generated for analysis: the first contained all matK sequences; the second all ITS sequences, and the third contained the combined matK and ITS data for samples with both regions sequenced (see Data S1, Supporting information). For each sequence (or combined sequences), a BLAST search was performed against the relevant database. The highest-scored match always included the original sequence, and sometimes included additional sequences, those matching with the same score to more than one sequence from multiple species were deemed failures of identification (Kress & Erickson 2007; Sass *et al.* 2007). Frequencies of correct assignment of genus as well as correct assignment to species were recorded.

Genetic distances

Pairwise distances were calculated using PAUP* v. 4b10 (Swofford 2002) based on the aligned sequence matrices. For distance calculations, appropriate models of sequence evolution were selected by MODELTEST version 3.7 (Posada & Crandall 1998). Models found were as follows *matK*: GTR+G, *ITS*: GTR+G.

Results

Data for each individual (species name, accession number and identification success/failure) for *ITS* and *matK* sequences are presented in Data S1 (Supporting information).

matK

This study sampled 645 *matK* sequences for 528 species. The aligned matrix, trimmed to the outermost barcoding

primers, was 1207 bp in length, 386 bp were constant and 527 bp were parsimony informative. Of the 528 species represented by *matK* sequences, 408 (77%) were identified uniquely using the BLAST search. Therefore, 23% of species were found to have sequences that matched to at least one sequence from another species, and so failed identification.

We note that a quarter of our *matK* sequences contain a substantial area of missing sequence (>100 bp) in the centre of the barcoding region, because of a failure of one of the internal primers. These regions were coded with 'N'. Data were analysed with and without these sequences, but excluding these sequences made no significant difference to the overall results. Excluding all sequences with one or more ambiguous base produced a similar identification rate 77% (n = 351), suggesting that the problem of missing sequence data did not affect the results. Subsequent numbers will be based on the complete data set (n = 528). Note that none of the primer binding sites were affected by this problem.

The majority of failed identifications were identified correctly to the generic level, and only 34 species (6%) matched to sequences from different genera (note that sequences with missing data were not significantly overrepresented in this data set). A generic level summary is given in Table 1. Generic level match rates vary from 0% (*Maihuenia*, *Pereskiopsis*) up to 100% for 21 genera, but we note that these extremes are found for groups with low species numbers; only two groups have more than 10 species sampled. All *matK* barcoding primers were located within the alignment, and a summary of the variation observed within these primer regions is shown in Table 2. Each primer site demonstrates variation from the published primer for at least four bases. Only four primers (1326R, 3.2R, 71R and 3F_KIM f(R)) show sufficiently frequent identical matches within our samples to show up on Table 4. The primer 3F_KIM f(R) is conserved for 84% of samples and is the most conserved primer of the set. However, variation in the remaining 16% of samples is high, with 20/25 bases showing variation in at least one sample.

ITS

Eighty-seven opuntioid species were sampled for the nuclear *ITS* region. An alignment was produced of length 696 bp (including indels), of which 495 bp were constant and 80 bp were parsimony informative. A generic summary can be seen in Table 3. Generic match rates range from 83% (Opuntia) up to 100% for nine genera; once again the extremes are found in groups with low species sampling (n < 10). In total, seventy-five opuntioid species (86%) were identifiable uniquely, and eight of these could not be identified by *matK* sequence. In contrast, six species that failed identification with *ITS* had unique *matK* sequences. In the case of *Ariocarpus*, PCR using standard primers resulted in the amplification of two bands in all species, a band of approximately 500 bp and another of 650 bp. As the difference in length between the two bands was approximately 150 bp, it was possible to gel purify the two bands. Sequencing of the purified bands showed that the indel event accounting for most of the length difference was an indel of 142 bp (pos. 518–659). Modifications of PCR conditions (DNA concentration and annealing temperature) did not result in single bands being produced, and alternative published ITS primers also resulted in the amplification of the two bands.

Combining regions for the Opuntioideae

Table 4 provides a summary of blast matching identification success for combinations of DNA sequence regions. Performing identifications based on both *matK* and *ITS* regions provides 98% success for species-level identification for the Opuntioideae. Although overall *ITS* has a higher identification success rate than *matK*, on this data *matK* (95%) performs marginally better than ITS (92%).

Discussion

If barcodes are to play an important role in conservation, it should be a priority that they can distinguish species of conservation importance, such as those listed on CITES Appendix I or the IUCN Red List. We have sampled 31 species listed on CITES Appendix I (http://www.cites.org/), of which 27 (87%) can be identified uniquely with *matK* barcodes. The four species that cannot be identified to the species level could be confused with species not listed on Appendix I. Additionally, we have sampled 38 species identified as vulnerable or more severely threatened on the IUCN Red List (IUCN 2008), of which 30 (79%) can be identified uniquely with *matK* barcodes. Six of these eight species failing identification were confused with species not on the Red List. Identification success rates vary between genera, indicating that some taxa are easier to differentiate than others. The extremes of the reported success rates tend to be observed in genera with lower species sampling, which reduces the certainty of these numbers. However, where species-level sampling is complete, these match rates can be regarded as maximum bounds. For example, both accepted *Maihuenia* species are sampled in this study, and their sequences cannot be distinguished, giving a 0% identification success. Increasing sampling to examine further specimens will not alter the fact that some *M. poeppigii* are indistinguishable from some *M. patagonica*. Species-level sampling for *Maihuenia*, *Polaskia* and *Stenocactus* is

Table 1 A generic level summary of sequence variation for the *matK* region. Unique sequences are defined as those matching uniquely to themselves in a blast match against all sequences. Genetic distances are based on infra-generic comparisons. Approximate total species per genus as Hunt *et al*. (2006)

		Blast matches			Infra-specific distances				
Genus	No. of species	Sampled species	Matched species	% match	Minimum	Maximum	Mean	SD	*n*
Ariocarpus	7	7	5	71	0	0.011	0.005	0.003	21
Astrophytum	6	5	5	100	0.006	0.041	0.018	0.014	15
Austrocylindropuntia	8	3	1	33	0	0.012	0.006	0.005	6
Aztekium	2	2	2	100	0.001	0.001	0.001	–	1
Browningia	8	2	2	100	0.001	0.001	0.001	–	1
Cephalocereus	3	3	3	100	0.002	0.007	0.005	0.003	3
Copiapoa	21	3	3	100	0.003	0.006	0.005	0.002	3
Corryocactus	12	2	2	100	0.003	0.003	0.003	–	1
Corynopuntia	14	11	11	100	0	0.035	0.008	0.008	78
Coryphantha	42	32	30	94	0	0.163	0.025	0.025	990
Cylindropuntia	33	24	22	92	0	1.884	0.025	0.167	378
Disocactus	11	3	3	100	0.003	0.006	0.004	0.002	3
Echinocactus	6	5	4	80	0	0.014	0.005	0.004	105
Echinocereus	67	48	34	71	0	0.054	0.007	0.008	1,326
Echinopsis	77	3	2	67	0.001	0.003	0.002	0.001	3
Epithelantha	2	2	2	100	0.002	0.019	0.009	0.007	10
Eriosyce	32	4	2	50	0	0.002	0.001	0.001	6
Escobaria	19	6	6	100	0	0.032	0.018	0.009	15
Ferocactus	28	25	18	72	0	1.884	0.018	0.088	465
Frailea	12	2	2	100	0.002	0.002	0.002	–	1
Hylocereus	14	3	1	33	0	0.012	0.006	0.006	3
Lophophora	3	2	2	100	0.04	0.04	0.04	–	1
Mailhuenia	2	2	0	0	0	0	0	–	1
Mammillaria	163	139	89	64	0	1.884	0.018	0.051	15,931
Myrtillocactus	4	3	2	67	0.003	0.012	0.009	0.005	3
Neobuxbaumia	8	5	5	100	0.004	0.017	0.011	0.005	10
Opuntia	75	40	35	88	0	1.884	0.017	0.103	1,326
Pachycereus	13	8	7	88	0	0.03	0.011	0.006	45
Parodia	58	6	6	100	0.006	0.017	0.011	0.004	15
Pelecyphora	2	2	2	100	0.013	0.013	0.013	–	1
Peniocereus	20	6	6	100	0.003	0.02	0.011	0.005	15
Pereskia	17	13	11	85	0	0.017	0.007	0.004	91
Pereskiopsis	6	2	0	0	0	0.003	0.002	0.002	3
Pfeiffera	9	3	3	100	0.003	0.007	0.005	0.002	3
Pilosocereus	41	5	5	100	0	0.031	0.011	0.012	15
Polaskia	2	2	1	50	0.006	0.006	0.006	–	1
Rhipsalis	35	2	2	100	0.004	0.004	0.004	–	1
Selenicereus	12	3	2	67	0	0.001	0.001	0.001	3
Stenocactus	8	8	3	38	0	0.017	0.004	0.006	66
Stenocereus	24	14	10	71	0	0.028	0.007	0.005	171
Tephrocactus	7	2	2	100	0.018	0.018	0.018	–	1
Thelocactus	14	9	7	78	0	0.012	0.004	0.003	66
Turbinicarpus	16	13	13	100	0.003	0.1	0.027	0.024	153

complete, so we may conclude that these are difficult groups to identify using matK sequences.

This study found that the *matK* barcoding region was successful for species-level identification for 77% of species. Other studies report varying rates of identification success for *matK* barcodes ranging from 49 to 86% (Kress & Erickson 2007; Lahaye *et al*. 2008; Newmaster *et al*. 2008). These studies use different methods of sequence identification and sample a wide range of taxa. Our study selected predominantly Mexican Cactaceae, which will include many sister species pairs, but when the sister species is not found in Mexico we are unlikely to have sampled these. Complete sampling of these sister species may reduce the number of unique identifications.

Table 2 Observed sequence variation for selected *matK* barcoding primers

Primer (position)	Sequence	*n* (%)
390F (0)	**CGATCTATTC–ATTCAATATTT–C**	
	.....A..................	377 (58)
	.....A.............C.....	182 (28)
	G....A.............C.....	45 (7)
	.A...A..................	21 (3)
	17 Others	30 (5)
consensus	SGATCAATTCCCATTCAATHTTTCC	
2.1F (93)	**CCT–ATCCATCTGGAAA–TCTTAG**	
	..C....T..............G.	333 (51)
	..C.................T..G.	113 (17)
	..C....T.....A.........G.	97 (15)
	..C...................G.	50 (8)
	27 Others	62 (9)
consensus	CCCCAATYCATCTVGAAAATNTTGG	
2.1aF (98)	**ATCCATCTGGAAA–TCTTAG–TTC**	
	..T..............G.....	339 (52)
	..............T..G.....	112 (17)
	..T.....A.........G.....	98 (15)
	..................G.....	50 (8)
	23 Others	56 (9)
consensus	ATYCATCTVGAAAATNTTGGKTTC	
1326R (1087)	**ACTTCGACTTTCGTGTGCTAGA**	
	...........T.........	276 (42)
	.....T......T.........	273 (42)
	....	13 (2)
	....AT......T.........	6 (1)
	30 Others	87 (13)
consensus	DCYKMDAYTTYHDWGTVBKRNM	
5R (1091)	**CGACTTTCTTGTGCTAGAAG**	
	C	279 (43)
	.T..............C	273 (42)
	........G.........C	12 (2)
	AT..............C	6 (1)
	30 Others	86 (13)
consensus	MDAYTTYHDWGTVBKRNMAC	
3.2R (1182)	**GAATTCTTTAC———AGAGGAAG**	
	....	362 (55)
	............TAC........	167 (25)
	A...........TAC........	10 (2)
	.........-.............	4 (1)
	28 Others	112 (17)
consensus	NAATTYTYKVHDWYTMCRGRRGRAR	
71R (925)	**G——TATT–AGGACATCCC–ATTA–G**	
	....	504 (77)
	..................G.....	47 (7)
	................T..G.....	23 (4)
	.........C..............	6 (1)
	35 Others	75 (11)
consensus	NATACYATTTNGGRCAYCCBGYVTTRGR	
1R_KIM r(F) (91)	**ACCCA–ATCCATCTGGAAA–TCTTAG–TTC**	
	....C....T..............G....	328 (50)
	....C..............T..G....	111 (17)
	....C....T.....A.........G....	97 (15)
	....C..................G....	49 (7)
	35 Others	70 (11)
consensus	ACCCCCAATYCATCTVGAAAATNTTGGKTTC	

Table 2 Continued

Primer (position)	Sequence	*n* (%)
3F_KIM f(R) (1116)	**CTCGTAAACACAAAAGTACTGTACG**	
	.....	553 (84)
	T.T.....T.A....	3 (0)
	..T.....T.....T....T.	2 (0)
	..T.....T.....T.....	2 (0)
	25 Others	95 (15)
consensus	YBYRHMRRYAHMAMAGYHYKVYVYG	

The bold sequence shows the original primer, below each primer are the four most frequently observed sequences. Dots show conservancy with primer, letters show variation. Indels are identified as '–' within the primer sequence. Primers 390F & 1326R are from Cuenoud *et al.* (2002), Kim 1R and Kim 3F are primers designed by Ki-Joong Kim and tested by Fazekas *et al.* (2008) and all others are from the Kew barcoding protocols website (http://www.kew.org/barcoding/protocols.html). Position of the primer indicates the starting position of the first base in the overall alignment. Consensus sequences summarise the variation across all samples.

Table 3 A generic level summary of sequence variation for the *ITS* region covering 11 genera of Opuntioideae

		Blast matches			Infra-generic genetic distances				
Genus	No. of species	Species Sampled	Matched Species	% match	Minimum	Maximum	Mean	SD	*n*
Austrocylindropuntia	8	1	1	100	–	–	–	–	–
Brasiliopuntia	1	1	1	100	–	–	–	–	–
Corynopuntia	14	9	9	100	0	0.027	0.008	0.007	35
Cylindropuntia	33	28	26	93	0	0.044	0.011	0.008	374
Grusonia	1	1	1	100	–	–	–	–	–
Maihueniopsis	7	1	1	100	–	–	–	–	–
Nopalea	4	2	2	100	–	–	–	–	–
Opuntia	75	59	48	83	0	0.112	0.015	0.013	1,688
Pereskia	17	1	1	100	–	–	–	–	–
Pereskiopsis	6	3	3	100	0.007	0.018	0.012	0.006	3
Tephrocactus	7	4	4	100	0.027	0.049	0.036	–	3

Table 4 Summary of blast matching success for each region and combinations of regions

Region(s)	*n*	ID	%
matK	528	408	77
ITS	87	75	86
Combined dataset			
matK	62	59	95
ITS	62	57	92
matK+ITS	62	61	98

n, number of species; ID, species successfully identified; %, proportion of species identified.
Data for *ITS* and combined regions refers to the Opuntioideae data sets. The combined dataset refers to specimens with both ITS and matK sequences.

Furthermore, our study selects individual specimens and treats these as representative of the full range of variation. Many species may show variation when examined in more detail; for example, the 10 specimens of *Echinocactus grusonii* sequenced for *matK* produced eight unique sequences (maximum genetic distance = 0.005, mean = 0.002, standard deviation = 0.001). It seems likely that sampling more individuals may close the 'barcoding gap' for some species and so reduce the number of species identifiable uniquely by the *matK* barcode. Our reported identification success rate is lower than the 80% objective proposed by some authors (Rubinoff *et al.* 2006), and much lower than the hoped for figure of 95% (Hebert & Gregory 2005); though, it compares favourably with the two-barcode identification rate of 72% reported by the CBOL Plant Working Group (2009).

A significant problem with the *matK* region is the variation displayed within the primers. Many of the primers favoured for by *matK* barcoding region appear to be unreliable for amplification, which is in agreement with previous findings (Kress & Erickson 2007). Fazekas *et al.* (2008) found it necessary to employ 10 different *matK* primer combinations in their study of 32 land

plant genera, including eight combinations for angiosperms. However, (Lahaye *et al.* 2008) report 100% amplification success using the 390F and 1326R primers on a wide selection of plant families from South Africa and Mesoamerica. Neither of these studies sampled Cactaceae, and it is possible that the levels of variation presented in this study could be peculiar to the Cactaceae. We found that the majority of sequences show small numbers of substitutions (commonly 1–3 bases per sample) from the published primers, but this low level of variation may not prevent primer binding. Where primers differ sufficiently to prevent binding, it should be possible to design at least genera specific primers. For example, although all *Mammillaria* sequences differ from the tested forward primers, they are constant at these sites, differing from the published primers by 3–4 fixed bases. Given the wide selection of primers available for the *matK* region, and the potential to modify them to optimize performance, we feel that some combination can be found to successfully study *matK* barcodes for most Cactaceae. However, the selection of primers can alter both the start and end of the *matK* barcoding region by more than 100 bp, which could lead to issues of comparability of barcodes if different researchers use different primers.

Although the nuclear *ITS* region showed higher levels of variation than *matK*, it is noted that where both *matK* and *ITS* data were available identification success was marginally better for *matK*. Additionally, PCR amplification of ITS sequences proved difficult for many opuntioid samples. Many samples that were straightforward to amplify for matK failed amplification for ITS, even after trial adjustments of annealing temperature and the addition of 0–1 μl bovine serum albumen. In the case of *Ariocarpus*, multiple bands were recovered by PCR. The failure to produce a single amplification products rules out direct sequencing for barcoding purposes, and so *ITS* is not a suitable barcode region for *Ariocarpus*, despite high levels of variation. There is some evidence that problems with multiple copies are not likely to be limited to the genus *Ariocarpus* or tribe Cacteae. Harpke & Peterson (2008) showed that in *Mammillaria*, another genus of tribe Cacteae have up to five different 5.8S rDNA types in one individual. In their study, 28 of 30 cloned *Mammillaria* 5.8S rDNA sequences were putative pseudogenes. In another study sampling two species in subfamily Cactoideae, tribe Echinocereeae, *Lophocereus schottii* (syn. *Pachycereus schottii sensu* Hunt *et al.* 2006) and *Pachycereus marginatus*, PCR using *ITS* primers resulted in two bands in all studied individuals (Hartmann *et al.* 2001). Alignment of *Ariocarpus* sequences to the *Pachycereus* sequences revealed sequence similarities, suggesting that the origin of the paralogous copies predates the divergence of the tribes (V. M. Rodríguez, unpublished). Thus, although in some plant groups the presence of multiple *ITS* copies may be a useful indicator of recent hybridization of introgression (e.g. Chase *et al.* 2003), this is not the case for the Cacteae.

One of the primary objectives of DNA barcoding is to provide tools for species-level identifications. Our findings suggest that the *matK* region, by itself, is unsuitable for use as the sole plant DNA barcode for Cactaceae, because it is insufficiently variable to be used as a standalone tool for identification, although we note that identification success is higher for species of conservation importance. Technical issues relating to primer selection and use may prove an obstacle for *matK* barcoding studies, and future studies to develop primers with higher universality should include Cactaceae. However, if primer issues were resolved, *matK* could be a useful constituent of a multiple region barcode. Combining *ITS* and *matK* produced identification above the 95% threshold for the Opuntioideae, but *ITS* cannot be a universal second barcode for the family because of paralogy issues, at least for the subfamily Cactoideae. Additional barcoding regions should be sought to complement *matK* if we are to achieve the goal of reliable species-level identification for Cactaceae.

Acknowledgements

This work was funded by the Darwin Initiative, DEFRA, UK, as part of the project 'Certification to Support Conservation of Endangered Mexican Cacti' (http://www.uaq.mx/ccma/; project code 14-059). We thank the personnel and curators of the following herbaria, MEXU, QMEX, ZSS, RSA, and also Carlos Gómez-Hinostrosa from the Instituto de Biología, UNAM and the staff of El Charco del Ingenio Botanic Gardens.

References

Anderson EF (2001) *The Cactus Family*. Timber Press Inc., Portland, Oregon.

Armstrong KF, Ball SL (2005) DNA barcodes for biosecurity: invasive species identification. *Philosophical Transactions of the Royal Society of London. Series B-Biological Sciences*, **360**, 1813–1823.

Baldwin BG (1993) Molecular Phylogenetics of Calycadenia (Compositae) Based on *ITS* Sequences of Nuclear Ribosomal DNA – Chromosomal and Morphological Evolution Reexamined. *American Journal of Botany*, **80**, 222–238.

Bárcenas-Luna RT (2003) Chihuahuan Desert Cacti in Mexico: an assessment of trade, management, and conservation priorities. In: *Prickly trade: trade and conservation of Chihuahuan Desert cacti* (ed. Robbins CS), pp. 1–65. Traffic, Washington, D.C.

Bárcenas-Luna RT (2006) Comercio de cactáceas mexicanas y perspectivas para su conservación. *Biodiversitas – Boletín Bimestral de la Comisión Nacional para el Conocimiento y Uso de la Biodiversidad*, **68**, 11–15.

Bárcenas RT, Yesson C, Hawkins JA (2011) Molecular systematics of the Cactaceae. *Cladistics*, doi: 10.1111/j.1096-0031.2011.00350.x.

CBOL Plant Working Group (2009) A DNA barcode for land plants. *Proceedings of the National Academy of Sciences of the United States of America*, **106**, 12794–12797.

Chase MW, Knapp S, Cox AV *et al.* (2003) Molecular systematics, GISH and the origin of hybrid taxa in *Nicotiana* (Solanaceae). *Annals of Botany*, **92**, 107–127.

Chase MW, Cowan RS, Hollingsworth PM *et al.* (2007) A proposal for a standardised protocol to barcode all land plants. *Taxon*, **56**, 295–299.

Columbus JT, Kinney MS, Pant R, Siqueiros-Delgado ME (1998) Cladistic parsimony analysis of internal transcribed spacer region (nrDNA) sequences of *Bouteloua* and relatives (Gramineae: Chloridoideae). *Aliso*, **17**, 99–130.

Cowan RS, Chase MW, Kress WJ, Savolainen V (2006) 300,000 species to identify: problems, progress, and prospects in DNA barcoding of land plants. *Taxon*, **55**, 611–616.

Cuenoud P, Savolainen V, Chatrou LW *et al.* (2002) Molecular phylogenetics of Caryophyllales based on nuclear 18S rDNA and plastid rbcL, atpB, and matK DNA sequences. *American Journal of Botany*, **89**, 132–144.

Drabkova L, Kirschner J, Vlcek C (2002) Comparison of seven DNA extraction and amplification protocols in historical herbarium specimens of Juncaceae. *Plant Molecular Biology Reporter*, **20**, 161–175.

Edwards D, Horn A, Taylor D, Savolain V, Hawkins J (2008) DNA barcoding of a large genus, *Aspalathus* L. (Fabaceae). *Taxon*, **57**, 1317–1327.

Fazekas AJ, Burgess KS, Kesanakurti PR *et al.* (2008) Multiple multilocus DNA barcodes from the plastid genome discriminate plant species equally well. *PLoS ONE*, **3**, e2802.

Hajibabaei M, Smith MA, Janzen DH *et al.* (2006) A minimalist barcode can identify a specimen whose DNA is degraded. *Molecular Ecology Notes*, **6**, 959–964.

Hardesty BD, Hughes SL, Rodríguez VM, Hawkins JA (2008) Characterization of microsatellite loci for the endangered cactus *Echinocactus grusonii*, and their cross-species utilization. *Molecular Ecology Resources*, **8**, 164–167.

Harpke D, Peterson A (2008) Extensive 5.8S nrDNA polymorphism in *Mammillaria* (Cactaceae) with special reference to the identification of pseudogenic internal transcribed spacer regions. *Journal of Plant Research*, **121**, 261–270.

Hartmann S, Nason J, Bhattacharya D (2001) Extensive ribosomal DNA genic variation in the columnar cactus *Lophocereus*. *Journal of Molecular Evolution*, **53**, 124–134.

Hebert PDN, Gregory TR (2005) The promise of DNA barcoding for taxonomy. *Systematic Biology*, **54**, 852–859.

Hughes SL, Rodríguez VM, Hardesty BD *et al.* (2008) Characterisation of microsatellite loci for the critically endangered cactus *Ariocarpus bravoanus*. *Molecular Ecology Resources*, **8**, 1068–1070.

Hunt D (1999) *CITES Cactaceae Checklist*, 2nd edn. Kew, London.

Hunt D, Taylor N, Charles G (2006) *The New Cactus Lexicon, Volumes I and II*. DH Books, Milborne Port.

IUCN (2008) IUCN Red List of Threatened Species. Available from http://www.iucnredlist.org (Downloaded on 12th August 2008).

Johnson LA, Soltis DE (1994) Matk DNA-Sequences and Phylogenetic Reconstruction in Saxifragaceae S-Str. *Systematic Botany*, **19**, 143–156.

Kress WJ, 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, Janzen DH (2005) Use of DNA barcodes to identify flowering plants. *Proceedings of the National Academy of Sciences of the United States of America*, **102**, 8369–8374.

Lahaye R, Van der Bank M, Bogarin D *et al.* (2008) DNA barcoding the floras of biodiversity hotspots. *Proceedings of the National Academy of Sciences of the United States of America*, **105**, 2923–2928.

Little DP, Stevenson DW (2007) A comparison of algorithms for identification of specimens using DNA barcodes: examples from gymnosperms. *Cladistics*, **23**, 1–21.

Newmaster SG, Fazekas AJ, Ragupathy S (2006) DNA barcoding in land plants: evaluation of rbcL in a multigene tiered approach. *Canadian Journal of Botany-Revue Canadienne De Botanique*, **84**, 335–341.

Newmaster SG, Fazekas AJ, Steeves RAD, Janovec J (2008) Testing candidate plant barcode regions in the Myristicaceae. *Molecular Ecology Resources*, **8**, 480–490.

Nyffeler R (2002) Phylogenetic relationships in the cactus family (Cactaceae) based on evidence from trnK/matK and trnL-trnF sequences. *American Journal of Botany*, **89**, 312–326.

Ortega-Baes P, Godínez-Alvarez H (2006) Global Diversity and Conservation Priorities in the Cactaceae. *Biodiversity and Conservation*, **15**, 817–827.

Posada D, Crandall K (1998) MODELTEST: testing the model of DNA substitution. *Bioinformatics*, **14**, 817–818.

Rubinoff D, Cameron S, Will K (2006) Are plant DNA barcodes a search for the Holy Grail? *Trends in Ecology & Evolution*, **21**, 1–2.

Sass C, Little DP, Stevenson DW, Specht CD (2007) DNA barcoding in the cycadales: testing the potential of proposed barcoding markers for species identification of cycads. *PLoS ONE*, **2**, e1154.

Shaw J, Lickey EB, Beck JT *et al.* (2005) The tortoise and the hare II: relative utility of 21 noncoding chloroplast DNA sequences for phylogenetic analysis. *American Journal of Botany*, **92**, 142–166.

Starr J, Naczi R, Chouinard B (2009) Plant DNA barcodes and species resolution in sedges (*Carex*, Cyperaceae). *Molecular Ecology Resources*, **9**, 151–163.

Swofford DL (2002) *PAUP*: Phylogenetic Analysis Using Parsimony (* and other methods), Version 4.0 b10*. Sinauar, Sunderland, Massachusetts.

Thompson JD, Higgins DG, Gibson TJ (1994) Clustal-W – Improving the Sensitivity of Progressive Multiple Sequence Alignment through Sequence Weighting, Position-Specific Gap Penalties and Weight Matrix Choice. *Nucleic Acids Research*, **22**, 4673–4680.

Supporting Information

Additional supporting information may be found in the online version of this article.

Data S1 A list of species and EMBL accession numbers used in this study.

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting information supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.