Mitochondrial DNA, December 2010; 21(S1): 51–59

informa
healthcare

Character-based, population-level DNA barcoding in Mexican species of *Zamia* L. (Zamiaceae: Cycadales)

FERNANDO NICOLALDE-MOREJÓN[1,2], FRANCISCO VERGARA-SILVA[3], JORGE GONZÁLEZ-ASTORGA[1], & DENNIS W. STEVENSON[4]

[1]*Laboratorio de Genética de Poblaciones, Red de Biología Evolutiva, Instituto de Ecología, A.C., Xalapa, Veracruz, Mexico,* [2]*Instituto de Investigaciones Biológicas, Universidad Veracruzana, Xalapa, Veracruz, Mexico,* [3]*Laboratorio de Sistemática Molecular (Jardín Botánico), Instituto de Biología, Universidad Nacional Autónoma de México, México D.F., Mexico, and* [4]*The New York Botanical Garden, Bronx, NY, USA*

(*Received 5 July 2010; revised 10 August 2010; accepted 7 November 2010*)

Abstract

Background and aims: With the recent proposal of *matK* and *rbcL* as core plant DNA barcoding regions by the Consortium for the Barcoding of Life Plant Working Group, the construction of reference libraries in the botanical DNA barcoding initiative has entered a new phase. However, in a recent DNA barcoding study in the three Mexican genera of the gymnosperm order Cycadales, we found that neither *matK* nor *rbcL* allow high levels of molecular identification of previously established species.
Materials and methods: Our data analysis in that study rested on the "Characteristic Attributes Organization System" (CAOS), a character-based algorithm for the definition of "DNA diagnostics." Here, we use CAOS to analyze a population-level molecular data set in *Zamia*, one of the three cycad genera occurring in Mexico, whose populations display contrasting biogeographic patterns. Our population-level study, which includes all species in the region formally known as Megamexico, is restricted to the genome region, which showed the best single-locus molecular identification performance in our previous study—namely, the noncoding intergenic chloroplast spacer *psbK-I*.
Results: Our comparison of single-individual vs. population-level *psbK-I* datasets in *Zamia* indicates that CAOS analyses are sensitive to slight alignment changes, which in turn derive from the different amounts of molecular variation present in each matrix type.
Conclusion: We, therefore, suggest that character-based studies that involve population-level data should contemplate this type of comparison between data matrices, before a set of DNA diagnostics in a given DNA barcoding reference library is considered definitive.

Keywords: *DNA barcoding, character-based methods,* psbK-I, Zamia, *population level, Megamexico*

Introduction

In 2009, the Consortium for the Barcoding of Life Plant Working Group (CBOL PWG) selected the chloroplast regions *matK* and *rbcL* as a "core barcode" for the land plants. Both gene-coding loci had been widely used already in plant molecular systematics studies, some of which helped to establish the research field as such (e.g. Chase et al. 1993; Hilu et al. 2003). Therefore, these regions were included in several plant DNA barcoding projects (Chase et al. 2005, 2007; Kress et al. 2005; Cowan et al. 2006; Kress and Erickson 2007; Erickson et al. 2008; Fazekas et al. 2008, 2009; Lahaye et al. 2008a,b; Ford et al. 2009), whose results ultimately led to the CBOL PWG consensus.

In a recent DNA barcoding (i.e. molecular identification) study in the three genera of cycads that occur in Mexico—*Ceratozamia* Brongn., *Dioon* Lindl. and *Zamia* L.—we tested the comparative performance of seven chloroplast coding regions and the nuclear internal transcribed spacer (ITS) (Nicolalde-Morejón et al. 2010). In contrast to the results presented in the CBOL PWG paper, we showed that both *matK* and

Correspondence: F. Nicolalde-Morejón, Instituto de Investigaciones Biológicas, Universidad Veracruzana, Avenida Luis Castelazo Ayala s/n, Colonia Industrial Ánimas, Xalapa 91190, Veracruz, Mexico. Tel: (52) 228 8418911. Fax: (52) 228 8418911. E-mail: f_nicolalde@yahoo.com

ISSN 1940-1736 print/ISSN 1940-1744 online © 2010 Informa UK, Ltd.
DOI: 10.3109/19401736.2010.539215

rbcL failed to comply, in any of these cycad genera, with the criterion of "discrimination," i.e. the third parameter according to which a candidate genome segment should be accepted as a consensus DNA barcoding region (CBOL PWG 2009, p. 12794). Our findings (Nicolalde-Morejón et al. 2010), based on single-individual sampling in every species in three Mexican cycad genera, echoed previous indications of the absence of useful variation for DNA barcoding purposes in *matK* and *rbcL*, in practically all cycad genera (Little and Stevenson 2007; Sass et al. 2007). Nicolalde-Morejón et al. (2010) analyzed the data with the "Character Attributes Organization System" (CAOS), a method that defines "DNA diagnostics," i.e. (molecular) character states shared by members of a given taxon and simultaneously absent from comparable groups. For a concise summary of the theoretical reasons behind our decision to use CAOS in a DNA barcoding context, see Nicolalde-Morejón et al. (2010, p. 11). For more extensive expositions of the rationale behind character-based methods in DNA barcoding and additional discussion concerning their advantages over phenetic (and other noncharacter-based) approaches, see DeSalle et al. (2005), DeSalle (2006, 2007), Rach et al. (2008), Sarkar et al. (2008), Bergmann et al. (2009), and Lowenstein et al. (2009).

Justification for the use of psbK-I *as the DNA barcoding region in the present study*

In our CAOS-based analysis of the Mexican cycad genera, we found that the chloroplast intergenic spacer *psbK-I* displayed the best overall single-region performance of all the tested loci (Nicolalde-Morejón et al. 2010, p. 9; see Table I). The measurement behind this qualitative estimation is the percentage of unique, correct species identification under the CAOS analytical regime. Whereas in *Dioon*, this percentage was 57 (i.e. eight out of 14 species), in the case of *Zamia*, one-half of the total number of species (i.e. 12 out of 24) was successfully identified. Only in *Ceratozamia* did the *psbK-I* region have low levels of discrimination between species (four out of 23, or 17%). However, this value is not very different from the best (*ITS2*) and the second best (*atpF-atpH*) performing regions in the same genus, indicating that *Ceratozamia* is a particularly difficult taxon to address from a DNA barcoding perspective.

Nicolalde-Morejón et al. (2010, p. 12) had already noted that the group of *Zamia* species located in the biogeographic zone known as "Megamexico"—a region that roughly covers Mexico, Guatemala, Belize, El Salvador, and part of Northern Nicaragua (Rzedowski 1991)—involves a set of taxa for which there is "an increasingly better understanding of taxonomy and systematics." This assertion rests upon

Table I. DNA barcoding in species of *Zamia* occurring in Megamexico: comparative assessment of the number of diagnostic sites for four chloroplast coding and non-coding genome regions and one nuclear region.

		Diagnostic sites				
Zamia species	Country of distribution	*psbK-I*	*atpF-H*	*ITS2*	*rpoC1*	*matK*
Zamia cremnophila Vovides, Schutzman & Dehgan⋆	Mexico	>1	–	–	–	5
Z. fischeri Miq.⋆	Mexico	–	2	–	–	–
Z. furfuracea L. f.⋆	Mexico	–	–	4	–	–
Z. herrerae Calderón & Standl.⋆	Mexico	–	1	–	–	–
Z. inermis Vovides, Rees & Vázq. Torres⋆	Mexico	1	–	–	–	2
Z. katzeriana (Regel) Rettig	Mexico	–	–	–	–	–
Z. lacandona Schutzman & Vovides⋆	Mexico	–	1	–	–	–
Z. loddigesii Miq.⋆	Mexico	2	–	–	1	–
Z. paucijuga Wieland	Mexico	–	–	–	–	–
Z. polymorpha Stevenson, Moretti & Vázq. Torres⋆	Mexico	3	–	–	–	–
Z. prasina W. Bull⋆	Belize	1	1	–	–	–
Z. purpurea Vovides, Rees & Vázq. Torres⋆	Mexico	–	–	–	1	–
Z. soconuscensis Schutzman, Vovides & Dehgan⋆	Mexico	1	1	–	–	–
Z. spartea A. DC.	Mexico	–	–	–	–	–
Z. standleyi Schutzman⋆	Honduras	2	–	3	–	–
Z. tuerckheimii Donn. Sm.⋆	Guatemala	–	–	1	–	–
Z. variegata Warsz.⋆	Mexico	–	–	1	–	–
Z. vazquezii Stevenson, Sabato, Moretti & De Luca	Mexico	–	–	–	–	–
Z. cunaria Dressler & Stevenson⋆	Panama	–	4	–	1	–
Z. elegantissima Schutzman, Vovides & Adams⋆	Panama	–	–	–	1	–
Z. integrifolia L. f.⋆	USA	1	–	3	–	–
Z. manicata Linden ex Regel⋆	Colombia	4	1	1	1	–
Z. pseudoparasitica Yates in Seem⋆	Panama	>50	–	–	–	–
Z. pygmea Sims	Cuba	–	–	–	–	–

The intergenic spacer *psbK-psbI* displays the best single-locus performance. Information compiled from Nicolalde-Morejón et al. (2010).
⋆The *Zamia* species diagnosable with these loci.

the contents of a recently published monograph (Nicolalde-Morejón et al. 2009), which includes detailed information about the sites of geographic occurrence of populations in every Megamexican *Zamia* species. In the present paper, we analyzed molecular data obtained from samples collected in the field, directly by our research group, for several of these *Zamia* populations. On the basis of the aforementioned best performance of *psbK-I* as a DNA barcoding region in the Mexican species of *Zamia* (Table II; see also Figure 1 and Table 2 in Nicolalde-Morejón et al. (2010)), our analyses are focused only on this region. Although reliance on single-gene matrices might seem restrictive at first sight, the use of the same character-based DNA barcoding strategy in our *Zamia* population-level data set allows us to establish comparisons with character-based results from a single-individual data set, for the same genomic region (i.e. the *Zamia psbK-I* matrix already analyzed in Nicolalde-Morejón et al. (2010)). This type of comparison is useful to establish whether character-based DNA barcoding analyses are sensitive to slight alignment differences caused by the introduction of new molecular variation through the addition of samples/sequence replicas in matrices. We consider that our results might be of general interest for DNA barcoding researchers interested in character-based methods, whose projects might involve the retrieval of molecular variation from diverse populations within species.

Materials and methods

Sampling of biological materials

We have sampled at least one population for each of the 21 *Zamia* species recognized from Megamexico, which represents the entire diversity of species for this genus in Megamexico (*sensu* Rzedowski 1991; see Nicolalde-Morejón et al. 2009, 2010). In total, 63 *Zamia* populations are represented in our study (for a quantitative description of the distribution of these populations per *Zamia* species and the total number of samples processed per population, see Table II). All materials were obtained either from living plants at the National Cycad Collection in the Jardín Botánico "Francisco Javier Clavijero" (JBC) that is administered by the Instituto de Ecología, A.C. (Xalapa, Veracruz, Mexico) or collected in the field (for a list of the collection sites for every *Zamia* population included in the present study, see Table II). Leaf tissues from *Zamia standleyi*, *Zamia tuerckheimii*, and *Zamia prasina* were obtained as a gift from the Montgomery Botanical Center (MBC, Miami, FL, USA).

Leaf genomic DNA extraction, PCR amplification, and DNA sequencing

Apart from the leaf samples brought into the laboratory from the field, freshly collected materials were used in total leaf genomic DNA extractions of materials from greenhouses at the JBC. For the extractions, we used either the DNeasy Plant Mini Kit (Qiagen, Valencia, California, USA) or a modified protocol based on a widely employed CTAB DNA extraction procedure (Doyle and Doyle 1987). PCR amplification experiments were carried out as reported in one of the pioneer cycad DNA barcoding studies (Sass et al. 2007); primers specific for the selected region in the present study—i.e. the intergenic spacer *psbK-I*—are the same that were used by Nicolalde-Morejón et al. (2010, p. 5; these primers are in turn based on Lahaye et al. (2008a,b)). Amplification products were observed and photographed after gel electrophoresis in 1% agarose gels stained with ethidium bromide. PCR products were purified directly, using the QIAquick PCR Purification Kit (Qiagen), and automated sequencing was carried out at Macrogen (Seoul, South Korea).

Sequence assembly and alignment

Electropherogram editing for assembly of *psbK-I* fragments into contigs was performed with the software program Sequencher 4.8 (Gene Codes Corp, Ann Arbor, Michigan, USA). Sequences were deposited on GenBank (see Table II for accession numbers). Sequence alignment of assembled contigs was carried out in BioEdit 7.0.9 (Hall 1999), using the ClustalX (Thompson et al. 1997) multiple alignment mode function. Aligned matrices were imported into Mesquite 2.73 (Maddison and Maddison 2010) and edited by hand after further visual inspection. Files were saved in Nexus format for subsequent analysis, and are available from the corresponding author upon request.

Character-based analysis of a population-level matrix of psbK-I *sequences from* Zamia *species distributed in Megamexico*

As stated in the Introduction, in the present work, we have used the software program CAOS (Sarkar et al. 2008) in order to define "DNA diagnostic characters" in our matrices and provided a basis for the molecular identification of samples (i.e. individuals) that potentially belong to the set of *Zamia* species described in the taxonomic revision by Nicolalde-Morejón et al. (2009). On the basis of the instruction manual for CAOS ("CAOS Documentation and Worked Examples"; Sarkar et al. 2008) and the methodological protocol described in Nicolalde-Morejón et al. (2010, pp. 4–5), we first built a neighbor-joining phenogram (Saitou and Nei 1987) in PAUP 4b10 (Swofford 2002), in order to have a starting branching diagram, which was then converted into the definitive "guide tree" using Mesquite 2.73 (Maddison and Maddison 2010). Evidently, the topology of the neighbor-joining tree had no effect on our subsequent analyses; at the same time, in this context, we emphasize that DNA barcoding studies are neither phenetic nor phylogenetic inference

Table II. Sequence data and geographical origin characteristics of sampled *Zamia* species.

Zamia species	Number of populations	Distribution (state, country)	Number of sequences per collection locality and haplotypes	
			psbk-psbI	Haplotype and GenBank accession numbers
Z. paucijuga	13	1. Nayarit, Mexico	9	H1: HQ454120
		2. Jalisco, Mexico	9	H1
		3. Jalisco, Mexico	10	H1
		4. Jalisco, Mexico	9	H1
		5. Jalisco, Mexico	10	H1
		6. Michoacán, Mexico	10	H1
		7. Guerrero, Mexico	5	H1
		8. Guerrero, Mexico	10	H1
		9. Guerrero, Mexico	9	H1
		10. Guerrero, Mexico	8	H11: HQ454130
		11. Oaxaca, Mexico	8	H11
		12. Oaxaca, Mexico	10	H11
		13. Oaxaca, Mexico	10	H1
Z. soconuscensis	1	1. Chiapas, Mexico	10	H1
Z. herrerae	2	1. Chiapas, Mexico	9	H4: HQ454123
		2. Chiapas, Mexico	7	H4
Z. loddigesii	7	1. Tabasco, Mexico	10	H1 (Ind. 1–3, 5–8, 10); H4 (Ind. 4, 9)
		2. Tamaulipas, Mexico	10	H4
		3. Veracruz, Mexico	4	H1 (Ind. 1, 3, 4); H4 (Ind. 2)
		4. Veracruz, Mexico	3	H1
		5. Veracruz, Mexico	3	H1
		6. Veracuz, Mexico	2	H1
		7. Oaxaca, Mexico	4	H4 (Ind. 3); H9: HQ454128 (Ind. 1, 2); H10: HQ454129 (Ind. 4)
Z. fischeri	2	1. San Luis Potosí, Mexico	10	H2: HQ454121
		2. San Luis Potosí, Mexico	8	H2 (Ind. 1, 4, 5, 6, 7); H3: HQ454122 (Ind. 2, 3, 8)
Z. vazquezii	1	1. Veracruz, Mexico	9	H8: HQ454127
Z. inermis	1	1. Veracruz, Mexico	8	H2 (Ind. 1, 3–8); H5: HQ454124 (Ind. 2)
Z. furfuracea	6	1. Veracruz, Mexico	10	H1
		2. Veracruz, Mexico	10	H1
		3. Veracruz, Mexico	8	H1
		4. Veracruz, Mexico	7	H1 (Ind. 1, 2, 4–7); H4 (Ind. 3)
		5. Veracruz, Mexico	9	H1
		6. Veracruz, Mexico	10	H1
Z. katzeriana	4	1. Chiapas, Mexico	7	H1
		2. Chiapas, Mexico	7	H4
		3. Chiapas, Mexico	2	H1
		4. Veracruz, Mexico	4	H1 (Ind. 1, 3); H4 (Ind. 2, 4)
Z. spartea	3	1. Oaxaca, Mexico	9	H1 (Ind. 1–3, 5–8); H4 (Ind. 4, 9)
		2. Oaxaca, Mexico	9	H1
		3. Oaxaca, Mexico	3	H1 (Ind. 3); H4 (Ind. 1, 2)
Z. purpurea	2	1. Veracruz, Mexico	8	H1
		2. Oaxaca, Mexico	10	H1
Z. cremnophila	1	1. Tabasco, Mexico	10	H1
Z. lacandona	3	1. Chiapas, Mexico	5	H1
		2. Chiapas, Mexico	1	H1
Z. polymorpha	11	1. Yucatán, Mexico	3	H12 (Ind. 1); H13 (Ind. 2, 3)
		2. Yucatán, Mexico	4	H12: HQ454131
		3. Yucatán, Mexico	4	H12
		4. Campeche, Mexico	3	H12
		5. Campeche, Mexico	4	H13: HQ454132
		6. Quintana Roo, Mexico	2	H12
		7. Quintana Roo, Mexico	2	H1 (Ind. 1); H12 (Ind. 2)
		8. Quintana Roo, Mexico	2	H12
		9. Quintana Roo, Mexico	2	H12
		10. Chiapas, Mexico	3	H12
		11. Tabasco, Mexico	4	H12
Z. variegata	1	1. Chiapas, Mexico	9	H1
Z. standleyi	1	1. Honduras	4	H1
Z. tuerckheimii	1	1. Guatemala	5	H1
Z. prasina	1	1. Belize	5	H14: HQ454133

Table II – *continued*

Zamia species	Number of populations	Distribution (state, country)	Number of sequences per collection locality and haplotypes	
			psbk-psbI	Haplotype and GenBank accession numbers
Z. cunaria	1	Panama	1	H1
Z. manicata	1	Colombia	1	H15: HQ454134
Z. pygmea	1	Cuba	1	H6: HQ454125
Z. integrifolia	1	USA	1	H7: HQ454126

studies *per se*, but rather an identification method. Given that our data sets involve population-level molecular information, we followed the CAOS instruction manual and collapsed all nodes basal to the groups of sequences corresponding to each population; subsequently, these "clades" were nested within higher-order groups, corresponding in each case to a single unique species name. The Mesquite-generated guide tree resulting from these manipulations (see Supplementary material) was then stored in Nexus format for subsequent analysis with the programs specific to the CAOS software package.

We operated the CAOS programs following the protocol described by Nicolalde-Morejón et al. (2010, pp. 4–5). As stated in that protocol, determination of DNA diagnostics required the manual revision of the "CAOS-attribute file" and "CAOS-group file" archives generated by the program P-Gnome (Sarkar et al. 2008). After this step, characters ("attributes") with confidence value of 1.00 were selected to construct the actual matrix of DNA diagnostics. Corroboration of attributes was achieved by visually comparing the information of the "CAOS-group file" archives with the original, Mesquite-edited matrices. The matrix of DNA diagnostics constructed after CAOS analyses of the population-level matrix of *psbK-I* sequences for *Zamia* species from Megamexico was then compared with the single-individual, DNA diagnostics matrix for the same chloroplast region, which is a subset of the multigene global matrix analyzed by Nicolalde-Morejón et al. (2010).

Results and discussion

Why are mitochondrial genome regions not suitable for plant DNA barcoding?

The availability of primer pairs for many different plant genome regions determined that the test for candidate land plant DNA barcodes was initially open to the inclusion of noncoding segments, as well as other coding sequences besides *matK* and *rbcL*. However, the interest that molecular biology-oriented botanists have had on the chloroplast genome as a source of DNA barcoding candidate regions did not derive only from the aforementioned record of success achieved with *matK* and *rbcL* in plant molecular systematics. An equally important reason behind that research focus has been the observation (made independently by plant molecular biologists not formally involved in plant DNA barcoding) that the cytochrome *c* oxidase subunit I gene—i.e. the mitochondrial region that had been already established as the "universal DNA barcode" in animals (Hebert et al. 2003a,b)—is not suitable for DNA-based identification purposes due to specificities in the molecular evolution of plant mitochondrial genomes (Chase et al. 2005). We believe that this context is of interest to readers of this journal, for obvious reasons. Nevertheless, we also think that the analysis of molecular variation in the mitochondrial genome of cycads might prove useful to understand plant genome evolution dynamics in general, and perhaps also have a limited utility for molecular identification purposes. The extent of such molecular variation in the cycads is, at any rate, practically unknown to date (for a rare example in which mitochondrial DNA variation was detected among populations of a cycad species, see Huang et al. 2001).

How useful is psbK-I *for DNA barcoding in plants?*

In the context of the international initiative to find plant DNA barcoding regions, the chloroplast genome intergenic spacer *trnH-psbA* was proposed early as a potentially ideal plant DNA barcode on the basis of its relatively small size and ease of amplification (Kress et al. 2005; for the proposal of *trnH-psbA* and *rbcL* as a potential "two-locus universal DNA barcode" in plants, see Kress and Erickson 2008). Later on, a couple of additional chloroplast spacers—namely, *atpF-atpH* and *psbK-I*—were proposed (by Korean botanist K.-J. Kim; see Pennisi 2007) and their utility discussed in international meetings. Ultimately, the three noncoding chloroplast regions just mentioned were included in the CBOL PWG (2009) publication, as a "supplementary set" of barcode sources.

In line with the findings of Sass et al. (2007) concerning the unsuitability of *matK* and *rbcL* as DNA barcoding loci in the gymnosperm order Cycadales, Nicolalde-Morejón et al. (2010) further established that *psbK-I* has the best single-locus performance for the Mexican cycads in terms of unique molecular species identification, in comparison with several other chloroplast regions that were entertained as

DNA barcoding region candidates in land plants. Interestingly, calculations carried out by the CBOL PWG for their selection of the core plant DNA barcoding regions had also indicated that *psbK-I* is a good single-gene DNA barcoding locus: this intergenic spacer had only a very slight deficit in terms of percentage discrimination success (i.e. the third criterion for DNA barcoding loci selection; CBOL PWG 2009, p. 12794) compared with *trnH-psbA*, which occupied the first place (see Figure 1c in CBOL PWG (2009)). Therefore, while our cycad results (Nicolalde-Morejón et al. 2010) in fact provide support for this CBOL PWG observation, they also suggest that the selection of *matK* and *rbcL* as core DNA barcode regions was perhaps premature. In Table I, we show a quantitative summary of the comparative performance of *psbK-I* as a (potential) DNA barcoding region in *Zamia* species from Megamexico, according to the results previously discussed by Nicolalde-Morejón et al. (2010).

Comparing DNA barcoding single-individual vs. population-level matrices in Zamia*: CAOS analyses can be sensitive to alignment*

As shown in Table III, our CAOS-based analysis of the population-level matrix for *psbK-I* in *Zamia* retrieves 14 DNA diagnostic sites—matching sites 1, 31, 51, 74, 83, 153, 202, 220, 253, 319, 361, 411, 622, and 664 of the corresponding alignment. The number of DNA diagnostics in this matrix represents four additional sites with respect to the total number of sites obtained with CAOS for a single-individual dataset (Nicolalde-Morejón et al. 2010; in Figure 1, the sites in common among the two matrices are marked with shading). Interestingly, out of the nonmatching sites between the two matrices, there is a single character that was present in the single-individual alignment (from Nicolalde-Morejón et al. (2010)), but disappeared in the population-level 1 (namely, character 618 in the former data set; this site contains the DNA diagnostic for *Zamia lacandona*). Further visual inspection of the combined matrix indicates, in addition, the presence of three dimorphic sites—two for *Zamia fischeri*, and one for *Zamia polymorpha*, which nevertheless contribute to the distinctive DNA barcode of these two species; and one site in *Zamia paucijuga* that distinguishes three populations (out of 13), leaving the remaining 10 with no DNA diagnostics to separate them from the most common combination of character states for the CAOS-defined DNA diagnostics.

In summary, our single-gene comparative evaluation of a single-individual vs. a population-level data set demonstrates that a character-based analytical regime can be sensitive to alignment discrepancies. We have observed that sequence alignments can change slightly with the addition of multiple sequence entries for the same populations and/or species; evidently, the incorporation of such information is a realistic possibility in any plant DNA barcoding project for which natural populations are available for study. Therefore, we suggest that the stipulation of a set of "definitive" DNA barcodes for reference in a DNA barcoding library in taxonomic genera, families, etc. should consider comparisons of at least two matrices of aligned sequences: on the one hand, the "minimal matrix" that only contains one sequence per species, and the matrix that contains the largest sampling at the population scale, on the other. In this way, researchers interested in using character-based methods in DNA barcoding projects can have high confidence that they will include most—if not all—naturally occurring nucleotidic variation in the regions they might have selected for the molecular identification of their study taxa.

Prospects for molecular identification and "integrative taxonomy" studies in Neotropical Zamia *species*

The present work constitutes only a step toward a comprehensive interpretation of the patterns of molecular variation existing in current populations in species of *Zamia* occurring in the Neotropics. Such interpretation could ultimately help us understand the evolution of the cycad genus *Zamia* in this biogeographic region. From a strictly taxonomic and systematic standpoint, it is important to keep in mind that differences in the intensity of botanical collections in the biogeographic subregions where *Zamia* populations occur might be biasing our estimation of such variation (Nicolalde-Morejón F, Stevenson DW, González-Astorga J, Vergara-Silva F, unpublished observations). In turn, we might be currently underestimating the utility of candidate DNA barcoding regions for the construction of reference libraries for various research and applied purposes; this is certainly an issue that should be addressed in subsequent attempts to refine DNA barcoding reference libraries in any cycad genus. In any event, restricting for the moment our scope to *Zamia*, and with the taxonomic data already at hand (Nicolalde-Morejón et al. 2009), we predict that future detailed analysis of molecular data sets for some of the species in this genus could suggest recircumscriptions and/or nomenclatural changes.

In our view, the prospects for these taxonomic advances should be addressed in the conceptual framework associated with the "taxonomic circle" inference procedure advocated by DeSalle et al. (2005). In this regard, some of the best candidates for such ulterior "integrative taxonomy" studies are *Zamia loddigesii*, *Z. paucijuga*, and *Z. polymorpha*, three species of *Zamia* from Megamexico with wide ranges of morphological and karyotypic variation and ample geographic distributions (Caputo et al. 1996; Nicolalde-Morejón et al. 2009). Likewise, the Caribbean species of *Zamia* that are putatively sister taxa to the Mexican

Table III. Comparison of DNA diagnostic sites between the single-individual and the population-level *psbK-psbI* data sets for *Zamia* species from Megamexico.

	Diagnostic sites for *psbI-psbK* (single-individual/population level)														
Zamia species	–/1	1/31	–/51	47/74	58/83	143/153	–/202	–/220	251/253	–/319	386/361	439/411	618/–	666/622	709/664
Z. cremnophila	G	A	T	T	T	A	C	C	A	–	–	–	–	C	C
Z. fischeri	**T/A**⋆	A	**C**	T	T	A	C	**A/C**⋆	A	C	T	T	T	C	C
Z. furfuracea	G	A	T	T	T	A	C	C	A	C	T	T	T	C	C
Z. herrerae	G	A	T	T	T	A	C	C	A	C	T	T	T	C	C
Z. inermis	G	A	T	T	T	A	C	C	A	C	T	T	T	**T**	C
Z. katzeriana	G	A	T	T	T	A	C	C	A	C	T	T	T	C	C
Z. lacandona	G	A	T	T	T	A	C	C	A	C	T	T	**C**	C	C
Z. loddigesii	G	A	T	T	T	A	C	C	A	C	T	T	T	C	C
Z. paucijuga	G	A	T	T	T	A	$C_{1-9,13}/A_{10-12}$	C	A	C	T	T	T	C	C
Z. polymorpha	G	A	T	**C**	T	**G/A**⋆	**C**	C	A	C	**C**	T	T	C	C
Z. prasina	G	A	T	T	**C**	A	C	C	A	C	T	T	T	C	C
Z. purpurea	G	A	T	T	T	A	C	C	A	C	T	T	T	C	C
Z. soconuscensis	G	**G**	T	T	T	A	C	C	A	C	T	T	T	C	C
Z. spartea	G	A	T	T	T	A	C	C	A	C	T	T	T	C	C
Z. standleyi	G	A	T	T	T	A	C	C	A	C	T	T	T	C	C
Z. tuerckheimii	G	A	T	T	T	A	C	C	A	C	T	T	T	C	C
Z. variegata	G	A	T	T	T	A	C	C	A	C	T	T	T	C	C
Z. vazquezii†	G	A	T	T	T	A	C	C	A	C	T	T	T	C	C
Z. cunaria†	G	A	T	T	T	A	C	C	A	C	T	T	T	C	C
Z. integrifolia†	G	A	T	T	T	A	C	C	A	**T**	T	T	T	C	C
Z. manicata†	G	A	T	T	T	A	C	C	**C**	C	T	**C**	T	C	**G**
Z. pygmea†	G	A	T	T	T	A	C	C	A	C	T	T	T	C	C

Diagnostic sites in common for the two matrices/alignments are indicated with shading. Notice the polymorphisms in *Z. fischeri*, *Z. polymorpha*, and particularly in *Z. paucijuga*. ⋆Only a few individuals displayed the nucleotidic variant; †Only one individual was studied in these species.

species *Z. fischeri* (Caputo et al. 2004) constitute another good candidate group for integrative taxonomic work. Most species of *Zamia* in the Caribbean clade exhibit morphological diversity between populations on any given island and from island to island. As a result there are, historically, over 35 validly published epithets for a currently recognized six to eight species (Eckenwalder 1980; Stevenson 1987; Géigel 2003). Currently, there are two competing and mutually exclusive hypotheses to explain this. In one scenario, all of the species on a given island have evolved on that island and thus similarities of morphologies from island to island are the result of parallel evolution to the same edaphic features. Thus, each island either has one very variable species or a set of unique species. The contrasting hypothesis is that there are several species, each with a unique distribution pattern among the islands either from vicariance or dispersal or both. As in the case of the *Zamia* species from Megamexico, we suggest that the key to selecting among these competing hypotheses for the Caribbean *Zamia*s might lie in an integrative approach to inference in which the DNA barcoding information could "break out of the taxonomic circle" (*sensu* DeSalle et al. 2005, p. 1908) already formed by the biogeographic, morphological—and possibly also the ecological—data points.

Acknowledgements

The authors thank Raúl Jiménez-Rosenberg and Martha Gual for their support during the planning stages of the project. They also acknowledge Alejandro Espinosa de los Monteros and Eduardo Ruíz for data analysis support, and Julia Hernández Villa and Janet Nolasco Soto for technical assistance in the laboratory. Finally, the authors are grateful to the staff at the JBC (Xalapa, Veracruz, Mexico) for access to living specimens in their cycad collections, and to the Montgomery Botanical Center, Miami, FL for supplying samples of some species.

This research project was supported by Mexican Comisión Nacional para el Uso y Conservación de la Biodiversidad (CONABIO) Grant GE004, and US NSF Grants BSR-8607049 and EF-0629817 (to D.W.S).

Declarations of interest: The authors report no conflicts of interest. The authors alone are responsible for the content and writing of the paper.

References

Bergmann T, Hadrys H, Breves G, Schierwater B. 2009. Character-based DNA barcoding: A superior tool for species identification. Berl Münch Tierärztl Wochenschr 122:446–450.

Caputo P, Cozzolino S, Gaudio L, Moretti A, Stevenson D. 1996. Karyology and phylogeny of some Mesoamerican species of *Zamia* (Zamiaceae). Am J Bot 83:1513–1520.

Caputo P, Cozzolino S, De Luca P, Moretti A, Stevenson D. 2004. Molecular phylogeny of *Zamia*. In: Walters T, Osborne R, editors. Cycad classification: Concepts and recommendations. Oxford: CABI Publishing. p 149–158.

CBOL PWG Hollingsworth PM, Forrest LL, Spouge JL, Hajibabaei M, Ratnasingham S, van der Bank M, et al. 2009. A DNA barcode for land plants. Proc Natl Acad Sci USA 106: 12794–12797.

Chase MW, Soltis DE, Olmstead RG, Morgan D, Les DH, Mishler BD, Duvall MR, Price RA, Hills HG, Qiu Y-L, Kron KA, Rettig JH, Conti E, Palmer JD, Manhart JR, Sytsma KJ, Michaels HJ, Kress WJ, Karol KG, Clark WD, Hedren M, Gaut BS, Jansen RK, Kim K-J, Wimpee CF, Smith JF, Furnier GR, Strauss SH, Qui-Yun Xiang, Plunkett GM, Soltis PS, Swensen SM, Williams SE, Gadek PA, Quinn CJ, Eguiarte LE, Golenberg E, Learn GH Jr, Graham SW, Barrett SCH, Dayanandan S, Albert VA1993. Phylogenetics of seed plants: An analysis of nucleotide sequences from the plastid gene *rbc*L. Ann Mo Bot Gard 80:528–580.

Chase MW, Salamin N, Wilkinson M, Dunwell JM, Kesanakurth RP, Haidar N, Savolainen V. 2005. Land plants and DNA barcodes: Short-term and long-term goals. Philos Trans R Soc Lond B Biol Sci 360:1889–1895.

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

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

DeSalle R. 2006. Species discovery versus species identification in DNA barcoding efforts: Response to Rubinoff. Conserv Biol 20: 1545–1547.

DeSalle R. 2007. Phenetic and DNA taxonomy; a comment on Waugh. BioEssays 29:1289–1290.

DeSalle R, Egan MG, Siddall ME. 2005. The unholy trinity: Taxonomy, species delimitation and DNA barcoding. Philos Trans R Soc Lond B Biol Sci 360:1905–1916.

Doyle JJ, Doyle JL. 1987. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. Phytochem Bull 19:11–15.

Eckenwalder JE. 1980. Taxonomy of the West Indian cycads. J Arnold Arbor Harv Univ 61:701–722.

Erickson DL, Spouge J, Resch A, Weigt LA, Kress WJ. 2008. DNA barcoding in land plants: Developing standards to quantify and maximize success. Taxon 57:1304–1316.

Fazekas AJ, Burgess KS, Kesanakurti PR, Graham SW, Newmaster SG, Husband BC, et al. 2008. Multiple multilocus DNA barcodes from the plastid genome discriminate plant species equally well. PLoS ONE 3:e2802.

Fazekas AJ, Kesanakurti PR, Burgess KS, Percy DM, Graham SW, Barrett SCH, et al. 2009. Are plant species inherently harder to discriminate than animal species using DNA barcoding markers? Mol Ecol Resour 1(Suppl):130–139.

Ford CS, Ayres KL, Toomey N, Haider N, van Alphen Stahl J, Kelly LJ. 2009. Selection of candidate coding DNA barcoding regions for use on land plants. Bot J Linn Soc 159:1–11.

Géigel LG. 2003. Zamiaceae. Flora Rep Cuba Ser A 8:1–21.

Hall TA. 1999. BioEdit: A user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symp Ser 41:95–98.

Hebert PDN, Cywinska A, Ball SL, deWaard JR. 2003a. Biological identifications through DNA barcodes. Proc R Soc London B Biol Sci 270:313–321.

Hebert PDN, Ratnasingham S, de Waard JR. 2003b. Barcoding animal life: Cytochrome *c* oxidase subunit 1 divergences among closely related species. Proc R Soc London B Biol Sci 270(Suppl):S96–S99.

Hilu KW, Borsch T, Müller K, Soltis DE, Soltis PS, Savolainen V, et al. 2003. Angiosperm phylogeny based on *matK* sequence information. Am J Bot 90:1758–1776.

Huang S, Chiang YC, Schaal BA, Chou CH, Chiang TY. 2001. Organelle DNA phylogeography of *Cycas taitungensis*, a relict species in Taiwan. Mol Ecol 10:2669–2681.

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. Proc Natl Acad Sci USA 102:8369–8374.

Lahaye R, Savolainen V, Duthoit S, Maurin O, van der Bank M. 2008a. A test of *psbK-psbI* and *atpF-atpH* as potential plant DNA barcodes using the flora of the Kruger National Park as a model system (South Africa), Nat Prec, hdl:10101/npre.2008.1896.1 (posted May 2008, retrieved July 2010).

Lahaye R, van der Bank M, Bogarin D, Warner J, Pupulin F, Gigot G, et al. 2008b. DNA barcoding the floras of biodiversity hotspots. Proc Natl Acad Sci USA 105:2923–2928.

Little DP, Stevenson DW. 2007. A comparison of algorithms for the identification of species using DNA barcodes: Examples for gymnosperms. Cladistics 23:1–21.

Lowenstein JH, Amato G, Kolokotronis SO. 2009. The real *maccoyii*: Identifying tuna fish with DNA barcodes—Contrasting characteristic attributes and genetic distances. PLoS ONE 4: e7866.

Maddison WP, Maddison DR. 2010. Mesquite: A modular system for evolutionary analysis. Version 2.73 [Online]. Available at: http://mesquiteproject.org

Nicolalde-Morejón F, Vovides AP, Stevenson DW. 2009. Taxonomic revision of *Zamia* in Mega-Mexico. Brittonia 61: 301–335.

Nicolalde-Morejón F, Vergara-Silva F, González-Astorga J, Stevenson DW, Vovides AP, Sosa V. 2010. A character-based approach in the Mexican cycads supports diverse multigene combinations for DNA barcoding Cladistics, in press. doi:10.1111/j.1096-0031.2010.00321.x.

Pennisi E. 2007. Wanted: A barcode for plants. Science 318: 190–191.

Rach J, DeSalle R, Sarkar IN, Schierwater B, Hadrys H. 2008. Character-based DNA barcoding allows discrimination of genera, species and populations in Odonata. Proc R Soc London B Biol Sci 275:237–247.

Rzedowski J. 1991. Vegetación de México. Mexico City: Limusa. p 431.

Saitou N, Nei M. 1987. The neighbor-joining method: A new method for reconstructing phylogenetic trees. Mol Biol Evol 4: 406–425.

Sarkar IN, Planet PJ, DeSalle R. 2008. CAOS software for use in character-based DNA barcoding. Mol Ecol Resour 8: 1256–1259.

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.

Stevenson DW. 1987. Again, the West Indian Zamias. Fairchild Trop Gard Bull 42:23–27.

Swofford DL. 2002. PAUP*. Phylogenetic analysis using parsimony (*and other methods). Sunderland, MA: Sinauer Associates.

Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG. 1997. The ClustalX windows interface: Flexible strategies for multiple sequence alignment aided by quality analysis tool. Nucleic Acids Res 25:4876–4882.