Annals of Botany **108**: 1445–1451, 2011
doi:10.1093/aob/mcr158, available online at www.aob.oxfordjournals.org

ANNALS OF BOTANY Founded 1887

PART OF A SPECIAL ISSUE ON PALM BIOLOGY

DNA barcoding: a new tool for palm taxonomists?

Marc L. Jeanson[1,2,*], Jean-Noël Labat[2] and Damon P. Little[1]

[1]*Lewis B. and Dorothy Cullman Program for Molecular Systematics, The New York Botanical Garden, Bronx, NY 10458-5126, USA and* [2]*Département systématique et évolution, UMR 7205, Muséum National d'Histoire Naturelle, Paris 75005, France*
* *For correspondence. E-mail* mjeanson@nybg.org

Received: 31 October 2010 Returned for revision: 17 December 2010 Accepted: 3 May 2011 Published electronically: 14 July 2011

• *Background and Aims* In the last decade, a new tool – DNA barcoding – was proposed to identify species. The technique of DNA barcoding is still being developed. The Consortium for the Barcode of Life's Plant Working Group (CBOL-PWG) selected two core markers (*matK* and *rbcL*) that now must be tested in as many taxa as possible. Although the taxonomy of palms (Arecaceae/Palmae) has been greatly improved in the past decades, taxonomic problems remain. Species complexes, for example, could significantly benefit from DNA barcoding. Palms have never before been subjected to a DNA barcoding test.

• *Methods* For this study, 40 out of the 48 species of the southeast Asian tribe Caryoteae (subfamily Coryphoideae) were included. In total, four DNA markers – three plastid encoded (*matK*, *rbcL* and *psbA-trnH*) and one nuclear encoded (nrITS2) – were analysed to determine if adequate variation exists to discriminate among species.

• *Key Results* The combination of three markers – *matK*, *rbcL* and nrITS2 – results in 92 % species discrimination. This rate is high for a barcoding experiment. The two core markers suggested by the CBOL-PWG, *rbcL* and *matK*, have a low species discrimination rate and need to be supplemented by another marker. In Caryoteae, nrITS2 should be chosen over *psbA-trnH* to supplement the two 'core' markers.

• *Conclusions* For the first time a test of DNA barcoding was conducted in Arecaceae. Considering that palms have highly variable mutation rates compared with other angiosperms, the results presented here are encouraging for developing DNA barcoding as a useful tool to identify species within this ecologically important tropical plant family.

Key words: DNA barcoding, Arecaceae, Caryoteae, *Caryota*, *Arenga*, *Wallichia*, *rbcL*, *matK*, *psbA-trnH*, nrITS2.

INTRODUCTION

Recent estimates state that as few as 4 % to approx. 40 % of the species on earth have been described in the scientific literature (Systematics Agenda, 1994; Pennisi, 2003). The intertropical zone shelters high levels of species diversity (Reaka-Kudla *et al.*, 1996; Kress and Erickson, 2008) but faces the strongest threats of extinction (Laurance and Peres, 2006), inducing an ever increasing rate of species loss (Pimm and Raven, 2000; Thomas *et al.*, 2004).

A huge amount of research is still needed to produce a satisfactory inventory of biodiversity. This, combined with the range and intensity of threats to this diversity, has served to emphasize the role of systematics in modern biology (Wilson, 1989; May, 1990). A new tool for identifying species called 'DNA barcoding' was proposed earlier this decade by Hebert and colleagues (Hebert *et al.*, 2003, 2004). They advocated the use of short DNA sequences for biological identification. This innovation gave rise to many controversial questions about the nature and purpose of systematics and appropriate funding levels for various sub-disciplines (Lipscomb *et al.*, 2003; DeSalle *et al.*, 2005; Rubinoff, 2005; DeSalle, 2006; Rubinoff *et al.*, 2006).

For the last several years, the plant barcoding community and especially the Plant Working Group of the Consortium for the Barcode of Life (CBOL-PWG) have focused on the identification of a universally informative plant barcode (CBOL Plant Working Group, 2009). In the end, *matK* and *rbcL* were selected by the CBOL-PWG as 'core' markers for plant identification (http://www.barcoding.si.edu/PDF/PlantWG/CBOL%20Decision%20-%20Plant%20Barcode%20Regions.pdf). It is imperative that these markers be tested in a broad range of laboratories across a diverse set of land plants. Although these markers have been widely used to construct phylogenies in Arecaceae, they have not been evaluated in a barcoding context.

Almost completely restricted to tropical and sub-tropical areas, palms (Arecaceae/Palmae) contain about 2400 species (Govaerts and Dransfield, 2005) and are the sixth largest monocot family. Although the taxonomy of this family has greatly improved at the generic level (Dransfield *et al.*, 2008), problems remain (Baker *et al.*, 2011). This is especially true at the species level. Several species complexes have been identified that are hard to resolve using only morphological characters [e.g. *Calamus* or *Geonoma* (Henderson and Martins, 2002; Henderson, 2011)]. Palms are considered study models in the field of tropical forest ecology (Dransfield, 1988; Kahn and De Granville, 1992; Dransfield *et al.*, 2008), but remain poorly represented in herbaria, leading to a severe lack of taxonomic advancement (Henderson *et al.*, 1995). One of the reasons for the relative scarcity of palms in herbaria is that they are more difficult to collect than most other plants due to their (often) large size. This lack of collection and taxonomic knowledge is even more problematic considering that many species in Arecaceae are known to be under major threat (Johnson and IUCN/SSC, 1996). In some palm groups, plastid DNA seems to evolve more slowly than in other monocots (Gaut *et al.*, 1992, 1996; Asmussen, 1999*a*, *b*; Hahn, 2002; Lewis and Doyle, 2002; Gunn, 2004; Roncal *et al.*, 2005; Thomas *et al.*, 2006), whereas in others the rate of evolution seems unexpectedly high (Bayton, 2005; Cuenca and Asmussen-Lange, 2007). This makes the analysis and the understanding of inter- and intraspecies relationships more complex. Thus, DNA barcoding in palms could be challenging, but the development of such a tool would be very helpful to taxonomists and ecologists (Valentini *et al.*, 2009). No DNA barcoding test in palms has ever been published.

Tribe Caryoteae (subfamily Coryphoideae) includes three genera, *Caryota*, *Arenga* and *Wallichia*, distributed from mainland Asia to the western Pacific and Australia. This group is being monographed by the first author. This tribe was selected for this experiment for different reasons. Amongst palms *Caryota* is one of the most difficult palm genera to collect (Dransfield, 1974), leading to a scarcity of material and thus the traditional alphataxonomic approach is difficult. Another reason is that *Arenga caudata* and *Arenga hookeriana* form a species complex for which the traditional morphological approach is uninformative (Hodel and Vatcharakorn, 1998; Henderson, 2009). Here, the usefulness of four DNA markers to discriminate among the species traditionally recognized within tribe Caryoteae were tested.

MATERIALS AND METHODS

Markers used

Four DNA regions were amplified. Following the requirements of the CBOL Plant Working Group (2009) parts of two plastid genes – *matK* and *rbcL* – were included. Two markers, *psbA-trnH*, a plastid intergenic spacer (Kress *et al.*, 2005), and the nuclear ribosomal internal transcribed spacer (nrITS2) (Kress *et al.*, 2005; Sass *et al.*, 2007) were added.

Taxon sampling

Species delimitation used here was based on previous work: for *Caryota* the taxonomic reference was Hahn's revision (Hahn, 1992), for *Arenga* we followed the checklist published by Govaerts and Dransfield (2005), and for *Wallichia*, Henderson's revision was our reference (Henderson, 2007). A total of 40 species out of the approx. 48 found in Caryoteae were included in this study (Table 1). For some species more than one accession was examined, resulting in 88 samples in total (Table 1; Supplementary Data Table S1, available online). Material for DNA extraction was obtained from field trips and living collections (FTG, K and Montgomery Botanical Center).

TABLE 1. *List of the species included in the Tribe Caryoteae sorted by genus*

Genus	Species	No. of individuals available for study	No. of samples for which all four markers were amplified	Total no. of sequences
Caryota	*bacsonensis*	2	0	5
	cumingii	0		
	gigas	2	1	7
	kiriwongensis	2	1	7
	maxima	5	3	16
	mitis	4	0	13
	monostachya	4	1	11
	no	2	2	8
	obtusa	4	3	15
	ophiopellis	1	1	4
	rumphiana	1	1	4
	sympetala	3	0	2
	urens	1	1	4
	zebrina	1	0	3
Arenga	*australasica*	1	0	3
	brevipes	2	1	6
	caudata	10	6	36
	distincta	1	1	4
	engleri	5	2	15
	hastata	2	2	5
	hookeriana	4	1	3
	listeri	4	1	4
	longicarpa	1	0	4
Arenga	*longipes*	0		
	micrantha	0		
	microcarpa	1	0	3
	mindorensis	0		
	obtusifolia	3	1	10
	pinnata	1	0	3
	plicata	0		
	porphyrocarpa	2	1	7
	retroflorescens	1	0	2
	ryukyuensis	2	1	7
	talamauensis	0		
	tremula	1	1	4
	undulatifolia	2	2	8
	westerhoutii	4	1	11
	wightii	1	1	4
Wallichia	*caryotoides*	1	0	3
	disticha	2	1	7
	gracilis	3	1	9
	lidiae	0		
	marianneae	1	0	3
	nana	0		
	oblongifolia	1	1	4
	triandra	0		

For each species, the number of individuals (samples) included in the study (0 indicates that the species was not sampled), the number of individuals (samples) that provided a sequence for all four markers (0 indicates that the species was not included in the analysis because all four markers could not be sequenced) and the total number of sequences obtained for each species (submitted to GenBank) are given.

DNA extraction, amplification and sequencing

DNA was extracted from silica gel-dried (0·01–0·05 g) leaves using a modified cetyl trimethyl ammonium bromide (CTAB) extraction method as described in Kang *et al.* (1998) or using DNeasy$^{®}$ Plant MiniKits (Qiagen Inc., Valencia, CA, USA). PCRs were conducted in a total reaction volume of 15 µL containing 8·13 µL of autoclaved ion-exchanged water, 1·5 µL of dNTP mixture (stock of 2·5 mM of each dNTP), 1·5 µL of bovine serum albumin (BSA; 0·25 µg μL^{-1} stock), 1·5 µL of buffer [200 mM Tris pH 8·8, 100 mM KCl, 100 mM $(NH_4)_2SO_4$, 20 mM $MgSO_4 \cdot 7H_2O$, 1 % (v/v) Triton X-100, 50 % (w/v) sucrose, 0·25 % (w/v) cresol red], 1 µL of each primer (0·67 µM final concentration), 0·3 µL of *Taq* polymerase and 0·25 µL of DNA. The amplicon size of *matK* ranges from 846 to 852 bp; primers used (5'–3') were CGTACAGTACTTTT GTGTTTACGAG and ACCCAGTCCATCTGGAAATCT TGGTTC (K. J. Kim, pers. com.), The amplicon size of *rbcL* is 654 bp; primers used (5'–3') were ATGTCACCACAAACAGAGACTAAAGC and GAAACGG TCTCTCCAACGCAT (Kress and Erickson, 2007; Fazekas *et al.*, 2008). The amplicon size of *psbA-trnH* ranges from 318 to 820 bp. This length variation is mostly the result of small scattered insertions/deletions without an apparent taxonomic pattern (Supplementary Data Alignment File, available online). In addition there is an approx. 40 bp insertion, composed mostly of A/T, present in *Caryota* as well as in *A. hastata* and *A. distincta*. The primers used (5'–3') were GTTATGCATGAACGTAATGCTC and CGCGCATGGTGG ATTCACAATCC (Sang *et al.*, 1997; Tate and Simpson, 2003). The amplicon size of nrITS2 ranges from 538 to 559 bp; primers used (5'–3') were ATGCGATAC TTGGTGTGAAT and GACGCTTCTCCAGACTACAAT (Shu-Jiau *et al.*, 2007). PCR for *matK* and *rbcL* was performed with an initial denaturation of 2 min 30 s at 94 °C followed by ten cycles under the following conditions: 94 °C for 30 s, 54 °C for 30 s, 72 °C for 30 s, and 25 cycles under the following conditions: 88 °C for 30 s, 54 °C for 30 s and 72 °C for 30 s, terminated by an extension of 72 °C for 10 min. PCR for *psbA-trnH* was performed with an initial denaturation of 2 min and 30 s at 95 °C followed by 35 cycles under the following conditions: 95 °C for 30 s, 58 °C for 30 s, 64 °C for 1 min, terminated by an extension of 72 °C for 7 min. PCR for nrITS2 was performed with an initial denaturation of 3 min at 94 °C followed by 35 cycles under the following conditions: 95 °C for 30 s, 56 °C for 30 s, 72 °C for 30 s, terminated by an extension of 72 °C for 7 min.

The PCRs were run in a Gene Amp® PCR system 9700 or on an Eppendorf vapo Protect Mastercycler pro S. PCR products were purified with ExoSAP-IT (exonuclease I and shrimp alkaline phosphatase). Sequencing was performed using the Big Dye terminator sequencing kit version 3·1 (Applied Biosystems, Carlsbad, CA, USA) and an ABI 3700 sequencer (sequencing conducted in the Department of Genome Sciences, University of Washington). The complete sequences were manually edited using Staden (http://staden.sourceforge.net/). Sequences and voucher information are archived in GenBank (accessions JF344793–JF345078; Supplementary Data Table S1)

Analysis of the data

Evaluation of the relative discriminatory power of the different markers was conducted with a suite of PERL scripts (CBOL Plant Working Group, 2009) available at http://www.nybg.org/files/scientists/dlittle/PWG.html. Statistical analyses used the R package (R v2·11·1; MASS v7·3-6; agricolae v1·0-9; http://cran.r-project.org/).

To make meaningful comparisons between markers, samples with missing sequence data were eliminated from the analyses, resulting in a matrix of 26 species and 39 sequences per marker. Micro- and macroinversions found in *psbA-trnH* were manually re-inverted to see if this could make a difference in the species discrimination. The original *psbA-trnH* sequences were analysed as one marker and the 'corrected' sequences were analysed as a separate marker.

For each marker, all possible pairwise global alignments were calculated using MUSCLE v3·6 (Edgar, 2004) and unambiguous nucleotide changes between sequences were tabulated (pairwise p-distances). A given species was considered distinct if all samples, for that species, could be unambiguously differentiated (i.e. one or more nucleotide changes resulting in an intraspecific distance greater than zero) from all other samples. Discrimination was calculated for marker combinations by summing the number of differences for each sample. An index of sequence depth (coverage) and quality (*B*) was calculated from manually edited contigs using QV30 as the minimum acceptable value (Little, 2010). Sequence quality was calculated for marker combinations using the mean value for each sample. Thus, estimates of average sequence quality for individual markers and all possible combinations of markers could be compared. Statistical differences in PCR success, species discrimination and sequence quality among markers were examined using the test of Scheffé (1953) at $P = 0{\cdot}05$. The binomial distribution was used for the Scheffé tests of PCR success and species discrimination, while the Gaussian distribution was used for the Scheffé test of sequence quality.

RESULTS

The amplification success was of 91·7 % for *psbA-trnH* (abbreviated as p), 90·58 % for *rbcL* (abbreviated as r), 88·1 % for nrITS2 (abbreviated as i) and 73 % for *matK* (abbreviated as m). PCR success was not statistically different among markers (Scheffé test, $P = 0{\cdot}05$) (Scheffé, 1953).

The best sequence quality (B_{30}; Little, 2010) was obtained from the 'core' markers with almost 0·9 for *matK* (m) and 0·93 for *rbcL* (r; Fig. 1). The B_{30} of these two markers, individually as well as combined (mr), is statistically different from that of all other markers and their combinations (Scheffé test, $P = 0{\cdot}05$) (Scheffé, 1953). There were no other unambiguously statistically distinct groupings based on sequence quality (with and without the inclusion of marker combinations). *PsbA-trnH* (p) has the lowest sequence quality value ($B_{30} = 0{\cdot}69$). The 'corrected' *psbA-trnH* sequences (d; 11 reverse complemented microinversions in positions 83–95 of the alignment; Supplementary Data Alignment File; Fig. 2) were plotted with sequence quality identical to the uncorrected sequences. Between the *rbcL/matK* (r/m) cluster and the *psbA-trnH*/'corrected' *psbA-trnH* (p/d) cluster is a cloud of different combinations of *matK*, *psbA-trnH* and *rbcL* (mpr, mp and pr). The sequence quality values are the average of *psbA-trnH* (p) and *matK/rbcL* (m/r), ranging from 0·79 to 0·82 (Fig. 1). For the nrITS2 (i) and combinations [nrITS2/*matK*/*psbA-trnH* (imp), nrITS2/*psbA-trnH*/*rbcL* (ipr),

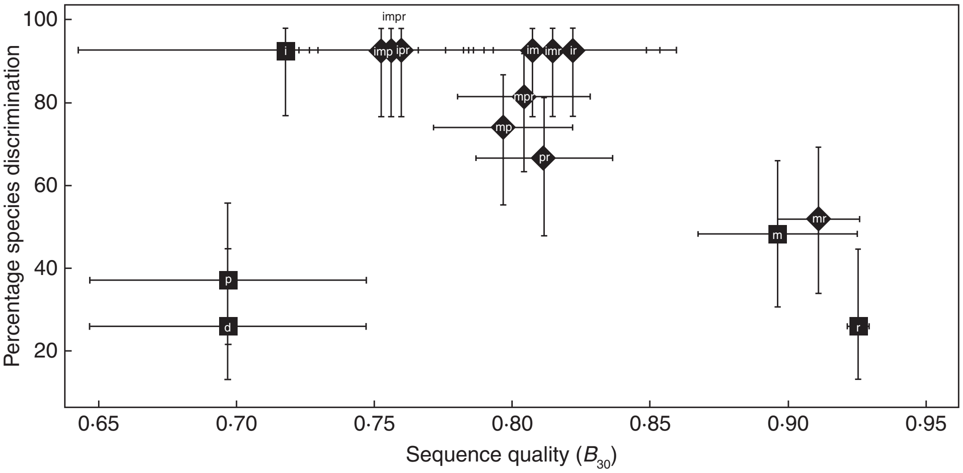

FIG. 1. Percentage species discrimination in relation to sequence quality. m = *matK*; r = *rbcL*; p = *psbA-trnH*; i = nrITS2; d= 'corrected' *psbA-trnH*. Squares represent single barcodes. Diamonds represent various combinations of markers. Errors bars are 95 % confidence intervals. Combinations with 'd' are not shown for clarity (they exactly mirror the relationship between 'd' and 'p').

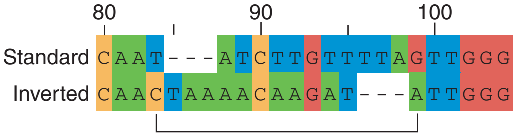

FIG. 2. An example of microinversions within *psbA-trnH* sequences. 'Inverted' sequences can be 'corrected' to the 'standard' (most frequently retrieved) sequence type by replacing the bracketed segment with its reverse complement. Nucleotide positions correspond to the full MUSCLE alignment (Supplementary Data Alignment File, available online).

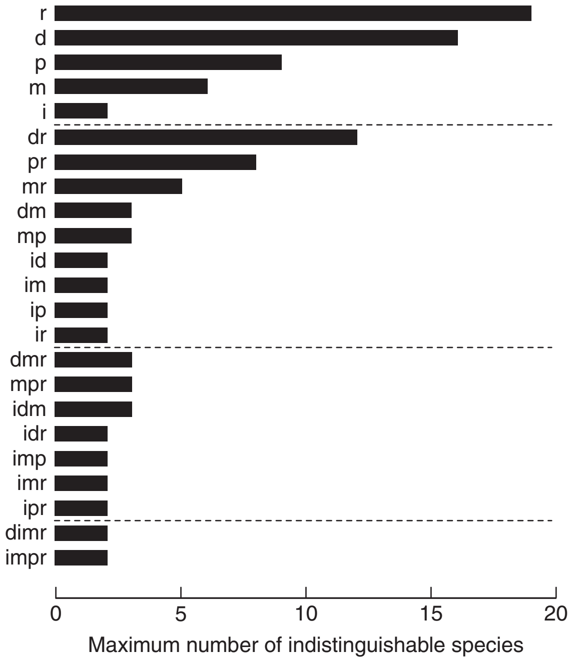

FIG. 3. Maximum number of indistinguishable species for each marker and combination of markers; d = *psbA-trnH* corrected, i = *nrITS2*, m = *matK*, p = *psbA-trnH*, r = *rbcL* (ir, for example, represents the combination of nrITS2 and rbcL).

nrITS2/*matK* (im), nrITS2/*matK*/*rbcL* (imr), nrITS2/*rbcL* (ir)] sequence quality ranges between 0·72 and 0·83.

The percentage species discrimination with *rbcL* (r) alone, *matK* (m) alone or their combination (mr) is either low [26 % for *rbcL* (r) alone, the lowest value] or average [48 and 51·8 % for *matK* (m) and *matK* + *rbcL* (mr), respectively]. Sequence quality (B_{30}) for *rbcL* (r) and *matK* (m) is uniformity high. *PsbA-trnH* (p) combines a low B_{30} with poor species discrimination (37 %) – the lowest species discrimination (26 %) is obtained with 'corrected' *psbA-trnH* sequences (d) and *rbcL* (r). Although not included in the final data set due to missing data for other markers, the *psbA-trnH* (p) sequences of *Caryota mitis* were almost entirely inverted. Different 5' and 3' ends in the two individuals sampled suggest that the inversions have independent origins. Within the cloud of various combinations of *matK* (m), *psbA-trnH* (p) and *rbcL* (r; Fig. 1), the percentage species discrimination ranges between 66 and 81 % – higher than the values obtained with the individual component markers. The highest species discrimination (92 %) is reached with nrITS2 (i) and its various combinations – although at the cost of lower sequence quality (B_{30}). For species discrimination, there were no unambiguously statistically distinct groupings (with, and without, the inclusion of marker combinations).

If identification fails, the size of the resulting ambiguously identified group is measured, for each marker, by the maximum number of indistinguishable species. The minimum value is 0 (meaning identification success) and the maximum value is the size of the study set (meaning an invariant marker). In Caryoteae, the maximum number of indistinguishable species ranges from two for nrITS2 (i) and its combinations (meaning that, at most, one will not be able to distinguish between a set of two species, using this marker) to 19 for *rbcL* (r; Fig. 3); meaning that if only this marker is used, in some cases, 19 different species are indistinguishable from another (Fig. 3). In contrast, *psbA-trnH* (p) alone has a maximum number of indistinguishable species of nine and, for some combinations in which it is involved, three [*matK*/*rbcL*/ *psbA-trnH* (mrp) and *matK*/*psbA-trnH* (mp)]. In combination with *psbA-trnH* (pr), *rbcL* still presents a high maximum number of indistinguishable species (eight). If used alone, 'corrected' *psbA-trnH* (d) has the next highest maximum number of indistinguishable species – 16 species are potentially unrecognizable. The combination of 'corrected' *psbA-trnH* with other markers improves the situation [*nrITS2*/'corrected' *psbA-trnH* (id) = 2; *rbcL*/'corrected' *psbA-trnH* (rd) = 3] except if the other marker is *rbcL* [*rbcL*/'corrected' *psbA-trnH* (rd) = 12]. *MatK* (m) has an intermediate value if used alone as a barcode – six species are potentially indistinguishable.

DISCUSSION

Palms are well known in the systematics community for highly variable rates of plastid DNA evolution (Baker *et al.*, 1999; Asmussen *et al.*, 2000; Baker *et al.*, 2000*b*; Dransfield *et al.*, 2008). We therefore expected that DNA barcoding would not be very serviceable for species discrimination, but, interestingly, our results indicate that DNA barcoding may be a very useful tool – at least within Caryoteae. Sequence quality (B_{30}) ranged between 0·69 and 0·92 (Fig. 1). Both the statistically higher B_{30} of *matK*, *rbcL* and their combination and the low B_{30} for *psbA-trnH* conform to prior expectations (CBOL Plant Working Group, 2009). For the animal *CO1* sequence, the minimum B_{20} value, defined in the BARCODE data standard, is 0·75 (http://www.barcoding.si.edu/PDF/DWG_data_standards-Final.pdf). The CBOL has not yet defined sequence quality criteria for plant barcode markers. The criteria used by the CBOL Plant Working Group (2009) mandated a minimum B_{30} of 0·375 (Little, 2010). To date, most plant barcoding studies have focused on species discrimination, and to a lesser extent marker universality. However, sequence quality must be a criterion in selecting a DNA barcode because the BARCODE data standard requires high-quality sequences (CBOL Plant Working Group, 2009). There is only one published analysis of plant barcode sequence quality: B_{30} values ranged between 0·40 and 0·85 (CBOL Plant Working Group, 2009; Little, 2010). In comparison, our results are slightly better (0·69–0·93). The discrepancy is likely to be the result of focused vs. diverse sampling (Caryoteae vs. angiosperms, gymnosperms and cryptogams). Given the CBOL Plant Working Group (2009) criteria, all the markers and their combinations tested here provide a reliable enough sequence quality (Fig. 1).

A good DNA barcode should have acceptable sequence quality with species discrimination as high as possible. The combination of *matK* and *rbcL* does not provide good species discrimination in the current study (Fig. 1), and a broad range of levels of discrimination success has been obtained in other studies published to date (Hollingsworth *et al.*, 2011). The low discrimination success of 'core' markers observed in Caryoteae is similar to what has been observed in *Geonomateae* and *Archontophoenicinae* (Conny B. Asmussen-Lange, pers. com.). However, the two 'core' regions selected by the CBOL should always be used for DNA barcoding, with supplemental markers, such as *psbA-trnH* or nrITS2, added as required. A uniform database of markers is required to maximize the utility of DNA barcoding. Therefore, the two 'core' makers (*matK* and *rbcL*) must be sequenced for all barcoded plants regardless of their utility within a given clade. Failure to create a uniform database of core marker sequences would require a barcoder first to identify a sample using morphology and then sequence the markers best represented in the database for that clade – completely negating the purpose of barcoding.

Taking into account sequence quality and species discrimination, three good options emerge from our data set: nrITS2 combined with *matK* (im), nrITS2 combined with *matK* and *rbcL* (imr) or nrITS2 combined with *rbcL* (ir; Fig. 1). Some combinations of markers resulted in species discrimination as high as 92 % [nrITS2/*matK*/*psbA-trnH* (imp), nrITS2/*psbA-trnH*/*rbcL* (ipr), nrITS2/*matK* (im), nrITS2/*matK*/*rbcL* (imr), nrITS2/*rbcL* (ir); Fig. 1], which is high for plant DNA barcoding. The results presented here show that the addition of a third marker to the two 'core' markers greatly improves species discrimination power. Our results also indicate that for Caryoteae nrITS2 is more informative than *psbA-trnH* and should be combined with *rbcL* and *matK*. The maximum number of indistinguishable species also justifies the choice of nrITS2 over *psbA-trnH* as a supplemental marker. Among the species included in this study, at most two cannot be distinguished (maximum number of indistinguishable species) if *matK*, *rbcL* and nrITS2 are used together (imr; Fig. 3) – making these markers quite accurate (although not 100 % reliable).

Multiple divergent nrITS copies are reported from Calamoid palms (Baker *et al.*, 2000*a*). In a DNA barcode context, this potentially leads to misidentification due to differential sampling of divergent paralogues. We did not observe more than one band for any marker (visualized on agarose gels) nor did any of the electropherograms show evidence of overlapping signals that would be suggestive of the presence of divergent paralogues. These observations lead us to believe that, in tribe Caryoteae, nrITS2 can be reliably used. Compared with other studies that included nrITS2 as a barcode marker in monocots (Yao *et al.*, 2010), our results are rather encouraging (74·2 % species discrimination in the monocots vs. 92 % in our study).

Other markers, or parts of markers, could be informative for barcoding in the palm family, such as *ndhF* (Hahn, 2002; Cuenca and Asmussen-Lange, 2007), *rpb2* or *PRK* (Thomas *et al.*, 2006; Trenel *et al.*, 2007). These potentially informative markers were not selected by the CBOL for different reasons: for example, *ndhF* is not found in gymnosperms and some pteridophytes so it cannot be used as a universal marker.

The complex molecular evolution of *psbA-trnH* makes it difficult to use as a barcode (Lahaye *et al.*, 2005; Whitlock *et al.*, 2010). Microinversions in *psbA-trnH* (p; Fig.2) have been reported from a variety of angiosperms (reviewed in Whitlock *et al.*, 2010). Whitlock *et al.* (2010) demonstrated that manually correcting inverted sequences has a positive impact on understanding of the relationship among sequences (i.e. phylogenies). In our case, 'correcting' the inversions resulted in lower species discrimination because the inversions are species specific in two-thirds of our sample (p vs. d in Fig. 1). Therefore, 'correcting' the microinversion harmonizes one-third of the species and destroys a diagnostic set of bases for two-thirds of the species.

To test if these promising results can be extended to the whole family, this experiment should be replicated with other groups of Arecaceae. If these results were to be applicable to the rest of the family, DNA barcoding could become an ideal supplementary tool for palm systematics and greatly enhance taxonomic knowledge especially within species complexes.

SUPPLEMENTARY DATA

Supplementary data are available online at www.aob.oxfordjournals.org and consist of the following. Table S1: samples, voucher information and GenBank accessions used in this study. Alignment File: MUSCLE alignment of *psbA-trnH* (.txt file). Positions identical to that of the first sequence are reported as points.

ACKNOWLEDGEMENTS

We thank William J. Baker and Edith Kapinos from the Royal Botanic Garden Kew for providing DNA aliquots. Thanks to Carl Lewis and Patrick Griffith for having helped us to have access to the living collections of Fairchild Tropical Botanical Garden and Montgomery Botanical Center. Thanks to Jeff Claussen from Christmas Island National Park for supplying DNA material of *Arenga listeri*. We thank Elisa Suganuma for her help in the laboratory, and Thomas Couvreur for commenting on previous versions of the manuscript. Thanks to four anonymous reviewers for their comments and suggestions. This work was supported by the Annette Kade Fellowship, Lewis B. and Dorothy Cullman Program for Molecular Systematics, the Alfred P. Sloan Foundation and a PPF Barcode grant from the Muséum National d'Histoire Naturelle, Paris, France.

LITERATURE CITED

Asmussen CB. 1999*a*. Relationships of the tribe Geonomae (Arecaceae) based on plastid *rps* 16 DNA sequences. *Acta Botanica Venezuelica* **22**: 65–76.

Asmussen CB. 1999*b*. Toward a chloroplast DNA phylogeny of the tribe Geonomeae (Palmae). In: Henderson A, Borchsenius F. eds. *Evolution, variation and classification of palms.* Bronx, The New York Botanical Garden Press, 121–129.

Asmussen CB, Baker WJ, Dransfield J. 2000. Phylogeny of the Palm family (Arecaceae) based on *rps*16 intron and *trn*L-*trn*F plastid DNA sequences In: Wilson KL, Morrison DA. eds. *Monocots: systematics and evolution.* Melbourne: CSIRO, 525–537.

Baker WJ, Asmussen CB, Barrow SC, Dransfield J, Hedderson TA. 1999. A phylogenetic study of the palm family (Palmae) based on chloroplast DNA sequences from the trnL-trnF region. *Plant Systematics and Evolution* **219**: 111–126.

Baker WJ, Dransfield J, Hedderson TA. 2000*a*. Phylogeny, character evolution, and a new classification of the calamoid palms. *Systematic Botany* **25**: 297–322.

Baker WJ, Hedderson TA, Dransfield J. 2000*b*. Molecular phylogenetics of subfamily Calamoideae (Palmae) based on nrDNA ITS and cpDNA rps16 intron sequence data. *Molecular Phylogenetics and Evolution* **14**: 195–217.

Baker WJ, Norup MV, Clarkson JJ, *et al*. 2011. Phylogenetic relationships among arecoid palms (Arecaceae: Arecoideae). *Annals of Botany* **108**: 1417–1432.

Bayton RP. 2005. *Borassus L. and the borassoid palms: systematics and evolution.* PhD Thesis. University of Reading, Reading, UK.

CBOL Plant Working Group. 2009. A DNA barcode for land plants. *Proceedings of the National Academy of Sciences, USA* **106**: 12794–12797.

Cuenca A, Asmussen-Lange CB. 2007. Phylogeny of the palm tribe Chamaedoreeae (Arecaceae) based on plastid DNA sequences. *Systematic Botany* **32**: 250–263.

DeSalle R. 2006. Species discovery versus species identification in DNA barcoding efforts: response to Rubinoff. *Conservation Biology* **20**: 1545–1547.

DeSalle R, Egan MG, Siddall M. 2005. The unholy trinity: taxonomy, species delimitation and DNA barcoding. *Philosophical Transactions of the Royal Society B: Biological Sciences* **360**: 1905–1916.

Dransfield J. 1974. Notes on *Caryota no* Becc. and other Malesian *Caryota* species. *Principes* **18**: 87–93.

Dransfield J. 1988. Forest palms. In: Cranbrook Earl of. ed. *Malaysia.* Oxford: Pergamon Press, 49–55.

Dransfield J, Uhl NW, Asmussen CB, Baker JW, Harley MM, Lewis CE. 2008. *Genera Palmarum: the evolution and classification of palms.* Richmond, UK: Kew Publishing.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* **32**: 1792–1797.

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. doi:10.1371/journal.pone.0002802.

Gaut BS, Muse SV, Clark WD, Clegg MT. 1992. Relative rates of nucleotide substitution at the rbcL locus of monocotyledonous plants. *Journal of Molecular Evolution* **35**: 292–303.

Gaut BS, Morton BR, McCaig BC, Clegg MT. 1996. Substitution rate comparisons between grasses and palms: synonymous rate differences at the nuclear gene Adh parallel rate differences at the plastid gene rbcL. *Proceedings of the National Academy of Sciences, USA* **93**: 10274–10279.

Govaerts R, Dransfield J. 2005. *World checklist of palms.* Kew, UK: Royal Botanic Gardens Kew.

Gunn FB. 2004. The phylogeny of the Cocoeae (Arecaceae) with emphasis on *Cocos nucifera. Annals of the Missouri Botanical Garden* **91**: 505–522.

Hahn WJ. 1992. *Biosystematics and evolution of the genus Caryota (Palmae: Arecoideae).* PhD Thesis. University of Wisconsin–Madison, Madison.

Hahn WJ. 2002. A phylogenetic analysis of the Arecoid line of palms based on plastid DNA sequence data. *Molecular Phylogenetics and Evolution* **23**: 189–204.

Hebert PDN, Ratnasingham S, deWaard JR. 2003. Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. *Proceedings of the Royal Society B: Biological Sciences* **270**: S96–S99.

Hebert PDN, Stoeckle MY, Zemlak TS, Francis CM. 2004. Identification of birds through DNA barcodes. *PLoS Biology* **2**: 1657–1663.

Henderson A. 2007. A revision of Wallichia. *Taiwania* **52**: 1–11.

Henderson A. 2009. *Palms of Southeast Asia.* Princeton, NJ: Princeton University Press.

Henderson A. 2011. A revision of Geonoma (Arecaceae). *Phylotaxa* **17**: 1–27.

Henderson A, Galeano G, Bernal R. 1995. *Field guide to the palms of the Americas.* Princeton, NJ: Princeton University Press.

Henderson A, Martins R. 2002. Classification of specimens in the *Geonoma stricta* (Palmae) complex: the problem of leaf size and shape. *Brittonia* **54**: 202–212.

Hodel DR, Vatcharakorn P. 1998. *The palms and cycads of Thailand.* Lawrence, KS: Allen Press Inc.

Hollingsworth PM, Graham SW, Little DP. 2011. Choosing and using a plant DNA barcode. *PLoS One* **6**: e19254. doi:10.1371/journal.pone.0019254.

Johnson DV, IUCN/SSC PSG. 1996. *Palms. Their conservation and sustained utilization. Status survey and conservation action plan.* Gland, Switzerland and London, UK: IUCN.

Kahn F, De Granville J-J. 1992. *Palms in forest ecosystems of Amazonia.* Berlin, Springer-Verlag.

Kang HW, Cho AG, Yoon UH, Eun MY. 1998. A rapid DNA extraction method for RFLP and PCR analysis from a single dry seed. *Plant Molecular Biology Reporter* **16**: 1–9.

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. doi:10.1371/journal.pone.0000508.

Kress WJ, Erickson DL. 2008. DNA barcoding – a windfall for tropical biology? *Biotropica* **40**: 405–408.

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, USA* **102**: 8369–8374.

Lahaye R, Civeyrel L, Speck T, Rowe NP. 2005. Evolution of shrub-like growth forms in the lianoid subfamily Secamonoideae (Apocynaceae s.l.) of Madagascar: phylogeny, biomechanics, and development. *American Journal of Botany* **92**: 1381–1396.

Laurance WF, Peres CA. 2006. *Emerging threats to tropical forests*. Chicago: The University of Chicago Press.

Lewis C, Doyle JJ. 2002. A phylogenetic analysis of tribe Areceae (Arecaceae) using two low-copy nuclear genes. *Plant Systematics and Evolution* **236**: 1–17.

Lipscomb D, Platnick N, Wheeler Q. 2003. The intellectual content of taxonomy: a comment on DNA taxonomy. *Trends in Ecology and Evolution* **18**: 65–66.

Little DP. 2010. A unified index of sequence quality and contig overlap for DNA barcoding. *Bioinformatics* **26**: 2780–2781.

May R. 1990. Taxonomy as destiny. *Nature* **347**: 129–130.

Pennisi E. 2003. Modernizing the tree of life. *Science* **300**: 1692–1697.

Pimm SL, Raven HP. 2000. Biodiversity: extinction by numbers. *Nature* **403**: 843–845.

Reaka-Kudla ML, Wilson DE, Wilson EO. 1996. *Biodiversity II*. Washington, DC: Joseph Henry Press.

Roncal J, Francisco-Ortega J, Asmussen CB, Lewis CE. 2005. Molecular phylogenetics of tribe geonomeae (Arecaceae) using nuclear DNA sequences of phosphoribulokinase and RNA polymerase II. *Systematic Botany* **30**: 275–283.

Rubinoff D. 2005. Utility of mitochondrial DNA barcodes in species conservation. *Conservation Biology* **20**: 1026–1033.

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

Sang T, Crawford DJ, Stuessy TF. 1997. Chloroplast DNA phylogeny, reticulate evolution, and biogeography of Paeonia (Paeoniaceae). *American Journal of Botany* **84**: 1120–1136.

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. doi:10.1371/journal.pone.0001154.

Scheffé H. 1953. A method for judging all contrasts in the analysis of variance. *Biometrika* **40**: 87–104.

Shu-Jiau C, Jui-Hung Y, Cheng-Li F, Hui-Ling C, Tsai-Yun L. 2007. Authentication of medicinal herbs using PCR-amplified ITS2 with specific primers. *Planta Med* **73**: 1421–1426.

Systematics Agenda. 1994. *Charting the biosphere. Technical report.* The Bronx: The New York Botanical Garden.

Tate JA, Simpson BB. 2003. Paraphyly of Tarasa (Malvaceae) and diverse origins of the polyploid species. *Systematic Botany* **28**: 723–737.

Thomas JA, Telfer MG, Roy DB, *et al*. 2004. Comparative losses of British butterflies, birds and plants and the global extinction crisis. *Science* **303**: 1879–1881.

Thomas MM, Garwood NC, Baker WJ, *et al*. 2006. Molecular phylogeny of the palm genus *Chamaedorea*, based on the low-copy nuclear genes *PRK* and *RPB2*. *Molecular Phylogenetics and Evolution* **38**: 398–415.

Trenel P, Gustafsson MHG, Baker WJ, *et al*. 2007. Mid-Tertiary dispersal, not Gondwanan vicariance explains distribution patterns in the wax palm subfamily (Ceroxyloideae: Arecaceae). *Molecular Phylogenetics and Evolution* **45**: 272–288.

Valentini A, Pompanon F, Taberlet P. 2009. DNA barcoding for ecologists. *Trends in Ecology and Evolution*, **24**: 110–117.

Whitlock BA, Hale AM, Groff PA. 2010. Intraspecific inversions pose a challenge for the *trnH-psbA* plant DNA barcode. *PLoS One* **5**: e11533. doi:10.1371/journal.pone.0011533.

Wilson EO. 1989. The coming pluralization of biology and the stewardship of systematics. *Bioscience* **39**: 242–245.

Yao H, Song JY, Liu C, *et al*. 2010. Use of ITS2 region as the universal DNA barcode for plants and animals. *PLoS One* **5**: e13102. doi:10.1371/journal.pone.0013102.