DNA Barcoding in the Cycadales: Testing the Potential of Proposed Barcoding Markers for Species Identification of Cycads

Chodon Sass1, Damon P. Little2, Dennis Wm. Stevenson2, Chelsea D. Specht1**

1 Department of Plant and Microbial Biology, University of California at Berkeley, Berkeley, California, United States of America, 2 Cullman Program for Molecular Systematics Studies, The New York Botanical Garden, Bronx, New York, United States of America

INTRODUCTION

Barcoding all described species is an enormous task with large sums being spent annually toward this end [1]. The proposed utility of the Barcode of Life project has been debated [2–8] and fundamental challenges have been acknowledged that focus on (a) the identification of DNA regions useful at the appropriate taxonomic level, (b) development of universal primers for these regions, and (c) the proper use of DNA barcodes as taxonomic identifiers. Proponents argue that molecular barcodes can be used to identify new species and eliminate the need for the complex taxonomic training that is currently required for species description and identification [9]—helping to ease the taxonomic crisis, especially in countries with high biodiversity and small numbers of practicing taxonomists. However the patterns of sequence variation make it logically impossible to use DNA barcodes for species circumscription as originally proposed [see 8 for an empirical example, see 10 for a theoretical example]. Although barcodes are appealing as a powerful tool to identify already described species, the cautious among us argue that the use of a single locus for identification may produce misleading results especially considering the different evolutionary histories of organellar and nuclear genomes within a single species [11]. Moreover, there is limited intraspecific sequence variation data for the proposed barcoding loci in plants. Others reject the use of barcodes for taxonomic purposes on the grounds that species description and identification requires full taxonomic revisions and that "phylogenies" produced by barcoding genes do not necessarily represent evolutionary history [2,12].

Ultimately the ability to identify a sample to species could be useful in cases where specimens are not of adequate quality to make accurate identifications (e.g. adult forms verses larval forms, sterile vouchers of plant specimens) and for ecologists and conservation biologists to rapidly assess biological diversity. In this sense, barcoding acts as a “forensic” tool for the accurate identification of a sample to species. The species, in this case, needs to be both described as unique (i.e. monographed) with a known range of morphological and sequence variation and be represented in a DNA barcoding database. This is an enormous task, requiring active participation of taxonomists, DNA sequencing facilities, database managers, and funding agencies to support monography, DNA sequencing, continuous specimen and database management, and potentially, the recircumscription of species as new data become available.

In order for a region of DNA to be operative as a barcode, it must simultaneously contain enough variability to be informative for identification (i.e. contain unique identifiers), be short enough to sequence in a single reaction, and contain invariant regions that...
can be used to develop universal primers [13]. Unfortunately, it is difficult to find a single region of DNA that has all three of these properties. For animals, the mitochondrial cytochrome oxidase I (cox) gene has been successfully used for identification [5,14,15] although there are exceptions [for example, see 8,16,17]. For land plants, the cox gene, and the mitochondrial genome in general, is not useful for identification at the species level because of low levels of primary sequence variability [18,19]. Other regions often used for phylogenetic analysis across large groups of plants [e.g., rbcL] do not usually contain enough variability to identify individual species [but see 20,21]. Developing a barcoding region for plants is further complicated by extensive genome-wide horizontal gene transfer, hybridization, and homoplasmy [6].

Despite these obstacles, several gene regions have recently been proposed for use in land plants [4,22,23]. One set of loci includes a nuclear region, the ribosomal internal transcribed spacer with embedded 5.8S (nrITS), and a chloroplast region, the psbA-trnH intergenic spacer [4]. The combination of these two regions to positively and accurately identify taxa to species was tested on a subset of plants in the published analysis, but the combination is predicted to yield difficulties at the species level because nrITS is extremely variable in length—making analysis potentially more difficult—and psbA-trnH is likely to provide insufficient variation to reliably identify an organism to species, especially in groups with low divergence.

A portion of the chloroplast encoded large subunit ribosomal DNA, that is potentially “universally” amplifiable (Universal Plastid Aplicon; UPA), has also been proposed as a barcode for photosynthetic organisms. Available data suggest that although UPA may be variable at the species level in some algal lineages, it is not particularly variable among land plants [23].

A consortium of institutions operating under the auspices of the Plant Working Group (PWG) of the Consortium for the Barcode of Life (CBoL) initially suggested five chloroplast gene regions for evaluation as potential barcodes: matK, rpoC1, rpoB, accD and YCF5, with ndhJ as a potential sixth region [http://www.kew.org/barcoding/]. These markers were proposed because of their potential for amplification with universal primers and because they may harbor sufficient sequence diversity, individually or in combination, to distinguish among species. In order for either of these criteria to be demonstrated, members of the community must devote time and effort to evaluating the proposed regions in the plant group they study with the goal of developing a defined “barcoding workflow” for the taxonomic group in question (Figure 1).

The last step of any barcoding workflow is to use newly generated sequence data in combination with a well-maintained database to positively identify the species in question. The BOLD identification system developed as part of the ongoing barcoding initiative at the University of Guelph (www.barcodinglife.org;[24]) uses a Hidden-Markov Model to align a query sequence to a reference database of coxI sequences generated for animal barcodes and then select the most similar sequence(s) as the identification. Unfortunately this algorithm is only applicable to sequences that can be globally aligned [24]. Some of the proposed plant barcodes are non-coding regions that cannot be sensibly aligned across land plants and thus could not benefit from BOLD-ID. Little and Stevenson [10] demonstrated that search algorithms can be successfully used on unaligned nucleotide sequence data, the most accurate and precise algorithms were, respectively, the commonly used local alignment search tool, BLAST [25] and a diagnostic method, DNA-BAR [26,27]. DNA-BAR was originally intended as an algorithmic tool to select oligonucleotides for identification of microorganisms by Southern hybridization, but DNA-BAR’s output file can be queried by a PERL script (DEGENBAR) that uses a simple matching algorithm to pick the most similar sequences(s) in the reference database. Provided that DNA-BAR is run on an input file containing each sequence and its reverse complement, both forward and reverse query sequences can be used to search the reference database.

The Cycadales are unique in their evolutionary position and importance for conservation, and as such are important to include in tests of proposed barcoding regions. Cycads are often thought of as “living fossils” and the extant taxonomic assemblage represents only a sampling of the ancient diversity. Most extant genera have representative fossils that date to the early Permian—indicating a minimum of 50-60 million years of morphological evolution that might enable us to observe greater nucleotide divergence than one would expect in more recently derived species [28–30]. Because of the relictual nature of the genera and their high value in illegal horticultural trade, cycads are an important focus for conservation efforts [31]. Most cycad genera are listed in CITES Appendix I and the remaining are listed in Appendix II [32]. An easy-to-use and inexpensive identification system would enable non-experts to identify illegally harvested individuals and help prevent the illegal trade of these species. Ideally it would be possible to identify an individual to species and perhaps even identify the population from whence the specimen was removed, allowing for proper repatriation of illegally harvested individuals.

The only way to determine if it is possible to use DNA barcodes across a wide variety of plant life is to test the proposed loci and search algorithms. In this study, we test the proposed barcoding regions in the members of the ancient gymnosperm order Cycadales in an effort to develop a functional barcoding workflow for this order.

**RESULTS**

**Proposed regions**

The primer pairs chosen using Ceratozamia hildae and Cycas chilensis (Figure 1) for ndhJ, rpoB and matK did not work well for the remaining taxa (Table 1): non-specific primer binding resulted in multiple bands or complete lack of amplification. Because the purpose of these experiments was to test the functionality and utility of the proposed barcoding conditions and primers (as per www.kew.org/barcoding) on cycads we did not to develop novel cycad-specific primers or reaction conditions. Further analyses were performed only on those primers that successfully generated single products under universal conditions: accD, YCF5 and rpoC1.

Sequences generated from these three regions were tested for their ability to provide unique species identifications using both BLAST and DNA-BAR/DEGENBAR. Neither algorithm was able to positively identify individuals to species due to a lack of unique species-specific sequence for all species tested. Both algorithms had some success with identification of individuals to genus with 63–93% of query sequences correctly identified depending on the marker used (Figure 2). Inspection of the alignment revealed that there were very few variable positions. Over the three tested DNA regions, approximately 10% of the bases were variable (93 of 917 total bases); for accD, 28 of 242 base pairs were variable; for rpoC1, 41 of 476 base pairs were variable; and for YCF5, 24 of 199 base pairs were variable.

**Secondary regions (nrITS and psbA-trnH)**

Because the chloroplast gene regions initially suggested by the Plant Working Group did not promise to distinguish among species even with our rather incomplete sampling, the alternative regions suggested by Kress et al. [4]—psbA-trnH intergenic spacer coupled with nrITS—were tested on the original 27–species set.
Figure 1. Barcoding optimization workflow. Step 1: genera used for testing all primer pairs, amplification products of each combination of primer pairs, and primer pair combination chosen for testing on more genera (highlighted in blue). Step 2: taxa subject to further testing and success of amplification with chosen primer pairs (highlighted in gray). Step 3: trimming all sequences to highly similar anchor regions. When possible, anchor regions were actually the primer binding sites. Step 4: each sequence entered into a database and used as a query sequence. The process is repeated for more species, with promising regions, or with new markers. The PWG suggested primer regions (www.kew.org/barcoding) are used as the example.

doi:10.1371/journal.pone.0001154.g001
The nrITS repeat (nrITS 1, 5.8S, and nrITS2) amplified cleanly in most species, but sequencing was difficult because nrITS in cycads (and other gymnosperms) is variable in length—approximately 1100 bp in most species, but around 1400 bp in Stangeria eriopus. In many species, the use of internal primers was required to generate contigs of the full sequence—making nrITS less desirable as a DNA barcode for cycads. A second potentially negative factor is the presence of long poly-G, poly-C, and poly-A repeats that are difficult to sequence through. Despite these issues, nrITS had sufficient variation to correctly identify individuals to species for the 27 individuals initially tested plus 4 additional species represented by sequences downloaded from GenBank (due to sequencing difficulties for Zamia and lack of fresh tissue samples for Bowenia).

Additional species were sampled from Dioon and Encephalartos to further test the functionality of nrITS for species-level identification. These genera were chosen because tissue samples were available that maximized the total percent coverage of species within each genus (7 out of 13 Dioon, 44 out of 65 Encephalartos). To further increase the number of species represented, available sequences from GenBank were included in the reference databases. For nrITS, a total of 96 sequences comprising 74 taxa were included in the ordinal-level analysis. Each genus was included in the ordinal analysis, and where possible more than one species from each genus was included (Table 2). In the ordinal-level database, all species were correctly identified to genus and 90.5% of queries correctly and uniquely identified the query sequence in the reference database. The success of self-identification is broken down by genus in Table 2. Genus-specific databases were made for Encephalartos, Cycas, and Macrozamia because some species could not be included in the ordinal-level database as the sequences did not contain the necessary anchor regions. In the generic-level databases the percent identification decreased: For Encephalartos, 26 of 44 (59.1%) species identified uniquely; for Cycas, 11 of 12 (91.7%) species identified uniquely; and for Macrozamia, 8 of 8 (100%) species identified uniquely. The nrITS locus had the

![Figure 2. Success of species and genus level identification using CBOL proposed gene regions. Performance (% correct identification) at genus and species levels is noted for each marker for each of the 10 genera tested. No bar indicates failure of identification (0% success). Values are identical for BLAST and DNA-BAR/DEGENBAR. doi:10.1371/journal.pone.0001154.g002](http://www.plosone.org/)

| Marker | Successful amplification (single bands) | Non-specific amplification (multiple bands) | No amplification | Used for identification |
|--------|----------------------------------------|---------------------------------------------|-----------------|------------------------|
| accD   | 26/27 = 96%                            | 1/27 = 4%                                   | 0/27 = 0%       | yes                    |
| YCF5   | 66/66 = 100%                           | 0/66 = 0%                                   | 0/66 = 0%       | yes                    |
| rpoC1  | 29/29 = 100%                           | 0/29 = 0%                                   | 0/29 = 0%       | yes                    |
| ndhJ   | 12/21 = 57% *                         | 6/21 = 29%                                  | 3/21 = 15%      | no                     |
| rpoB   | 7/21 = 33%                             | 14/21 = 67%                                 | 0/21 = 0%       | no                     |
| matK   | 5/21 = 24%                             | 11/21 = 52%                                 | 5/21 = 24%      | no                     |

*Only markers with near universal amplification success were sequenced and tested for identification. An * indicates very weak bands.*
highest success rate of the any of the markers tested; even though not all species could be correctly identified. Because of variation in length and sequence, complete alignments were not generated and the number of variable characters was not counted.

The *psbA-trnH* spacer primers and reaction conditions specified by Kress et al. [4] yielded distinct double bands in all but *Cycas* species (Figure 3A). Even with greatly increased annealing temperature, double bands were still present (Figure 3B). The utility of this region for barcoding was tested by sequencing the larger of the two fragments (after gel excision) from species that could not be uniquely identified in the nrITS database. The addition of *psbA-trnH* sequence data did not further resolve the non-specific identifications made by nrITS for the species tested (Table S1). Of 322 total characters, including gaps, in the cursory

| Genera    | Number of species analyzed | Total number of named species [40] | Percent of species that are represented and success rate of unique identification in ordinal level and generic level analyses |
|-----------|----------------------------|----------------------------------|--------------------------------------------------------------------------------|
|          | Order level | Genus level | Ordinal level | Generic level |
|           | Percent represented | Success | Percent represented | Success |
| Cycas     | 11          | 12        | 99             | 11.1          | 11/11 = 100% | 12.1          | 11/12 = 91.7% |
| Zamia     | 8           | –         | 59             | 13.6          | 8/8 = 100%  | –             | –             |
| Chigua    | 2           | –         | 2              | 100           | 2/2 = 100%  | –             | –             |
| Ceratozamia | 8           | –         | 21             | 38.1          | 8/8 = 100%  | –             | –             |
| Macrozamia | 7           | 8         | 40             | 17.5          | 7/7 = 100%  | 20            | 8/8 = 100%    |
| Stangeria | 1           | –         | 1              | 100           | 1/1 = 100%  | –             | –             |
| Encephalartos | 25    | 44         | 65             | 38.5          | 18/25 = 72% | 67.7          | 26/44 = 59.1% |
| Lepidozamia | 2           | –         | 2              | 100           | 2/2 = 100%  | –             | –             |
| Microcycas | 1           | –         | 1              | 100           | 1/1 = 100%  | –             | –             |
| Bowenia   | 2           | –         | 2              | 100           | 2/2 = 100%  | –             | –             |
| Dioon     | 7           | –         | 13             | 53.8          | 7/7 = 100%  | –             | –             |
| TOTAL     | 74          | 94        | 305            | 24.2          | 67/74 = 90.1% | 32.5          | 76/94 = 80.9% |

Success indicates results for both BLAST and DNA-BAR/DEGENBAR, which were identical.

doi:10.1371/journal.pone.0001154.t002

![Figure 3](image)

**FIGURE 3**

**Figure 3. Amplification of PCR producing using published *psbA-trnH* primers.** A: All genera except Cycas showed double bands, some genera had more prominent smaller fragments (e.g. *Dioon*), while others had more prominent larger fragments (e.g. *Macrozamia*). B: When more stringent reaction conditions were applied by running the amplification with the primer annealing temperature at 62ºC, double bands were still evident.

doi:10.1371/journal.pone.0001154.g003

DNA Barcoding in Cycads
Species are not included in the reference database [33]. Finally, sequencing using cloned PCR products will be necessary to ensure missed. If nrITS is to be used as a barcoding locus, further not be evident and potential variation within a species will be genome DNA, rare alleles (less than 10% of the amplicon) may sequences are generated directly from amplifications of whole genomic DNA, so it remains unclear whether both genes are being amplified from the chloroplast genome or if the second fragment could be nuclear DNA that was transferred from the chloroplast, a well-documented phenomenon [37]. Problems with amplification aside, psbA-trnH does not show promise as a barcoding locus for cycads because of its inability to provide specific identification for taxa that could not be distinguished with nrITS (Table S1).

Algorithm comparisons: BLAST and DNA-BAR/DEGENBAR
For our data sets, there was no difference between BLAST and DNA-BAR/DEGENBAR. For optimization, BLAST offers several advantages: It generates a more detailed output and is readily available and downloadable from NCBI. For use in barcoding in practice, either method seems to be similarly successful [10]. Standardization of an algorithm used for database searches as part of the DNA barcoding workflow should be promoted in order to provide maximally consistent results.

Conclusions
The goal of finding universal primer pairs and reaction conditions with unique internal sequence for all land plants remains elusive—not surprising given the complex history of land plant genomes. At least in cycads, the chloroplast regions tested do not have sufficient variability to provide the unique sequences (characters or combinations of characters) necessary to identify an individual to species. Nuclear regions may provide more usable variability, but such regions have not yet been identified. Perhaps a set of primers designed for each of the major clades of land plants (such as gymnosperms, pteridophytes, angiosperms, mosses, etc.) could be used simultaneously if universal tails were added to the primers so that although only one set of primers would amplify an unknown sample, the amplicon could be sequenced using a primer that matched the tail sequence. This approach would be especially useful in situations where little morphological information is available from the sample (e.g., determination of diet based on scat collections, identification of degraded, fragmented or sterile tissue). Alternatively, improved technology such as sequencing long regions of DNA (e.g., whole or partial chloroplast genomes) may enable identification based on both genome architecture and
Table 3. Primers and reaction conditions used in this study.

| Gene     | Primer      | Sequence 5'-3' | Rxn conditions |
|----------|-------------|----------------|----------------|
| matK     | 2.1 forward | CCTATCCACCTCGGAATCTTAG | 94°C-4 min |
|          | 2.1a forward| ATCCATTGGAATCTTAGTTC | 94°C-30 sec |
|          | 5 reverse   | GTTCTACACGAAAGAATTGC | 53°C-40 sec |
|          | 3.2 reverse | CTTCCTCTGTAAGAAATTTC | 72°C-40 sec |
|          |             |                 | 72°C-7 min |
| ndhJ     | 1 forward   | CATAGATCTTTGGGCTTYGA | 94°C-4 min |
|          | 2 forward   | TTGGGCTTCGATTACCAAGG | 94°C-30 sec |
|          | 3 reverse   | TTAGAAAACATATTAAACGAGG | 53°C-40 sec |
|          | 4 reverse   | CGATCCAGCTATCTGTAGATGG | 72°C-40 sec |
|          |             |                 | 72°C-7 min |
| rpoC1    | 1 forward   | GGGTAGACACTCTCTGATAATGG | 94°C-4 min |
|          | 2 forward   | GGCAAAGAGGGAAAGATACCG | 94°C-30 sec |
|          | 3 reverse   | TATCATGGAAATAGGTTAGGG | 53°C-40 sec |
|          | 4 reverse   | CTCAATGAGCATCTTGTATTTC | 72°C-40 sec |
|          |             |                 | 72°C-7 min |
| rpoB     | 1 forward   | AATGCATATCGTTGGAACTGG | 94°C-4 min |
|          | 2 forward   | ATGGACAGCTCAAGATCTTCC | 94°C-30 sec |
|          | 3 reverse   | TCAGGGATGAAATGGAATACG | 53°C-40 sec |
|          | 4 reverse   | GAATCCAGCTATCTGTAGATGG | 72°C-40 sec |
|          |             |                 | 72°C-7 min |
| accD     | 1 forward   | AGATGCGATCGTCTCGTACG | 94°C-4 min |
|          | 2 forward   | GTTGAGGAGAGGAAAGATACCG | 94°C-30 sec |
|          | 3 reverse   | CCCATGCACGATATCTGCACAG | 53°C-40 sec |
|          | 4 reverse   | GTTCCAGTCATCTGATCATGG | 72°C-40 sec |
|          |             |                 | 72°C-7 min |
| YCF5     | 1 forward   | GATATGGAATCGTCTCGTACG | 94°C-4 min |
|          | 2 forward   | ATGGGAGAGGAAAGATACCG | 94°C-30 sec |
|          | 3 reverse   | CGATCCAGCTATCTGTAGATGG | 53°C-40 sec |
|          | 4 reverse   | CCATATCTGATCTGATATGG | 72°C-40 sec |
|          |             |                 | 72°C-7 min |
| psbA-trnH [4] | fwd | GTTATGCATGAACTGCTAGTCTC | 94°C-5 min |
|           | rev        | CGCGCATGTTGGATTTCC | 94°C-1 min |
|           |            |                 | 55°C-1 min |
|           |            |                 | 72°C-1.5 min |
|           |            |                 | 72°C-7 min |
| psbA-trnH (including protein coding region) | fwd | CGAGCTGTGTTCTGGCTTC | 98°C-3 min |
|           | Rev (short-fragment) | GGGGTGGTGGTGGAGCAGT | 98°C-10 sec |
|           | Rev (long-fragment) | GGGGTGGTGGTGGAGCAGT | 98°C-10 sec |
|           |            |                 | 50°C-1 min |
|           |            |                 | 72°C-1.5 min |
| nrITS    | 5a fwd     | CCTCATTCTTGGAGGAAAGGAG | 94°C-5 min |
|           | 4 rev      | TCTCCGCTTATGATATGAG | 94°C-1 min |
|           |            |                 | 72°C-1.5 min |
|           |            |                 | 72°C-7 min |

Conditions for chloroplast markers from the Plant Working Group (www.kew.org/barcoding/protocols.html); conditions for psbA-trnH adopted from Kress et al. 2005. doi:10.1371/journal.pone.0001154.t003
additional variation captured by simply increasing the total amount of sequence.

MATERIALS AND METHODS

Taxon sampling and primer testing strategy
For each region, the Plant Working Group designed 4 primers (2 forward and 2 reverse) in their Phase I trials (www.kew.org/barcoding) in an attempt to increase the likelihood of finding a working combination. The primer pairs were first tested in all combinations on two species—Cycas ophiolitica and Ceratozamia hidalgoi—chosen based on their distant placement in the Cycadales phylogeny. Primers were considered successful if they amplified a single product. If a single band was obtained by more than one primer pair, the pair that generated the largest and brightest (highest PCR yield) of the bands was chosen. If amplification was successful in only one of the two species, the pair generating the brightest band for that species was selected. The best working primers were then tested for a set of 21 species representing 10 of 11 cycad genera. Gene regions with universal or near universal success in amplification were sequenced. Gene regions with variability that enabled specific positive identification were tested on additional species within each genus to further test the region’s ability to provide identification at the species level. This workflow is outlined in Figure 1.

Plant collection, DNA extraction, and amplification
Leaflets were clipped from live plants, dried in silica gel, and then stored at −80°C. Whole genomic DNA was extracted using DNeasy Plant Mini Kits (Qiagen, Valencia, CA) or a modified CTAB method [38] from fresh or frozen tissue. PCR amplification was performed from genomic DNA according to instructions on Kew’s website (www.kew.org/barcoding) for the 6 chloroplast regions or following Kress et al. [4] (Table 3). Some modifications were made to accommodate the use of iProof® High-Fidelity DNA polymerase (Bio-Rad, Hercules, CA). Amplified products were inspected on 1% agarose/TAE gels. Amplicons was cleaned through gel extraction using the QiaQuick® Gel Extraction Kit (Qiagen, Valencia, CA). Cycle sequencing was performed using AmpliTag™ (Amersham, Piscataway, NJ) or BigDye® v3.1 (Applied Biosystems, Foster City, CA) sequencing chemistry and an ABI PRISM® 3100 sequencer (Applied Biosystems, Foster City, CA).

Sequence alignment and determination of barcoding regions
Sequence editing and contig generation were performed using Sequencher (Gene Codes Corp., Ann Arbor, MI). Additional sequences for nrITS were downloaded from GenBank (see Table S1). If sequences did not include primer regions, all sequences were trimmed to an area with highly similar (>98% identity) sequence regions at the ends of the sequence reads—anchor regions (Figure 1). This was only necessary for nrITS and psbA sequences. Sequences from these loci were longer and more variable than other regions and as a result primer regions were not always sequenced. Sequences were used for further analysis only if they contained the anchor regions—ensuring that identification success was due to internal variability and not arbitrary factors such as sequence read length. In order to identify and trim sequences to the anchor regions, nrITS and psbA-trnH regions were aligned using CLUSTAL W [39] and then manually adjusted using MacClade (Sinauer Associates, Inc., Sunderland, MA). After elimination of sequences with ambiguous nucleotides and non-anchor-containing sequences, databases were created and individual sequences were queried against the databases with BLAST and DNA-BAR/DEGENBAR. Sequences excluded from the ordinal level database due to the absence of anchor regions were included in secondary databases that contained only a single genus provided that genus-specific anchors could be identified from sequences that were not long enough to be included in the ordinal-level database (this was only the case for nrITS sequences from Encephalartos, Cycas and Macrozamia; Table 2).

Comparability of results with different algorithms:
BLAST and DNA-BAR/DEGENBAR
The same sets of sequences were used both to generate databases and as query sequences for both BLAST and DNA-BAR/DEGENBAR [10]. BLAST queries were run without filtering. Before generating the database with DNA-BAR the sequences were run through a PERL script that added a reverse complement for each sequence in order to ensure that query sequences would match the database in either the forward or the reverse orientation. To test for unique species-specific barcodes that could be used for a species level identification, the sequence belonging to each species was copied from the database and used as a query sequence. If the query sequence returned an exact match only to itself, this was scored as a positive identification at the species level. If the query sequence returned an exact match to itself and other members of the same genus, this was scored as a negative identification at the species level, but a positive identification to the genus level. DNA-BAR/DEGENBAR returns only the highest scoring match(es), so the cutoff for genus and species identification is straightforward. For BLAST, an additional constraint was added: to positively score an identification at the genus level the best match as well as the next most similar sequence had to match the genus of the query sequence. If any other genus was included in the top two hits, the result was not considered genus specific.

SUPPORTING INFORMATION

Table S1 Supplemental Data: List of GenBank ID numbers of taxa used in tests for barcoding utility. Numbers in bold showed unique identifications in both BLAST and DNA-BAR/DEGENBAR, numbers in plain text were not identified uniquely. Results shown are those from the database with the most inclusive species sampling (i.e., the genus level database in the case of Encephalartos, Cycas and Macrozamia for nrITS).

ACKNOWLEDGMENTS

The authors thank the University of California Botanical Garden at Berkeley for access to cycad specimens and especially C. Carmichael, H. Forbes, and J. Finn (UCBG) for aiding in sampling efforts. We also thank H. Driscoll for discussions, laboratory and computational support, H. Driscoll and M. Bartlett for assistance in collecting tissue, and T. Gregory for ongoing support of cycad research at the UCBG and UCB.

Author Contributions
Conceived and designed the experiments: CDS CS DS. Performed the experiments: CS. Analyzed the data: CDS CS DL. Contributed reagents/materials/analysis tools: CDS DL. Wrote the paper: CDS CS DL.
REFERENCES

1. Ebach MC, Holderegge C (2005) More Taxonomy, Not DNA Barcoding. BioScience 55: 823–824.

2. Will KW, Rubinstein D (2004) Myth of the molecule: DNA barcodes for species cannot replace morphology for identification and classification. Cladistics 20: 47–55.

3. Rubiolo D (2006) Utility of Mitochondrial DNA Barcodes in Species Conservation. Conservation Biology 20: 1026–1033.

4. 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.

5. Hebert PDN, Alina C, Shelley LB, Jeremy Rd (2003) Biological identifications through DNA barcodes. Proceedings of the Royal Society B: Biological Sciences 270: 313–323.

6. Chase M, Salamin N, Wilkinson M, Dunwell J, Kesanakurthi R, et al. (2005) Land plants and DNA barcodes: short-term and long-term goals. Philosophical Transactions of the Royal Society B: Biological Sciences 360: 1089–1095.

7. Hebert PDN, Stoeckle MY, Zemlak TS, Francis CM (2004) Identification of Birds through DNA Barcodes. PLoS Biology 2: e312.

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

9. Blaxter M (2004) The promise of a DNA taxonomy. Philosophical Transactions of the Royal Society B: Biological Sciences 359: 669–679.

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

11. Moritz C, Cicero C (2004) DNA Barcoding: Promise and Pitfalls. Public Library of Science Biology 2: e508.

12. Will K, Mishler B, Wheeler Q (2005) The Perils of DNA Barcoding and the Need for Integrative Taxonomy. Systematic Biology 54: 844–851.

13. Stoeckle M (2003) Taxonomy, DNA, and the Bar Code of Life. BioScience 53: 796–797.

14. Hebert PDN, Penton EH, Burns JM, Janzen DH, Hallwachs W (2004) Ten species in one: DNA barcoding reveals cryptic species in the neotropical skipper butterfly Anthepetes furcifer. Proceedings of the National Academy of Sciences USA 101: 14612–14617.

15. Ward RD, Zemlak TS, Innes BH, Last PR, Hebert PDN (2005) DNA barcoding Australia’s fish species. Philosophical Transactions: Biological Sciences 360: 1847–1857.

16. Gompert Z, Nice CC, Fordyce JA, Forister ML, Shapiro AM (2006) Identifying units for conservation using molecular systematics: the cautionary tale of the Karner blue butterfly. Molecular Ecology 15: 1759–1768.

17. Meyer CP, Paulay G (2005) DNA Barcoding: Error Rates Based on Comprehensive Sampling. Public Library of Science Biology 3: e422.

18. Palmer JD, Herbon LA (1988) Plant mitochondrial DNA evolved rapidly in Australia’s fish species. Philosophical Transactions: Biological Sciences 360: 1847–1857.

19. Meyer CP, Herbon LA, (1988) Plant mitochondrial DNA evolved rapidly in Australia’s fish species. Philosophical Transactions: Biological Sciences 360: 1847–1857.

20. 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.

21. Driscoll HE, Barrington DS (2007) Origin of Hawaiian Polystichum (Dryopteridaceae) in the context of a world phylogeny. Am J Bot 94: 000–000.

22. Chase MW, Cowan RS, Hollingsworth PM, van den Berg C, Madrihan S, et al. (2007) A proposal for a standardised protocol to barcode all land plants. Taxon 56: 000–000.

23. Presting GG (2006) Identification of conserved regions in the plastid genome: implications for DNA barcoding and biological function. Canadian Journal of Botany 84: 1434–1443.

24. Ramasingham S, Hebert PJD (2007) BOLD: The Barcode of Life Data System (http://www.barcodinglife.org). Molecular Ecology Notes 7: 355–364.

25. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215: 403–410.

26. DasGupta B, Konwar KM, Mandouli I, Shvartsman AA (2005) DNA-BAR distinguished selection for DNA barcoding. Bioinformatics 21: 3424–3426.

27. Lüttgau BM, Konwar KM, Mandouli I, Shvartsman AA (2005) Highly scalable algorithms for robust string barcoding. International Journal of Bioinformatics Research and Applications 1: 145–161.

28. Whitehouse ML (2002) The Cycads. Portland, OR: Timber Press, Inc. 374 p.

29. Jones DL (1993) Cycads of the world; Young J, ed. Smithsonian Institution Press. 312 p.

30. Hermon EJ, Taylor TN, Taylor EL, Stevenson DW (2006) Catasetums of the Middle Triassic cycad Antarcticeycas schopfii and new insights into cycad evolution. American Journal of Botany 93: 724–738.

31. Donaldson JS (2003) Cycads: status survey and conservation action plan. IUCN–World Conservation Union.

32. CITES (2007) Convention on International Trade in Endangered Species of Wild Fauna and Flora. May.

33. Jakob SS, Blattert F (2006) A chloroplast genealogy of herdreum (poaceae): Long-term persisting haplotypes, incomplete lineage sorting, regional extinction, and the consequences for phylogenetic inference. Mol Biol Evol 23: 1602–1612.

34. Wolf PG, Rowe CA, Sinclair RB, Hasebe M (2003) Complete Nucleotide Sequence of the Chloroplast Genome from a Leptosporangiate Fern, Adiantum capillus-veneris L. DNA Research 10: 59.

35. Lüttgau BM, Konwar KM, Mandouli I, Shvartsman AA (2005) Highly scalable algorithms for robust string barcoding. International Journal of Bioinformatics Research and Applications 1: 145–161.

36. Lüttgau BM, Konwar KM, Mandouli I, Shvartsman AA (2005) Highly scalable algorithms for robust string barcoding. International Journal of Bioinformatics Research and Applications 1: 145–161.

37. 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.

38. Doyle JJ, Doyle JL (1987) A rapid DNA isolation procedure for small quantities of fresh leaf tissue. Phytochemistry Bulletin 19: 11–15.

39. Timmins JN, Ayliffe MA, Huang CY, Martin W (2004) Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes. Nature Reviews Genetics 5: 123–135.

40. Doyle JJ, Doyle JL (1987) A rapid DNA isolation procedure for small quantities of fresh leaf tissue. Phytochemistry Bulletin 19: 11–15.

41. 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.

42. Hill KD, Chase MW, Stevenson DW, Hills HG, Schutzman B (2003) The families and genera of cycads: a molecular phylogenetic analysis of cycadophyta based on nuclear and plastid DNA sequences. International Journal of Plant Systematics and Evolution 213: 199–205.

43. Hill KD, Chase MW, Stevenson DW, Hills HG, Schutzman B (2003) The families and genera of cycads: a molecular phylogenetic analysis of cycadophyta based on nuclear and plastid DNA sequences. International Journal of Plant Systematics and Evolution 213: 199–205.

44. Delaire C, Donoghue MJ, Stuessy TF (2004) A molecular phylogeny for the conifer order Taxodiaceae inferred from chloroplast matK gene and nuclear rDNA ITS region. Mol Phylogenet Evol 37: 214–234.

45. Hill KD, Chase MW, Stevenson DW, Hills HG, Schutzman B (2003) The families and genera of cycads: a molecular phylogenetic analysis of cycadophyta based on nuclear and plastid DNA sequences. International Journal of Plant Systematics and Evolution 213: 199–205.

46. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215: 403–410.

47. Wolf PG, Rowe CA, Sinclair RB, Hasebe M (2003) Complete Nucleotide Sequence of the Chloroplast Genome from a Leptosporangiate Fern, Adiantum capillus-veneris L. DNA Research 10: 59.

48. Lüttgau BM, Konwar KM, Mandouli I, Shvartsman AA (2005) Highly scalable algorithms for robust string barcoding. International Journal of Bioinformatics Research and Applications 1: 145–161.

49. Lüttgau BM, Konwar KM, Mandouli I, Shvartsman AA (2005) Highly scalable algorithms for robust string barcoding. International Journal of Bioinformatics Research and Applications 1: 145–161.

50. Lüttgau BM, Konwar KM, Mandouli I, Shvartsman AA (2005) Highly scalable algorithms for robust string barcoding. International Journal of Bioinformatics Research and Applications 1: 145–161.

51. Lüttgau BM, Konwar KM, Mandouli I, Shvartsman AA (2005) Highly scalable algorithms for robust string barcoding. International Journal of Bioinformatics Research and Applications 1: 145–161.