

DNA barcoding the floras of biodiversity hotspots

Renaud Lahaye*, Michelle van der Bank*, Diego Bogarin†, Jorge Warner†, Franco Pupulin†, Guillaume Gigot‡, Olivier Maurin*, Sylvie Duthoit*, Timothy G. Barraclough§, and Vincent Savolainen*§¶

*Department of Botany and Plant Biotechnology, APK Campus, University of Johannesburg, P.O. Box 524, Auckland Park 2006, Johannesburg, South Africa; †Lancker Botanical Garden, University of Costa Rica, P.O. Box 1031-7050 Cartago, Costa Rica; ‡Royal Botanic Gardens, Kew, Richmond TW9 3DS, United Kingdom; and §Imperial College London, Silwood Park Campus, Buckhurst Road, Ascot SL5 7PY, United Kingdom

Edited by Daniel H. Janzen, University of Pennsylvania, Philadelphia, PA, and approved December 17, 2007 (received for review October 18, 2007)

DNA barcoding is a technique in which species identification is performed by using DNA sequences from a small fragment of the genome, with the aim of contributing to a wide range of ecological and conservation studies in which traditional taxonomic identification is not practical. DNA barcoding is well established in animals, but there is not yet any universally accepted barcode for plants. Here, we undertook intensive field collections in two biodiversity hotspots (Mesoamerica and southern Africa). Using >1,600 samples, we compared eight potential barcodes. Going beyond previous plant studies, we assessed to what extent a “DNA barcoding gap” is present between intra- and interspecific variations, using multiple accessions per species. Given its adequate rate of variation, easy amplification, and alignment, we identified a portion of the plastid *matK* gene as a universal DNA barcode for flowering plants. Critically, we further demonstrate the applicability of DNA barcoding for biodiversity inventories. In addition, analyzing >1,000 species of Mesoamerican orchids, DNA barcoding with *matK* alone reveals cryptic species and proves useful in identifying species listed in Convention on International Trade of Endangered Species (CITES) appendices.

CITES | Kruger National Park | Mesoamerica

DNA barcoding is a diagnostic technique for species identification, using a short, standardized DNA region, i.e., the “DNA barcode” (www.barcoding.si.edu). It is, however, challenging to find a suitable genomic region for DNA barcoding a wide range of taxa. Indeed, for DNA barcoding to work, sequence variation must be high enough between species so that they can be discriminated from one another; however, it must be low enough within species that a clear threshold between intra- and interspecific genetic variations can be defined. Although the use of DNA barcoding for identification and taxonomy has been controversial (1, 2), a growing scientific community has embraced DNA barcoding as a practical tool for biodiversity studies, for example to facilitate inventories of very diverse but taxonomically poorly known regions (3–6). DNA barcoding, using the mitochondrial *cox1* gene (COI) (7–10), is now well established for animals, but the quest for a universal DNA barcode in plants is still disputed (11, 12).

Kress *et al.* (13) proposed originally that the *trnH-psbA* plastid region would be a suitable universal barcode for land plants. Concurrently, the newly established “plant working group” from the consortium for the barcoding of life tested a series of other genomic regions at first disregarding *trnH-psbA* because of its complex molecular evolution (14). It was also proposed that, because the plastid genome is evolving so slowly relative to other genomes, more than one barcode may be necessary to provide enough variation for this technique to work (15–17). However, several competing proposals have so far been put forward, which need thorough evaluations. Kress and Erickson (16) proposed to combine the original *trnH-psbA* barcode from Kress *et al.* (13) with *rbcL*, following analyses from Newman *et al.* (17). By contrast, Chase *et al.* (15) proposed either to combine *rpoc1*, *rpoB*, and *matK* or *rpoc1*, *matK*, and *trnH-psbA*, whereas Taberlet *et al.* (18) suggested the *trnL* intron as a suitable plant barcode. Furthermore, tests of potential DNA barcodes have been based on a taxonomic coverage approach, necessarily encompassing just a few represen-

tatives from a wide range of distantly related groups of land plants (13, 15–17). However, the critical test of evaluating the applicability of DNA barcoding for biodiversity inventories in species-rich geographic areas has been lacking.

Here, we focus on two biodiversity hotspots (19, 20), Mesoamerica and Maputaland–Pondoland–Albany in southern Africa, in which we analyze >1,600 plant specimens. We test eight potential DNA barcodes, six of which were made publicly available at the plant working group’s website [www.kew.org/barcoding (15)], whereas a further two were proposed by Kress and Erickson (16). Our study sites have been chosen for their exceptional plant diversity and contrasting habitats. Costa Rica comprises tropical forests and has one of the richest orchid floras in the world. Although there is a well developed network of protected areas in Costa Rica, the orchid flora remains under constant threat from deforestation and illegal trade. Orchids are also well known to be difficult to identify, particularly when they are sterile, which makes them an ideal model group in which to test DNA barcoding techniques. In southern Africa, we have undertaken our study in the Kruger National Park (KNP), one of the largest protected areas in the world. The KNP is renowned for its large game animals but less for its flora, which is under continuous pressure from mega-herbivores. Home to ≈600 species of trees and shrubs (21), the KNP area has the highest tree diversity of any of the world’s temperate regions.

During 2005–2007, we conducted extensive fieldwork to collect samples for this study. We used several metrics to evaluate the various potential barcoding regions. Intra- and interspecific genetic divergences were assessed by using pairwise calculations (22). Statistical tests were used to compare divergences. Phylogenetic analyses were performed to look for species monophyly. Genetic clustering algorithms (23, 24) were applied to test whether the coalescent process in a given barcode matched species delimitation.

Results and Discussion

PCRs were generally successful with all potential barcodes, except *ndhJ* and *ycf5*, which did not amplify efficiently in orchids. It is known that *rbcL* is not variable enough in orchids (25), so we did not sequence this gene in this group. The *rbcL* and *trnH-psbA* regions did not amplify in the achlorophyllous *Hydnora johannis* but amplified in other parasitic plants. A portion of the *matK* exon amplified easily by using primers 390F and 1326R from Cuénoud

Author contributions: M.v.d.B., J.W., and V.S. designed research; R.L., D.B., F.P., G.G., O.M., and S.D. performed research; T.G.B. contributed new reagents/analytic tools; R.L. analyzed data; and R.L., M.v.d.B., and V.S. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

Data deposition: The sequences reported in this paper have been deposited in the GenBank database (accession nos. EU254252–EU254410 and EU213263–EU214530).

See Commentary on page 2761.

¶To whom correspondence should be addressed. E-mail: v.savolainen@imperial.ac.uk.

This article contains supporting information online at www.pnas.org/cgi/content/full/0709936105/DC1.

© 2008 by The National Academy of Sciences of the USA

et al. (26). Alignment of sequences was straightforward, except for *trnH-psbA* that required the addition of several gaps. In orchids and amaryllids, we also found that *trnH-psbA* hosts a well conserved exon, which corresponds to an extra copy of the *rps19* gene (14).

We assessed genetic divergences within and between species, using various metrics (22). We comment here on calculations, using the best-fit models for each barcode (Table 1). For comparison purposes with other studies, we also provide as SI the results based on other distances [supporting information (SI) Tables 6 and 7]. A suitable barcode must exhibit high interspecific but low intraspecific divergence. Here, the highest interspecific divergence is provided by *trnH-psbA* (KNP and combined datasets; Table 1). The next most variable barcode at interspecific level is *matK* for all datasets. Three different metrics were used to characterize intraspecific divergence: (i) average of all pairwise distances between all individuals sampled within those species that had at least two representatives; (ii) “mean theta,” with theta being the average pairwise distances calculated for each species that have more than one representative, thereby eliminating biases associated with uneven sampling among taxa; and (iii) average coalescent depth, i.e., the maximum distance from tips of a node linking all sampled extend members of a species, “book-ending” intraspecific variability (see also SI Table 8). The results from these calculations of intraspecific differences do not show a clear pattern. In orchids, the barcodes exhibiting the lowest intraspecific divergence are *rpoC1* (average mean divergence), *accD/matK* (mean theta) and *matK* (coalescent depth). In the KNP, the lowest intraspecific divergence is provided by *ndhJ* with all three metrics. Wilcoxon signed rank tests on combined data show that *trnH-psbA* is the most variable barcode at interspecific level, followed by *matK* (Table 2). At intraspecific level, Wilcoxon signed rank tests show *rpoC1* and *accD* having the lowest level of divergence, whereas the highest is provided by *trnH-psbA* (Table 3). Based on these results alone, it is difficult to decide on which barcode is the most suited for plants.

Ideally, barcodes must exhibit a “barcoding gap” between inter-versus intraspecific divergences (22). To evaluate whether such a gap is present, we looked at the distribution of divergences in classes of 0.001 distance units (Fig. 1). Median and Wilcoxon two-sample tests were significant in each case, i.e., the distribution and mean of intraspecific differences were lower than that of interspecific divergences, with the highest significances found for *matK* (Wilcoxon two-sample test, $P < 0.0001$), followed by *trnH-psbA* (Wilcoxon two-samples test, $P < 0.0001$; SI Table 9). We did not find, however, any large barcoding gap typical of *cox1* in animals (22), although with *matK* in the Mesoamerican orchids matrix the distributions of intra- versus interspecific divergence are relatively well separated (Fig. 1I).

We evaluated for each barcode whether species are recovered as monophyletic, using phylogenetic techniques and bootstrap resampling. We compared the performance of potential barcodes in recovering species as monophyletic, using maximum parsimony (MP), likelihood, Bayesian, and distance methods. The *trnH-psbA* and *matK* barcodes both recovered the highest value of species monophyly [highest score with unweighted pair group method with arithmetic mean (UPGMA), 90.9%; Table 4]. These two barcodes also recovered the highest percentage of species monophyly with other tree building techniques than UPGMA but with lower percentages (Table 4). When we combined *trnH-psbA* with *matK*, the percentage of species monophyly did not increase notably, except with MP (+7%). Similarly when all barcodes were combined, the percentage of monophyly did not show much increase (93.1% recovered). Combining all potential barcodes did not provide 100% of species monophyly, and for example *Faurea* (Proteaceae), *Ficus glumosa*, and *Ficus abutilifolia* (Moraceae) were always polyphyletic, and the multiple accessions of the palm *Hyphaene coriacea* and orchid *Prosthechea radiata* did not cluster as single species.

Table 1. Measures of inter- and intraspecific divergences for eight potential barcodes sampled in Costa Rica and in the KNP of South Africa

Dataset	Mean	Potential Barcode							
		<i>trnH-psbA</i>	<i>matK</i>	<i>ycf5</i>	<i>rbcl</i>	<i>rpoB</i>	<i>ndhJ</i>	<i>accD</i>	<i>rpoC1</i>
KNP	All interspecific distances	0.0271 ± 0.0258	0.013 ± 0.0126	0.0104 ± 0.0092	0.0082 ± 0.0066	0.0061 ± 0.006	0.0029 ± 0.004	0.0033 ± 0.0037	0.0019 ± 0.0033
	All intraspecific distances	0.0012 ± 0.0021	0.0016 ± 0.0057	0.0002 ± 0.001	0.0005 ± 0.001	0.0007 ± 0.0021	0	0.0005 ± 0.0017	0.0002 ± 0.0009
	Theta	0.0015 ± 0.0033	0.0012 ± 0.0012	0.0005 ± 0.0017	0.0003 ± 0.0006	0.0004 ± 0.0013	0.00003 ± 0.0001	0.0004 ± 0.0013	0.0001 ± 0.0005
Costa Rican	Coalescent depth	0.0024 ± 0.0049	0.0017 ± 0.0017	0.0009 ± 0.0027	0.0004 ± 0.001	0.0007 ± 0.002	0.00009 ± 0.0005	0.0008 ± 0.0025	0.0002 ± 0.0008
	All interspecific distances	0.0082 ± 0.0069	0.0079 ± 0.0086	—	—	0.0067 ± 0.0086	0.0163 ± 0.0211	0.0071 ± 0.0069	0.0022 ± 0.003
	All intraspecific distances	0.0033 ± 0.0034	0.0016 ± 0.0022	—	—	0.0038 ± 0.0096	0.0077 ± 0.0146	0.0017 ± 0.0038	0.0014 ± 0.002
Combined	Theta	0.0024 ± 0.0025	0.001 ± 0.001	—	—	0.0067 ± 0.0133	0.0132 ± 0.0138	0.001 ± 0.0019	0.0015 ± 0.0018
	Coalescent depth	0.0034 ± 0.0037	0.0015 ± 0.0015	—	—	0.0081 ± 0.0146	0.0174 ± 0.02	0.0018 ± 0.0035	0.0021 ± 0.0027
	All interspecific distances	0.0236 ± 0.0246	0.0121 ± 0.0121	0.0104 ± 0.0092	0.0082 ± 0.0066	0.0062 ± 0.0065	0.0046 ± 0.0095	0.0039 ± 0.0047	0.002 ± 0.0033
	All interspecific distances	0.0023 ± 0.0031	0.0016 ± 0.0042	0.0002 ± 0.001	0.0005 ± 0.001	0.0023 ± 0.0071	0.0037 ± 0.0107	0.0012 ± 0.003	0.0008 ± 0.0017
	Theta	0.0018 ± 0.0031	0.001 ± 0.001	0.0005 ± 0.0017	0.0003 ± 0.0007	0.0021 ± 0.0074	0.003 ± 0.0084	0.0006 ± 0.0015	0.0005 ± 0.0012
	Coalescent depth	0.0027 ± 0.0046	0.0014 ± 0.0014	0.0009 ± 0.0027	0.0004 ± 0.001	0.0027 ± 0.0082	0.0039 ± 0.0116	0.0011 ± 0.0028	0.0007 ± 0.0017

Table 2. Wilcoxon signed rank tests of inter-specific divergence among loci

W+	W-	Relative Ranks, <i>n</i> , <i>P</i> value	Result
<i>trnH-psbA</i>	<i>matK</i>	W+ = 21,089, W- = 3,001, <i>n</i> = 219, <i>P</i> ≤ 5.886 × 10 ⁻²²	<i>trnH-psbA</i> ≫ <i>matK</i>
<i>trnH-psbA</i>	<i>ycf5</i>	W+ = 13,226, W- = 802, <i>n</i> = 167, <i>P</i> ≤ 3.276 × 10 ⁻²³	<i>trnH-psbA</i> ≫ <i>ycf5</i>
<i>trnH-psbA</i>	<i>rbcl</i>	W+ = 15,878, W- = 1,327, <i>n</i> = 185, <i>P</i> ≤ 2.001 × 10 ⁻²³	<i>trnH-psbA</i> ≫ <i>rbcl</i>
<i>trnH-psbA</i>	<i>rpoB</i>	W+ = 23,403, W- = 2,022, <i>n</i> = 225, <i>P</i> ≤ 7.967 × 10 ⁻²⁸	<i>trnH-psbA</i> ≫ <i>rpoB</i>
<i>trnH-psbA</i>	<i>ndhJ</i>	W+ = 20,363, W- = 2,642, <i>n</i> = 214, <i>P</i> ≤ 1.546 × 10 ⁻²²	<i>trnH-psbA</i> ≫ <i>ndhJ</i>
<i>trnH-psbA</i>	<i>rpoc1</i>	W+ = 23,709, W- = 162, <i>n</i> = 218, <i>P</i> ≤ 1.55 × 10 ⁻³⁶	<i>trnH-psbA</i> ≫ <i>rpoc1</i>
<i>trnH-psbA</i>	<i>accD</i>	W+ = 23,669, W- = 1,756, <i>n</i> = 225, <i>P</i> ≤ 3.828 × 10 ⁻²⁹	<i>trnH-psbA</i> ≫ <i>accD</i>
<i>matK</i>	<i>ycf5</i>	W+ = 6,833, W- = 6,862, <i>n</i> = 165, <i>P</i> ≤ 0.9818	<i>matK</i> = <i>ycf5</i>
<i>matK</i>	<i>rbcl</i>	W+ = 12,312, W- = 4,893, <i>n</i> = 185, <i>P</i> ≤ 3.673 × 10 ⁻⁷	<i>matK</i> > <i>rbcl</i>
<i>matK</i>	<i>rpoB</i>	W+ = 21,020, W- = 3,290, <i>n</i> = 220, <i>P</i> ≤ 6.803 × 10 ⁻²¹	<i>matK</i> > <i>rpoB</i>
<i>matK</i>	<i>ndhJ</i>	W+ = 19,554, W- = 2,812, <i>n</i> = 211, <i>P</i> ≤ 4.287 × 10 ⁻²¹	<i>matK</i> > <i>ndhJ</i>
<i>matK</i>	<i>rpoc1</i>	W+ = 24,054, W- = 477, <i>n</i> = 221, <i>P</i> ≤ 3.17 × 10 ⁻³⁵	<i>matK</i> > <i>rpoc1</i>
<i>matK</i>	<i>accD</i>	W+ = 22,666, W- = 2,087, <i>n</i> = 222, <i>P</i> ≤ 6.824 × 10 ⁻²⁷	<i>matK</i> > <i>accD</i>
<i>rbcl</i>	<i>ycf5</i>	W+ = 4,564, W- = 10,487, <i>n</i> = 173, <i>P</i> ≤ 7.186 × 10 ⁻⁶	<i>rbcl</i> < <i>ycf5</i>
<i>rbcl</i>	<i>rpoB</i>	W+ = 11,985, W- = 5,220, <i>n</i> = 185, <i>P</i> ≤ 3.536 × 10 ⁻⁶	<i>rbcl</i> > <i>rpoB</i>
<i>rbcl</i>	<i>ndhJ</i>	W+ = 14,475, W- = 576, <i>n</i> = 173, <i>P</i> ≤ 6.202 × 10 ⁻²⁶	<i>rbcl</i> > <i>ndhJ</i>
<i>rbcl</i>	<i>rpoc1</i>	W+ = 14,908, W- = 143, <i>n</i> = 173, <i>P</i> ≤ 4.702 × 10 ⁻²⁹	<i>rbcl</i> > <i>rpoc1</i>
<i>rbcl</i>	<i>accD</i>	W+ = 15,215, W- = 1,438, <i>n</i> = 182, <i>P</i> ≤ 3.803 × 10 ⁻²²	<i>rbcl</i> > <i>accD</i>
<i>ycf5</i>	<i>rpoB</i>	W+ = 10,796, W- = 2,899, <i>n</i> = 165, <i>P</i> ≤ 1.338 × 10 ⁻¹⁰	<i>ycf5</i> > <i>rpoB</i>
<i>ycf5</i>	<i>ndhJ</i>	W+ = 11,259, W- = 987, <i>n</i> = 156, <i>P</i> ≤ 1.037 × 10 ⁻¹⁹	<i>ycf5</i> > <i>ndhJ</i>
<i>ycf5</i>	<i>rpoc1</i>	W+ = 11,952, W- = 294, <i>n</i> = 156, <i>P</i> ≤ 6.297 × 10 ⁻²⁵	<i>ycf5</i> > <i>rpoc1</i>
<i>ycf5</i>	<i>accD</i>	W+ = 10,755, W- = 1,026, <i>n</i> = 153, <i>P</i> ≤ 8.128 × 10 ⁻¹⁹	<i>ycf5</i> > <i>accD</i>
<i>rpoB</i>	<i>ndhJ</i>	W+ = 11,709, W- = 6,057, <i>n</i> = 188, <i>P</i> ≤ 0.0001556	<i>rpoB</i> > <i>ndhJ</i>
<i>rpoB</i>	<i>rpoc1</i>	W+ = 16,047, W- = 3,456, <i>n</i> = 197, <i>P</i> ≤ 3.984 × 10 ⁻¹⁵	<i>rpoB</i> > <i>rpoc1</i>
<i>rpoB</i>	<i>accD</i>	W+ = 11,614, W- = 5,777, <i>n</i> = 186, <i>P</i> ≤ 7.227 × 10 ⁻⁵	<i>rpoB</i> > <i>accD</i>
<i>ndhJ</i>	<i>rpoc1</i>	W+ = 11,859, W- = 5,346, <i>n</i> = 185, <i>P</i> ≤ 8.037 × 10 ⁻⁶	<i>ndhJ</i> > <i>rpoc1</i>
<i>ndhJ</i>	<i>accD</i>	W+ = 7,469, W- = 6,392, <i>n</i> = 166, <i>P</i> ≤ 0.3857	<i>ndhJ</i> = <i>accD</i>
<i>rpoc1</i>	<i>accD</i>	W+ = 3,891, W- = 14,064, <i>n</i> = 189, <i>P</i> ≤ 1.447 × 10 ⁻¹¹	<i>rpoc1</i> < <i>accD</i>

Finally, we used coalescence analyses to compare the branching patterns along trees and identify distinct genetic clusters (24). The highest number of independent clusters was found by using UPGMA with *matK* (SI Fig. 2), followed by *rpoB* and *trnH-psbA* (Table 5). With *matK*, 41 clusters were identified, of which 30 fully correspond to previously recognized taxonomic species, 4 partially matched taxonomic species (i.e., failed to group all representatives

into a single cluster), and 7 mixed species together (Table 5 and SI Fig. 2). With *rpoB*, 36 clusters were identified, of which 20 fully correspond to taxonomic species; whereas with *trnH-psbA*, 34 clusters were identified, which a slightly higher proportion corresponding to previously recognized species (i.e., 19 clusters).

Altogether, our results indicate that either *matK* or *trnH-psbA* are the most suitable regions for plant DNA barcoding. In this sense,

Table 3. Wilcoxon signed rank tests of intraspecific difference among loci

W+	W-	Relative Ranks, <i>n</i> , <i>P</i> value	Result
<i>trnH-psbA</i>	<i>matK</i>	W+ = 1,949, W- = 826, <i>n</i> = 74, <i>P</i> ≤ 0.002509	<i>trnH-psbA</i> > <i>matK</i>
<i>trnH-psbA</i>	<i>ycf5</i>	W+ = 327, W- = 108, <i>n</i> = 29, <i>P</i> ≤ 0.01843	<i>trnH-psbA</i> > <i>ycf5</i>
<i>trnH-psbA</i>	<i>rbcl</i>	W+ = 436, W- = 92, <i>n</i> = 32, <i>P</i> ≤ 0.001342	<i>trnH-psbA</i> > <i>rbcl</i>
<i>trnH-psbA</i>	<i>rpoB</i>	W+ = 1,113, W- = 483, <i>n</i> = 56, <i>P</i> ≤ 0.01031	<i>trnH-psbA</i> > <i>rpoB</i>
<i>trnH-psbA</i>	<i>ndhJ</i>	W+ = 973, W- = 567, <i>n</i> = 55, <i>P</i> ≤ 0.08976	<i>trnH-psbA</i> ≅ <i>ndhJ</i>
<i>trnH-psbA</i>	<i>rpoc1</i>	W+ = 1,596, W- = 234, <i>n</i> = 60, <i>P</i> ≤ 5.464 × 10 ⁻⁷	<i>trnH-psbA</i> > <i>rpoc1</i>
<i>trnH-psbA</i>	<i>accD</i>	W+ = 1,579, W- = 437, <i>n</i> = 63, <i>P</i> ≤ 9.399 × 10 ⁻⁵	<i>trnH-psbA</i> > <i>accD</i>
<i>matK</i>	<i>ycf5</i>	W+ = 260, W- = 175, <i>n</i> = 29, <i>P</i> ≤ 0.3638	<i>matK</i> = <i>ycf5</i>
<i>matK</i>	<i>rbcl</i>	W+ = 299, W- = 197, <i>n</i> = 31, <i>P</i> ≤ 0.3224	<i>matK</i> = <i>rbcl</i>
<i>matK</i>	<i>rpoB</i>	W+ = 695, W- = 790, <i>n</i> = 54, <i>P</i> ≤ 0.6857	<i>matK</i> = <i>rpoB</i>
<i>matK</i>	<i>ndhJ</i>	W+ = 585, W- = 640, <i>n</i> = 49, <i>P</i> ≤ 0.7883	<i>matK</i> = <i>ndhJ</i>
<i>matK</i>	<i>rpoc1</i>	W+ = 1,220, W- = 491, <i>n</i> = 58, <i>P</i> ≤ 0.004829	<i>matK</i> > <i>rpoc1</i>
<i>matK</i>	<i>accD</i>	W+ = 1,059, W- = 594, <i>n</i> = 57, <i>P</i> ≤ 0.06529	<i>matK</i> > <i>accD</i>
<i>rbcl</i>	<i>ycf5</i>	W+ = 66, W- = 124, <i>n</i> = 19, <i>P</i> ≤ 0.2579	<i>rbcl</i> = <i>ycf5</i>
<i>rbcl</i>	<i>rpoB</i>	W+ = 104, W- = 127, <i>n</i> = 21, <i>P</i> ≤ 0.7022	<i>rbcl</i> = <i>rpoB</i>
<i>rbcl</i>	<i>ndhJ</i>	W+ = 96, W- = 9, <i>n</i> = 14, <i>P</i> ≤ 0.004028	<i>rbcl</i> > <i>ndhJ</i>
<i>rbcl</i>	<i>rpoc1</i>	W+ = 98, W- = 22, <i>n</i> = 15, <i>P</i> ≤ 0.03015	<i>rbcl</i> > <i>rpoc1</i>
<i>rbcl</i>	<i>accD</i>	W+ = 66, W- = 105, <i>n</i> = 18, <i>P</i> ≤ 0.4171	<i>rbcl</i> = <i>accD</i>
<i>ycf5</i>	<i>rpoB</i>	W+ = 94, W- = 96, <i>n</i> = 19, <i>P</i> ≤ 0.9843	<i>ycf5</i> = <i>rpoB</i>
<i>ycf5</i>	<i>ndhJ</i>	W+ = 44, W- = 1, <i>n</i> = 9, <i>P</i> ≤ 0.007812	<i>ycf5</i> > <i>ndhJ</i>
<i>ycf5</i>	<i>rpoc1</i>	W+ = 68, W- = 10, <i>n</i> = 12, <i>P</i> ≤ 0.021	<i>ycf5</i> > <i>rpoc1</i>
<i>ycf5</i>	<i>accD</i>	W+ = 46, W- = 59, <i>n</i> = 14, <i>P</i> ≤ 0.7148	<i>ycf5</i> = <i>accD</i>
<i>rpoB</i>	<i>ndhJ</i>	W+ = 297, W- = 406, <i>n</i> = 37, <i>P</i> ≤ 0.4153	<i>rpoB</i> = <i>ndhJ</i>
<i>rpoB</i>	<i>rpoc1</i>	W+ = 496, W- = 207, <i>n</i> = 37, <i>P</i> ≤ 0.02982	<i>rpoB</i> > <i>rpoc1</i>
<i>rpoB</i>	<i>accD</i>	W+ = 465, W- = 438, <i>n</i> = 42, <i>P</i> ≤ 0.8709	<i>rpoB</i> = <i>accD</i>
<i>ndhJ</i>	<i>rpoc1</i>	W+ = 243, W- = 82, <i>n</i> = 25, <i>P</i> ≤ 0.03135	<i>ndhJ</i> > <i>rpoc1</i>
<i>ndhJ</i>	<i>accD</i>	W+ = 322, W- = 174, <i>n</i> = 31, <i>P</i> ≤ 0.1498	<i>ndhJ</i> = <i>accD</i>
<i>rpoc1</i>	<i>accD</i>	W+ = 276, W- = 427, <i>n</i> = 37, <i>P</i> ≤ 0.2579	<i>rpoc1</i> = <i>accD</i>

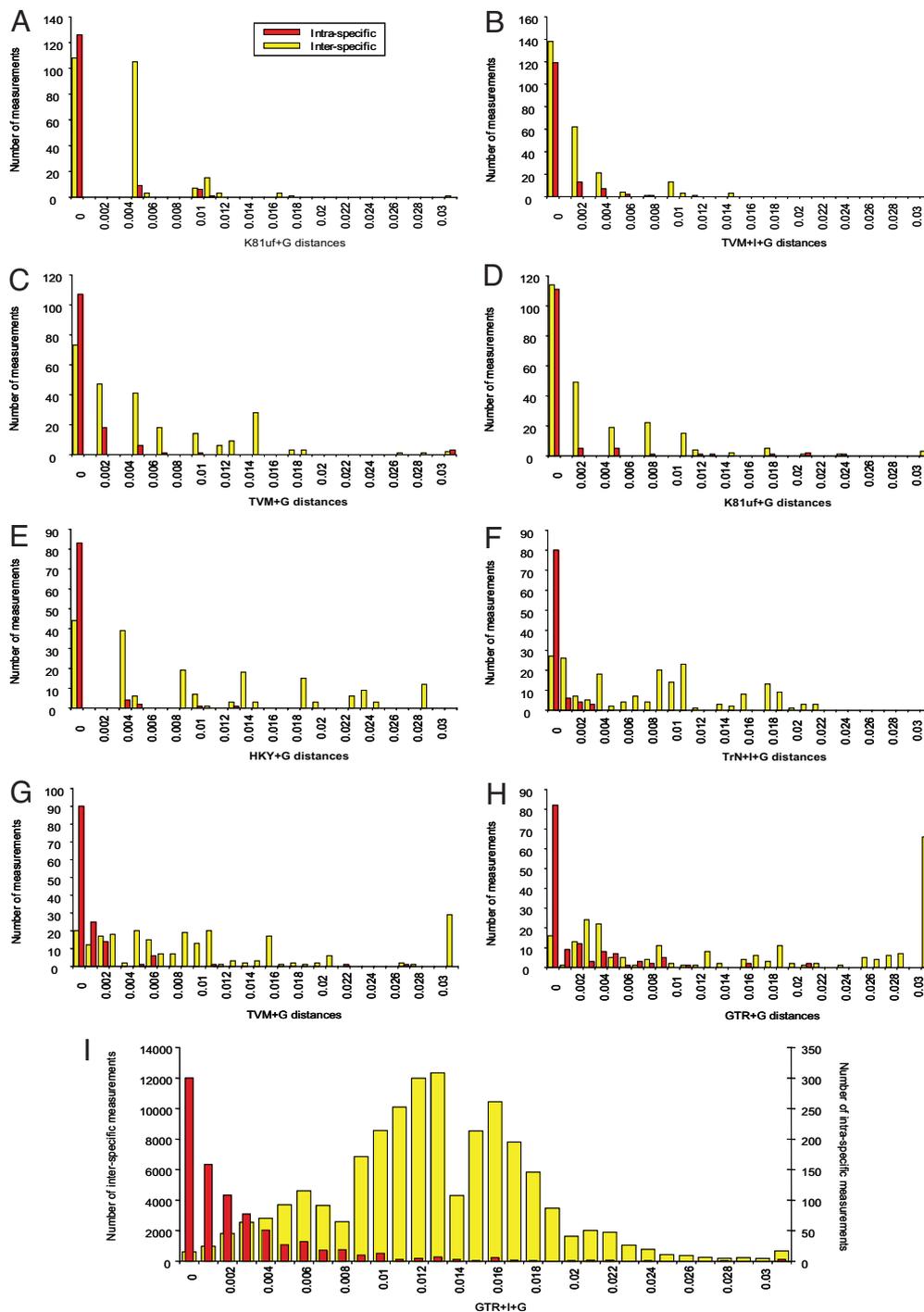


Fig. 1. Relative distribution of inter-specific divergence between congeneric species (yellow) and intra-specific distances (with best fit-model; red) for eight loci. (A) *accD*. (B) *rpoC1*. (C) *rpoB*. (D) *ndhJ*. (E) *ycf5*. (F) *rbcl*. (G) *matK* reduced matrix. (H) *trnH-psbA*. (I) *matK* expanded Mesoamerican orchids matrix. x axis, increments of 0.001; y axis, number of occurrences. Barcoding gaps were assessed with Median tests and Wilcoxon two-samples tests, and all were highly significant ($P < 0.0001$).

we agree with previous relatively small-scale studies that focused on the nutmeg family [Myristicaceae (27)] or the 50-acre forest of the New York Botanical Garden (28). Because several *matK* sequences were already available in GenBank for orchids, we expanded our Costa Rican sampling and compiled a large matrix for Mesoamerican species, assembling 1,566 DNA sequences. Coalescence analyses from the UPGMA tree identified 212 genetic clusters, of which 86 fully matched previously recognized species and a further 25 partially matched taxonomic species (SI Fig. 3). An examination of these clusters reveals cryptic species, which need further taxonomic work. For example, we sequenced four accessions of *Lycaste tricolor* (Klotzsch.) Rehb. (numbered 841, 840, 838, and 1011 in SI Table 10). *Lycaste cf. tricolor* 1011 does not cluster with the other three

accessions and taxonomists had indeed suspected it could be another, separate, species. *Lycaste cf. tricolor* 1011 grows on the Pacific slopes of Costa Rica, whereas the other three representatives (i.e., the “typical” *L. tricolor*) grow on the Atlantic side. There are also discrete morphological differences. The pollinarium of *Lycaste cf. tricolor* 1011, like all other representatives of that species in the Pacific slope, have long stipes, a twisted column and a hairy anther cap, whereas the typical *tricolor* have short stipes, a straight column and a smooth anther cap (SI Fig. 4). These differences in column are probably also involved in reproductive isolation whereby *L. cf. tricolor* 1011 would deposit pollinia on the shoulder of the pollinating bees and the “true” *tricolor*, with their straight column, would deposit pollinia on the back of the bees.

Table 4. Proportion (%) of monophyletic species recovered with different phylogenetic techniques and loci

Dataset	UPGMA	NJ	MP	ML	BI
<i>accD</i>	56.8 (36.3)	45.4 (29.5)	29.5 (27.2)	31.8 (29.5)	29.5 (29.5)
<i>rpoc1</i>	63.6 (38.6)	40.9 (27.2)	34 (29.5)	34 (29.5)	31.8 (34)
<i>ndhJ</i>	63.4 (39)	51.2 (36.5)	39 (26.8)	34.1 (26.8)	34.1 (29.2)
<i>ycf5</i>	80 (66.6)	60 (43.3)	50 (46.6)	53.3 (46.6)	53.3 (53.3)
<i>rpoB</i>	72.7 (56.8)	61.3 (50)	54.5 (50)	59 (50)	56.8 (54.5)
<i>rbcl</i>	87.5 (78.1)	65.6 (75)	68.7 (68.7)	71.8 (68.7)	71.8 (71.8)
<i>trnH-psbA</i>	90.6 (65.1)	53.4 (32.5)	76.7 (62.7)	72 (60.4)	69.7 (69.7)
<i>matK</i>	90.6 (76.7)	79 (76.7)	79 (79)	79 (79)	79 (79)
<i>matK + trnH-psbA</i>	90.9 (86.3)	79.5 (70.4)	86.3 (81.8)	81.8 (75)	81.3 (79.5)
All barcodes combined	93.1 (84)	72.7 (77.2)	88.6 (86.3)	88.6 (86.3)	88.6 (88.6)

Proportions supported by posterior probabilities or bootstrap >50% are in brackets.

Our sampling is more comprehensive than previous studies on DNA barcoding in plants. Kress *et al.* (13) used 19 species with duplicates/triplicates and a further 83 species with only one representative per species. Kress and Erickson (16) used 48 pairs of species, each represented by one sample. Cowan *et al.* (29) and Chase *et al.* (15) report that the plant working group started with 96 pairs of taxa but narrowed it down to fewer species. Cameron used 343 species from within the botanical garden in New York (28). We used here 86 species in which all barcodes were tested and a further 1,036 orchid species in the dataset restricted to *matK*. Because the assessment of intraspecific variability is crucial for

deciding on a suitable barcode, we included 44 species in which there were at least two and up to seven representatives per species. Our results are robust and all point toward the same pair of loci. Given that the second half (5' end) of the *matK* exon is easy to amplify and align, we propose that *matK* is used as a preferred universal DNA barcode for flowering plants. The *trnH-psbA* region performs nearly equally well, although its pattern of molecular evolution is complex. Therefore, we propose that *trnH-psbA* is used as either an alternative to *matK* or a complementary barcode to *matK*. When combined, these loci achieve only moderate improvement, as shown by our analyses of recovering species monophyly.

The use of *matK* as a barcode has been criticized mainly because no universal primers were available (15), hence it had the lowest amplification success in Kress and Erickson (16). However, we found that primers 390F and 1326R from Cuénoud *et al.* (26) amplify the same region with a 100% success. The use of *trnH-psbA* has been criticized because of the difficulty in the alignment due to extensive length variation and because certain species host a pseudogene (15). Although in certain cases *trnH-psbA* might indeed be problematic, we found here that it was one of the most useful regions across a wide range of angiosperms.

Using *matK* alone or in combination with *trnH-psbA*, our tests of monophyly reach >90% of correct species identification. If our sampling was restricted to sister species rather than natural geographic assemblages of species, we may have found this value to drop. However, our samples do include very closely related species, given that Costa Rica and southern Africa both have experienced extensive rapid radiations (30, 31).

Apart from combining *matK* and *trnH-psbA*, we found that adding the other barcodes did not improve species identification by >3% and therefore was not worth pursuing if one balances gains in identification versus sequencing efforts. It is possible that some regions yet untested here may be useful as a complementary barcode, and we await further studies. Alternatively, we may need to accept that no more than ≈90% of species will be identified with universal plastid barcodes and that those difficult lineages will need "case-by-case" analyses, using, for example, nuclear population genetic markers and taking advantage of recent developments in DNA sequencing technology (32).

Our results differ from the proposal of Kress and Erickson (16) in the sense that we advocate *matK* rather than *rbcl*, although we agree with the utility of *trnH-psbA*. As explained above, the amplification of *matK* is not problematic, as Kress and Erickson thought before, and the pattern of variation in its second half (5' end) is particularly appropriate for its use as a DNA barcode, as exemplified by our large-scale analysis in orchids. The *matK* gene also presents another advantage: its first half (3' end) was useful to reconstruct the phylogeny of angiosperms (33), and therefore the complete sequence of this gene can be used as dual barcode-phylogenetic marker. The *matK* gene has an unusual mode and tempo of evolution; it is the only putative chloroplast-encoded group II intron maturase, and its function relates to the regulation

Table 5. Coalescence analyses indicating the number of independent genetic clusters and their correspondence with taxonomically recognized species

	Dataset	No. of Genetic Clusters	Full match	Partial match	No match	
UPGMA	Combined	31	23	3	5	
	<i>matK + trnHpsbA</i>	33	12	19	2	
	<i>accD</i>	33	17	3	13	
	<i>matK</i>	41	30	4	7	
	<i>ndhJ</i>	28	15	3	10	
	<i>rbcl</i>	28	20	5	3	
	<i>rpoB</i>	36	20	5	11	
	<i>rpoC1</i>	29	13	3	13	
	<i>trnH-psbA</i>	34	19	12	3	
	<i>ycf5</i>	16	7	0	9	
	MP branch lengths plus NPRS	Combined	8	3	4	1
		<i>matK + trnHpsbA</i>	11	7	3	1
		<i>accD</i>	20	13	3	4
<i>matK</i>		20	16	1	3	
<i>ndhJ</i>		20	16	1	3	
<i>rbcl</i>		21	17	3	1	
<i>rpoB</i>		17	14	0	3	
<i>rpoC1</i>		16	10	0	6	
<i>trnH-psbA</i>		16	13	3	0	
<i>ycf5</i>		15	10	3	2	
ML plus NPRS		Combined	32	26	2	4
		<i>matK + trnHpsbA</i>	11	9	1	1
		<i>accD</i>	20	13	3	4
	<i>matK</i>	23	17	2	4	
	<i>ndhJ</i>	16	9	2	5	
	<i>rbcl</i>	20	16	3	1	
	<i>rpoB</i>	19	14	1	4	
	<i>rpoC1</i>	18	10	1	7	
	<i>trnH-psbA</i>	16	15	1	0	
	<i>ycf5</i>	14	9	3	2	

of plant development. Analyses of the expression of this gene suggested that “genetic buffers” are in operation and constrain its evolution, which may explain why relatively low intraspecific but high interspecific variation is found and therefore why it fits DNA barcoding purposes so well. We disagree with Kondo *et al.* (34), who argued that *matK* on its own was not useful for species identification, but their study focused exclusively on species of liquorices in the legume family. We also disagree with the proposal of Chase *et al.* (15), because we found that neither *rpoC1* nor *rpoB* were performing well as a barcode (Tables 1–5). These two loci amplify easily in non-angiosperms (15), but we found that they were too conserved in angiosperms. It might in fact not be so important to design primer pairs or barcodes that work universally from ferns, mosses, to seed plants. Several of the DNA barcoding applications (e.g., rapid inventories for conservation) may not need to identify non-seed plants at the species level, and alternatively if this was required then moss- and fern-specific primers or barcodes could be used in complement to seed plant barcodes. In the meantime, we propose that DNA barcoding with *matK* is used on a large scale.

DNA barcoding with *matK* alone (or *matK* plus *trnH-psbA* combined) has the potential to speed up the exploration and preservation of plant life on Earth by facilitating considerably biodiversity inventories beyond South Africa and Costa Rica. In addition, new methods are now being developed in which DNA barcoding data can be used in conservation (35). As an example, we illustrate how customs officers could use DNA barcoding to identify plant fragments from species in which trade is controlled by the Convention on International Trade of Endangered Species (CITES). All orchids are in Appendix 2 of CITES [i.e., a special permit is required for their trade (www.cites.org)], but a few species, such as the lady’s slipper orchids in Mesoamerica (genus *Phragmipedium*), are so threatened in the wild that their trade is prohibited altogether (i.e., they are listed in Appendix 1 of CITES). We included in our large *matK* matrix one sequence of *Phragmipedium* as a reference and ran a UPGMA analysis with all 1,500+ orchids with 10 additional *Phragmipedium* sequences representing another seven species (GenBank accession nos. AY918826–31, AJ581442, AY557204). All species of *Phragmipedium* clustered together correctly. This means that in our theoretical case, using our proposed DNA barcode, the custom services would have positively identified species from CITES Appendix 1 (i.e., the lady’s slipper orchids) from species in Appendix 2 (i.e., the other orchids) and those not listed by CITES (here, the species from the KNP).

To ensure even longer-term benefit of the DNA barcoding efforts, it is also essential to put in place DNA banking strategies (36) so that complementary barcodes to the ones identified here can be produced in the future. More importantly, if DNA barcoding is to achieve its goals, it must urgently become available to countries rich in biodiversity but poor in resources through efficient capacity building and judicious funding programs.

Methods

Sampling. In total, 1,667 taxa were sampled (SI Table 10). In the KNP, we collected 101 specimens of trees, shrubs, and achlorophyllous parasites, including 32 species in which we have more than one representative per species. The first dataset of Costa Rican orchids comprises 71 specimens representing 48 species in which 12 have more than one representative per species. A second orchid dataset was assembled with *matK* only, but with a much increased taxon sampling with a total of 1,566 specimens representing 1,084 species from Mesoamerica in which 295 have at least two representatives.

DNA Sequencing. Total DNA was extracted by using the method of Doyle and Doyle (37). We amplified and sequenced *accD*, *rpoC1*, *rpoB*, *ndhJ*, *matK*, and *ycf5*, following guidelines from the plant working group. For *matK*, additional primers 390F and 1326R (26) were used. Primers *trnHf* and *psbA3’f* were used for *trnH-psbA* (13). For the first half of the *rbcl* exon, primers 1F and 724R were used following Kress *et al.* (13). DNA sequences were aligned in PAUP4b10 (38).

Genetic and Phylogenetic Analyses. Inter- and intraspecific genetic divergences were calculated following Meyer and Paulay (22). Pairwise distances were calculated with PAUP4b10 (38) and the best-fitting model as given by applying MODELTEST 3.7 (39). Wilcoxon signed rank tests were performed to compare intra- and interspecific variability for every pair of barcodes following Kress and Erickson (16). We evaluated DNA barcoding gaps by comparing the distribution of intra- versus interspecific divergences (22). To evaluate whether species were recovered as monophyletic with each barcode, we used standard phylogenetic techniques: MP, maximum likelihood (ML), neighbor joining (NJ), and UPGMA with PAUP4b10 (38). Bayesian statistical inferences (BI) were performed with MrBayes software, Version 3.1.2 (40). The parsimony analysis of the large *matK* matrix of Mesoamerican orchids was performed by using the parsimony ratchet method (41). We identified genetic clusters by coalescence analyses, using methods developed by Pons *et al.* (23) and Fontaneto *et al.* (24). Details are available from the corresponding author upon request.

ACKNOWLEDGMENTS. We thank the KNP, South African National Parks, H. Eckhardt, I. Smit, G. Zambatis, T. Khoza, Ministerio de Ambiente y Energía, and Sistema Nacional de Areas de Conservación, for granting access to the park and sharing data; M. Chase, R. Cowan, M. Powell, H. van Niekerk, two anonymous reviewers, and the editor for comments; and T. Rikombe, R. Bryden, T. Mhlongo, H. van der Bank for fieldwork. This work was supported by the South African National Research Foundation, the University of Johannesburg, the United Kingdom Darwin Initiative, The Royal Society (U.K.), and the European Commission (HOTSPOTS Consortium).

- Ebach MC, Holdrege C (2005) *Nature* 434:697.
- Will KW, Mishler BD, Wheeler QD (2005) *Syst Biol* 54:844–851.
- Blaxter ML (2004) *Philos Trans R Soc London Ser B* 359:669–679.
- Hajibabaei M, de Waard JR, Ivanova NV, Ratnasingham S, Dooph RT, Kirk SL, Mackie PM, Hebert PDN (2005) *Philos Trans R Soc London Ser B* 360:1959–1967.
- Janzen DH (2004) *Philos Trans R Soc London Ser B* 359:731–732.
- Janzen DH, Hajibabaei M, Burns JM, Hallwachs W, Remigio E, Hebert PDN (2005) *Philos Trans R Soc London Ser B* 360:1835–1845.
- Hebert PDN, Cywinska A, Ball SL, De Waard JR (2003) *Proc R Soc Biol Sci Ser B* 270:313–321.
- Smith MA, Fisher BL, Hebert PDN (2005) *Philos Trans R Soc London Ser B* 360:1825–1834.
- Vences M, Thomas M, Bonnett RM, Vieites DR (2005) *Philos Trans R Soc London Ser B* 360:1859–1868.
- Ward RD, Zemlak TS, Innes BH, Last PR, Hebert PDN (2005) *Philos Trans R Soc London Ser B* 360:1847–1857.
- Rubioff D, Cameron S, Will K (2006) *Trends Ecol Evol* 21:1–2.
- Pennisi E (2007) *Science* 318:190–191.
- Kress JW, Wurdack KJ, Zimmer EA, Weigt LA, Janzen DH (2005) *Proc Natl Acad Sci USA* 102:8369–8374.
- Chang CC, Lin HC, Lin IP, Chow TY, Chen HH, Chen WH, Cheng CH, Lin CY, Liu SM, Chang CC, *et al.* (2006) *Mol Biol Evol* 23:279–291.
- Chase MW, Cowan RS, Hollingsworth PM, van den Berg C, Madrinan S, Petersen G, Seberg O, Jorgensen T, Cameron KM, Carine M, *et al.* (2007) *Taxon* 56:295–299.
- Kress WJ, Erickson DL (2007) *PLoS One* 2:e508.
- Newmaster SG, Fazekas AJ, Ragupathy S (2006) *Can J Bot* 84:335–341.
- Taberlet P, Coissac E, Pompanon F, Gielly L, Miquel C, Valentini A, Vermet T, Corthier G, Brochmann C, Willerslev E (2007) *Nucleic Acids Res* 35:e1–e8.
- Myers N, Mittermeier RA, Mittermeier CG, da Fonseca GAB, Kent J (2000) *Nature* 403:853–858.
- Myers N (2003) *Bioscience* 53:916–917.
- van der Schijff HP (1969) *Publikasies van die Universiteit van Pretoria, Nuwe reeks* 53:1–100.
- Meyer CP, Paulay G (2005) *PLoS Biol* 3:2229–2238.
- Pons J, Barraclough TG, Gomez-Zurita J, Cardoso A, Duran DP, Hazell S, Kamoun S, Suhlman WD, Vogler AP (2006) *Syst Biol* 55:595–609.
- Fontaneto D, Herniou EA, Boschetti C, Caprioli M, Melone G, Ricci C, Barraclough TG (2007) *PLoS Biol* 5:914–921.
- Cameron KM, Chase MW, Whitten WM, Kores PJ, Jarrell DC, Albert VA, Yukawa T, Hills HG, Goldman DH (1999) *Am J Bot* 86:208–224.
- Cuénoud P, Savolainen V, Chatrou LW, Powell M, Grayer RJ, Chase MW (2002) *Am J Bot* 89:132–144.
- Newmaster SG, Fazekas AJ, Steeves RAD, Janovec J (2008) *Mol Ecol Notes*, 10.1111/j.1471-8286.2007.02002.x.
- Cameron K (2007) in *Botany and Plant Biology 2007 Joint Congress*, ed The Botanical Society of America (The Botanical Society of America, Chicago).
- Cowan RS, Chase MW, Kress JW, Savolainen V (2006) *Taxon* 55:611–616.
- Linder HP (2003) *Biol Rev* 78:597–638.
- Gravendeel B, Smithson A, Slik FJW, Schuitman A (2004) *Philos Trans R Soc London Ser B* 359:1523–1535.
- Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen ZT, *et al.* (2005) *Nature* 437:376–380.
- Hilu K, Borsch T, Müller K, Soltis DE, Soltis PS, Savolainen V, Chase MW, Alice L, Evans R, Sauquet H, *et al.* (2003) *Am J Bot* 90:1758–1776.
- Kondo K, Shiba M, Yamaji H, Morota T, Zhengmin C, Huixia P, Shoyama Y (2007) *Biol Pharm Bull* 30:1497–1502.
- Faith DP, Baker A (2006) *Evol Bioinform Online* 2:70–77.
- Savolainen V, Reeves G (2004) *Science* 304:1445.
- Doyle JJ, Doyle JL (1987) *Phytochem Bull* 19:11–15.
- Swofford DL (2001) *PAUP* 4.0: Phylogenetic Analysis Using Parsimony (* and other methods)* (Sinauer Associates, Sunderland, MA).
- Posada D (2006) *Nucleic Acids Res* 34:W700–W703.
- Ronquist F, Huelsenbeck JP (2003) *Bioinformatics (Oxford)* 19:1572–1574.
- Nixon KC (1999) *Cladistics* 15:407–414.