Comparative analysis of a large dataset indicates that internal transcribed spacer (ITS) should be incorporated into the core barcode for seed plants

China Plant BOL Group¹, De-Zhu Li², Lian-Ming Gao³, Hong-Tao Li³, Hong Wang⁴, Xue-Jun Ge⁵, Jian-Quan Liu⁶, Zhi-Duan Chen⁷, Shi-Liang Zhou⁸, Shi-Lin Chen⁹, Jun-Bo Yang⁸, Cheng-Xin Fu⁸, Chun-Xia Zeng¹⁰, Hai-Fei Yan¹¹, Yong-Shuai Sun¹¹, Si-Yun Chen¹¹, Lei Zhao¹¹, Kun Wang¹¹, Tuo Yan¹¹, and Guang-Wen Du¹¹

A two-marker combination of plastid rbcL and matK has previously been recommended as the core plant barcode, to be supplemented with additional markers such as plastid trnH-psbA and nuclear ribosomal internal transcribed spacer (ITS). To assess the effectiveness and universality of these barcode markers in seed plants, we sampled 6,286 individuals representing 1,757 species in 141 genera of 75 families (42 orders) by using four different methods of data analysis. These analyses indicate that (i) the three plastid markers showed high levels of universality (87.1–92.7%), whereas ITS performed relatively well (79%) in angiosperms but not so well in gymnosperms; (ii) in taxonomic groups for which direct sequencing of the marker is possible, ITS showed the highest discriminatory power of the four markers, and a combination of ITS and any plastid DNA marker was able to discriminate 69.9–79.1% of species, compared with only 49.7% with rbcL + matK; and (iii) where multiple individuals of a single species were tested, ascensions based on ITS and plastid DNA barcodes were incongruent in some samples for 45.2% of the sampled genera (for genera with more than one species sampled). This finding highlights the importance of both sampling multiple individuals and using markers with different modes of inheritance. In cases where it is difficult to amplify and directly sequence ITS in its entirety, just using ITS2 is a useful backup because it is easier to amplify and sequence this subset of the marker. We therefore propose that ITS/ITS2 should be incorporated into the core barcode for seed plants.

After a joint international effort, the two-marker combination of rbcL + matK was proposed as the core barcode for land plants in August 2009 (12). However, this recommendation was based on the study of only a relatively small number of species in which multiple individuals were sampled from multiple congeneric species. Subsequent to this study, internal transcribed spacer 2 (ITS2) was also suggested as a novel barcode for both plants and animals (13, 14). At the Third International Barcoding of Life Conference in Mexico City in November 2009, it was stressed that complementary markers to the proposed core barcode of rbcL and matK should continue to be assessed from both the plastid genome (e.g., trnH-psbA) and the nuclear genome (e.g., ribosomal DNA ITS or ITS2). The CBOL Plant Working Group urged the international plant barcoding community to make an effort to further evaluate these plant barcodes within 18 mo and ultimately to standardize a DNA barcode for plants (15).

As a response to this call, a coordinated effort was made among research groups in China. China is a megadiverse country with 28,600 species (in ~3,200 genera) of seed plants and contains 4 of the 34 recognized global biodiversity hotspots: the mountains of Central Asia, the Himalayas, the Indo-Myanmar region, and the mountains of Southwest China (16, 17). China is also the center of distribution of many endemic-rich temperate genera, such as Pedicularis, Primula, and Rhododendron, and is the location of a unique evergreen broadleaf forest ecosystem dominated by subtropical species of Fagaceae, Lauraceae, Magnoliaceae, and Theaceae (18). Thus, a coordinated plant DNA-barcoding effort in China is of great significance in a global context.

The project involved 46 research groups from 17 research institutes and universities in China, all with longstanding experience in taxonomy and extensive collections of plant material. In total, 6,286 individuals were sampled, representing 1,757 species in 141 genera of 75 families (42 orders) of seed plants, mainly from China. All selected species could unambiguously be identified to species based on morphology and geography. We amplified and sequenced four DNA-barcoding regions, i.e., plastid rbcL, matK, trnH-psbA, and nuclear ribosomal (nr)ITS. Using combinations of the datasets and

1. China Plant BOL Group


A complete list of the China Plant BOL Group can be found in Table S4.

*China Plant BOL Group


The authors declare no conflict of interest.


Data deposition: The sequences reported in this paper have been deposited in the GenBank database (accession nos. are available in Table S4).

*See Commentary on page 19451.


2. Authors should be corresponded to. E-mail: dlz@mail.kib.ac.cn.

3. This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1104551108/-/DCSupplemental.

Supporting Information includes a complete list of the China Plant BOL Group, a list of plant species sampled, and comprehensive data for each species included in this study. This information is available online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1104551108/-/DCSupplemental.

www.pnas.org/cgi/doi/10.1073/pnas.1104551108

PNAS | December 6, 2011 | vol. 108 | no. 49 | 19641–19646

www.pnas.org/cgi/doi/10.1073/pnas.1104551108


**Data deposition:** The sequences reported in this paper have been deposited in the GenBank database (accession nos. are available in Table S4).

**See Commentary on page 19451.**

**1.** A complete list of the China Plant BOL Group can be found in SI Appendix and online at: http://english.kib.csrs.cri/images/2011-10-28.pdf.

**2.** To whom correspondence should be addressed. E-mail: dlz@mail.kib.ac.cn.

**This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1104551108/-/DCSupplemental.**
following the standards and guidelines of the CBOL Plant Working Group, we tested the effectiveness and universality of the core, complementary, and additional “novel” plant barcodes as proposed at, and subsequent to, the Mexico City conference.

Results

Universalit.

All 6,286 samples were used to test universality. By using single- or multiple-primer sets as necessary, PCR success levels for rbcL, matK, trnH-psbA, and ITS in angiosperms were 94.5%, 91.0%, 90.2%, and 88.0%, respectively. For gymnosperms, the success levels were 98.7% (rbcL), 94.6% (matK), 98.5% (tmH-psbA), and 57.6% (ITS). Overall seed plant success levels were 94.8% (rbcL), 91.2% (matK), 90.7% (tmH-psbA), and 86.1% (ITS). Sequencing success rates were 97.7% (rbcL), 95.3% (matK), 97.5% (tmH-psbA), and 89.8% (ITS) in angiosperms and 99.2% (rbcL), 97.6% (matK), 99.0% (tmH-psbA), and 67.0% (ITS) in gymnosperms. Overall sequencing success rates for seed plants were 97.8% (rbcL), 95.5% (matK), 97.6% (tmH-psbA), and 88.9% (ITS) (Fig. 1A and B). Overall, the total numbers of barcode sequences generated were 5,826 for rbcL, 5,471 for matK, 5,566 for tmH-psbA, and 4,810 for ITS (Table S1). Amplification success rates were highest when using a single set of primers, as recommended by the CBOL Plant Working Group, with 97.7% for rbcL, 95.3% for matK, 97.5% for tmH-psbA, and 89.8% for ITS in angiosperms and 99.2% for rbcL, 97.6% for matK, and 99.0% for tmH-psbA, and 67.0% for ITS in gymnosperms. Overall sequencing success rates for seed plants were 97.8% (rbcL), 95.5% (matK), 97.6% (tmH-psbA), and 88.9% (ITS) (Fig. 1A and B). Overall, the total numbers of barcode sequences generated were 5,826 for rbcL, 5,471 for matK, 5,566 for tmH-psbA, and 4,810 for ITS (Table S1).

Sequence Quality.

Examination of sequence quality and coverage indicated that rbcL, matK, and ITS routinely generated high-quality bidirectional sequences. The percentage of samples from which high-quality sequences were obtained was 60.2% for rbcL, 60.2% for matK, and 58.6% for ITS; however, the sequence quality of tmH-psbA was only 40% (Fig. 1B). The mean coverage of bidirectional reads for the four candidate markers can be ranked as ITS (93.6%), matK (93.5%), rbcL (93.2%), and tmH-psbA (90.3%). Problems were encountered in assembly of the bidirectional sequences with a few ambiguous bases in tmH-psbA, which often had sequence runs interrupted by mononucleotide repeats. Similar problems were also found in matK for some taxonomic groups.

Discriminatory Power.

In total, we obtained 21,673 barcode sequences from all samples, with 18,820 sequences from 5,583 individuals of 1,349 species (at least 2 individuals per species) in 141 genera of 75 families (42 orders) of seed plants, including 121 individuals of 38 species from outside China. Coverage (Table S3) included 4 genera with >50 species, 16 genera with 20–49 species, 23 genera with 10–19 species, 72 genera with 2–9 species, and 26 genera with 1 species (17 of which are monotypic). Forty-three of the sampled genera were represented by at least 50% of their global species, and 17 genera were represented by 30–50% of their global species. Sixty-eight sampled genera were represented by at least 50% of their Chinese species, and a further 23 genera were represented by 30–50% of the Chinese species. In total, an estimated 6.1% of species and 4.4% of genera of seed plants in China were covered. The total number of barcoding sequences used for species discrimination was 5,118 (representing 1,276 species) for rbcL, 4,814 (1,197 species) for matK, 4,884 (1,206 species) for tmH-psbA, and 4,004 (1,018 species) for ITS. To evaluate the discriminatory power of the ITS2 portion of ITS, a duplicate set of these ITS sequences was made and truncated at the end of the 5.8S gene, and these ITS2 sequences were included in the assessments of discriminatory power.

Two datasets were analyzed. The first (Dataset A) comprised 5,583 samples (representing 1,349 species in 141 genera of 42 orders) with at least two sampled individuals per species to quantify discriminatory power based on the maximum data. A subdataset was extracted excluding monotypic genera and those with one sampled species (5,484 individuals representing 1,323 species in 115 genera). The second (Dataset B) comprised the 3,011 samples (representing 765 species in 83 genera of 30 orders) where at least two species were sampled per genus and all four markers were successfully sequenced to make the levels of species discrimination compatible with those of the CBOL.

Fig. 1. Comparison of the performance of four barcoding markers, rbcL, matK, rbcL, and ITS. (A) Universalit. assessment for PCR and sequencing success. PCR success was based on 6,286 samples representing 1,757 species (5,897 angiosperm samples and 389 gymnosperm samples); sequencing success was based on 5,412 samples for ITS, 5,702 samples for tmH-psbA, 5,732 samples for matK, and 5,957 samples for rbcL. (B) Assessment of species discrimination success and sequence quality based on 3,011 individuals representing 765 species, where at least 2 species were sampled per genus and all four markers were successfully sequenced. Assessment of sequence quality with QV of ≥30 (see Materials and Methods for trace-quality criteria).
Plants Working Group. Discriminatory power was generally higher for Dataset B with the exception of ITS2 alone and in combination, which showed slightly higher species discrimination in Dataset A. This trend was stable except that \textit{matK} showed a slightly higher species discrimination than \textit{trnH-psbA} did in Dataset B (45.2\% versus 44.8\%) compared with Dataset A (37.1\% versus 38.2\%) (Fig. S1). Because both datasets produced similar trends in discrimination for all markers and combinations of markers, our analyses focus on Dataset B because it is most directly comparable across markers (Fig. 2).

We calculated levels of species discrimination based on the same datasets by using four different analytical methods currently used in DNA barcoding (Materials and Methods): (i) Tree-Building and (ii) Distance (both of which are based on within-genera multispecies alignments), (iii) Blast, and (iv) PWG-Distance, the distance method adopted by the CBOL Plant Working Group that uses pairwise alignments. Among these methods, Blast tended to give higher discrimination rates, without exception. The lowest rates were found when using Tree-Building except that \textit{rbcL}, \textit{matK}, and \textit{trnH-psbA} showed slightly lower rates with Distance (Fig. S2).

It is noted that, with Blast, species discrimination ranged from 29.9\% (\textit{rbcL}) to 81.1\% (ITS) with the proposed core barcode; \textit{matK} + \textit{rbcL} provided 60.8\% discrimination. To ensure that our results are comparable with the CBOL Plant Working Group, the PWG-Distance method was hereafter adopted for discussion of discriminatory power.

Of the four single-marker barcodes, ITS showed the highest discriminatory power, with 67.2\% of all species being discriminated. Its partial sequence, ITS2, also had a high identification rate (54.6\%). \textit{rbcL} showed the lowest discrimination rate (26.4\%). Among the four genera with more than 50 sampled species tested, \textit{Primula} showed the highest discrimination rate (88.2\% with ITS; 41.5\% with \textit{rbcL}), followed by \textit{Pedicularis} (86.2\% with ITS; 46.0\% with \textit{rbcL}), with \textit{Rhododendron} being the lowest (15.3\% with ITS; 10.3\% with \textit{rbcL}). Two-marker combinations led to higher rates of species discrimination, with the highest being obtained with \textit{trnH-psbA} + ITS (79.1\%; compared with that of \textit{trnH-psbA} + ITS2, which was 69.7\%), followed by \textit{matK} + ITS (75.3\%; \textit{matK} + ITS2 was 66.1\%), and \textit{rbcL} + ITS (69.9\%; \textit{rbcL} + ITS2 was 58.5\%). The lowest rate (49.7\%) for pairwise combinations of markers was obtained by using the proposed core barcode, \textit{matK} + \textit{rbcL}. A combination of ITS and any plastid DNA marker achieved 69.9\%–79.1\% species discrimination (any plastid marker + ITS2 was 58.5\%–69.7\%). Three-marker combinations generated higher discrimination when ITS was included: \textit{matK} + \textit{trnH-psbA} + ITS was the highest with 81.8\% species discrimination (\textit{matK} + \textit{trnH-psbA} + ITS2 was 75.0\%). \textit{rbcL} + \textit{matK} + ITS gave 77.4\% discrimination (\textit{rbcL} + \textit{matK} + ITS2 was 68.5\%), whereas the three plastid DNA markers (\textit{rbcL} + \textit{matK} + \textit{trnH-psbA}) together produced only 62.0\% species discrimination. The four-way combined barcode of \textit{rbcL} + \textit{matK} + \textit{trnH-psbA} + ITS gave 82.8\% discrimination (77.2\% when ITS2 was used instead of ITS).

Based on our dataset, the four markers performed differently in different orders of angiosperms. Of the 30 orders covered by Dataset B, 6 were represented by fewer than five sampled species (Alismatales and Solanales, both with four sampled species, and Aquifoliiales, Crossoxomatales, Malpighiales, and Myrtales, each with two sampled species); these orders are not discussed because of this inadequate sampling. Laurales was the most intractable order, with very low species discrimination when using all four markers (1.8\%–14.3\%). ITS generally performed well for the major orders of seed plants, with lowest discrimination success in Ranunculales (6.7\%) and Laurales (14.3\%). \textit{trnH-psbA} performed well in Saxifragales, relatively well in Brassicales, Caryaophyllales, Celastrales, and Sapindales, but worse in Dicosorecales, Poales, and Apiales. \textit{matK} performed better in Saxifragales and Asparagales but poorly in Poales, Laurales, and Dioecocarles (Fig. 3).

Incongruence between nuclear ITS and plastid DNA barcode markers. When comparing the results based on nuclear ITS and plastid DNA markers applied to multiple individuals within morphologically defined species, incongruence was observed in some samples for 52 of 115 (45.2\%) sampled genera (excluding monotypic genera and genera with only one sampled species). This incongruence may take three forms: first, all individuals of a single species were grouped as such by the ITS sequences but were divided into two or more different entities (species) by plastid DNA sequences [22 genera, or 19.1\%, e.g., \textit{Morinda} (Rubiacae); Fig. S3]; second, all individuals of a single species were grouped into a species by the plastid DNA sequences but were divided into two or more different species by ITS data [23

Fig. 2. Comparison of discrimination success for the four markers (plus ITS2, the partial sequence of ITS) and all 2- to 4-marker combinations based on 3,011 individuals representing 765 species, where at least 2 species were sampled per genus and all four markers were successfully sequenced (\textit{I}, ITS; \textit{I2}, ITS2; \textit{M}, \textit{matK}; \textit{P}, \textit{trnH-psbA}; \textit{R}, \textit{rbcL}).
Fig. 3. Discrimination success at the ordinal level (1 order of gymnosperms, 23 orders of angiosperms) for four markers (plus ITS2, the ITS partial sequence) and all possible 2- to 4-marker combinations, based on 3,011 individuals representing 765 species, where at least 2 species were sampled per genus and all four markers were successfully sequenced (I, ITS; I2, ITS2; M, matK; P, trnH–psbA; R, rbcL). Sequence of angiosperm orders is according to the Angiosperm Phylogeny Group (APG) III (42).

Discussion

Primer universality is an important criterion for an ideal DNA barcode. Among the three plastid markers, rbcL showed the highest level of universality in both angiosperms and gymnosperms, and matK and trnH–psbA performed better in gymnosperms than in angiosperms. However, nrITS performed relatively well in angiosperms, with moderately high universality (PCR: 88%; sequencing: 89.8%) but with lower success in gymnosperms (PCR: 57.6%; sequencing: 67%). Overall, we detected a modest frequency of multiple-copy sequences from ITS (7.4% individuals, including species in genera such as Castanopsis and Fagus [Fagaceae]) and only rare cases of fungal contamination (2.5% of individuals in total). The greatest problems with ITS were encountered in gymnosperms where the great variability in length and lack of universal primers hampered PCR and sequencing success, although some of the problems may be alleviated with use of additional primers (only one pair of ITS primers, IT1 and IT1S4, was used in this study). Furthermore, in cases where ITS is difficult to amplify and performs unsatisfactorily, ITS2 represents a useful alternative for gymnosperms, or even for other seed plants (13), because of the relative ease of amplification with a single set of universal primers in all green plants (21).

The proposed core barcode, rbcL + matK, discriminated only 49.7% of the sampled species in the present study, much lower than the 72% figure previously reported (12). There are two possible explanations for this discrepancy. The most obvious reason is that the focus of the study by the CBOL Plant Working Group was to assess relative, rather than absolute, discriminatory power of the tested barcode markers. In the present study, we sampled many more closely related species within single genera. It is clear that rbcL and matK discriminate well at the genus level; however, their identification power decreases at infrageneric levels. The second explanation is that these two plastid DNA regions have high species identification power at the species level in some taxonomic groups (e.g., Orchidaceae), as suggested by previous studies (22) and confirmed by the present study, but do not perform well in other groups such as Poales, Laurales, Dioscoreales, Apiales, and Zygophyllales (Fig. 3). The inclusion of well-sampled genera in certain families undoubtedly reduced the discriminatory power of these two markers, alone and in combination.

Our study found that, of the four single markers and the combined plastid DNA markers, for taxonomic groups in which direct sequencing of this marker is possible, ITS had the highest overall discriminating power (Fig. 2). This finding is consistent with numerous previous studies showing that this nrDNA region evolves rapidly, leading to genetic changes that can differentiate closely related, congeneric species (9, 19, 23). This ITS region, or a portion of it (ITS2), has already been suggested as a potential DNA barcode for plants (9, 13, 14, 19). However, because of the incomplete concerted evolution of this nuclear multiple-copy region caused by hybridization or other factors, it is difficult to amplify and directly sequence the region in some taxa (20). Our results also confirm that ITS had lower amplification and sequencing success compared with the three plastid DNA regions, particularly in gymnosperms. Conversely, in 5–10% of the sampled angiosperm species, we found that PCR amplification of the three plastid DNA regions failed when the amplification and direct sequencing of ITS performed well.

The argument as to whether ITS or ITS2 should be a universal or a local plant barcode has been profound (24, 25) and continual since it was first proposed as a candidate barcode. The limitations of ITS have been well-documented in general terms.

genera, or 20%, e.g., Thladiantha (Cucurbitaceae); Fig. S4]; and third, species could be identified and differentiated from closely related species by the ITS sequences but could not be distinguished by plastid DNA data [15 genera, or 13%, e.g., Pugionium (Brassicaceae); Fig. S5]. The former two scenarios clearly suggest hybridization and introgression between closely related species or shared ancestral polymorphisms; the third scenario indicates either a lower mutation rate in plastid DNA compared with ITS or possible hybridization and introgression, as has been found by numerous previous studies (9, 19, 20).
Key concerns regarding ITS are (i) incomplete concerted evolution can lead to divergent paralogous copies within individuals, (ii) fungal contamination, and (iii) difficulties in amplifying and sequencing the marker in diverse sample sets (15). However, there have been few formal empirical estimates of the number of plant groups in which these problems are likely to occur. In our analyses of a large dataset with 6,286 individuals of 1,757 species in 141 genera, direct sequencing of single-copy ITS sequences were successful in 75.5% of sampled species, whereas multiple copies within individuals were limited to 7.4% of the sampled individuals, and fungal contamination was detected in only 1.8% of the sampled species. It seems that the extent of the problems concerning ITS as a standard core plant DNA barcode is not as pervasive as previously estimated. In cases where it is difficult to amplify and directly sequence ITS in its entirety, ITS2 could be an alternative because it is shorter and easier to sequence than ITS (21, 26). Our study revealed that the discriminatory power of ITS2 is higher than that of plastid markers, although it is generally 10% lower than ITS per se. Given the existing bioinformatics support, coupled with the relative ease of obtaining comparable data and the benefits of a secondary-structure approach (27, 28), ITS2 does, however, represent a useful back-up where obtaining the entire ITS region is not possible.

An ideal DNA barcode should be universal, reliable, and cost-effective and show good discriminatory power (12). Because none of the proposed barcodes perfectly meets all these criteria, it is generally considered necessary to use more than one marker to barcode plants (8, 10). However, all previous protocols suggested the combination of two or three plastid DNA markers, i.e., rbcL + matK or rbcL + matK + trnH-psbA (8, 12). Although high-quality sequences of rbcL are easily retrievable in major lineages of seed plants, our analyses suggest that the proposed core barcode, rbcL + matK, or these together with plastid trnH-psbA, produces lower levels of discrimination than ITS alone or the combination of ITS with any plastid DNA markers (Fig. 2). Considering the tradeoffs between universality, sequence quality, discrimination, rate of throughput, and cost efficiency, we propose that ITS/ITS2 should be incorporated into the core barcode for land plants, as suggested by earlier (9, 19) and more recent (13, 25) studies. If a three-marker combination is adopted, ITS/ITS2 should be added to the proposed core barcode (i.e., rbcL + matK + ITS). This combination solves the average sequencing on the existing system, and, in many plant groups, researchers are already sequencing ITS anyway as a supplementary barcode. If a two-marker barcode is preferred, our analyses suggest that the best two-marker option is matK + ITS, which produced 75.3% species discrimination, higher than rbcL + ITS (69.9%), while maintaining higher sequence quality than trnH-psbA + ITS. For taxa where matK cannot be amplified and sequenced (a rare scenario according to our data and previous reports), rbcL could be used as a back-up marker to replace matK in a two-marker strategy. This suggestion, using rbcL + matK + ITS/ITS2 as the standard plant DNA barcode, represents a practical tradeoff solution among the various criteria. During barcoding of unidentified material, if both ITS and matK sequences can be obtained, it should enable maximal identification power, even for recently diverged or cryptic species. However, if only one sequence, or one plus rbcL, can be obtained, material may still be identified to a rough taxonomic position (for example, species group or genus). This approach does require initial population of a reference database with all three markers to a sufficient density to enable identification to the level of species discrimination afforded by each.

The inclusion of ITS/ITS2 as part of the core barcode is critically important to the application of DNA barcoding in seed plants, particularly angiosperms, for the following three reasons. First, one extensive application of DNA barcoding is in recovering unidentified or cryptic species (29, 30), which are often related closely to existing described species. Furthermore, because parapatric speciation is suggested to predominate in plants (31, 32), these recently diverged species may tend to occur in the same geographical areas as their sister species. The previously proposed barcode of rbcL + matK alone may not show adequate discriminatory power for this task. Second, DNA barcoding has the potential to help identify the origin of plants and plant products in international trade and transport, for example, protected or weedy species. However, such species may be congeneric with nonweedy or nonthreatened species (33). Without ITS or ITS2, it may be difficult to differentiate between such closely related species. Finally, a combination of DNA markers from different genomes, which have different modes of inheritance and track different evolutionary histories, will further our understanding of species delimitation and evolutionary processes of speciation, another important aim of DNA barcoding (6) that may also be highly useful for the applications described above as well as in monitoring community dynamics (34).

In this study, we found that incongruent species ascriptions between plastid DNA and ITS barcodes for multiple individuals of the same morphological species occurred in some samples for nearly half of the sampled genera. Further study is needed to obtain an accurate figure at the species level. The incongruence may result from hybridization and introgression or incomplete lineage sorting (4, 20, 23). All of these phenomena are known to occur frequently in plants (35, 36). It is now clear that using only plastid DNA markers may not enable discrimination between closely related species. In addition, our findings suggest that using only plastid DNA markers may be highly misleading when establishing a barcode database that uses a single individual for each species. Although it is not feasible, at least in the short term, to assess genetic variation within and between closely related seed plant species, the multiple-sampling strategy, as recommended by the CBOL Plant Working Group (12), will therefore be essential in establishing a reference database. Sampling multiple individuals with markers from different genomes will also allow taxonomists to double-check identifications and previous species delimitations. The follow-up and redefinition of species boundaries will refine the barcode reference database and, in turn, will lead to increased identification accuracy by DNA barcoding.

Materials and Methods

Plant Materials. Data were pooled from research groups enrolled in the DNA Barcoding Chinese Plants project in September 2009 (37). A total of 6,286 samples from 1,757 species (including 5,897 samples of 1,675 angiosperm species and 389 samples of gymnosperm species) was used to test the universality of the four markers. Only those species for which sequences were obtained for at least two individuals were used for further analysis. Thus, 5,583 samples of 1,349 species (1,257 angiosperms and 82 gymnosperms) representing the major lineages of seed plants (40 orders, 70 families, and 131 genera of angiosperms and 2 orders, 5 families, and 10 genera of gymnosperms) were used to evaluate the four candidate barcoding markers. Most of the samples were collected from China. A list of the plant samples used and their GenBank accession details are provided in Table S4.

Universality. To obtain statistics on the universality of primers and recoverability of the different markers, we assembled data on amplification and sequencing success across all research groups for all plant taxa studied. Different primer sets [trnH2/P2R for rbcL; KIM_3F/KIM_1R, 390F/1526R, and Xf/FR for matK; trnH1/PsbA1A for trnH-psbA, and ITS1 (or ITS2)/ITS4 for angiosperms and ITS-Leu/ITS4 for gymnosperms for ITS] were used for barcoding in different taxa as proposed by the CBOL Plant Working Group (12). Other alternative primers for the four markers were also used in some taxa (Table S5). The universality of PCR was assessed simply by recording whether the PCR products showed a clear single band on an agarose gel. Sequencing success was measured as whether sequence data were obtained, regardless of the amount of material required for the single-directional read. If the ITS sequence was “messy,” or showed polymorphism within a single individual by a direct PCR-based sequencing approach, we treated the ITS sequence as a sequencing failure.
Sequence Quality and Coverage. To assess suitability for bidirectional sequencing, a requirement for manual editing of sequences, we followed the methods outlined in Plant Barcoding Working Group (12), using a window size of 20 bp and starting reading from 40 bp. Sequence traces with >2 bp showing a quality value (QV) of <20 were trimmed. The amount of high-quality sequence data recovered was defined such that both the forward and reverse reads had a minimum length of 100 bp and a minimum average QV of 30 and the lengths after trimming were >50% of the original sequence length. The assembled contig was defined as having >50% overlap in alignment between the forward and reverse reads, with <1% internal gaps and substitutions when aligning the forward and reverse reads. These quality-control criteria were selected as a pragmatic set of thresholds to discriminate higher-quality sequences from lower-quality sequences. Different parameters were tested but resulted in the same general trends, i.e., rbcl, matK, and ITS performed relatively well, whereas lower sequence quality was obtained for trnH-psbA.

Species Discrimination. To evaluate species discrimination success, we applied four different methods (PWG-Distance, Distance, Blast, and Tree-Building) to the single markers to and all possible 2- to 4-marker combinations. The PWG-Distance method (simple pairwise matching for DNA barcoding) recommended by the CBOL Plant Working Group (12) employs distances calculated from pairwise alignments counting unambiguous base substitutions only. This method was used for comparison throughout the subsequent analyses (38). For Distance analysis, sequences were aligned within genera by using MUSCLE v3.6 (39), and neighbor-joining trees were constructed with p-distances in PAUP* 4.0b10 (40). Species were considered discriminated if all individuals of a species formed a monophyletic group (11). General assessment of species discrimination success followed the rationale outlined by the CBOL Plant Working Group (12). Thus, for all four methods, we used only species for which multiple individuals were sampled from multiple congeneric species (Dataset B: 3,011 individuals of 765 species). Monophasic genera and genera with only a single sampled species were not counted as potential sources of discrimination failure but were included to serve as sequence success statistics (17 monotypic genera and 9 other genera with only one sampled species). We evaluated species discrimination for multiple markers by summing the components of all possible 2- to 4-marker combinations and recording the success of each multimarker combination. Species discrimination assessments were also repeated on samples from which all four markers were successfully sequenced and multiple individuals were sampled from multiple congeneric species (Dataset B: 3,011 individuals of 765 species) by using the PWG-Distance approach. Meanwhile, we also used ITS2 (extracted from the ITS dataset) in place of ITS to conduct the same analyses to assess the discriminatory power of ITS2 by using the PWG-Distance approach.

ACKNOWLEDGMENTS. We are indebted to Ms. Yan Du, Dr. Zong-Xin Ren, and the national network in China for banking rare, endangered, and endemic species plant material. We also thank Dr. Pete M. Hollingsworth and Dr. Alexandra H. Wortley of the Royal Botanic Garden Edinburgh and Dr. W. John Kress of the Smithsonian Institution for critical reading of earlier versions of the manuscript. This work was funded by the Chinese Academy of Sciences through a Large-Scale Scientific Facilities Research Project (2009-LSF-GB005-01) and the Basic Research Program of China (973 Program no. 2007CB114100).

2. Linnaeus C (1735) Species Plantarum (Impensis Laurenti Salvii, Stockholm), 1st Ed.
Comparative analysis of a large dataset indicates that internal transcribed spacer (ITS) should be incorporated into the core barcode for seed plants

By China Plant BOL Group1

Recommended citation: China Plant BOL Group (2011) Comparative analysis of a large dataset indicates that internal transcribed spacer (ITS) should be incorporated into the core barcode for seed plants. Proc Natl Acad Sci USA 108: 19641–19646.

A complete list of authors (and their contributions) of the China Plant BOL Group can be found in SI Appendix and online at:


1Chinese Plant BOL group: De-Zhu Li, Jian-Quan Liu, Lian-Ming Gao, Hong-Wang, Jun-Bo Yang and Cheng-Xin Fu designed the research; Li Chen, Shi-Lin Chen, Xue-Jun Ge, Xing-Hui Kong, Hai-Fei Yan, Ying-Jie Zhu, Yong-Shuai Sun, Si-Yun Chen, Lei Zhao, Jing-Yuan Song, Qing-Feng Wang, Lu Lu, Rui-Jiang Wang, Ding Wu, Yue-Yu Hang, Xiu-Qin Ci, Jian-Wen Zhang, Shu-Dong Zhang, Bao-Qing Ren, Yu-Min Shui, Kun Wang, Tuo Yang, Guang-Wen Duan, Jie Liu, Wen-Bin Yu, Dian-Xiang Zhang, Qin-Ru Yang, Ying Meng, Chang-Qin Zhang, Lu-Qi Huang, Yun-Heng Ji, Xiao-Hua Jin, Yong-Hong Zhou, Hong-Qing Li, Yue-Zhi Pan, Chun-Feng Song, Xiao-Yu Dong, Ting-Shuang Yi, Gang Hao, Song-Dong Zhou, Chun-Ying Xue, Pei-Gen Xiao, Hai-Fei Yan, Jun-Bo Yang, Qin-Ru Yang, Xiang-Yun Yang, Ting-Shuang Yi, Wen-Bin Yu, Chun-Xia Zeng, Chang-Qin Zhang, Dian-Xiang Zhang, Jian-Wen Zhang, Ling Zhang, Shu-Dong Zhang, Yu-Xiao Zhang, Shu-Dong Zhou, Shi-Liang Zhou and Yong-Hong Zhou performed the research; Hong-Tao Li, Lian-Ming Gao, Chun-Xia Zeng, Hai-Fei Yan, Ying-Jie ZHU, Yong-Shuai SUN, Si-Yun CHEN, Lei ZHAO, Kun WANG, Tuo YANG, Guang-Wen DUAN and De-Zhu Li analyzed data; and De-Zhu Li, Lian-Ming GAO, Jian-Quan LIU, Hong WANG, Zhi-Duan CHEN, Xue-Jun GE, and Shilin CHEN prepared the paper.

Author list (by institution):

Page 1 of 4
a. Plant Germplasm and Genomics Center, Germplasm Bank of Wild Species, Kunming Institute of Botany, Chinese Academy of Sciences (12)

De-Zhu Li\textsuperscript{a,b}, Hong-Tao Li\textsuperscript{a}, Jun-Bo Yang\textsuperscript{a}, Chun-Xia Zeng\textsuperscript{a}, Si-Yun Chen\textsuperscript{a}, Lei Zhao\textsuperscript{a}, Shu-Dong Zhang\textsuperscript{a}, Jie Liu\textsuperscript{a}, Wen-Bin Yu\textsuperscript{a}, Ting-Shuang Yi\textsuperscript{a}, Yu-Xiao Zhang\textsuperscript{a}, Xiang-Yun Yang\textsuperscript{a}

b. Key Laboratory of Biodiversity and Biogeography, Kunming Institute of Botany, Chinese Academy of Sciences (8)

Lian-Ming Gao\textsuperscript{b}, Hong Wang\textsuperscript{b}, Lu Lu\textsuperscript{b}, Jian-Wen Zhang\textsuperscript{b}, Yu-Min Shui\textsuperscript{b}, Yun-Heng Ji\textsuperscript{b}, Chun-Ying Xue\textsuperscript{b}, Chun-Lei Xiang\textsuperscript{b},

c. South China Botanical Garden, Chinese Academy of Sciences (7)

Xue-Jun Ge\textsuperscript{c}, Hang-Hui Kong\textsuperscript{c}, Hai-Fei Yan\textsuperscript{c}, Rui-Jiang Wang\textsuperscript{c}, Dian-Xiang Zhang\textsuperscript{c}, Qin-Er Yang\textsuperscript{c}, Ye Sun\textsuperscript{c}

d. Lanzhou University (3)

Jian-Quan Liu\textsuperscript{d}, Yong-Shuai Sun\textsuperscript{d}, Kun Wang\textsuperscript{d},

e. Institute of Botany, Chinese Academy of Sciences (6)

Zhi-Duan Chen\textsuperscript{e}, Shi-Liang Zhou\textsuperscript{e}, Bao-Qing Ren\textsuperscript{e}, Tuo Yang\textsuperscript{e}, Xiao-Hua Jin\textsuperscript{e}, Xiao-Yu Dong\textsuperscript{e},

f. Chinese Academy of Medical Sciences, Peking Union Medical College (4)

Shilin Chen\textsuperscript{f}, Ying-Jie Zhu\textsuperscript{f}, Jing-Yuan Song\textsuperscript{f}, Pei-Gen Xiao\textsuperscript{f}

g. Zhejiang University (2)

Cheng-Xin Fu\textsuperscript{g}, Zhe-Chen Qi\textsuperscript{g},

h. Sichuan University (2)

Xing-Jin He\textsuperscript{h}, Song-Dong Zhou\textsuperscript{h},

i. Xishuangbanna Tropical Botanical Garden, Chinese Academy of Sciences (2)

Ling Zhang\textsuperscript{i}, Xiu-Qin Ci\textsuperscript{i},

j. Wuhan Botanical Garden, Chinese Academy of Sciences (2)

Qing-Feng Wang\textsuperscript{j}, Li Chen\textsuperscript{j},

k. Jingdezhen College (1)

Ding Wu\textsuperscript{k},

Page 2 of 4
l. Institute of Botany, Jiangsu Province and Chinese Academy of Sciences (2)
   Yue-Yu Hang\textsuperscript{l}, Chun-Feng Song\textsuperscript{l},

m. Key Laboratory of Economic Plants and Biotechnology, Kunming Institute of Botany, Chinese Academy of Sciences (5)
   Guang-Wen Duan\textsuperscript{m}, Chang-Qin Zhang\textsuperscript{m}, Yue-Zhi Pan\textsuperscript{m}, Yuan Huang\textsuperscript{m}, Yu-Hua Wang\textsuperscript{m},

n. Institute of Tibetan Plateau Research at Kunming, Chinese Academy of Sciences (1)
   Ying Meng\textsuperscript{n}

o. Institute of Chinese Materia Medica, China Academy of Chinese Medical Sciences (1)
   Lu-Qi Huang\textsuperscript{o},

p. Sichuan Agricultural University (1)
   Yong-Hong Zhou\textsuperscript{p},

q. East China Normal University (1)
   Hong-Qing Li\textsuperscript{q},

r. South China Agricultural University (1)
   Gang Hao\textsuperscript{r},

s. Shanghai Jiao Tong University (1)
   Zhong Liu\textsuperscript{s},

Full institutional addresses

\textsuperscript{a}Plant Germplasm and Genomics Centre, Germplasm Bank of Wild Species, Kunming Institute of Botany, Chinese Academy of Sciences, Kunming, Yunnan 650204, China;

\textsuperscript{b}Key Laboratory of Biodiversity and Biogeography, Kunming Institute of Botany, Chinese Academy of Sciences, Kunming, Yunnan 650204, China;

\textsuperscript{c}Key Laboratory of Plant Resources Conservation and Sustainable Utilization, South China Botanical Garden, Chinese Academy of Sciences, Guangzhou 510650, China;

\textsuperscript{d}Key Laboratory of Arid and Grassland Ecology, College of Life Science, Lanzhou University, Lanzhou 730000, Gansu, China;
¢State Key Laboratory of Systematic and Evolutionary Botany, Institute of Botany, Chinese Academy of Sciences, Beijing 100093, China;

£Institute of Medicinal Plant Development, Chinese Academy of Medical Sciences, Peking Union Medical College, Beijing 100193, China;

£Laboratory of Systematic and Evolutionary Botany, Department of Biology, College of Life Science, Zhejiang University, Hangzhou 310058, China;

School of Life Sciences, Sichuan University, Chengdu, Sichuan 610064, China;

Key Laboratory of Tropical Forest Ecology, Xishuangbanna Tropical Botanical Garden, Chinese Academy of Sciences, Menglun, Yunnan 666303, China;

Department of Taxonomy and Systematics, Wuhan Botanical Garden, Chinese Academy of Sciences, Wuhan, Hubei 430074, China;

Jingdezhen College, Jingdezhen, Jiangxi 333000, China;

Institute of Botany, Jiangsu Province and Chinese Academy of Sciences, Nanjing, 210014, China;

Key Laboratory of Economic Plants and Biotechnology, Kunming Institute of Botany, Chinese Academy of Sciences, Kunming, Yunnan 650204, China;

Institute of Tibetan Plateau Research at Kunming, Chinese Academy of Sciences, Kunming, Yunnan 650204, China;

Laboratory for Molecular Pharmaceutics, Institute of Chinese Materia Medica, China Academy of Chinese Medical Sciences, Beijing 100700, China;

Triticeae Research Institute, Sichuan Agricultural University, Wenjiang, Chengdu, Sichuan 611180, China;

School of Life Science, East China Normal University, Shanghai 200062, China;

College of Life Sciences, South China Agricultural University, Guangzhou 510642, China; School of Pharmacy, Shanghai Jiao Tong University, Shanghai, 200240, China

*To whom correspondence should be addressed. E-mail: dzl@mail.kib.ac.cn.*
Supporting Information

Li et al. 10.1073/pnas.1104551108

Fig. S1. Comparison of discrimination success for the four markers and all possible 2- to 4-marker combinations. (A) Based on 5,583 samples representing 1,349 species where at least 2 individuals were sampled per species (monotypic genera and those with 1 sampled species were excluded). (B) Based on 3,011 individuals representing 765 species where at least 2 species were sampled per genus and all four markers were successfully sequenced (I, ITS; M, matK; P, trnH-psbA; R, rbcL).

Fig. S2. Comparison of discrimination success for four markers and all possible 2- to 4-marker combinations based on four different analytical methods (for codes for the four analytical methods, see Materials and Methods, Species Discrimination) for 5,583 samples of 1,349 species. I, internal transcribed spacer (ITS); I2, ITS2; M, matK; P, trnH-psbA; R, rbcL.
Fig. S3. Topologies of *Morinda* (Rubiaceae) based on *rbcL*, *matK*, *trnH-psbA*, *rbcL + matK + trnH-psbA*, and ITS sequences.
Fig. S4. Topologies of Thladiantha (Cucurbitaceae) based on rbcl, matK, trnH-psbA, rbcl + matK + trnH-psbA, and ITS sequences.
Fig. S5. Topologies of Pugionium (Brassicaceae) based on rbcL, matK, trnH-psbA, rbcL + matK + trnH-psbA, and ITS sequences.
Table S1. Universality statistics based on all 6,286 samples

Table S2. Statistics of ITS sequences based on 5,583 samples

Table S3. Coverage of the 141 sampled genera in this study

Table S4. Sample details with voucher information and GenBank accession nos. for four markers for all 5,583 samples of 1,349 species

Table S5. (A) Primer information for the four markers used in 141 sampled genera. (B) Primer sequences applied in this study. F, forward; R, reverse.

Other Supporting Information Files

SI Appendix (PDF)