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 Group[1], De-Zhu Li[a,2], Lian-Ming Gao[a], Hong-Tao Li[a], Hong Wang[a], Xue-Jun Ge[a], Jian-Quan Liu[a], Zhi-Duan Chen[a], Shi-Liang Zhou[a], Shi-Lin Chen[a], Jun-Bo Yang[a], Cheng-Xin Fu[a], Chun-Xia Zeng[a], Hai-Fei Yan[a], Ying-Jie Zhu[a], Yong-Shuai Sun[a], Si-Yun Chen[a], Lei Zhao[a], Kun Wang[a], Tuo Yang[a], and Guang-Wen Duan[a]

[a]China Plant BOL Group

Edited* by Daniel H. Janzen, University of Pennsylvania, Philadelphia, PA, and approved August 25, 2011 (received for review March 23, 2011)

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, ascriptions 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.

land plants | species identification | nuclear ribosomal (nr) DNA

The seed plants account for some 90% of land plant diversity, dominating terrestrial ecosystems and providing food, timber, drugs, fibers, fuels, and ornamentals for human use (1). Identification is an essential step for humans in using and conserving plants. Since the time of Linnaeus, botanists have used a range of character sources as taxonomic evidence for documenting plant biodiversity (2), including gross morphology, anatomy, embryology, palynology, pollination biology, chromosomes, proteins, secondary metabolites, and ad hoc use of DNA sequence data (3). However, it can still be difficult to rapidly and accurately identify plant species. In part, this is because of the huge diversity of plant species and the fact that identifications are often attempted from suboptimal material that lacks the key diagnostic characters. It is especially difficult in the case of closely related species where recent radiation, frequent hybridization, and high intraspecific variation can compound identification problems (4, 5).

DNA barcoding, an approach to identify species based on sequences from a short, standardized DNA region, opens up a unique avenue for the identification of organisms (6, 7). Although *CO1*, a mitochondrial marker, is known to work relatively consistently in animal barcoding, this region has not been adopted for plants because of low substitution rates in the plant mitochondrial genome (8). A number of DNA regions, the majority taken from the plastid genome, have instead been tested for universality and discriminatory power in barcoding plants (8–11). 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 broadleaved 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

Author contributions: D.-Z.L., L.-M.G., H.W., Z.-D.C., X.-J.G., S.-L.Z., S.-L.C., J.-B.Y., and C.-X.F. designed research; China Plant BOL Group performed research; D.-Z.L., L.-M.G., H.-T.L., C.-X.Z., H.-F.Y., Y.-J.Z., Y.-S.S., S.-Y.C., L.Z., K.W., T.Y., and G.-W.D. analyzed data; and D.-Z.L., L.-M.G., J.-Q.L., H.W., Z.-D.C., X.-J.G., and S.-L.C. wrote the paper.

The authors declare no conflict of interest.

*This Direct Submission article had a prearranged editor.

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.cas.cn/images/2011-10-28.pdf.

[2]To whom correspondence should be addressed. E-mail: dzl@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

Universality. 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% (*trnH–psbA*), and 57.6% (ITS). Overall seed plant success levels were 94.8% (*rbcL*), 91.2% (*matK*), 90.7% (*trnH–psbA*), and 86.1% (ITS). Sequencing success rates were 97.7% (*rbcL*), 95.3% (*matK*), 97.5% (*trnH–psbA*), and 89.8% (ITS) in angiosperms and 99.2% (*rbcL*), 97.6% (*matK*), 99.0% (*trnH–psbA*), and 67.0% (ITS) in gymnosperms. Overall sequencing success rates for seed plants were 97.8% (*rbcL*), 95.5% (*matK*), 97.6% (*trnH–psbA*), and 88.9% (ITS) (Fig. 1 *A* and *B*). Overall, the total numbers of barcode sequences generated were 5,826 for *rbcL*, 5,471 for *matK*, 5,566 for *trnH–psbA*, and 4,810 for ITS (Table S1). Amplification success rates when using a single set of primers, as recommended by the CBOL Plant Working Group, were 89.2% for *rbcL* and 79.5% for *matK*; a single set of primers was used to amplify all *trnH–psbA* sequences.

A single set of ITS primers, ITS5 (ITS1 or ITS-Leu)/ITS4, was tested on 82.9% of samples. Direct sequencing of single-copied ITS sequences was successful in 71.7% of individuals and 75.5% of species, whereas multiple copies within individuals were limited to 7.4% of individuals and 9.3% of species. Fungal contamination was detected in only 2.5% of individuals and 1.8% of the sampled species. In addition, 18.4% of individuals and 13.4% of species, mainly gymnosperms, were not successfully sequenced for ITS (Table S2).

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 60.3% for *rbcL*, 60.2% for *matK*, and 58.6% for ITS; however, the sequence quality of *trnH–psbA* was only 40% (Fig. 1*B*). 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 *trnH–psbA* (90.3%). Problems were encountered in assembly of the bidirectional sequences with a few ambiguous bases in *trnH–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 *trnH–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

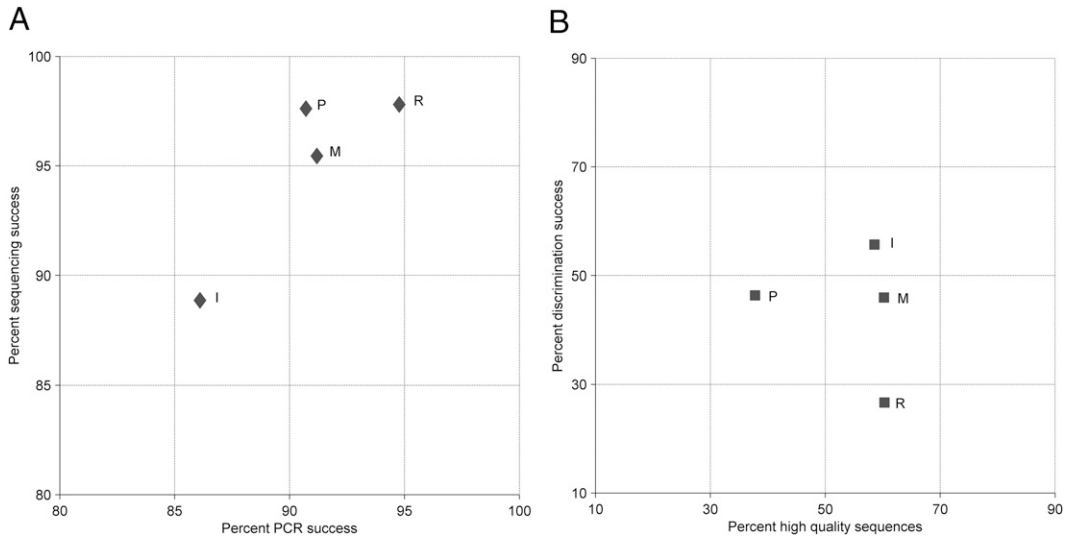

Fig. 1. Comparison of the performance of four barcoding markers (I, ITS; M, *matK*; P, *trnH–psbA*; R, *rbcL*). (*A*) Universality assessment for PCR and sequencing success. PCR success was based on 6,286 samples representing 1,757 species (5,897 angiosperm samples and 389 gymnosperm samples); sequencing success was based on 5,412 samples for ITS, 5,702 samples for *trnH–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).

Plant 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 *matK* showed a slightly higher species discrimination than *trnH–psbA* did in Dataset B (45.2% versus 44.8%) compared with Dataset A (37.1% vs. 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 *rbcL*, *matK*, and *trnH–psbA* showed slightly lower rates with Distance (Fig. S2). It is noted that, with Blast, species discrimination ranged from 29.9% (*rbcL*) to 81.1% (ITS) with the proposed core barcode; *matK* + *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%). *rbcL* showed the lowest discrimination rate (26.4%). Among the four genera with more than 50 sampled species tested, *Primula* showed the highest discrimination rate (88.2% with ITS; 41.5% with *rbcL*), followed by *Pedicularis* (86.2% with ITS; 46.0% with *rbcL*), with *Rhododendron* being the lowest (15.3% with ITS; 10.3% with *rbcL*). Two-marker combinations led to higher rates of species discrimination, with the highest being obtained with *trnH–psbA* + ITS (79.1%; compared with that of *trnH–psbA* + ITS2, which was 69.7%), followed by *matK* + ITS (75.3%; *matK* + ITS2 was 66.1%), and *rbcL* + ITS (69.9%; *rbcL* + ITS2 was 58.5%). The lowest rate (49.7%) for pairwise combinations of markers was obtained by using the proposed core barcode, *matK* + *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: *matK* + *trnH–psbA* + ITS was the highest with 81.8% species discrimination (*matK* + *trnH–psbA* + ITS2 was 75.0%), *rbcL* + *matK* + ITS gave 77.4% discrimination (*rbcL* + *matK* + ITS2 was 68.5%), whereas the three plastid DNA markers (*rbcL* + *matK* + *trnH–psbA*) together produced only 62.0% species discrimination. The four-way combined barcode of *rbcL* + *matK* + *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 Aquifoliales, Crossosomatales, 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%). *trnH–psbA* performed well in Saxifragales, relatively well in Brassicales, Caryophyllales, Celastrales, and Sapindales, but worse in Dioscoreales, Poales, and Apiales. *matK* performed better in Saxifragales and Asparagales but poorly in Poales, Laurales, and Dioscoreales (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., *Morinda* (Rubiaceae); 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

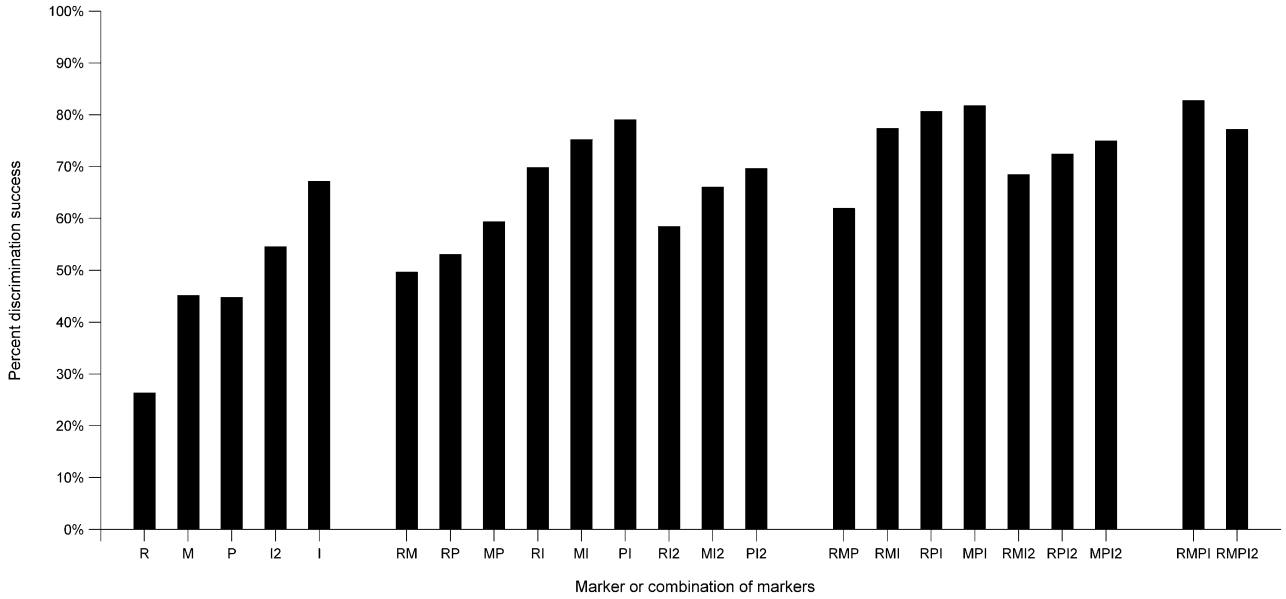

Fig. 2. Comparison of discrimination success for the four markers (plus ITS2, the partial sequence of ITS) and all 2- to 4-marker combinations based on 3,011 individuals representing 765 species, where at least 2 species were sampled per genus and all four markers were successfully sequenced (I, ITS; I2, ITS2; M, *matK*; P, *trnH–psbA*; R, *rbcL*).

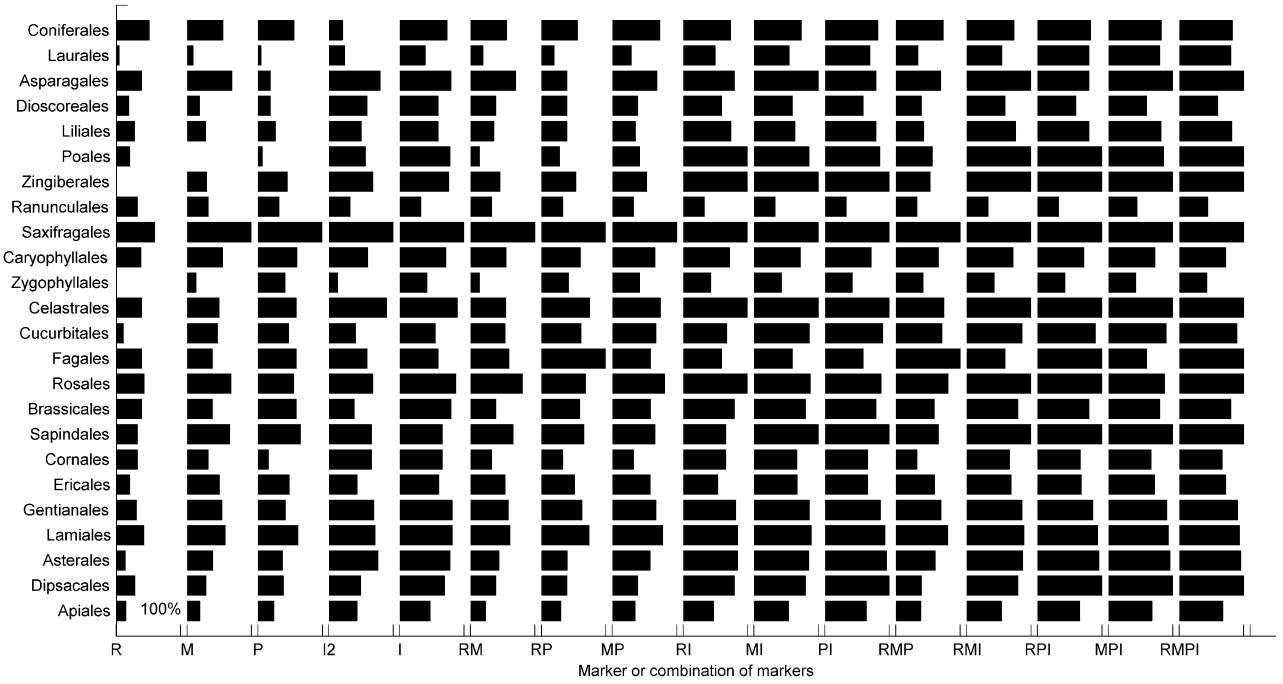

Fig. 3. Discrimination success at the ordinal level (1 order of gymnosperms, 23 orders of angiosperms) for four markers (plus ITS2, the ITS partial sequence) and all possible 2- to 4-marker combinations, based on 3,011 individuals representing 765 species, where at least 2 species were sampled per genus 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).

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).

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 primer, ITS-Leu/ITS4, 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.

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/ITS2). This solution has the advantage of building 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 identifications 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 *all* closely related seed plant species, the multiple-sampling strategy, as recommended by the CBOL Plant Working Group (12), will therefore to 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 82 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 [1F/724R for *rbcL*; KIM_3F/KIM_1R, 390F/1326R, and XF/5R for *matK*; *trn*H2/*psb*Af for *trnH–psbA;* and ITS1 (or ITS5)/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 manual trace editing required or the extent of the bidirectional 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 method used by the CBOL Plant 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% low-quality bases (<20 QV) and <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 and to 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 p-distances were calculated with PAUP* 4.0b10 (40). For both of the measures based on distance only, we considered discrimination to be successful if the minimum uncorrected interspecific p-distance involving a species was larger than its maximum intraspecific distance. For the Blast method, all sequences of the four markers and possible combinations of 2–4 markers were used as query sequences with an *E* value $<1 \times 10^{-5}$, and the Blast program (v2.2.17) was used to query the reference database with each sample in turn to establish whether the closest hit was a conspecific species and to provide statistics for species discrimination (the query sequence itself was excluded from the list of top hits). Species discrimination was considered successful if all individuals of a species had a top matching hit of only a conspecific individual (41). When using the Tree-Building method, 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 A: 5,484 samples of 1,323 species). Monotypic 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 seeds for 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-GBOWS-01) and the Basic Research Program of China (973 Program no. 2007CBI411600).

1. Mabberley DJ (2008) *Mabberley's Plant-book: A Portable Dictionary of Plants, Their Classifications and Uses* (Cambridge Univ Press, Cambridge, UK), 3rd Ed.
2. Linnaeus C (1753) *Species Plantarum* (Impensis Laurentii Salvii, Stockholm), 1st Ed.
3. Judd WS, Campbell CS, Kellogg EA, Stevens PF, Donoghue MJ (2008) *Plant Systematics, a Phylogenetic Approach* (Sinauer, Sunderland, MA), 3rd Ed.
4. Rieseberg LH, Wood TE, Baack EJ (2006) The nature of plant species. *Nature* 440: 524–527.
5. Stebbins GL (1950) *Variation and Evolution in Plants* (Columbia Univ Press, New York), p xix.
6. Hebert PD, Cywinska A, Ball SL, deWaard JR (2003) Biological identifications through DNA barcodes. *Proc Biol Sci* 270:313–321.
7. Hebert PDN, Gregory TR (2005) The promise of DNA barcoding for taxonomy. *Syst Biol* 54:852–859.
8. Fazekas AJ, et al. (2008) Multiple multilocus DNA barcodes from the plastid genome discriminate plant species equally well. *PLoS ONE* 3:e2802.
9. Kress WJ, Wurdack KJ, Zimmer EA, Weigt LA, Janzen DH (2005) Use of DNA barcodes to identify flowering plants. *Proc Natl Acad Sci USA* 102:8369–8374.
10. Kress WJ, Erickson DL (2007) A two-locus global DNA barcode for land plants: The coding *rbcL* gene complements the non-coding *trnH-psbA* spacer region. *PLoS ONE* 2: e508.
11. Hollingsworth ML, et al. (2009) Selecting barcoding loci for plants: Evaluation of seven candidate loci with species-level sampling in three divergent groups of land plants. *Mol Ecol Resour* 9:439–457.
12. CBOL Plant Working Group (2009) A DNA barcode for land plants. *Proc Natl Acad Sci USA* 106:12794–12797.
13. Yao H, et al. (2010) Use of ITS2 region as the universal DNA barcode for plants and animals. *PLoS ONE* 5:e13102.
14. Chen SL, et al. (2010) Validation of the ITS2 region as a novel DNA barcode for identifying medicinal plant species. *PLoS ONE* 5:e8613.
15. Hollingsworth PM, Graham SW, Little DP (2011) Choosing and using a plant DNA barcode. *PLoS ONE* 6:e19254.
16. Raven PH (2011) Plant conservation in the future: New challenges, new opportunities. *Plant Diversity Resour* 33:1–9.
17. Mittermeier RA, et al. (2005) *Hotspots Revisited-Earth's Biologically Richest and Most Endangered Terrestrial Ecoregions* (Univ of Chicago Press, Chicago).
18. Li DZ (2008) Floristics and plant biogeography in China. *J Integr Plant Biol* 50:771–777.
19. Sass C, Little DP, Stevenson DW, Specht CD (2007) DNA barcoding in the Cycadales: Testing the potential of proposed barcoding markers for species identification of cycads. *PLoS ONE* 2:e1154.
20. Alvarez I, Wendel JF (2003) Ribosomal ITS sequences and plant phylogenetic inference. *Mol Phylogenet Evol* 29:417–434.
21. White TJ, Bruns T, Lee S, Taylor J (1990) Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics. *PCR Protocols*, eds Innis MA, Gelfand DH, Sninsky JJ, White TJ (Academic, San Diego), pp 315–322.
22. Lahaye R, et al. (2008) DNA barcoding the floras of biodiversity hotspots. *Proc Natl Acad Sci USA* 105:2923–2928.
23. Nieto Feliner G, Rosselló JA (2007) Better the devil you know? Guidelines for insightful utilization of nrDNA ITS in species-level evolutionary studies in plants. *Mol Phylogenet Evol* 44:911–919.
24. Chase MW, et al. (2007) A proposal for a standardised protocol to barcode all land plants. *Taxon* 56:295–299.
25. Buchheim MA, et al. (2011) Internal transcribed spacer 2 (nu ITS2 rRNA) sequence-structure phylogenetics: Towards an automated reconstruction of the green algal tree of life. *PLoS ONE* 6:e16931.
26. Coleman AW (2003) ITS2 is a double-edged tool for eukaryote evolutionary comparisons. *Trends Genet* 19:370–375.
27. Schultz J, Wolf M (2009) ITS2 sequence-structure analysis in phylogenetics: A how-to manual for molecular systematics. *Mol Phylogenet Evol* 52:520–523.
28. Keller A, et al. (2010) Including RNA secondary structures improves accuracy and robustness in reconstruction of phylogenetic trees. *Biol Direct* 5:4.
29. 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 *Astraptes fulgerator*. *Proc Natl Acad Sci USA* 101:14812–14817.
30. Bickford D, et al. (2007) Cryptic species as a window on diversity and conservation. *Trends Ecol Evol* 22:148–155.
31. Schluter D (2009) Evidence for ecological speciation and its alternative. *Science* 323: 737–741.
32. Abbott RJ, Ritchie MG, Hollingsworth PM (2008) Introduction. Speciation in plants and animals: Pattern and process. *Philos Trans R Soc Lond B Biol Sci* 363:2965–2969.
33. DeSalle R, Amato G (2004) The expansion of conservation genetics. *Nat Rev Genet* 5: 702–712.
34. Kress WJ, et al. (2009) Plant DNA barcodes and a community phylogeny of a tropical forest dynamics plot in Panama. *Proc Natl Acad Sci USA* 106:18621–18626.
35. Arnold ML (1997) *Natural Hybridization and Evolution* (Oxford Univ Press, New York).
36. Abbott RJ, Hegarty MJ, Hiscock SJ, Brennan AC (2010) Homoploid hybrid speciation in action. *Taxon* 59:1375–1386.
37. Li DZ, et al. (2011) Plant DNA barcoding in China. *J Syst Evol* 49:165–168.
38. Little DP (2009) Simple pairwise matching for DNA barcoding, Available at http://www.nybg.org/files/scientists/dlittle/PWG.html.
39. Edgar RC (2004) MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 32:1792–1797.
40. Swofford DL (2003) *PAUP*: Phylogenetic Analysis Using Parsimony* (**And Other Methods*) (Sinauer, Sunderland, MA).
41. Ross HA, Murugan S, Li WLS (2008) Testing the reliability of genetic methods of species identification via simulation. *Syst Biol* 57:216–230.
42. The Angiosperm Phylogeny Group (2009) An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG III. *Bot J Linn Soc* 161:105–121.