Refining the DNA barcode for land plants

Peter M. Hollingsworth[1]
Royal Botanic Garden Edinburgh, Edinburgh EH3 5LR, Scotland

The goal of DNA barcoding is conceptually simple: Find one or a few regions of DNA that will distinguish among the majority of the world's species, and sequence these from diverse sample sets to produce a large-scale reference library of life on earth (1). This approach can then be used as a tool for species identification and to help in the discovery of new species (2). Since the first DNA barcoding study in 2003 (1), the "animal barcode," a portion of the mitochondrial gene *Cytochrome Oxidase 1*, has proved remarkably effective at discriminating among species in diverse groups such as birds, fishes, and insects. In contrast, finding a robust and effective barcode for plants has been more difficult. In 2009, a large consortium of researchers, the "Consortium for the Barcode of Life (CBOL) Plant Working Group," proposed portions of two coding regions from the plastid (chloroplast) genome—*rbcL* and *matK*—as a "core barcode" for plants, to be supplemented with additional regions as required (3). This recommendation was accepted by the international Consortium for the Barcode of Life, but with the important qualifier that further sequencing of additional markers should be undertaken during a "trial period" (4). This trial period was driven by concerns that routine use of a third (or even a fourth) marker may be necessary to obtain adequate discriminatory power and to guard against sequencing failure for one of the markers (*matK* can be difficult to amplify and sequence). In PNAS, the China Plant Barcode of Life (BOL) Group provides an impressive dataset tackling this question (5) and assesses the potential benefits of supplementing the core barcode for land plants.

A total of 46 research teams, coordinated by Li De-Zhu from the Kunming Institute of Botany (Kunming, Yunnan, China), assembled a sequence matrix from 6,286 samples from 1,757 species from 75 seed plant families in China. They sequenced the core barcoding markers (*rbcL* and *matK*) along with two other markers: the plastid region, *trnH-psbA*, and the internal transcribed spacers of nuclear ribosomal DNA (nrDNA ITS). This dataset was used to assess universality (the ease of recovery of barcode sequences), sequence quality (how good the sequences were), and discriminatory power (how effective the sequences were in distinguishing among species). Their findings on the performance of the three plastid markers broadly match the 2009 CBOL Plant Working Group study (3), albeit here based on much larger sample sizes. Each gene region had different strengths and weaknesses: good recovery and sequence quality but low species discrimination for *rbcL*, better and broadly equal discriminatory power for *trnH-psbA* and *matK*, with more efficient recovery of *trnH-psbA* and better sequence quality for *matK*.

The use of ITS may be necessary to tip resolution levels from "too low to be useful" to "acceptable" in many situations.

The major unique finding from this study relates to the assessment of the performance of nrDNA ITS, which has so far been absent from large-scale comparative assessments of plant barcoding markers (4). The universality of this marker (~76.5%) was lower than that of the three plastid regions (~87–93%). However, it offered a significant increase in discriminatory power. Focusing on a dataset in which samples from all four loci were recovered, adding ITS to the plant barcode took levels of species discrimination success from 50-62% for two or three marker plastid barcodes, to between 77% and 82% when ITS was combined with two plastid markers.

Challenges for the Use of ITS as a Barcode

Previously, many researchers have been concerned about the use of nrDNA ITS as a standard barcode. The potential increase in resolving power is not unexpected, on the basis of its performance in phylogenetic studies (6), but there has been a residual nervousness stemming from three major potential problems:

a) Fungal contamination: The primers used for amplification and sequencing of nrDNA ITS in plants and fungi are similar enough such that fungal DNA is often inadvertently amplified from plant samples. This outcome can obviously lead to some spectacularly misleading sample identifications. However, in the study by the China Plant BOL Group, in silico searches of the data for fungal ITS motifs suggest that overall, the extent of this problem was limited, with only 2–3% of samples showing evidence of fungal contamination.

b) Paralogous gene copies: The nrDNA ITS region is present in multiple copies within each cell. These copies generally evolve in a concerted fashion, leading to a single detectable sequence per plant. However, in some plant groups, divergent copies co-occur within individuals (7). This can lead to messy sequences (attributable to the presence of multiple different variants being simultaneously sequenced) or, worse, inadvertent differential sequencing of different variants among samples. This process can lead to members of the same species being given different identifications depending on which variant was sequenced. This "paralogy problem" was not tested directly in the current study but—indirectly—the results suggest that its impacts are not as severe as might have been feared. Only a relatively modest number of sequences were unreadable in a fashion easily attributable to the presence of multiple divergent copies (7.4%), and if differential sampling of paralogous copies has occurred, it is has not happened to such an extent that it has compromised the identification ability of the region, compared with the other tested markers.

c) Recovery: The main limitation for nrDNA ITS is that it is sometimes simply difficult to amplify and sequence (8). There are a multitude of potential reasons for this, and in the current study, the recovery rate of 76.5% was some 10–15% lower than for the other barcode markers. However, this recovery rate is not spectacularly low compared with the other regions, and the authors present a "back-up plan": If obtaining full ITS is difficult, one can amplify up just half of the region (just ITS2) (9). This partial region is often much easier to amplify and sequence than the entire region, but can still provide

Author contributions: P.M.H. wrote the paper.
The author declares no conflict of interest.
See companion article on page 19641.
[1]E-mail: p.hollingsworth@rbge.org.uk.

appreciable gains in discriminatory power beyond that of plastid regions alone. Although the recovery rates of ITS2 were not tested directly in this study, in silico analyses of its resolving power (by truncating the full ITS sequences) show a still significant gain over and above that of plastid barcodes, making it a useful option when obtaining full ITS sequences is problematic.

What are the Implications for the Plant Barcode?

The findings of this paper make a persuasive case for the consideration of ITS as a routine addition to the plant barcode and provide a useful empirical estimate of the strengths and weaknesses of the region. A 20% gain in discriminatory power for full ITS and 10–15% gain even when just the ITS2 region is used is an appreciable benefit. As seen in other recent smaller-scale studies (10, 11), the use of ITS may be necessary to tip resolution levels from "too low to be useful" to "acceptable" in many situations. Basically the authors make a strong case that the benefits of using ITS outweigh its limitations. The increased resolving power of ITS also matches a recent theoretical prediction that the ability of a marker to track species limits may be associated with its dispersal ability (12). The nuclear inheritance of nrDNA ITS (transmitted in both pollen and seed) may be a contributing factor to its increased resolving power compared with the predominantly maternally inherited plastid DNA markers (which are transmitted by seed alone, and overall are expected to be more poorly dispersed, and hence show lower discrimination success) (4).

Future Prospects

In the longer term, it is desirable to increase the levels of species discrimination beyond those achievable by combining even all four of the markers tested here. Options for simultaneously sequencing entire plastid genomes and nrDNA ITS from phylogenetically disparate sample sets are becoming closer to routine (13), and when these are as cost effective as Sanger sequencing a few loci when dealing with thousands of samples and/or are widely accessible to the many smaller laboratories involved in DNA barcoding, such approaches may overtake current methods. However, these criteria have yet to be satisfied, and ultimately these approaches still do not address the crux challenge, which is obtaining sequence data from multiple unlinked single-copy nuclear markers to enable high-resolution species discrimination that will cope even with closely related species assemblages (14). The technical and analytical framework to deliver on this problem remains a pressing challenge.

However, lest there be too much angst about current imperfections of plant barcodes for species-level resolution, it is a salutary observation that there are still many plant genera lacking DNA sequences and, at present, there is no universal database populated with DNA sequence data to provide a robust genus-level identification system across land plants—supported by links to high-quality digitized reference specimens of the samples that were sequenced. And even the currently imperfect level of resolution from plant barcodes is useful for many applications, ranging from the discovery of new species (15), to "ecological forensic" insights into community structure (16), to practical outcomes such as detecting that ~30% of tested commercial tea products showed the presence of nonlabel ingredients (17). And finally, the rate-limiting step in building a high-quality reference library of DNA sequences of plant life on earth will be the collection and assembly of well-identified sample sets, amenable for DNA sequence analysis. Once these sample sets are assembled, their subsequent resequencing for additional loci in light of technical improvements will be relatively straightforward.

In summary, the paper by Li et al. (5) represents another step forward toward routine use of DNA sequence data as a tool for species-level plant taxonomy and identification. The China Plant BOL Group presents the argument that the benefits of using nrDNA ITS in terms of species resolution are likely to outweigh the problems of using this region and that in the short-to-medium term, this approach will improve our ability to distinguish among plant species.

1. Hebert PDN, Cywinska A, Ball SL, deWaard JR (2003) Biological identifications through DNA barcodes. *Proc Biol Sci* 270:313–321.
2. 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.
3. CBOL Plant Working Group (2009) A DNA barcode for land plants. *Proc Natl Acad Sci USA* 106:12794–12797.
4. Hollingsworth PM, Graham SW, Little DP (2011) Choosing and using a plant DNA barcode. *PLoS ONE* 6: e19254.
5. Li D-Z, et al. (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.
6. Baldwin BG (1992) Phylogenetic utility of the internal transcribed spacers of nuclear ribosomal DNA in plants: An example from the compositae. *Mol Phylogenet Evol* 1:3–16.
7. Álvarez I, Wendel JF (2003) Ribosomal ITS sequences and plant phylogenetic inference. *Mol Phylogenet Evol* 29:417–434.
8. 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.
9. Chen S, et al. (2010) Validation of the ITS2 region as a novel DNA barcode for identifying medicinal plant species. *PLoS ONE* 5:e8613.
10. Roy S, et al. (2010) Universal plant DNA barcode loci may not work in complex groups: A case study with Indian *berberis* species. *PLoS ONE* 5:e13674.
11. Muellner AN, Schaefer H, Lahaye R (2011) Evaluation of candidate DNA barcoding loci for economically important timber species of the mahogany family (Meliaceae). *Mol Ecol Resour* 11:450–460.
12. Petit RJ, Excoffier L (2009) Gene flow and species delimitation. *Trends Ecol Evol* 24:386–393.
13. Steele PR, Pires JC (2011) Biodiversity assessment: State-of-the-art techniques in phylogenomics and species identification. *Am J Bot* 98:415–425.
14. Chase MW, et al. (2005) Land plants and DNA barcodes: Short-term and long-term goals. *Philos Trans R Soc Lond B Biol Sci* 360:1889–1895.
15. Bell D, et al. (2011) DNA barcoding European *Herbertus* (Marchantiopsida, Herbertaceae) and the discovery and description of a new species. *Mol Ecol Resour*, 10.1111/j.1755-0998.2011.03053.x.
16. Kesanakurti PR, et al. (2011) Spatial patterns of plant diversity below-ground as revealed by DNA barcoding. *Mol Ecol* 20:1289–1302.
17. Stoeckle MY, et al. (2011) Commercial teas highlight plant DNA barcode identification successes and obstacles. *Nat Sci Rep* 1:42.