npg

NEWS AND COMMENTARY

DNA barcoding plants in biodiversity hot spots

Progress and outstanding questions

PM Hollingsworth

DNA barcoding in animals is now routinely used for organismal identification and has contributed toward the discovery of new species. Although the approach has received strong criticisms, a number of studies have illustrated how sequencing just a single organelle region (mitochondrial cytochrome oxidase 1, *CO1*) can serve as a powerful high-throughput tool for biodiversity research (Hajibabaei *et al.*, 2007). In plants, progress has been hampered by slow substitution rates in mitochondrial DNA, and the search for an analogous region to animal *CO1* has focused on chloroplast DNA. A number of different chloroplast regions have been proposed, but a consensus remains elusive (Pennisi, 2007; Ledford, 2008). The plant barcoding regions suggested at the Second International Barcode of Life Conference in Taipei (September 2007) are summarised in Table 1.

A recent paper published by Lahaye *et al.* (2008) reports on the application of DNA barcoding in plants and tackles two substantive issues. Firstly, the authors provide new data to contribute toward this ongoing debate regarding the most appropriate DNA regions for barcoding in plants and secondly, they apply one candidate barcoding region to the flora of a global biodiversity hot spot.

To assess the comparative performance of different barcoding regions, 71 specimens of 48 Costa Rican orchid species and 101 samples of 38 species from the Kruger National Park in South Africa were examined with eight candidate barcoding regions. These included *rbcL*, *rpoC1*, *rpoB*, *trnH-psbA* and *matK* that feature in the preferred barcode solutions of different research groups described above, with *accD*, *nhdJ* and *ycf5* that have been considered previously as potential loci by the consortium led by RBG Kew. The *atpF-H* and *psbK-I* spacers very recently proposed by Kim *et al.* (Table 1) were not included.

Of the regions Lahaye *et al.* (2008) analyzed, *matK* was their preferred option. Their results support observations from other groups that *matK* has a rapid substitution rate compared to other chloroplast coding regions (Chase *et al.*, 2007). Critically, however, the authors report high levels of amplification success (100%) from a single primer pair, a result, which, to date, has not been obtained by other groups. This gene has a reputation for being one of the more difficult chloroplast regions to routinely amplify and sequence across divergent lineages, so the success rate reported by Lahaye *et al.* is notable. They used primers described by Cuenoud *et al.* (2002) targeting a region up to approximately 900 bp in length in the middle of the gene (forward: 5′-CGATCTATTCATTCAA TATTTC-3′; reverse: 5′-TCTAGCACAC GAAAGTCGAAGT-3′). Further testing of these primers on a broader sample set for barcoding applications is now needed to assess whether the success rate of 100% is generalizable beyond the taxa examined here.

The other region favored by Lahaye *et al.* was *trnH-psbA*. This is one of the most rapidly evolving chloroplast spacers, and the study by Kress and Erickson (2007) also highlighted the potential power of this intergenic spacer as a barcoding locus. Direct comparative evaluation of these regions, with the full set of other recently proposed candidate barcoding loci in Table 1, is now a priority to enable a standard barcoding solution to be agreed in plants.

Resolving power of plant DNA barcodes

Using *matK* alone, or in combination with *trnH-psbA*, Lahaye *et al.* reported that over 90% of species could be discriminated (multiple individuals of species resolved as monophyletic). This figure is based on the 44 species from the Kruger National Park and Costa Rica from which multiple accessions were sampled. This is an encouragingly high success rate for plant barcoding using only organelle genes. However, this involves many comparisons in which just a single species was sampled from a genus or family, and where multiple congeneric species have been sampled, it does not necessarily include the closest sister species. Although the ability to distinguish among species in a restricted sample set has many potential applications, a desirable trait for DNA barcoding is to be able to distinguish among the different species within a genus. This is the performance measure that perhaps most will be interested in. Re-examination of the data shows a total of 17 genera from which multiple species (2 or 3) have been sampled. On the basis of the Unweighted Pair Group Method with Arithmetic mean (UPGMA) analysis of *matK* (Lahaye *et al.*, Supplementary Figure S1), species level discrimination was achieved in 10/17 genera (reciprocal monophyly of species where multiple conspecific individuals were sampled, and non-zero length branches between samples where just single individuals represent species). In the seven other genera, there were examples of non-monophyletic species topologies or identical sequences shared between species.

The discriminatory abilities of *matK* was followed up in the second part of the Lahaye *et al.* paper, which describes the first published application of plant DNA barcoding for inventory work in a floristic hot spot. The authors generated and compiled *matK* sequences from an impressive data set of 1566 specimens representing 1084 orchid species from Mesoamerica. The sequences were used to see if a 'barcode gap' is present in plants (a discontinuity between intra- and interspecific variation). There was, as expected, greater interspecific than intraspecific sequence divergence. However, there were more than 500 interspecific comparisons with zero differences between species, and no clear discontinuity between intra- and interspecific divergences. The UPGMA tree of these data (Lahaye *et al.*, Supplementary

Table 1 Plant barcoding regions proposed at the Second International Barcode of Life Conference

Research group	Proposed plant barcode
Chase *et al.* (consortium led by Royal Botanic Garden Kew, UK)	*rpoC1+rpoB+matK* or *rpoC1+matK+trnH-psbA*
Kim *et al.* (Korea University, Korea)	*matK+atpF-H+psbK-I* or *matK+atpF-H+trnH-psbA*
Kress and Erickson (Smithsonian Institute, USA)	*rbcL+trnH-psbA*

Figure S2) also illustrates the high frequency with which species cannot be distinguished with *matK*, especially when multiple congeneric species are considered. Species-level discrimination in this larger data set is much lower than the 90% reported for the smaller data set described above. The UPGMA tree is replete with examples of identical sequences shared between species (and genera), and a lack of reciprocal monophyly for species with multiple accessions sampled.

Of course, in undertaking biodiversity inventory work in species-rich hot spots, there is no 'perfect' taxonomy to serve as a baseline for performance measures. Lack of coincidence of *matK* sequence clusters with species boundaries may reflect problems with DNA barcoding in the group in question (either recent divergence or hybridization as biological causes, or contamination when carrying out large-scale molecular surveys). However, it may also, in part, be attributable to the current taxonomy needing updating. Lahaye *et al.* noted high levels of divergence among accessions of one particular orchid species. The divergent sequences coincided with morphological and geographical differences and represent an example of barcoding approaches identifying potential cryptic species warranting further taxonomic investigation.

Although both the final choice of a barcoding region and the percentage of plant species that will be distinguishable by organelle barcoding remain to be determined, this study provides useful data toward these topics. The authors also report how even 'genus-level' resolution from DNA barcoding can have practical applications. *MatK* sequences were able to distinguish samples of the orchid genus *Phragmipedium*, from 1500 samples of other Mesoamerican orchids. All *Phragmipedium* species are listed on Convention on International Trade in Endangered Species (CITES) Appendix 1 (trade completely forbidden). Being able to distinguish these, from orchid species for which trade is permissible with permits, provides a simple practical example of how the methodology could be used by customs agencies to assess the legitimacy of samples without needing specialist knowledge of orchid biology.

PM Hollingsworth is at the Royal Botanic Garden, 20 Inverleith Row, Edinburgh EH3 5LR, UK.

e-mail: P.Hollingsworth@rbge.org.uk

Chase MW, Cowan RS, Hollingsworth PM, van den Berg C, Madriñán S, Petersen G *et al.* (2007). A proposal for a standardised protocol to barcode all land plants. *Taxon* **56**: 295–299.

Cuenoud P, Savolainen V, Chatrou LW, Powell M, Grayer RJ, Chase MW (2002). Molecular phylogenetics of Caryophyllales based on nuclear 18S rDNA and plastid *rbcL*, *atpB*, and *matK* DNA sequences. *Am J Bot* **89**: 132–144.

Hajibabaei M, Singer GAC, Hebert PDN, Hickey DA (2007). DNA barcoding: how it complements taxonomy, molecular phylogenetics and population genetics. *Trends Genet* **23**: 167–172.

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.

Lahaye R, van der Bank M, Bogarin D, Warner J, Pupulin F, Gigot G *et al.* (2008). DNA barcoding the floras of biodiversity hotspots. *Proc Natl Acad Sci USA* **105**: 2923–2928. http://www.pnas.org/cgi/content/abstract/105/8/2923?etoc.

Ledford H (2008). Botanical identities: DNA barcoding for plants comes a step closer. *Nature* **451**: 616.

Pennisi E (2007). Taxonomy. Wanted: a barcode for plants. *Science* **318**: 190–191.

Editor's suggested reading

Dasmahapatra KK, Mallet J (2006). DNA barcodes: recent successes and future prospects. *Heredity* **97**: 254–255.

Shiu S-H, Borevitz JO (2008). The next generation of microarray research: applications in evolutionary and ecological genomics. *Heredity* **100**: 141–149.

Meirmans PG, Den Nijs HJCM, Van Tienderen PH (2006). Male sterility in triploid dandelions: asexual females vs asexual hermaphrodites. *Heredity* **96**: 45–52.

McGrath S, Hodkinson TR, Barth S (2007). Extremely high cytoplasmic diversity in natural and breeding populations of *Lolium* (Poaceae). *Heredity* **99**: 531–544.