NEW TRENDS IN PLANT SYSTEMATICS

A proposal for a standardised protocol to barcode all land plants

Mark W. Chase[1], Robyn S. Cowan[1], Peter M. Hollingsworth[2], Cassio van den Berg[3], Santiago Madriñán[4], Gitte Petersen[5], Ole Seberg[5], Tina Jørgsensen[5], Kenneth M. Cameron[6], Mark Carine[7], Niklas Pedersen[7], Terry A.J. Hedderson[8], Ferozah Conrad[9], Gerardo A. Salazar[10], James E. Richardson[2], Michelle L. Hollingsworth[2], Timothy G. Barraclough[11], Laura Kelly[1] & Mike Wilkinson[12]

[1] *Jodrell Laboratory, Royal Botanic Gardens, Kew, Richmond, Surrey, TW9 3DS, U.K. m.chase@kew.org (author for correspondence)*
[2] *Royal Botanic Garden, Inverleith Row, Edinburgh, EH3 5LR, U.K.*
[3] *Departamento de Ciências Biológicas-DCBIO, Universidade Estadual de Feira de Santana-UEFS, BR116 Km3 Campus Universitário, Feira de Santana – BA, CEP 44031-460, Brazil*
[4] *Departamento de Ciencias Biológicas, Universidad de los Andes, Apartado Aéreo 4976, Bogotá, D.C., Colombia*
[5] *The Natural History Museum of Denmark, Sølvgade 83, Opg. S, 1307 Copenhagen K, Denmark*
[6] *The Lewis B. and Dorothy Cullman Program for Molecular Systematics Studies, The New York Botanical Garden, Bronx, New York 10458-5126, U.S.A.*
[7] *Department of Botany, The Natural History Museum, Cromwell Road, London SW7 5BD, U.K.*
[8] *Bolus Herbarium, Department of Botany, University of Cape Town, 7701 Rondebosch, South Africa*
[9] *Leslie Hill Molecular Systematics Laboratory, Kirstenbosch Research Centre, Private Bag X7, Claremont 7735, Cape Town, South Africa*
[10] *Departamento de Botánica, Instituto de Biología, Universidad Nacional Autonóma de México, Apartado Postal 70-367, 04510 México, D.F., México*
[11] *Imperial College London and Royal Botanic Gardens, Kew, Division of Biology, Silwood Park Campus, Ascot, Berkshire SL5 7PY, U.K.*
[12] *Institute of Biological Sciences, Edward Llwyd Building, The University of Wales, Aberystwyth, Ceredigion, SY23 3DA, U.K.*

We propose in this paper to use three regions of plastid DNA as a standard protocol for barcoding all land plants. We review the other markers that have been proposed and discuss their advantages and disadvantages. The low levels of variation in plastid DNA make three regions necessary; there are no plastid regions, coding or non-coding, that evolve as rapidly as mitochondrial DNA generally does in animals. We outline two, three-region options, (1) *rpoC1*, *rpoB* and *matK* or (2) *rpoC1*, *matK* and *psbA-trnH* as viable markers for land plant barcoding.

KEYWORDS: ITS nrDNA, land plant barcoding, *matK*, mitochondrial DNA, plastid DNA, *psbA-trnH*, *rbcL, rpoC1, rpoB, trnL*

At present there is no standard protocol for DNA barcoding of land plants. This is mainly because the DNA region being used as the official barcode for animals, a portion of the mitochondrial gene *coxI* (sometimes known as CO1), will not succeed due to generally low levels of variability in the mitochondrial DNA of land plants. The aim of this paper is to keep the plant barcoding community abreast of ongoing research towards finding a standard DNA barcoding protocol for land plants. We describe the attributes of the plant barcoding regions that have been proposed to-date and suggest two options for standardising on a plant DNA barcoding protocol. One option is to use portions of three plastid genes, *rpoC1*, *rpoB* and *matK;* the other is to combine portions of two plastid genes, *rpoC1* and *matK*, and a plastid intergenic spacer, *psbA-trnH*. Protocols and primers for these two options can be found at: http://www.rbgkew.org.uk/barcoding/update.html. The reasons why both of these options involve three DNA regions and why we offer two options instead of one are discussed below.

The slow evolutionary rate of plant mitochondrial DNA is well known (with some notable exceptions; Palmer, 1992; Palmer & al., 2000; Parkinson & al., 2005; Bakker & al., 2006), so when zoologists selected a mitochondrial gene, *coxI* (or CO1), as the standard barcode for animals (Hebert & al., 2003a, b), it was clear from the outset that an alternative solution would be required for plants. The closest equivalent source of a plant barcoding region is the

plastid genome. This genome shares many of the desirable attributes of animal mitochondrial DNA for barcoding, such as conserved gene order and high copy number in each cell enabling easy retrieval of DNA for PCR and sequencing. One problem with plastid DNA, however, is its generally slow rate of evolution, and the challenge has been to find a plastid region that is sufficiently variable for DNA barcoding. A suitable region should ideally show enough variation within it to discriminate among species, yet be conserved enough to be present and routinely retrievable across the > 400 million years of evolutionary divergence represented by extant land plant diversity. This is a non-trivial problem; finding a marker (or perhaps set of markers) for which primer binding sites are conserved but which shows high levels of variability across all groups of land plants represents a set of contradictory targets. If markers have highly conserved primer binding-sites, they tend to also be internally more conservative, whereas for the most variable regions it is difficult to identify sites for reasonably conserved primers.

An additional desirable trait for a potential barcoding region is to have a reading frame so that the presence of nonsense substitutions could be used as a criterion to evaluate how good sequencing reactions/editing have been. This is of course the case for *coxI*, which is an exon (coding region). Although not specifically a requirement, it is viewed as desirable for the universal plant barcoding marker(s) to be coding regions so that alignment is easily accomplished (if necessary by conversion to amino acids), thus making the selected markers also useful for studies of phylogenetic relationships and molecular evolution. For a non-coding region (intron or intergenic spacer) to represent a viable alternative it is necessary for it to have (1) universal primers and standard PCR protocols, (2) consistently higher variation than coding regions, and (3) a non-complicated pattern of molecular evolution (see below for an example of one such problem).

The protein encoding plastid gene *rbcL* has been proposed as a potential plant barcode by several sets of researchers (Chase & al., 2005; Newmaster & al., 2006), usually in conjunction with one or more other markers. One benefit of this region is the large amount of existing information—there are more than 10,000 *rbcL* sequences already in GenBank (Chase & al., 2005; Newmaster & al., 2006). However, many of these are unvouchered or erroneously identified, and none has electropherogram trace files available, so all of these would have to be repeated to meet the standards for an official "DNA barcode" designation in GenBank. Furthermore, studies by Chase & al. (2005) and Newmaster & al. (2006), which demonstrated a fair degree of success in discriminating species, used nearly entire *rbcL* sequences (at least 1300 bp long). An ideal DNA barcoding region should be short enough to amplify from degraded DNA and analysed via single-pass sequencing. One possibility is to develop primer sets for short portions of this gene to produce a barcode of appropriate length, but our attempts to develop universal primers to achieve this have been unsuccessful to date.

Another plastid DNA region proposed is the non-coding *psbA-trnH* spacer (Kress & al., 2005; Shaw & al., 2007). This region is one of the most variable non-coding regions of the plastid genome in angiosperms in terms of having the highest percentages of variable sites (Shaw & al., 2007). This variation means that this inter-genic spacer can offer high levels of species discrimination (Kress & al., 2005; Shaw & al., 2007). However, there are enormous problems with alignment for this locus caused by high rates of insertions/deletions; alignment of the *psbA-trnH* spacer across most larger families of angiosperms is highly ambiguous. It appears that even within closely related taxa, great length differences exist, such that at greater taxonomic distances no shared sequence remains. Furthermore, in some groups of plants, the *psbA-trnH* spacer is exceedingly short (less than 300 bp; Kress & al., 2006; K. Cameron, unpubl.), whereas in others, such as orchids, it is much longer because it contains copies of *rpl22* and *rps19* (which makes it greater than 1,000 bp; Chang & al., 2006). The *rps19* gene or pseudogene is also found between *trnH* and *psbA* in maize, rice and wheat (Chang & al., 2006). In BLAST searches of monocot *psbA-trnH* sequences in GenBank, we have found that representatives of Commelinales, Dioscoreales, Liliales and Zingiberales also have a copy of *rps19* in this position, but we could not detect any *rps19* sequence in this spacer for representatives of Acorales or Alismatales. Chang & al. (2006) suggested that the *rps19-trnH* cluster was duplicated early in the evolution of monocots, but that in some monocots the copy of *rps19* positioned between *trnH* and *psbA* is apparently a truncated pseudogene. It is possible that bioinformatic solutions to deal with such enormous length variation and molecular evolution could be developed and implemented. However, the use of this region as the standard plant DNA barcode alone is problematical. Its long length in some species (> 1000 bp) represents a problem for retrieval from degraded tissue such as highly processed material (as in folk medicines) or from forensic samples. Furthermore, founding a massive plant barcoding database that at its core has sequence data unalignable between families clearly places massive constraints on the benefits that can be gained from future comparative analyses. A last point here is that although non-coding regions such as *psbA-trnH* are generally more rapidly evolving than genes, this is not always the case; in some groups of mosses, *matK* and *rpoC1* both contain more variable positions (T. Hedderson, unpubl.). This was also true for *matK* in *Crocus* and *Hordeum* (O. Seberg & G. Petersen, unpubl.). Shaw & al. (2007) made this point as well; these

comparisons of rates of change in plastid regions are only generalisations, and deviations are common.

The plastid intron in *trnL* has also been suggested as an appropriate region for DNA barcoding (Taberlet & al., 2007). However, resolution with this region is obviously far too low due to its slow rate of molecular evolution (Shaw & al., 2005). For a non-coding region, it is surprisingly conserved, perhaps due to its highly conserved secondary structure.

Aside from the three regions of plastid DNA that have been presented, the one other widely used and more variable region that has been proposed as a barcoding region is the internal transcribed spacer portions of nuclear ribosomal DNA or nrITS (Chase & al., 2005; Kress & al., 2005). ITS has a long record of use (Baldwin, 1992), and in most groups of flowering plants it has performed well as a phylogenetic marker. In nearly all cases, it has produced results similar to those found with plastid DNA, but it often has 3–4 times more variable sites that evolve up to four times more rapidly (van den Berg & al., 2000). For example, in a set of eight *Protea* species (Proteaceae) that span the basal nodes of the genus, we detected only eight variable positions in total from sequences of the plastid genes *rpoC1*, *rpoB*, and *matK*, and only four of these were unique to single accessions; only six positions were variable in *psbA-trnH* for this same set of species, and only three species had unique sequences. For nrITS in these same eight species, we recovered 16 variable positions and a unique sequence for each species (L. Valente & M. Chase, unpubl.). However, in yet other taxa, e.g., *Scalesia* (Asteraceae), no variation is found in any of the plastid regions investigated here or in nrITS (G. Petersen & O. Seberg, unpubl.). Clearly there will have to be special protocols if DNA barcoding is to be viable in certain groups of plants. Although not often noted, the same is true for certain groups of animals and perhaps some fungi.

Nuclear nrITS is also subject in most organisms to gene conversion/concerted evolution (Wendel & al., 1995; Chase & al., 2003), so that a single copy type is maintained (variation among the thousands of copies does occur, but one general consensus copy predominates). However, in some land plants, multiple copies are maintained, and even in some groups of angiosperms several divergent and still functional copies of ITS are routinely detected (Rapini & al., 2006). Presence of multiple, divergent copies makes nrITS unacceptable as a standard barcoding region across all land plants. In addition, nrITS also requires different PCR conditions and additives than the plastid regions selected (dimethyl sulfoxide, DMSO), so multiplexing PCR reactions with ITS and most plastid DNA regions is not possible. There are several groups of angiosperms for which nrITS can still make a valuable additional contribution as a "local barcode" when low levels of plastid DNA variation are encountered, but the problems related to its molecular evolution in many groups makes it undesirable to include nrITS in our proposal for a standardised protocol for all land plants.

Based on our assessment of the existing literature, there is no currently available, universally usable region of the nuclear genome, and because the overall rate of plastid DNA evolution in plants is much slower in general than mitochondrial DNA in most animal groups, there is no single plastid DNA marker, coding or non-coding, that can alone stand as the plant barcode in all groups of land plants. One solution to this problem is to use a combination of plastid regions that together represent a viable plant barcode.

The Alfred P. Sloan Foundation and the Gordon and Betty Moore Foundation (both of the U.S.A.) recently funded a project that compared performance of a range of plastid regions with the aim of finding suitable DNA barcoding regions that could be used as "the universal land plant barcoding protocol". This consortium project, centred at the Royal Botanic Gardens, Kew, but including researchers from 11 institutes in seven countries (Brazil, Colombia, Denmark, Mexico, South Africa, U.K., and U.S.A.), undertook first to screen more than 100 potential coding and non-coding plastid DNA regions to (1) identify those that could be amplified with a simple and standardised set of primers and protocols, and (2) evaluate the 5–6 most promising regions on a broad set of land plant taxa. To make comparisons of variation between loci quick and easy in the first phase of our project, DNA of 96 pairs of closely related, often sister taxa, from across the land plants was mixed together for PCR and then sequenced, which permitted us to estimate numbers of polymorphic sites by which these pairs differed. By doing comparisons of PCR success and levels of variation, we quickly narrowed our search down to those loci with the greatest potential for universality and variability.

The full results of this study will be published elsewhere, and analyses of the efficacy of intensively trialled regions in part 2 are still ongoing. However, as a "research update", we summarise the current state of play. Two plastid gene regions, partial *rpoC1* and *rpoB*, performed well as barcoding regions in terms of being amplifiable with a limited range of PCR conditions and primer sets and, although not particularly rapidly evolving, were able to discriminate among species in many groups of organisms. A third gene region, *matK*, showed much higher levels of sequence variation and provided better species discrimination, but work is still underway to improve PCR primer sets to enhance its 'universality'. The greatest level of species discrimination was achieved when all three regions are combined (*rpoC1*, *rpoB* and *matK*), and this represents one option as a standard DNA barcode for plants.

A second option we present is *rpoC1*, *matK,* and *psbA-trnH.* This option substitutes for the relatively conserved

coding region, *rpoB,* the previously mentioned, highly length-variable, non-coding intergenic spacer, *psbA-trnH.* The benefits to doing this are that additional species level resolution may be obtained, while at least part of the plant barcode (the sequences from *rpoC1* and *matK*) will be comparable and alignable across broad evolutionary distances. The downside of this approach relates to the introduction of bioinformatics challenges and problems with degraded tissue due to the variation and often larger size of the *psbA-trnH* spacer.

If a standardised protocol is to be adopted so that complete unknowns can be identified, the evolutionary dynamics of plastid DNA mean that a multi-locus procedure is necessary to be able to retrieve an accurate identification from the database. Of course, producing three regions will cost more than one, but there is no universally variable, single plastid DNA region that can serve this purpose. Furthermore, the general principle of having a multi-locus barcode has been accepted by CBOL (Consortium for the Barcode of Life; D. Schindel, pers. comm.). DNA sequencing costs are now low enough to make three loci feasible, and with future improvements costs will be even less of an issue. When several (8–12) appropriate, nuclear, low-copy regions have been identified and become feasible (Chase & al., 2005), multi-locus barcodes will be better able to deal with the biological complexities of species distinctions that many people worry will not be addressed by the use of just plastid or mitochondrial DNA. Until we reach this point with exploration and development of nuclear loci as barcodes, it is in the meantime important to get plants into the barcoding effort. Even imperfect systems, such as the ones proposed here, are likely to make a major impact on many areas of research and are sufficient for many applications. For example, to address the flora of a specific geographic region, a barcode need only deal with a limited number of the possible taxonomic entities that exist within a genus, and for these applications the currently proposed methods are highly successful (K. Cameron & al., unpubl); sometimes even a single, relatively conserved DNA region will work well (Taberlet & al., 2007).

Any proposal for assignment of the keyword "barcode" to a sequence region in GenBank must be accepted by CBOL before GenBank will accord this recognition. Thus far, only *coxI* has that designation, but it is clear that plants necessitate several loci and thus a different approach.

Methods and protocols for the regions selected can be found on the Royal Botanic Gardens, Kew, website (www.rbgkew.org.uk/barcoding/rationale.html). Primer sequences for some taxa are still under development because those currently in use are not as robust and broadly applicable as we feel they should be. Nonetheless, progress has been significant enough that we are able to say that, providing these remaining technical issues can be overcome, the two, three-region options, (1) *rpoC1*, *rpoB* and *matK,* or (2) *rpoC1*, *matK* and *psbA-trnH*, represent the best viable options for the use of plastid DNA as barcodes for all land plants. A proposal presenting these two options as candidate barcoding regions for plants and exploring the strengths and weaknesses of the two will be submitted to CBOL shortly.

LITERATURE CITED

Bakker, F.T., Breman, F. & Merckx, V. 2006. DNA sequence evolution in fast evolving mitochondrial DNA *nad1* exons in Geraniaceae and Plantaginaceae. *Taxon* 55: 887–896.

Baldwin, B.G. 1992. Phylogenetic utility of the internal transcribed spacers of nuclear ribosomal DNA in plants: an example from the Compositae. *Molec. Phylog. Evol.* 1: 3–16.

Chang, C.-C., Lin, H.-C., Lin, I.-P., Chow, T.-Y., Chen, H.-H., Chen, W.-H., Cheng, C.-H., Cheng, C.-Y., Lin, C.-Y., Liu, S.-M., Chang, C.-C. & Chaw, S.-M. 2006. The chloroplast genome of *Phalaenopsis aphrodite* (Orchidaceae): comparative analysis of evolutionary rate with that of grasses and its phylogenetic implications. *Molec. Biol. Evol.* 23: 279–291.

Chase, M.W., Knapp, S., Cox, A.V., Clarkson, J.J., Butsko, I.Y., Joseph, J., Savolainen, V. & Parokonny, A.S. 2003. Molecular systematics, GISH and the origin of hybrid taxa in *Nicotiana* (Solanaceae). *Ann. Bot.* 92: 107–127.

Chase, M.W., Salamin, N., Wilkinson, M., Dunwell, J.M., Kesanakurthi, R.P., Haider, N. & Savolainen, V. 2005. Land plants and DNA barcodes: short-term and long-term goals. *Philos. Trans., Ser. B* 360: 1889–1895.

Hebert, P.D.N., Cywinska, A., Ball. S.L. & De Waard, J.R. 2003a. Biological identifications through DNA barcodes. *Proc. Roy. Soc. London, Ser. B, Biol. Sci.* 270: 313–321.

Hebert, P.D.N., Ratnasingham, S. & De Waard, J.R. 2003b. Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related animals. *Proc. Roy. Soc. London, Ser. B, Biol. Sci.* 270: S96–S99.

Kress, W.J., Wurdack, K.J., Zimmer, E.A., Weigt, L.A. & Janzen, D.H. 2005. Use of DNA barcodes to identify flowering plants. *Proc. Natl. Acad. Sci. U.S.A.* 102: 8369–8374.

Newmaster, S.G., Fazekas, A.J. & Ragupathy, S. 2006. DNA barcoding in land plants: evaluation of *rbcL* in a multigene tiered approach. *Canad. J. Bot.* 84: 335–341.

Palmer, J.D. 1992. Mitochondrial DNA in plant systematics: applications and limitations. Pp. 36–49 in: Soltis, P.S., Soltis, D.E. & Doyle, J.J. (eds.), *Molecular Systematics of Plants*. Chapman and Hall, New York.

Palmer, J.D., Adams, K.L., Cho, Y., Parkinson, C.L. & Song, K. 2000. Dynamic evolution of plant mitochondrial genomes: mobile genes and introns and highly variable mutation rates. *Proc. Natl. Acad. Sci. U.S.A.* 97: 6960–6966.

Parkinson, C.L., Mower, J.P., Qiu, Y.-L., Shrik, A.J., Song, K., Young, N.D., de Pamphilis, C.W. & Palmer, J.D. 2005. Multiple major increases and decreases in mitochon-

drial substitution rates in plant family Geraniaceae. *BMC Evol. Biol.* 5: 73.

Rapini, A., Chase, M.W. & Konno, T.U.P. 2006. Phylogenetics of the New World Asclepiadeae (Apocynaceae). *Taxon* 55: 119–124.

Shaw, J., Lickey, E.B., Beck, J.T., Farmer, S.B., Liu, W., Miller, J., Siripun, K.C., Winder, C.T., Schilling, E.E. & Small, R.L. 2005. The tortoise and the hare II: relative utility of 21 noncoding chloroplast DNA sequences for phylogenetic analysis. *Amer. J. Bot.* 92: 142–166.

Shaw, J., Lickey, E.B., Schilling, E.E. & Small, R.L. 2007. Comparison of whole chloroplast genome sequences to choose noncoding regions for phylogenetic studies in angiosperms. The tortoise and the hare III. *Amer. J. Bot.* 94: 275–288.

Taberlet, P., Coissac, E., Pompanon, F., Gielly, L., Miquel, C., Valentini, A., Vermat, T., Corthier, G., Brochmann & Willerslev, E. 2007. Power and limitations of the chloroplast *trnL* (UAA) intron for plant DNA barcoding. *Nucleic Acid. Res.* 35: e14.

Van den Berg, C., Higgins, W.E., Dressler, R.L., Whitten, W.M., Soto Arenas, M.A., Culham, A. & Chase, M.W. 2000. A phylogenetic analysis of Laeliinae (Orchidaceae) based on sequence data from nuclear internal transcribed spacers (ITS) of ribosomal DNA. *Lindleyana* 15: 96114.

Wendel, J.F., Schnabel, A. & Seelanan, T. 1995. Bidirectional interlocus concerted evolution following allopolyploid speciation in cotton (*Gossypium*). *Proc. Natl. Acad. Sci. U.S.A.* 92: 280–284.