PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B

Phil. Trans. R. Soc. B (2005) **360**, 1889–1895
doi:10.1098/rstb.2005.1720
Published online 12 September 2005

Land plants and DNA barcodes: short-term and long-term goals

Mark W. Chase[1,*], Nicolas Salamin[2], Mike Wilkinson[3], James M. Dunwell[3], Rao Prasad Kesanakurthi[3], Nadia Haidar[3] and Vincent Savolainen[1]

[1]*Jodrell Laboratory, Royal Botanic Gardens, Kew, Richmond, Surrey TW9 3DS, UK*
[2]*Department of Ecology and Evolution, University of Lausanne, 1015 Lausanne, Switzerland*
[3]*School of Plant Sciences, University of Reading, Reading RG6 6AS, UK*

Land plants have had the reputation of being problematic for DNA barcoding for two general reasons: (i) the standard DNA regions used in algae, animals and fungi have exceedingly low levels of variability and (ii) the typically used land plant plastid phylogenetic markers (e.g. *rbcL*, *trnL-F*, etc.) appear to have too little variation. However, no one has assessed how well current phylogenetic resources might work in the context of identification (versus phylogeny reconstruction). In this paper, we make such an assessment, particularly with two of the markers commonly sequenced in land plant phylogenetic studies, plastid *rbcL* and internal transcribed spacers of the large subunits of nuclear ribosomal DNA (ITS), and find that both of these DNA regions perform well even though the data currently available in GenBank/EBI were not produced to be used as barcodes and BLAST searches are not an ideal tool for this purpose. These results bode well for the use of even more variable regions of plastid DNA (such as, for example, *psbA-trnH*) as barcodes, once they have been widely sequenced. In the short term, efforts to bring land plant barcoding up to the standards being used now in other organisms should make swift progress. There are two categories of DNA barcode users, scientists in fields other than taxonomy and taxonomists. For the former, the use of mitochondrial and plastid DNA, the two most easily assessed genomes, is at least in the short term a useful tool that permits them to get on with their studies, which depend on knowing roughly which species or species groups they are dealing with, but these same DNA regions have important drawbacks for use in taxonomic studies (i.e. studies designed to elucidate species limits). For these purposes, DNA markers from uniparentally (usually maternally) inherited genomes can only provide half of the story required to improve taxonomic standards being used in DNA barcoding. In the long term, we will need to develop more sophisticated barcoding tools, which would be multiple, low-copy nuclear markers with sufficient genetic variability and PCR-reliability; these would permit the detection of hybrids and permit researchers to identify the 'genetic gaps' that are useful in assessing species limits.

Keywords: BLAST; molecular taxonomy; plant DNA barcoding; phylogenetics; population genetics

1. INTRODUCTION

Efforts to produce DNA barcodes (the so-called 'DNA-taxonomy' of Tautz *et al.* 2003) are proceeding apace for animals and fungi using a standard DNA region, the mitochondrial *cox1* gene (which codes for subunit 1 of cytochrome oxidase). It appears that this region can also be used in at least some groups of 'algae' (Saunders 2005; note that algae, even those that are multi-cellular, belong to various distantly related clades), but in land plants *cox1* sequences are highly invariant and therefore unsuitable for use as DNA barcodes. Alternative single regions in the plastid genome have seen wide use in phylogenetic studies (e.g. exons such as *rbcL*, *atpB*, *ndhF* and *matK* and non-coding regions such as the *trnL* intron and *trnL-F* intergenic spacer), but these have 'appeared' not to be variable enough to be useful as barcodes because in phylogenetic studies results from individual loci have been highly unresolved due to too few phylogenetically informative sites. Another commonly sequenced region for land plant phylogenetic studies is nuclear ribosomal ITS (the internal transcribed spacers of the large subunit of ribosomal DNA), but this region does not work well in at least some groups of plants due to problems of paralogy and other factors associated with the complex concerted evolution of this highly repeated part of the nuclear genome. Thus, to at least some workers, barcoding in land plants with an approach similar to that being used in other organisms (employing *cox1*) has appeared to be 'on hold' until more appropriate plastid markers have been developed (but see Kress *et al.* 2005). However, lack of utility in a phylogenetic context does not necessarily mean that standard phylogenetic markers could not function well as identity codes; in the latter case, autapomorphies (unique single substitutions) are important whereas in

* Author for correspondence. (m.chase@kew.org).

One contribution of 18 to a Theme Issue 'DNA barcoding of life'.

the former these are uninformative. Therefore, the first task is to make an attempt to see how such markers would perform in the altered context of DNA barcoding. We provide here an assessment of autapomorphies in several plastid DNA regions for two large South African genera, *Moraea* (approximately 200 species; Goldblatt *et al.* 2003) and *Protea* (approximately 85 species; Reeves *et al.* in press). Then, in a second phase, we provide an estimate of the utility of the two most abundant plant sequences in GenBank/EBI, the plastid gene *rbcL* and nuclear ribosomal ITS, using the BLAST procedure (Altschul *et al.* 1990) to make the 'identification'. These DNA sequences and BLAST were never intended to be used in this manner, but they should provide some insights into how well we can expect perhaps more appropriate plastid DNA regions to perform as barcodes.

2. USERS OF PLANT BARCODES

There are two general categories of potential users of DNA barcodes: plant taxonomists/systematists, who wish to use these methods/markers as tools to elucidate species limits, and scientists in other fields, who are 'end users' of taxonomic concepts developed by taxonomists/systematists. For the latter category, there is an urgent need to establish at least a crude system of barcoding, and for this purpose plastid DNA regions are perfectly suited. By 'crude', we mean an easily developed but sometimes coarse system that is based on a uniparentally inherited marker, which makes it a less than perfect system. In some, perhaps many instances, this sort of marker will not provide an accurate identification, but there is still a great deal of utility in developing such a system. Many applications require DNA markers that can be easily amplified from degraded DNA samples, particularly in forensics and economic uses, such as traditional-drug authentication efforts. However, the incidence of hybridization, introgression and (allo)polyploidy in land plants is well documented, and to improve the taxonomic base upon which DNA barcoding efforts rest there is also a need to assess variation in multiple nuclear DNA regions. This also applies to algae, animals and fungi, although perhaps to a lesser degree due to a lower incidence of hybridization in such groups compared to higher plants.

Population genetic studies typically have large numbers of freshly collected specimens at their disposal, so DNA quality is a lesser concern, and it is upon such high-quality DNA samples that more accurate barcoding techniques would depend. To improve species concepts, we need to develop a more sophisticated approach to barcoding, which would ideally include sequences from multiple (perhaps six to eight) independent markers, a multi-locus barcode, and specific inference tools that could be used to explore species limits and identify genetic 'gaps'. This second type of barcode would improve the information base upon which the cruder plastid and mitochondrial DNA barcodes depend. We will in the last part of this paper propose and describe in more detail our vision of how such a system could be developed.

3. AN ASSESSMENT OF UNIQUENESS OF CURRENTLY USED PHYLOGENETIC MARKERS

To be used in phylogenetic studies, markers must exhibit sufficient variability to link species and groups of species by possessing shared (synapomorphic) substitutions; unique substitutions or autapomorphies are not used in assessing phylogenetic relationships of species and other taxa (but note that they are used in dating of phylogenetic trees, i.e. in molecular clock studies, and establishing overall genetic distances between species). Therefore it is not appropriate to determine utility of such markers for use in DNA barcoding efforts by comparing how well individual loci perform in phylogenetic studies; lack of resolution (the production of many equally optimal trees) caused by low levels of informative characters is not a useful measure when evaluating markers for use as barcodes, which requires unique substitutions that provide 'species markers'. Therefore, we provide here an assessment of uniqueness in several markers commonly used in phylogenetic studies. We selected for this purpose two large genera emblematic of the flora of South Africa (two of the 34 global diversity hotspots, Cape and Succulent Karoo): (i) *Moraea* (peacock irises, approximately 250 species; Iridaceae) studied by Goldblatt *et al.* (2003) using three plastid DNA regions, the *rbcL* gene, the *trnL-F* intron/intergenic spacer and the *rps16* intron; and (ii) *Protea* (proteas, approximately 85 species; Proteaceae) studied by Reeves *et al.* (in press) using three plastid genes and a low-copy nuclear region, glutamine synthetase (the copy expressed in plastids, *ncpGS*; Emshwiller & Doyle 2002). We determined the degree to which species could be separated by unique changes (table 1) and found that in *Moraea*, which is approximately 25 million years old (Goldblatt *et al.* 2003), even the 'relatively conserved' *rbcL* exon exhibited sufficient numbers of autapomorphies to separate more than 99% of the 170 species in our DNA matrices. In contrast, for *Protea*, which is of similar age (Reeves *et al.* in press), these same sorts of standard plastid markers did not fare so well (table 1): the best plastid region exhibited enough unique variation to separate only 65% of the 82 species in our matrices. However, the fragment of *ncpGS* sequenced, which contains three introns (about 80% of its length), separated >99% of the species. There are at least two factors that could make these statistics less meaningful than they appear: some of the unique changes could be sequencing errors (which by some estimates, Kristensen *et al.* 1992, are as high as 4%, a figure that we would dispute in this case) and bad taxonomy, such that variants of the same biological entity have been given two or more names. These two phenomena would to a degree compensate for each other. These studies did not include many accessions of the same species, so we cannot assess the degree of intraspecific variation for these markers. Thus it appears that in the case of *Moraea* a single plastid marker would be highly useful as a tool for barcoding whereas in *Protea* two or more plastid markers would be required; *ncpGS*, however, could be successful on its own for *Protea*. In any case, this result demonstrates that although not sufficient as phylogenetic markers these DNA regions contain

Table 1. Probability (p) of identifying the correct species based on DNA sequences.
(Pair-wise distance matrices of absolute numbers of differences were computed using PAUP* 4.0b10 (Swofford 2001). Note that the probability of identifying the correct species was calculated as the proportion of comparisons in which at least one nucleotide difference was found between species pairs (in practice we would aim at targeting genes that have more than just one nucleotide difference between species).)

taxa	DNA regions (number of base pairs, bp)	p
Moraea ($n=170$)	*trnL-F* intron/spacer (1229 bp)	>0.99
	rbcL (1354 bp)	>0.99
	rps16 intron (992 bp)	>0.99
Protea ($n=88$)	*trnL-F* (1074 bp)	0.60
	atpB-rbcL spacer (842 bp)	0.85
	rps16 intron (832 bp)	0.96
	ncpGS (854 bp)	>0.99

unique changes that could serve as DNA barcodes (provided that as follow-up studies intraspecific variation would be assessed to determine species limits).

In a second case study, we utilized the two most abundant DNA sequences for plants in GenBank/EBI, plastid *rbcL* (6 741 sequences) and nuclear ribosomal ITS (total 33 508, some treating ITS1, 5.8S and ITS2 as separate entries). All available entries for these two DNA regions were extracted from the Euphyllophyta dataset in GenBank release 144. In turn, each ITS and *rbcL* entry was used as a query for a BLAST search against the entire Euphyllophyta GenBank dataset. First, the percentages of identity, as returned by the BLAST algorithm (Altschul *et al.* 1990), were calculated between and within several taxonomic levels (species, genus, family and order). The genera were assigned to families and orders following the APGII classification (APG 2003; list of genera available from MWC and VS), but those genera present in GenBank but not recognized by APGII were removed from the BLAST analyses. Second, for each query sequence, the percentages of incorrect assignment were calculated at the genus and species level. These proportions represent the number of BLAST hits not identical to the query at the genus or species level but with a better hit score than the first and last correct hit for the query. At the species level, we examined the number of best hits before finding with BLAST the same sequence used as input for the search, plus any other entries of the same species. At the genus level, we looked at all entries for a target genus and recorded the number of incorrect hits (i.e. species belonging to other genera) that appeared above the lowest ranked entry of a species correctly assigned to the genus. GenBank/EBI accessions for ITS contain either the complete ITS region or part of it. The results for this DNA region were therefore split according to which part represented the query sequence (table 2).

One can imagine several reasons why this procedure is a less than ideal measure of the barcode potential of these markers, ranging from incorrect name assignments in GenBank/EBI to the use of BLAST as a tool for which it was never designed, and we did not expect this procedure to work well. Hence, evaluating the feasibility of barcoding with the available data and tools is likely to produce an unfavourable outcome; such methods should perform much better when the reference database is more complete, when intraspecific levels of variation have been determined and included in the procedure and when a more appropriate search tool is employed (e.g. string barcoding, the use of particular motifs or combinations of particular bases; Rash & Gusfield 2002; DeSalle *et al.* 2005). For ITS, the two more variable parts, ITS1 and ITS2, performed better than the more conserved 5.8S region (table 2). On average, with an ITS1 probe, the cluster of best BLAST hits formed by all sequences from the same species contained 6.79% of sequences from other species (table 2). At the genus level, the cluster of genus-specific sequences contained *ca* 40% of other genera (table 2), which is some cases could be the result of current generic limits being unnatural rather than BLAST not being able to discriminate between natural (i.e. monophyletic) genera. For ITS2, these percentages were of 33.70 and 51.68% for the species and genus level, respectively (table 2). However, ITS1 had a higher percentage identity at all taxonomic levels compared to ITS2. In our comparisons, the intraspecific levels of variation were not an important factor for correct assignment of species. Although *rbcL* sequences had higher levels of taxonomic fidelity (i.e. getting an appropriate match), the proportions of erroneous assignment were *ca* 17 and 68% for species and genus level, respectively (table 2). We are encouraged by the relatively high levels of the target species being in the highest BLAST categories. This bodes well for using markers similar to those already widely sequenced as phylogenetic markers as barcodes for land plants.

4. DNA BARCODING AS A TOOL FOR GENETIC DELIMITATION OF BIOLOGICAL SPECIES

Most potential users of DNA barcoding are not taxonomists (or systematists, which we here consider synonymous, although many would distinguish between these two, with the former being equivalent to nomenclaturists and the latter population and/or evolutionary biologists). These users in other fields need a quick, easily used and accurate system of identification, and in many cases a relatively crude diagnosis would be acceptable. The need for such a system is immediate and cannot wait for something more sophisticated to be developed. Taxonomists have worried that DNA barcoding will be less successful than it could be because species limits in many groups of organisms are merely statements of what we think rather than what we know, and therefore the old adage applies: 'rubbish in, rubbish out.' We think that taxonomists are overly critical of their work and focus on the gaps in their knowledge, which is in many ways admirable (users of taxonomic data would prefer to be told of a lacuna in knowledge rather than that it is perfect when it is not). Many of the situations in which barcodes would be applied can accept the application of a broad species concept (i.e. identification to an aggregated species, a group of species for which limits

Table 2. Percentage of identity, as measured by the BLAST algorithm, at different taxonomic levels (inter/intra), and proportion of incorrect assignment of sequences at the specific and generic level (see text for details).

			wrong assignment (%)	
DNA regions	taxonomic levels	identity (%)	first occurrence	last occurrence
ITS1+5.8S+ITS2	species	95.75/97.20	6.59	45.73
	genus	94.95/95.65	0.30	43.66
	family	94.35/95.24		
	order	92.98/93.12		
ITS1+5.8S	species	97.24/97.43	0.84	43.65
	genus	94.69/95.76	0	40.30
	family	93.02/94.27		
	order	92.07/92.54		
5.8S+ITS2	species	95.48/97.21	3.57	40.77
	genus	93.50/93.72	0.90	34.66
	family	91.38/93.21		
	order	90.75/91.79		
ITS1+ITS2	species	95.16/96.08	0.47	49.86
	genus	93.98/94.42	1.13	46.28
	family	93.20/92.03		
	order	91.86/92.09		
ITS1	species	96.43/96.72	1.11	6.79
	genus	93.39/95.77	0.02	39.64
	family	91.69/92.51		
	order	91.11/91.34		
ITS2	species	93.98/95.28	7.08	33.70
	genus	93.78/93.10	0.10	51.68
	family	91.79/89.46		
	order	88.89/89.24		
5.8S	species	97.18/99.80	32.76	62.94
	genus	96.30/98.45	36.04	64.62
	family	95.32/97.25		
	order	93.25/95.28		
rbcL	species	97.62/99.73	3.69	16.95
	genus	95.75/97.13	0.23	67.71
	family	91.43/95.92		
	order	88.81/92.65		

are not clear, or to one of a closely related set of species, a species complex, rather than a single species) and identification of a hybrid or introgressed plant as its maternal parent (because plastid DNA are maternally inherited in most plants) would not be hugely problematic on a practical level. Moreover, an ecologist trying to identify sterile plants, perhaps seedlings, in his plots or the forensic scientist trying to tie a vehicle to a particular location where a rare plant grows will not be overly concerned about the effects of hybridization, introgression or polyploidy on the results of the barcoding effort. This is not a denial that such phenomena are factors in the identification of species, but rather that a high degree of precision is not always required; just getting the field of possibilities narrowed to this extent provides such immense benefits that a degree of imprecision can be easily tolerated.

In contrast, a specialist working on a particular group of organisms (perhaps one for which the specialist is the world-recognised expert) worries a great deal about all the phenomena that can make identification of a particular accession problematic. The working taxonomist is concerned about the effects of hybridization, introgression and (allo)polyploidy, and some groups are notoriously difficult in these respects (e.g. *Nicotiana* in Solanaceae, Chase *et al.* 2003; Clarkson *et al.* 2004; and some orchids, e.g. *Dactylorhiza*, Pillon *et al.* in press), so it is not surprising that many taxonomists have viewed the issue of DNA barcoding with a great deal of suspicion and scepticism. Even if they were willing to recognize that it could be done, they are simultaneously critical of how good the results would be simply due to the ambiguities of species limits that they know exist within their groups of particular interest and hence by extrapolation to all others as well. DNA barcodes based on uniparentally inherited markers can never reflect the complexity that exists in nature, and many taxonomists have by and large ignored or been highly critical of the barcoding movement as a waste of time and money with at best the prospect of dubious results.

However, we can easily imagine a more sophisticated barcoding technique that would provide taxonomists with a new tool to investigate species limits and identify the genetic gaps that result when gene flow has become negligible. An example of these gaps is presented by Richardson *et al.* (2003) for a group of *Phylica* species (Rhamnaceae) using AFLPs (a genetic fingerprinting technique not suitable for barcoding because of difficulties in inferring the homology of amplicons on the basis of shared fragment size and to a lesser extent the need to find PCR primers appropriate

for each group, i.e. because it cannot be made universal; Vos *et al.* 1995). Genetic gaps do not necessarily reflect species limits and hence cannot always be used as a guide to the application of names (i.e. there is not a 1 : 1 relationship of genetic gaps to taxonomic names), but knowing where such gaps exist is extremely useful information in the quest to make meaningful taxonomic decisions. Such a more accurate barcode would have to be based on highly variable markers in the nuclear genome, and development of such a method would have to utilize multiple loci (hybridization and introgression, for example, cannot routinely be diagnosed by sequencing only a single locus).

The need for this more sophisticated tool ('gold standard') should not be seen as an argument to delay implementation of single-locus barcoding ('silver standard') until a time when we have developed such methods. We advocate the immediate development in land plants of a silver standard system based on one or two plastid DNA regions plus perhaps ITS (in those groups for which it has been demonstrated to work well, e.g. orchids). This system would serve the needs of the wider scientific community that needs rapid and reasonably accurate identification of unknowns. While this phase is being implemented, another effort should be made to develop this more sophisticated gold standard method (see below).

We can imagine that in fact a two-step barcoding system with a traffic light approach might eventually be routinely used (we thank Kenneth Cameron, New York Botanical Garden for this idea). For many groups of organisms and many applications, the first step would be production of the uniparentally inherited barcode, and there the process would end with this crude answer. If the result was a name for which we know the taxonomy is simple and robust, then the name would be produced with a 'green light' beside it, meaning that this identification is uncomplicated and clear, whereas if a 'yellow light' appears then the identification is from a group that has some problems, and the user can decide if a more detailed investigation is necessary to satisfy the level of precision required. If the third situation is encountered, then a 'red light' appears, and the user would be informed that the identification is highly likely to be inaccurate because it belongs to a species complex in which phenomena such as hybridization are commonplace. In such a situation, the user could still stop there if there might be no need for a more accurate identification, but in many cases greater precision would be desirable, in which case the more diagnostic procedure could take place, provided that high quality DNA is available. Of course, for most taxonomists, the only reasonable barcode would be the one based on multiple nuclear DNA loci, the gold standard barcode.

5. DEVELOPMENT OF A NUCLEAR MULTI-LOCUS BARCODE AS A TOOL FOR INVESTIGATING SPECIES LIMITS

To develop this multi-locus barcode (MBC) system, there would need first to be an effort to identify conserved sites flanking regions containing variable sequences, most notably introns of appropriate size. These conserved sites would serve as universal PCR priming locations in all land plants and could enable amplification of these variable regions in products of an appropriate size for single-pass sequencing reactions. This phase of the project would take advantage of the completely sequenced genomes of *Arabidopsis, Populus* and *Oryza* and the EST libraries of a diverse range of plants that are now populating GenBank. An algorithmic approach to the identification of potential sites could be automated to develop over 1000 such sites that will then be evaluated on a set of DNA samples representing all major clades of land plants. For instance, Kozik & Michelmore (2003) identified a provisional list of over 3000 putative single copy genes for *Arabidopsis*. Although an unknown proportion of these candidates will inevitably prove to be incorrectly classified, we are in the process of refining this list to assemble a first set of primers for potential use on all higher plants. Testing these on DNA samples representing pairs of closely related species that have previously been studied with some of the standard phylogenetic markers, such as the *Moraea* and *Protea* studies described above, would provide evidence of the relative levels of sequence variability in the newly developed loci and permit selection of those appropriate for further testing. To be practical, these putatively low-copy markers would need to satisfy a set of requirements similar to those already widely discussed, particularly with respect to length, which is even more critical if such nuclear loci are to be routinely amplified from DNA samples of highly variable quality.

The development and subsequent application of the MBC as a tool to investigate species limits will utilize the large numbers of intact DNA samples that are routinely associated with the field of population genetics. Unlike plastid and nuclear rDNA regions that can be amplified from highly degraded DNA samples due to their highly iterative nature, the MBC method will be much more subject to failure to amplify from even moderately degraded DNA samples, but this should not be seen as a fatal flaw. This system would not be a replacement for the simpler, more robust, but less sophisticated methods. If six to eight appropriately variable regions are identified, then the MBC might consist of 50–100 base pairs of DNA sequence from each region, perhaps then sequenced with some of the newer techniques, such as pyrosequencing (Pacey-Miller & Henry 2003; Mochida *et al.* 2004) that do not produce such long sequences as do standard methods but do so with much greater speed.

The reason why the MBC has to be multi-locus is because detection of hybridization/introgression cannot be reliably accomplished by examination of a single nuclear DNA region. An F_1 hybrid on the other hand can be detected with DNA sequences from a single locus because it will be heterozygous at most if not all loci, but F_2 hybrids or backcrossed progeny (to one or the other parent) will be homozygous at many loci, so dependence on a single locus could be misleading.

It has been suggested that once barcodes have been established for most species a quick way to identify these taxa would be array-based rather than sequence

based. If one were studying seed germination and seedling establishment in set of previously studied forest plots in which all adult trees had been barcoded, then an array could be developed with dots for each species, and one could expect then to be able to see which species were present as bright spots on the array. If an ecologist were instead working on many previously unstudied plots, e.g. in the tropics, then it would be important to have a barcoding system from which it could be expected that species previously unknown from those plots could be identified. Ultimately, sequence-based barcodes will have greater power and more applications than array-based technologies, but we acknowledge that there may be specific high-throughput applications where an array-based approach offers a quick and practical solution.

6. A PLEA FOR DNA BANKING

To make as rapid progress as possible, nearly all early efforts to establish a set of reference standards have focused their attention on using herbarium or museum specimens (Kress *et al*. 2005, for example). It is impractical to expect, particularly in the short run, that a representative of each species on Earth can be newly collected. Such efforts would have to be global in scale, and there are many political as well as practical difficulties that make such efforts unrealistic. Therefore, reliance upon already collected material (often with highly degraded DNA) has been viewed as an imperative. This is one of the reasons why plastid and mitochondrial DNA have been viewed as the most appropriate regions to sequence: these highly repeated genomes are the most likely to survive reasonably intact within herbarium and museum specimens. In the rush to get a barcoding system up and running quickly, procurement of more intact materials has been ignored, but this is something to which more attention needs to be paid. In the course of work on barcoding, it should be possible to assemble DNA banks from those samples that are of higher quality. These high quality samples would be the basis upon which the MBC can be established. Given the high rates of global extinction, it is imperative that DNA banks of high-quality DNAs be established (Savolainen & Reeves 2004; Savolainen *et al*. in press). There are a number of large DNA banks already in operation, such as the one at the Royal Botanic Gardens, Kew, which currently holds 23 000 DNAs (www.rbgkew.org.uk/data/dnaBank/homepage.html). These DNAs are available to researchers worldwide, and many of these samples could become the standards used in the development of the MBC system for plants. The Royal Botanic Gardens, Kew, is willing to hold tissue and extract DNAs from plants collected anywhere in the world, with no charge to the collectors, and handling of these samples is done in accordance with the international Convention of Biological Diversity, so that the rights of countries of origin to profits derived from exploitation of their genetic resources are taken into account. Such partnerships have been already successfully put in place between the Royal Botanic Gardens, Kew, and South Africa under the umbrella of the UK Darwin Initiative for the Survival of Species. Before it is too late, DNA samples from as wide a range of organisms as possible should be assembled and curated. If, during the rush to barcode every organism, we lose sight of the need to document their genetic makeup, it would be most tragic. At the least, we seem to be well on our way to banking DNA from a large percentage of land plants, and this effort must be expanded as quickly as possible.

REFERENCES

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. 1990 Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410. (doi:10.1006/jmbi.1990.9999.)

Chase, M. W., Knapp, S., Cox, A. V., Clarkson, J. J., Butsko, Y., Joseph, J. S. V. & Parokonny, A. S. 2003 Molecular systematics, GISH and the origin of hybrid taxa in *Nicotiana* (Solanaceae). *Ann. Bot.* **92**, 107–127. (doi:10.1093/aob/mcg087.)

Clarkson, J. J., Knapp, S., Garcia, V., Olmstead, R. G. & Chase, M. W. 2004 Phylogenetic relationships in *Nicotiana* (Solanaceae) inferred from multiple plastid DNA regions. *Mol. Phylogenet. Evol.* **33**, 75–90. (doi:10.1016/j.ympev.2004.05.002.)

DeSalle, R., Egan, M. G. & Siddall, M. 2005 The unholy trinity: taxonomy, species delimitation and DNA barcoding. *Phil. Trans. R. Soc. B* **360**, 1905–1916. (doi:10.1098/rstb.2005.1722.)

Emshwiller, E. & Doyle, J. J. 2002 Origins of domestication and polyploidy in oca (*Oxalis tuberosa*: Oxalidaceae). 2. Chloroplast-expressed glutamine synthetase data. *Am. J. Bot.* **89**, 1042–1056.

Goldblatt, P., Savolainen, V., Porteous, O., Sostaric, I., Powell, M., Reeves, G., Manning, J. C. & Barraclough, T. G. 2002 Radiation in the Cape flora and the phylogeny of peocock irises *Moraea* (Iridaceae) based on four plastid DNA regions. *Mol. Phylogenet. Evol.* **25**, 341–360. (doi:10.1016/S1055-7903(02)00235-X.)

Kozik, A. & Michelmore, R. 2003 Graphical representation of BLAST search lettuce, sunflower, tomato, soybean, maize and rice ESTs against *Arabidopsis* genome. Potential conserved orthologs (single copy genes in Arabidopsis). Genes are web links to corresponding entries in CGP database. Available at: http://cgpdb.ucdavis.edu/COS_Arabidopsis/arabidopsis_single_copy_genes_2003.html. Last visited 20th April 2005.

Kress, J. W., 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. USA* **102**, 8369–8374. (doi:10.1073/pnas.0503123102.)

Kristensen, T., Lopez, R. & Prydz, H. 1992 An estimate of the sequencing error frequency in the DNA sequence databases. *DNA Sequences* **2**, 343–346.

Mochida, K., Yamazaki, Y. & Ogihara, Y. 2004 Discrimination of homoeologous gene expression in hexaploid wheat by SNP analysis of contigs grouped from a large number of expressed sequence tags. *Mol. Genet. Genom.* **270**, 371–377. (doi:10.1007/s00438-003-0939-7.)

Pacey-Miller, T. & Henry, R. 2003 Single-nucleotide polymorphism detection in plants using a single-stranded pyrosequencing protocol with a universal biotinylated primer. *Anal. Biochem.* **317**, 165–170. (doi:10.1016/S0003-2697(03)00089-7.)

Pillon, Y., Fay, M. F., Shipunov, A. B. & Chase, M. W. In press. Species diversity versus phylogenetic diversity: a practical study in the taxonomically difficult genus *Dactylorhiza* (Orchidaceae). *Biol. Cons.*

Rash, S. & Gusfield, D. 2002 String barcoding: uncovering optimal virus signatures. *6th Conference on Computational Biology*, pp. 254–261.

Reeves, G., Barraclough, T. G., Rebelo, T. G., Fay, M. F. & Chase, M. W. In press. Molecular phylogenetic of African Protea: evidence from DNA sequences and AFLP markers for a Cape origin. *Ann. Missouri Bot. Gard.*

Richardson, J. E., Fay, M. F., Cronk, Q. C. B. & Chase, M. W. 2003 Species delimitation and the origin of populations in island representatives of Phylica (Rhamnaceae). *Evolution* **57**, 816–827.

Saunders, G. W. 2005 Applying DNA barcoding to red macroalgae: a preliminary appraisal holds promise for future applications. *Phil. Trans. R. Soc. B* **360**, 1879–1888. (doi:10.1098/rstb.2005.1719.)

Savolainen, V. & Reeves, G. 2004 A plea for DNA banking. *Science* **304**, 1445. (doi:10.1126/science.304.5676.1445b.)

Savolainen, V., Powell, M. P., Davies, K., Cothals, A. & Reeves, G. (ed.). In press. *DNA banking for biodiversity and conservation*. Kew Publishing and IUCN.

Swofford, D. L. 2001 *PAUP* 4.0: Phylogenetic analysis using parsimony (* and other methods)*. Sunderland, Massachusetts: Sinauer Associates.

Tautz, D., Arctander, P., Minelli, A., Thomas, R. H. & Vogler, A. P. 2003 A plea for DNA taxonomy. *Trends Ecol. Evol.* **18**, 70–74. (doi:10.1016/S0169-5347(02)00041-1.)

Vos, P., Hogers, R., Bleeker, M., Rijans, M., Van de Lee, T., Hornes, M., Frijters, A., Pot, J., Kuiper, M. & Zabeau, M. 1995 AFLP: a new technique for DNA fingerprinting. *Nucl. Acids Res.* **23**, 4407–4414.