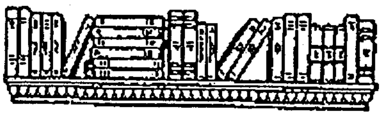


Molecular Ecology Resources (2009) 9, 439–457

doi: 10.1111/j.1755-0998.2008.02439.x

DNA BARCODING

Selecting barcoding loci for plants: evaluation of seven candidate loci with species-level sampling in three divergent groups of land plants

MICHELLE L. HOLLINGSWORTH,* ALEXANDRA CLARK,* LAURA L. FORREST,* JAMES RICHARDSON,* R. TOBY PENNINGTON,* DAVID G. LONG,* ROBYN COWAN,† MARK W. CHASE,† MYRIAM GAUDEUL‡ and PETER M. HOLLINGSWORTH*

**Royal Botanic Garden, 20 Inverleith Row, Edinburgh EH3 5LR, UK, †Jodrell Laboratory, Royal Botanic Gardens Kew, Richmond, TW9 3DS, UK, ‡Département Systématique et Evolution, Museum National d'Histoire Naturelle, 16 Rue Buffon, F-75005 Paris, France*

Abstract

There has been considerable debate, but little consensus regarding locus choice for DNA barcoding land plants. This is partly attributable to a shortage of comparable data from all proposed candidate loci on a common set of samples. In this study, we evaluated the seven main candidate plastid regions (*rpoC1, rpoB, rbcL, matK, trnH-psbA, atpF-atpH, psbK-psbI*) in three divergent groups of land plants [*Inga* (angiosperm); *Araucaria* (gymnosperm); *Asterella* s.l. (liverwort)]. Across these groups, no single locus showed high levels of universality and resolvability. Interspecific sharing of sequences from individual loci was common. However, when multiple loci were combined, fewer barcodes were shared among species. Evaluation of the performance of previously published suggestions of particular multilocus barcode combinations showed broadly equivalent performance. Minor improvements on these were obtained by various new three-locus combinations involving *rpoC1, rbcL, matK* and *trnH-psbA*, but no single combination clearly outperformed all others. In terms of absolute discriminatory power, promising results occurred in liverworts (e.g. *c.* 90% species discrimination based on *rbcL* alone). However, *Inga* (rapid radiation) and *Araucaria* (slow rates of substitution) represent challenging groups for DNA barcoding, and their corresponding levels of species discrimination reflect this (upper estimate of species discrimination = 69% in *Inga* and only 32% in *Araucaria;* mean = 60% averaging all three groups).

Keywords: *Araucaria*, *Asterella*, *Inga*, plant barcode

Received 17 June 2008; revision accepted 12 September 2008

Introduction

The working principle of DNA barcoding is the coordinated use of sequencing technologies to facilitate characterization of biodiversity (Hebert *et al.* 2003). In many animal groups, sequences of the mitochondrial cytochrome oxidase I gene (COI) provide species-level discrimination with potential for high throughput, automated identification of unknown samples when queried against an appropriately established reference set. The methodology can also contribute toward taxon discovery by highlighting samples with divergent sequences, which are then candidates for further taxonomic investigation.

Although DNA barcoding does not provide species-level resolution in all animal groups (e.g. Whitworth *et al.* 2007; Shearer & Coffroth 2008), it has been successful in many (e.g. Hebert *et al.* 2003, 2004; Smith *et al.* 2006), and a number of large-scale projects are underway in taxa such as birds, fishes and mosquitoes (http://www.barcoding.si.edu/major_projects.html). In plants, however, a lack of consensus on the most appropriate barcoding locus has impeded progress (Pennisi 2007; Ledford 2008). Compared to animals, land plant mitochondrial DNA has slower substitution rates and shows intramolecular recombination (Mower *et al.* 2007). This has impelled the search for alternative

Correspondence: Peter Hollingsworth, Fax: 44 (0) 131 248 2901; E-mail: P.Hollingsworth@rbge.org.uk

DNA barcoding regions outwith the mitochondrial genome (Kress *et al.* 2005; Chase *et al.* 2007).

The two most important traits of DNA barcoding loci are: (i) conserved flanking regions to enable routine amplification across highly divergent taxa; and (ii) sufficient internal variability to enable species discrimination. Additional factors to be considered include: (iii) length (short enough to routinely sequence, even in sub-optimal material); (iv) lack of heterozygosity enabling direct polymerase chain reaction (PCR) sequencing without cloning; (v) ease of alignment enabling the use of character-based data analysis methods; and (vi) lack of problematic sequence composition, such as regions with several microsatellites, that reduces sequence quality.

A section of plant DNA that fulfils all of these criteria has proved elusive. Table 1 summarizes the empirical studies published to date that have involved comparisons of multiple regions in a barcoding context. Also included is reference to two other large-scale unpublished comparative studies from which summary information is available. One of the first regions to be considered was the internal transcribed spacers (ITS) of nuclear ribosomal DNA (Chase *et al.* 2005; Kress *et al.* 2005). This is the most rapidly evolving 'off the shelf' region routinely used in plant molecular systematics. Although ITS works well in many plant groups and may be a useful supplementary locus, numerous cases of incomplete concerted evolution and intra-individual variation make it unsuitable as a universal plant barcode. The other regions proposed have been from the plastid genome and include a mixture of coding and noncoding regions.

From the broad pool of loci initially considered, the seven candidate loci that have emerged as front runners are sections of *rpoB*, *rpoC1* and *rbcL* (all conserved, easy-to-align coding regions), a section of *matK* (a rapidly evolving coding region, but with reported amplification problems), and *trnH-psbA*, *atpF-atpH* and *psbK-psbI* (three rapidly evolving but length variable intergenic spacers). Different research groups have proposed different combinations of these loci (some with mutually exclusive combinations). However, there is a shortage of published empirical studies comparing all regions on a common sample set. In this study, we provide such a comparison by evaluating performance of these seven candidate barcoding loci in three divergent groups of land plants. Specifically, we address the following questions:

1 Which of the proposed barcodes show the greatest universality?
2 Which of the proposed barcodes shows the greatest level of species discrimination?
3 What are the benefits in terms of species discrimination of using different combinations and different numbers of loci in a multilocus barcoding approach?
4 What percentage of plant species in these three groups of land plants can be discriminated by plastid barcoding?
5 Is there any evidence for a 'barcode gap' in plants (a discontinuity between intra- and interspecific sequence divergence)?

Table 1 Summary of studies comparing DNA barcoding regions in plants

Study	Regions compared	Sampling strategy	Universality (% success)	Sequence divergence/variation	Barcode recommendation
Kress *et al.* 2005	*atpB-rbcL*, ITS, *psbM-trnD*, *trnC-ycf6*, *trnH-psbA*, *trnL-F*, *trnk-rps16*, *trnV-atpE*, *rpl36-rps8*, *ycf6-psbM*	19 individuals/19 species from 7 angiosperm families	*trnH-psbA*, *rpl136-rpf8*, *trnL-F* = 100% *trnC-ycf6*, *ycf6-psbM* = 90% Other regions = 73–80%	% sequence divergences: ITS (2.81%) *trnH-psbA* (1.24%) *trnH-psbA* had ≈ 2 × sequence divergence of other plastid regions	ITS and *trnH-psbA*
	ITS, *rbcL**, *trnH-psbA*	83 individuals/83 species from 50 families	*trnH-psbA* = 100% *rbcL* = 95% ITS ≤ 88%	*trnH-psbA* >> *rbcL*	
Kress & Erickson 2007	*accD*, ITS1, *ndhJ*, *matk*, *trnH-psbA*, *rbcL*, *rpoB*, *rpoC1*, *ycf5*	96 individuals/96 species from 43 families of land plants	*trnH-psbA*, †*rbcL* = 95% †*rpoC1* = 90% *accD*, *rpoB* ≈ 80% *ndhJ* = 70%, ITS1 = 60% *ycf5* = 50% *matK* = 40%	ITS (5.7%) *trnH-psbA* (2.69%) *rpoB* (2.05%) *ycf5* (1.55%) *rpoC1* (1.38%) *rbcL* (1.29%) *accD* (1.2%) *matK* (1.13%) *ndhJ* (0.2%)	*rbcL* and *trnH-psbA*

Table 1 *Continued*

Study	Regions compared	Sampling strategy	Universality (% success)	Sequence divergence/variation	Barcode recommendation
Sass *et al.* 2007	*accD*, ITS, *ndhJ*, *matk*, *trnH-psbA*, *rpoB*, *rpoC1*, *ycf5*	21 individuals/21 species from 10 genera of cycads, more individuals for some regions (up to 96)	*rpoC1*, *ycf5* = 100% ITS = 100%?‡ *accD* = 96% *ndhJ* = 57% *rpoB* = 33% *matK* = 24% *trnH-psbA* double-banded in most samples	ITS most variable; quantitative figures on relative sequence divergence of other regions not given, other than *c.* 10% of bases variable in each region	ITS (considered most promising)
Newmaster *et al.* 2008	*accD*, *matK*, *trnH-psbA*, *rbcL*, *rpoB*, *rpoC1*, UPA	40 individuals/8 species in Myristicaceae	*trnH-psbA*, *rpoC1*, UPA = 100% *rpoB*, *accD*, ≥ 95% *rbcL* = 90% *matK* required primer redesign, then ≈ 98%	Uncorrected interspecific *p*-distances: *trnH-psbA* (0.060) *matK* (0.042) *accD* (0.003) *rpoC1* (0.002) *rbcL* (0.002) *rpoB* (0.001) UPA (0.001)	*matK* and *trnH-psbA*
Lahaye *et al.* 2008a	*accD*, *ndhJ*, *matK*, *trnH-psbA*, *rbcL*§, *rpoB*, *rpoC1*, *ycf5*	172 individuals/86 species total (consisting of 71 individuals/48 orchid species + 101 individuals/38 species from 13 angiosperm families)	In the orchids, *ycf5* did not amplify and *ndhJ* amplification was patchy. All other regions = 95–100%	K2P interspecific sequence divergence: *trnH-psbA* (0.0216) *matK* (0.0125) *ycf5* (0.01) *rbcL* (0.0079) *rpoB* (0.0061) *ndhJ* (0.0046) *accD* (0.0038) *rpoC*1 (0.0019)	*matK* (or *matK* and *trnH-psbA*)¶
Fazekas *et al.* 2008	*cox1*, 23S rDNA, *rpoB*, *rpoC1*, *rbcL*, *matK*, *trnH-psbA*, *atpF-atpH*, *psbK-psbI*	251 individuals/92 species from 32 genera of land plants	23S rDNA = 100% *rbcL* = 100% (2 primer pairs used) *trnH-psbA* = 99% *rpoC1* = 95% (3 primer pairs used) *rpoB* = 92% (5 primer pairs used) *matK* = 88% (10 primer pairs used) *psbK-psbI* = 85% *cox1* = 72% *atpF-atpH* = 65%	No. of parsimony-informative characters: *matK* (386) *trnH-psbA* (350) *atpF-atpH* (308) *psbK-psbI* (263) *rbcL* (242) *rpoB* (179) *cox1* (146) *rpoC1* (134) 23S rDNA (19)	Broadly equivalent performance from various combinations of loci; suggested selecting 3–4 regions from: *rbcL*, *rpoB*, *matK*, *trnH-psbA*, *atpF-atpH*
Chase *et al.* 2007	In preparation, empirical data currently unpublished				*rpoC1*, *rpoB* and *matK* or *rpoC1*, *matK* and *trnH-psbA*
Kim *et al.* cited in Pennisi 2007	In preparation, empirical data currently unpublished				*matK*, *atpF-atpH* and *psbK-psbI* or *matK*, *atpF-atpH* and *trnH-psbA*

*full *rbcL* sequences, rather than the shorter partial section proposed in more recent papers.

†corrected figure obtained from authors.

‡reported as 'amplified cleanly in most species'.

§*rbcL* not tested in the orchid samples.

¶Lahaye *et al.* recently posted 'online' results adding *atpF-atpH* and *psbK-psbI* to this comparison (Lahaye *et al.* 2008b) using just the 101 individuals/38 species data set; the preferred region reported following this analysis was *matK*.

Materials and methods

Sampling strategy

Various approaches have been taken to compare the performance of plant barcoding loci. The 'species pairs' approach involves taking pairs of related species from multiple phylogenetically divergent genera. This provides a sound assessment of universality of regions, but only limited insights into species-level resolution, as individual genera are not sampled in sufficient depth to provide assessments of the percentage of species that can be discriminated. The 'floristic' approach involves sampling multiple species within a given geographical area. This again can provide a sound assessment of universality and also represents an example of how barcoding might be applied in practice. One weakness, however, is that the 'floristic' approach inevitably includes samples of various levels of relatedness, but does not necessarily include the closest relatives of each species. An absence of sister-species sampling or multiple cases of single species sampled per genus may lead to overestimates of levels of species discrimination. Finally, the taxon-based approach involves sampling multiple species within a given taxonomic group. This provides limited insights into universality, but offers more definitive information on levels of species discrimination.

To date, the species pairs (e.g. Kress *et al.* 2005; Kress & Erickson 2007), the floristic (e.g. Fazekas *et al.* 2008; Lahaye *et al.* 2008a) and the 'taxon-based' (e.g. Newmaster *et al.* 2008) approaches to barcoding have all provided useful insights into the behaviour of varying combinations of barcoding loci. Our approach here combines wide phylogenetic coverage with taxon-based sampling within individual groups. We have selected a group of liverworts (*Asterella* P.Beauv.), a genus of flowering plants (*Inga* Mill), and a gymnosperm genus (*Araucaria* Juss.). Within each group, we sequenced 40–44 samples including multiple representatives of individual species. As levels of species discrimination were notably higher for the liverworts than for *Inga* or *Araucaria*, we screened a total of 98 individuals from 39 species for the best performing barcoding loci for this group (Appendix S1, Supporting Information).

This sampling strategy enables us to assess the relative performance of the candidate loci in three disparate plant groups. Sampling is not exhaustive within groups, and we make no claim that any one of these groups is necessarily typical of the larger taxonomic group it was drawn from (it is indeed debatable whether there is such a thing as a 'typical' genus). Instead, our sampling strategy is designed (i) to have sufficient density of sampling within groups, such that sets of closely related species are included (which some loci will distinguish, and others will not), and (ii) by including three very divergent genera, to ensure that our conclusions are not susceptible to atypical behaviour of a given locus in one particular clade. This approach was chosen as a pragmatic trade-off between phylogenetic coverage and species-level sampling. It allows us to establish, in these three genera, the *relative* performances of the candidate barcoding loci.

Study taxa

Inga (*Leguminosae; angiosperm*). *Inga* is a genus of *c.* 300 South American tropical tree species and a significant contributor to the high levels of species diversity observed in many Neotropical forests (Pennington 1996). It is a classic example of a recent radiation, with evidence for many species arising within the last 10 million years (Richardson *et al.* 2001). Species-rich genera are important targets for DNA barcoding approaches as they often present significant identification challenges. Forty-four individuals representing 26 species were sampled (Appendix S1).

Araucaria (*Araucariaceae; gymnosperm*). *Araucaria* is a genus of 19 coniferous tree species, of which 13 are endemic to New Caledonia, whereas the other species have more scattered distributions (2 species in South America, 1 species on Norfolk Island, 1 species in Australia, 1 species in New Guinea and 1 species in both Australia and New Guinea). The genus has a fossil record dating back to the Jurassic and includes extant sections that are considered to have diverged during the Cretaceous, along with assemblages such as the New Caledonian species that are of more recent origin (Setoguchi *et al.* 1998). In spite of its great age, low levels of sequence variability have been reported (Kranitz 2005). A total of 42 individuals representing 17 species were sampled, along with one individual from each of two species of the related genus *Agathis* Salisbury (Appendix S1).

Asterella *s.l.* (*Aytoniaceae; liverwort*). *Asterella* is a paraphyletic genus of approximately 45–48 species (Long 2006); all others named Aytoniaceae genera are nested within it (Long *et al.* 2000), namely *Reboulia* Raddi (1 species: Bischler 1998), *Mannia* Opiz (7–8 species, Schill 2006), *Plagiochasma* Lehm. & Lindenb. (16 species, Bischler 1998) and *Cryptomitrium* Austin ex Underw. (3 species, Bischler 1998). Given the paraphyly of *Asterella*, we have included all of the constituent genera in our study. For convenience, we refer to this combined set of taxa as *Asterella* s.l. A total of 41 individuals representing 26 species were sampled for assessments of universality. A further 57 individuals, adding 13 more species, were sampled for the best performing barcoding loci for this genus (*rpoC1*, *rbcL*, *trnH-psbA*, *psbK-psbI*) to provide statistics on levels of species discrimination (Appendix S1). The final matrix comprised: *Asterella*, 39 accessions representing 20 species (*c.* 42% of total); *Reboulia*, 18 accessions representing 1 species (100% of total); *Mannia*, 23 accessions representing 5 species (*c.* 70% of total); *Plagiochasma*, 12 accessions representing 9 species (56% of total); *Cryptomitrium*, 2 accessions representing 2 species (66%

of total). One accession from Cleveaceae and three from Targioniaceae were also included due to initial plant misidentifications/mixed collections (subsequently identified and corrected using our barcoding data followed by morphological re-examination).

Locus screening overview

Seven candidate plastid barcoding loci were evaluated (*atpF-H, matK, rbcL, rpoB, rpoC1, psbK-psbI* and *trnH-psbA*). Initially, we used a test set consisting of 5–10 individuals each from *Inga, Araucaria* and *Asterella* s.l. These were trialled on all seven regions using available sets of primers (Appendix S2, Supporting Information). Optimal primer combinations were then selected and used for screening the full sample set.

DNA extraction and PCR

Total DNA was extracted from silica dried plant leaf/thallus material using QIAGEN's Plant DNeasy kits. Details of optimal PCR conditions and the primers tested for the seven candidate barcoding regions are given in Appendices S2 and S3. PCR products were cleaned using illustra DNA and Gel Band purification kits (GE Healthcare) and eluted in 20–30 μL type 4 elution buffer. Cycle sequencing was performed with 2–5 μL PCR product and 2 μL DTCS (Beckman Coulter) in a 10 μL reaction, and cleaned by ethanol precipitation. Sequences were analysed on a Beckman Coulter CEQ 8800 or 8000 Genetic Analysis System and edited using CEQ Genetic Analysis System software (version 8.0) before being assembled with Sequencher 4.6 (GeneCodes Corporation). All sequences were deposited in GenBank (Appendix S1).

Data analyses

Which of the proposed barcodes show the greatest universality? Our criterion for assessing universality simply involves establishing which regions could be routinely amplified and sequenced in the maximum number of samples in the three different plant groups, with the minimal set of PCR conditions. To facilitate interpretation of successes and failures, we have listed the primer combinations tested and the amount of PCR optimization required, with notes on the performance of each locus in Appendix S4, Supporting Information (including information on whether failures were due to PCR or sequencing problems).

Which of the proposed barcodes shows the greatest level of species discrimination? Sequences were exported as aligned NEXUS files from Sequencher. Alignments for noncoding loci were then optimized manually in Se-Al version 2.0a11 (Rambaut 2002). Separate alignments were made for *Inga, Araucaria* and the liverworts, with no attempt to align data between these groups. Evaluation of comparative levels of variation and discrimination was then undertaken in several ways. First, PAUP* 4.0b10 (Swofford 2003) was used to generate Kimura 2-parameter (K2P) distance matrices for each locus, and graphs comparing intrageneric divergences for each pair of loci were produced. The significance of divergence differences were tested with Wilcoxon signed-rank tests using PAST 1.81 (Hammer *et al.* 2001). Second, we assessed levels of species discrimination more directly. Although there are many potential ways of doing this, we opted for a simple characterization of the data into the following categories to form the basis of our comparisons among loci.

1 If any accession of a species has an identical DNA barcode sequence to an individual from another species, those species are considered nondistinguishable.
2 Species where just a single sample is included in the study are considered potentially distinguishable if the sequence from that sample is unique (i.e. there is the potential that successful species-level discrimination may be achieved, but further sampling is required to establish this).
3 Where multiple accessions are sampled per species, if all conspecific individuals of a species have more similar sequences (smallest K2P distances) compared to any heterospecific comparisons, then this is considered as successful discrimination for that locus, for that species. For convenience, we refer to this as conspecific individuals 'grouping together'.
4 Conversely, where multiple accessions are sampled for a given species but at least one conspecific K2P distance is greater than the smallest heterospecific distance involving that species, then this is considered an identification failure for the species in question (referred to as conspecific individuals 'not grouping together').

K2P distances were used following guidelines from the Consortium for the Barcoding of Life for evaluating performance among barcoding loci (http://www.barcoding.si.edu/protocols.html). Uncorrected P distances were also examined; the biological conclusions were identical for both models. Calculations assessing levels of species discrimination were only carried out in cases where a given region produced sequence data in > 50% of the samples for a given taxonomic group. This is to avoid spurious inflation of species discrimination statistics caused by simply having fewer species to discriminate. Cases where less than 50% of samples were sequenced for a given region in a given taxonomic group are thus considered as failures for the purposes of our analyses (0% success). This 50% sequencing success is an arbitrary threshold, but it does at least provide a consistent method for avoiding inflation of success statistics due to patchy sampling, and preliminary analyses of the data without this correction showed clear examples of artefactually increased species discrimination.

Tree-based analyses were also used to evaluate species discrimination and provide a convenient method of viewing the data. Neighbour-joining (NJ), unweighted pair group method with arithmetic mean (UPGMA) and maximum parsimony (MP) trees were generated in PAUP*. For NJ and UPGMA, both uncorrected P and K2P distances were used, with the 'break ties randomly' option. Parsimony searches were conducted using Fitch parsimony, gaps coded as missing data, with 100 random taxon addition replicates, saving no more than 15 trees per replicate.

What are the benefits in terms of species discrimination of using different combinations and different numbers of loci in a multi-locus barcoding approach? To evaluate potential benefits of multilocus barcodes over a single-locus barcode, we examined multiple combinations of the barcoding regions *within* each taxonomic group, and recorded levels of species discrimination afforded by each as described above. When loci were combined, individual samples that were missing for any one locus were excluded from the analyses; this results in minor differences in sample sizes for different combinations. The combinations tested included previously proposed multilocus barcode combinations (see Pennisi 2007), along with other combinations which looked promising based on the performance of individual loci. Up to 14 different multilocus combinations were evaluated in each genus.

What percentage of plant species in these three groups of land plants can be discriminated by plastid barcoding? Using the various methods for assessing levels of species discrimination described above, we estimated *overall success* of organelle barcoding regions in discriminating or potentially discriminating among species in the total data set. For single-locus statistics, this involves taking average values over all three taxonomic groups. In cases where a region has worked in all three taxonomic groups, this is straightforward. In cases where a given region was not successfully sequenced in > 50% of samples from a given taxonomic group, it simply contributes a 0% success rate to the average value. When combinations of loci were considered, variable success rates of loci across groups again becomes an issue. In cases where one locus in a combination failed in one or more taxonomic groups, those taxonomic groups are simply represented by the locus (or loci) that worked. For example, for the combination *rpoC1* + *trnH-psbA*, both regions produced sequence data from > 50% of samples in both *Inga* and the liverworts, but only *rpoC1* produced sequence data from > 50% of *Araucaria* individuals. In this case, data used to produce discrimination success values for this two-locus combination are *rpoC1* + *trnH-psbA* for *Inga* and the liverworts and *rpoC1* from *Araucaria.*

Is there any evidence for a 'barcode gap' in plants (a discontinuity between intra- and interspecific sequence divergence)? Taxon DNA (Meier *et al.* 2006) was used to generate intraspecific divergences and interspecific, congeneric divergences for the *Inga*, *Araucaria* and *Asterella* s.l. matrices. The inter- and intraspecific divergences were assigned into bins, and histograms of distance vs. abundance were generated to assess whether there was discontinuity between intra- and interspecific distances. As the liverwort genus *Asterella* resolves as paraphyletic in phylogenetic analyses (Long *et al.* 2000 and unpublished data; Schill 2006), the interspecific congeneric distances were estimated in a conservative fashion, in which the data set was broken down into monophyletic genera in a manner consistent with unpublished molecular and morphological evidence: *Reboulia*, *Plagiochasma*, *Mannia* (including *Asterella gracilis*), *Cryptomitrium*, AsterellaB (= *A. californica*), AsterellaC (= *A. grollei* and *A. palmeri*), and *Asterella* (all other *Asterella* species).

Results

Which of the proposed barcodes show the greatest universality?

The potential universality of these barcoding regions is summarized in Table 2 and presented in detail in Appendix S4. Only one barcoding locus (*rpoC1*) was routinely amplified and sequenced using a single primer pair and reaction conditions in all samples in all taxonomic groups. High quality sequences were reliably obtained from the forward primer enabling assembly of a character matrix of *c.* 450 bps, but success was more intermittent for the reverse primer. The second most universal region was *rbcL*. PCR and sequencing was straightforward in *Araucaria* and the liverworts using the Kress & Erickson (2007) barcoding primers. However, in *Inga*, although some samples worked well, others persistently failed using these primers, and an additional primer set was required to complete the matrix (Appendix S4).

The other barcoding loci all had low success rates in one group or another (Table 2, Appendix S4). We were unable to get *rpoB* to work in *Asterella* s.l. and needed different primer sets in *Inga* and *Araucaria.* In *Inga*, *matK* amplified easily, although internal sequencing primers were required for seven samples due to short reads; a different primer set was needed for *Araucaria*, and we were unsuccessful with *Asterella* s.l. Of the three spacer regions, *trnH-psbA* worked well in *Inga*, and most *Asterella* s.l. samples amplified and sequenced well. Success in *Araucaria* was more variable, and the length of the intergenic spacer varied greatly in size between *Araucaria* (*c.* 970–1120 bp) and the two species from the related genus *Agathis* (*c.* 380 bp). In *Araucaria*, *atpF-atpH* worked well, but PCR and sequencing success was lower in *Inga* (32%) and *Asterella* s.l. (21%); *psbK-psbI* did not amplify in *Araucaria*, but performed reasonably well in *Inga* (70%), although homopolymer repeats hampered sequencing success. In the small *Asterella* s.l. data matrix, we had promising initial success (78%, Table 4); however, unlike

Table 2 Summary of the proportion of individuals successfully amplified and sequenced from seven candidate barcoding regions in three groups of land plants using (a) all tested primers, and (b) the best single performing primer pairs. Full details are presented in Appendix S4. Where different primers produce broadly the same success but in different taxonomic groups, the primer pairs that worked in angiosperms are listed. The letter in square brackets following primer names corresponds to a citation reference in Appendix S2

	rpoC1	% success	*rpoB*	% success	*matK*	% success	*rbcL*	% success	*trnH-psbA*	% success	*atpF-atpH*	% success	*psbK-psbI*	% success
(a) With all tested primers														
Araucaria	44/44	100	44/44	100	44/44	100	41/44	93	21/44	48	41/44	93	0/44	0
Inga	44/44	100	42/44	95	42/44	95	43/43	100	44/44	100	14/44	32	31/44	70
Liverworts	40/40	100	0/40	0	0/40	0	40/41	98	36/41	88	5/24	21	32/41	78
Totals	128/128		86/128		86/128		124/128		101/129		60/112		63/129	
Mean % success		100.0		65.2		65.2		96.9		78.5		48.6		49.5
(b) With best single primer pair														
Best primers	2 and 4 (A)		1 and 3 (A)		FX and 3.2 (A)		a_f and a_r (E)		*psbAF* and *trnH2* (G&H)		*atpF* and *atpH* (C)		*psbK* and *psbI* (C)	
Araucaria	44/44	100	0/44	0	0/44	0	41/44	93	21/44	48	41/44	93	0/44	0
Inga	44/44	100	42/44	95	35/44	80	19/43	44	44/44	100	14/44	32	31/44	70
Liverworts	40/40	100	0/40	0	0/40	0	40/41	98	36/41	88	5/24	21	32/41	78
Totals	128/128		42/128		35/128		100/128		101/129		60/112		63/129	
Mean % success		100.0		31.8		26.7		78.3		78.5		48.6		49.5

Table 3 Levels of potential species discrimination and failure using single-locus and multilocus combinations of seven candidate DNA barcoding regions in three groups of land plants. *Results for *psbK-psbI* for *Asterella* s.l. (liverworts) are based on reduced data set compared to other regions (see text for details)

						Species in which no sampled accessions have identical sequences to other species				
							Species with > 1 sampled accessions			
Taxon	Locus	Total no. of accessions from which sequence data was obtained	Total no. of species where > 1 accession was sampled	Total no. of species	% species with at least one accession with identical sequence to another species	% of species represented by single samples which have unique sequences	% of species in which all accessions group together	% of species in which all accessions do not group together	Total % identification failures	% potential success (1-failure rate)
Araucaria	*matK*	44	14	19	78.9	15.8	5.3 (1/14)	0.0	78.9	21.1
	rpoB	44	14	19	73.7	21.1	5.3 (1/14)	0.0	73.7	26.3
	rpoC1	44	14	19	89.5	10.5	0 (0/14)	0.0	89.5	10.5
	rbcL	41	13	19	78.9	15.8	5.3 (1/14)	0.0	78.9	21.1
	atpF-H	41	13	18	83.3	16.7	0 (0/13)	0.0	83.3	16.7
	rpoC1 + *matK*	44	14	19	78.9	15.8	5.3 (1/14)	0.0	78.9	21.1
	rbcL + *matK*	41	13	19	78.9	15.8	5.3 (1/13)	0.0	78.9	21.1
	rbcL + *rpoC1*	41	13	19	78.9	15.8	5.3 (1/13)	0.0	78.9	21.1
	matK + *atpF-H*	41	13	18	72.2	27.8	0 (0/13)	0.0	72.2	27.8
	rpoC1, rpoB + *matK*	44	14	19	68.4	26.3	5.3 (1/14)	0.0	68.4	31.6
	rpoC1, rbcL + *matK*	41	13	19	78.9	15.8	5.3 (1/13)	0.0	78.9	21.1
	rpoC1, rbcL, rpoB, matK + *atpF-H*	38	12	18	72.2	27.8	0 (0/12)	0.0	72.2	27.8

Table 3 *Continued*

						Species in which no sampled accessions have identical sequences to other species				
							Species with > 1 sampled accessions			
Taxon	Locus	Total no. of accessions from which sequence data was obtained	Total no. of species where > 1 accession was sampled	Total no. of species	% species with at least one accession with identical sequence to another species	% of species represented by single samples which have unique sequences	% of species in which all accessions group together	% of species in which all accessions do not group together	Total % identification failures	% potential success (1-failure rate)
Inga	*matK*	42	7	26	65.4	26.9	3.8 (1/7)	3.8	69.2	30.8
	rpoB	42	9	24	91.7	8.3	0 (0/9)	0.0	91.7	8.3
	rpoC1	44	9	26	84.6	15.4	0 (0/9)	0.0	84.6	15.4
	rbcL	43	8	26	84.6	15.4	0 (0/8)	0.0	84.6	15.4
	trnH-psbA	44	9	26	73.1	23.1	3.8 (1/9)	0.0	73.1	26.9
	rpoC1 + *matK*	42	7	26	46.2	46.2	3.8 (1/7)	3.8	50.0	50.0
	rbcL + *matK*	42	7	26	42.3	46.2	3.8 (1/7)	7.7	50.0	50.0
	matK + *trnH-psbA*	42	7	26	34.6	53.8	3.8 (1/7)	7.7	42.3	57.7
	rpoC1 + *rbcL*	43	8	26	57.7	34.6	7.7 (2/8)	0.0	57.7	42.3
	rpoC1 + *trnH-psbA*	44	9	26	61.5	34.6	3.8 (1/9)	0.0	61.5	38.5
	rbcL + *trnH-psbA*	43	8	26	50.0	46.2	0 (0/8)	3.8	53.8	46.2
	rpoC1, *rbcL* + *matK*	42	7	26	26.9	61.5	7.7 (2/7)	3.8	30.8	69.2
	rpoC1, *rpoB* + *matK*	40	7	24	50.0	41.7	4.2 (1/7)	4.2	54.2	45.8
	rpoC1, *matK* + *trnH-psbA*	42	7	26	34.6	53.8	3.8 (1/7)	7.7	42.3	57.7
	rbcL, *matK* + *trnH-psbA*	42	7	26	23.1	61.5	7.7 (2/7)	7.7	30.8	69.2
	rpoC1, *rbcL* + *trnH-psbA*	43	8	26	34.6	53.8	3.8 (1/8)	7.7	42.3	57.7
	rpoC1, *rbcL*, *matK* + *trnH-psbA*	42	7	26	23.1	61.5	3.8 (1/7)	11.5	34.6	65.4
	rpoC1, *rbcL*, *rpoB*, *matK* + *trnH-psbA*	40	7	24	25.0	58.3	8.3 (2/7)	8.3	33.3	66.7
Asterella s.l.	*rpoC1*	96	17	39	30.8	35.9	33.3 (13/17)	0.0	30.8	69.2
	rbcL	95	17	38	7.9	47.4	42.1 (16/17)	2.6	10.5	89.5
	trnH-psbA	86	17	38	26.3	34.2	39.5 (15/17)	0.0	26.3	73.7
	**psbK-psbI*	32	5	23	8.7	73.9	17.4 (4/5)	0.0	8.7	91.3
	rpoC1 + *rbcL*	93	17	38	7.9	47.4	42.1 (16/17)	2.6	10.5	89.5
	rbcL + *trnH-psbA*	85	17	36	8.1	45.9	43.2 (16/17)	2.7	10.8	89.2
	rpoC1 + *trnH-psbA*	85	17	38	23.7	36.8	39.5 (15/17)	0.0	23.7	76.3
	rpoC1, *rbcL* + *trnH-psbA*	84	17	36	8.1	45.9	45.9 (17/17)	0.0	8.1	91.9

Table 4 Summary statistics indicating % levels of species discrimination for seven candidate DNA barcoding regions in three groups of land plants. Where a region worked in < 50% of individuals in a given group, it is given 0% discrimination rate to avoid sparse sampling inflating success statistics. *Results for *psbK-psbI* for *Asterella* s.l. (liverworts) are based on reduced data set compared to other regions (see text for details)

Region		*rpoC1*	*rpoB*	*matK*	*rbcL*	*trnH-psbA*	*atpF-atpH*	*psbK-psbI*
% species discrimination using all tested primers	*Araucaria*	10.5	26.3	21.1	21.1	0	16.7	0
	Inga	15.4	8.3	30.8	15.4	26.9	0	4.8
	Asterella s.l.	69.2	0	0	89.5	73.7	0	91*
Mean		31.7	11.5	17.3	42.0	33.5	5.6	32.0
% species discrimination using single best primer pair		31.7	8.8	7.0	36.84	33.5	5.6	32.0

rbcL, *trnH-psbA* and *rpoC1*, PCR failure meant this success was not robust to increasing sampling to the full 98 sample matrix.

Taking the mean percentage of samples for which sequence data was recovered from the three groups, universality of the loci can be ranked as *rpoC1* (100%), *rbcL* (97%), *trnH-psbA* (79%), *rpoB*/*matK* (both 65%), *psbK-psbI* (50%), *atpF-atpH* (49%). This is based on using a range of primer sets for some loci (Table 2a). Using just the best performing primer set for each locus (Table 2b), the rank order becomes *rpoC1* (100%), *trnH-psbA* (79%), *rbcL* (78%), *psbK-psbI* (50%), *atpF-atpH* (49%), *rpoB* (32%), *matK* (27%).

Which of the proposed barcodes shows the greatest level of species discrimination?

Description of divergence levels. Within individual data sets, the least number of variable characters was three, for *rpoC1* for the *Araucaria* matrix, whereas the highest number of variable characters for a locus was 116 in the small liverwort matrix for *psbK-psbI*. Graphs comparing K2P distances between individuals for each of the seven potential barcode regions are shown in Fig. 1. Due to the range of distances between coding and noncoding regions, the graphs presented are drawn to three separate scales, with K2P axes distances of 0.04, 0.12 and 0.2 according to the loci being compared. Interpretation of the results is complicated as not all regions worked in each group. In comparisons between noncoding and coding regions, the noncoding region always had the larger distances, although for the comparison between *matK* and *atpF-atpH*, the divergence levels were similar. Among the coding regions, *matK* showed higher pairwise divergences than the other regions, and *rpoB* showed higher divergence values in pairwise comparisons with *rbcL* and *rpoC1* (although for several of these comparisons, the differences were small, particularly for *rpoB* and *rbcL*, and for *matK* and *rbcL*). For noncoding regions, no clear picture arises, although divergence in the *psbK-psbI* region was greater than in *trnH-psbA*, and *trnH-psbA* showed higher divergences than *atpF-atpH*. Across all taxonomic groups, using pairwise distances, the seven loci can be broadly ranked as follows: *psbK-psbI* > *trnH-psbA* > *atpF-atpH* > *matK* > *rpoB* > *rpoC1* > *rbcL*.

Assessments of levels of species discrimination. The *Araucaria* data set showed low levels of sequence divergence between most samples for all barcode regions. The most extreme example was from *rpoC1* in the *Araucaria* matrix, in which the single individuals sampled of *Araucaria araucana* and *A. hunsteinii* had unique sequences differing from each other by a single base change; all other *Araucaria* species shared a sequence, and the two *Agathis* species shared a different sequence. Thus, only 10.5% of the species are potentially distinguishable (Tables 3 and 4). The most successful region was *rpoB* with 21% of singleton sampled species potentially distinguishable (4 species had unique sequences), and 1 of 14 species from which multiple accessions were sampled 'grouped together' (having all accessions more similar to each other than to accessions from any other species). All New Caledonian *Araucaria* species shared sequences with at least one other species for *rpoB*. Thus, even for the most successful region, the highest potential level of species discrimination for a single locus barcode in the total *Araucaria* matrix is 26% (5 out of 19 species; Tables 3 and 4).

In the *Inga* data set, many species have identical sequences at any given barcoding locus. The best performing was *matK*, for which 1 out of the 7 species with multiple accessions sampled group together (Table 3), and 7 out of 17 species from which single individuals were sampled had unique sequences (65% of the species sampled had at least one accession with a sequence identical to another species). The highest potential level of discrimination for a single locus barcode in this group is 31% (comprised of *c.* 27% species having unique sequences based on singleton samples, and *c.* 4% of species with multiple accessions grouping together). *trnH-psbA* performed almost as well as *matK* (Tables 3 and 4). The most poorly performing loci were *rpoB* and *psbK-psbI* in which 92% and 95% of species, respectively, had at least one accession with an identical sequence to another species; performances of *rpoC1* and *rbcL* were intermediate (Tables 3 and 4).

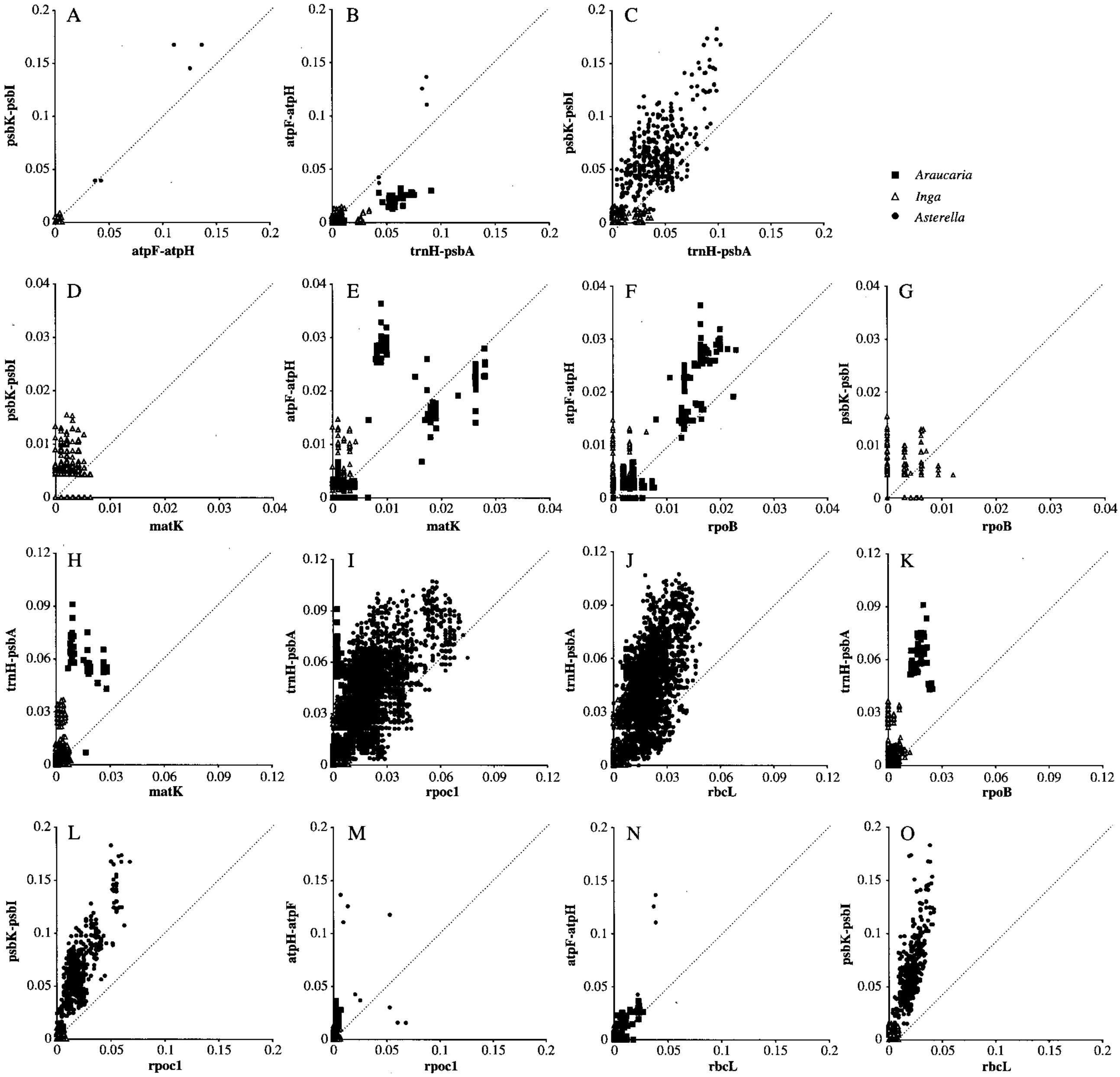
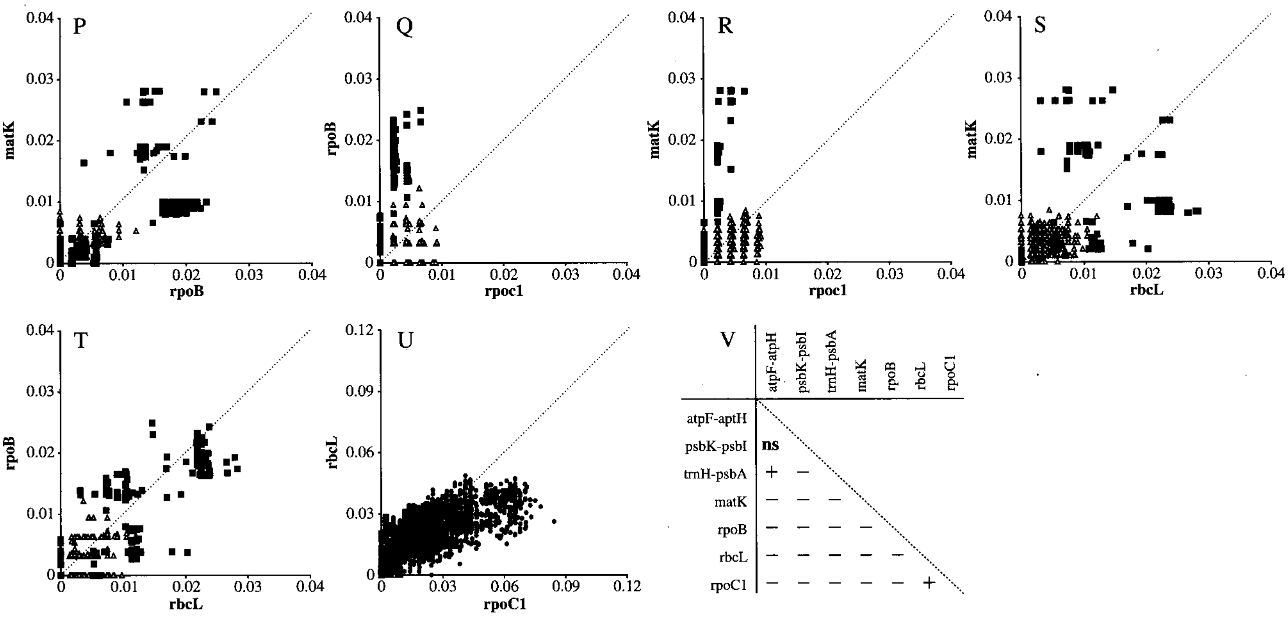


Fig. 1 K2P pairwise genetic distances for all two-locus permutations based on seven candidate DNA barcoding regions in three groups of land plants. (A) *psbK-psbI* and *atpF-atpH*, based on *Inga* and *Asterella* s.l.; (B) *atpF-atpH* and *trnH-psbA*, based on *Araucaria*, *Inga* and *Asterella* s.l.; (C) *psbK-psbI* and *trnH-psbA*, based on *Inga* and *Asterella* s.l.; (D) *psbK-psbI* and *matK*, based on *Inga*; (E) *atpF-atpH* and *matK*, based on *Araucaria* and *Inga*; (F) *atpF-atpH* and *rpoB*, based on *Araucaria* and *Inga*; (G) *psbK-psbI* and *rpoB*, based on *Inga*; (H) *trnH-psbA* and *matK*, based on *Araucaria* and *Inga*, (I) *trnH-psbA* and *rpoC1*, based on *Araucaria*, *Inga* and *Asterella* s.l., (J) *trnH-psbA* and *rbcL*, based on *Araucaria*, *Inga* and *Asterella* s.l.; (K) *trnH-psbA* and *rpoB*, based on *Araucaria* and *Inga*; (L) *psbK-psbI* and *rpoC1*, based on *Inga* and *Asterella* s.l.; (M) *atpF-atpH* and *rpoC1*, based on *Araucaria*, *Inga* and *Asterella* s.l., (N) *atpF-atpH* and *rbcL*, based on *Araucaria*, *Inga* and *Asterella* s.l.; (O) *psbK-psbI* and *rbcL*, based on *Inga* and *Asterella* s.l.; (P) *matK* and *rpoB*, based on *Araucaria* and *Inga*; (Q) *rpoB* and *rpoC1*, based on *Araucaria* and *Inga*; (R) *matK* and *rpoC1*, based on *Araucaria* and *Inga*, (S) *matK* and *rbcL*, based on *Araucaria* and *Inga*, (T) *rpoB* and *rbcL*, based on *Araucaria* and *Inga*; (U) *rbcL* and *rpoC1*, based on *Araucaria*, *Inga* and *Asterella* s.l. Scale: D–G, P–T: K2P distances up to 0.04; A–C, L–O: K2P distances up to 0.2; H–K, U: K2P distances up to 0.12, A–C represent comparisons between noncoding loci, D–O represent comparisons between coding and noncoding loci, P–U represent comparisons between coding loci. (V) Results of Wilcoxon signed-rank tests for each locus combination: ns, not significant; +, locus on vertical (*y*) axis significantly faster; –, locus on vertical axis significantly slower.



	atpF-atpH	psbK-psbI	trnH-psbA	matK	rpoB	rbcL	rpoC1
atpF-aptH							
psbK-psbI	**ns**						
trnH-psbA	+	–					
matK	–	–	–				
rpoB	–	–	–	–			
rbcL	–	–	–	–	–		
rpoC1	–	–	–	–	–	+	

Fig. 1 *Continued*

In *Asterella* s.l., of the four regions that worked well in the initial screen of 41 samples (*rpoC1, rbcL, trnH-psbA, psbK-psbI*), three worked well when the 57 further samples were added. In the larger sample screen, *psbK-psbI* amplified poorly, and thus, the results presented for this locus are based on the smaller data set.

Much higher levels of resolution were obtained in the *Asterella* s.l. data set than in the other two (Tables 3 and 4). For *rbcL* 16 of 17 species from which multiple individuals were sampled 'grouped together', and only 8% of species shared sequences with another species (Table 3). The second and third highest levels of resolution were shown by *trnH-psbA* and *rpoC1* with 15 of 17 and 13 of 17 species 'grouping together', and 26% and 31% of species sharing sequences, respectively (Table 3). *psbK-psbI* showed good discrimination in the smaller data set, but a confounding variable here is the smaller number of species present in the analysis.

Tree-based analyses. Parsimony and distance analyses were carried out on all data sets, and unsurprisingly mirrored the results described above in terms of the distribution of sequence variation among species. Representative trees are shown in Fig. 2 illustrating the highest levels of species discrimination achieved in each taxonomic group for a single locus, namely *rpoB* in *Araucaria*, *matK* in *Inga* and *rbcL* in *Asterella* s.l.

What are the benefits in terms of species discrimination of using different combinations and different numbers of loci in a multilocus barcoding approach?

For *Araucaria*, when all loci that produced sequence data from > 50% of samples were combined (*rpoC1, rpoB, matK, rbcL, atpF-atpH*), 72% of species still shared a sequence with at least one other species (Table 3). This is virtually identical to the performance of the best single locus (*rpoB*). For this group, there are no benefits to adding loci. A marginally better performance was achieved from the combination of *rpoC1*, *rpoB* and *matK* (68% of samples sharing sequences), but this is just attributable to a slightly different sample set being considered rather than an improvement in performance.

For *Inga*, a different picture emerges. As further loci are added, the percentage of species that share sequences declines. There is thus an increase in the percentage of species that are potentially distinguishable. The biggest two-locus effect comes from the combined use of *matK* and *trnH-psbA*, which takes the percentage of species sharing sequences from 65% for the best single locus solution to 35% (Table 3). Adding a third locus, the combination of *rpoC1*, *rbcL* and *matK* drops this to 27%, and with four loci (*matK, rbcL, rpoC1, trnH-psbA*) to 23% (Table 3). However, the species from which multiple individuals have been sampled do not show a corresponding increase in the frequency of multiple accessions grouping together. Depending on the locus combination, the maximum increase is from one species to two (Table 3).

In *Asterella* s.l., the potential for improvement in species resolution by adding regions is limited due to the high performance of *rbcL* alone (Table 3). Adding *rpoC1* and *trnH-psbA* results in 17 of 17 species with multiple individuals sampled having all intraspecific accessions grouping together, and only 8% of samples sharing sequences (*rbcL* alone gave 16 of 17 and 8%, respectively).

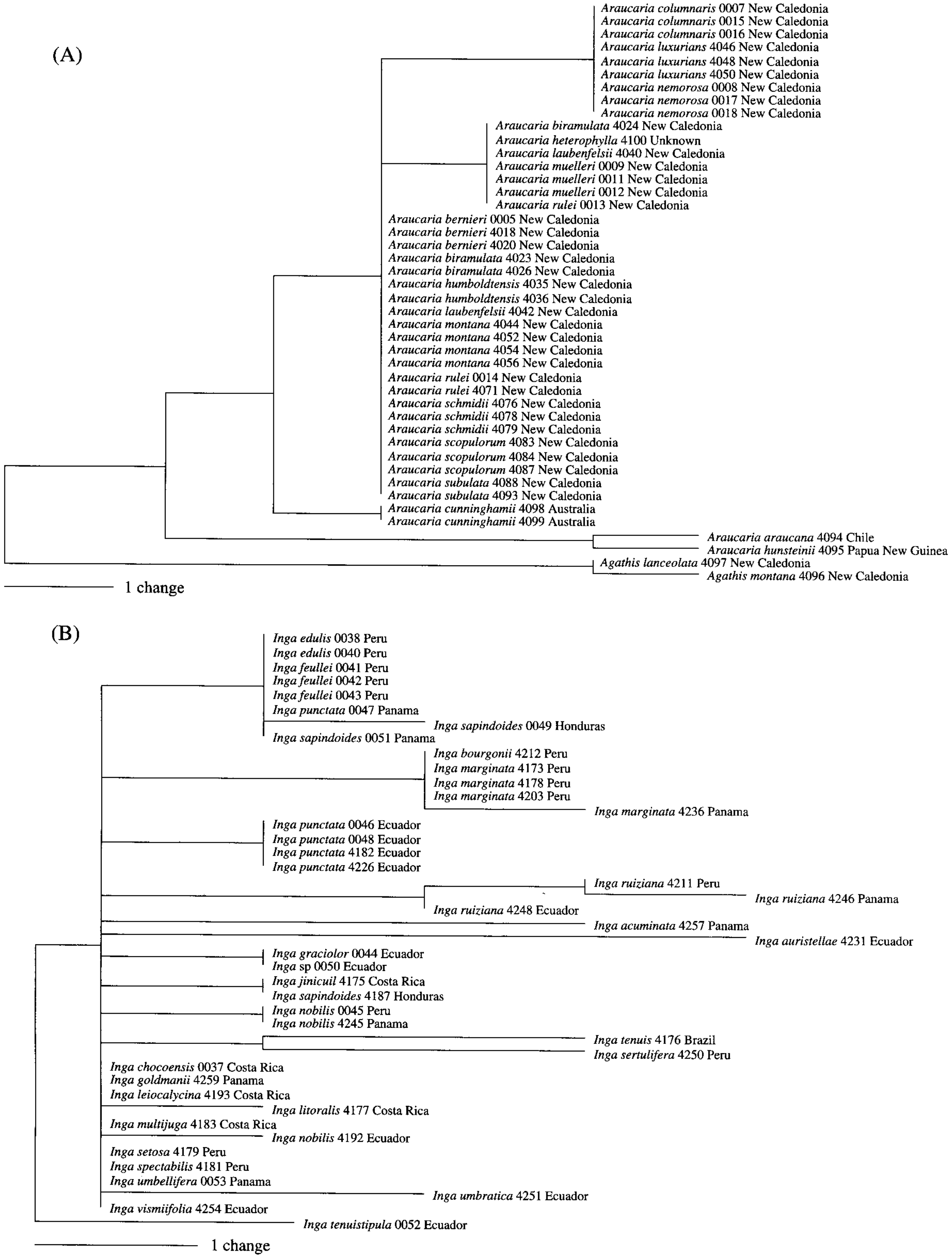

(A)
Araucaria columnaris 0007 New Caledonia
Araucaria columnaris 0015 New Caledonia
Araucaria columnaris 0016 New Caledonia
Araucaria luxurians 4046 New Caledonia
Araucaria luxurians 4048 New Caledonia
Araucaria luxurians 4050 New Caledonia
Araucaria nemorosa 0008 New Caledonia
Araucaria nemorosa 0017 New Caledonia
Araucaria nemorosa 0018 New Caledonia
Araucaria biramulata 4024 New Caledonia
Araucaria heterophylla 4100 Unknown
Araucaria laubenfelsii 4040 New Caledonia
Araucaria muelleri 0009 New Caledonia
Araucaria muelleri 0011 New Caledonia
Araucaria muelleri 0012 New Caledonia
Araucaria rulei 0013 New Caledonia
Araucaria bernieri 0005 New Caledonia
Araucaria bernieri 4018 New Caledonia
Araucaria bernieri 4020 New Caledonia
Araucaria biramulata 4023 New Caledonia
Araucaria biramulata 4026 New Caledonia
Araucaria humboldtensis 4035 New Caledonia
Araucaria humboldtensis 4036 New Caledonia
Araucaria laubenfelsii 4042 New Caledonia
Araucaria montana 4044 New Caledonia
Araucaria montana 4052 New Caledonia
Araucaria montana 4054 New Caledonia
Araucaria montana 4056 New Caledonia
Araucaria rulei 0014 New Caledonia
Araucaria rulei 4071 New Caledonia
Araucaria schmidii 4076 New Caledonia
Araucaria schmidii 4078 New Caledonia
Araucaria schmidii 4079 New Caledonia
Araucaria scopulorum 4083 New Caledonia
Araucaria scopulorum 4084 New Caledonia
Araucaria scopulorum 4087 New Caledonia
Araucaria subulata 4088 New Caledonia
Araucaria subulata 4093 New Caledonia
Araucaria cunninghamii 4098 Australia
Araucaria cunninghamii 4099 Australia
Araucaria araucana 4094 Chile
Araucaria hunsteinii 4095 Papua New Guinea
Agathis lanceolata 4097 New Caledonia
Agathis montana 4096 New Caledonia
1 change
(B)
Inga edulis 0038 Peru
Inga edulis 0040 Peru
Inga feullei 0041 Peru
Inga feullei 0042 Peru
Inga feullei 0043 Peru
Inga punctata 0047 Panama
Inga sapindoides 0049 Honduras
Inga sapindoides 0051 Panama
Inga bourgonii 4212 Peru
Inga marginata 4173 Peru
Inga marginata 4178 Peru
Inga marginata 4203 Peru
Inga marginata 4236 Panama
Inga punctata 0046 Ecuador
Inga punctata 0048 Ecuador
Inga punctata 4182 Ecuador
Inga punctata 4226 Ecuador
Inga ruiziana 4211 Peru
Inga ruiziana 4246 Panama
Inga ruiziana 4248 Ecuador
Inga acuminata 4257 Panama
Inga auristellae 4231 Ecuador
Inga graciolor 0044 Ecuador
Inga sp 0050 Ecuador
Inga jinicuil 4175 Costa Rica
Inga sapindoides 4187 Honduras
Inga nobilis 0045 Peru
Inga nobilis 4245 Panama
Inga tenuis 4176 Brazil
Inga sertulifera 4250 Peru
Inga chocoensis 0037 Costa Rica
Inga goldmanii 4259 Panama
Inga leiocalycina 4193 Costa Rica
Inga litoralis 4177 Costa Rica
Inga multijuga 4183 Costa Rica
Inga nobilis 4192 Ecuador
Inga setosa 4179 Peru
Inga spectabilis 4181 Peru
Inga umbellifera 0053 Panama
Inga umbratica 4251 Ecuador
Inga vismiifolia 4254 Ecuador
Inga tenuistipula 0052 Ecuador
1 change

Fig. 2 Maximum parsimony phylograms illustrating sample relationships and branch lengths; four-digit identification numbers refer to voucher details in Appendix S1. (A) *Araucaria* based on *rpoB* sequence data, (B) *Inga* using *matK* sequence data, (C) *Asterella* s.l. using *rbcL* sequence data, with multi-accession species picked out in colour. Distance analysis of these data groups 16 of 17 multi-accessioned species together. 15 of these cluster here; *Reboulia hemisphaerica* groups together in distance analysis but resolves as paraphyletic here, while *Asterella mussuriensis* fails in both.

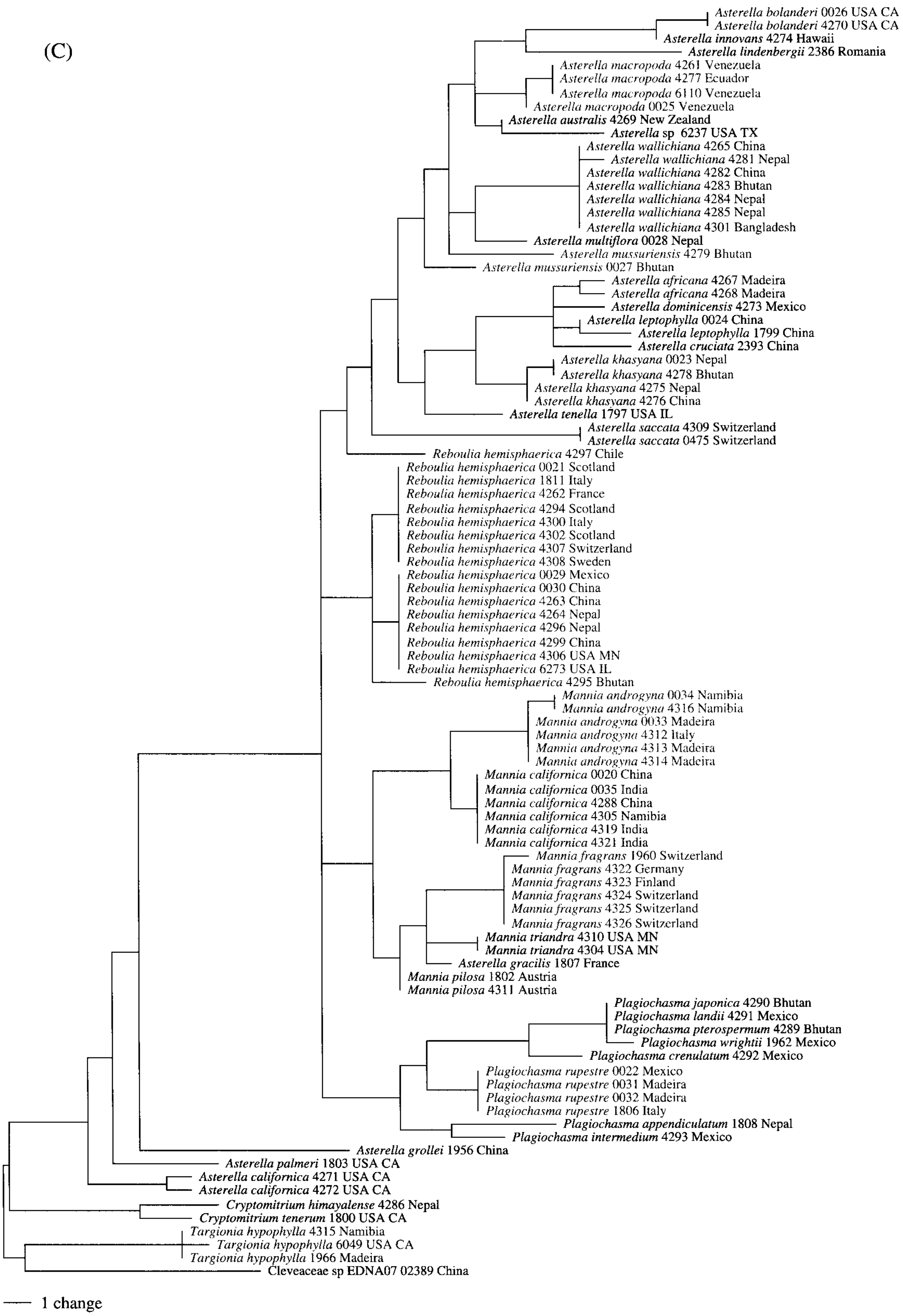

(C)
Asterella bolanderi 0026 USA CA
Asterella bolanderi 4270 USA CA
Asterella innovans 4274 Hawaii
Asterella lindenbergii 2386 Romania
Asterella macropoda 4261 Venezuela
Asterella macropoda 4277 Ecuador
Asterella macropoda 6110 Venezuela
Asterella macropoda 0025 Venezuela
Asterella australis 4269 New Zealand
Asterella sp 6237 USA TX
Asterella wallichiana 4265 China
Asterella wallichiana 4281 Nepal
Asterella wallichiana 4282 China
Asterella wallichiana 4283 Bhutan
Asterella wallichiana 4284 Nepal
Asterella wallichiana 4285 Nepal
Asterella wallichiana 4301 Bangladesh
Asterella multiflora 0028 Nepal
Asterella mussuriensis 4279 Bhutan
Asterella mussuriensis 0027 Bhutan
Asterella africana 4267 Madeira
Asterella africana 4268 Madeira
Asterella dominicensis 4273 Mexico
Asterella leptophylla 0024 China
Asterella leptophylla 1799 China
Asterella cruciata 2393 China
Asterella khasyana 0023 Nepal
Asterella khasyana 4278 Bhutan
Asterella khasyana 4275 Nepal
Asterella khasyana 4276 China
Asterella tenella 1797 USA IL
Asterella saccata 4309 Switzerland
Asterella saccata 0475 Switzerland
Reboulia hemisphaerica 4297 Chile
Reboulia hemisphaerica 0021 Scotland
Reboulia hemisphaerica 1811 Italy
Reboulia hemisphaerica 4262 France
Reboulia hemisphaerica 4294 Scotland
Reboulia hemisphaerica 4300 Italy
Reboulia hemisphaerica 4302 Scotland
Reboulia hemisphaerica 4307 Switzerland
Reboulia hemisphaerica 4308 Sweden
Reboulia hemisphaerica 0029 Mexico
Reboulia hemisphaerica 0030 China
Reboulia hemisphaerica 4263 China
Reboulia hemisphaerica 4264 Nepal
Reboulia hemisphaerica 4296 Nepal
Reboulia hemisphaerica 4299 China
Reboulia hemisphaerica 4306 USA MN
Reboulia hemisphaerica 6273 USA IL
Reboulia hemisphaerica 4295 Bhutan
Mannia androgyna 0034 Namibia
Mannia androgyna 4316 Namibia
Mannia androgyna 0033 Madeira
Mannia androgyna 4312 Italy
Mannia androgyna 4313 Madeira
Mannia androgyna 4314 Madeira
Mannia californica 0020 China
Mannia californica 0035 India
Mannia californica 4288 China
Mannia californica 4305 Namibia
Mannia californica 4319 India
Mannia californica 4321 India
Mannia fragrans 1960 Switzerland
Mannia fragrans 4322 Germany
Mannia fragrans 4323 Finland
Mannia fragrans 4324 Switzerland
Mannia fragrans 4325 Switzerland
Mannia fragrans 4326 Switzerland
Mannia triandra 4310 USA MN
Mannia triandra 4304 USA MN
Asterella gracilis 1807 France
Mannia pilosa 1802 Austria
Mannia pilosa 4311 Austria
Plagiochasma japonica 4290 Bhutan
Plagiochasma landii 4291 Mexico
Plagiochasma pterospermum 4289 Bhutan
Plagiochasma wrightii 1962 Mexico
Plagiochasma crenulatum 4292 Mexico
Plagiochasma rupestre 0022 Mexico
Plagiochasma rupestre 0031 Madeira
Plagiochasma rupestre 0032 Madeira
Plagiochasma rupestre 1806 Italy
Plagiochasma appendiculatum 1808 Nepal
Plagiochasma intermedium 4293 Mexico
Asterella grollei 1956 China
Asterella palmeri 1803 USA CA
Asterella californica 4271 USA CA
Asterella californica 4272 USA CA
Cryptomitrium himayalense 4286 Nepal
Cryptomitrium tenerum 1800 USA CA
Targionia hypophylla 4315 Namibia
Targionia hypophylla 6049 USA CA
Targionia hypophylla 1966 Madeira
Cleveaceae sp EDNA07 02389 China
1 change

Fig. 2 *Continued*

What percentage of plant species in these three groups of land plants can be discriminated by plastid barcoding?

We assessed the percentage of species potentially distinguishable as the sum of singleton sampled species that had unique sequences plus the species represented by multiple accessions for which samples group together. Considering single-locus approaches, the highest potential level of species discrimination was 90% from *rbcL* in the liverworts (Table 4). Taken as an average across the three groups, *rbcL* was the most successful locus with an upper estimate of 42% of species discriminated. The second most successful locus was *trnH-psbA* (34%). This assumes all groups are treated equally. If we just consider the angiosperm group *Inga*, *matK* gives the single greatest resolution (31%), followed closely by *trnH-psbA*.

If we recalculate these figures based on using the best single primer combination for each region (Table 4), *rbcL* is again the most successful at 37% (it is considered a failure in *Inga* as the standard barcoding primers needed supplementing to get this region to work, but as the percentage of species successfully resolved in *Inga* is so low anyway this makes little difference). Both *trnH-psbA* and *rpoC1* give similar results at 34% and 32%, respectively (the former with high resolution but patchy amplification, the latter with consistent amplification but low resolution). While *psbK-psbI* gives a similar result of 32% in these analyses, this is primarily based on its success in the small 41 sample liverwort data set, which was not repeatable when we scaled up to the large 98 sample matrix.

When combinations of regions are considered, there are gains in levels of overall potential species discrimination. Tables 3 and 5 summarize the results of various combinations of loci. A key point that emerges is that all major competing multilocus combinations published to date produced virtually identical levels of success, with *c.* 50% success based on the average of these three data sets (Table 5). The best performing combinations were *rbcL* + *trnH-psbA* + *matK* and also *rpoC1* + *rbcL* + *matK* with about 60% of species potentially discriminated. The amount of missing data cells that a given primer combination produced are summarized in Table 5 (a missing data cell = failure to obtain sequence data from > 50% of individuals from an individual barcode region for a given taxonomic group). The number of missing cells range from 1/9 to 4/9 for the locus combinations listed (11–44%). The locus combinations that produced the smallest number of missing data cells based on the best performing single primer pairs were *rpoC1* + *trnH-psbA* (1/6 cells missing), and *rpoC1* + *rbcL* + *trnH-psbA* (2/9 cells missing).

Is there any evidence for a 'barcode gap' in plants?

The frequency distribution of intraspecific and interspecific sequence divergences is shown in Fig. 3 (all interspecific distances are 'between-species within-genera'). As expected,

Table 5 Summary statistics indicating levels of resolvability of nine different multilocus combinations of seven candidate DNA barcoding regions in three groups of land plants. Combinations suggested by other studies are indicated by column headings; see Table 1 for further details. A data cell (an individual barcode region for a given taxonomic group) is considered missing if sequence data was not obtained for > 50% of samples for a given taxonomic group

Combination suggested by other studies (Pennisi 2007)		Chase *et al.*	Chase *et al.*	—	Kress & Erickson	—	Kim	Kim	—	—
Multilocus combination		*rpoC1*, *rpoB* + *matK*	*rpoC1*, *matK* + *trnH-psbA*	*rpoC1* + *trnH-psbA*	*rbcL* + *trnH-psbA*	*rbcL*, *trnH-psbA* + *matK*	*matK*, *atpF-atpH* + *psbK-psbI*	*matK*, *atpF-atpH* + *trnH-psbA*	*rpoC1*, *rbcL* + *matK*	*rpoC1*, *rbcL* + *trnH-psbA*
% species discrimination using all tested primers	*Araucaria*	31.6	21.1	10.5	21.1	21.1	27.8	27.8	21.1	21.1
	Inga	45.8	57.7	38.5	46.2	69.2	30.8	57.7	69.2	57.7
	Asterella s.l.	69.2	76.3	76.3	89.2	89.2	91.3	73.7	89.5	91.9
Mean (%)		48.9	51.7	41.8	52.1	59.8	50.0	53.1	59.9	56.9
Missing data cells		2 of 9	2 of 9	1 of 6	1 of 6	2 of 9	4 of 9	4 of 9	1 of 9	1 of 9
Missing data cells with best single primer pair		4 of 9	3 of 9	1 of 6	2 of 6	4 of 9	6 of 9	5 of 9	3 of 9	2 of 9

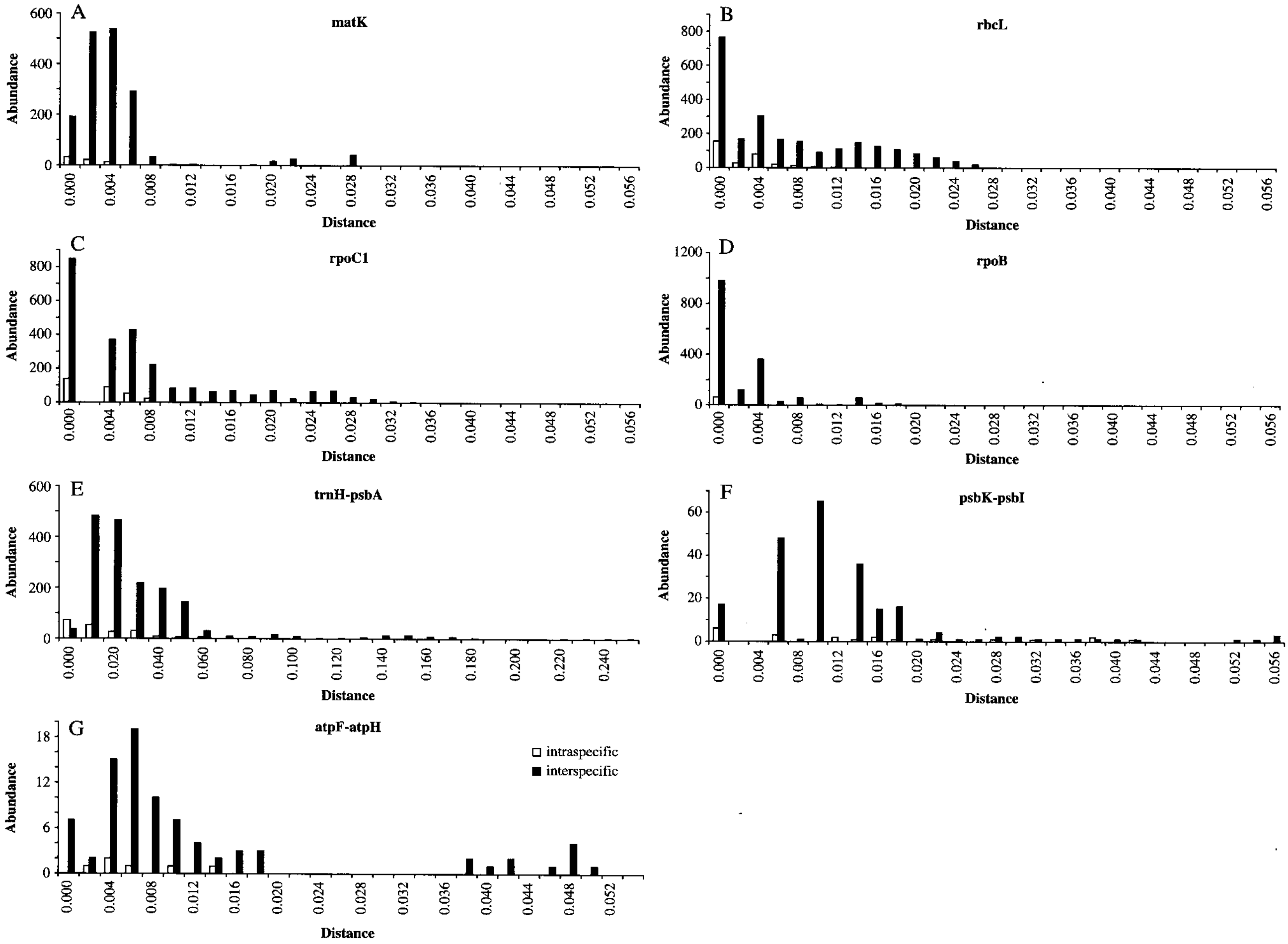


Fig. 3 Intraspecific vs. interspecific K2P distances from seven candidate DNA barcoding regions in three groups of land plants. (A) *matK*, generated using *Araucaria* and *Inga* sequence data; (B) *rbcL*, generated using *Araucaria*, *Inga* and *Asterella* s.l. sequence data; (C) *rpoC1*, generated using *Araucaria*, *Inga* and *Asterella* s.l. sequence data; (D) *rpoB*, generated using *Araucaria* and *Inga* sequence data; (E) *trnH-psbA*, generated using *Araucaria*, *Inga* and *Asterella* s.l. sequence data; (F) *psbK-psbI*, generated using *Inga* and *Asterella* s.l. sequence data; (G) *atpF-atpH*, generated using *Araucaria* and *Inga* sequence data.

interspecific divergences are generally larger overall than intraspecific values. However, there is no discontinuity between intra- and interspecific divergences, and the graphs reflect the many cases of interspecific distances of zero. Of the coding regions, *matK* is the only one in which the most abundant interspecific distance class is not zero. For *trnH-psbA*, the zero distance class is dominated by intraspecific comparisons, but even here, there is considerable overlap between intra- and interspecific distances.

Results summary

The key points which emerge from this set of analyses are as follows:

1 *rpoC1* was the most universal locus and amplified well across all three groups; *trnH-psbA* showed greatest universality of the noncoding regions.

2 Higher levels of sequence divergence were detected using noncoding regions, but in individual taxonomic groups, for species discrimination, the best performing locus was in each case a coding locus, albeit a different locus in each group.

3 DNA barcoding worked well in *Asterella* s.l., with high levels of species discrimination (90% from *rbcL* alone).

4 Species discrimination success in the two groups of seed plants was much lower with 26% (*Araucaria*) and 31% (*Inga*) based on single loci, and 32% (*Araucaria*) and 69% (*Inga*) based on multilocus combinations.

5 In the angiosperm *Inga*, *matK* showed the greatest levels of species discrimination (31%), followed by *trnH-psbA* (27%).

6 The main previously published suggestions for multilocus barcoding combinations performed approximately equally in this study.

7 There was no evidence for a clear disjunction between intra- and interspecific divergences in the three groups analysed here (i.e. no DNA barcode gap).

Discussion

This study provides comparative assessments of universality, resolvability and benefits of combining loci for seven candidate plastid barcoding loci for land plants. Of these loci, only *rpoC1* worked across all three taxonomic groups with a single set of PCR conditions. This is an impressive performance given the range of taxonomic diversity encompassed. However, the trade-off in the universality of *rpoC1* is its relatively low levels of species discrimination. For each taxonomic group, there were two to three other regions that showed greater levels of resolution. The conflicting requirements of universality and resolvability mean that no one region performs well in all cases: *rpoB* was the best region in *Araucaria*, but had less variation in *Inga* and did not amplify in *Asterella* s.l. In *Inga*, *matK* was the best performing region (but required internal sequencing primers in a small number of samples), it failed in *Asterella* s.l., and showed intermediate levels of success in *Araucaria* despite the combined *matK* amplicon size being *c.* 1000 bp. *rbcL*, worked well in *Asterella* s.l., showed intermediate success in *Araucaria*, but required additional primers in *Inga*. Of the noncoding regions, *trnH-psbA* was the most universal and worked well in *Inga*, but sequencing was difficult in *Araucaria* (in part due to the large size of the region), and it showed lower levels of species discrimination in *Asterella* s.l. than *rbcL*. We had relatively limited success with *atpF-atpH* and *psbK-psbI*. For both of these regions, despite a modest number of optimization attempts, we did not obtain sequence data for some taxonomic groups, and when these regions worked, they did not show high levels of species discrimination.

So, where does this leave us in the search for a standard approach to DNA barcoding in land plants? As a starting point, we evaluate our results in light of the other proposed barcoding locus solutions for plants. Lahaye *et al.* (2008a, b) recommended *matK* alone as a universal barcode for flowering plants. There is clear congruence in studies to date that *matK* is the most variable plastid coding region, and in the angiosperm group examined here, *matK* was the single most successful region in terms of species discrimination. However, other laboratories have reported difficulties in getting this region to work routinely with limited primer sets, even in studies just focusing on angiosperms (Chase *et al.* 2007; Fazekas *et al.* 2008; D. Erickson, Smithsonian Institute, personal communication, 2008; K. James, Natural History Museum London, personal communication, 2008). In the current study, over all three taxonomic groups and using multiple primer sets, the use of *matK* alone gave us 17% species discrimination, and suggests that a *matK*-only barcoding solution for land plants is likely to involve a high proportion of PCR/sequencing failures with current protocols and low resolution in some groups.

Data from our study support the notion that a multilocus barcoding solution is more appropriate than focusing on a single locus. By incorporating loci that perform well over broad phylogenetic distances (high universality), all samples can be given an approximate identification (to at least a group of species). Additional barcoding loci can then increase the proportion of cases in which species-level discrimination is achieved. In our data sets, potential levels of species discrimination are higher for multilocus barcoding solutions than for any single locus. In *Inga*, for instance, 65% of species shared an identical sequence with at least one other species for the best performing locus (*matK*). For some of the three-locus combinations, this dropped to 23%.

Of the multilocus solutions that have been proposed (Table 1), the percentage of species potentially distinguishable was almost identical, ranging from 48 to 53%. Slightly enhanced levels of success come from novel combinations of the loci from the existing proposals, and three 'mix and match' combinations (*rbcL* + *trnH-psbA* + *matK*, *rpoC1* + *rbcL* + *trnH-psbA*, and *rpoC1* + *rbcL* + *matK*) all have corresponding success values between 57% and 60%. Of these combinations, the one with the lowest proportion of missing data (either as total missing data, or the amount of missing data if just a single primer pair had been used for each locus) is *rpoC1* + *rbcL* + *trnH-psbA*. The combination of *rpoC1* + *rbcL* + *matK* had only marginally more missing data, but a slightly higher level of discrimination, and being entirely coding avoids complications associated with highly length-variable regions. One point worth stressing is that the generally high level of species discrimination for regions that worked, but greater PCR/sequencing failure rates in *Asterella* s.l. compared to the other groups, means that a successful 'liverwort locus' contributes disproportionately to the final totals.

Our data suggest some combination of *rbcL*, *rpoC1*, *matK* and *trnH-psbA* as the land plant barcoding solution. The inclusion of a locus like *rpoC1* would act as a strong universal tag, from which all samples will get an approximate identification. Certainly, it will make the management of a multilocus barcode database for plants easier if at least one of the loci is easily recoverable from almost every sample with standard conditions. In groups such as *Inga*, a three-locus system offered greater potential resolution than the best two-locus solution, and at this stage, a three-locus solution may prove a pragmatic insurance policy should any of the selected loci prove to be completely recalcitrant in as yet unstudied taxonomic groups (a point also made by Fazekas *et al.* 2008). However, it is important to stress that there is simply no perfect solution. Based on these data, there is no clear evidence to argue that any one single option is much better than several other possibilities. Essentially, all these loci are sub-optimal in one way or another, and a number of different locus combinations would probably end up with similar performance from the system. Similar conclusions were reached by Fazekas *et al.* (2008). The most important point is for the plant barcoding community to settle on a consensus solution and to follow this up with targeted investment enhancing laboratory protocols and informatics tools for the regions that this involves. Several other research groups are currently comparing this same set of regions, and efforts are underway to compile the findings into an overarching review. This will enable the final decision to be based on evaluation of a broader set of samples than the three genera considered here. Some points to be followed up from the best performing regions identified here include (i) assessing whether the patchy performance of the *rbcL* barcoding primers within *Inga* is a problem for other angiosperm groups, (ii) quantification of the extent of universality problems for *matK*, and (iii) quantification of the frequency with which length variation and microsatellites necessitate extensive manual editing of electropherograms and/or leads to partial reads in the noncoding regions.

Levels of species discrimination and DNA barcode 'gaps' in land plants

Prior to evaluating the percentage of plant species distinguishable, it is worth making a brief comment on the success criteria used. Interspecific sharing of identical sequences or failure of conspecific individuals to 'group together' are considered as straightforward failures. Conversely, where all individuals of a species group together exclusively, this is treated as successful discrimination. Where just a single individual was sampled from a species, and the sequence obtained was unique, this is treated as potentially distinguishable and included as a 'success'. However, unique substitutions in single samples do not necessarily translate to an ability to discriminate species, and our estimates of success should be considered as upper estimates; percentages may fall with further intraspecific sampling. In addition, sampling of additional species within each of these groups may lower the overall percentage of species distinguishable.

We encountered considerable heterogeneity in levels of species discrimination among the three groups. In *Inga* and *Araucaria*, the success rate was limited and identical sequences were frequently recovered from species that are clearly distinct on morphological grounds. Levels of discrimination are < 30% based on single loci in these two groups and upper estimates of 32% in *Araucaria* and 69% in *Inga* when multiple loci are used. However, the improvement in *Inga* when multiple loci are used comes entirely from the transition from singleton-sampled-species for which sequences were shared based on individual loci, shifting to being unique as more loci are added. There was no increase in grouping together of the accessions from species from which multiple individuals were sampled. The success rate in these cases stayed resolutely at one to two species out of seven regardless of how many loci were added, suggesting that the 69% success rate is an upper estimate. With additional intraspecific sampling (and representation of an increased number of species), we would expect this figure to fall.

Inga to some extent represents a known 'difficult challenge'. The genus has undergone much speciation within the last 10 million years (Richardson *et al.* 2001). In *Araucaria*, the barcode discrimination problems are primarily among the New Caledonian species, which have previously been reported as showing low levels of *rbcL* divergence (Setoguchi *et al.* 1998). Dating the divergence of the New Caledonian species is complicated by the relatively slow rates of molecular evolution in this group (Kranitz 2005). However, although the underlying causes may differ in *Inga* and *Araucaria*, the end result is the same. The rate of speciation outstrips the rate of accumulation of species-specific differences. In *Asterella* s.l., figures were much more encouraging (> 90%). Compared to higher plants, lower plant groups in general are character-poor and have received less taxonomic attention. There is an expectation that species limits may be broader, and one prediction is that DNA barcoding will be particularly useful in helping to identify 'cryptic' species in such groups.

Over all three taxonomic groups, our best locus combinations gave an upper estimate of *c.* 60% species discrimination. There are few empirical figures in the literature with which to compare this. Newmaster *et al.* (2008) were able to distinguish all tested individuals from six out of eight sampled species of *Compsoneura* (DC) Warb. (nutmegs), using *matK* and *trnH-psbA*. Fazekas *et al.* (2008) found that various combinations of up to seven plastid barcoding loci gave an upper limit of *c.* 70% of species distinguishable in a Canadian floristic study based on 92 species from 32 genera. Lahaye *et al.* (2008a) report species discrimination figures of 90% and above, based on their analysis of orchids and the flora of the Kruger National Park. They noted that 'we may need to accept that no more than ~90% of species will be identified with universal plastid barcodes' (p. 2927). However, this 90% relates to data sets with limited sampling of multiple species from the same genus. When Lahaye *et al.* extended their sampling to a large group of Mesoamerican orchids with extensive intra-generic sampling, levels of species discrimination were much lower (Hollingsworth 2008; Lahaye *et al.* 2008a).

At this point, it is difficult to come up with a reliable global estimate of how barcoding will perform in land plants given the small number of reports to date. We do, however, predict that the percentage of plant species distinguishable by barcoding will be lower that the 90%

suggested by Lahaye *et al.* (2008a). Recently diverged species will often show a lack of intraspecific coalescence/shared haplotypes; plant species frequently hybridize (Mallet 2005; Stace 1975), and there are many examples of plastid introgression (Rieseberg & Soltis 1991; Rieseberg & Carney 1998). Based on the available data, we expect that the final figure for species-level discrimination using plastid barcodes in plants will be < 70%, and it is clear from our results and others (e.g. Lahaye *et al.* 2008a; Newmaster *et al.* 2008) that there is no evidence of a clear discontinuity between intra- and interspecific divergences.

Thus, levels of species discrimination from a plastid barcode system in plants will not be perfect. However, it is not our intention to be negative. In some groups, the approach will work well, and in others, the best that will be achieved is identification to a group of species (Chase *et al.* 2007; Hollingsworth 2008). In many cases, this latter level of resolution will be adequate. Where it is not, additional data sources (such as ITS in taxa in which it is suitable) will be required to achieve species-level resolution for specific applications, but the key point is that adoption of a standardized barcoding approach now marks the first stage of the coordinated use of DNA sequence data at the species level for plants. This involves routine collection of DNA-ready material for herbaria (including intraspecific sampling), establishing informatics systems capable of handling the data, and implementing appropriate data standards for these systems. Future technological developments will undoubtedly enhance the levels of species discrimination achievable and after a period of relative stasis, sequencing technologies are undertaking quantum leaps (Ellegren 2008; Hudson 2008). Thus, the system should be expected to evolve and change as new technologies come on stream, but for now, existing technologies are adequate to commence the process of routinely incorporating DNA sequence data into an automatable, scalable system for plant taxonomy and plant identifications.

Acknowledgements

We are grateful to Jane Squirrell and Vimi Lomax for assistance in the laboratory, Terry Pennington, Tania Brenes, Phyllis Coley, Tom Kursar, Martin Gardner, Chris Kettle, Mai-lan Kranitz and Phil Thomas for assistance in the field, Daniela Schill for provision of liverwort DNAs, Ki-Joong Kim and Mike Wilkinson for unpublished primer sequences, and to the Sloan and Moore foundations and the Scottish Government Rural and Environment Research and Analysis Directorate for funding. We are very grateful to the reviewers of this paper for their thoughtful and helpful comments.

References

Bischler H (1998) Systematics and evolution in the genera of the Marchantiales. *Bryophytorum Bibliotheca*, **51**, 1–201.

Chase MW, Salamin N, Wilkinson M *et al.* (2005) Land plants and DNA barcodes: short-term and long-term goals. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **360**, 1889–1895.

Chase MW, Cowan RS, Hollingsworth PM *et al.* (2007) A proposal for a standardised protocol to barcode all land plants. *Taxon*, **56**, 295–299.

Ellegren H (2008) Sequencing goes 454 and takes large-scale genomics into the wild. *Molecular Ecology*, **17**, 1629–1631.

Fazekas AJ, Burgess KS, Kesanakurti PR *et al.* (2008) Multiple multilocus DNA barcodes from the plastid genome discriminate plant species equally well. *PLoS ONE*, **3**, e2802.

Hammer Ø, Harper DAT, Ryan PD (2001) PAST: paleontological statistics software package for education and data analysis. *Palaeontologia Electronica*, **4**, 9. http://palaeo-electronica.org/2001_1/past/issue1_01.htm.

Hebert PDN, Cywinska A, Ball SL, deWaard JR (2003) Biological identifications through DNA barcodes. *Proceedings of the Royal Society B: Biological Sciences*, **270**, 313–321.

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*. *Proceedings of the National Academy of Sciences, USA*, **101**, 14812–14817.

Hollingsworth PM (2008) DNA barcoding plants in biodiversity hotspots: progress and outstanding questions. *Heredity*, **101**, 1–2.

Hudson ME (2008) Sequencing breakthroughs for genomic ecology and evolutionary biology. *Molecular Ecology Resources*, **8**, 3–17.

Kranitz M-L (2005) *Systematics and evolution of New Caledonian* Araucaria. Unpublished PhD Thesis, University of Edinburgh and the Royal Botanic Garden Edinburgh, Edinburgh, UK .

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.

Kress WJ, Wurdack KJ, Zimmer EA, Weigt LA, Janzen DH (2005) Use of DNA barcodes to identify flowering plants. *Proceedings of the National Academy of Sciences, USA*, **102**, 8369–8374.

Lahaye R, van der Bank M, Bogarin D *et al.* (2008a) DNA barcoding the floras of biodiversity hotspots. *Proceedings of the National Academy of Sciences, USA*, **105**, 2923–2928.

Lahaye R, Savolainen V, Duthoit S, Maurin O, van der Bank M (2008b) A test of *psbK-psbI* and *atpF-atpH* as potential plant DNA barcodes using the flora of the Kruger National Park (South Africa) as a model system. Available from *Nature Precedings* <http://hdl.handle.net/10101/npre.2008.1896.1>.

Ledford H (2008) Botanical identities: DNA barcoding for plants comes a step closer. *Nature*, **415**, 616.

Long DG (2006) Revision of the genus *Asterella* P. Beauv. in Eurasia. *Bryophytorum Bibliotheca*, **63**, 1–299.

Long DG, Möller M, Preston J (2000) Phylogenetic relationships of *Asterella* (Aytoniaceae, Marchantiopsida) inferred from chloroplast DNA sequences. *The Bryologist*, **103**, 625–644.

Mallet J (2005) Hybridization as an invasion of the genome. *Trends in Ecology & Evolution*, **20**, 229–237.

Meier R, Shiyang K, Vaidya G, Ng PKL (2006) DNA barcoding and taxonomy in Diptera: a tale of high intraspecific variability and low identification success. *Systematic Biology*, **55**, 715–728.

Mower J, Touzet P, Gummow J, Delph L, Palmer J (2007) Extensive variation in synonymous substitution rates in mitochondrial genes of seed plants. *BMC Evolutionary Biology*, **7**, 135.

Newmaster SG, Fazekas AJ, Steeves RAD, Janovec J (2008) Testing candidate plant barcode regions in the Myristicaceae. *Molecular Ecology Notes*, **8**, 480–490.

Pennington TD (1996) *The Genus* Inga*: Botany*. Royal Botanic Gardens, Kew, Kew, UK.

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

Rambaut A (2002) Se-Al: sequence alignment editor version 2. http://tree.bio.ed.ac.uk/software/seal/.

Richardson JE, Pennington RT, Pennington TD, Hollingsworth PM (2001) Recent and rapid diversification of a species rich Neotropical rain forest tree genus. *Science*, **293**, 2242–2245.

Rieseberg LH, Carney SE (1998) Plant hybridization. *New Phytologist*, **140**, 599–624.

Rieseberg LH, Soltis DE (1991) Phylogenetic consequences of cytoplasmic gene flow in plants. *Evolutionary Trends in Plants*, **5**, 65–84.

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.

Schill DB (2006) *Taxonomy and phylogeny of the liverwort genus* Mannia *(Aytoniaceae, Marchantiales)*. Unpublished PhD Thesis, University of Edinburgh and the Royal Botanic Garden Edinburgh, Edinburgh, UK.

Setoguchi H, Osawa TA, Pintaud C, Jaffré T, Veillon J-M (1998) Phylogenetic relationships within Araucariaceae based on *rbcL* gene sequences. *American Journal of Botany*, **85**, 1507–1516.

Shearer TL, Coffroth MA (2008) Barcoding corals: limited by interspecific divergence, not intraspecific variation. *Molecular Ecology Resources*, **8**, 247–255.

Smith MA, Woodley NE, Janzen DH, Hallwachs W, Hebert PDN (2006) DNA barcodes reveal cryptic host-specificity within the presumed polyphagous members of a genus of parasitoid flies (Diptera: Tachinidae). *Proceedings of the National Academy of Sciences, USA*, **103**, 3657–3662.

Stace CA (1975) *Hybridization and the Flora of the British Isles*. Academic Press, London.

Swofford DL (2003) *PAUP* 4.0 b10 Phylogenetic Analysis Using Parsimony (*and Other Methods)*, Version 4. Sinauer Associates, Sunderland, Massachusetts.

Whitworth TL, Dawson RD, Magalon H, Baudry E (2007) DNA barcoding cannot reliably identify species of the blowfly genus *Protocalliphora* (Diptera: Calliphoridae). *Proceedings of the Royal Society B: Biological Sciences*, **274**, 1731–1739.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Appendix S1 Plant material, collection details and GenBank accession numbers of material used for comparative evaluation of seven candidate DNA barcoding regions in three groups of land plants. All vouchers are housed at E (herbarium, Royal Botanic Garden Edinburgh) unless otherwise indicated; – indicates no sequence was obtained

Appendix S2 PCR primers used for evaluation of seven candidate DNA barcoding regions in three groups of land plants. Reference codes are referred to in Table 2 and Appendix S4

Appendix S3 PCR conditions used for evaluation of seven candidate DNA barcoding regions in three groups of land plants. See Appendix S4 for which conditions were successful in which taxon/primer combinations

Appendix S4 Universality assessment of seven candidate DNA barcoding regions in three groups of land plants. PCR protocol names correspond to details of the reaction conditions given in Appendix S3, Supporting Information. The letter in square brackets following primer names corresponds to a citation reference in Appendix S2. PCR optimization was classified as: low, used a single set of PCR conditions; medium, 2–5 attempts made, varying PCR conditions; high, > 5 attempts made, extensive optimisation attempted. Trimmed matrix character number refers to the total number of characters considered based on an aligned matrix

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.