DNA barcoding in plants: taxonomy in a new perspective

K. Vijayan[1,2,*] and C. H. Tsou[1]

[1]Institute of Plant and Microbial Biology, Academia Sinica, Taipei, Taiwan-115, ROC
[2]Present address: Central Silk Board, BTM Layout, Madiwala, Bangalore 560 068, India

DNA barcoding is the process of identification of species based on nucleotide diversity of short DNA segments. It is well established in animals with the introduction of cytochrome *c* oxidase subunit 1 (COI) as a standard barcode. In plants, however, due to the difficulty in finding a universally acceptable barcode, it is yet to be well established. Based on the relative efficacy testing, the Consortium for the Barcode of Life–Plant Working Group has recently identified a few loci as potential barcode candidates and from them a two-locus standard barcode (*rbcL* + *matK*) has been recommended for initiating the barcoding process of plant species. With 70% species discriminatory power, this two-locus barcode is capable of serving many projects, but for better resolution additional loci need to be used. This article provides an overview of the technical details, and merits and demerits of these loci as plant barcodes.

Keywords. Barcoding, DNA sequence, locus, plants, taxonomy.

DNA BARCODING is a relatively new concept that has been developed for providing rapid, accurate and automatable species identification using standardized DNA sequences as tags[1–3]. In fact, it started with the seminal work of Hebert *et al.*[1], who demonstrated that individual species from a collection of 200 closely allied species of lepidopterans could be identified with 100% accuracy using the mitochondrial gene cytochrome *c* oxidase subunit I (COI). Barcoding is now a well-established technique for species identification in animals. In DNA barcoding, the unique nucleotide sequence patterns of small DNA fragments (400–800 bp) are used as specific reference collections to identify specimens and to discover overlooked species[4,5]. Thus, the initial goal of the DNA barcoding process is to construct on-line libraries of barcode sequences for all known species that can serve as a standard to which DNA barcodes of any identified or unidentified specimens can be matched[6]. This can alleviate several inherent problems associated with traditional taxonomic identification, based on morphological characters, such as wrong identification of species due to phenotypic plasticity and genotypic variability of the characters, overlooking cryptic taxa, difficulty in finding reliable characters due to long maturity periods, etc.[5]. DNA barcoding, thus, can provide the taxonomists, conservationists and others who need the identification of species, a cost-effective and efficient tool, much as a barcode that identifies supermarket products[7,8]. It is especially of much use in areas where species identification with morphological charterers is not practicable due to extensive damage or delayed expression[9]. In the context of the rapid economic developments and the consequent impacts on the flora and fauna of various nations, especially in the tropical and subtropical regions, identification of species using a faster method is essential to evaluate the biodiversity of these regions in order to preserve the rare endemic and endangered species. However, it should be borne in mind that DNA barcoding is not an alternative to taxonomy and it cannot replace taxonomy as such, but is a useful tool that creates information on unknown taxa[10]. In order to promote the use of DNA barcoding for all eukaryotic life in this planet, a Consortium for the Barcode of Life (CBOL) was established in May 2004, which currently includes more than 120 organizations from 45 nations[6]. With the support of CBOL, the effort of DNA barcoding has been slowly progressing with controversies and intense debates[10–12]. A brief history of the Barcode of Life initiative is available at http://www.dnabarcodes.org/pa/ge/history_of_boli. Other useful information on DNA barcoding is also available at www.barcodinglife.org, http://barcoding.si.edu/, http://www.dnabarcoding.ca, http://www.kew.org/barcoding and http://www.ibolproject.org.

Basic features of barcoding sequences

The most important characteristic features of a DNA barcode are its universality, specificity on variation and easiness on employment. This means that the gene segment used as a barcode should be suitable for a wide range of taxa, should have high variation between species but should be conserved within the species, so that the intra-specific variation will be insignificant[4,5,12]. Consequently, an ideal DNA barcode should also be routinely retrievable with a single primer pair, and should be amenable to bidirectional sequencing with little requirement for manual editing of sequence traces[4,5,12]. Additionally,

*For correspondence. (e-mail: kvijayan01@yahoo.com)

Table 1. DNA segments tested for their suitability and recommendations of DNA markers for barcoding in land plants

DNA segment tested for suitability	Proposed/recommended	Reference
nrITS, *atpB–rbcL*, *psbM–trnD*, *trnC–ycf6*, *trnH–psbA*, *trnL–F*, *trnK–rps16*, *trnV–atpE rpl36–rps8*, *ycf6–psbM*	*nrITS* and *trnH–psbA*	4
nrITS, *rbcL*, *trnH–psbA*	*nrITS* and *trnH–psbA*	4
ITS1, *accD*, *ndhJ*, *matK*, *trnH-psbA*, *rbcL*, *rpoB*, *rpoC1*, *ycf5*	*rbcL* and *trnH–psbA*	41
atpF–atpH, *atpH–atpI*, *rps15–ycf1*, *ndhG–ndhI*, *psbK–psbI*, *petA–psbJ*, *trnH-psbA*	*atpF–atpH* + *psbK–psbI*	74
rpoC1, *rpoB*, *matK*, *trnH-psbA*, *nr*ITS, *trnL-F*	*rpoC1* + *rpoB* + *matK* or *rpoC1* + *matK* + *trnH-psbA* (on the basis of available merits and demerits)	14
nrITS, *accD*, *ndhJ*, *matK*, *trnH-psbA*, *rpoB*, *rpoC1*, *ycf5*	*NrITS*	15
trnL (UAA) intron	*trnL9UAA* intron	3
accD, *matK*, *trnH–psbA*, *rbcL*, *rpoB*, *rpoC1*, UPA	*matK* and *trnH–psbA*	77
matK, *trnH-psbA*, *psbK-psbI*, atpF-atpH	*matK* or *matK* + *trnH–psbA* + *psbK–psbI*	72
accD, *matK*, *trnH-psbA*, *rbcL*, *rpoB*, *rpoC1*, *ycf5*, *ndhJ*	*matK* or *matK* + *trnH–psbA*	9
Cox1, *23SrDNA*, *rpoB*, *rpoC1*, *rbcL*, *matK*, *trnH–psbA*, *atpF–atpH*, *psbK-psbI*	*rbcL*, *rpoB*, *matK*, *trnH–psbA*, *atpF–atpH*	57
CO1, *rpoC*, *rpoB*, *rbcL-a*, *matK*, *trnH–psbA*	*trnH–psbA* + *rbcL-a* [based on probability of correct identification (PCI)	78
atpF–atpH, *rpoB*, *rpoC1*, *rbcL*, *matK*, *psbK–psbI*, *trnH–psbA*	*rbcL* + *rpoC1* + *matK* + *trnH–psbA*	32
accD, *matK*, *ndhA*, *ndhJ*, *ndhK*, *rpl22*, *rpoB*, *rpoC1*, *rpoC2*, *ycj2*, *ycf5*, *ycf9*	*matK*, *rpoB*, *rpoC*1, *ndhJ*, *ycf*5 and *accD*	50
matK, *rbcL*, *rpoB*, *rpoC1*, *trnH-psbA*	*matK*	51
atpF–atpH, *matK*, *rbcL*, *rpoB*, *rpoC1*, *psbK–psbI*, *trnH–psbA*	*rbcL* + *matK*	5

it is often required that degraded DNA is to be used for PCR amplification. Therefore, the gene sequences used for barcoding should be short enough to be PCR-amplified easily. In general, DNA barcoding is based on the use of a short, standard region that enables cost-effective species identification. DNA barcoding is well-established in animals using COI in mtDNA, which is 648 nucleotide base pairs long and is flanked by regions of conserved sequences, making it relatively easy to PCR amplify, sequence and analyse. A growing number of studies have shown that COI sequence variability is low within species (generally less than 1–2%), but differs by several per cent in closely related species in animals, thus making it possible to identify species with high confidence[1]. However, in plants, the mtDNA has low substitution rates and a rapidly changing gene content and structure, which makes COI unsuitable for barcoding in plants[13]. Therefore, it was necessary to search for suitable gene sequences for barcoding the plant species. Based on the information available from phylogenetic studies, a number of gene sequences, both coding and non-coding sequences from the chloroplast DNA along with a gene from nuclear DNA have been examined for their suitability as barcodes (Table 1). The characteristic features of the 12 loci proposed by different authors, including CBOL–Plant Working Group for DNA barcoding in plants are described below.

Nuclear genome sequence

Although barcoding based on the biparentally inherited nuclear DNA segment is expected to provide more information on species identity, including hybridization events, till date internal transcribed spacer (ITS) regions of the ribosomal DNA (rDNA) are the only nuclear DNA that have been tested for suitability as barcodes in plants[4,14,15]. The difficulty in obtaining high universality of the PCR amplification of single or low-copy genes, especially from degraded and low-quality DNA and the low species discriminatory power due to conservation of functional genes across large lineages could be the major reasons why such limited number of genes are being tested.

Internal transcribed spacer regions of nuclear ribosomal cistron

The rDNA cistron is a multigene family encoding the nucleic acid core of the ribosome. Within the cell, the rDNA is arranged as tandemly repeated units containing *18S*, *5.8S*, *26S* coding regions and two internal transcribed spacers (*ITS1* and *ITS2*) present on either side of *5.8S* region (Figure 1). Generally, the rDNA units are reiterated thousands of times and are organized into large blocks in the chromosome called the nucleolar organizer regions[16,17]. One of the most remarkable features of the rDNA is that

the individual unit of this multiple gene family does not evolve independently; instead all the units evolve in a concerted manner such that higher level of overall sequence homogeneity exists among copies of the rDNA within a species, but differs among different species[18,19]. This high sequence homogeneity is achieved through a process initially termed as horizontal evolution or coincidental evolution[18], but later renamed as concerted evolution[20], which involves unequal crossing over and gene conversions. Currently, *nrITS* is considered as one of the most useful phylogenetic markers for both plants and animals, because of its ubiquitous nature, biparental inheritance, and comparatively higher evolutionary changes due to less functional constrains[21–24]. Likewise, species-level discrimination and technical ease have also contributed to its wider acceptability as a powerful phylogenetic marker. Another advantage is that the ITS1 and ITS2 regions can be PCR-amplified separately by anchoring primers in the conserved coding genes. This facilitates easy amplification of ITS even from poor quality or degraded DNA[4]. Universal primers are also available for amplification of ITS1 and ITS2 regions (Table 2). However, some of the recent reports in tree plants and asexually propagated plants revealed the presence of some degree of intra-individual variations among the copies of ITS1 and ITS2 sequences[25–30]. Various reasons such as recent hybridization, lineage sorting, recombination among copies, high mutation rate and pseudogene formation of cistrons are considered to be the reasons for such variations[21,23,24,26,31]. Nevertheless, *nrITS* is still considered to be a powerful phylogenetic tool at the species level[29]. When tested for its suitability as barcode in plants, *nrITS* along with nine other loci from the chloroplast genome, showed better universality (88%) and species discrimination than the chloroplast loci[4] (Table 1). Considering the availability of universal primers, presence of multiple copies in cells, high universality and good species discriminatory power, *nrITS* was proposed as a potential candidate for barcoding in plants[4,15]. Hollingsworth *et al.*[32] later endorsed this earlier view by suggesting that *nrITS* can be considered for barcoding of species that have limited variations in the plastid genome. However, perhaps the problems arising from paralogous sequences, pseudogenes and intragenomic variability and difficulties in direct sequencing of PCR products, the CBOL-Plant Working Group[5] has not regarded *nrITS* suitable for a universal plant DNA barcode, but as a supplementary locus for taxonomic groups which have less resolution with cpDNA and where direct sequencing of *nrITS* is possible. It has recently been used as a barcode for identifying a reproductively isolated and cryptic species of *Asimitellaria* from its close relatives[33]. On the basis of this observation, it was suggested that nrDNA can be of use for accurate and efficient delimitation of plant biological species in lineages with various life-history traits (annuals, perennials, trees, aquatics) and evolutionary backgrounds (recent and old radiations, oceanic island endemics)[33]. A recent report on the use of ITS2 to identify medicinal plants and their close relatives again proved the potential of this nuclear gene as a useful barcode for plants[34].

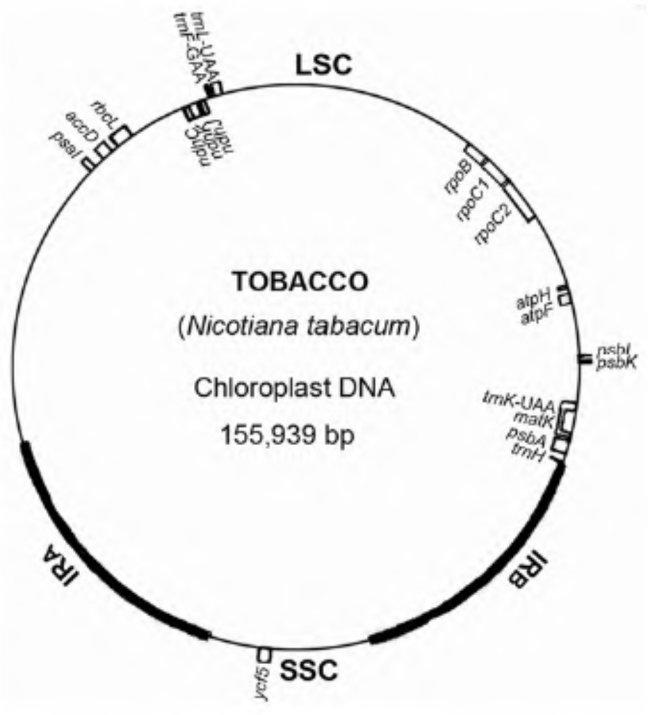

Figure 1. Structural organization of the chloroplast genome of *Nicotiana tabacum*. Locations of the seven loci proposed by different investigating group as potential candidates for barcoding in plants have been shown as described by Shinozaki *et al.*[36]. Genes shown outside the circle are encoded on the A strand and transcribed counter-clockwise. Genes shown inside are encoded on the B strand and transcribed clockwise. LSC, Large single copy region; IR, Inverted repeat; SSC, Small single copy region.

Chloroplast genome sequences

The chloroplast genome shares several attributes of mitochondrial genomes such as conserved gene order, high copy number per cell, amenability to PCR amplification and availability of universal primers. Hence, chloroplast genes could be considered as analogous to the mitochondrial gene that has been used for DNA barcoding in animals. However, compared to mtDNA genes in animals, chloroplast genes in plants have slower rate of evolution; therefore, finding suitable gene sequences with sufficient species discriminatory power is a great challenge[13]. Nonetheless, due to the nature of uniparental inheritance, nonrecombination and structural stability in both the genic and intergenic regions of the chloroplast, many genes have been examined carefully for their potentiality as barcodes in plants[7,35].

Chloroplast genome of higher plants is a circular structure with a size of 120–160 k bp (Figure 1). The general

Table 2. Universal primers for PCR amplification of ITS1–5.8S–ITS2 region of the nrDNA cistron

Primer	Sequence (5′ → 3′)	Comments
ITS1	TCCGTAGGTGAACCTGCGG	Forward; anchoring at the 5′ region of 18S (ref. 98)
ITSleu1	GTCCACTGAACCTTATCATTTAG	Forward; anchoring at the 5′ region of 18S (ref. 99)
ITS2	GCTGCGTTCTTCATCGATGC	Forward; amplifying 5.8S (ref. 98)
ITS3	GCATCGATGAAGAACGCAGC	Reverse; amplifying 5.8S (ref. 98)
ITS4	TCCTCCGCTTATTGATATGC	Reverse; anchoring at the 5′ region of 26S (ref. 98)
ITS5	GGAAGTAAAAGTCGTAACAAGG	Forward; anchoring at the 5′ region of 18S (ref. 98)
ITS1-F	CTTGGTCATTTAGAGGAAGTAA	Forward; anchoring at the 5′ region of 18S (ref. 100)
ITS4-B	CAGGAGACTTGTACACGGTCCAG	Reverse; anchoring at the 5′ region of 26S (ref. 100)

architecture of the chloroplast genomes is represented by a large and a small single-copy region (LSC and SSC) intervened by two copies of a large inverted repeat (IRa and IRb). The chloroplast genome contains all the rRNA genes (four genes in higher plants), tRNA genes (35 genes) and other genes for those proteins synthesized in the chloroplast (~100 genes) that are essential for its existence[36]. On the basis of the considerable amount of information available from phylogenetic studies and recent testing with limited number of taxa, potentially useful genic and intergenic loci were initially selected as potential candidates for testing as barcodes for the land plants (Table 1). Efficacy of these sequences as barcodes has been examined individually and in combination with other loci on a large number of samples from a wide range of species covering all the major taxonomic lineages. The following loci were proposed by different investigating groups. Primers suitable for these genes are available in Table 3.

rbcL gene sequence

Among the plastid genes, *rbcL* is the best characterized gene sequence. Therefore, most of the investigating groups tested its suitability in barcoding (Table 1). It encodes the large subunit of rubilose-1,5-bisphosphate carboxylase/oxygenase (RUBISCO). As RUBISCO is a critical photosynthetic enzyme, *rbcL* was the first gene that was sequenced from the plants[37]. *rbcL* has been used so extensively in plant phylogenetic studies that more than 10,000 *rbcL* sequences are already available in GenBank[14,38]. Because of this wide utility, various aspects of the molecular evolution of *rbcL* have also been studied in detail[39]. Most of the phylogenetic studies suggest that *rbcL* is best suited to reconstruct the relationships down to the generic levels, but is not useful for specific levels[40]. Furthermore, in order to obtain enough species discrimination, the entire ~1430 bp needs to be sequenced, which acts as a limiting factor for its use as a barcoding sequence because an ideal DNA barcoding region should be short enough to amplify from degraded DNA and analysed via single-pass sequencing[14]. One solution for this was to amplify short sequences with enough variability. Primers for PCR amplification and sequencing for such short sequence within the *rbcL* gene have been developed accordingly for most of the taxa (Table 3)[41,42]. Owing to the ease in PCR amplification across a wide range of plant groups and the availability of sequence information in many plant groups, the CBOL–Plant Working Group[5] has recently recognized *rbcL* as one of the most potential gene sequences for DNA barcoding in plants. However, because of the low species discrimination, most of the investigating groups are of the opinion that *rbcL* should be used in conjunction with other markers[5,14,32,40]. Therefore, the CBOL–Plant Working Group[5] recommended a combination of *rbcL* and *matK* as the standard two-locus barcode for plants, because this combination of genes appears to be a pragmatic solution to a complex trade-off among universality, sequence quality discrimination and cost.

matK gene sequence

Among the chloroplast genes, *matK* is one of the most rapidly evolving genes. It has a length[43] of about 1550 bp and encodes the enzyme maturase which is involved in the splicing of type-II introns from RNA transcripts[44]. Since *matK* is embedded in the group II intron of the lysine gene *trnK*, it can be easily PCR-amplified with a primer set designed from the conserved regions of the genes *trnK*, *rps16* and *psbA*. *matK* has been used as a marker to construct plant phylogenies because of its rapid evolution and the ubiquitous presence in plants[45,46]. However, failure of PCR amplification of *matK* in some taxonomic groups was also reported[13]. In order to circumvent this problem, new sets of primers were developed, which work well in most of the major taxonomic groups (Table 3)[47]. This primer set amplifies a DNA fragment of ~930 bp between positions 429 and 1313 of the *matK* sequence[47,48]. Phylogenetically, the rate of evolution of *matK* was found suitable for resolving intergeneric as well as interspecies relationships in many angiosperms[40,49]. Considering the high evolutionary rate of *matK*, it has been tested by several workers for suitability as a

Table 3. Suitable primer sets for the candidate regions selected for development as potential barcoding markers by different investigating groups (forward →; reverse ←)

Gene	Primer	Sequence 5′ → 3′	Type	Comments	Reference
rbcL	a–f	ATGTCACCACAAACAGAGACTAAAGC	→	For 550–600 bp from the 5′ region in all major	41
	a–r	CTTCTGCTACAAATAAGAATCGATCTC	←	taxonomic groups	
	1F	ATGTCACCACAAACAGAAAC	→	734 bp from the 5′ region in angiosperms	42
	724R	TCGCATGTACCTGCAGTAGC	←		
accD	accD 1	AGTATGGGATCCGTAGTAGG	→	170–381 bp is amplified with primers	50
	accD 2	GGRGCACGTATGCAAGAAGG	→	1 + 3/1 + 4/2 + 3/2 + 4. None is universal, but	
	accD 3	TTTAAAGGATTACGTGGTAC	←	suitable for major taxonomic groups	
	accD 4	TCTTTTACCCGCAAATGCAAT	←		
matK	matK 2.1	CCTATCCATCTGGAAATCTTAG	→	734–821 bp is amplified with 2.1 + 3.2/2.1 +	50
	matK 2.1a	ATCCATCTGGAAATCTTAGTTC	→	5/2.1a + 3.2/2.1a + 5/X + 5. None is universal, but	
	matK X	TAATTTACGATCAATTCATTC	→	suitable for major taxonomic groups, except	
	matK 3.2	CTTCCTCTGTAAAGAATTC	←	*Araucaria* and *Diospyros*	
	matK 5	GTTCTAGCACAAGAAAGTCG	←		
	390F	CGATCTATTCATTCAATATTTC	→	Amplify ~930 bp between positions 429 and	47
	1326R	TCTAGCACACGAAAGTCGAAGT	←	1313 of the *matK* gene. Suitable for all angiosperms	
	matK_1F	GAACTCGTCGGATGGAGTG	→	1F + 1R amplifies sequences between 1530 and	101
	matK_1R	GAGAAATCTTTTTCATTACTACAGTG	←	1928 bp. 2F + 2R between 2017 and 2567 bp of	
	matK_2F	CGTACTTTTATGTTTACAGGCTAA	→	*matK* gene. Suitable for conifers	
	matK_2R	TAAACGATCCTCTCATTCACGA	←		
rpoB	rpoB 1	AAGTGCATTGTTGGAACTGG	→	298–510 bp is amplified by 1 + 3/2 + 3/1 + 4/2 + 4.	50
	rpoB 2	ATGCAACGTCAAGCAGTTCC	→	Not suitable for *Equisetum*	
	rpoB 3	CCGTATGTGAAAAGAAGTATA	←		
	rpoB 4	GATCCCAGCATCACAATTCC	←		
rpoC1	rpoC1 1	GTGGATACACTTCTTGATAATGG	→	1 + 3/2 + 3/1 + 4/2 + 4 amplifies 467–564 bp.	50
	rpoC1 2	GGCAAAGAGGGAAGATTTCG	→	None is universal, but suitable for major taxonomic	
	rpoC1 3	TGAGAAAACATAAGTAAACGGGC	←	groups, except *Equisetum*	
	rpoC1 4	CCATAAGCATATCTTGAGTTGG	←		
ycf5	ycf5 1	GGATTATTAGTCACTCGTTGG	→	1 + 3/2 + 3/1 + 4/2 + 4 amplifies 221–382 bp in all	50
(ccsA)	ycf5 2	ACTTTAGAGCATATATTAACTC	→	the major taxonomic groups, except *Equisetum*	
	ycf5 3	ACTTACGTGCATCATTAACCA	←	and *Araucaria*	
	ycf5 4	CCCAATACCATCATACTTAC	←		
ndhJ	ndhJ 1	CATAGATCTTTGGGCTTYGA	→	1 + 3/2 + 3/1 + 4/2 + 4 amplifies 221–382 bp in all	48
	ndhJ 2	TTGGGCTTCGATTACCAAGG	→	the major taxonomic groups, except *Pinus*	
	ndhJ 3	ATAATCCTTACGTAAGGGCC	←		
	ndhJ 4	TCAATGAGCATCTTGTATTTC	←		
trnH–	trnH2	CGCGCATGGTGGATTCACAATCC	→	trnH2 + psbAF/trnH (GUG) + psb A works well in	102–
psbA	psbAF	GTTATGCATGAACGTAATGCTC	←	most of the major taxonomic groups to	104
	trn H (GUG)	ACTGCCTTGATCCACTTGGC	→	amplify 296–1120 bp fragment	
	psb A	CGAAGCTCCATCTACAAATGG	←		
atpF–	atpF	ACTCGCACACACTCCCTTTCC	→	196–573 bp good quality sequence from major	32
atpH	atpH	GCTTTTATGGAAGCTTTAACAAT	←	taxonomic groups	
psbK–	psbK	TTAGCCTTTGTTTGGCAAG	→	Not suitable for *Araucaria*	32
psbI	psbI	AGAGTTTGAGAGTAAGCAT	←		
trnL–F	trnL-c	CGAAATCGGTAGACGCTACG	→	trnL-c + trnL-f amplifies whole gene, intron and	83
	trnL-d	GGGGATAGAGGGACTTGAAC	←	intergenic regions. trnL-c + trnL-d amplifies	
	trnL-e	GGTTCAAGTCCCTCTTATCCC	→	introns (254–767 bp)	
	trnL-f	ATTTGAACTGGTGACACGAG3’	←	trnL-g + trnL-h amplifies the p6 loop of *trnL*(UAA)	3
	trnL-g	GGGCAATCCTGAGCCAA	→	intron (10–143 bp)	
	trnL-h	CCATTGAGTCTCTGCACCTATC	←		

plant barcode and has been proposed either alone or in combination with other loci (Table 1). For example, Ford *et al.*[50], after testing *matK* along with 11 other cpDNA loci in 98 land plant taxa, proposed a combination of *rpoC1* + *rpoB* + *matK* as the most promising combination for barcoding of land plants. Recently, Starr *et al.*[51], after testing *matK*, *rbcL*, *rpoB*, *rpoC1* and *trnH–psbA* as barcodes in Cyperaceae, also advocated the use of *matK* alone as a universal barcode for land plants. The CBOL–Plant Working Group[5] tested *matK* in nearly 550 plant species and found that nearly 90% of the angiosperm samples were easily amplified and sequenced using a single primer pair, though the success was limited in gymnosperms (83%) and much worse in cryptogams (10%). Because of this high universality and species discrimination, the CBOL–Plant Working Group[5] recommended *matK* in combination with *rbcL* as the standard two-locus barcode for plants.

rpoB and rpoC1 gene sequences

Genes *rpoB*, *rpoC1* and *rpoC2* encode three out the four subunits of the chloroplast RNA polymerase[36,52]. Genome-wide substitution analysis in a family like Geraniaceae revealed that *rpoB*, *rpoC1* and *rpoC2* are accumulating higher amount of nonsynonymous substitutions, indicating either positive or relaxed selection, and this high substitution rate makes these genes highly suitable for phylogenetic studies[53]. Likewise in Dipterocarpaceae, *rpoC* was found well suited for phylogenetic analysis[54]. Currently, *rpoB* has been considered as the core gene for phylogenetic analyses and identification of bacteria, especially when studying closely related isolates[55]. Together with the 16S rRNA gene, *rpoB* helps delineate new bacterial species and refine bacterial community analysis[55]. Logacheva *et al.*[56] also found that *rpoA*, *rpoB*, *rpoC1* and *rpoC2* as a group, are ideal for phylogenetic studies, but *rpoB* and *rpoC1* alone may not generate good results. Primers suitable for barcoding of *rpoB* and *rpoC1* are given in Table 3. After extensive studies, these genes have been proposed for barcoding either individually or in combination by various groups[14,32,50,57], though it was noted that PCR amplification of *rpoB* failed in *Araucaria*, *Ephedra*, *Equesetum*, *Isoetes*, *Lycopodium* and *Mannia*[50]. The CBOL–Plant Working Group[5], on the other hand, observed that the species discrimination of *rpoC1* was the least (43%) among the seven loci tested. Considering the low species discriminatory power, *rpoB* and *rpoC1* have been eliminated from further consideration for barcoding in plants by the CBOL–Plant Working Group, even though they have high universality and yield high-quality sequences[5]. Nonetheless, recently, *rpoC1* has been found highly useful for barcoding the bryophytes (mosses)[58]. Thus, further research on these gene sequences is required for deciding their suitability as a barcode.

accD gene sequence

The plastid *accD* gene encodes the β-carboxyl transferase subunit of acetyl-CoA carboxylase and is present in most flowering plants, except in grasses[59]. The *accD* gene sequence has been used for several phylogenetic studies in plants[60]. The 5′ coding region of *accD* evolved about five times faster than the *rbcL* coding region in some groups of plants like *Fagopyrum*[60]. Therefore, *accD* has been frequently tested for its suitability in DNA barcoding[9,15,50,41,61]. However, only Ford *et al.*[50] proposed it for DNA barcoding in some groups of plants, perhaps due to the complete absence of this gene in grasses.

ycf5 gene sequence

Ycf5 is the only gene from the small single-copy region being seriously studied for its suitability in DNA barcoding (Figure 1). *Ycf5* encodes a protein containing 313 amino acids[60]. This gene is conserved across all land plants[62] and has been tested for its suitability for DNA barcoding by several groups[9,15,41,63]. Although it yielded a high proportion of polymorphic sites, due to poor universality and problems in aligning sequences[50], the gene has not received much support. However, the barcoding site of the Royal Botanical Gardens, Kew (http://www.kew.org/barcoding/protocols.html) suggests that *ycf5* has the potential for barcoding some plants, if the problems in obtaining high-quality sequences are resolved. Considering the current advances in sequencing technologies, it is envisaged that this sequence may find its place in the list of genes useful for barcoding the plant species.

ndhJ gene sequence

The plastid *ndh* gene complex, identified originally from the tobacco plastid genomes[35] and liverwort[64], codes for subunits of a functional respiratory protein complex of size ~ 550 kDa within the mature chloroplast[65]. This gene complex consists of 11 genes, *ndhA–ndhK*. Among them, *ndhJ*[66], which was formerly called ORF159, was found together with *ndhC* and *ndhK* as a single operon (Figure 1)[65]. *ndhJ* encodes NADH dehydrogenase 30 kDa subunit. Since this gene was found useful for phylogenetic studies in plants[67], its suitability for barcoding was tested[9,15,50,60]. PCR amplification of *ndhJ* with new sets of primers (Table 3) was highly successful in many angiosperm groups, but it was reported that *ndhJ* is absent in *Pinus*[68] and *Cuscuta*[69]. Further, it was reported that all *ndh* genes are absent across Gnetales and Pinaceae, but present in other groups of gymnosperms[70]. Difficulties in PCR amplification of *ndhJ* in orchids were also reported[9]. These findings have significant implications on the universality of *ndhJ* as a barcode for land plants. Additionally, Lahaye *et al.*[9] observed low species discriminatory

power for *ndhJ*, but Ford *et al.*[50] found better species pair differences in liverworts, pteridophytes, gymnosperms, monocotyledons and other angiosperms. Similarly, a combination of *matK* and *ndhJ* generated the highest proportion of variable sites compared with 10 other chloroplast loci[50]. Therefore, Ford *et al.*[50] proposed it as a supplementary locus for barcoding of certain groups of plants. However, the CBOL–Plant Working Group[5] did not select this locus for evaluation, which could be due to the absence of *ndhJ* in several economically important plant groups in conifers[69].

atpF–atpH intergenic sequence

The genes *atpF* and *atpH* encode ATP synthase subunits CFO I and CFO III respectively[71]. Testing of the intergeneric spacer between these two genes as barcode in the flora of the Kruger National Park, South Africa[72], revealed that PCR amplification was easier but alignment of sequences was considerably difficult due to significant length variations, 218–847 bp. Thus, it was found useful only as a supplementary locus in combination with *matK* for barcoding in plants[57,71]. The CBOL–Plant Working Group[5] also observed its high universality but less species discriminatory power compared to that of *rbcL* and *matK* genes and other intergenic loci. Thus, it has only been listed as one of the supplementary loci suitable for combining with the standard two-locus barcode for better resolution on specific projects and taxonomic groups.

trnH–psbA intergenic sequence

This intergenic spacer is one of the most variable genome segments in the chloroplast of angiosperms. It has an average length of approximately 450 bp, but varying from 296 to 1120 bp based on available data[5,14,32,73]. Because of the high species discriminatory power exhibited by this small segment of DNA, Kress *et al.*[4] proposed it along with *nrITS* for DNA barcoding in plants. The *trnH–psbA* locus has been successfully PCR amplified from a wide range of angiosperms and gymnosperms, ferns, mosses, and wild liverworts using the primers given in Table 3 (refs 8, 74). However, problems were observed for obtaining high-quality bidirectional sequences and alignment of sequences in certain taxa, due to the high length variations. Additionally, *trnH–psbA* has an exceedingly short (~300 bp) length in many angiosperms. Thus, it may not have enough sequence variation for species discrimination of those taxa[8]. Furthermore, due to the presence of the *rps19* gene or pseudogene within the *trnH–psbA* region[75], the length of this spacer may go beyond expectation. The length of the *trnH–psbA* locus in some monocot[14] and conifer[32] species may go up to ~1000 bp, which can lead to problems in obtaining bidirectional sequences without using taxon-specific internal sequencing primers. Owing to these problems, Devey *et al.*[76] suggested that developing amplification strategies for the *matK* gene is better than solving the problem of the *trnH–psbA*, especially those caused by mononucleotide repeats. Nonetheless, a majority of the teams proposed the use of *trnH–psbA* as a barcode for plants, mostly in combination with *matK*[14,57,77,78,79] (Table 1). The CBOL–Plant Working Group[5] found the species discriminatory power of *trnH–psbA* to be the highest (69%) among the seven loci tested and thus proposed it as the most preferred supplementary locus. Therefore, *trnH–psbA* can be used in a three-locus barcode system wherever the two-locus barcode system fails to provide adequate resolution[5].

psbK–psbI intergenic sequences

psbK and *psbI* genes encode two low molecular mass polypeptides, K and I respectively, for the photosystem II (ref. 80) and are conserved from algae to land plants[81,82]. The potential of the *psbK–psbI* intergenic region as a barcode for plants was tested in the flora of the Kruger National Park, South Africa[72]. The results revealed its high PCR amplification and sequencing performances (98% of taxa) and ease in the alignment of sequences. Therefore, *psbK–psbI* was proposed in combination with other loci such as *matK*, *trnH–psbA* and *atpF–atpH* for barcoding of plants[72,74]. The CBOL–Plant Working Group[5] also observed that the species discriminatory power of this locus was better than that of *matK* and other loci, except *trnH–psbA*. However, due to the inconsistency in getting bidirectional unambiguous sequences, this has only been considered as a supplementary locus that can be used for better resolution as the need arises[5].

trnL (UAA)–trnF(GAA): genic, intron and intergenic sequences

The *trnL* (UAA)–*trnF*(GAA) locus contains the *trnL* (UAA) gene, its intron and the intergenic region between *trnL* (UAA) and *trnF* (GAA). Taberlet *et al.*[83] employed *trnL* intron for the first time in plant systematic studies. Since then it has been used for molecular phylogenetic studies of several taxa at various taxonomic levels[84–86]. It is widely believed that *trnL* (UAA)–*trnF*(GAA) is not the most variable non-coding region of the chloroplast DNA[73], but it has certain unique advantages. It has a conserved secondary structure with alternation of conserved and variable regions[87]. This facilitates designing of primers that harbour conserved regions and amplify short variable regions between them; thus several universal primers for amplifying the *trnL-F* region in land plants are available (Table 3)[3,83], and depending on the primers, different regions can be amplified. From an investigation

on the suitability of *trnL* (UAA) intron as a barcode in 100 plant species, Taberlet *et al.*[3] concluded that despite the low resolution of the *trnL* intron, it can be used as a barcode for plants, as universal primers are available. Although the CBOL–Plant Working Group has not tested this locus for its potentiality as a barcode, use of *trnL*(UAA) intron as a supplementary locus was advocated for those projects that involved PCR amplification of DNA from highly degraded tissues[5].

Present status of barcoding in plants

The focus of barcoding studies in plants, thus far, was mostly on assessing the relative efficiency of molecular markers that had been used in various phylogenetic studies. From this relative assessment it became clear that none of the DNA segments tested so far had all the qualities essential for a standard barcode for plants. Although some of the loci tested had many promising characters, they had several limitations as well. For instance, *rbcL* and *trnL* (UAA)–*trnF*(GAA) have higher universality, but they lack adequate species discriminatory power. *matK* and *trnH–psbA* have higher species resolution, but problems remain with PCR amplification and sequencing. *rpoC1*, *rpoB*, *atpF–atpH* and *psbK–psbI* have problems on either species discrimination or PCR amplification across all major plant groups. This promotes the proposal of using locus combinations, which can complement each other, for designation as a standard barcode. Considering all the available data on universality, sequence quality retrieved with a single pair of primers, difficulties in sequence alignment, and species discriminatory power along with cost of sequencing and other analysis, the CBOL–Plant Working Group[5] preferred a two-locus barcode combination consisting of *rbcL* and *matK* genes as the standard barcode for land plants. This two-locus combination will act as the universal barcode for land plants. The selection was based on the fact that *rbcL* has long been used in phylogenetic studies, and protocols for high-quality sequence can be retrieved across phylogenetically divergent lineages. *rbcL* also performs well in discrimination tests in combination with other loci. Likewise, the *matK* gene sequence, as stated above, has the highest evolution rate among plastid genes, and thus has high species discriminatory power. Furthermore, recently developed primers[9,50] have improved its PCR amplification and sequencing in a wide range of angiosperms. Thus, the CBOL–Plant Working Group considers that '*rbcL*+*matK* combination represents a pragmatic solution to a complex trade-off between universality, sequence quality, discrimination and cost'[5]. Further, this two-locus barcode offers the opportunity to make use of the high-throughput DNA sequencing facilities to establish a universal framework for DNA barcoding in plants. Using this combination it is possible to achieve approximately 70% species discrimination. This capacity of species level identification may be sufficient for several applications such as investigating plant–animal interactions, identifying protected plant species, and large-scale biodiversity surveys and seedling discrimination in forest regeneration programmes[9,88,89], but higher level of discrimination is essential for absolute identification of species for taxonomic purposes.

In this context, it is also equally important to note some of the major criticisms against the currently proposed barcode system. Spooner[90], after investigating the efficacy of *psbA–trnH*, *matK* and *nrITS* as barcodes on 104 accessions from 63 species of wild potatoes, reported that sequences of *psbA–trnH*, *matK* and *nrITS* failed to provide species-specific markers, especially in the section *Petota*. The plastid genes failed to provide adequate differentiation, whereas the *nrITS* sequences exhibited high intraspecific variations. Similar difficulties were also observed earlier in many genera of the subfamily Magnolioideae, family Magnoliaceae[91], and in another family Lauraceae[92] with *matK* gene for elucidating interspecifc relationships. Two of the strong proponents of employment of DNA barcoding system in plants[14,32], also concluded from detailed studies on three divergent plant groups, viz. angiosperms, gymnosperms and liverworts, that a combination of plastid genomic loci such as *rbcL* + *rpoC1* + *matK* + *trnH–psbA* is useful as a barcoding system only for identification of broad groups of species. Thus, for projects like identification or circumscription of species which require high resolution, the presently proposed two-locus standard barcode is not sufficient[32]. Furthermore, it is also a known fact that extensive intraspecific and intrapopulation variations were reported from many angiosperms, especially in sympatric zones undergoing hybridization and backcrossing[93,94]. Based on these results, it was argued that variations in one or two plastid gene sequences are not adequate to highlight species boundaries[92]. Likewise, the use of a universal barcode system for species determination in all the major lineages is further impeded by complicating biological problems such as allotetraploidy, aneuploidy, apomixes, introgression, lineage sorting, convergency and rapid morphological evolution. Therefore, different DNA regions are required to identify species in different groups.

In the meantime some researchers, without waiting for a perfect barcoding system to emerge, have already started the barcoding of plants, and the results emerging from some of these studies are promising. For instance, using *trnH–psbA* as barcodes, floating pennywort (*Hydrocotyle ranunculoides* L.f.) was distinguished from its most closely related congeners[79]. In another study using *matK* and *trnH–psbA* as DNA barcodes, Raghupathy *et al.*[95] discriminated a new cryptic species of grass *Tripogon cope*, as deciphered by the hill tribes, from its close relatives in the Western Ghats and part of the Nilgiri Biosphere Reserve in India. Further, using *rbcL*, *matK* and

trnH–psbA as barcodes, a new genus *Vachellia* Wight & Arn. was discriminated from its closest relative *Acacia* Mill.[96]. In this study, the DNA barcodes not only discriminated the sister species within either genus, but also displayed biogeographical patterns among populations from India, Africa and Australia. Subsequent morphometric analysis confirmed the cryptic nature of these sister species and the limitations of the existing classification based on 'phenetic' data. Results of some of these studies demonstrated that the DNA barcoding system has the potential to resolve some of the taxonomic problems which cannot be resolved by morphology-based taxonomy alone.

Availability of bioinformatics resources

Parallel to the efforts in finding a standard barcoding system for all organisms in this planet, efforts have also been made to develop adequate bioinformatics resources to support the barcoding of life. The Barcode of Life Data System (BOLD; http://www.barcodinglife.org/views/login.php) was the result of such efforts made by CBOL to facilitate easy deposition and retrieval of data on barcodes. BOLD provides an integrated bioinformatics platform for all phases of the analytical pathway from specimen collection to tightly validated barcode library. BOLD aligns multiple sequences using hidden Markov models[97]. It generates varied distance matrix to construct a neighbour–joining tree labelling the terminal branches with taxonomic information, locality data and/or sequence length. Provisions are also provided for converting the results of analyses into PDF format for transmission[4]. Unknown specimens in the samples can also be identified using BOLD. At present, the query sequence must be at least 300 bp in length. The query sequence is aligned quickly to the global alignment through the hidden Markov model followed by a linear search of the reference library. In this way it tries to identify the possible species. If the species-level search fails, it tries to search possible genus or higher levels. A copy of the data submitted to BOLD will also be deposited at the National Centre for Biotechnology Information (NCBI), USA and its sister genomic repositories. However, it should be kept in mind that to submit a sequence as 'barcode' to a sequence region in GenBank, it must be derived from a gene or genome region that is accepted by CBOL[14]. The sequence should also meet certain quality standards and must be derived from a specimen whose taxonomic assignment can be reviewed, ordinarily through linkage to a specimen that is held in a major collection. NCBI has also provided a web-based barcode submission tool (BarSTool) for submitting sequences of barcodes, but currently it accepts sequences of CO1 only and other sequences should be submitted through Bankit or Sequin.

Conclusion and prospects

The efforts on DNA barcoding in plants have so far mostly focused on assessing the relative discriminatory power among selected loci, rather than estimating absolute discriminatory power of the loci. Based on the available information, the CBOL–Plant Working Group[5] proposed a two-locus (*rbcL* + *matK)* standard barcode for all land plants. Although empirical data have shown the inadequacies of this as a universal barcode for all land plants, the proposal of such a barcode for all plants is an important step towards establishing a centralized plant barcode database, just like the DNA database in GenBank, for its effective and easy use in taxonomy, biodiversity assessment and conservation, prevention of illegal plant exports and imports, identification of endemic and endangered plant species and other activities where identification of plant species is essential. The projected species discriminatory power (ca. 70%) of the proposed barcode is not sufficient for species identification as the gold standard of any taxonomic system is its ability to deliver accurate species identification. Thus, the currently proposed barcode system needs further refinement. The CBOL–Plant Working Group[5] is also aware of these limitations. Thus, it has proposed a supplementary list of loci that can be used wherever needed. It is envisaged that the rapid advances in genomics technologies will enable routine use of multiple unlinked nuclear loci in a barcoding context (Hollingsworth, pers. commun.). Nowadays, a number of EST and UniGene sequences of several plant species are being generated and available in public database. These gene sequences can also be used for identifying potential sequences for barcoding the plant species. Until the more useful barcodes are indentified, the presently proposed DNA barcodes (*rbcL* + *matK)* can be used to initiate barcoding of all land plants. For those plant groups which have limited variation in chloroplast genome, the rapidly evolving *nrITS* can be used as a supplementary barcode if direct sequencing after PCR amplification of this locus is successful[5]. Thus, by recommending a two-locus barcode, the CBOL has initiated a new era in plant taxonomy.

1. Hebert, P. D. N., Cywinska, A., Ball, S. L. and deWaard, J. R., Biological identifications through DNA barcodes. *Proc. R. Soc. London, Ser. B*, 2003, **270**, 313–321.
2. Hebert, P. D. N. and Gregory, T. R., The promise of DNA barcoding for taxonomy. *Syst. Biol.*, 2005, **54**, 852–859.
3. Taberlet, P. *et al.*, Power and limitations of the chloroplast *trn*L (UAA) intron for plant DNA barcoding. *Nucleic Acids Res.*, 2007, **35**, e14.
4. Kress, J. W., Wurdack, K. J., Zimmer, E. A., Weigt, L. A. and Janzen, D. H., Use of DNA barcodes to identify flowering plants. *Proc. Natl. Acad. Sci. USA*, 2005, **102**, 8369–8374.
5. CBOL–Plant Working Group, A DNA barcode for land plants. *Proc. Natl. Acad. Sci. USA*, 2009, **106**, 12794–12797.
6. Ratnasingham, S. and Hebert, P. D. N., BOLD: The Barcode of Life Data System (www.barcodinglife.org). *Mol. Ecol. Notes*, 2007, **7**, 355–364.

7. Blaxter, M., Counting angels with DNA. *Nature*, 2003, **42**, 122–124.
8. Schindel, D. E. and Miller, S. E., DNA barcoding a useful tool for taxonomists. *Nature*, 2005, **435**, 177.
9. Lahaye, R. *et al.*, DNA barcoding the floras of biodiversity hotspots. *Proc. Natl. Acad. Sci. USA*, 2008, **105**, 2923–2928.
10. Ebach, M. C. and Holrege, C., DNA barcoding is no substitute for taxonomy. *Nature*, 2005, **434**, 697.
11. Gregory, T. R., DNA barcoding does not compete with taxonomy. *Nature*, 2005, **434**, 1067.
12. Pennisi, E., Taxonomy. Wanted: A barcode for plants. *Science*, 2007, **318**,190–191.
13. Wolfe, K. H., Li, W. H. and Sharp, P. M., Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast and nuclear DNAs. *Proc. Natl. Acad. Sci. USA*, 1987, **84**, 9054–9058.
14. Chase, M. W. *et al.*, A proposal for a standardized protocol to barcode all land plants. *Taxon*, 2007, **56**, 295–299.
15. Sass, C., Little, D. P., Stevenson, D. W. and Specht, C. D., DNA barcoding in the Cucadales: testing the potential of proposed barcoding markers for species indentification of cycades. *PLoS ONE*, 2007, **2**, e1154.
16. Appels., R. and Honeycutt, R. L., rDNA: evolution over a billion years. In *DNA Systematics Vol. II. Plants* (ed. Dutta S. K.), CRC Press, Boca Raton, FL, 1986, pp. 81–135.
17. Hemleben, V., Ganal, M., Gerstner, J., Schiebel, K. and Torres, R. A., Organization and length heterogeneity of plant ribosomal RNA genes. In *Architecture of Eukaryotic Genes* (ed. Kahl, G.), VCH, Weinhein, Germany, 1988, pp. 371–383.
18. Hershkovitz, M. A. and Zimmer, E. A., Conservation pattern in angiosperm rDNA–ITS2 sequences. *Nucleic Acids Res.*, 1996, **24**, 2857–2867.
19. Brown, D., Wensink, P. C. and Jordan, E., Comparison of the ribosomal DNA's of *Xenopus laecis* and *Xenopus muller*: the evolution of tandem genes. *J. Mol. Biol.*, 1972, **63**, 57–73.
20. Hood, L., Campbell, J. H. and Elgin, S. C. R., The organization, expression and evolution of antibody genes and other multigene families. *Annu. Rev. Genet.*, 1975, **9**, 305–353.
21. Zimmer, E. A., Martin, S. L., Beverley, S. M., Kan, Y. W. and Wilson, A. C., Rapid duplication and loss of genes coding for the alpha chains of hemoglobin. *Proc. Natl. Acad. Sci. USA*, 1980, **77**, 2158–2162.
22. Baldwin, B. G., Sanderson, M. J., Porter, J. M., Wojciechowski, M. F., Campbell, C. S. and Donoghue, M. J., The ITS regions of nuclear ribosomal DNA: a valuable source of evidence on angiosperm phylogeny. *Ann. Mo. Bot. Gard.*, 1995, **82**, 247–277.
23. Alvarez, I. and Wendel, J. F., Ribosomal ITS sequences and plant phylogenetic inference. *Mol. Phylogenet. Evol.*, 2003, **29**, 417–434.
24. Rogers, S. O. and Bendich, A. J., Ribosomal RNA genes in plants: variability in copy number and in the intergenic spacer. *Plant Mol. Biol.*, 1987, **6**, 339–345.
25. Bailey, C. D., Carr, T. G., Harris, S. A. and Hughes, C. E., Characterization of angiosperm nrDNA polymorphism, paralogy, and pseudogenes. *Mol. Phylogenet. Evol.*, 2003, **29**, 435–455.
26. Buckler, E. S., Ipplito, S. A. and Holtsford, T. P., The evolution of ribosomal DNA: divergent paralogues and phylogenetic implications. *Genetics*, 1997, **145**, 111–125.
27. Campbell, C. S., Wojciechowski, M. F., Baldwin, B. G., Alice, L. A. and Donoghue, M. J., Persistent nuclear ribosomal DNA sequence polymorphism in the *Amelanchier* agamic complex (Rosaceae). *Mol. Biol. Evol.*, 1997, **14**, 81–90.
28. Feliner, G. N. and Rosselló, J. A., Better the devil you know? Guidelines for insightful utilization of *nrDNA* ITS in species-level evolutionary studies in plants. *Mol. Phylogenet. Evol.*, 2007, **44**, 911–919.
29. Vijayan, K., Zhang, W. J. and Tsou, C. H., Molecular taxonomy of *Camellia* (Theaceae) as inferred from *nrITS* sequences. *Am. J. Bot.*, 2009, **96**, 1348–1360.
30. Muir, G., Fleming, C. C. and Schlötterer, C., Three divergent rDNA clusters predate the species divergence in *Quercus petraea* (Matt.) Liebl. and *Quercus robur* L. *Mol. Biol. Evol.*, 2001, **18**, 112–119.
31. Wendel, J. F., Schnabel, A. and Seelanan, T., Bidirectional interlocus concerted evolution following allopolyploid speciation in cotton (*Gossypium*). *Proc. Natl. Acad. Sci. USA*, 1995, **92**, 280–284.
32. Hollingsworth, M. L. *et al.*, Selecting barcoding loci for plants: evaluation of seven candidate loci with species-level sampling in three divergent groups of land plants. *Mol. Ecol. Res.*, 2009, **9**, 439–457.
33. Okuyama, Y. and Kato, M., Unveiling cryptic species diversity of flowering plants: successful biological species identification of Asian Mitella using nuclear ribosomal DNA sequences. *BMC Evol. Biol.*, 2009, **9**, 105.
34. Chen, S. *et al.*, Validation of the ITS2 region as a model DNA barcode for identifying medicinal plants species. *PLoS ONE*, 2010, **5**, e8613.
35. Ledford, H., Botanical identities: DNA barcoding for plants comes a step closer. *Nature*, 2008, **451**, 616.
36. Shinozaki, K. *et al.*, The complete nucleotide sequence of the tobacco chloroplast genome: its gene organization and expression. *EMBO J.*, 1986, **5**, 2043–2049.
37. Zurawski, G., Perrot, B., Bottomley, W. and Whitefield, P. R., The structure of the gene for the large subunit of ribulose-1,5-bisphosphate carboxylase from spinach chloroplast DNA. *Nucleic Acids Res.*, 1981, **9**, 3251–3270.
38. Newmaster, S. G., Fazekas, A. J. and Ragupathy, S., DNA barcoding in land plants: evaluation of *rbcL* in a multigene tiered approach. *Can. J. Bot.*, 2006, **84**, 335–341.
39. Albert, V. A., Backlund, A., Bremer, K., Chase, M. W., Manhart, J. R., Mishler, B. D. and Nixon, K. C., Functional constraints and rbcL evidence for land plant phylogeny. *Ann. Mo. Bot. Gard.*, 1994, **81**, 534–567.
40. Soltis, D. E. and Soltis, P. S., Choosing an approach and an appropriate gene for phylogenetic analysis. In *Molecular Systematics of Plants II: DNA Sequencing* (eds Soltis, D. E., Soltis, P. S. and Doyle, J. J.), Kluwer, Dordrecht, 1998, pp. 21–24.
41. Kress, W. J. and Erickson, D. L., A two-locus global DNA barcode for land plants: The coding *rbcL* gene complements the non-coding *trnH–psbA* spacer region. *PLoS ONE*, 2007, **2**, e508.
42. Fay, M. F., Swensen, S. M. and Chase, M. W., Taxonomic affinities of *Medusagyne oppositifolia* (Medusagynaceae). *Kew Bull.*, 1997, **52**, 111–120.
43. Wolfe, K. H., Protein-coding genes in chloroplast DNA: compilation of nucleotide sequences, data base entries and rates of molecular evolution. In *Cell Culture and Somatic Cell Genetics of Plants* (ed. Vasil, K.), Academic Press, San Diego, 1991, vol 7BI, pp. 467–482.
44. Neuhaus, H. and Link, G., The chloroplast tRNA(UUU) gene from mustard (*Sinapsis alba*) contains a class II intron potentially coding for a maturase-related polypeptide. *Curr. Genet.*, 1987, **11**, 251–257.
45. Hilu, K. W. and Liang, H., The *matK* gene: sequence variation and application in plant systematics. *Am. J. Bot.*, 1997, **84**, 830–839.
46. Kelchner, S. A., The evolution of noncoding chloroplast DNA and its application in plant systematics. *Ann. Mo. Bot. Gard.*, 2000, **87**, 482–498.
47. Cuénoud, P., Savolainen, V., Chatrou, L. W., Powell, M., Grayer, R. J. and Chase, M. W., Molecular phylogenetics of caryophyllales based on nuclear 18S rDNA and plastid *rbcL, atpB,* and *matK* DNA sequences. *Am. J. Bot.*, 2002, **89**, 132–144.
48. Schmitz-Linneweber, C., Maier, R. M., Alcaraz J. P., Cottet, A., Herrmann, R. G. and Mache, R., The plastid chromosome of spinach (*Spinacia oleracea*): complete nucleotide sequence and gene organization. *Plant Mol. Biol.*, 2001, **45**, 307–315.

49. Johnson, L. S. and Soltis, D. E., Phylogenetic inference in Saxifragaceae *sensu strict* and *Gilia* (Polemoniaceae) using *matK* sequences. *Ann. Mo. Bot. Gard.*, 1995, **82**, 149–175.
50. Ford, C. S. *et al.*, Selection of candidate coding DNA barcoding regions for use on land plants. *Bot. Linn. Soc.*, 2009, **159**, 1–11.
51. Starr, J. R., Naczi, R. F. C. and Chouinard, B. N., Plant DNA barcodes and species resolution in sedges (*Carex*, Cyperaceae). *Mol. Ecol. Resour.*, 2009, **9**(Suppl. 1), 151–163.
52. Serino, G. and Maliga, P., RNA polymerase subunits encoded by the plastid *rpo* genes are not shared with the nucleus-encoded plastid enzyme. *Plant Physiol.*, 1998, **117**, 1165–1170.
53. Guisinger, M. M., Kuehl, J. V., Boore, J. L. and Jansen, R. K., Genome-wide analyses of Geraniaceae plastid DNA reveal unprecedented patterns of increased nucleotide substitutions. *Proc. Natl. Acad. Sci. USA*, 2008, **105**, 18424–18429.
54. Tsumura, Y., Kawahara, T., Wickneswari, R. and Yoshimura, K., Molecular phylogeny of Dipterocarpaceae in Southeast Asia using RFLP of PCR-amplified chloroplast genes. *Theor. Appl. Genet.*, 1996, **93**, 22–29.
55. Adekambi, T., Drancourt, M. and Raoult, D., The *rpoB* gene as a tool for clinical microbiologists. *Trends Microbiol.*, 2008, **17**, 37–46.
56. Logacheva, M. D., Penin, A. A., Samigullin, T. H., Vallejo-Roman, C. M. and Antonov, A. S., Phylogeny of flowering plants by the chloroplast genome sequences: in search of a 'Lucky Gene'. *Biochemistry*, 2007, **72**, 1324–1329.
57. Fazekas, A. J. *et al.*, Multiple multilocus DNA barcodes from the plastid genome discriminate plant species equally well. *PLoS ONE*. 2008, **3**, e2802.
58. Liu, Y., Yan, H. F., Cao, T. and Ge, X. J., Evaluation of 10 plant barcodes in bryophyte (mosses). *J. Syst. Evol.*, 2010, **48**, 36–46.
59. Hajdukiewicz, P. T. J., Allison, L. A. and Maliga, P., The two RNA polymerases encoded by the nuclear and the plastid compartments transcribe distinct groups of genes in tobacco plastids. *EMBO J.*, 1997, **16**, 4041–4048.
60. Yasui, Y. and Ohnishi, O., Interspecific relationships in *Fagopyrum* (Polygonaceae) revealed by the nucleotide sequences of the *rbcL* and *accD* genes and their intergenic region. *Am. J. Bot.*, 1998, **85**, 1134–1142.
61. Devey, D. S., Chase, M. W. and Clarkson, J. J., A stuttering start to plant DNA barcoding: microsatellites present a previously overlooked problem in non-coding plastid regions. *Taxon*, 2009, **58**, 7–15.
62. Wakasugi, T., Tsudzuki, T. and Sugiura, M., The genomic of plant chloroplasts: gene content and alteration of genomic information by RNA editing. *Photosyn. Res.*, 2001, **70**, 107–118.
63. Tsuruya, K., Suzuki, M., Plader, W., Sugita, C. and Sugita, M., Chloroplast transformation reveals that tobacco *ycf5* is involved in photosynthesis. *Acta Physiol. Plant.*, 2006, **28**, 365–371.
64. Ohyama, K. *et al.*, Chloroplast gene organisation deduced from complete sequence analysis of liverwort *Marchantia polymorpha* chloroplast DNA. *Nature*, 1986, **322**, 572–574.
65. Burrows, P. A., Sazanov, L. A., Svab, Z., Maliga, P. and Nixon, P. J., Identification of a functional respiratory complex in chloroplasts through analysis of tobacco mutants containing disrupted plastid *ndh* genes. *EMBO J.*, 1998, **17**, 868–876.
66. Nakazono, M. and Hirai, A., Identification of the entire set of transferred chloroplast DNA sequences in the mitochondrial genome of rice. *Mol. Gen. Genet.*, 1993, **236**, 341–346.
67. Yamagishi, H., Terachi, T., Ozaki, A. and Ishibashi, A., Inter- and intraspecific sequence variations of the chloroplast genome in wild and cultivated *Raphanus*. *Plant Breed.*, 2009, **128**, 172–177.
68. Wakasugi, T., Tsudzuki, J., Ito, S., Nakashima, K., Sudzuki, T. and Sugiura, M., Loss of all *ndh* genes as determined by sequencing the entire chloroplast genome of the black pine *Pinus thunbergii*. *Proc. Natl. Acad. Sci. USA*, 1994, **91**, 9794–9798.
69. Haberhausen, G. and Zetsche, K., Functional loss of all *ndh* genes in an otherwise relatively unaltered plastid genome of the holoparasitic flowering plant *Cuscuta reflexa*. *Plant Mol. Biol.*, 1994, **24**, 217–222.
70. Braukmann, T. W. A., Kuzmina, M. and Stefanović, S., Loss of all plastid *ndh* genes in Gnetales and conifers: Extent and evolutionary significance for the seed plant phylogeny. *Curr. Genet.*, 2009, **55**, 323–337.
71. Drager, R. G. and Hallick, R. B., A novel *Euglena gracilis* chloroplast operon encoding 4 ATP synthase subunits and 2 ribosomal-proteins contains 17 Introns. *Curr. Genet.*, 1993, **23**, 271–280.
72. Lahaye, R., Savolainen, V., Duthoit, S., Maurin, O. and van der Bank, M., A test of *psbK–psbI* and *atpF–atpH* as potential plant DNA barcodes using the flora of the Kruger National Park as a model system (South Africa). *Nature Proc.*, 2008; hdl:10101/npre.1896.1
73. Shaw, J. *et al.*, The tortoise and the hare II: relative utility of 21 noncoding chloroplast DNA sequences for phylogenetic analysis. *Am. J. Bot.*, 2005, **92**, 142–166.
74. Lee, H. L., Yi, D. K., Kim, J. S. and Kim, K. J., Development of plant DNA barcoding markers from the variable noncoding regions of chloroplast genome. In Abstract presented at the Second International Barcode of Life Conference, Academia Sinica, Taipei, Taiwan, September 2007, pp. 18–20; http://www.bolinfonet.org/conferences/assets/files/confrence_abstract_book.pdf
75. Chang, C. C. *et al.*, The chloroplast genome of *Phalaenopsis aphrodite* (Orchidaceae): comparative analysis of evolutionary rate with that of grasses and its phylogenetic implications. *Mol. Biol. Evol.*, 2006, **23**, 279–291.
76. Devey, D. S., Chase, M. W. and Clarkson, J. J., A stuttering start to plant DNA barcoding: Microsatellites present a previously overlooked problem in non-coding plastid regions. *Taxon*, 2009, **58**, 7–15.
77. Newmaster, S. G., Fazekas, A. J., Steeves, R. A. D. and Janovec, J., Testing candidate plant barcode regions in the Myristicaceae. *Mol. Ecol. Notes*, 2008, **8**, 480–490.
78. Erickson, D. L., Spouge, J., Resch, A., Weigt, L. A. and Kress, J. W., DNA barcoding in land plants: Developing standards to quantify and maximize success. *Taxon*, 2008, **57**, 1304–1316.
79. van de Wiel, C. C. M., van der Schoot, J., van Valkenburg, J. L., Duistermaat, C. H. and Smulders, M. J. M., DNA barcoding discriminates the noxious invasive plant species, floating pennywort (*Hydrocotyle ranunculoides* L.f.), from non-invasive relatives. *Mol. Ecol. Resour.*, 2009, **9**, 1086–1091.
80. Meng, B. Y., Wakasugi, T. and Sugiura, M., Two promoters within the *psbK–psbI–trnG* gene-cluster in tobacco chloroplast DNA. *Curr. Genet.*, 1991, **20**, 259–264.
81. Knauf, U. and Hachtel, W., The genes encoding subunits of ATP synthase are conserved in the reduced plastid genome of the heterotrophic alga *Prototheca wickerhamii*. *Mol. Genet. Genome*, 2002, **267**, 492–497.
82. McNeal, J. R., Kuehl, J. V., Boore, J. L. and de Pamphilis, C. W., Complete plastid genome sequences suggest strong selection for retention of photosynthetic genes in the parasitic plant genus *Cuscuta*. *BMC Plant Biol.*, 2007, **7**.
83. Taberlet, P., Gielly, L., Pautou, G. and Bouvet, J., Universal primers for amplification of three non-coding regions of the chloroplast DNA. *Plant Mol. Biol.*, 1991, **17**, 1105–1109.
84. Besendahl, A., Qiu, Y.-L., Lee, J., Palmer, J. D. and Bhattacharya, D., The cyanobacterial origin and vertical transmission of the plastid tRNALeu group-I-intron. *Curr. Genet.*, 2000, **37**, 12–23.
85. Simon, D., Fewer, D., Friedl, T. and Bhattacharya, D., Phylogeny and self-splicing ability of the plastid tRNA-Leu group I intron. *J. Mol. Evol.*, 2003, **57**, 710–720.
86. Quandt, D. and Stech, M., Molecular systematics of bryophytes in context of land plant phylogeny. In *Plant Genome Phanerogams* (eds Sharma, A. K. and Sharma, A.), Oxford and IBH Publishing, New Delhi, 2003, vol. 1, pp. 267–295.

87. Quandt, D., Müller, K., Stech, M., Frahm, J. P., Frey, W., Hilu, K. W. and Borsch, T., Molecular evolution of the chloroplast *trnL–F* region in land plants. *Monogr. Syst. Bot. Mo. Bot. Gard.*, 2004, **98**, 13–37.
88. Jurado-Rivera, J. A., Vogler, A. P., Reid, C. A. M., Petitpierre, E. and Gomez-Zurita, J., DNA barcoding insect–host plant association. *Proc. R. Soc. Biol. Sci. Ser. B*, 2009, **276**, 639–648.
89. Little, D. P. and Stevenson, D. W., A comparison of algorithms for the identification of specimens using DNA barcodes: Examples from gymnosperms. *Cladistics*, 2007, **23**, 1–21.
90. Spooner, D. M., DNA barcoding will frequently fail in complicated groups: an example in wild potatoes. *Am. J. Bot.*, 2009, **96**, 1177–1189.
91. Shi, S., Jin, H., Zhong, Y., He, X., Huang, Y., Tan, F. and Boufford, D. E., Phylogenetic relationships of the Magnoliaceae inferred from cpDNA *matK* sequences. *Theor. Appl. Genet.*, 2000, **101**, 925–930.
92. Rohwer, J. G., Towards a phylogenetic classification of the Lauraceae: Evidence from *matK* sequences. *Syst. Bot.*, 2000, **25**, 60–71.
93. Soltis, D. E., Soltis, P. S. and Milligan, B. G., Intraspecific chloroplast DNA variation: systematic and phylogenetic implications. In *Molecular Systematics of Plants* (eds Soltis, P. S., Soltis, D. E. and Doyle, J. J.), Chapman and Hall, New York, USA, 1992, pp. 117–150.
94. Okuyama, Y. *et al.*, Nonuniform concerted evolution and chloroplast capture: Heterogeneity of observed introgression patterns in three molecular data partition phylogenies of Asian *Mitella* (Saxifragaceae). *Mol. Biol. Evol.*, 2005, **22**, 285–296.
95. Raghupathy, S., Newmaster, S. G., Murugesan, M. and Balasubramaniam, V., DNA barcoding discriminates a new cryptic grass species revealed in an ethnobotany study by the hill tribes of the Western Ghats in southern India. *Mol. Ecol. Resources*, 2009, **9** (*Suppl. 1*), pp. 164–171.
96. Newmaster, S. G. and Raghupathy, S., Testing plant barcoding in a sister species complex of pantropical *Acacia* (Mimosoideae, Fabaceae). *Mol. Ecol. Res.*, 2009, **9** (Suppl. 1), 172–180.
97. Durbin, R., Eddy, S., Krogh, A. and Mitchison, G., *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*, Cambridge University Press, Cambridge, UK, 1998.
98. White, T. J., Bruns, T., Lee, S. and Taylor, J. W., Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics. In *PCR Protocols: A Guide to Methods and Applications* (eds Innis, M. A. *et al.*), Academic Press, New York, 1990, pp. 315–322.
99. Urbatsch, L. E., Baldwin, B. G. and Donoghue, M. J., Phylogeny of the coneflowers and relatives (Heliantheae: Asteraceae) based on nuclear rDNA internal transcribed spacer (ITS) sequences and chloroplast DNA restriction site data. *Syst. Bot.*, 2000, **25**, 539–565.
100. Gardes, M. and Bruns, T. D., ITS primers with enhanced specificity for basidiomycetes – application to the identification of mycorrhizae and rusts. *Mol. Ecol.*, 1993, **2**, 113–118.
101. Wang, X., Tsumura, Y., Yoshimaru, H., Nagasaka, K. and Szmidt, A. E., Phylogenetic relationships of Eurasian pines (*Pinus*, Pinaceae) based on chloroplast *rbcL*, *matK*, *rpl20–rps18* spacer, and *trnV* intron sequences. *Am. J. Bot.*, 1999, **86**, 1742–1753.
102. Tate, J. A. and Simpson, B. B., Paraphyly of *Tarasa* (Malvaceae) and diverse origins of the polyploid species. *Syst. Bot.*, 2003, **28**, 723–737.
103. Sang, T., Crawford, D. J. and Stuessy, T. F., Chloroplast DNA phylogeny, reticulate evolution, and biogeography of *Paeonia* (Paeoniaceae). *Am. J. Bot.*, 1997, **84**, 1120–1136.
104. Hamilton, M. B., Four primer pairs for the amplification of chloroplast intergenic regions with intraspecific variation. *Mol. Ecol.*, 1999, **8**, 513–525.

ACKNOWLEDGEMENT. K.V. received financial support from the Institute of Plant and Microbial Biology, Academia Sinica, Taipei, Taiwan, ROC.

Received 20 November 2009; revised accepted 12 October 2010