Plant DNA Barcoding and Phylogenetics, Pages 1-14
M. Ajmal Ali, Gyulai Gábor and Fahad Al-Hemaid (Eds.)

1 Introduction to Plant DNA Barcoding

M.A. Ali, A.K. Pandey, G. Gyulai, S.H. Park, C. Lee, S.Y. Kim and J. Lee

Introduction

Till date taxonomists have described approximately 1.7 million species, but this figure might be a gross under-estimate of the true biological diversity on Earth (Blaxter 2003; Wilson, 2003). Although taxonomists can identify most organisms with which they are familiar, an ever-growing community requires taxonomic information for a broad range of taxa. DNA barcoding is a novel system designed to provide rapid, accurate, and automatable species identifications by using short, standardized gene regions as internal species tags (Hebert et al., 2005). The genomes of living organisms are analogous to bar-codes. The use of short DNA sequences for biological identifications was first proposed by Paul Herbert and colleagues in 2003. The role of barcodes is to provide a tool to assign unidentified specimens to already characterized species (Hebert et al. 2003). Building upon the idea of the 'universal product code', known as 'barcodes', a few DNA nucleotides (e.g. the sequences of a short DNA fragment) may provide an immediate diagnosis for species. As with commercial barcodes, the use of these 'species barcodes' first require the assembly of a comprehensive library that links barcodes and organisms. DNA barcodes consist of a short sequence of DNA between 400 and 800 base pairs long that can be easily extracted and characterized for all species on this planet. These

genetic barcodes will be accessed through a digital library and used to identify unknown plants in the field or garden.

DNA barcoding follows the same principle as does the basic taxonomic practice of associating a name with a specific reference collection in conjunction with a functional understanding of species concepts (i.e., interpreting discontinuities in interspecific variation). In DNA barcoding the complete data can be obtained from single specimens irrespective of sexual morph or life stage. Morphologically indistinguishable taxa can be diagnosed without the need for live material, particular morphs or population measures. Barcode sequences can be generated from type specimens (holotype, paratype or neotype). A specimen barcode can be compared with sequences derived from other molecular taxonomy initiatives. If a close match is found to a named taxon, recourse can be made to traditional monographs and keys to understand the biological properties of the identified MOTU (molecular operational taxonomic units) and their close relatives (Floyd et al. 2002). Molecular phylogenetic analyses can be used to generate testable hypotheses of MOTU interrelatedness. The core idea of DNA barcoding is based on the fact that short pieces of DNA can be found that vary only to a very minor degree within species, such that this variation is much less than between species.

Whether or not actual species can be identified with DNA, the number of distinct DNA sequences in environmental sampling and reconstruction of phylogenetic trees to place these sequences into an evolutionary context have been used in several inventories of cryptic biodiversity (e.g. soil bacteria or marine/freshwater micro-organisms). Initially referred to as DNA typing or profiling, the DNA barcoding initiative has taken this step forward, and several taxa have now been surveyed in their natural habitats using this technique. Such an approach has been particularly useful for marine organisms (Shander and Willassen, 2005), including fishes (Mason, 2003; Ward et al., 2005), soil meiofauna (Blaxter et al., 2004), freshwater meiobenthos (Markmann and Tautz, 2005) and even extinct birds (Lambert et al., 2005). In the rainforests, rapid DNA-based entomological inventories have been performed so efficiently (Monaghan et al., 2005; Smith et al., 2005) that tropical ecologists have been among the most active advocates of DNA barcoding (Janzen, 2004).

Plant DNA barcoding markers

The use of DNA sequences to identify organisms has been proposed as a more efficient approach than traditional taxonomic practices (Blaxter et al, 2004; Tautz et al., 2003). **T**he identification of animal biological diversity by using molecular markers has recently been proposed and

demonstrated on a large scale through the use of a short DNA sequence the mitochondrial cytochrome oxidase subunit 1 (cox1, usually referred to as COI in barcoding studies), was proposed to be a good candidate for barcoding animal species (Hebert et al., 2003). The availability of broad-range primers for amplification of mitochondrial COI from diverse invertebrate phyla establishes this gene as a particularly promising target for species identification in animals (Folmer et al., 1994). Plants have relatively little sequence variation in their mitochondrial DNA, perhaps because of hybridization and introgression. A chloroplast gene such as matK (maturase K) or a nuclear gene such as ITS (internal transcribed spacer) may be an effective target for barcoding in plants (Kress et. al., 2005). Kress et al. (2005) have demonstrated the effectiveness of "DNA barcoding" in angiosperms using nrDNA and non-coding cpDNA sequences.

In flowering plants another approach has been put forward. On one hand several plastid loci do discriminate between species, e.g. the trnH-psbA intergenic spacer (Kress et al., 2005) and some more typical phylogenetic markers such as rbcL and trnL-F (Chase et al., 2005), but on the other hand multiple genetic loci might be necessary to account for the common hybridization and polyploidy events in angiosperms. Ribosomal DNA (e.g. ITS in orchids) could be used to complement plastid genes, and shorter low-copy nuclear markers are being discovered that might in the future be used to provide a more sophisticated multiple component barcode for species diagnosis and delimitation (Chase et al., 2005). The sequences used thus for molecular barcoding are the nuclear small subunit ribosomal RNA gene (SSU, also known as 16S in prokaryotes, and 18S in most eukaryotes), the nuclear large-subunit ribosomal RNA gene (LSU, also known as 23S and 28S; the highly variable expansion loops that are flanked by conserved stem sequences are particularly useful), the highly variable internal transcribed spacer section of the ribosomal RNA cistron (ITS, separated by the 5S ribosomal RNA gene into ITS1 and ITS2 regions), the mitochondrial cytochrome *c* oxidase 1 (CO1 or COX1) gene and the chloroplast ribulose bisphosphate carboxylase large subunit (rbcL) gene. Kress et al. (2005) have suggested that the nuclear internal transcribed spacer region and the plastid trnH-psbA intergenic spacer as potentially usable DNA regions for applying barcoding to flowering plants. The internal transcribed spacer is the most commonly sequenced locus used in plant phylogenetic investigations at the species level and shows high levels of interspecific divergence (Pandey and Ali, 2006). The trnH-psbA spacer, although short (≈450-bp), is the most variable plastid region in angiosperms and is easily amplified across a broad range of land plants (Kress et al., 2005).

The primary reason that barcoding has not been applied to plants is that plant mitochondrial genes, because of their low rate of sequence change, are poor candidates for species-level discrimination. The divergence of CO1 coding regions among families of flowering plants has been documented to be only a few base pairs across 1.4 kb of sequence. Furthermore, plants rapidly change their mitochondrial genome structure; thereby precluding the existence of universal intergenic spacers that otherwise would be appropriately variable unique identifiers at the species level. The ITS region has shown broad utility across photosynthetic eukaryotes (with the exception of ferns) and fungi and has been suggested as a possible plant barcode locus. Species-level discrimination and technical ease have been validated in most phylogenetic studies that employ ITS, and a large body of sequence data already exists for this region. An advantage of the ITS region is that it can be amplified in two smaller fragments (ITS1 and ITS2) adjoining the 5.8S locus, which has proven especially useful for degraded samples. The quite conserved 5.8S region in fact contains enough phylogenetic signals for discrimination at the level of orders and phyla, although identification at this taxonomic level is not the concern of barcoding. The 5.8S locus can serve as a critical alignment-free anchor point for search algorithms that make sequence comparisons for both phylogenetic and barcoding purposes. The utility of conserved regions such as 5.8S to generate a pool of nearest neighbors for refined comparisons will be critical for effective database searches, especially when comparing a sequence that has no identical match in a sequence library.

For phylogenetic investigations, the plastid genome has been more readily exploited than the nuclear genome and may offer for plant barcoding what the mitochondrial genome does for animals. It is a uniparentally inherited, nonrecombining, and, in general, structurally stable genome. Universal primers are available for a number of loci and intergenic spacers that are evolving at a variety of rates. The plastid locus most commonly sequenced by plant systematists for phylogenetic purposes is rbcL, followed by the trnL-F intergenic spacer, matK, ndhF, and atpB-rbcL has been suggested as a candidate for plant barcoding, even though it has generally been used to determine evolutionary relationships at the generic level and above. Besides rbcL and atpB, all of the latter plastid loci have been used at the species level with various degrees of success. Most of them (except the trnL-F spacer) require full-length sequences of >1 kb to yield enough sequence length to discriminate species. Most relevant to plant barcoding, no region of the plastid genome has been found to have the high level of variation seen in most animal CO1 barcodes, although a few intergenic spacers have shown more promise than any plastid locus now in general use. Kress et al. (2005) have compared plastid genomes of *Atropa* and *Nicotiana*, and

recorded that nine intergenic spacers trnK-rps16, trnH-psbA, rp136-rps8, atpB-rbcL, ycf6-psbM, trnV-atpE, trnC-ycf6, psbM-trnD, and trnL-F met the barcode criteria. By comparison, ITS had a much higher divergence value (13.6%) than any of the plastid regions, and rbcL was by far the lowest in divergence (0.83%). Although three spacers (atpB-rbcL, ycf6-psbM, and psbM-trnD) were slightly to moderately longer than our 800-bp cutoff.

Besides ITS, those single-copy nuclear genes or their introns that are gaining prominence in species-level molecular systematics studies (e.g., *leafy, waxy, pistillata, and RPB2*), also have been considered. The significantly greater length of *rbcL* (usually 1,428 bp) causes problems because it is necessary to use four primers for double-stranded sequencing of the entire gene. It has been suggested that the trnH-psbA intergenic spacer is the best plastid option for a DNA barcode sequence that has good priming sites, length, and interspecific variation. In their trials across a diverse set of genera in seven plant families, Kress et al. (2005) reported that three plastid regions (trnH-psbA, rp136-rpf8, and trnL-F) ranked highest with respect to amplification success and appropriate sequence length, but trnH-psbA demonstrated nearly three times the percentage sequence divergence of these other two regions. By applying barcode criteria (i.e., length considerations and universality) to the framework of their study, it has been concluded that trnH-psbA has greater potential for species-level discrimination than any other locus (Kress et al., 2005).

Despite this high level of interspecific variation, trnH-psbA has found only limited use in species-level phylogenetic reconstruction because of the short length as well as the difficulty of alignments resulting from a high number of indels (deletions). In contrast with the problems of indels for phylogenetic construction, it is suspected that indels will ultimately enhance the information needed for species identifications, once the appropriate informatics tools for barcoding are developed. Both ITS and trnH-psbA are good starting points for large-scale testing of DNA barcoding across a large sample of angiosperms.

Basic steps in DNA barcoding

DNA barcoding, a new method for the quick identification of any species based on extracting a DNA sequence from a tiny tissue sample of any organism, is now being applied to taxa across the tree of life. As a research tool for taxonomists, DNA barcoding assists in identification by expanding the ability to diagnose species by including all life history stages of an organism. As a biodiversity discovery tool, DNA barcoding helps to flag species that are potentially new to science. As a biological

tool, DNA barcoding is being used to address fundamental ecological and evolutionary questions, such as how species in plant communities are assembled. The process of DNA barcoding entails two basic steps: (1) building the DNA barcode library of known species and (2) matching the barcode sequence of the unknown sample against the barcode library for identification. Although DNA barcoding as a methodology has been in use for less than a decade, it has grown exponentially in terms of the number of sequences generated as barcodes as well as its applications (Kress and Erickson, 2012).

DNA is a relatively stable molecule, and can be isolated from museum collections, including specimens preserved in formalin (Fang et al., 2002). The extraction of DNA from specimens in herbarium collections can easily be made. This success may be due to the specimens having been air-dried and in a good state of preservation as evidenced by the generally green appearance of the leaves selected for extraction. Plant voucher specimens vary in how and when they are dried after being pressed. If specimen-drying facilities are not immediately available, especially in humid tropical climates, botanists often treat pressed specimens with ethanol to temporarily preserve them against fungal attack and degradation. Alcohol has been shown to be detrimental to recovering high-quality DNA, although how it will affect the short sequences needed for barcoding is unknown. It is encouraging that museum specimens of insects dried from ethanol storage readily yield CO1 sequences. A more thorough investigation and optimization of methods to extract high-quality barcode DNA from herbarium collections in a high-throughput format will be critical to efficiently build a sequence-database library for plant DNA barcodes. Positive results have been obtained by using well preserved specimens which indicate that the *a priori* selection of apparently under graded plant samples will be an important determinant of success. Fortunately, herbaria often have more than one specimen per species among which to select for successful DNA barcoding.

Recent advances

Global DNA barcoding efforts have resulted in the formation of the Consortium for the Barcode of Life (CBOL). In January 2013, the Barcode of Life Database (BOLD) contained more than 2.7 million specimen records, with 2 million having barcodes belonging to over 170,000 species (Ratnasingham and Hebert, 2007; BOLD Systems, 2013). Smaller databases, containing sequences of specialized groups, also exist [for example, Fungal Database (Crous et al., 2004), Genome Database for Rosaceae, GDR (Jung et al., 2008)].

The main DNA barcoding bodies and resources are (1) Consortium for the Barcode of Life (CBOL) http://www.barcodeoflife.org established in 2004. CBOL promotes DNA barcoding through over 200 member organizations from 50 countries, operates out of the Smithsonian Institution's National Museum of Natural History in Washington, (2) International Barcode of Life (iBOL) http://www.ibol.org Launched in October 2010, iBOL represents a not-for-profit effort to involve both developing and developed countries in the global barcoding effort, establishing commitments and working groups in 25 countries. The Biodiversity Institute of Ontario is the project's scientific hub and its director, (3) The Barcode of Life Datasystems (BOLD) http://www.boldsystems.org. The Barcode of Life Datasystems is an online workbench for DNA barcoders, combines a barcode repository, analytical tools, interface for submission of sequences to GenBank, a species identification tool and connectivity for external web developers and bioinformaticians. The Consortium for the Barcode of Life (CBOL) Plant Working Group (2009) recommended rbcL + matK as a core two-locus combination. However, as these loci encode conserved functional traits, it is not clear whether they provide sufficiently high species resolution. One of the challenges for plant barcoding is the ability to distinguish closely related or recently evolved species.

The classical way of practice of plant taxonomy for the identification of species lead the discipline many a times to a subject of opinion; the plant DNA barcoding is now transitioning the epitome of species identification (Ali et al., 2014). One of the most important uses of the DNA barcoding is in the medicinal plant authentication. Recently ITS, trnH-psbA, rbcL, matK and trnL–trnF gene sequence have successfully been used for DNA barcoding of several plant species. In addition with the above, Chen et al. (2010) tested the discrimination ability of ITS2 in more than 6600 plant samples belonging to 4800 species from 753 distinct genera and found that the rate of successful identification with the ITS2 was 92.7% at the species level. Yao et al. (2010) also evaluated 50,790 plant and 12,221 animal ITS2 sequences downloaded from GenBank, and propose that the ITS2 locus should be used as a universal DNA barcode for identifying plant species and as a complementary locus for CO1 to identify animal species.

Benefits

Traditionally, taxonomic identification has relied upon morphological characters. In the last two decades, molecular tools based on DNA sequences of short standardized gene fragments, termed DNA barcodes, have been developed for species discrimination. The most

common DNA barcode used in animals is a fragment of the cytochrome c oxidase (COI) mitochondrial gene, while for plants, two chloroplast gene fragments from the RuBisCo large subunit (rbcL) and maturase K (matK) genes are widely used. Information gathered from DNA barcodes can be used beyond taxonomic studies and will have far-reaching implications across many fields of biology, including ecology (rapid biodiversity assessment and food chain analysis), conservation biology (monitoring of protected species), biosecurity (early identification of invasive pest species), medicine (identification of medically important pathogens and their vectors) and pharmacology (identification of active compounds). However, it is important that the limitations of DNA barcoding are understood and techniques continually adapted and improved as this young science matures (Fisˇer and Buzan, 2014)

DNA barcodes are likely to play a major role in the future of taxonomy. The build-up of DNA databases has great potential for the identification and classification of organisms and for supporting ecological and biodiversity research programmes (Tautz et al., 2002). As a uniform, practical method for species identification, it appears to have broad scientific applications. DNA-based species identification offers enormous potential benefits for the biological scientific community, educators, and the interested public. It will help open the treasury of biological knowledge and increase community interest in conservation biology and understanding of evolution. A rapid and accurate method is now being developed for the quick identification of plant species based on extracting DNA from a tiny tissue sample of a leaf, flower, or fruit.

The direct benefits of DNA barcoding is to make the outputs of systematics available to a large number of end-users by providing standardized and high-tech identification tools, e.g. for biomedicine (parasites and vectors), agriculture (pests), environmental assays and customs (trade in endangered species). It will provide a bio-literacy tool for the general public. DNA based species identification will help open the treasury of biological knowledge, which is currently underused partly because taxonomic expertise for species identification is relatively inaccessible. DNA barcoding will also relieve the enormous burden of identifications from taxonomists, so they can focus on more pertinent duties such as delimiting taxa, resolving their relationships and discovering and describing new species. It will also help in pairing up various life stages of the same species (e.g. seedlings, larvae). The most important aspect of DNA barcoding is that it will facilitate basic biodiversity inventories (Savolainen et al., 2005).

DNA barcoding can be likened to aerial photography, in that it provides an efficient method for mapping the extent of species, though in sample space rather than physical space. The “aerial map” of DNA barcodes will help investigators explore the biological world and make

full use of the enormous knowledge that has been built on 250 years of classical taxonomy. As sequencing costs decrease, DNA-based species identification will become available to an increasingly wide community. When costs are low enough, science teachers and backyard naturalists will be able to use DNA barcoding for in depth examination of local ecosystems.

Limitations

DNA-based species identification depends on distinguishing intraspecific from interspecific genetic variation. The ranges of these types of variation are unknown and may differ between groups. It may be difficult to resolve recently diverged species or new species that have arisen through hybridization. There is no universal DNA barcode gene, no single gene that is conserved in all domains of life and exhibits enough sequence divergence for species discrimination. The validity of DNA barcoding therefore depends on establishing reference sequences from taxonomically confirmed specimens. This is likely to be a complex process that will involve cooperation among a diverse group of scientists and institutions.

Sequencing is essentially equally easy for all DNA fragments barring extreme base composition biases, polynucleotide runs and stable secondary structures. However, the ITS region often varies by insertions or deletions within an individual, making sequencing very difficult as two independent sequence types are being analysed simultaneously (Elbadri et al. 2002). ITS sequences are also difficult to align as they tend to evolve by insertion and deletion rather than substitution making the secondary steps of phylogenetic reconstruction problematic. SSU, LSU, COX1 and rbcL are each relatively simple to align and analyse, though exceptions do occur. It may be suggested that any barcoding system should aim to acquire data for at least a nuclear and an organellar gene from single specimens. For specimen-independent, 'environmental DNA' based surveys, any target may do, but the universality of SSU and LSU primer sets recommends them. The most common criticism of 18S rDNA, as a source of phylogenetic information, has been that it is not sufficiently variable for phylogenetic reconstruction within the angiosperms and that it is highly prone to insertion and deletion, making sequence alignment difficult. 18S rDNA provides a sufficient number of characters for broad scale phylogenetic reconstruction of the angiosperms.

Where species are simply unknown or no attempts have been made to delimit them, the barcode approach as originally intended would be limited in its applicability. However, it is a widely accepted fact that

species, however defined, are variable for most DNA markers including the widely used cox1 gene. Hence, the analogy to commercial barcodes presumes that the variation within these species is smaller than between them.

Barcoding has created some controversy in the taxonomy community (Mallet and Willmott, 2003; Lipscomb et al., 2003; Seberg et al., 2003; DeSalle et al., 2005; Lee, 2004; Ebach and Holdrege, 2005; Will et al, 2005; Gregory, 2005). Traditional taxonomists use multiple morphological traits to delineate species. Today, such traits are increasingly being supplemented with DNA-based information. In contrast, the DNA barcoding identification system is based on what is in essence a single complex character (a portion of one gene, comprising ~650 bp from the first half of the mitochondrial cytochrome c oxidase subunit I gene sometimes called COXI or COI), and barcoding results are therefore seen as being unreliable and prone to errors in identification (Dasmahapatra and Mallet, 2006). Although the mitochondrial cytochrome oxidase subunit I (CO1) is a widely used barcode in a range of animal groups (Hebert et al., 2003), this locus is unsuitable for use in plants due to its low mutation rate (Kress et al., 2005; Cowen et al., 2006; Fazekas et al., 2008). In addition, complex evolutionary processes, such as hybridization and polyploidy, are common in plants, making species boundaries difficult to define (Rieseberg et al., 2006; Fazekas et al., 2009). The number and identity of DNA sequences that should be used for barcoding is a matter of debate (Pennisi, 2007; Ledford, 2008).

In conclusion, methods for identifying species by using short orthologous DNA sequences, known as “DNA barcodes”. In DNA barcoding the complete data can be obtained from single specimens irrespective of sexual morph or life stage. Morphologically indistinguishable taxa can be diagnosed without the need for live material. The core idea of DNA barcoding is based on the fact that short pieces of DNA can be found that vary only to a very minor degree within species, such that this variation is much less than between species. More pragmatically, DNA barcodes have proved useful in biosecurity, e.g. for surveillance of disease vectors (Besansky et al., 2003) and invasive insects (Armstrong and Ball, 2005), as well as for law enforcement and primatology (Lorenz et al., 2005). These “DNA barcodes” show promise in providing a practical, standardized, species-level identification tool that can be used for biodiversity assessment, life history and ecological studies, and forensic analysis.

References

Ali, M.A., Gábor, G., Norbert, H., Balázs, K., Al-Hemaid, F.M.A., Pandey, A.K. and Lee, J. (2014) The changing epitome of species

Plant DNA Barcoding and Phylogenetics, Pages 15-36
M. Ajmal Ali, Gyulai Gábor and Fahad Al-Hemaid (Eds.)

2 Molecular Markers for Plant DNA Barcoding

M.R. Enan

Introduction

Traditionally the macroscopic and microscopic identifications are performed to authenticate plant materials at the species level. There are an estimated 300000 plant species in the world (International Union for Conservation of Nature; IUCN, 2012), the accurate classification and identification of this large number of species remains a challenge even for specialist taxonomists. The emergence of DNA barcoding has had a positive impact on biodiversity classification and identification (Gregory, 2005). Several universal systems for molecular systematic analyses were used for lower taxa but were not successfully applied for broader range. The 'Barcode of Life' project aims to create a universal system for a eukaryotic species based on a standard molecular approach. It was initiated in 2003 by researchers at the University of Guelph in Ontario, Canada (http://www.barcoding.si.edu) and promoted in 2004 by the international initiative 'Consortium for the Barcode of Life' (CBOL). The DNA Barcode of Life Data System (BOLD, http://www.boldsystems.org) has progressively been developed since 2004 and was officially established in 2007 (Ratnasingham and Hebert, 2007). This data system enables the storage, analysis and publication of DNA barcode records.

Sample collection and DNA preservation

Total genomic DNA extraction from the collected plant tissue sample is the first step followed by amplification of desired region using barcode primer using PCR. The amplified sequence (amplicon) is then subjected to sequencing in one or both directions. The tools of bioinformatics are then used for the analyses of generated sequences (Figure1).

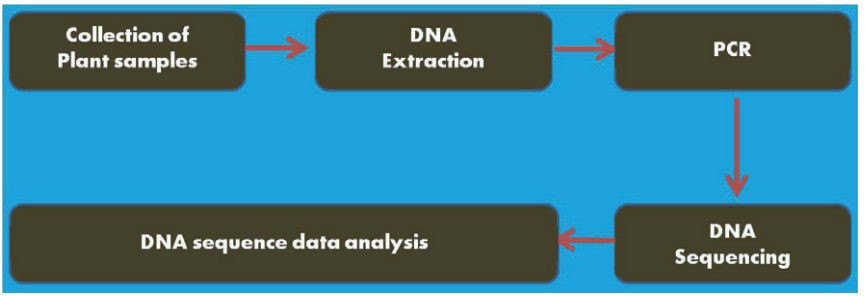

Fig.1: Flow chart demonstrating practical steps involved in plant DNA barcoding.

A close match quickly identifies a species that is already represented in the database. Three methods have been developed to preserve DNA in plant samples collected in the field (Kress, 2004; Gonzalez et al., 2009; Webb et al., 2010). One employs silica gel as a desiccant to rapidly dry the tissue, which reduces degradation in most specimens (Kress, 2004). However, it does not eliminate degradation, and DNA yields are low for some tissues (Condit, 1998). The second method uses a saturated NaCl-CTAB (cetyltrimethylammonium bromide) solution. The high salt partially dehydrates the tissues and the CTAB can complex with nucleic acids, proteins and carbohydrates to slow the degradation processes. However, high degrees of degradation have been noted in some cases with this method, and occasionally low yields of DNA result (Condit, 1998). The third method uses an absorbent paper for preserving the DNA (Webb et al., 2010). Pieces of plant tissues are mashed onto the paper, and then allowed to dry. Almost all methods to extract nucleic acids must be performed in a laboratory (Fine and Ree, 2006; Dick and Kress, 2009). Generally fresh tissues are used for extraction of the nucleic acids, because degradation and other biochemical processes begin immediately after the tissue has been removed from the organism or from its natural substrate. DNA in many species of plants has been detected in dried tissues from months to centuries after the organism has

died (Ratnasingham and Hebert, 2007; Dick and Heuertz, 2008). When the samples cannot be effectively sampled, preserved and transported rapidly to the laboratory, then alternatively the laboratory equipment and solutions can be transported to the target specimens in their natural environments in order to extract the DNA (*in situ*), a possible alternative that can minimize degradation and maximize yield. Degradation can be monitored by gel electrophoresis because as the DNA is broken down the higher molecular weight bands become more diffuse and smaller fragments of DNA are seen as increasingly bright smears of fluorescence extending into lower molecular weight regions of the gel. Simultaneously, other degradative processes also occurs, resulting in losses of sequence information (Ratnasingham and Hebert, 2007). The most common changes are losses of bases by hydrolytic attack of the glycosidic bonds. Depurination occurs at a higher rate, but depyrimidization occurs at a lower rate. When these DNAs are used as templates for PCR, approximately 75% of the time, an inaccurate base will be incorporated at those sites, causing a potential loss in sequence accuracy.

Plant DNA barcode primers

In every species, primer information is the most vital in starting the screening of various reported candidate genes towards their suitability. The forward and reverse sequences should be carefully combined (Table 1). Several universal primers for amplifying noncoding spacers of the chloroplast genome have been reported (Demesure et al., 1995). Most of the primers were designed for amplifying spacers between tRNA genes which have been proved variable among species (Demesure et al., 1996). Plant nuclear genes often occur in multiple copies and are highly variable, making the design of universal primers difficult (Yu et al., 2011).

Plant DNA barcode elements

For DNA barcoding to work, sequence variation must be high enough between species so that they can be discriminated from one another; however, it must be low enough within species that a clear threshold between intra- and inter-specific genetic variations can be defined. The two most important traits of DNA barcoding loci are the presence of conserved flanking regions to enable routine amplification across highly different taxa and sufficient internal variability to facilitate species discrimination but with a relatively low level of intra-specific variation.

Additional factors are short length facilitated routine sequencing, even with sub-optimal material, lack of heterozygosity enabling direct polymerase chain reaction followed by sequencing without cloning, ease of alignment that enables the use of character-based data analysis methods, lack of problematic sequence composition, such as regions with several microsatellites, that reduces sequence quality, universal capability to get amplified/sequenced with standardized primers, easy align ability and capability to get recovered easily from herbarium samples and other degraded DNA samples (Hollingsworth et al*.,*2009).

Types of plant DNA barcode markers

A total of 17 barcode regions (matK, rbcL, ITS, ITS2, psbA-trnH, atpF-atpH, ycf5, psbK-I, psbM-trnD, rps16, coxI, nad1, trnL-F, rpoB, rpoC1, atpF-atpH, rps16) of medicinal plants were reported to aid in the authentication and identification of medicinal plant materials. The majority of barcoding regions stated in the literature were the matK, ITS region, rbcL, and psbA-trnH. Although many studies have searched for a universal plant barcode, none of the available loci work across all species (Chase and Fay, 2009; Chen et al., 2010). The Consortium for the Barcode of Life-Plant Working Group (CBOL) recently recommended the two-locus combination of matK + rbcL as the best plant barcode with a discriminatory efficiency of only 72% (CBOL Plant Working Group, 2009). Taxonomists have suggested that a multi-locus method may be necessary to discriminate plant species (Hebert et al., 2004; Chase et al., 2007; Kress and Erickson, 2007; Erickson et al., 2008; Lahaye et al., 2008; Kane et al., 2012). However, CBOL demonstrated that the use of multiple loci did not clearly improve the species-level discriminatory ability of these techniques (CBOL Plant Working Group, 2009). Researchers have recently proposed the use of the whole-plastid genome sequence in plant identification (Erickson et al., 2008; Sucher and Carles, 2008; Parks et al., 2009; Nock et al., 2011; Yang et al., 2013). However this concept has not yet been universally accepted. One of the main concerns is the high sequencing cost and difficulties involved in obtaining complete plastid genome sequences in comparison to the use of single-locus barcodes. Hollingsworth et al. (2011) argued that the full plastid haplotype is not a good marker because it does not always track species boundaries. To date, it is still unclear whether plastid genomes can be regarded as a suitable barcode.

Table 1: Primer sequences for the candidate genes for barcoding in plants.

Barcode markers	**Primer sequence (5'-3')**	**Reference**
ITS2 (The second internal transcribed spacer of nuclear ribosomal DNA)	ITS3-F 5'- GCATCGATGAAGAACGTAGC-3' ITS4-R 5'- TCCTCCGCTTATTGATATGC-3'	White et al. (1990)
matK (Maturase coding gene)	matK472F 5'-CCCRTYCATCTGGAAATCTTGGTTC-3' matK1248R 5'-GCTRTRATAATGAGAAAGATTTCTGC-3'	Yu et al. (2011)
	3f-KIM-F 5'-CGTACAGTACTTTTGTGTTTACGAG-3' 1R KIM-R 5'-ACCCAGTCCATCTGGAAATCTTGGTTC-3'	CBOL Plant Working Group (2009)
	matK_1F 5'-GAACTCGTCGGATGGAGTG-3' matK_12R 5'-GAGAAATCTTTTTCATTACTACAGTG-3'	Wang et al. (1999)
	matK_2F 5'-CGTACTTTTATGTTTACAGGCTAA-3' matK_2R 5'-TAAACGATCCTCTCATTCACGA-3'	Wang et al. (1999)
rbcL (Ribulose1,5-biphosphate carboxylase oxygenase large subunit)	rbcL-af 5'- ATGTCACCACAAACAGAAAC-3 rbcL-724r 5'- TCGCATGTACCTGCAGTAGC-3	Kress and Erickson, 2007; Fay et al., 1997
rpoC1 (RNA polymerase γ-subunit gene)	rpoC1-F 5'-GGCAAAGAGGGAAGATTTCG-3 rpoC1-R 5'- CCATAAGCATATCTTGAGTTGG-3	Sass et al. 2007

trnH- psbA (Chloroplast intergenic spacer region)	psbA03_F 5'-GTTATGCATGAACGTAATGCTC-3 trnH-05_R 5'-CGCGCATGGTGGATTCACAATCC-3	Sang et al., 1997; Tate and Simpson, 2003
atpF-atpH (chloroplast intergenic spacer region)	atpF-F 5'-ACTCGCACACACTCCCTTTCC-3' atpH-R 5'-GCTTTTATGGAAGCTTTAACAAT-3'	Lahaye et al. (2008)
psbK-psbI (chloroplast intergenic spacer region)	psbK-F 5'-TTAGCCTTTGTTTGGCAAG-3' PsbI-R 5'-AGAGTTTGAGAGTAAGCAT-3'	Lahaye et al. (2008)
accD (Carboxytransferase-β-subunit)	accD-F 5'-AGTATGGGATCCGTAGTAGG-3' AccD-R 5'-TTTAAAGGATTACGTGGTAC-3'	Sass et al. (2007)
rpoB (RNA polymerase β-subunit gene)	rpoB-F 5'-AAGTGCATTGTTGGAACTGG-3' RpoB-R 5'-CCGTATGTGAAAAGAAGTATA-3'	Sass et.al. 2007
ndhJ (NADH Dehydrogenase subunit)	ndhj-F 5'-CATAGATCTTTGGGCTTYGA-3' Ndhj-R 5'-ATAATCCTTACGTAAGGGCC-3'	Sass et.al. 2007
ycf5 (Chloroplast intergenic spacer region)	ycf5-F 5'-GGATTATTAGTCACTCGTTGG-3' ycf5-R 5'-ACTTACGTGCATCATTAACCA-3'	Sass et.al. 2007

MaturaseK (matK)

The matK coding region is one of the most rapidly evolving regions in chloroplasts and shows a high level of species discrimination among angiosperms (Fazekas et al., 2008; Lahaye et al., 2008).The advantages of this gene are that it is easy to amplify, sequencing and alignment in most land plants and is a good DNA barcoding region for plants at the family and genus levels. Although the matK region is useful to determine species identity and the geographical origin of medicinal herbs, the success rate for the amplification and sequencing of matK region of some plant groups, such as cryptogams, is unsatisfactory and the universality of the amplification primers requires improvement (CBOL Plant Working Group, 2009). However, there are a few reports that some of the barcodes are universally useful for plants, it still remains mandatory to screen out the suited barcode for any new species (Rubinoff et al., 2006; Pennisi, 2007; Ledford, 2008). In general, the genes used in angiosperms are matK, rpoC1, rpoB, accD ,YCF5 and ndhJ whereas in non-angiosperms matK, rpoC1, rpoB, accD, and ndhJ are used (http://www.rbgkew.org.uk/barcoding/index.html). With higher potential to identify the variation, easy amplification and alignment, a portion of the plastid matK gene was proposed as a universal DNA barcode for flowering plants (Lahaye et al., 2008). The choice of rbcL+matK as a core barcode was based on the straightforward recovery of the rbcL region and the discriminatory power of the matk region. The matK gene is one of the most rapidly evolving coding sections of the plastid genome (Hilu and Liang, 1997). Studies by Newmaster et al. (2008) in Myristicaceae and Seberg and Petersen (2009) in *Crocus* have confirmed matK and the intergenic spacer trnH-psbA as suitable land plant barcodes. The matK gene has a high evolutionary rate, suitable length and obvious interspecific divergence as well as a low transition/transversion rate (Min and Hickey, 2007; Selvaraj et al., 2008). But the matK is difficult to amplify universally using currently available primer sets. The CBOL Plant Working Group (2009) revealed nearly 90% success rate in amplifying angiosperm DNA using a single primer pair. However, the success was limited in gymnosperms (83%) and much worse in cryptogams (10%) even with multiple primer sets. The matK gene can discriminate more than 90% of species in the Orchidaceae (Kress and Erickson, 2007) but less than 49% in the nutmeg family (Newmaster et al., 2008). Fazekas et al. (2008) attempted the identification of 92 species from 32 genera using the matK barcode but only achieved a success rate of 56%. These findings demonstrate that the matK barcode alone is not a suitable universal barcode.

Ribulose 1,5-biphosphate carboxylase oxygenase large subunit (rbcL)

The large subunit of ribulose-bisphosphate carboxylase, rbcL region is a chloroplast gene coding region that has a high amplification success rate in a broad range of flowering plant, gymnosperm and cryptogam species, plus high sequence quality among seven loci tested (CBOL Plant Working Group, 2009). However, the rbcL region showed the lowest divergence (0.83%) among 11 potential barcoding loci tested for the differentiation of two species in Solanaceae, (Kress et al., 2005). Low interspecific variation was also observed between other herbal medicinal materials and their adulterants. However, rbcL sequences evolve slowly and this locus has by far the lowest divergence of plastid genes in flowering plants (Kress et al., 2005). Consequently, it is not suitable at the species level due to its modest discriminatory power (Fazekas et al., 2008; Lahaye et al., 2008; CBOL Plant Working Group, 2009; Chen et al., 2010). Despite these limitations, rbcL was still suggested as one of the best potential candidate plant DNA barcodes based on the straightforward recovery of the gene sequence (Blaxter, 2004; CBOL Plant Working Group, 2009; Hollingsworth et al., 2011). Although rbcL by itself does not meet the desired attributes of a DNA barcoding locus, it is accepted that rbcL in combination with various plastid or nuclear loci can make accurate identifications (Newmaster et al., 2006; Chase et al., 2007; Kress and Erickson, 2007; CBOL Plant Working Group, 2009; Hollingsworth et al., 2009). CBOL demonstrated that the use of seven candidate loci did not significantly improve species-level discriminatory ability compared to rbcL + matK. Thus, the combinations of candidate loci cannot eliminate the inherent deficiencies of current DNA barcoding of plants.

Nuclear barcode marker (ITS)

A variety of loci have been suggested as DNA barcodes for plants, including coding genes and non-coding spacers in the nuclear and plastid genomes (Figure 2). The internal transcribed spacer (ITS) region comprises the ITS1 intergenic spacer, 5.8S rDNA, and the ITS2 intergenic spacer (ITS1-5.8S-ITS2), with size ranging from 400 to 1000 bp in total. This is the most frequently sequenced region for plant phylogenetic studies because of its high species discrimination and technical ease of amplification (Alvarez and wendel, 2003; Kress et al., 2005). Although the ITS region and ITS2 intergenic spacer can help identify herbal medicinal materials by DNA sequencing, these regions sometimes require cloning because of the presence of multiple copies

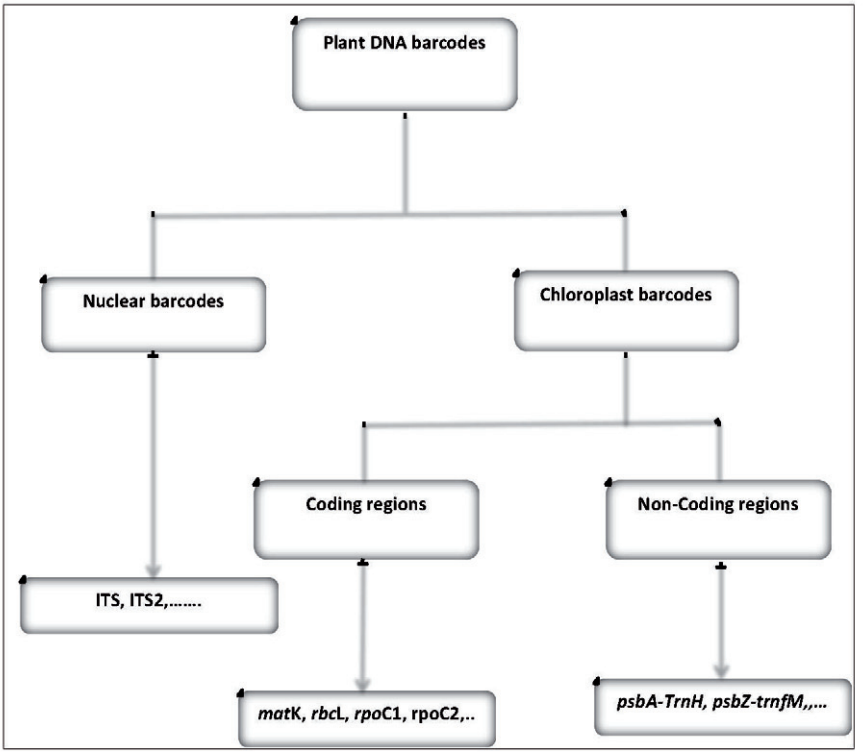

Fig. 2: Schematic illustration of employed DNA barcode markers

-and the problems of secondary structure resulting in poor-quality sequence data (Baldwin et al., 1995; Alvarez and Wendel, 2003). Fungal -contamination is common in herbal medicinal materials that are improperly processed and stored. Fungal ITS sequences are readily amplified using universal primers, generating false-positive PCR results. To overcome this issue, plant-specific primers need to be designed (Zhang et al., 1997; Cullings and Vogler, 1998). The greater discriminatory power of ITS over plastid regions at low taxonomic levels has been widely studied leading to it also being suggested as a plant barcode (Stoeckle, 2003; Kress et al., 2005; Sass et al., 2007), especially in parasitic plants which offer less resolution from plastid barcodes (Hollingsworth et al., 2011). However, CBOL has only regarded ITS as a supplementary locus (CBOL Plant Working Group, 2009). Some limitations prevent it from being a core barcode: incomplete concerted evolution, fungal contamination and difficulties of amplification

and sequencing (Hollingsworth et al., 2011). Plant BOL Group recently argued that when direct sequencing was possible, the ITS region should be incorporated into the core barcodes because of higher discriminatory power than plastid barcodes (CBOL Plant Working Group, 2011). To resolve the difficulties involved in sequencing the entire ITS, they suggested ITS2 as a backup because of its conserved sequence characters which reduce amplification and sequencing problems. It was accepted that ITS2 could be used as a novel universal barcode for the identification of a broader range of plant taxa (Chen et al., 2010; Gao et al., 2010a,b; Pang et al., 2010) even from herbarium specimens with degraded DNA (Chiou et al., 2007). Song et al. (2012) recently showed that the ITS2 intra-genomic distances were markedly smaller than those of the intra-specific or inter-specific variants in a wide range of plant families. Internal transcribed spacer regions of nuclear ribosomal DNA (ITS) is commonly recommended based on the facts that these are often highly variable in angiosperms at the generic and species level and divergent copies are often present within single individuals (Kress et al., 2005). 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.

TrnH-psbA spacer

TrnH-psbA is currently the most widely used plastid DNA barcode marker. The size of the trnH–psbA region of most flowering plants ranges between 340 and 660 bp. This region shows the highest amplification success rate (100%) and discrimination rate (83%) among nine loci tested (Kress et al., 2005; Kress and Erickson, 2007). Therefore, this intergenic spacer appears to be a useful region for the differentiation of medicinal plants from their adulterants. The presence of highly conserved coding sequences on both sides make the design of universal primers feasible (Shaw et al., 2005), with a single primer pair likely to amplify nearly all angiosperms (Shaw et al., 2007). The non-coding intergenic region exhibits most sequence divergence and has high rates of insertion/deletion (Kress and Erickson, 2007). These attributes make trnH-psbA highly suitable as a plant barcode for species discrimination (Kress and Erickson, 2007; Shaw et al., 2007).

Alignment of the trnH-psbA spacer can be highly ambiguous because of its complicated molecular evolution, considerable length variation (Chang et al., 2006), and high rates of insertion/deletion in larger families of angiosperms (Chase et al., 2007). Furthermore, due to the presence of duplicated loci and a pseudogene, the trnH-psbA sequence is much

longer >1000 base pairs in some conifers and monocots (Chase et al., 2007; Hollingsworth et al., 2009) while it is exceedingly short, less than 300 bp, in other groups (Kress et al., 2005) and shorter than 100 bp in Bryophytes (Stech and Quandt, 2010). One of the key problems associated with the use of trnH-psbA as a standard barcode is the frequent inversions in some plant lineages, which may lead to large overestimates of genetic divergence and to incorrect phylogenetic assignment (Whitlock et al., 2010). Additionally, because of the premature termination of sequencing reads caused by mononucleotide repeats, longer trnH-psbA regions can be difficult to retrieve without taxon-specific internal sequencing primers designed to obtain high-quality bi-directional sequences (Devey et al., 2009; Ebihara et al., 2010). Shorter trnH-psbA spacers may not have adequate sequence variation for species discrimination. As a consequence, Kress et al. (2005) and Chase et al. (2007), respectively, proposed that trnH-psbA can be used in two-locus or three-locus barcode systems to provide adequate resolution. Kress et al. (2005) also proposed that the trnH-psbA plastid inter-genic spacer region would be a suitable universal barcode for land plants.

Multilocus plant DNA barcoding approaches

Despite extensive efforts to identify a universal plant DNA barcode comparable to CO1 in animals, the task has proved difficult due to the lack of adequate variation within single loci (Kress et al., 2005; Newmaster et al., 2006; Chase et al., 2007; Kress and Erickson, 2007; Sass et al., 2007; Fazekas et al., 2008; Lahaye et al., 2008). Many researchers have suggested that a multi-locus method will be required to obtain adequate species discrimination (Hebert et al., 2004; Kress and Erickson, 2007; Erickson et al., 2008; Kane and Cronk, 2008; Lahaye et al., 2008; CBOL Plant Working Group, 2009; Chase and Fay, 2009). Various combinations of plastid loci have been proposed including rbcL + trnH-psbA (Kress and Erickson, 2007), rpoC1 + rpoB + matK or rpoC1 + matK + trnH-psbA (Chase et al., 2007) and matK + atpF-atpH + psbK-psbI or matK + atpF-atpH + trnH-psbA (Pennisi, 2007). These combined barcodes exhibit higher species discrimination than single-locus approaches. Different research groups have tested different combinations using different taxa while attempting to select a universal barcode, however universal agreement is yet to be reached. Fazekas et al. (2008) compared these barcode combinations using the same large-scale taxonomic samples, but none could identify more than 70% of tested species. de Boer et al. (2014) demonstrated that combining psbA-trnH, rpoC1, and ITS allowed the majority of the market samples to be