DNA barcoding of medicinal plant material for identification

Natascha Techen¹, Iffat Parveen¹, Zhiqiang Pan² and Ikhas A Khan¹,³

Because of the increasing demand for herbal remedies and for authentication of the source material, it is vital to provide a single database containing information about authentic plant materials and their potential adulterants. The database should provide DNA barcodes for data retrieval and similarity search. In order to obtain such barcodes, several molecular methods have been applied to develop markers that aid with the authentication and identification of medicinal plant materials. In this review, we discuss the genomic regions and molecular methods selected to provide barcodes, available databases and the potential future of barcoding using next generation sequencing.

Addresses
¹ National Center for Natural Products Research and Research Institute of Pharmaceutical Sciences, Department of Pharmacognosy, School of Pharmacy, University of Mississippi, P.O. Box 1848, MS 38677, USA
² USDA-ARS-NPURU, P.O. Box 8048, University, MS 38677, USA
³ School of Pharmacy, King Saud University, Saudi Arabia

Corresponding author: Khan, Ikhas A (ikhan@olemiss.edu)

Current Opinion in Biotechnology 2014, 25:103–110
This review comes from a themed issue on Analytical biotechnology
Edited by Frank L Jaksch and Savaş Tay

0958-1669/$ – see front matter, © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.copbio.2013.09.010

Introduction

The traditional system of medicine utilizes medicinal plants to cure various ailments but the herbal industry suffers from substitution and adulteration of medicinal herbs with closely related species. The efficacy of the drug decreases if it is adulterated, and in some cases, can be lethal if it is substituted with toxic adulterants. Hence, the correct formulation is important for the medicinal herb to be effective. The traditional methods of medicinal plant identification include organoleptic methods (identification by the senses: taste, sight, smell, touch), macroscopic and microscopic methods (identification by shape, colour, texture) and chemical profiling (e.g. TLC, HPLC-UV, HPLC-MS). However, neither method can identify the related species easily in processed products because the former method requires trained personal for macroscopic and microscopic examinations. In the latter method, chemical profiles or markers may be affected by physiological and storage conditions. Authentication at the DNA level provides more reliability because, in contrast to RNA, DNA is a stable macromolecule that is not affected by external factors and is found in all tissues. Therefore, development of DNA-based markers is important for authentication of medicinal plants.

The novel technique of identifying biological specimens using short DNA sequences from either nuclear or organelle genomes is called DNA barcoding. The term ‘DNA barcode’ as taxon identifiers was first proposed by Paul Hebert of University of Guelph in 2003 [1]. He recommended that the 5’ end of cytochrome c oxidase 1 (CO1) from the mitochondrial genome was sufficient to generate DNA barcodes for the identification of animals [1–4]. On the basis of this initial success with animals, CO1 was suggested as the locus that could provide recognition tags for all organisms. They further emphasized that DNA barcoding not only helps in the identification of species but can also define species boundaries, flagging of new species and species delimitation [2,3]. However, in plants the mitochondrial genes are slowly evolving, with very low substitution rates and were not suitable for barcoding. Therefore, the search for plant barcode shifted to chloroplast and nuclear genomes with high substitution rates. Following initial in-silico and laboratory-based assessment of different loci from chloroplast and nuclear genomes led to the conclusion that no single locus plant barcode exists, and soon it was realized that multi-locus barcodes are requisite for plant barcoding. Subsequently a number of loci were being tested for their suitability as plant barcodes and many multi-locus combinations were suggested. The Consortium for the Barcode of Life Plant Working Group (CBOL) [5] evaluated seven chloroplast genomic regions across the plant kingdom and proposed a combination of matK and rbcL as plant barcodes. High universality but less species resolution is provided by rbcL whereas matK affords high resolution but less universality. A combination of these two can help to achieve maximum species discrimination. Nevertheless, in closely related species, the discriminating ability of these two markers is low [6,7]. Therefore, the China Plant BOL Group [8] proposed the addition of nuclear ITS (Internal Transcribed Spacer) to the matK + rbcL combination as plant barcode in order to achieve maximum identification rates even in closely related species.

The aim of this review is to assess the progress made so far in the field of DNA barcoding in relation to the identification of botanicals. In the current paper, we review the genomic regions selected as possible barcodes for medicinal plants and the emerging new methods for rapid generation of barcodes. We also discuss the challenges of barcoding and what databases are available to retrieve barcodes of medicinal plants, their substitutes and adulterants.
Loci suggested as plant barcodes

At the Fourth International Barcode of Life Conference (http://www.dnabarcodes2011.org/) the option of a three-locus barcode (matK + rbcl + psbA-trnH) versus a two-locus barcode was discussed. The two-locus barcode was preferred to avoid the increased costs of sequencing three loci rather than two in very large sample sets, and to prevent further delays in implementing a standard barcode for land plants. The barcode combination rbcl + matK was the preferred choice.

A search of the literature in SciFinder (a chemical abstracts service database) from 2010 to 2013 resulted in 60 publications (Figure 1 and Table 1). In the literature analyzed in this review, a total of 17 barcode regions (matK, rbcl, ITS, ITS2, psbA-trnH, atpF-atpH, ycf5, psbK-1, psbM-trnD, rps16, coxl, nad1, trnL-F, rpoB, rpoC1, atpF-atpH, rpoA) of medicinal plants were reported to aid in the authentication and identification of medicinal plant materials. The majority of barcoding regions mentioned in the literature were the ITS region (26 references), psbA-trnH (21 references), matK (19 references), and rbcl (14 references). Further genomic regions used for barcoding were ITS2 (9 references), rpoC1 (6 references), rpoB (4 references), and trnL-F (3 references).

Besides using known genomic regions, other PCR-based methods have been applied to develop markers that help with the authentication and identification of medicinal plant material: RAPD, RFLP, microsatellites, ISSRs, SNPs, and ARMS (12, 3, 1, 2, 2, 2 publications, respectively). SCAR markers (17 references) have been developed from RAPD, ISSR and a variety of genomic regions.

Zuo et al. [14] analyzed 12 genomic regions of 95 ginseng samples, representing all the species in the genus. They demonstrated that the combination of psbA-trnH and ITS was sufficient for the identification of all species and species clusters in the genus. For various samples analyzed, the combination of up to three genomic regions (matK, rbcl, ITS, psbA-trnH) provided the required information for identification [15, 35, 38].

Chen et al. [11] analyzed >6600 plant samples belonging to 4800 species from 753 distinct genera using the genomic regions psbA-trnH, matK, rbcl, rpoC1, ycf5, ITS, and ITS2. Their data suggested that the second internal transcribed spacer (ITS2) of nuclear ribosomal DNA represents the most suitable region for DNA barcoding applications. The percentage of successful identification with ITS2 at the species level was 92.7%. He et al. and Selvaraj et al. [9, 18] also analyzed multiple genomic barcode regions and came to a similar conclusion that ITS or ITS2 showed the highest discrimination rate among the samples analyzed. In contrast, the highest discrimination rate was found to be from psbA-trntrnH [10, 12, 39] or from matK [41, 48].

Theodoridis et al. [36] analyzed medicinal plants of the Labiatae (Lamiaceae) family from Chios Island (Greece) and the adjacent Cesme-Karaburun Peninsula (Turkey) using all three barcode loci matK, rbcl, psbA-trnH and...
<table>
<thead>
<tr>
<th>Ref.</th>
<th>Analyzed material</th>
<th>Genomic regions analyzed/molecular methods applied</th>
</tr>
</thead>
<tbody>
<tr>
<td>[12]</td>
<td>Loniceria spp.</td>
<td>matK, rbcL, ITS, p5bA-trnH, trnL-F</td>
</tr>
<tr>
<td>[16]</td>
<td>Various medicinal roots</td>
<td>matK, ITS, p5bA-trnH, rpoC1</td>
</tr>
<tr>
<td>[17]</td>
<td>Various medicinal plants</td>
<td>ITS, ITS2</td>
</tr>
<tr>
<td>[20]</td>
<td>Various (old) medicinal plants</td>
<td>ITS, ITS2</td>
</tr>
<tr>
<td>[21]</td>
<td>Rubus spp.</td>
<td>ITS, p5bA-trnH, tmL-F, SCAR, SNP</td>
</tr>
<tr>
<td>[22]</td>
<td>Hypericum spp.</td>
<td>ITS, SCAR</td>
</tr>
<tr>
<td>[23]</td>
<td>Ochradenus spp.</td>
<td>ITS, rpoB, rpoC1</td>
</tr>
<tr>
<td>[24]</td>
<td>Rehmannia spp.</td>
<td>ITS, RAPD, SCAR</td>
</tr>
<tr>
<td>[25]</td>
<td>Medicinal vines (22 genera)</td>
<td>ITS</td>
</tr>
<tr>
<td>[26]</td>
<td>Dipsacus spp.</td>
<td>ITS</td>
</tr>
<tr>
<td>[27]</td>
<td>Dendrobium spp.</td>
<td>ITS, ARMS</td>
</tr>
<tr>
<td>[28]</td>
<td>Paris spp.</td>
<td>ITS, RFLP</td>
</tr>
<tr>
<td>[29]</td>
<td>Citrus spp.</td>
<td>ITS, SNP</td>
</tr>
<tr>
<td>[30]</td>
<td>Dendrobium spp.</td>
<td>ITS</td>
</tr>
<tr>
<td>[31]</td>
<td>Ruta spp.</td>
<td>ITS, rpoB, rpoC1</td>
</tr>
<tr>
<td>[32]</td>
<td>Astragalus spp.</td>
<td>ITS</td>
</tr>
<tr>
<td>[33]</td>
<td>Various medicinal plants</td>
<td>ITS</td>
</tr>
<tr>
<td>[34]</td>
<td>Meconopsis spp.</td>
<td>ITS</td>
</tr>
<tr>
<td>[36]</td>
<td>Lamiaceae</td>
<td>matK, rbcL, p5bA-trnH</td>
</tr>
<tr>
<td>[37]</td>
<td>Various medicinal plants</td>
<td>matK, rbcL, p5bA-trnH</td>
</tr>
<tr>
<td>[38]</td>
<td>Sibasia spp.</td>
<td>matK, rbcL, p5bA-trnH</td>
</tr>
<tr>
<td>[39]</td>
<td>Pteridophytes</td>
<td>matK, rbcL, p5bA-trnH, rpoB, rpoC1</td>
</tr>
<tr>
<td>[40]</td>
<td>Vitex spp.</td>
<td>matK, p5bA-trnH, RFLP</td>
</tr>
<tr>
<td>[41]</td>
<td>Sideritis spp.</td>
<td>matK, p5bA-trnH</td>
</tr>
<tr>
<td>[42]</td>
<td>Paris spp. and adulterants</td>
<td>p5bA-trnH</td>
</tr>
<tr>
<td>[43]</td>
<td>Senna spp.</td>
<td>p5bA-trnH</td>
</tr>
<tr>
<td>[44]</td>
<td>Smlax spp.</td>
<td>p5bA-trnH</td>
</tr>
<tr>
<td>[45]</td>
<td>Phyllanthus spp.</td>
<td>p5bA-trnH</td>
</tr>
<tr>
<td>[46]</td>
<td>Cistance spp.</td>
<td>p5bA-trnH</td>
</tr>
<tr>
<td>[47]</td>
<td>Galpemia spp.</td>
<td>matK, rbcL, rpoC1</td>
</tr>
<tr>
<td>[48]</td>
<td>Dendrobium spp.</td>
<td>matK, rbcL</td>
</tr>
<tr>
<td>[49]</td>
<td>Rhubarb</td>
<td>matK, SCAR</td>
</tr>
<tr>
<td>[50]</td>
<td>Pueraria candolbei, Butea superb, Mucuna colletti</td>
<td>matK, RFLP</td>
</tr>
<tr>
<td>[51]</td>
<td>Orobanche spp. and adulterants</td>
<td>ITS2</td>
</tr>
<tr>
<td>[52]</td>
<td>Taraxacum and adulterants</td>
<td>ITS2, SCAR</td>
</tr>
<tr>
<td>[53]</td>
<td>Cuscuta spp. and adulterants</td>
<td>RAPD, SCAR</td>
</tr>
<tr>
<td>[54]</td>
<td>Artemisia spp.</td>
<td>RAPD, SCAR</td>
</tr>
<tr>
<td>[55]</td>
<td>Linope spp., Ophiopogon spp.</td>
<td>RAPD, SCAR</td>
</tr>
<tr>
<td>[56]</td>
<td>Rheum spp.</td>
<td>RAPD, SCAR</td>
</tr>
<tr>
<td>[57]</td>
<td>Glycerhiza spp.</td>
<td>SCAR</td>
</tr>
<tr>
<td>[58]</td>
<td>Zzyphus spp.</td>
<td>RAPD, SCAR</td>
</tr>
<tr>
<td>[59]</td>
<td>Crocus spp. and adulterants</td>
<td>RAPD, SCAR</td>
</tr>
<tr>
<td>[60]</td>
<td>Prunella spp.</td>
<td>RAPD, SCAR</td>
</tr>
<tr>
<td>[61]</td>
<td>Viola spp. and adulterants</td>
<td>RAPD, SCAR</td>
</tr>
<tr>
<td>[62]</td>
<td>Srocophila spp.</td>
<td>SCAR, ISSR</td>
</tr>
<tr>
<td>[63]</td>
<td>Panax ginseng</td>
<td>RAPD, SCAR</td>
</tr>
<tr>
<td>[64]</td>
<td>Uncaria spp.</td>
<td>RAPD, SCAR</td>
</tr>
<tr>
<td>[65]</td>
<td>Marsdenia and adulterants</td>
<td>RAPD, SCAR</td>
</tr>
<tr>
<td>[66]</td>
<td>Phyllanthus spp.</td>
<td>ISSR</td>
</tr>
<tr>
<td>[67]</td>
<td>Asparagus and adulterants</td>
<td>ARMS</td>
</tr>
<tr>
<td>[68]</td>
<td>Dendrobium spp.</td>
<td>MS</td>
</tr>
</tbody>
</table>

MS = microsatellites.
tested them either as single-region or as multi-region barcodes. They showed that \textit{matK} and \textit{psbA-trnH} were as useful in discriminating species of the \textit{Labiatae} as any multi-region combination. For the \textit{Labiatae} species analyzed \textit{matK} and \textit{psbA-trnH} could serve as single-region barcodes. Schori and Showalter [37] analyzed the \textit{rbcL}, \textit{matK}, \textit{psbA-trnH} loci of fourteen species of medicinal plants from Pakistan and found that, depending on plant to be identified, one region was preferred over the other to aid with species identification.

Single barcode-regions for identification have been reported for (i) \textit{matK} [41,49,50], (ii) ITS [22*,25–30,32–34], (iii) ITS2 [17,20*,51,52], and (iv) \textit{psbA-trnH} [42–46].

Not only genomic regions, but also the usefulness of RAPD-based DNA markers has been reported for medicinal plant identification. RAPD and SCAR markers have been developed for the identification of several medicinal plant materials [53–56,58–61,62].

The review by Heubl [69**] focused on the most commonly used DNA-based technologies (RAPD, RFLP, ARMS, CAPS, AFLP, DAF, ISSR, SSR, sequencing, hybridization and microarrays) and new achievements in the field of DNA barcoding and DNA chip technology used to identify traditional Chinese medicinal materials. This particular review provides a critical view of the methods presented concerning sensitivity, reproducibility, and dependability.

**Future developments**

Although the traditional DNA barcoding techniques remain an effective DNA method for identification of medicinal plants, the more advanced and newly developed high throughput sequencing, specifically next-generation sequencing (NGS) technologies [70], could be adopted and potentially revolutionize the process. Even though DNA barcoding usually targets short regions of DNA molecule within the genome and does not require full genome-scale data, the potential of using NGS to simultaneously identify multiple species or bulk samples of organisms by sequencing DNA barcodes is enormous.

The continuing improvement in NGS technologies and the massive expansion of reference sequence databases have made the NGS approach promising. However, only one publication using NGS for medicinal plant DNA barcoding has appeared to date in which it was utilized to identify potential nuclear genomic regions for barcoding [71]. But most recently, there have been a number of cases published in which NGS was employed, for example, to explore the intragenomic divergence of Dikarya [72], and the taxonomic characterization of freshwater diatom communities [73]. Presently NGS and bioinformatics are used to probe and deduce the transcriptome of medicinal plants to provide a robust resource for gene discovery and downstream metabolic pathway discovery [74–76]. Since medicinal plants tend to have a large genome, the combination of targeted genomic enrichment [77] and NGS could ultimately make the technologies applicable in DNA barcoding of medicinal plant material (Figure 2).

**Challenges and limitations of barcoding**

The isolation of pure, high molecular weight DNA is critical for the successful application of molecular methods. This can be quite a challenge since in processed medicinal plant material the DNA is often highly degraded or the plant material contains high amounts of polysaccharides, polyphenols and other secondary metabolites, such as, alkaloids and flavonoids. Various commercial kits and modified traditional methods are available to yield in good quality DNA from raw and powdered medicinal plant material, herbarium specimens, capsules, tablets, or tinctures for downstream applications [22*,78,79**,80,81]. Särkinen et al. [79**] found a strong negative correlation between amplicon size and PCR success, indicating that shorter fragments are easier to amplify from herbarium DNA. The relatively large genomic region \textit{rbcL} (670 bp) was amplified only in 10% of the analyzed samples, the medium sized LEAFY intron 3 (260 bp) amplified in 24%, while the small

![Figure 2](https://example.com/figure2.png)

Current and proposed barcoding applications of botanicals.
genomic region \textit{trn}L P6 loop (10–143 bp) was successfully amplified in 78% of the samples. For heavily damaged DNA, it is therefore recommended to develop conditions to produce very short amplicons, which are easier to amplify, or perform DNA ‘repair reactions’ [79**,82], and that are also available as commercial kits. Särkinen \textit{et al.} [79**] believes that fragmented DNA of medicinal plants is going to be far less of a problem in the near future using the NGS approach since this technique requires fragmented (and ligated) DNA as starting material.

**Availability of data**

It is desirable to have access to a single barcode library for all medicinal material used (including fungal and animal species). Currently, however several barcode libraries are freely accessible (see also review by Taylor and Harris [83]):

(i) BOLD (The barcode of life data system) [84] (http://www.barcodinglife.com) was created and is maintained by the University of Guelph in Ontario. It offers researchers a way to collect, manage, and analyze DNA barcode data. The goal is, over the next 20 years, to provide a barcode library for all eukaryotic life.

(ii) CBOL (Consortium for the barcode of life) (http://www.barcodeoflife.org/) is a public reference library of species identifiers which could be used to assign unknown specimens to known species. CBOL was founded in 2004 and promotes barcoding through working groups, networks, workshops, conferences, outreach, and training. CBOL has 200 member organizations from 50 countries and operates from a Secretariat Office located in the Smithsonian Institution’s National Museum of Natural History in Washington, DC [5].

(iii) iBOL (International Barcode of Life project) (http://www.ibol.org/) consists of a group of hundreds of scientists from 25 nations working together to construct a DNA barcode reference library that will be the foundation for a DNA-based identification system for all multi-cellular life. Their five-year (2010–2015) goal is to barcode five million specimens representing 500,000 species.

(iv) The GenBank online genetic sequence database (http://www.ncbi.nlm.nih.gov/genbank/) [85] is possibly one of the most important repositories of genetic information. GenBank contains over 108 million entries for over 260,000 named organisms and is one of the most frequently used databases for genomic authentication [86]. With the BLAST program [87] an unknown DNA sequence can be rapidly and accurately compared to known and well characterized sequences.

(v) MMDBD (Medicinal Materials DNA Barcode Database) (http://137.189.42.34/mherbsdb/index.php) is a website that includes DNA sequences and information and key references of the medicinal materials recorded in the Pharmacopoeia of the People’s Republic of China, American Herbal Pharmacopoeia and other related references. This database, updated in May 2012 with 1658 species and 31,468 sequences available, provides information material for distinguishing medicinal materials (plant, animal, and fungi) from their common substitutes and adulterants [88**].

(vi) The GDR (Genome Database for Rosaceae), founded in 2009, provides genetic markers and ESTs of \textit{Rosaceae}. A large number of species in \textit{Rosaceae} or rose family have a medicinal value (http://www.rosaceae.org/).

**Conclusion**

Molecular barcoding methods are reliable tools for the identification of medicinal plants, their substitutes and adulterants at the genus and species level. The methods discussed provide consistent and reliable results regardless of the age, plant part, or environmental factors of the sample.

Based on the literature analyzed in this review, it appears that, although the Barcode of Life Plant Working Group [5] recommends the genomic regions \textit{rbcL} + \textit{matK} for barcoding, often other genomic regions could be more useful for medicinal material identification. Furthermore, depending on the material analyzed, one or the combination of up to three genomic regions was necessary to provide the required information for identification.

Because of the increasing demand for herbal remedies, authentication of the medicinal plant material is important; therefore it is vital to provide a sole, extensive database with DNA data for easy identification.

Future sequencing advances to analyze large scale nucleic acid sequences have a great potential for genotyping and taxon identification, even for damaged or fragmented template DNA. In the near future, full genomic sequence data from medicinal plants will be available. The challenge will be to provide sufficient storage space for all the data produced.

**Acknowledgements**

This research was funded in part by the Food and Drug Administration and USDA. We thank Jon Parcher for his revision of the manuscript and suggestions.

**References and recommended reading**

Papers of particular interest, published within the period of review, have been highlighted as:

- of special interest
- of outstanding interest


Herbarium voucher specimen, some 90 years old, were used to compare amplification success of the ITS (average length 634 bp) and ITS2 (average length 233 bp) region. Efficiency was 23% and 91% respectively. 12,861 plant sequences were used to compare identification efficiency of ITS and ITS2. Results show ITS2 can serve as minibarcode.


This article reports the identification technique for Hypericum perforatum, St. John’s Wort. A PCR based method has been applied to DNA extracted from capsules, tablets, and tinctures. Extracting good quality DNA from the samples is a challenge. Only four out of 13 products yielded in a PCR product due to low quality of DNA.


34. Li R, Dao Z: Identification of Meconopsis species by a DNA barcode sequence: the nuclear internal transcribed spacer


This article provides a critical evaluation of currently applied molecular biology methods in the field of barcoding. It gives a critical view concerning sensitivity, reliability, reproducibility and running costs. Recent achievements in the next generation technologies are briefly outlined.


The comparison of various DNA extraction methods from Herbarium samples often stored under DNA degrading conditions. PCR amplification success is measured by using three differentially sized markers. The shorter the PCR product, the higher the amplification success rate.


This article introduces a new database: Medicinal Materials DNA Barcode Database MMDBD for data retrieval and similarity search. The database contains information about the Chinese medicinal material including known adulterants, photographs, and primers used for obtaining the DNA barcodes.