

RESEARCH ARTICLE

Unidentifiable by morphology: DNA barcoding of plant material in local markets in Iran

Abdolbaset Ghorbani^{1,2}, Yousef Saeedi², Hugo J. de Boer^{1,3,4*}

1 Department of Organismal Biology, Evolutionary Biology Centre, Uppsala University, Uppsala, Sweden, **2** Traditional Medicine and Materia Medica Research Center, Shahid Beheshti University of Medical Sciences, Tehran, Iran, **3** Naturalis Biodiversity Center, Leiden, The Netherlands, **4** The Natural History Museum, University of Oslo, Oslo, Norway

* Hugo.DeBoer@ebc.uu.se



Abstract

Local markets provide a rapid insight into the medicinal plants growing in a region as well as local traditional health concerns. Identification of market plant material can be challenging as plants are often sold in dried or processed forms. In this study, three approaches of DNA barcoding-based molecular identification of market samples are evaluated, two objective sequence matching approaches and an integrative approach that coalesces sequence matching with *a priori* and *a posteriori* data from other markers, morphology, ethnoclassification and species distribution. Plant samples from markets and herbal shops were identified using morphology, descriptions of local use, and vernacular names with relevant floras and pharmacopoeias. DNA barcoding was used for identification of samples that could not be identified to species level using morphology. Two methods based on BLAST similarity-based identification, were compared with an integrative identification approach. Integrative identification combining the optimized similarity-based approach with *a priori* and *a posteriori* information resulted in a 1.67, 1.95 and 2.00 fold increase for ITS, *trnL-F* spacer, and both combined, respectively. DNA barcoding of traded plant material requires objective strategies to include data from multiple markers, morphology, and traditional knowledge to optimize species level identification success.

OPEN ACCESS

Citation: Ghorbani A, Saeedi Y, de Boer HJ (2017) Unidentifiable by morphology: DNA barcoding of plant material in local markets in Iran. PLoS ONE 12(4): e0175722. <https://doi.org/10.1371/journal.pone.0175722>

Editor: Rainer Bussmann, Missouri Botanical Garden, UNITED STATES

Received: May 18, 2016

Accepted: March 30, 2017

Published: April 18, 2017

Copyright: © 2017 Ghorbani et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: The research was supported by a grant to AG from the International Foundation for Science, Sweden (<http://www.ifs.se/>) and a grant to HdB from Carl Tryggers Foundation, Sweden (<http://www.carltryggersstiftelse.se/>). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Introduction

In places where traditional medicine plays a major role in delivering health care, local markets can provide a rapid insight into medicinal plants harvested and traded in a region [1]. Marketplaces do not only reflect the diversity and prevalence of medicinal plants growing in the region, but also yield an understanding of the plants' utilization, seasonal availability, and chain of commercialization, as well as local health and disease concerns and the importance of traditional medicine among local people [1,2].

Medicinal plants in markets and traditional herbal shops are normally traded as dried leaves, roots and barks, or in processed forms such as powdered parts, mixtures, or extracts. Such plant material lacks many of the morphological characteristics necessary for accurate identification by retailers and customers. Aerial parts may lose important diagnostic characters

Competing interests: The authors have declared that no competing interests exist.

necessary for taxonomic identification, and the identification of roots is always challenging due to a lack of distinctive morphology [3,4]. Moreover, taxonomic identification using macro- and micro-morphological and organoleptic methods can be time-consuming, error-prone and requires expertise and reliable references [5,6].

Many studies of market ethnobotany include vernacular names as part of the identification diagnostics, combined with morphological or molecular identification [3,7–10]. In ethnopharmacology, pharmacognosy, and pharmacovigilance there is also strong reliance on vernacular names from products, interviews, pharmacopoeias, and literature for identification of plant material [11,12]. However, vernacular names are prone to ambiguity as one species may have multiple vernacular names or conversely one vernacular name may be applied to many, and often widely divergent, species [13,14]. Vernacular names also vary across geographical regions and languages, resulting in many orthographic variants of the same name [13,15–17]. An example of under-differentiation is the common name “Avishan,” which is used for any of 14 species of *Thymus* L. in different parts of Iran [18]. Likewise, an example of homonymy is the vernacular name “Goosh Fiel,” which is given to the unrelated species *Arctium lappa* L., *Colocasia* spp. and *Caladium* spp. [19]. Vernacular names are important in the preservation of traditional knowledge and recording ethnotaxa, but relying on them as a basis for scientific identification can result in erroneous identification that may invalidate research findings [17,20].

A further potential complication results from the substitution and adulteration of herbal medicinal products. The retailers of herbal products are often not the producers of the products or the harvesters of the raw materials, and this chain of commercialization often involves many middlemen between the harvest of the plant and its final purchase by the consumer. For example, vendors in *Attari* (traditional herbal shops in Iranian bazaars) typically buy dried plant material from medicinal plant collectors, middlemen or wholesalers [14,21]. Vendors are often able to identify herbal products from their morphological characteristics, but lack the ability to identify the living plants [9]. Plant collectors in turn might have a broad knowledge of plant diversity, but might not be able to tell the medicinal species from other similar species in the same genus [22]. Moreover, morphological similarities among some plant species and dried plant parts, scarcity of medicinal species in nature, careless collection practices and a lack of standard identification and control system are all factors contributing to both accidental and intentional substitution [22,23]. The substitution and adulteration of plant ingredients in herbal products can cause health and safety concerns and has public health implications [24,25]. Ize-Ludlow et al. [26] report that the Chinese star anise (*Illicium verum* Hook f.) in herbal teas is substituted with the morphologically similar but toxic Japanese star anise (*I. anisatum* L.), and can result in neurotoxic effects in infants. In Thailand, *Thunbergia laurifolia* Lindl., a common Thai herbal medicine, and *Crotalaria spectabilis* Roth share the same common name [27], but pyrrolizidine alkaloids in *C. spectabilis* cause pulmonary arteritis [28] and acute hepatotoxicity [29]. A method to accurately identify plant materials in traditional herbal medicines is essential in order to guarantee consumer safety, and it is also important for quality control and product authentication [12,25]. Moreover, taxonomic identification of plant samples is crucial in ethnobotanical and ethnopharmacological studies, as species identity connects local classifications to scientific names [13,15,16].

DNA barcoding can provide an accurate and reliable alternative to morphological identification for biological material and is often used when identification using macroscopic or microscopic methods is challenging [30]. It can be used to identify and discriminate species in any developmental or processed stage from which DNA can be extracted [30,31], and from minute amounts of material such as those found in dung [32], pollen [33], degraded herbarium vouchers [34], permafrost preserved subfossils [35], and ancient sediment cores [36,37]. Plant

DNA barcoding has been applied in molecular systematics [38,39], biodiversity inventories [40], wildlife forensics and bio-piracy [41,42], and authentication of herbal products [3,25,43]. Degraded material can include modern material that is no longer fresh or old samples, faeces samples, samples exposed to contamination from DNA of other organisms, samples that have been heated or dried at high temperatures, have been wet, damp or dried improperly, have been processed intensively, or exposed to insect infestations or fungal infection, but also ancient samples such as those listed above.

In DNA barcoding, a short standardized region of genomic DNA from an unidentified sample is queried against a reference sequence database and the identity of the matching sequence is assigned to the sample [30,31,44]. Several genetic regions have been proposed as standard barcodes for land plants, and the ideal barcode needs to be both easily amplifiable and efficiently retrieved from any of the 300,000+ species of plants [44,45]. A single barcoding locus combining these two traits has not been found for plants, and the focus has shifted to a combination of two or more loci to approach a satisfactory level of species discrimination and universality [46]. Most studies now employ a tiered approach, which is based on the use of a common, easily amplified and aligned region such as *rbcl*, *rpoC1*, or *trnL-F* spacer that can act as a scaffold on which to place data from a more variable noncoding region such as *matK*, *trnH-psbA*, nrITS1, nrITS2 or the full ITS1-5.8S-ITS2 (nrITS). The CBOL Plant Working Group recommends the use of *rbcl* and *matK* [47], and together with nrITS these are plant barcoding markers curated by BOLD [48]. Using a tiered approach including at least one of these markers, most species (approx. 75–85%) can be identified, and the subsequent addition of surrogate regions can increase barcoding success to over 90% in some floras [49–52]. The barcoding marker *matK* can be difficult to amplify especially from degraded material [3,9,53,54], but by using target group specific primers amplification success can be greatly increased [55–58].

Sequence similarity [59], tree-based criteria [60], or character-based methods [61,62] can be used for matching unknown query sequences with a reference sequence database for barcode identification. Recent studies show that similarity-based and diagnostic methods that include taking into account non-sequence information significantly outperform tree-based methods [39]. However, the success rate of any method used to assign sequences to a certain taxon is ultimately dependent on the taxonomic coverage of the reference database [3,9].

This study tests the hypothesis that DNA barcode identification by sequence matching can be enhanced by including *a priori* and *a posteriori* data. It employs a two-tiered identification approach using two markers suitable for degraded medicinal plant material, nrITS and *trnL-F* spacer [3,6,9,63,64]. Two methods based on BLAST similarity-based identification, a simple method taking the top hit and an optimized method putting extra weight on the identity value of the query-reference comparison and the deviation of other hits from the top hit, were compared with an integrative identification approach that coalesces sequence matching with *a priori* and *a posteriori* data from other molecular markers, morphology-based identification, pharmacopoeias, scientific literature, vernacular names, and informant identifications. The performances of these respective approaches are evaluated in terms of species level identifications. The hypothesis is tested using a large dataset compiled from medicinal plant samples purchased from herbal vendors in markets throughout northern Iran.

Methods

Market samples

Medicinal plant samples were purchased from 17 traditional herbal shops (*Attari*) in six cities in Northern Khorasan province, Iran, as part of an ethnobotany project on medicinal plants

from local markets (S1 Table). In total 229 medicinal plant samples were collected (S2 Table). Information regarding vernacular names, plant parts, their uses, and the processing and preparation methods were recorded (S3 Table). Samples were sold as leaves, fruit, seeds, flower petals, roots, stems or gums, and were either sold whole, crushed or powdered (S3 Table). All plant samples were purchased as single ingredient products. The plant samples were deposited in the Herbarium of Traditional Medicine and Materia Medica Research Center (HTMRC), Shahid Beheshti University of Medical Sciences, Tehran, Iran and the Herbarium of Uppsala University (UPS), Uppsala, Sweden.

Sample identification

All samples were assigned when possible to family, genus and species based on existing morphological characteristics by professional botanists and pharmacognocists trained in morphological and micro-morphological identification of medicinal plants identification using available scientific literature [65–68] (S2 Table). Each sample was tentatively identified to species based on matching vernacular names from herbal pharmacopoeia and reference literature [65–68] (S3 Table). Sixty-eight out of 229 samples were not identifiable beyond genus level, and were subsequently selected for identification through DNA barcoding. Nomenclature of plant names follows The Plant List [69].

DNA extraction, amplification and sequencing

Total genomic DNA of 68 market samples was extracted using a CTAB protocol [70]. The extracted DNA was purified using the GE Illustra GFX™ PCR DNA and Gel Band Purification kit following the manufacturer's protocol (GE Healthcare, Little Chalfont, United Kingdom). Two markers were amplified and sequenced for DNA barcoding, the nuclear ribosomal internal transcribed spacer (nrITS), and a plastid marker (*trnL-F* spacer). nrITS was amplified using primers ITS5 (5' - GGAAGTAAAAGTCGTAACAAGG - 3') and ITS4 (5' - TCCTCCGCT TATTGATATGC - 3') [71] and *trnL-F* spacer using primers *trnL_c2* (5' - GGATAGGTGCAGA GACTCAAT - 3') [72] and *trnL_f* (5' - ATTTGAAGTGGTGACACGAG - 3') [73]. PCR amplification was performed in 50 µl reactions containing 5 µl reaction buffer IV (Qiagen NV, Venlo, Netherlands) (10x), 5 µl MgCl₂ (25mM), 1 µl dNTP (10 µM), 0.25 µl Taq-polymerase (Qiagen NV, Venlo, Netherlands) (5 U/µl), 0.5 µl BSA, 1 µl of each primer (10 mM) and 1 µl of template DNA. The PCR protocol for nrITS was an initial 3 min of denaturation at 95°C, followed by 35 cycles of 20 sec of denaturation at 95°C, 1 min of annealing at 55°C and 2 min of elongation at 72°C, and a final elongation of 10 min at 72°C. For *trnL-F*, the PCR protocol started with an initial 3 min denaturation at 95°C, followed by 35 cycles of 15 sec denaturation at 95°C, 50 sec of annealing at 55°C, and 4 min of elongation at 72°C, and a final elongation of 8 min at 72°C. Sequencing was performed by Macrogen Europe Inc. (Amsterdam, the Netherlands) on an ABI3730XL automated sequencer (Applied Biosystems, Waltham, Massachusetts, USA). Primers used for PCR amplification were also used for sequencing reactions. Sequence trace files were assembled using Pregap4 and Gap4 in the Staden Package [74]. All plant sequences were submitted to BOLD and linked to NCBI GenBank (S3 Table).

DNA barcode identification

Three approaches were used for DNA barcode identification, two sequence similarity-based methods using BLAST [59] and an integrative method that coalesces similarity-based results with *a priori* and *a posteriori* data. The two methods based on BLAST similarity-based identification were a simple method taking the top hit and an optimized method putting extra weight on the identity value of the query-reference comparison. For both methods sequences were

sequentially queried using megablast [59] online at NCBI nucleotide BLAST against the nucleotide database. For the simple method all top hits within 10 points deviation down of the max score were considered: if the max score (-10 points) included only a single species then a species level identification was assigned; if the max score (-10 points) included multiple species in the same genus then a genus level identification was assigned; and if the max score (-10 points) included multiple species in different genera in the same family then a family level identification was assigned. However, the length of the query coverage and the identity between the query and the reference sequence influences the max score in BLAST, and hits with low identity but high query coverage can have higher max scores than hits with high identity but low coverage. For the optimized method a similarity score was calculated for up to 100 BLAST hits if the query cover was 70% or higher: $\text{max score} * (\text{query cover} / \text{identity})$. Subsequently all hits were ordered by this score, and the deviation for each similarity score value from the highest similarity score was calculated (S4 Table). Identifications were assigned based on a combination of the identity score (High identity: $i \geq 95\%$; Medium identity: $90\% \leq i < 95\%$; Low identity: $i < 90\%$) and the number of species within 1% deviation of the calculated similarity score. High identity and one species within 1% deviation was assigned species-level confidence; high identity and more than one species was assigned genus-level confidence; medium identity and one or more species within the same genus was assigned genus-level confidence; medium identity and species from more than one genus was assigned family-level confidence; and low identity was assigned family-level confidence (S5 Table).

Integrative approach for identification

The integrative approach coalesced the optimized BLAST-based sequence matching results with *a priori* data from morphological characteristics of the material, interview data on vernacular names and studies of literature and pharmacopeias for these names, along with *a posteriori* data from multiple molecular markers and data on traditional use, occurrence, and distribution of putative species in the study area. For example, *a priori* data suggests that the inflorescences Kh111 are *Amaranthus caudatus* L. based on literature, *Amaranthus* sp. based on ethnobotanical interview data and *Amaranthus* sp. based on morphology (lacking spiny hairs). The BLAST hits suggest that the sequence matches with either *Amaranthus hybridus* L. (based on ITS) or *Amaranthus spinosus* L. (based on *trnL*-F spacer). *A posteriori* data gives two additional clues that aid the identification process: 1) Consulting the Flora of Iran and other literature shows us that *Amaranthus spinosus* L. does not occur in Iran; and 2) *Amaranthus spinosus* L. has spiny hairs on the inflorescence which are absent in this sample. As a result, using an integrative approach the identification of this sample would be *Amaranthus hybridus* L. For a full example of this process see S1 Text.

Results

Sequencing success and BLAST matching

The amplification success of market samples for nrITS was 96% (65 samples). However, 17 samples (25%) yielded sequences of fungal DNA due to contamination of the original market samples. After exclusion of these contaminated samples the sequencing success rate for nrITS was 71% (48). The amplification success of market samples for the *trnL*-F spacer was also 96% (65 samples). For the *trnL*-F spacer five products (7%) failed to yield usable sequences, and thus the sequencing success rate for the *trnL*-F spacer was 88% (60 samples). Out of the 68 samples, there were 40 with both nrITS and *trnL*-F spacer sequences, 20 with only the *trnL*-F spacer and 8 with only nrITS.

The simple and optimized BLAST results based on sequence matching as well as the putative species identification for each of the 68 tested samples are included in [S3 Table](#). The identification success was dependent on the marker and availability of reference sequences in GenBank. For some putative species, reference sequences in GenBank were available for only one of the two markers. The BLAST sequence matching method included 60 *trnL*-F spacer and 48 nrITS query sequences. The simple *trnL*-F spacer BLAST search results identified 18% (11 samples) to species level, 53% (32) to genus level, and 28% (17) to family level. The optimized *trnL*-F spacer BLAST search results identified 33% (20 samples) to species level, 45% (27) to genus level, and 16% (10) to family level and 5% (3) could not be identified. The simple nrITS spacer BLAST search results identified 35% (17 samples) to species level, 58% (28) to genus level, and 6% (3) to family level. The optimized nrITS spacer BLAST search results identified 37% (18 samples) to species level, 50% (24) to genus level, and 6% (3) to family level and 6% (3) could not be identified. Combined data from both markers using the simple BLAST sequence matching method identified 32% (22 samples) to species level, 47% (32) to genus level, and 21% (14) to family level. Combined data from both markers using the optimized methods identified 38% (26 samples) to species level, 40% (27) to genus level, and 19% (13) to family level and 3% (2) could not be identified.

The integrative approach coalesces sequence matching results with *a priori* and *a posteriori* data. This approach resulted in the identification of 65% (39 samples) to species level, 27% (16) to genus level, and 8% (5) to family level for the *trnL*-F spacer. For the second marker, nrITS, the integrative approach resulted in the identification of all samples to either species or genus level, with 62% (30 samples) identified to species level, and 38% (18) identified to genus level. Combining data from both markers resulted in 77% (52 samples) species level identification and 23% (16) genus level identification. [Fig 1](#) shows the results for the two sequence matching approaches: simple and optimized sequence matching; and the integrative approach where *a priori* and *a posteriori* information is incorporated in the identification process. Evaluating both methods, the integrative approach gives a 1.67, 1.95 and 2.00 fold increase in species level identification rates for nrITS, the *trnL*-F spacer, and combined markers respectively ([Table 1](#)).

Discussion

DNA barcode identification, morphology and herbal pharmacopeia

Fifty-eight of the 68 samples that could not be identified to species-level based on morphology, were identified to genus and another ten to family only. Applying the integrative approach outlined here, 52 (76%) of these samples could be assigned a species-level identification ([Table 1](#), [Fig 1](#) and [S3 Table](#)). Combining this integrative approach with the CBOL PWG recommended barcoding markers *rbcL* and *matK* [47] in addition to nrITS could have further increased the identification rate. In this study we choose not to include these as *rbcL* has low sequence variation [6,63] and *matK* has low primer universality [3,53,54]. Primer universality is important when identifying unknown species such as biodiversity samples or market products, but could be mitigated using a tiered amplification approach with multiple primers sets. Different studies have reported different species identification rates for the combination of *rbcL* and *matK*, but the most comprehensive study to date reports an identification rate of as low as 49.7% [6]. Two approaches that are relevant for the type of degraded and contaminated material included in this study that could have improved amplification success and limited amplification of fungal contaminants would have been targeting the shorter nrITS2 fragment of nrITS and the use of novel plant specific primers such as those published by Cheng et al. [64].

The identification results from DNA barcoding compared with putative species names derived from herbal pharmacopeia showed inconsistency. Agreement in identification at the

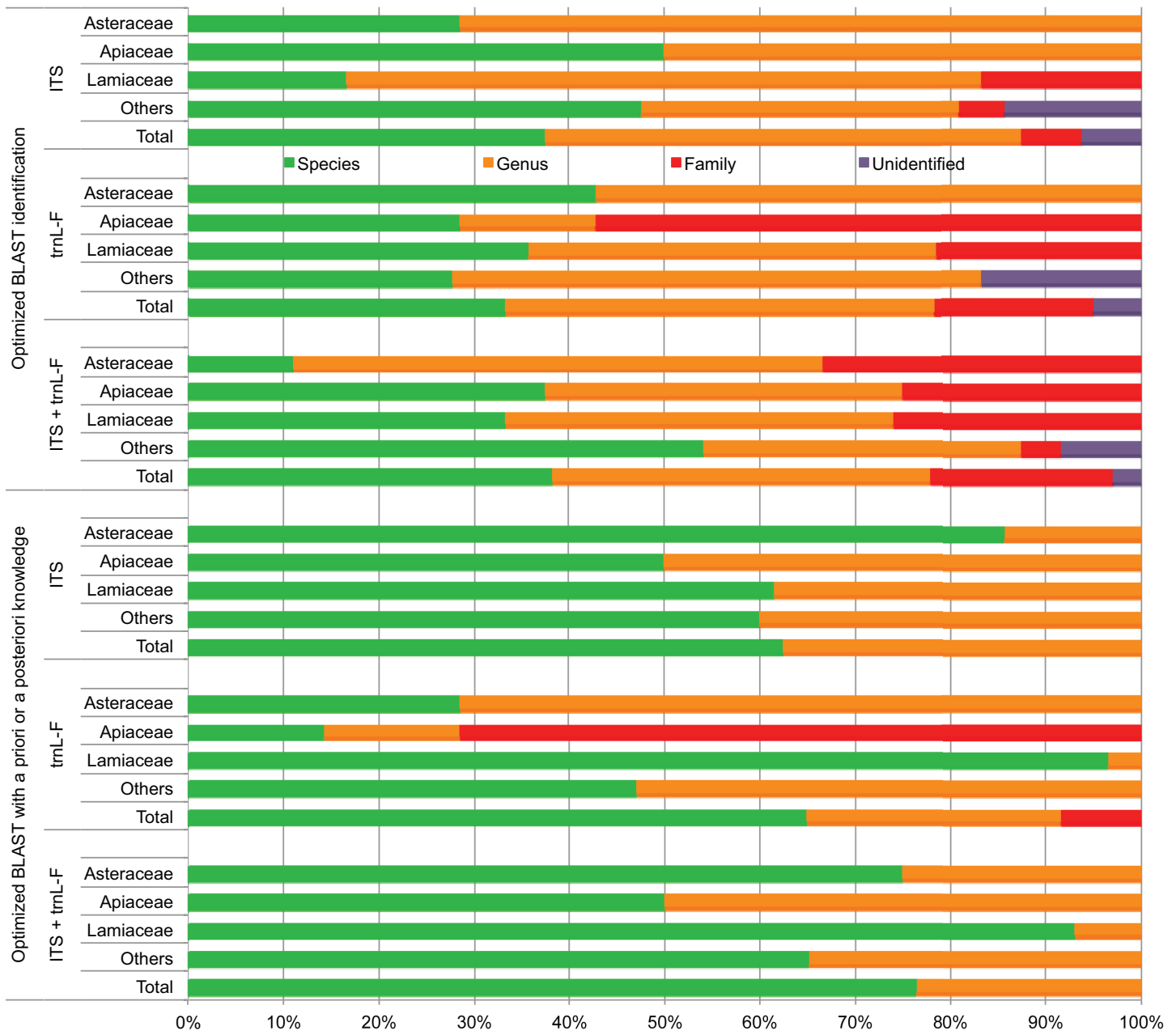


Fig 1. Comparison of identification success rates among different plant families 1) when relying on sequence matching alone, and 2) when *a priori* and *a posteriori* information is incorporated in the identification process for each of the two markers separately and when they are combined.

<https://doi.org/10.1371/journal.pone.0175722.g001>

species level between the herbal pharmacopoeia and the integrative barcoding approach was found in 71% of cases (48 samples), whereas in 24% (18 samples) identification using herbal pharmacopoeia resulted in erroneous identifications (S6 Table). It should be noted however, that matches were higher at genus and family level, respectively 72% and 89%. It is unlikely that these mismatches are due to incomplete sequence reference libraries as accessions of most species and all genera from the herbal pharmacopoeia assignments were present in NCBI GenBank. These findings imply that ethnopharmacological studies should be careful about relying on herbal pharmacopoeia as this may result in incorrect identifications. Accurate species

Table 1. Comparison of species level identification rates for optimized BLAST similarity-based and those using the outlined integrative approach.

Sequencing success (out of 68)		Species level identification					
		Simple BLAST		Optimized BLAST		Integrative approach	
Marker	No.	Absolute	Relative	Absolute	Relative	Absolute	Relative
<i>nrITS</i>	48	17	35%	18	38%	30	63%
<i>trnL-F</i>	60	11	18%	20	33%	39	65%
<i>trnL-F + ITS</i>	68	22	32%	26	38%	52	76%
Sequencing success (out of 68)		Comparing approaches					
		Integrative approach improvement from simple BLAST			Integrative approach improvement from optimized BLAST		
Marker	No.	Absolute	Relative	Increase	Absolute	Relative	Increase
<i>nrITS</i>	48	13	27%	1.76	12	25%	1.67
<i>trnL-F</i>	60	28	47%	3.55	19	32%	1.95
<i>trnL-F + ITS</i>	68	30	44%	2.36	26	38%	2.00

<https://doi.org/10.1371/journal.pone.0175722.t001>

identification of samples helps ethnopharmacologists and ethnobotanists to relate ethnobiological information about species to scientific literature for further research. Most studies on molecular identification of herbal products focus on specific families and genera because building a comprehensive sequence reference database is more feasible for defined groups than for entire families or all plants [12,75,76]. However, when dealing with completely unknown samples from a broad range of taxa, incomplete reference database coverage can be problematic. In such cases the best way to overcome this problem is an integrative approach in which samples are identified using a total evidence approach that includes multi-marker DNA barcoding, morphology, distribution, pharmacopoeias, and literature. The results from molecular identification of samples have a low identification rate if only BLAST is used and no other evidence is incorporated. However, this can be augmented using an integrative approach (Table 1 and Fig 1). Fig 2 outlines the integrative approach, as used in this study, such that it can be adopted for similar investigations that aim to identify samples of unknown taxonomic identity or authentication of herbal products by pharmaceutical manufacturers or pharmaceutical quality control agencies (see S1 Text for an example). Combining the three identification methods used in this study (sequence similarity matching, morphological classification, and ethnoclassification matching) in an integrative way enables single species assignments for 52 out of 68 samples, whereas unambiguous species assignments are only possible in 22–26 samples using similarity matching (Table 1). It must be noted that these identifications cannot be verified independently. However the objective here was to see whether a combination of similarity matching and *a priori* and *a posteriori* data could reduce ambiguity and enable limiting the number of putative species to one and thus a species level identification.

Substitution and adulteration

The total evidence results revealed that 26% (18 samples) of the medicinal plants sampled from the markets did not match the intended species recorded in the herbal pharmacopeia [68] (S6 Table). These results could suggest that the species intended for medicinal use, as identified by the herbal pharmacopeia, are substituted locally for other species within the same genus or for different species altogether. Alternatively, the traditional products may consist of different species than those mentioned in the pharmacopeia. Similar research in Morocco also showed a high level of substitution and suggested that, in addition to the options above,



Fig 2. The strengths of DNA barcoding outweigh its weaknesses, but an integrative approach is necessary to optimize identification of unknown plant material.

<https://doi.org/10.1371/journal.pone.0175722.g002>

gradual substitution over time of one species for another could explain the discrepancy [9,22]. Some examples of cases where DNA barcoding showed samples to be completely different from the putative species in this study are “Marzeh” and “Maryam goli.” The vernacular name “Marzeh” refers to either *Satureja laxiflora* C.Koch or *S. hortensis* L. (Lamiaceae) according to the literature, but molecular identification showed samples with this name to be *Urtica dioica* L. (Urticaceae), an unrelated species from a different family. “Maryam goli” refers to either *Salvia sclarea* L. or *S. officinalis* L. (Lamiaceae) based on herbal pharmacopeias [67,68], but molecular identification showed the sample analyzed under this name to be *Althaea cannabina* L. (Malvaceae).

In this study, we applied DNA barcoding only to samples lacking morphological characteristics for identification. One can assume that identification of these plant products is similarly challenging for traders and consumers. Quite a few samples (6%) belonged to families other than those expected based on the putative identifications, and many more (27%) belonged to different genera than those expected, within the same family. S6 Table lists samples where family and genus identifications using the integrative molecular identification approach did not match putative species identifications from the official herbal pharmacopeia. A list of all samples and identifications are given in S3 Table.

Conclusions

The present study shows that molecular identification through DNA barcoding is a very useful tool for the identification of traded medicinal plant products originating from a wide range of taxonomic groups. The identification rates of samples that were unidentifiable by morphology alone show the importance of having a complete sequence reference databases for DNA barcoding. Species assignments using DNA barcoding are limited by the comprehensiveness of the sequence reference database, and at the moment no such complete database exists. It is essential to support efforts by iBOL [International Barcode of Life (<http://www.ibol.org>)], CBOL [Consortium for the Barcode of Life (<http://www.barcodeoflife.org>)], GBIF [Global Biodiversity Information Facility (<http://www.gbif.org>)] and NCBI GenBank [National Center for Biotechnology Information (<http://www.ncbi.nlm.nih.gov>)] to expand barcode reference databases. Identification of plant material does not necessarily need to rely on sequence matching alone, but can also take into account *a priori* data from morphological characteristics of the material, interview data on vernacular names and traditional knowledge, studies of old and current literature, and pharmacopeias for these names, and *a posteriori* data from multiple molecular markers and data on traditional use, flowering time, occurrence and distribution of putative species. The evaluation of the sequence matching approaches shows that DNA barcode identification rates can be enhanced by relying on an integrative approach that combines *a priori* and *a posteriori* data. Although these identifications cannot be verified independently, the method shows that a combination of similarity matching and *a priori* and *a posteriori* data reduces ambiguity and make it possible to assign a single species identification to an increased number of samples. These results do not advocate a novel method of identification, but rather highlight the risk of using automated identification based on sequence similarity for identification of unknown material. As DNA barcoding becomes more and more mainstream, researchers and non-academic professionals without a strong background in the studied organism group, rely on automated identifications without taking other evidence into account. Other resources and data for proper identification require expertise in taxonomy and manual intervention in the identification process. This means that a person should have a good overview of the plant group that they are working on for samples under investigation.

Supporting information

S1 Table. List of herbal shops and their locations visited for this study in Iran.

(PDF)

S2 Table. Medicinal plant samples from Northern Khorasan province, Iran.

(PDF)

S3 Table. List of samples, vernacular names, putative species identification, Genbank accession numbers, simple and optimized BLAST similarity identifications and final identifications based on the integrative approach.

(PDF)

S4 Table. Optimized BLAST similarity identifications. BLAST hits sorted by max score* (identity/cover). Colored by deviation (d) from highest hit: $d = < 1\%$: green, $1\% < d = < 2\%$: orange, $2\% < d = < 3\%$: red, $3\% < d$: no color.

(PDF)

S5 Table. Optimized BLAST similarity identifications per sample. Identifications are assigned based on a combination of the identity score (High identity: $i \geq 95\%$; Medium identity: $90\% \leq i < 95\%$; Low identity: $i < 90\%$) and the number of species within 1% deviation of the calculated similarity score.

(PDF)

S6 Table. Samples for which family and genus identifications based on DNA barcoding results did not match putative species identifications from the official herbal pharmacopeia.

(PDF)

S1 Text. An example illustrating the identification process using the integrative approach coalescing *a priori* and *a posteriori* data.

(PDF)

Acknowledgments

The authors are grateful to all *Attars* who participated in this study and shared their knowledge on medicinal plants of Khorasan province with the researchers. Saeideh Ghafari is acknowledged for her help in extracting putative species names from herbal pharmacopoeias, and Vincent Manzanilla for the graphic design and production of [Fig 2](#).

Author Contributions

Conceptualization: AG HdB.

Data curation: AG.

Formal analysis: AG HdB.

Funding acquisition: AG HdB.

Investigation: AG YS.

Methodology: AG HdB.

Project administration: AG HdB.

Resources: AG YS.

Supervision: HdB.

Visualization: HdB.

Writing – original draft: AG HdB.

Writing – review & editing: AG HdB.

References

1. Cunningham AB. Applied ethnobotany: people, wild plant use and conservation. London, UK: Earthscan Publications Ltd; 2001.
2. Mati E, de Boer HJ. Ethnobotany and trade of medicinal plants in the Qaysari Market, Kurdish Autonomous Region, Iraq. *J Ethnopharmacol*. 2011; 133: 490–510. <https://doi.org/10.1016/j.jep.2010.10.023> PMID: 20965241
3. Kool A, de Boer HJ, Krüger Å, Rydberg A, Abbad A, Björk L, et al. Molecular identification of commercialized medicinal plants in Southern Morocco. *PLOS ONE*. 2012; 7: e39459. <https://doi.org/10.1371/journal.pone.0039459> PMID: 22761800
4. Chen S, Pang X, Song J, Shi L, Yao H, Han J, et al. A renaissance in herbal medicine identification: from morphology to DNA. *Biotechnol Adv*. 2014; 32: 1237–1244. <https://doi.org/10.1016/j.biotechadv.2014.07.004> PMID: 25087935
5. Schlick-Steiner BC, Steiner FM, Seifert B, Stauffer C, Christian E, Crozier RH. Integrative taxonomy: a multisource approach to exploring biodiversity. *Annu Rev Entomol*. 2010; 55: 421–438. <https://doi.org/10.1146/annurev-ento-112408-085432> PMID: 19737081
6. Li DZ, Gao LM, Li HT, Wang H, Ge XJ, Liu JQ, et al. Comparative analysis of a large dataset indicates that internal transcribed spacer (ITS) should be incorporated into the core barcode for seed plants. *Proc Natl Acad Sci*. 2011; 108: 19641–19646. <https://doi.org/10.1073/pnas.1104551108> PMID: 22100737
7. Williams VL, Balkwill K, Witkowski ET. Unraveling the commercial market for medicinal plants and plant parts on the Witwatersrand, South Africa. *Econ Bot*. 2000; 54: 310–327.
8. Bussmann RW, Sharon D, Vandebroek I, Jones A, Revene Z. Health for sale: the medicinal plant markets in Trujillo and Chiclayo, Northern Peru. *J Ethnobiol Ethnomedicine*. 2007; 3: 37.
9. De Boer HJ, Ouarghidi A, Martin G, Abbad A, Kool A. DNA barcoding reveals limited accuracy of identifications based on folk taxonomy. *PLOS ONE*. 2014; 9: e84291. <https://doi.org/10.1371/journal.pone.0084291> PMID: 24416210
10. Towns AM, Quiroz D, Guinee L, de Boer HJ, van Andel T. Volume, value and floristic diversity of Gabon's medicinal plant markets. *J Ethnopharmacol*. 2014; 155: 1184–1193. <https://doi.org/10.1016/j.jep.2014.06.052> PMID: 24995835
11. Farah MH, Olsson S, Bate J, Lindquist M, Edwards R, Simmonds MS, et al. Botanical nomenclature in pharmacovigilance and a recommendation for standardisation. *Drug Saf*. 2006; 29: 1023–1029. PMID: 17061908
12. De Boer HJ, Ichim MC, Newmaster SG. DNA Barcoding and Pharmacovigilance of Herbal Medicines. *Drug Saf*. 2015; 38: 611–620. <https://doi.org/10.1007/s40264-015-0306-8> PMID: 26076652
13. Berlin B. *Ethnobiological classification: Principles of categorization of plants and animals in traditional societies*. Princeton, NJ: Princeton University Press; 1992.
14. Amiri MS, Joharchi MR. Ethnobotanical investigation of traditional medicinal plants commercialized in the markets of Mashhad, Iran. *Avicenna J Phytomedicine*. 2013; 3: 254–271.
15. Linares E, Bye RA. A study of four medicinal plant complexes of Mexico and adjacent United States. *J Ethnopharmacol*. 1987; 19: 153–183. PMID: 3613608
16. Otieno J, Abihudi S, Veldman S, Nahashon M, van Andel T, de Boer HJ. Vernacular dominance in folk taxonomy: A case study of ethnospecies in medicinal plant trade in Tanzania. *J Ethnobiol Ethnomedicine*. 2015; 11.
17. Bennett BC, Balick MJ. Does the name really matter? The importance of botanical nomenclature and plant taxonomy in biomedical research. *J Ethnopharmacol*. 2014; 152: 387–392. <https://doi.org/10.1016/j.jep.2013.11.042> PMID: 24321863
18. Naghibi F, Mosaddegh M, Mohammadi Motamed M, Ghorbani A. Labiatae family in folk medicine in Iran: from ethnobotany to pharmacology. *Iran J Pharm Res*. 2005; 4: 63–79.
19. Aynehchi Y. *Pharmacognosy and medicinal plants of Iran*. Tehran Univ. 1986;

20. Rivera D, Allkin R, Obón C, Alcaraz F, Verpoorte R, Heinrich M. What is in a name? The need for accurate scientific nomenclature for plants. *J Ethnopharmacol*. 2014; 152: 393–402. <https://doi.org/10.1016/j.jep.2013.12.022> PMID: 24374235
21. Moradi-Lake M, Saeidi M, Naserbakht M. Knowledge of medicinal plants properties among Attars of Tehran city. *Q J Payesh*. 2009; 7: 321–328.
22. Ouarghidi A, Powell B, Martin GJ, de Boer HJ, Abbad A. Species substitution in medicinal roots and possible implications for toxicity in Morocco. *Econ Bot*. 2012; 66: 370–382.
23. Joharchi MR, Amiri MS. Taxonomic evaluation of misidentification of crude herbal drugs marketed in Iran. *Avicenna J Phytomedicine*. 2012; 2: 105.
24. Posadzki P, Watson L, Ernst E. Contamination and adulteration of herbal medicinal products (HMPs): an overview of systematic reviews. *Eur J Clin Pharmacol*. 2013; 69: 295–307. <https://doi.org/10.1007/s00228-012-1353-z> PMID: 22843016
25. Newmaster SG, Grguric M, Shanmughanandhan D, Ramalingam S, Ragupathy S. DNA barcoding detects contamination and substitution in North American herbal products. *BMC Med*. 2013; 11: 222. <https://doi.org/10.1186/1741-7015-11-222> PMID: 24120035
26. Ize-Ludlow D, Ragone S, Bruck IS, Bernstein JN, Duchowny M, Pena BMG. Neurotoxicities in infants seen with the consumption of star anise tea. *Pediatrics*. 2004; 114: e653. <https://doi.org/10.1542/peds.2004-0058> PMID: 15492355
27. Suwanchaikasem P, Chaichantipyuth C, Amnuoypol S, Sukrong S. Random amplified polymorphic DNA analysis of *Thunbergia laurifolia* Lindl. and its related species. *J Med Plants Res*. 2012; 6: 2955–2961.
28. Meyrick B, Reid L. Development of pulmonary arterial changes in rats fed *Crotalaria spectabilis*. *Am J Pathol*. 1979; 94: 37. PMID: 153714
29. Smith LW, Culvenor CCJ. Plant sources of hepatotoxic pyrrolizidine alkaloids. *J Nat Prod*. 1981; 44: 129–152. PMID: 7017073
30. Hebert PDN, Cywinska A, Ball S, de Waard J. Biological identifications through DNA barcodes. *Proc R Soc B*. 2003; 270: 313–322. <https://doi.org/10.1098/rspb.2002.2218> PMID: 12614582
31. Hajibabaei M, Singer GA, Hebert PD, Hickey DA. DNA barcoding: how it complements taxonomy, molecular phylogenetics and population genetics. *Trends Genet*. 2007; 23: 167–172. <https://doi.org/10.1016/j.tig.2007.02.001> PMID: 17316886
32. Hibert F, Taberlet P, Chave J, Scotti-Saintagne C, Sabatier D, Richard-Hansen C. Unveiling the diet of elusive rainforest herbivores in next generation sequencing era? The tapir as a case study. *PLOS ONE*. 2013; 8: e60799. <https://doi.org/10.1371/journal.pone.0060799> PMID: 23560107
33. Richardson RT, Lin C-H, Sponsler DB, Quijia JO, Goodell K, Johnson RM. Application of ITS2 metabarcoding to determine the provenance of pollen collected by honey bees in an agroecosystem. *Appl Plant Sci*. 2015; 3: apps.1400066.
34. Särkinen T, Staats M, Richardson JE, Cowan RS, Bakker FT. How to open the treasure chest? Optimising DNA extraction from herbarium specimens. *PLOS ONE*. 2012; 7: e43808. <https://doi.org/10.1371/journal.pone.0043808> PMID: 22952770
35. Van Geel B, Aptroot A, Baittinger C, Birks HH, Bull ID, Cross HB, et al. The ecological implications of a Yakutian mammoth's last meal. *Quat Res*. 2008; 69: 361–376.
36. Haile J, Holdaway R, Oliver K, Bunce M, Gilbert MTP, Nielsen R, et al. Ancient DNA chronology within sediment deposits: are paleobiological reconstructions possible and is DNA leaching a factor? *Mol Biol Evol*. 2007; 24: 982–989. <https://doi.org/10.1093/molbev/msm016> PMID: 17255121
37. Bellemain E, Davey M, Kauserud H, Epp L, Boessenkool S, Coissac E, et al. Fungal palaeodiversity revealed using high-throughput metabarcoding from arctic permafrost. *Environ Microbiol*. 2013; 15: 1176–1189. <https://doi.org/10.1111/1462-2920.12020> PMID: 23171292
38. Liu JIE, Moeller M, Gao LM, Zhang D., Li DZ. DNA barcoding for the discrimination of Eurasian yews (*Taxus* L., Taxaceae) and the discovery of cryptic species. *Mol Ecol Resour*. 2011; 11: 89–100. <https://doi.org/10.1111/j.1755-0998.2010.02907.x> PMID: 21429104
39. Van Velzen R, Weitschek E, Felici G, Bakker FT. DNA barcoding of recently diverged species: relative performance of matching methods. *PLOS ONE*. 2012; 7: e30490. <https://doi.org/10.1371/journal.pone.0030490> PMID: 22272356
40. Thompson KA, Newmaster SG. Molecular taxonomic tools provide more accurate estimates of species richness at less cost than traditional morphology-based taxonomic practices in a vegetation survey. *Biodivers Conserv*. 2014; 23: 1411–1424.
41. Deguilloux M-F, Pemonge M-H, Petit RJ. Novel perspectives in wood certification and forensics: dry wood as a source of DNA. *Proc R Soc B*. 2002; 269: 1039–46. <https://doi.org/10.1098/rspb.2002.1982> PMID: 12028761

42. Baker CS, Steel D, Choi Y, Lee H, Kim KS, Choi SK, et al. Genetic evidence of illegal trade in protected whales links Japan with the US and South Korea. *Biol Lett*. 2010; 6: 647–650. <https://doi.org/10.1098/rsbl.2010.0239> PMID: 20392716
43. Coghlan M, Haile J, Houston J, Murray D, White N, Moolhuijzen P, et al. Deep sequencing of plant and animal DNA contained within traditional chinese medicines reveals legality issues and health safety concerns. *PLOS Genet*. 2012; 8: e1002657. <https://doi.org/10.1371/journal.pgen.1002657> PMID: 22511890
44. Kress WJ, Wurdack KJ, Zimmer EA, Weigt LA, Janzen DH. Use of DNA barcodes to identify flowering plants. *Proc Natl Acad Sci*. 2005; 102: 8369–8374. <https://doi.org/10.1073/pnas.0503123102> PMID: 15928076
45. Fazekas AJ, Burgess KS, Kesanakurti PR, Graham SW, Newmaster SG, Husband BC, et al. Multiple multilocus DNA barcodes from the plastid genome discriminate plant species equally well. *PLOS ONE*. 2008; 3: e2802. <https://doi.org/10.1371/journal.pone.0002802> PMID: 18665273
46. Kress WJ, Erickson DL. 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. <https://doi.org/10.1371/journal.pone.0000508> PMID: 17551588
47. CBOL Plant Working Group. A DNA barcode for land plants. *Proc Natl Acad Sci*. 2009; 106: 12794–12797. <https://doi.org/10.1073/pnas.0905845106> PMID: 19666622
48. Ratnasingham S, Hebert PD. BOLD: The Barcode of Life Data System (<http://www.barcodinglife.org>). *Mol Ecol Notes*. 2007; 7: 355–364. <https://doi.org/10.1111/j.1471-8286.2007.01678.x> PMID: 18784790
49. Ebihara A, Nitta JH, Ito M. Molecular species identification with rich floristic sampling: DNA barcoding the pteridophyte flora of Japan. *PLOS ONE*. 2010; 5: e15136. <https://doi.org/10.1371/journal.pone.0015136> PMID: 21170336
50. Burgess KS, Fazekas AJ, Kesanakurti PR, Graham SW, Husband BC, Newmaster SG, et al. Discriminating plant species in a local temperate flora using the *rbcL*+*matK* DNA barcode. *Methods Ecol Evol*. 2011; 2: 333–340.
51. De Vere N, Rich TC, Ford CR, Trinder SA, Long C, Moore CW, et al. DNA barcoding the native flowering plants and conifers of Wales. *PLOS ONE*. 2012; 7: e37945. <https://doi.org/10.1371/journal.pone.0037945> PMID: 22701588
52. Liu J, Yan H-F, Newmaster SG, Pei N, Ragupathy S, Ge X-J. The use of DNA barcoding as a tool for the conservation biogeography of subtropical forests in China. *Divers Distrib*. 2015; 21: 188–199.
53. Sass C, Little DP, Stevenson DW, Specht CD. DNA barcoding in the cycadales: testing the potential of proposed barcoding markers for species identification of cycads. *PLOS ONE*. 2007; 2: e1154. <https://doi.org/10.1371/journal.pone.0001154> PMID: 17987130
54. Piredda R, Simeone MC, Attimonelli M, Bellarosa R, Schirone B. Prospects of barcoding the Italian wild dendroflora: oaks reveal severe limitations to tracking species identity. *Mol Ecol Resour*. 2011; 11: 72–83. <https://doi.org/10.1111/j.1755-0998.2010.02900.x> PMID: 21429102
55. Palhares RM, Drummond MG, Brasil B dos SAF, Cosenza GP, Brandão M das GL, Oliveira G. Medicinal plants recommended by the world health organization: DNA barcode identification associated with chemical analyses guarantees their quality. *PLOS ONE*. 2015; 10: e0127866. <https://doi.org/10.1371/journal.pone.0127866> PMID: 25978064
56. Wallace LJ, Boilard SM, Eagle SH, Spall JL, Shokralla S, Hajibabaei M. DNA barcodes for everyday life: Routine authentication of Natural Health Products. *Food Res Int*. 2012; 49: 446–452.
57. Mahadani P, Ghosh SK. DNA Barcoding: A tool for species identification from herbal juices. *DNA Barcodes*. 2013; 1: 35–38.
58. Purushothaman N, Newmaster SG, Ragupathy S, Stalin N, Suresh D, Arunraj DR, et al. A tiered barcode authentication tool to differentiate medicinal *Cassia* species in India. *Genet Mol Res*. 2014; 13: 2959–2968. <https://doi.org/10.4238/2014.April.16.4> PMID: 24782130
59. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol*. 1990; 215: 403–410. [https://doi.org/10.1016/S0022-2836\(05\)80360-2](https://doi.org/10.1016/S0022-2836(05)80360-2) PMID: 2231712
60. Munch K, Boomsma W, Willerslev E, Nielsen R. Fast phylogenetic DNA barcoding. *Philos Trans R Soc B Biol Sci*. 2008; 363: 3997–4002.
61. DasGupta B, Konwar KM, Muandou II, Shvartsman AA. DNA-BAR: distinguisher selection for DNA barcoding. *Bioinformatics*. 2005; 21: 3424–3426. <https://doi.org/10.1093/bioinformatics/bti547> PMID: 15961439
62. Rach J, DeSalle R, Sarkar IN, Schierwater B, Hadrys H. Character-based DNA barcoding allows discrimination of genera, species and populations in Odonata. *Proc R Soc B Biol Sci*. 2008; 275: 237–247.
63. Chen S, Yao H, Han J, Liu C, Song J, Shi L, et al. Validation of the ITS2 region as a novel DNA barcode for identifying medicinal plant species. *PLOS ONE*. 2010; 5: 1–8.

64. Cheng T, Xu C, Lei L, Li C, Zhang Y, Zhou S. Barcoding the kingdom Plantae: new PCR primers for ITS regions of plants with improved universality and specificity. *Mol Ecol Resour.* 2016; 16: 138–149. <https://doi.org/10.1111/1755-0998.12438> PMID: 26084789
65. Rechinger KH, editor. *Flora Iranica*. Published volumes. Vienna, Austria: Naturhistorisches Museum; 1963.
66. Amin GR. *Popular Medicinal Plants of Iran: 1*. Iranian Research Institute of Medicinal Plants; 1991.
67. Mozafarian V. *A dictionary of Iranian plant names (Latin-English-Persian)* Farhang Moaser Publication. Tehran; 1996.
68. Editorial committee for Iranian herbal pharmacopoeia. *Iranian herbal pharmacopoeia*. Tehran, IR Iran: Ministry of Health and Medical Education; 2002.
69. The Plant List. Version 1.1. [Internet]. 2013. Available: <http://www.theplantlist.org/>
70. Doyle JJ, Doyle JL. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochem Bull.* 1987; 19: 11–15.
71. White TJ, Bruns TD, Lee S, Taylor J. Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics. In: Innis MA, Gelfand DH, Shinsky JJ, White TJ, editors. *PCR Protocols: A Guide to Methods and Applications*. San Diego: Academic Press; 1990. pp. 315–322.
72. Bellstedt DU, Linder HP, Harley EH. Phylogenetic relationships in *Disa* based on non-coding trnL-trnF chloroplast sequences: evidence of numerous repeat regions. *Am J Bot.* 2001; 88: 2088–2100. PMID: 21669640
73. Taberlet P, Gielly L, Pautou G, Bouvet J. Universal primers for amplification of three non-coding regions of chloroplast DNA. *Plant Mol Biol.* 1991; 17: 1105–1109. PMID: 1932684
74. Staden R. The Staden sequence analysis package. *Mol Biotechnol.* 1996; 5: 233–241. PMID: 8837029
75. Baker DA. DNA barcode identification of black cohosh herbal dietary supplements. *J AOAC Int.* 2012; 95: 1023–1034. PMID: 22970567
76. Xu S, Li D, Li J, Xiang X, Jin W, Huang W, et al. Evaluation of the DNA Barcodes in *Dendrobium* (Orchidaceae) from Mainland Asia. *PLOS ONE.* 2015; 10: e0115168. <https://doi.org/10.1371/journal.pone.0115168> PMID: 25602282