

RESEARCH

Development and validation of a multi-locus DNA metabarcoding method to identify endangered species in complex samples

Alfred J. Arulandhu^{1,2,†}, Martijn Staats^{1,†}, Rico Hagelaar¹, Marleen M. Voorhuijzen¹, Theo W. Prins¹, Ingrid Scholtens¹, Adalberto Costessi³, Danny Duijsings³, François Rechenmann⁴, Frédéric B. Gaspar⁵, Maria Teresa Barreto Crespo⁵, Arne Holst-Jensen⁶, Matthew Birck⁷, Malcolm Burns⁸, Edward Haynes⁹, Rupert Hochegger¹⁰, Alexander Klingl¹¹, Lisa Lundberg¹², Chiara Natale¹³, Hauke Niekamp¹⁴, Elena Perri¹⁵, Alessandra Barbante¹⁵, Jean-Philippe Rosec¹⁶, Ralf Seyfarth¹⁷, Tereza Sovová¹⁸, Christoff Van Moorlehem¹⁹, Saskia van Ruth^{1,2}, Tamara Peelen²⁰ and Esther Kok^{1,*}

¹RIKILT Wageningen University & Research, P.O. Box 230, 6700 AE Wageningen, The Netherlands, ²Food Quality and Design Group, Wageningen University and Research, P.O. Box 8129, 6700 EV Wageningen, The Netherlands, ³Baseclear B. V, Einsteinweg 5, 2333 CC Leiden, The Netherlands, ⁴GenoStar Bioinformatics Solutions, 60 rue Lavoisier, 38330 Montbonnot Saint Martin, France, ⁵iBET, Instituto de Biologia Experimental e Tecnológica, Apartado 12, 2780-901 Oeiras, Portugal, ⁶Norwegian Veterinary Institute, Ullevaalsveien 68, P.O. Box 750 Sentrum, 0106 Oslo, Norway, ⁷U.S. Customs and Border Protection Laboratory, 1100 Raymond Blvd Newark, NJ 07102 USA, ⁸LGC, Queens Road, Teddington, Middlesex, TW11 0LY, UK, ⁹Fera, Sand Hutton, York, YO41 1LZ, UK, ¹⁰Austrian Agency for Health and Food Safety, Spargelfeldstrasse 191, 1220 Vienna, Austria, ¹¹Generalzolldirektion, Direktion IX, Bildungs- und Wissenschaftszentrum der Bundesfinanzverwaltung, Dienstort Hamburg, Baumacker 3, D-22523 Hamburg, Germany, ¹²Livsmedelsverket, Att. Lisa Lundberg, Strandbodgatan 4, SE 75323 Uppsala, Sweden, ¹³AGENZIA DELLE DOGANE E DEI MONOPOLI, Laboratori e servizi chimici – Laboratorio Chimico di Genova, 16126 Genova, Via Rubattino n. 6, Italy, ¹⁴Eurofins GeneScan GmbH, Engesserstrasse 4 79108 Freiburg, Germany, ¹⁵CREA-SCS sede di Tavazzano - Laboratorio via Emilia, Km 307, 26838 Tavazzano, Italy, ¹⁶Service Commun des Laboratoires, Laboratoire de Montpellier, Parc Euromédecine, 205 rue de la Croix Verte, 34196 Montpellier Cedex 5, France, ¹⁷Biolytix AG, Benkenstrasse 254, 4108 Witterswil, Switzerland, ¹⁸Crop Research Institute, Department of Molecular Genetics, Drnovská 507,

Received: 8 May 2017; Revised: 21 June 2017; Accepted: 15 August 2017

© The Author 2017. Published by Oxford University Press. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

161 06 Prague, Czech Republic, ¹⁹Laboratory of Customs & Excises, Blijde Inkomststraat 20, B-3000 Leuven, Belgium and ²⁰Dutch Customs Laboratory, Kingsfordweg 1, 1043 GN, Amsterdam, The Netherlands

*Correspondence address: Esther Kok, RIKILT Wageningen University & Research, P.O. Box 230, 6700 AE Wageningen, The Netherlands; Tel: +31317481978/+31317480283; Fax: +31317417717; E-mail: esther.kok@wur.nl

[†]Equal contribution

Abstract

DNA metabarcoding provides great potential for species identification in complex samples such as food supplements and traditional medicines. Such a method would aid Convention on International Trade in Endangered Species of Wild Fauna and Flora (CITES) enforcement officers to combat wildlife crime by preventing illegal trade of endangered plant and animal species. The objective of this research was to develop a multi-locus DNA metabarcoding method for forensic wildlife species identification and to evaluate the applicability and reproducibility of this approach across different laboratories. A DNA metabarcoding method was developed that makes use of 12 DNA barcode markers that have demonstrated universal applicability across a wide range of plant and animal taxa and that facilitate the identification of species in samples containing degraded DNA. The DNA metabarcoding method was developed based on Illumina MiSeq amplicon sequencing of well-defined experimental mixtures, for which a bioinformatics pipeline with user-friendly web-interface was developed. The performance of the DNA metabarcoding method was assessed in an international validation trial by 16 laboratories, in which the method was found to be highly reproducible and sensitive enough to identify species present in a mixture at 1% dry weight content. The advanced multi-locus DNA metabarcoding method assessed in this study provides reliable and detailed data on the composition of complex food products, including information on the presence of CITES-listed species. The method can provide improved resolution for species identification, while verifying species with multiple DNA barcodes contributes to an enhanced quality assurance.

Keywords: Endangered species; CITES; Traditional medicines; DNA metabarcoding; Customs agencies; COI; *matK*; *rbcl*; *cyt b*; mini-barcodes

Background

The demand for endangered species as ingredients in traditional medicines (TMs) has become one of the major threats to the survival of a range of endangered species such as sea-horse (*Hippocampus* sp.), agarwood (*Aquilaria* sp.), and Saiga antelope (*Saiga tatarica*) [1–3]. The Convention on the International Trade in Endangered Species of Wild Fauna and Flora (CITES) is one of the best-supported conservation agreements to regulate trading of animal and plant species [4] and thereby conserve biodiversity. Currently, ~35 000 species are classified and listed by CITES in 3 categories based on their extinction level (CITES Appendix I, II and III) by which the trade in endangered species is regulated. The success of CITES is dependent upon the ability of customs inspectors to recognize and identify components and ingredients derived from endangered species, for which a wide range of morphological, chromatographic, and DNA-based identification techniques can be applied [5, 6].

Recent studies have shown the potential of DNA metabarcoding for identifying endangered species in TMs and other wildlife forensic samples [5–8]. DNA metabarcoding is an approach that combines DNA barcoding with next-generation sequencing (NGS), which enables sensitive high-throughput multispecies identification on the basis of DNA extracted from complex samples [9]. DNA metabarcoding uses more or less universal polymerase chain reaction (PCR) primers to mass-amplify informative DNA barcode sequences [10, 11]. Subsequently, the obtained DNA barcodes are sequenced and compared to a DNA sequence reference database from well-characterized species for taxonomic assignment [9, 11]. The main advantage of DNA metabarcoding over other identification techniques is that it permits the identification of all animal and plant species within samples that are composed of multiple ingredients, which would not

be possible through morphological means and would be time-consuming with traditional DNA barcoding [5–7]. Furthermore, the use of mini-barcode markers in DNA metabarcoding facilitates the identification of species in highly processed samples containing heavily degraded DNA [6, 7]. Such a molecular approach could aid the Customs Authorities to identify materials derived from endangered species in a wide variety of complex samples, such as food supplements and TMs [12].

Before routine DNA metabarcoding can be applied, there are some key issues that need to be taken into account. First, complex products seized by Customs, such as TM products, may contain plant and animal components that are highly processed and from which the isolation of good-quality DNA is challenging. Second, the universal DNA barcodes employed may not result in amplification of the related barcode for each species contained in a complex sample, due to DNA degradation or the lack of PCR primer sequence universality. For plants, for example, different sets of DNA barcodes have been suggested for different fields of application (i.e., general taxonomic identification of land plants, identification of medicinal plants, etc.), and none of them meets the true requirements of universal barcodes [13]. Also, whilst PCR primers can be designed to accommodate shorter DNA barcode regions for degraded DNA samples, such mini-barcodes contain less information and their primers are more restrictive, often making them unsuitable for universal species barcoding [5, 14]. The third challenge is the reference sequence database quality and integrity, which is particularly problematic for law enforcement issues, where high quality and reliability are essential. The current underrepresentation of DNA barcodes from species protected under CITES and closely related species critically hampers their identification. The fourth challenge is that a dedicated bioinformatics pipeline is necessary to process raw NGS data for accurate and sensitive identification of CITES-listed

species [10]. Finally, studies using the DNA metabarcoding approach are scarce, and none of these methods have been truly validated [10, 15]. Therefore, before implementing DNA metabarcoding by Customs and other enforcement agencies, the above-mentioned challenges need to be thoroughly assessed to ensure accurate taxonomic identifications.

The objective of this research was to develop a multi-locus DNA metabarcoding method for (endangered) species identification and to evaluate the applicability and reproducibility of this approach in an international interlaboratory study. The research was part of a larger programme on the development of advanced DNA-based methods from the DECATHLON project [16], within the European Union's Framework Programme 7. In the process of establishing the standard operating procedure (SOP) for multi-locus DNA metabarcoding, all important aspects of the procedure (i.e., DNA isolation procedure, DNA barcode marker, barcode primers, NGS strategy, and bioinformatics) were evaluated. The challenges concerning the quality and integrity of the DNA reference database(s) are discussed. The first step was aimed at identifying an ideal DNA isolation method to extract DNA from complex mixtures consisting of both animal and plant tissues. Second, animal and plant DNA barcode markers and corresponding primer sets were identified from literature that allowed good resolution for identifying (endangered) species from a wide taxonomic range. Third, a panel of universal plant and animal DNA barcodes was selected, and a single optimal PCR protocol was identified for efficient amplification of a panel of DNA barcode markers. Finally, the suitability of the Illumina MiSeq NGS technology was evaluated, and a bioinformatics pipeline with a user-friendly web-interface was established to allow stakeholders to perform the NGS data analysis without expert bioinformatics skills.

The DNA metabarcoding method was developed and tested based on data generated for 15 well-defined complex mixtures. The use of well-characterized mixtures allowed for optimizing the bioinformatics procedure and subsequent robustness testing of multiple parameter settings and thresholds. The practical performance and reproducibility of the DNA metabarcoding strategy was assessed in an international validation trial by 16 laboratories from 11 countries, on the basis of 8 other newly composed complex mixtures and 2 seized TMs, which were suspected to contain ingredients derived from CITES species. In this study, the multi-locus DNA metabarcoding method is presented, and we assess whether the method can improve the compositional analysis of complex and real-life samples by enabling the sensitive and reproducible identification of CITES-listed taxa by enforcement agencies and other laboratories.

Data Description

To constitute well-defined complex mixtures, 46 reference specimens were commercially purchased from shops or were provided by the Dutch Custom Laboratory. In addition, 2 TMs that were suspected to comprise endangered species material were also obtained from the Dutch Customs Laboratory. Each reference specimen was identified morphologically. Genomic DNA was extracted from 29 animal and 17 plant reference species for DNA barcoding. Standard cytochrome c oxidase I (COI) barcodes for all animal specimens were generated and individually sequenced using the Sanger method, and they were compared against the Barcode of Life Data Systems and NCBI database for taxonomic confirmation. For plant species, the DNA barcodes *rbcl* and *matK* were sequenced to confirm species identity. For a

number of plant and animal species, the generated barcode sequence information was deposited in the European Nucleotide Archive (ENA) under accession numbers LT009695–LT009705 and LT718651 (Additional file 1; Table S1).

The complex mixtures for the pilot study and interlaboratory validation trial were prepared with 2 to 11 taxonomically well-characterized species present in relative concentrations (dry mass: dry mass) from 1% to 47%. For all experimental mixtures in the interlaboratory trial, internal control species were used to verify the efficiency of homogenization and to check for possible sample cross-contamination using species-specific quantitative PCR (qPCR) assays. DNA was isolated from the complex mixtures, and the concentration and purity of extracted DNA were determined using spectrophotometer (NanoDrop 1000, Thermo Fisher Scientific Inc.). Subsequently, PCR amplifications using 12 DNA barcode primer sets were performed. The pooled and purified amplicons of each sample were sequenced using an Illumina MiSeq paired-end 300 technology, following the manufacturer's instructions (Illumina, Inc.). The NGS datasets were analysed using the CITESpecies-Detect pipeline. All raw NGS datasets from both analyses were deposited in ENA under accession numbers ERS1545972–ERS1545988, ERS1546502–ERS1546533, ERS1546540–ERS1546619, ERS1546624–ERS1546639, ERS1546742–ERS1546757, ERS1546759–ERS1546774, and study number PRJEB18620 (Additional file 3; Table S1). A web-interface was developed for the CITESpecies-Detect pipeline to allow stakeholders to perform the NGS data analysis of their own samples. The web-interface can be globally accessed via the SURFsara high-performance computing and data infrastructure [17].

Analyses

Establishing a laboratory procedure for multi-locus DNA barcode amplification

Based on the previous studies on DNA isolation for TMs [5, 18] and from the comparison between modified Qiagen DNeasy plant mini kit [19] and cetyltrimethylammonium bromide (CTAB) isolation [20] (unpublished results), we identified that the CTAB isolation method in general yields better DNA purity and provides better PCR amplification success. Therefore, the CTAB DNA isolation method was selected for successive experiments.

The DNA barcode markers included in this study were selected based on Staats et al. [10], supplemented with additional primers from literature (Table 1) [14]. DNA barcode markers were selected based on the availability of universal primer sets and DNA sequence information in public repositories [10]. Important considerations in selecting suitable primer sets were that, preferably, they are used in DNA barcoding campaigns and studies and, as such, have demonstrated universal applicability across a wide range of taxa. Furthermore, primer sets for both the amplification of full-length barcodes and their respective mini-barcodes (i.e., short barcode regions <300 nt within existing ones) were selected when available. This was done to facilitate PCR amplification from a range of wildlife forensic samples containing relatively intact DNA (using full-length barcodes) and/or degraded DNA (mini-barcodes). Based on these criteria, PCR primer sets for the following animal DNA barcodes were selected: regions of the mitochondrial genes encoding the 16S rRNA gene (16S), COI, and cytochrome *b* (*cyt b*). For plant species identification, primer sets for the following DNA barcodes were selected: regions of the plastidial genes encoding maturase K (*matK*), ribulose-1,5-bisphosphate carboxylase (*rbcl*),

Table 1: Overview of the PCR primer sets used in this study for amplifying plant and animal DNA barcodes and mini-barcodes.

DNA marker	Primer name	Primer sequence 5'-3'	Amplicon length (nt)	Reference
Universal animal DNA barcodes and mini-barcodes				
16S	16sar-L	TCGTCGGCAGGCTCAGATGTGTATAAGAGACAGCGCCCTGTTTATCAAAACAT	500-600	Palumbi [41]
	16sar-H	GTCCTCGGGCTGGGATGTGTATAAGAGACAGCGCGTCTGAACCTCAGATCACGT		
Mini-16S	16S-forward	TGGTCGGCAGGCTCAGATGTGTATAAGAGACAGAYAGACGAGAAAGACCC	250	Sarri et al. [42]
	16S-reverse	GTCCTCGGGCTGGGATGTGTATAAGAGACAGGATTGGCTGTATTCC		
COI ^a	LepF1.t1	TGGTCGCAGGCTCAGATGTGTATAAGAGACAGATTCAACCAATCATAAAGATATTGG	648 ^c	Modified from Ivanova et al. [21]
	VF1.t1	TGGTCGCAGGCTCAGATGTGTATAAGAGACAGTTCTCAACCAACCAAAAGACATTGG		
	VF1d.t1	TGGTCGCAGGCTCAGATGTGTATAAGAGACAGTTCTCAACCAACCAARGAYATYGG		
	VF1i.t1	TGGTCGCAGGCTCAGATGTGTATAAGAGACAGTTCTCAACCAACCAIAAIGAIATIGG		
	LepR1.t1	GTCCTGCGGCTGGGAGATGTGTATAAGAGACAGTAACTTCTGGATGTCAAAATAATCA		
	VR1d.t1	GTCCTGCGGCTGGGAGATGTGTATAAGAGACAGTAGACTTCTGGTGGCCAAAGAAATCA		
	VR1.t1	GTCCTGCGGCTGGGAGATGTGTATAAGAGACAGTAGACTTCTGGTGGCCAAAGAAATCA		
	VR1i.t1	GTCCTGCGGCTGGGAGATGTGTATAAGAGACAGTAGACTTCTGGTGGCCAAAGAAATCA		
Mini-COI	mICOIntF	TGGTCGCAGGCTCAGATGTGTATAAGAGACAGCGGACGTTGACWGTWYAYCCYCC	313	Leray et al. [43], Geller et al. [44]
	jgHCO2198	GTCCTGCGGCTGGGAGATGTGTATAAGAGACAGTAACTTCTGGATGTCAAAATAATCA		
cyt b	L14816	TGGTCGCAGGCTCAGATGTGTATAAGAGACAGCCATCCAACTCTCAGCATGATAA	743	Palumbi [41], Parson et al. [45]
	CB3-H	GTCCTGCGGCTGGGAGATGTGTATAAGAGACAGGGCAAATAGGAARTATCAITTC		
Mini-cyt b	L14816	TGGTCGCAGGCTCAGATGTGTATAAGAGACAGCCATCCAACTCTCAGCATGATAA	357	Parson et al. [45]
	H15173	GTCCTGCGGCTGGGAGATGTGTATAAGAGACAGCCCTCAGAATGATATTTGCTCTCA		
Universal plant DNA barcodes and mini-barcodes				
matK	matK-KIM1R	TGGTCGCAGGCTCAGATGTGTATAAGAGACAGACCCAGTCCATCTGGAATCTTGGTTC	656-889	Fazekas et al. [46]
	matK-KIM3F	GTCCTGCGGCTGGGAGATGTGTATAAGAGACAGCGTACAGTACTTTTGTGTTTACGAG		
matK ^b	matK-390f	TGGTCGCAGGCTCAGATGTGTATAAGAGACAGCGGATTAITTCATTAATATTTTC	656-889	Cuénoud et al. [47]
	matK-1326r	GTCCTGCGGCTGGGAGATGTGTATAAGAGACAGTCTAGCACAGAAAGTCCGAAGT		
rbcL	rbcLa-F	TGGTCGCAGGCTCAGATGTGTATAAGAGACAGATGTCAACCAACACAGACTAAAGC	654	Levin et al. [48]
	rbcLa-R	GTCCTGCGGCTGGGAGATGTGTATAAGAGACAGGTAATAATCAAGTCCACRCG		Kress and Erickson [49]
rbcL ^b	rbcL a-F	TGGTCGCAGGCTCAGATGTGTATAAGAGACAGATGTCAACCAACACAGACTAAAGC	607	Levin et al. [48]
	rbcLajf634R	GTCCTGCGGCTGGGAGATGTGTATAAGAGACAGAAACGGTCTCTCCAAGCCAT		Fazekas et al. [50]
Mini-rbcL	F52	TGGTCGCAGGCTCAGATGTGTATAAGAGACAGTTGGATTCAAAGCTGGTGTTA	140 ^c	Little [14]
	R193	GTCCTGCGGCTGGGAGATGTGTATAAGAGACAGCVGTCCAMACAGTWGTCCATGT		Taberlet et al. [51]
trnL (UAA)	c	TGGTCGCAGGCTCAGATGTGTATAAGAGACAGCGAAATCGGTAGACGGTACG	767	Taberlet et al. [51]
	d	GTCCTGCGGCTGGGAGATGTGTATAAGAGACAGGGGATAGAGGACTTGAAC		
trnL (P6 loop)	g	TGGTCGCAGGCTCAGATGTGTATAAGAGACAGGGGCAATCCTGAGCCAA	10-143 ^c	Taberlet et al. [51]
	h	GTCCTGCGGCTGGGAGATGTGTATAAGAGACAGCCATTTGAGTCTCGCACCTATC		
ITS2	S2F	TGGTCGCAGGCTCAGATGTGTATAAGAGACAGATGCCATCTTGGTGTGAAT	160-320 ^c	Chen et al. [52]
	S3R	GTCCTGCGGCTGGGAGATGTGTATAAGAGACAGACGGCTTCTCCAGACTACAAT		
psbA-trnH ^b	psbAf	TGGTCGCAGGCTCAGATGTGTATAAGAGACAGTTATGCATCAACGTAATGTC	264-792	Sang et al. [53], Tate and Simpson [54]
	trnH2	GTCCTGCGGCTGGGAGATGTGTATAAGAGACAGCGCGCATGGTGTGATTCAACAATCC		

The shaded text represents the sequence of the Illumina overhang adapters.

^aModified COI cocktail primers without M13-tails were used [21].

^bThe primers were not included in the final panel of DNA barcodes.

^cAmplicon length excluding primers.

tRNA^{Leu} (UAA) intron sequence (*trnL* (UAA)), *psbA-trnH* intergenic spacer region (*psbA-trnH*), and the nuclear internal transcribed spacer 2 (ITS2) region (Table 1). The selected primers sets were modified to include the Illumina adapter sequence at the 5' end of the locus-specific sequence to facilitate efficient NGS library preparation. A gradient PCR experiment was performed to identify the optimal PCR annealing temperature. While the selected PCR primer sets had previously been published with their own annealing temperatures and conditions, the identification of a single optimal annealing temperature for all PCR primer sets would allow for increased efficiency of analysis. Initially, a thermal gradient of 49.0°C to 55.0°C was tested on the *Bos taurus* reference material with the primer sets for COI, 16S, mini-16S, and *cyt b*. The amplification efficiency across the PCR primers sets was determined by comparing the intensity of the amplicons across the thermal gradient. An optimal annealing temperature of 49.5°C was identified, but additional non-specific amplicons were observed with some primers (not shown). To reduce the amounts of non-specific amplification products, the PCR program was modified to increase the annealing temperature after 5 cycles from 49.5°C to 54.0°C [21], and it was tested on all 15 PCR primer sets (Table 1). It was observed that certain PCR primer combinations still produced non-specific products (for *psbA-trnH* gene) or less intense PCR products (for the *rbcl* gene with primers *rbclA-F* and *rbclLajf634R*, and the *matK* gene with primers *matK-390f* and *matK-1326r*). Consequently, these PCR primer sets were excluded from subsequent experiments.

Next, the selected PCR thermocycling protocol was evaluated with the remaining 12 PCR primer sets on a panel of 29 animal and 17 plant species, representing a phylogenetically wide range of taxa (Mammalia, Actinopterygii, Malacostraca, Bivalvia, Aves, Reptilia, Amphibia, Insecta, Angiospermae, and Cycadopsida) (Additional file 1; Table S2 and S3). The overall PCR amplification success rates varied across reference species and across DNA barcode markers (Additional file 1; Table S2). For instance, no PCR amplification was observed with *cyt b* for the CITES-listed species *Balaenoptera physalus*, whereas intense amplification was seen for the same species with 16S, COI, mini-16S, and mini-COI (Additional file 1; Table S2). Overall, at least 1 DNA barcode marker could successfully be amplified for each of the 46 plant and animal species (Additional file 1; Table S2 and S3). For a number of plant and animal species, the generated barcode sequence information was deposited in the ENA under accession numbers LT009695–LT009705 and LT718651 (Additional file 1; Table S1).

Development and pre-validation of the CITESpeciesDetect bioinformatics pipeline

A dedicated bioinformatics pipeline, named CITESpeciesDetect, was developed for the purpose of rapid identification of CITES-listed species using Illumina paired-end sequencing technology. Illumina technology was selected because it produces NGS data with very low error rates compared to other technologies [2, 22]. Furthermore, the Illumina MiSeq platform enables paired-end read lengths of up to 300 nt, allowing relatively long DNA barcode regions of up to ~550 nt to be assembled. Also, the multiplexing capabilities of Illumina technology are well developed, allowing for simultaneous sequencing of multiple samples in 1 run, thereby enabling more cost-efficient NGS. While NGS data analysis pipelines exist that allow processing of Illumina DNA metabarcoding datasets (e.g., CLOTU, QIIME, Mothur), the majority have been developed for

specifically studying microbial communities using the 16S rRNA gene region. CITESpeciesDetect, developed in this study, extends on the frequently-used software tools developed within the USEARCH [22] and BLAST+ packages [23], and additionally includes dedicated steps for quality filtering, sorting of reads per barcode, and CITES species identification (Fig. 1). The CITESpeciesDetect is composed of 5 linked tools, and data analysis passes through 3 phases: (i) pre-processing of paired-end Illumina data involving quality trimming and filtering of reads, followed by sorting by DNA barcode, (ii) operational taxonomic unit (OTU) clustering by barcode, and (iii) taxonomy prediction and CITES identification.

It was found that with the current setup of the pipeline, reads generated for *cyt b* and mini-*cyt b* could not be separated based on the forward PCR primer as the forward primers are identical. It was therefore decided to combine (pool) the overlapping reads of *cyt b* and mini-*cyt b* during pre-processing (primer selection) of reads to prevent reads from being double selected. This means that the results of *cyt b* and mini-*cyt b* are presented by the CITESpeciesDetect pipeline as *cyt b*. The same issue was found for COI barcode and mini-barcode markers, for which the results are presented as COI.

A parameter scan was performed in order to assess the effect of software settings on the ability to identify species. The evaluation allowed for the identification of important parameters and their effects on the sensitivity, specificity, and robustness of the procedure. Changing the base quality score has a major impact on the number of reads per barcode (Additional file 1; Table S4). Increasing the strictness of the base quality score resulted in decreasing numbers of reads per barcode. Quality score values other than the default values (Q20 for 95% of bases) did not yield better identifications. When applying strict quality filtering settings (Q20 for 100% of bases, or Q30 for 99% of bases) the species *Pieris brassicae* and *Anguilla anguilla* could not be detected with *cyt b* and/or mini-COI, indicating that these settings were too strict (Additional file 1; Table S5). This is likely due to the resulting overall low read numbers for *cyt b* and mini-COI when applying these strict quality filtering settings (Additional file 1; Table S4).

An OTU abundance threshold is generally applied to make DNA metabarcoding less sensitive to (potential) false-positive identifications. False-positives may occur, e.g., as contaminants during pre-processing of samples (DNA extraction, PCR) or as cross-contamination during Illumina sequencing. Applying an OTU abundance threshold higher than 0 generally results in loss of sensitivity. We have found, however, that applying an OTU abundance threshold of higher than 0 may help in reducing noisy identifications and potential false-positive identifications (results not shown). It should be noted that applying filtering thresholds may always lead to false-negative or false-positive identifications. In this study, an OTU abundance threshold of 0.2% was set as default; however, the OTU abundance threshold may need re-evaluation for samples with expected very low species abundances (<1% dry weight).

The effect of applying a minimum DNA barcode length revealed that allowing DNA barcodes of ≥ 10 nt did not lead to additional identification of species, compared with default settings (e.g., ≥ 200 nt). Increasing the minimal DNA barcode length to 250 nt, however, resulted in a failure to identify most plant species with mini-*rbcl* and *rbcl*. We implemented a minimum DNA barcode length of 200 nt, except for DNA barcodes with a basic length shorter than 200 nt, in which case the minimum expected DNA barcode length is set to 100 nt for ITS2, 140 nt for mini-*rbcl*, and 10 nt for the *trnL* (P6 loop) marker.

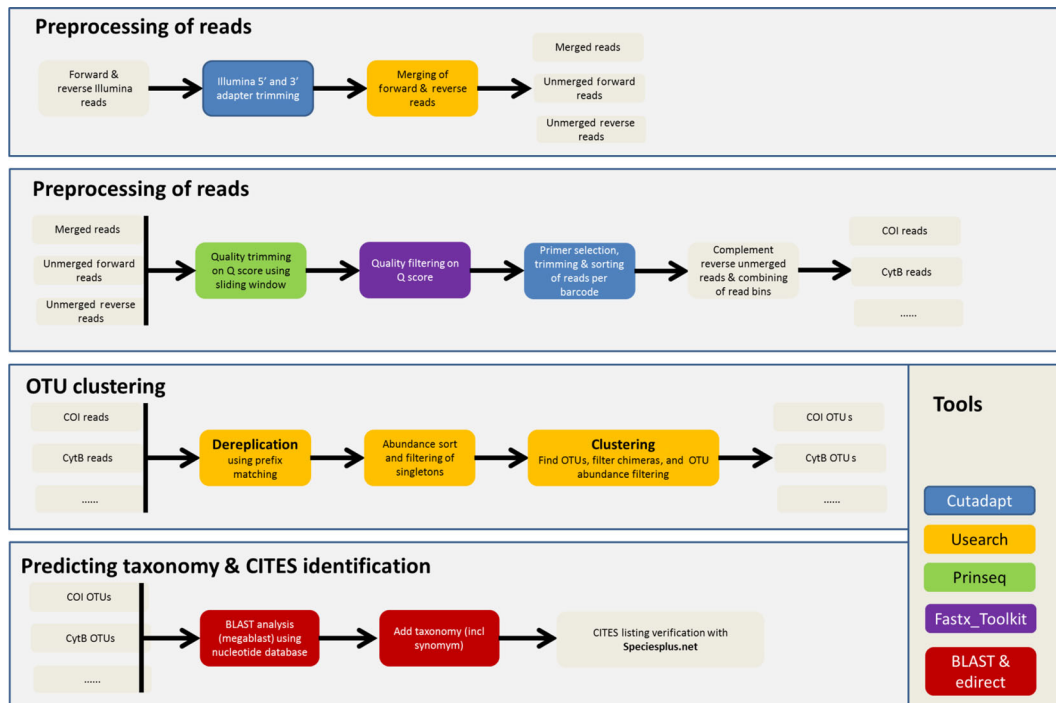


Figure 1: Schematic representation of the CITESpeciesDetect pipeline.

The parameter scan resulted in specifying recommended parameter values (default setting) for analysing DNA metabarcoding datasets using the CITESpeciesDetect pipeline (see the Bioinformatics analysis section). An online version of the CITESpeciesDetect pipeline with a user-friendly web-interface was developed for skilled analysts with basic, but no expert-level knowledge in bioinformatics [17].

Pilot study to assess the performance of the DNA metabarcoding procedure using experimental mixtures

The DNA metabarcoding procedure was assessed in a pilot study, for which 15 complex mixtures (EM1–EM15) were prepared containing from 2 to 10 taxonomically well-characterized species with DNA barcode reference sequences available in the NCBI reference database (Table 2). The experimental mixtures 10 and 11 (EM10 and EM11) were independently analysed twice to verify repeatability of the method (DNA isolation, barcode panel analysis, and pooling). Only mixtures were used with well-characterised species (DNA Sanger barcoded and taxonomically verified) ingredients, at known dry weight concentrations, and with high-quality DNA that would allow for an assessment of the performance of the DNA metabarcoding method under optimal conditions.

A total of 2.37 Gb of Illumina MiSeq sequencing data was generated for the 17 complex samples (15 complex mixtures along with the 2 replicates). On average, 464 648 raw forward and reverse Illumina reads were generated per sample, with minimum and maximum read numbers ranging between 273 104 (mixture EM4) and 723 130 (mixture EM10R) (Table 3). During raw data pre-processing with the default settings of the CITESpeciesDetect pipeline, the reads were first quality filtered, and overlapping paired-end Illumina reads were merged into pseudo-reads

(Fig. 1). The samples contained on average 269 099 quality-controlled (QC) unmerged (forward and reverse) reads and merged pseudo-reads, collectively named (pseudo-)reads. On average 88.27% (min = 77.38%, max = 96.26%) of raw reads passed the quality filtering and pre-processing steps, indicating that the overall quality of the Illumina data was high (not shown).

Next, the (pseudo-)reads were assigned to DNA barcodes based on PCR primer sequences. On average, 96.44% (min = 88.78%, max = 98.21%) of QC pre-processed reads were assigned to DNA barcodes, indicating a high percentage of reads containing the locus-specific DNA barcode primers (Table 3). After this, the (pseudo-)reads were clustered by 98% sequence similarity into OTUs. On average, 82.26% (min = 75.11%, max = 90.63%) of the DNA barcodes assigned reads were clustered into OTUs (Table 3). It was assumed that the small fraction of reads that was not assigned to OTUs contained non-informative (e.g., non-specific fragments, chimeras) sequences that may have been generated during PCR amplification and were filtered out during clustering.

For taxonomy prediction, OTUs were assigned to dataset sequences using BLAST when aligning with at least 98% sequence identity, a minimum of 90% query coverage, and an E-value of at least 0.001. Generally, the best match (“top hit”) is used as a best estimate of species identity. However, species identification using BLAST requires careful weighting of the evidence. To minimize erroneous taxonomic identifications, a more conservative guideline was used that allowed a species to be assigned only when the best 3 matches identified the species. If the bit scores do not decrease after the top 3 hits, or if other species have identical bit scores, then identification was considered inconclusive. In such cases, OTUs were assigned to higher taxonomic levels (genus, family, or order). All animal ingredients, except *Parapenaopsis* sp., could be identified at the species level with 1

Table 2: Pilot study: composition of the experimental mixtures and taxa identified using the default setting of the CITESpeciesDetect pipeline.

Experimental mixtures		EM1	EM2	EM3	EM4	EM5	EM6	EM7	EM8	EM9	EM10	EM10R	EM11	EM11R	EM12	EM13	EM14	EM15
Species/genus	Common name	99% (S)	90% (S)	1% (S)	10% (S)	99% (S)	95% (S)	85% (S)			10% (S)	10% (S)	46% (S)	46% (S)	95% (S)	85% (S)	1%	
<i>Bos taurus</i>	Cattle																	
<i>Parapenaeopsis</i> sp.	Shrimp						1%	3%			10%	10%	1%	1%	95% (S)		1%	3%
<i>Anguilla anguilla</i> ^a	European eel						1%	3%			10% (S)	10% (S)	1% (S)	1% (S)			1% (S)	3% (S)
<i>Crocodylus niloticus</i> ^a	Nile crocodile						1% (S)	3% (S)			10% (S)	10% (S)	1% (S)	1% (S)			1% (S)	3% (S)
<i>Gallus gallus</i>	Domestic chicken						1% (S)	3% (S)			10% (S)	10% (S)	1% (S)	1% (S)			1% (S)	3% (S)
<i>Pteris brassicae</i>	Large white (caterpillar)						1% (S)	3% (S)			10% (S)	10% (S)	1% (S)	1% (S)			1% (S)	3% (S)
<i>Echinocactus</i> sp. ^a	Barrel cactus						1% (S)	3% (S)			10% (S)	10% (S)	1% (S)	1% (S)			1% (S)	3% (S)
<i>Euphorbia</i> sp. ^a	Spurge						1% (F)	3% (F)			10% (F)	10% (F)	1% (F)	1% (F)			1% (F)	3% (F)
<i>Aloe variegata</i> ^{a,b}	Tiger aloe			1% (F)			1% (F)	3% (F)			10% (F)	10% (F)	1% (F)	1% (F)			1% (F)	3% (F)
<i>Dendrobium</i> sp. ^a	Dendrobium (orchid)						1% (F)	3% (F)			10% (F)	10% (F)	1% (F)	1% (F)			1% (F)	3% (F)
<i>Cycas revoluta</i> ^a	Sago palm						1%	3%			10% (G)	10% (G)	1%	1%			1%	3%
<i>Lactuca sativa</i>	Lettuce	1% (S)	10% (S)	99% (S)	90% (S)		95% (S)	85% (S)			10% (S)	10% (S)	46% (S)	46% (S)	1% (G)	3% (G)	95% (S)	85% (S)

Taxa were identified at the species level unless otherwise indicated in parentheses. Cells highlighted in grey indicate that taxa that were not identified. Identified taxa listed by CITES are highlighted in bold.

The symbol next to percentage indicates the taxonomic resolution of the identified taxon: (F): family level; (G): genus level; (S): species level.

^aSpecies listed by CITES.

^b*Aloe variegata* (synonym *Gonialoe variegata*) was recently assigned to the genus *Gonialoe* [55].

or more DNA barcode markers using the default settings of the CITESpeciesDetect pipeline (Tables 4 and 5). For plants, *Lactuca sativa* could be identified at the species level using the trnL (P6 loop). All other plant taxa were identified at the genus or higher level (Tables 4 and 5).

Putative contaminating species were observed in most of the experimental mixtures from multiple markers; detailed information about the identified cross-contained species in a sample and the related markers are specified in Additional file 2; Table S1. Even with the default OTU abundance threshold in place, the species *L. sativa*, *B. taurus*, and *Gallus gallus* were identified in mixtures that were not supposed to contain these species. To verify whether these putative contaminations occurred during DNA isolation or Illumina sequencing, qPCR assays for the specific detection of *B. taurus* and *G. gallus* were performed on selected DNA extracts. The high Cq values above 39 indicated the presence of these species, however, in low copy number, which suggests that for some experimental mixtures (EM8, EM9, and EM14) cross-contamination had occurred during sample preparation or DNA isolation, while for other experimental mixtures (EM15) cross-contamination may have occurred during PCR, Illumina library preparation, or sequencing. In addition to these contaminants, a species of *Brassica* was identified in experimental mixtures containing *P. brassica*. This result is most likely not a false-positive, because the caterpillars used for this study had been fed on cabbage.

The DNA metabarcoding method was found to be sensitive enough to identify most plant and animal taxa at 1% (dry mass: dry mass) in mixtures of both low (EM1, EM3, and EM5) (Table 2) and relatively high complexity (EM6, EM8, EM11, EM12, and EM14) (Table 2). The exception being *Parapenaeopsis* sp. (all mixtures), *A. anguilla* in EM6, and *Cycas revoluta* in EM8 and EM11. Careful inspection of the NGS data revealed that in nearly all cases OTUs related to *Parapenaeopsis* sp., *A. anguilla*, and *C. revoluta* were present, but that these sequences had been filtered out by the CITESpeciesDetect pipeline because their cluster sizes did not fulfil the 0.2% OTU abundance threshold. There appeared to be no trend as to the type and length of DNA barcode marker that had been filtered out by the CITESpeciesDetect pipeline. For instance, *Parapenaeopsis* sp. was detected below the OTU threshold with *cyt b*, mini-16S, COI, and 16S markers (not shown). Lowering the OTU abundance threshold, however, would lead to (more) false-positive identifications, and this was therefore not implemented.

The repeatability of the laboratory procedure (excluding NGS) was assessed by analysing the experimental mixtures 10 and 11 (EM10R and EM11R) (Table 2), which was independently performed twice, i.e., DNA isolation and PCR barcode amplification, but NGS was performed on the same MiSeq flow cell as the other samples of the pilot study. From the comparison, it was observed that the percentage of QC reads was nearly twice as high in the replicate analyses (Table 3). Also, the percentage of QC reads assigned to DNA barcodes varied among replicate analyses (Fig. 2). Most notable were the observed differences among replicate analyses in the percentage reads assigned to *matK* and the trnL (P6 loop). For example, the percentages of QC reads assigned to *matK* were 6.11% (14 081 reads) and 0.02% (97 reads) in EM10 and EM10R, respectively (Fig. 2). The low number of reads assigned to *matK* limited its use for taxonomy identification in EM10R (Table 4). The multi-locus approach, however, allowed for the repeatable identification of taxa in EM10 and EM11, though not in all cases with all DNA barcode markers (Tables 4 and 5).

Table 3: Pilot study: average number of Illumina MiSeq reads, the average number of (pseudo-)reads that passed QC, and the percentage of QC (pseudo-)reads that were assigned to DNA barcodes and OTUs generated per sample.

Experimental mixture	Number of raw reads	Percentage of QC (pseudo-)reads ^a	Percentage DNA barcode assigned (pseudo-)reads ^a	Percentage OTU clustered (pseudo-)reads ^a
EM1	466 108	88.07	95.68	83.86
EM2	448 428	86.04	97.24	84.04
EM3	496 328	87.46	96.61	84.34
EM4	273 104	77.38	95.74	80.54
EM5	582 254	96.26	97.84	90.63
EM6	442 574	92.81	97.54	81.48
EM7	394 354	93.04	97.14	80.70
EM8	455 172	79.62	95.66	82.35
EM9	434 326	86.23	97.30	83.60
EM10	387 816	87.73	97.00	75.11
EM10R	723 130	95.59	98.02	87.39
EM11	363 374	84.44	96.74	78.63
EM11R	635 304	91.11	98.21	87.01
EM12	355 634	92.55	97.54	76.54
EM13	405 742	89.46	96.49	77.31
EM14	480 772	85.74	95.98	81.91
EM15	554 602	87.05	88.78	82.98
Average ^b	464 648	88.27	96.44	82.26

^a(Pseudo-)reads are the combined QC pseudo-reads and the QC processed unmerged forward and reverse reads.

^bAveraged across the 17 Illumina MiSeq datasets.

Table 4: Taxonomic resolution provided by each DNA barcode marker for EM10 and EM10R.

Species/genus	Species	Genus	Family
<i>Anguilla anguilla</i>	<i>cyt b</i>	mini-16S	
<i>Parapenaepsis</i> sp.			
<i>Bos taurus</i>	16S, mini-16S, cyt b, COI		
<i>Gallus gallus domesticus</i>	mini-16S, cyt b, COI		
<i>Pieris brassicae</i>	COI		
<i>Echinocactus</i> sp.			matK, rbcL, mini-rbcL, ITS2
<i>Euphorbia</i> sp.		rbcL, mini-rbcL	ITS2
<i>Aloe variegata</i>			matK, rbcL, mini-rbcL, trnL (UAA)
<i>Cycas revoluta</i>		rbcL-mini, trnL (P6 loop)	
<i>Lactuca sativa</i>	trnL (P6 loop)	matK, trnL (UAA), ITS2	rbcL, mini-rbcL

Highlighted in bold are DNA barcodes with the same taxonomic resolution in both samples.

Table 5: Taxonomic resolution provided by each DNA barcode marker for EM11 and EM11R.

Species/genus	Species	Genus	Family
<i>Anguilla anguilla</i>	cyt b		
<i>Parapenaepsis</i> sp.			
<i>Bos taurus</i>	16S, mini-16S, cyt b, COI		
<i>Gallus gallus domesticus</i>	cyt b, COI		
<i>Pieris brassicae</i>	COI		
<i>Echinocactus</i> sp.			matK, rbcL, ITS2
<i>Euphorbia</i> sp.		rbcL, mini-rbcL	
<i>Aloe variegata</i>			matK, rbcL, mini-rbcL, trnL (UAA)
<i>Cycas revoluta</i>		mini-rbcL, trnL (P6 loop)	
<i>Lactuca sativa</i>	trnL (P6 loop)	matK, rbcL, trnL (UAA), ITS2	rbcL, mini-rbcL

Highlighted in bold are DNA barcodes with the same taxonomic resolution in both samples.

Based on the results obtained from the pilot study, precautions were taken when grinding the freeze-dried materials and mixing to avoid cross-contamination during the laboratory handling of samples, which were used to improve the SOP for the interlaboratory trial (see the protocols in

[24]). Also, control species were added to experimental mixtures that were prepared for the inter-laboratory trial to allow better confirmation of sample homogeneity and to verify that no cross-contamination had occurred during sample preparation.

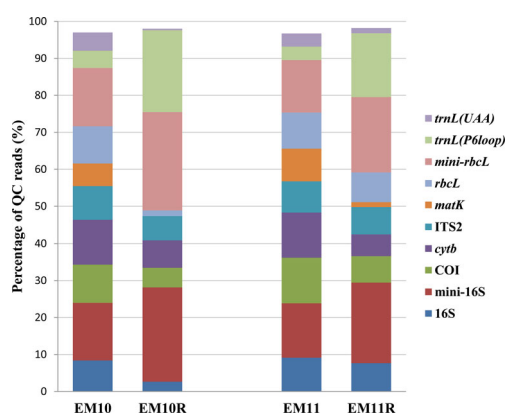


Figure 2: The percentage of QC reads assigned to DNA barcodes for samples EM10, EM10R, EM11, and EM11R of the pilot study.

Table 6: Laboratories participating in the interlaboratory trial.

Laboratory	City and country
Agenzia delle Dogane E dei Monopoli	Genoa, Italy
AGES	Vienna, Austria
BaseClear BV	Leiden, The Netherlands
Biolytix AG	Witterswil, Switzerland
CREA-SCS sede di Tavazzano—Laboratorio	Tavazzano, Italy
Crop Research Institute	Prague, Czech Republic
Dutch Customs Laboratory	Amsterdam, The Netherlands
Eurofins GeneScan GmbH	Freiburg, Germany
Fera	Sand Hutton, United Kingdom
Generalzolldirektion	Hamburg, Germany
Laboratoire de Montpellier	Montpellier, France
Laboratorium Douane Accijnzen	Leuven, Belgium
LGC	Middlesex, United Kingdom
Livsmedelsverket	Uppsala, Sweden
RIKILT Wageningen University & Research	Wageningen, The Netherlands
US Customs and Border Protection Laboratory	Newark, United States

Assessment of interlaboratory reproducibility of the DNA metabarcoding procedure

Altogether 16 laboratories from 11 countries (all experienced, well-equipped, and proficient in advanced molecular analysis work), including 2 of the method developers, participated in the inter-laboratory trial (Table 6). The laboratories received 10 anonymously labelled samples, each consisting of 250 mg of powdered material. Two of the samples, labelled S3 and S8, were authentic TM products seized by the Dutch Customs Laboratory while the other 8 samples were well-characterized mixtures of specimens from carefully identified taxa in relative dry weight concentrations from 1% to 47% (Table 7). In all experimental mixtures, 1% of *Zea mays* was added as quality control for homogeneity, which was confirmed with maize-specific high-mobility group gene (*hmg*) qPCR [19]. Also, tests performed with species-specific qPCR assays indicated that cross-contamination did not occur during sample preparation (Additional file 1; Table S6). The qPCR assay for the detection of *Brassica napus*, however, also gave a positive signal for other *Brassica* sp. in the mixtures.

Together with the sample materials, reagents for DNA extraction, and the complete set of barcode primers, the participants received an obligatory SOP. Any deviations from the SOP had to be reported. The participants were instructed to extract DNA, perform PCR using the barcode primers, purify the amplified DNA by removal of unincorporated primers and primer dimers, and assess the quality and quantity of the amplification products by gel electrophoresis and UV spectrophotometry. The purified PCR products were then collected by the coordinator of the trial (RIKILT Wageningen University & Research, the Netherlands) and shipped to a sequencing laboratory (BaseClear, the Netherlands) for Illumina sequencing using MiSeq PE300 technology. The sequencing laboratory performed Index PCR and Illumina library preparation prior to MiSeq sequencing, as specified in the Illumina 16S metagenomics sequencing library preparation guide. The altogether 160 PCR samples were sequenced using 2 Illumina flow cells with MiSeq reagent kit v. 3.

The interlaboratory trial should ideally have included the use of the online version of the pipeline, but unfortunately this was not possible due to shortage of time. Therefore, a single (developer) laboratory performed these bioinformatics analyses. The 160 individual samples contained on average 269 057 raw reads, and more than 150 000 reads per sample in 95% of the samples (Additional file 1; Table S7). One sample contained less than 100 000 reads (51 750), which was considered more than sufficient for reliable species identification. After pre-processing, the samples contained on average 142 938 (pseudo-)reads. On average, 94.66% of the reads (min = 88.12%, max = 98.02%) passed the quality filtering, indicating that the overall quality of the sequence data was consistently high across the 160 datasets.

OTU clustering at 98% sequence similarity on average assigned 78.14% of the pre-processed and DNA barcode assigned reads into OTUs (Additional file 1; Table S7). Only 2 samples, both from the same laboratory, had a slightly lower percentage of the (pseudo-)reads assigned to OTUs (66.02% and 66.05%). This indicates that the pipeline correctly removed PCR artefacts in the clustering phase.

For taxonomy prediction, an OTU would be assigned to a database hit if they aligned with $\geq 98\%$ sequence identity and $\geq 90\%$ query coverage and yielded an expected value (E-value) of at least 0.001. The BLAST output of the NGS data was interpreted by participants according to the guidelines in the SOP. Variation was observed among laboratories in interpreting the BLAST output: some laboratories consistently scored the top hits, irrespective of bitscore, while other labs selected all hits belonging to the top 3 bitscores or interpreted only the first OTU of each DNA barcode, leading to large differences in identified taxa. Because of these inconsistencies, the BLAST results were re-interpreted by RIKILT Wageningen University & Research following the established guideline, as mentioned in the SOP. These re-interpreted data are the data referred to in the following sections.

With 1 exception, all taxa mixed in at $\geq 1\%$ (dry mass: dry mass) were reproducibly identified by at least 13 (81%) laboratories (Table 7). *Beta vulgaris* in sample S6 could only be identified by 4 out of 16 (25%) laboratories. *Beta vulgaris*-specific sequences were present in all remaining datasets, but at very low read counts. So these clusters did not fulfil the 0.2% OTU abundance threshold (Additional file 2; Table S2). In order to provide insight into what alternative setting of the CITESpeciesDetect pipeline may have been better suited for identifying *Beta vulgaris*, 3 datasets with relatively low (S6 – laboratory 13), medium (S6 – laboratory 14), and high (S6 – laboratory 6) data volumes were

Table 7: Interlaboratory trial study: composition of the complex mixtures and taxa identified using the default setting of the CITESpeciesDetect pipeline.

Species/genus	Common name	Homogenized mixtures								
		S1	S2	S4	S5	S6	S7	S9	S10	
<i>Zea mays</i>	Maize	1% (13) Poaceae	1% (14) Poaceae	1% (14) Poaceae	1% (15) Poaceae	1% (16) Poaceae	1% (15) Poaceae	1% (15) Poaceae	1% (14) Poaceae	1% (16) Poaceae
<i>Glycine max</i>	Soy bean	1% (16) Glycine sp.								
<i>Gossypium hirsutum</i>	Cotton		1% (16) <i>Gossypium</i> sp.							
<i>Brassica napus</i>	Canola			1% (16) <i>Brassica</i> sp.						
<i>Triticum aestivum</i>	Wheat				1% (15) Poaceae					
<i>Beta vulgaris</i>	Sugar beet					1% (4) <i>Beta</i> sp.				
<i>Meleagris gallopavo</i>	Turkey						1% (16)			
<i>Carica papaya</i>	Papaya							1% (16)		
<i>Solanum lycopersicum</i>	Tomato									1% (16)
<i>Aloe variegata</i>	Tiger aloe	1% (16)	2% (16)	3% (16)	4% (16)	1% (16)	2% (16)	3% (16)	4% (16)	1% (16)
<i>Dendrobium</i>	<i>Dendrobium</i>	Xanthorrhoeaceae	Xanthorrhoeaceae	Xanthorrhoeaceae	Xanthorrhoeaceae	Xanthorrhoeaceae	Xanthorrhoeaceae	Xanthorrhoeaceae	Xanthorrhoeaceae	Xanthorrhoeaceae
sp. ^a	orchid	1% (16)	2% (16)	3% (16)	4% (16)	1% (16)	2% (16)	3% (16)	4% (16)	1% (16)
<i>Huso dauricus</i> ^a	Sturgeon/Kaluga	<i>Dendrobium</i> sp.	<i>Dendrobium</i> sp.	<i>Dendrobium</i> sp.	<i>Dendrobium</i> sp.	<i>Dendrobium</i> sp.	<i>Dendrobium</i> sp.	<i>Dendrobium</i> sp.	<i>Dendrobium</i> sp.	<i>Dendrobium</i> sp.
<i>Crocodylus niloticus</i> ^a	Nile crocodile	1% (16)	2% (16)	3% (16)	4% (16)	1% (14)	2% (16)	3% (16)	4% (16)	1% (16)
<i>Lactuca sativa</i>	Lettuce	1% (14)	2% (14)	3% (15)	4% (16)	1% (9)	2% (15)	3% (15)	4% (15)	1% (16)
<i>Brassica oleracea</i>	White cabbage	47% (16)	45% (16)	43% (16)	41% (16)	10% (16)	10% (16)	10% (16)	10% (16)	10% (16)
<i>Sus scrofa</i>	Pig					32% (16)	30% (16)	28% (16)	26% (16)	26% (16)
<i>Bos taurus</i>	Cattle	47% (16)	45% (16)	43% (16)	41% (16)	10% (16)	10% (16)	10% (16)	10% (16)	10% (16)
<i>Pleuronectes platessa</i>	European plaice					32% (16)	30% (16)	28% (16)	26% (16)	26% (16)

Taxa were identified at the species level unless otherwise indicated. The number of laboratories that have identified a taxon at the species level or higher is provided in parentheses. Identified taxa listed by CITES are highlighted in bold.

^aSpecies listed by CITES.

^b*Aloe variegata* (synonym *Goniaboe variegata*) was recently assigned to the genus *Goniaboe* [55].

reanalysed using a range of different settings for the OTU minimum cluster size and OTU abundance threshold (Additional file 2: Tables S3–S5). Setting the OTU minimum cluster size to 2, 4, or 6 has no effect on taxon identification, and *Beta vulgaris* is not identified at the species or higher taxonomic level in the datasets of laboratories 6 and 13. Setting the OTU abundance threshold to 0 allows the identification of *Beta vulgaris* in all 3 samples, but at the expense of many false-positive identifications. Applying an OTU abundance threshold of 0.1% (default is 0.2%) allows the identification of *Beta vulgaris* at the species or genus level irrespective of any differences in data volume between the 3 samples.

All 6 animal species could be identified to the species level with at least 1 barcode marker (COI), while only 4 of the 12 plant species (*Brassica oleracea*, *Carica papaya*, *Gossypium hirsutum*, and *L. sativa*) could be identified to the species level (Additional file 2; Table S6). All other plant species were identified at the genus level or higher. For plants, no single barcode marker was best, and the most reliable data were obtained by combining the plant barcodes.

Three taxa that were misidentified or not intentionally included in the mixtures were reproducibly identified across all laboratories. *Acipenser schrenckii* co-occurred in all samples containing *Huso dauricus*. We have confirmed with DNA metabarcoding that the caviar used for preparing the experimental mixtures contains both *H. dauricus* and *A. schrenckii* (results not shown). Furthermore, *Brassica rapa* was identified by ITS2 in sample S4 by all 16 (100%) laboratories, instead of *Brassica napus*. We confirmed by Sanger sequencing *rbcl* and *matK* that our reference specimen is indeed *Brassica napus*, but that its ITS2 sequence is identical to *Brassica rapa* (LT718651). Finally, a taxon of the plant family Phellinaceae was reproducibly identified (by all laboratories) using the mini-*rbcl* marker in all samples containing *L. sativa* (S6, S7, S9, S10). Species of the family Phellinaceae and *L. sativa* both belong to the order Asterales. The evidence for Phellinaceae was not strong; i.e., the family-level identification was based on a single NCBI reference sequence only (GenBank: X69748). We therefore suspect a misidentification during the interpretation of the BLAST results.

Taxa that were identified to be the result of possible contaminations were scarcely observed; i.e., these were found in isolated cases and could possibly be explained by cross-sample contamination that may have occurred during any step of sample processing (DNA isolation, PCR, NGS library preparation, or NGS). For example, a contamination with *Gossypium* sp. was observed using *trnL* (P6 loop) in sample S1 of 1 of the participating labs. A total of 6 such suspected cases of incidental cross-contaminations were observed (not shown).

For the authentic TMs S3 and S8, it was observed that few labelled ingredients could reproducibly be identified (Tables 8 and 9). For sample S3 (Ma pak leung sea-dog), only the listed ingredients *Cuscuta* sp. (Chinese dodder seed) and *Astragalus danicus* (*Astragalus* root) could be identified. For sample S8 (Cobra performance enhancer), only the listed ingredients *Epimedium* sp. (Horny goat weed; Berberidaceae) and *Panax ginseng* (Korean ginseng; Araliaceae) and the species of the plant families Arecaceae (*Serenoa repens*) and Rubiaceae (*Pausinystalia johimbe*) could be identified. While most declared taxa were not identified, many non-declared taxa were identified. For sample S3, the animal species *B. taurus* and the plants *Cullen* sp. (Fabaceae), *Melilotus officinalis* (Fabaceae), *Medicago* sp. (Fabaceae), *Bupleurum* sp. (Apiaceae), and *Rubus* sp. (Rosaceae) were identified by at least 14 (88%) laboratories (Table 8). Furthermore, the fungi *Aspergillus fumigatus* (Aspergillaceae) and

Fusarium sp. (Nectriaceae) were reproducibly identified, of which the former is also a known human pathogenic fungus. For sample S8, the animal species *B. taurus* and *Homo sapiens*, the plant species *Sanguisorba officinalis* and *Eleutherococcus sessiliflorus*, members of the plant genera *Croton* and *Erythroxylum*, and families Meliaceae and Asteraceae were reproducibly identified (Table 9).

Discussion

In this study, a DNA metabarcoding method was developed using a multi-locus panel of DNA barcodes for the identification of CITES-protected species in highly complex products such as TMs. As a first step, a CTAB DNA isolation method was selected for efficiently extracting high-quality DNA from pure plant and animal reference materials as well as from complex mixtures. DNA isolation can be very difficult to standardize and optimize because of the complexity and diversity of wild life forensic samples, and a more systematic comparison of different DNA extraction methods is required. Second, a single PCR protocol, suitable for all the barcodes included, i.e., multiple universal plant and animal barcode and mini-barcode markers, was identified. This facilitated the design of a multi-locus panel of DNA barcodes. Furthermore, the developed DNA metabarcoding method includes a dedicated bioinformatics workflow, named CITESspeciesDetect, that was specifically developed for the analysis of Illumina paired-end reads. The developed pipeline requires skilled experts in bioinformatics and applies scripts for command-line processing. NGS data analysis pipelines may provide a lot of flexibility to the user as modifications are easily implemented by expert users. The design of the pipeline prevented *cyt b* and COI full-length barcodes from being separated from their corresponding mini-barcodes as they have identical forward primers. Since the 300 PE reads can read through the *cyt b* and COI mini-barcodes, and therefore contain both 5' primer and 3' primer information, separation should be feasible.

To simplify the inter-laboratory validation of the pipeline, a user-friendly and intuitive web-interface with associated “Help” functions and “FAQs” was developed for the CITESspeciesDetect pipeline. The web-interface was, however, not available in the course of the inter-laboratory trial. Therefore, the sequence data generated in the inter-laboratory study could not be analysed by the individual laboratories using the CITESspeciesDetect pipeline. A single (developer) laboratory, therefore, performed these analyses. Upon the availability of the online web-interface, individual participants were later given the opportunity to re-analyse their DNA metabarcoding data. Observations made in this part demonstrated concordance of results with those obtained by the developing laboratory, reinforcing the perception of CITESspeciesDetect as a user-friendly and reliable pipeline that may readily be used by enforcement agencies and other laboratories.

The performance of the DNA metabarcoding method was assessed in an interlaboratory trial in which the method was found to be highly reproducible across laboratories and sensitive enough to identify species present at 1% dry weight content in experimental samples containing up to 11 different species as ingredients. However, not all laboratories could identify all specified ingredients (species) in the analysed experimental samples. From the current study, we demonstrate that diverse animal taxa could be identified at the species level, which highlights the object of the method to target a wide range of

Table 8: Sample S3 ingredients list and taxa (species, genus, family, order) identified.

Ingredients label	Common name	Species/genus	Family	(Infra)order
Herba Cistanches	Cistanche extract	Cistanche sp.	Orobanchaceae	Lamiales
Cauda cervi	Mature deer tail	Cervus sp.	Cervidae	Pecora
Radix Rehmanniae praeparata	Processed <i>Rehmannia</i> root	Rehmanniae sp.	Rehmanniaceae	Lamiales
Radix Ginseng	Dried root of <i>Panax ginseng</i>	<i>Panax ginseng</i>	Araliaceae	Apiales (8)
Radix morindae Officinalis	Morinda root	<i>Morinda officinalis</i>	Rubiaceae	Gentianales
Semen Cuscutae	Chinese dodder seed	<i>Cuscuta</i> sp. (14)	Convolvulaceae (2)	Solanales
Radix Achyranthis bidentatae	Dried root of <i>Achyranthis bidentatae</i>	<i>Achyranthes bidentatae</i>	Amaranthaceae	Caryophyllales
Rhizoma Cibotii	Root of <i>Cibotium barometz</i>	<i>Cibotium barometz</i>	Cibotiaceae	Cyatheales
Semen Platycladi	Dry ripe kernel of <i>Platycladus orientalis</i>	<i>Platycladus orientalis</i>	Cupressaceae	Cupressales
Cortex Eucommiae	Bark of <i>Eucommia ulmoides</i>	<i>Eucommia ulmoides</i>	Eucommiaceae	Garryales
Radix Astragali	Astragalus root	<i>Astragalus danicus</i> (16)	Fabaceae (16)	Fabales
Fructus Schisandrae chinensis	Chinese magnolia-vine fruit	<i>Schisandra chinensis</i>	Schisandraceae	Austrobaileyales
Cortex Cinnamomi	Dried inner bark of <i>Cinnamomum</i> sp.	<i>Cinnamomum</i> sp.	Lauraceae	Laurales
Cornu Cervi Pantotrichum	Antler of <i>Cervus</i> sp.	<i>Cervus</i> sp.	Cervidae	Pecora
Undeclared identified taxa ^a		<i>Bos taurus</i> (16) Cullen sp. (16) <i>Melilotus officinalis</i> (15) <i>Medicago</i> sp. (16) <i>Bupleurum</i> sp. (15) <i>Aspergillus fumigatus</i> (15) <i>Rubus</i> sp. (15) <i>Fusarium</i> sp. (15)		

The number of laboratories that have identified a taxon is provided in parentheses. Species marked in grey are listed by CITES.

^aSpecies identified by at least 14 laboratories that were not mentioned in the ingredients list.

Table 9: Sample S8 ingredients list and taxa (species, genus, family, order) identified.

Ingredients label	Common name	Species/genus	Family	(Infra)order
Kola nut	Fruit of kola nut	<i>Cola</i> sp.	Malvaceae	Malvales
Siberian ginseng	Siberian ginseng	<i>Eleutherococcus senticosus</i>	Araliaceae	Apiales
Horny goat weed	Horny goat weed	<i>Epimedium</i> sp. (16)	Berberidaceae (16)	Ranunculales
Catuaba	Catuaba bark	<i>Calophyllum antillanum</i>	Calophyllaceae	Malpighiales
Muria puama	Marapuama, potency wood	<i>Ptychopetalum</i> sp.	Olacaceae	Santalales
Korean ginseng	Korean ginseng	<i>Panax ginseng</i> (16)	Araliaceae (16)	Apiales
Damiana	Damiana leaves	<i>Turnera diffusa</i>	Passifloraceae	Malpighiales
Saw palmetto	Extract of fruit the of <i>Serenoa repens</i>	<i>Serenoa repens</i>	Arecaceae (16)	Arecales
Yohimbe	Extract from the bark of <i>Pausinystalia johimbe</i>	<i>Pausinystalia johimbe</i>	Rubiaceae (16)	Gentianales
Magnesium stearate				
Undeclared identified taxa ^a		<i>Bos taurus</i> (16) <i>Homo sapiens</i> (15) <i>Eleutherococcus sessiliflorus</i> (16) <i>Croton</i> sp. (16) <i>Erythroxylum</i> sp. (15) <i>Sanguisorba officinalis</i> (15)	Asteraceae (16) Meliaceae (16)	

The number of laboratories that have identified a taxon is provided in parentheses. Species marked in grey are listed by CITES.

^aSpecies identified by at least 14 laboratories that were not mentioned in the ingredients list.

animal species. COI (full-length COI and mini-COI) was found to be the most effective DNA barcode marker for animal species identification. This is not surprising considering that COI is the standard barcode for almost all animal groups [25]. Nearly all animal species identifications were supported by multiple DNA barcodes, thereby giving strong confidence to the correctness of the animal species identifications. In contrast, plants could mainly be identified at the family level, and no single DNA barcode marker was found to provide the best resolution for identifying plant taxa. Ideally, adequate plant species discrimination would require the combined use of multiple DNA barcode markers, e.g., *rbcl* + *matK* [26], but this is technically not possible due to the nature of the target samples (heavily processed) and with

the current Illumina MiSeq technology. For the identification of plant taxa listed by CITES, the use of DNA barcodes with relatively modest discriminatory power at the genus or higher taxonomic level can still be useful as it is often an entire plant genus or family that is listed by CITES, rather than individual plant species. This was the case for, e.g., Orchidaceae and Cactaceae in this study. Yet, for some plant species (e.g., *Aloe variegata*), the resolution provided by the used plant DNA barcodes may still be too low for unambiguous CITES identification. It is important to note that the maximum achievable Illumina NGS read length limits the taxonomic resolution of DNA barcodes that are longer than ~550 nt. This particularly limited the discriminatory power of the full-length plant barcodes *matK* and *rbcl*.

The DNA metabarcoding method may therefore benefit from (currently unavailable) Illumina read lengths longer than 300 nt, or other long-read sequencing technologies. Alternatively, full-length barcodes may be resolved using an advanced bioinformatics strategy (SOAPBarcode) to assemble Illumina shotgun sequences of PCR amplicons [27]. Single barcodes in several cases failed to amplify or provide resolution. The latter is likely to be caused mainly by database incompleteness, lack of genetic variability within some loci/target sequences, and sample composition. However, combining multiple barcodes into a multi-locus metabarcoding method mitigated the problems observed for individual barcodes. A high degree of confidence in the taxonomic assignments based on the combined barcodes was therefore observed, providing for enhanced quality assurance compared to the use of single barcodes.

While the use of well-characterized experimental mixtures allowed for an assessment of the performance of the DNA metabarcoding method under ideal conditions, the amplifiable DNA content of real-life samples encountered in routine diagnostic work is often of an unpredictable and variable quality. An analysis of 2 authentic TM products seized by the Dutch Customs Laboratory demonstrated that few ingredients listed on the labels could be reproducibly identified. This does not mean that the undetected species were not used as ingredients. Ingredients may have been processed in such a way that the DNA is either degraded or effectively removed. This is, e.g., the case with refined oils or cooked ingredients [28]. A PCR-free targeted DNA capturing approach coupled with shotgun sequencing was recently proposed for biodiversity assessments, which may potentially also be suitable for enhancing species identification in difficult wildlife forensic samples [27, 29]. The quality of the sequence reference database also strongly affects the ability to correctly identify species. Without correct references that also exhibit the necessary intraspecific variation, it is not possible to match and discriminate sequence reads correctly. It is well known that accurate DNA barcoding depends on the use of a reference database that provides good taxonomic coverage [6, 10]. The current underrepresentation of DNA barcodes from species protected by CITES and closely related species critically hampers their identification. We estimate that only 18.8% of species on the CITES list contain 1 or more DNA barcodes (COI for animals, and *matK* or *rbcL* for plants). This will improve as DNA barcoding campaigns continue, in particular through initiatives such as the Barcode of Wildlife Project (BWP) [30]. Only by expansion of the sequence reference database of endangered and illegally traded species can DNA barcoding provide the definitiveness required in a court of law.

A noteworthy observation was that most species that were reproducibly identified did not appear on the ingredients lists on the labels of the analysed TMs. This is possibly due to mislabelling. If the identifications are correct, this also indicates that consumption may pose health risks. These findings corroborate earlier reports that DNA metabarcoding may provide valuable information about the quality and safety of TMs [6, 7].

Potential implications

Overall, our findings demonstrate that the multi-locus DNA metabarcoding method assessed in this study can provide reliable and detailed data on the composition of highly complex food products and supplements. This study highlights the necessity of a multi-locus DNA metabarcoding strategy for species identification in complex samples since the use of multiple barcode markers can enable an increased resolution and quality

assurance, even in heavily processed samples. The developed robust bioinformatics pipeline for Illumina data analysis with user-friendly web-interface allows the method to be directly applied in various fields, such as (i) food mislabelling and fraud in the food industry [31], (ii) environmental monitoring of species [32], and (iii) wildlife forensics [33]. Furthermore, the pipeline can be readily used to analyse different types of Illumina paired-end datasets, even the future Illumina datasets (read length > 300 nt). Additionally, the web-interface provides an opportunity for the global audience with limited expertise in bioinformatics to analyse their own data. It also provides the liberty to select different primer sets and customize the settings for the selected purposes. As a result, the range of potential applications of the method to identify plant and animal species is diverse and the pipeline is versatile and adjustable to the user's needs, thus providing a powerful tool for research as well as enforcement purposes.

Methods

Reference materials and preparation of experimental mixtures

All reference specimens were obtained from a local shop in the Netherlands or provided by the Dutch Customs Laboratory (Additional file 1; Table S2 and Table S3). The reference specimens were taxonomically characterized to the finest possible taxonomic level. For each species, it was checked whether reference sequences were present in the NCBI GenBank. For taxonomic confirmation, standard COI barcodes for all animal specimens were generated and individually Sanger sequenced, then compared against the NCBI and BOLD nucleotide database. For plant species, the DNA barcodes *rbcL* and *matK* were Sanger sequenced to confirm species identity. For a number of plant and animal species, the generated barcode sequence information was deposited in the ENA under accession numbers LT009695–LT009705 and LT718651 (Additional file 1; Table S1).

For the initial pilot study, in which the SOP for the DNA metabarcoding approach was established and tested, 15 well-defined complex mixtures were artificially prepared (Table 2). These experimental mixtures were prepared with 2 to 10 taxonomically well-characterized species (Table 2). The ingredients were mixed based on dry weight ratio, for which individual materials were freeze-dried for 78 hours. The lyophilized ingredients were ground using an autoclaved mortar and pestle or blender in a cleaned fume hood, and they were subsequently stored at -20°C . The individual ingredients of each complex mixture were weighted and mixed thoroughly using a tumbler (Heidolph Reax 2) for 20 hours and stored at -20°C until further use.

For the interlaboratory validation trial, in which the applicability and reproducibility of the DNA metabarcoding method was assessed, 8 additional well-characterized mixtures were artificially prepared using the above procedure. These complex mixtures were prepared with 8 to 11 taxonomically well-characterized species present at dry weight concentrations from 1% to 47% (Table 7). These complex mixtures were prepared in such a way that the efficiency of homogenization and possibility of sample cross-contamination could be verified using species-specific qPCR assays. In all samples, 1% of *Zea mays* was added as quality control for homogeneity. The presence of *Z. mays* was checked after sample mixing using maize-specific *hmg* qPCR along with a positive and negative control. A unique species was added at 1% dry weight to each mixture (S1-Glycine max, S2-Gossypium sp., S4-Brassica napus, S5-Triticum aestivum, S6-Beta

vulgaris, S7-*Meleagris gallopavo*, S9-*Carica papaya*, S10-*Solanum lycopersicum*) (Table 7). Species-specific qPCR was performed in duplex (together with positive and negative controls) in all samples to check for possible cross-contamination between samples after sample preparation. Information about the qPCR primers and probes and qPCR procedure can be found in Additional file 1; Tables S8–S10. In addition to the 8 experimental mixtures, 2 TMs were included that were obtained from the Dutch Customs Laboratory: (i) Ma pak leung sea-dog hard capsules (MA PAK LEUNG CO, LTD, Hong Kong) was labelled to contain, among others, rhizoma Cibotii (*Cibotium barometz*, CITES appendix II) and Herba Cistanches (*Cistanche* sp., CITES appendix II); and (ii) Cobra performance enhancer hard capsules (Gold caps, USA) was labelled to contain, among others, Siberian ginseng (*Eleutherococcus senticosus*) and Korean ginseng (*Panax ginseng*). In both TMs, the medicine powder was encapsulated in a hard-capsule shell. All capsules were opened, and the powder inside the capsules was stored in air-sealed and sterilized containers. The powdered medicines were thoroughly mixed using tumbler (Heidolph Reax 2) for 20 hours and stored at -20°C until further use.

DNA isolation method

A cetyltrimethylammonium bromide (CTAB) extraction method [20] was assessed for its ability to efficiently extract DNA from a range of plant and animal materials (SOP). In brief, the CTAB method consists of an initial step to separate polysaccharides and organic soluble molecules using a CTAB extraction buffer (1X CTAB, 1.4 M NaCl, 0.1 M Tris-HCl [pH 8.0], and 20 mM Na_2EDTA) and chloroform. Next, the DNA was precipitated with 96% ethanol and purified with 70% ethanol, and the obtained DNA was stored at 4°C until further use. DNA was extracted from 100-mg reference materials (plant and animal), artificially made complex mixtures, and real-life samples (TMs), along with an extraction control. The concentration and purity ($\text{OD}_{260/280}$ and $\text{OD}_{260/230}$ ratios) of the obtained DNA was determined by spectrophotometer (NanoDrop 1000 instrument, Thermo Fisher Scientific Inc.). $\text{OD}_{260/280}$ ratios between 1.7 and 2.0 were considered to indicate purity of the obtained DNA. In cases where the extraction control contained DNA, the DNA isolation procedure was repeated.

Barcode markers

Candidate universal DNA barcode and mini-barcode markers and primer sets were identified using the information provided in Staats et al. (2016) [10], supplemented with additional primer sets from the literature (Table 1). The PCR primer sets were modified to have an additional Illumina tail sequence at the 5' end of the primers (Table 1).

PCR

A gradient PCR was performed with all PCR primer combinations using 10 ng of DNA. The following PCR conditions were applied: 95°C for 15 minutes, 5 cycles of 94°C for 30 seconds, annealing range ($49\text{--}55^{\circ}\text{C}$) for 40 seconds, and 72°C for 60 seconds, followed by 35 cycles of 94°C for 30 seconds, 54°C for 40 seconds, and 72°C for 60 seconds, with a final extension at 72°C for 10 minutes. The total volume of the PCR mixture was 25 μl , which included 12.5 μl of HotStarTaq Master Mix (Qiagen), 0.5 μl of 10 μM each sense and antisense primer, 7 μl of RNase-free water (Qiagen), and 5 μl of 10 ng/ μl of represented species DNA. PCR was performed in the CFX96 thermal cycler (Bio-Rad) and the amplified products from all the analysed

reference specimens, artificially made complex mixtures, and real-life samples (TMs), together with the positive and negative control reactions, were visualized on 1% agarose gels. If amplification was observed in the negative control, the PCR analysis was repeated. Prior to NGS library preparation, 8 μl of PCR product of each target (12 in total) per sample was pooled and mixed. Next, the pooled PCR products were purified using the QIAquick PCR purification kit (Qiagen) according to manufacturer's protocol, and the purified amplicons were visualized on 1% agarose gels for all the artificially made complex mixtures and real-life samples (TMs).

Next-generation sequencing

The pooled and purified PCR amplicons were sequenced using Illumina MiSeq paired-end 300 technology. Prior to MiSeq sequencing, Index PCR and Illumina library preparation were performed as specified in the Illumina 16S metagenomics sequencing library preparation guide [34]. All the DNA barcode amplicons of each sample were treated as 1 sample during library preparation; i.e., all DNA barcode amplicons of each sample were tagged with the addition of the same unique identifier or index sequence during library preparation. The Index PCR was performed to add dual indices (multiplex identifiers) and Illumina sequencing adapters using the Nextera XT Index Kit (Illumina, FC-131–1001). The prepared Illumina libraries from each sample were quantified using the Quant-iT dsDNA broad range assay (Life Technologies). Furthermore, the normalized library pools were prepared, and their concentration was quantified using the KAPA library quantification kit (KAPA Biosystems) and pooled prior to MiSeq sequencing using the MiSeq reagent kit v. 3.

Bioinformatics analysis

The raw demultiplexed Illumina reads with Illumina 1.8+ encoding were processed using a bioinformatics pipeline called CITESpeciesDetect. The CITESpeciesDetect is composed of 5 linked tools with data analysis passing through 3 phases: (i) preprocessing of paired-end Illumina data involving quality trimming and filtering of reads, followed by sorting by DNA barcode, (ii) OTU clustering by barcode, and (iii) taxonomy prediction and CITES identification (Fig. 1).

During preprocessing of reads, the 5' and 3' Illumina adapter sequences are trimmed using Cutadapt v. 1.9.1 (cutadapt, [RRID:SCR.011841](#)) [35] using the respective substrings TGTGTATAAGAGACAG and CTGTCTCTTATACACA. After Illumina adapter trimming, reads ≤ 10 bp are removed using Cutadapt. Then, the forward and reverse reads are merged to convert a pair into a single pseudo-read containing 1 sequence and 1 set of quality scores using USEARCH v. 8.1.1861 [22].

Next, the merged pseudo-reads, unmerged forward reads, and unmerged reverse reads are processed separately during quality filtering using a sliding window method implemented in PRINSEQ (PRINSEQ, [RRID:SCR.005454](#)) [36]. During this procedure, low-quality bases with Phred scores lower than 20 are trimmed from the 3' end using a window size of 15 nt and a step size of 5 nt. After PRINSEQ, reads with a minimum of 95% per base quality ≥ 20 are kept, while the remaining reads are removed using FASTX.Toolkit v. 0.0.14 [37]. Then, reads are successively selected, trimmed, and sorted per DNA barcode marker using Cutadapt [35]. The following steps are followed for each DNA barcode marker separately during this procedure. First, reads containing an anchored 5' forward primer or anchored 5' reverse primer (or their reverse complement) are selected with a maximum error tolerance of 0.2 (=20%) and with the overlap

parameter specified to 6 to ensure specific selection of reads. Also, reads ≤ 10 nt are removed. The anchored 5' primer sequences are subsequently trimmed. Second, primer sequences that are present at the 3' end of the selected reads are also removed. For each DNA barcode, the primer-selected and unmerged reverse reads are reverse complemented and combined with primer-selected merged and unmerged forward reads.

The following procedure is used to cluster the quality trimmed reads of each DNA barcode into OTUs using the UPARSE pipeline implemented in USEARCH [22] with the following modifications: reads are dereplicated using the `derep_prefix` command. Also, singleton reads and reads with minimum cluster size smaller than 4 are discarded. Representative OTUs are generated using an OTU radius of 2 (98% identity threshold) and 0.2% OTU abundance threshold with minimum barcode length per primer set. Filtering of chimeric reads is performed using the default settings of the UPARSE-REF algorithm implemented in the `cluster_otus` command of USEARCH.

To assign OTUs to taxonomy, standalone BLASTn megablast searches (BLASTN, RRID:SCR.001598) [23] of representative OTUs are performed on the National Centre for Biotechnology Information (NCBI) GenBank nucleotide database using an E-value threshold of 0.001 and a maximum of 20 aligned sequences. OTUs are assigned to the database sequence to which they align, based on bit score, having at least 98% sequence identity and a minimum of 90% query coverage. To identify putative CITES-listed taxa, the taxon ID first was matched against the NCBI taxonomy database using Entrez Direct (`edirect`) functions (available at [38]) to retrieve the scientific name (species, genus, family, order, and synonym name). The scientific, synonym, and/or family names are then matched against a local CITES database that is retrieved from Species+ [39]. The final results are presented as a tab-separated values file containing the BLAST hit metadata (i.e., bit-score, e-value, accession numbers, etc.), the scientific name, synonym name, and in case a CITES-listed taxon was found, also the CITES appendix listing and taxonomic group (i.e., species, genus, family, or order name) under which the taxon is listed by CITES.

The BLAST output was interpreted by following guidelines: first, to minimize the chance of erroneous species identifications, the same species should have at least 3 top hits, i.e., highest bit scores. Second, if multiple hits were obtained with identical quality results, but with different assigned species or with less than 3 top hits with same species designation, the OTU fragment was considered to lack the discriminatory power to refer the hit to species level. In such cases, the OTU would then be downgraded to a genus-level identification. Third, if multiple hits were obtained with identical quality results, but with different assigned genera, the OTU fragment was considered to lack the discriminatory power to describe the hit to genus level. In such cases, the OTU was then be downgraded to a family-level identification. An online web-interface-based application for the CITESpeciesDetect pipeline was developed [17]. The web-interface facilitates intuitive BLAST identification of species listed by speciesplus.net by highlighting species on the CITES appendix I in red. Species listed on CITES appendix II and III are highlighted in orange and yellow, respectively.

Pre-validation in-house of the CITESpeciesDetect pipeline

A parameter scan was performed in order to assess the effect of software settings on the ability to identify species. This evaluation allowed for identification of important parameters

and their effects on the sensitivity, specificity, and robustness of the procedure. This in turn resulted in specified, recommended (default) parameters values for analysing DNA metabarcoding datasets using the CITESpeciesDetect pipeline. The effects of the following parameters were assessed: base quality scores, error tolerance for primer selection, OTU radius, and OTU abundance threshold, expect E-value and query coverage threshold, percentage identity threshold, minimum DNA barcode length, and BLAST database. The parameter scan was performed on experimental mixture 11 of the pilot study (Table 2). This mixture was selected because of its (relatively) high sample complexity, making it the most challenging complex mixture to analyse. Furthermore, the parameter scan was limited to 4 barcode primer sets: full-length cytochrome-B (*cyt b*), COI mini barcode (mini-COI), *rbcL* mini barcode (mini-*rbcL*), and the full-length *rbcL* (*rbcL*) barcode.

Inter-laboratory validation trial: participants and method

To assess the overall performance of the developed DNA metabarcoding approach, 16 laboratories from 11 countries participated in an international inter-laboratory validation. Only laboratories that regularly perform molecular analyses and have well-equipped laboratory facilities were selected to participate (Table 6). The majority are governmental or semi-official institutes and are considered highly authoritative within each respective country. Participants were requested to follow the SOP [24] and were asked to document any deviations that were made. The chemicals and reagents that were provided to the laboratories were 10 samples (8 experimental mixtures and 2 TMs), *B. taurus* and *L. sativa* positive control DNA, CTAB extraction and precipitation buffer, 1.2 M NaCl solution, 12 universal plant and animal barcode and mini-barcode primer sets (Table 1), Qiagen HotStarTaq master mix, and Qiagen PCR purification kits. All reagents and samples were provided in quantities corresponding to $\times 2.5$ the amounts required for the planned experiments. After following the SOP from DNA isolation to purification of the amplified products, all the purified samples from all the laboratories ($n = 160$) were collected and sequenced using Illumina MiSeq paired-end 300 technology (at BaseClear, Leiden, NL, USA). The Index PCR and Illumina library preparation were performed according to the guideline, and all 160 samples were sequenced on 2 Illumina flow cells. After the Illumina MiSeq run, the raw NGS data were processed using the default settings of the CITESpeciesDetect pipeline. BLAST outputs for the samples were distributed back to the participating laboratories for interpretation of results. The laboratories interpreted the BLAST output based on the guideline provided in the SOP.

Availability of supporting data

All the sequence data obtained from the pilot study and the international interlaboratory validation trial, the CITESpeciesDetect pipeline, and access to the web-interface are freely available. The generated barcode sequence information for some animal and plant species was deposited in GenBank under the accession numbers LT009695–LT009705 and LT718651 (Additional file 1; Table S1). The Illumina PE300 MiSeq data obtained from the pilot study and the international interlaboratory validation trial ($n = 177$) were deposited to ENA with study ID PRJEB18620. The script for the CITESpeciesDetect pipeline is available at GitHub.

The web-interface for the CITESpeciesDetect pipeline [17]. The access to analysis via the web-interface will be provided on request. SOP protocols are available from protocols.io [24], and snapshots of the code and example results are available from the GigaScience database, GigaDB [40].

Availability and requirements

Project name: CITESpeciesDetect

Project home page: <https://github.com/RIKILT/CITESpeciesDetect>

Operating system(s): Linux

Programming language: Python and Bash

Other requirements: none

License: BSD 3-Clause License

Any restrictions to use by non-academics: none.

Additional files

Additional file 1: Table S1: Accession numbers of DNA barcode sequences of plant and animal species. Table S2: PCR success rate for animal reference species. Table S3: PCR success rate for plant reference species. Table S4: Statistics of different quality filtering settings for 4 DNA barcodes. Table S5: BLAST identification of species with different quality filtering settings for 4 DNA barcodes. Table S6: Results of species-specific qPCR performed on the experimental mixtures prepared for the interlaboratory validation trial. Table S7: Interlaboratory trial study: average number of Illumina reads per sample, the average number of (pseudo-)reads that passed QC, and the percentage of QC (pseudo-)reads that were assigned to DNA barcodes and OTUs. Table S8: qPCR primer and probe information. Table S9: qPCR reagent composition. Table S10: qPCR thermocycling program (*.docx).

Additional file 2: Table S1: Pilot study: composition of the experimental mixtures and taxa identified using the default settings of the CITESpeciesDetect pipeline. Table S2: Interlaboratory trial: *Beta vulgaris* observed in the sample S6 datasets generated by the 16 laboratories. Table S3–S5: Interlaboratory trial: assessment of the effect of different settings (OTU cluster size, OTU abundance threshold) of the CITESpeciesDetect pipeline on the identification of taxa using different data volumes (low, medium, and high) generated by 3 laboratories for S6. Table S6: Interlaboratory trial: the taxonomic resolution provided by each DNA barcode marker for 8 experimental mixtures (*.xlsx).

Additional file 3: Table S1: ENA accession numbers of all raw NGS datasets obtained in this study (*.xlsx).

Abbreviations

16S rDNA: 16S ribosomal DNA; BLAST: Basic Local Alignment Search Tool; CITES: Convention on International Trade in Endangered Species of Wild Fauna and Flora; COI: cytochrome c oxidase subunit I; CTAB: cetyltrimethylammonium bromide; *cyt b*: cytochrome *b* gene; ITS2: internal transcribed spacer region 2; *matK*: maturase K gene; NGS: next-generation sequencing; OTU: operational taxonomic unit; *rbcL*: ribulose-1,5-bisphosphate carboxylase large subunit gene; SOP: standard operating procedure; TMs: traditional medicines.

Competing interests

The authors declare that they have no competing interests.

Funding

The DECATHLON project has been funded with support from the European Commission in the context of the Seventh Framework Programme (FP7). This publication and all its contents reflect the views only of the authors, and the Commission cannot be held responsible for any use that may be made of the information contained therein.

Author contributions

A.J.A. and M.S. share the first authorship. A.J.A., M.S., M.V., T.P., A.C., and E.K. conceived and designed the experiments for the pilot study. A.J.A. performed the experiments for the pilot study. M.S., R.H., A.J.A. developed the CITESpeciesDetect pipeline. A.J.A., M.S., and R.H. analysed the NGS data obtained from the pilot study. A.J.A., M.S., M.V., T.P., T.W.P., I.S., E.K., F.G., M.T.B.C., and A.H.J. were involved in establishing the SOP for the validation trial. A.J.A., M.S., M.V., T.P., and E.K. conceived and designed the experiments for the validation trial. F.G., M.T.B.C., A.H.J., A.J.A., and M.S. were involved in coordinating the trial. A.J.A. and M.V. prepared the samples and materials for the validation trial and distributed to the participating laboratories. F.R., M.S., and R.H. were involved in developing the web-interface. M.S., T.P., D.D., M.B.I., M.B.U., E.H., R.H.O., A.K., L.L., C.N., H.N., E.P., J.P.R., R.S., T.S., and C.V.M. took part in the validation trial. A.J.A., M.S., R.H., and M.V. analysed the NGS data obtained from the validation trial. A.J.A., M.S., R.H., M.V., S.V.R., and E.K. contributed to the writing of the manuscript. All authors read and approved the final manuscript.

Acknowledgements

This work was supported by the DECATHLON project, which was funded by the European Commission under the Seventh Framework Programme (FP7).

References

1. Chang C, Jang-Liaw N, Lin Y et al. Authenticating the use of dried seahorses in the traditional Chinese medicine market in Taiwan using molecular forensics. *J Food Drug Anal* 2013;21:310–6.
2. Lee SY, Ng WL, Mahat MN et al. DNA barcoding of the endangered aquilaria (*Thymelaeaceae*) and its application in species authentication of agarwood products traded in the market. *PLoS One* 2016;11:e0154631.
3. Milner-Gulland EJ, Bukreeva OM, Coulson T et al. Conservation: reproductive collapse in saiga antelope harems. *Nature* 2003;422:135.
4. CITES. www.cites.org. Accessed 12 September 2017.
5. Cheng X, Su X, Chen X et al. Biological ingredient analysis of traditional Chinese medicine preparation based on high-throughput sequencing: the story for Liuwei Dihuang Wan. *Sci Rep* 2015;4:5147.
6. Coghlan ML, Haile J, Houston J 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.

7. Coghlan ML, Maker G, Crighton E et al. Combined DNA, toxicological and heavy metal analyses provides an auditing toolkit to improve pharmacovigilance of traditional Chinese medicine (TCM). *Sci Rep* 2015;5:17475.
8. Ivanova NV, Kuzmina ML, Braukmann TWA et al. Authentication of herbal supplements using next-generation sequencing. *PLoS One* 2016;11:e0156426.
9. Taberlet P, Coissac E, Pompanon F et al. Towards next-generation biodiversity assessment using DNA metabarcoding. *Mol Ecol* 2012;21:2045–50.
10. Staats M, Arulandhu AJ, Gravendeel B et al. Advances in DNA metabarcoding for food and wildlife forensic species identification. *Anal Bioanal Chem* 2016;408:4615–30.
11. Fahner NA, Shokralla S, Baird DJ et al. Large-scale monitoring of plants through environmental DNA metabarcoding of soil: recovery, resolution, and annotation of four DNA markers. *PLoS One* 2016;11:e0157505.
12. Arulandhu AJ, Staats M, Peelen T et al. DNA metabarcoding of endangered plant and animal species in seized forensic samples. *Genome* 2015;188–9.
13. Taylor HR, Harris WE. An emergent science on the brink of irrelevance: a review of the past 8 years of DNA barcoding. *Mol Ecol Resour* 2012;12:377–88.
14. Little DP. A DNA mini-barcode for land plants. *Mol Ecol Resour* 2014;14:437–46.
15. Parveen I, Gafner S, Techen N et al. DNA barcoding for the identification of botanicals in herbal medicine and dietary supplements: strengths and limitations. *Planta Med* 2016;82:1225–35.
16. RIKILT Wageningen University & Research. www.decathlon-project.eu. Accessed 12 September 2017.
17. RIKILT Wageningen University & Research. <http://decathlon-fp7.citespipe-wur.surf-hosted.nl:8080/>. Accessed 12 September 2017.
18. Chen R, Dong J, Cui X et al. DNA based identification of medicinal materials in Chinese patent medicines. *Sci Rep* 2012;2:958.
19. Scholtens I, Laurensse E, Molenaar B et al. Practical experiences with an extended screening strategy for genetically modified organisms (GMOs) in real-life samples. *J Agric Food Chem* 2013;61:9097–109.
20. Murray MG, Thompson WF. Rapid isolation of high molecular weight plant DNA. *Nucl Acids Res* 1980;8:4321–6.
21. Ivanova NV, Zemlak TS, Hanner RH et al. Universal primer cocktails for fish DNA barcoding. *Mol Ecol Notes* 2007;7:544–8.
22. Edgar RC. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 2010;26:2460–1.
23. Altschul S. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl Acids Res* 1997;25:3389–402.
24. Arulandhu AJ, Staats M, Hagelaar R. Development and validation of a multi-locus DNA metabarcoding method to identify endangered species in complex samples SOP. [protocols.io](http://dx.doi.org/10.17504/protocols.io.ixbcfn). 2017. <http://dx.doi.org/10.17504/protocols.io.ixbcfn>.
25. Hebert PDN, Cywinska A, Ball SL et al. Biological identifications through DNA barcodes. *Proc Royal Soc B Biol Sci* 2003;270:313–21.
26. Hollingsworth PM, Forrest LL, Spouge JL et al. A DNA barcode for land plants. *Proc Natl Acad Sci U S A* 2009;106:12794–7.
27. Liu S, Li Y, Lu J et al. SOAP Barcode: revealing arthropod biodiversity through assembly of Illumina shotgun sequences of PCR amplicons. *Methods Ecol Evol* 2013;4:1142–50.
28. Gryson N. Effect of food processing on plant DNA degradation and PCR-based GMO analysis: a review. *Anal Bioanal Chem* 2010;396:2003–22.
29. Tang M, Hardman CJ, Ji Y et al. High-throughput monitoring of wild bee diversity and abundance via mitogenomics. *Methods Ecol Evol* 2015;6:1034–43.
30. CBOL. www.barcodeofwildlife.org. Accessed 12 September 2017.
31. Galimberti A, De Mattia F, Losa A et al. DNA barcoding as a new tool for food traceability. *Food Res Int* 2013;50:55–63.
32. Taberlet P, Coissac E, Hajibabaei M et al. Environmental DNA. *Mol Ecol* 2012;21:1789–93.
33. Iyengar A. Forensic DNA analysis for animal protection and biodiversity conservation: a review. *J Nat Conserv* 2014;22:195–205.
34. 16s Metagenomic Sequencing Library Preparation. Illumina document 15044223. https://support.illumina.com/content/dam/illumina-support/documents/documentation/chemistry_documentation/16s/16s-metagenomic-library-prep-guide-15044223-b.pdf.
35. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J* 2011;17:10–12.
36. Schmieder R, Edwards R. Quality control and preprocessing of metagenomic datasets. *Bioinformatics* 2011;27:863–4.
37. FASTX-Toolkit v. 0.0.14, <http://hannonlab.cshl.edu/fastx-toolkit/>.
38. <ftp://ftp.ncbi.nlm.nih.gov/entrez/entrezdirect/>.
39. UNEP-WCMC. <https://speciesplus.net>. Accessed 12 September 2017.
40. Arulandhu AJ, Staats M, Hagelaar R et al. Supplementary data for “Development and validation of a multi-locus DNA metabarcoding method to identify endangered species in complex samples.” *GigaScience Database* 2017. <http://dx.doi.org/10.5524/100330>.
41. Palumbi S, Martin A, Romano S et al. The simple fool’s guide to PCR, version 2.0. Privately published document compiled by Palumbi S.. Honolulu, HI: Dept. of Zoology, University of Hawaii; 1991:96822.
42. Sarri C, Stamatis C, Sarafidou T et al. A new set of 16S rRNA universal primers for identification of animal species. *Food Control* 2014;43:35–41.
43. Leray M, Yang JY, Meyer CP et al. A new versatile primer set targeting a short fragment of the mitochondrial COI region for metabarcoding metazoan diversity: application for characterizing coral reef fish gut contents. *Front Zool* 2013;10:34.
44. Geller J, Meyer C, Parker M et al. Redesign of PCR primers for mitochondrial cytochrome c oxidase subunit I for marine invertebrates and application in all-taxa biotic surveys. *Mol Ecol Resour* 2013;13:851–61.
45. Parson W, Pegoraro K, Niederstätter H et al. Species identification by means of the cytochrome b gene. *Int J Legal Med* 2000;114:23–28.
46. Fazekas AJ, Kuzmina ML, Newmaster SG et al. DNA barcoding methods for land 39. *Methods Mol Biol* 2012;858:223–52.
47. Cuenoud P, Savolainen V, Chatrou LW et al. Molecular phylogenetics of Caryophyllales based on nuclear 18S rDNA and plastid *rbcL*, *atpB*, and *matK* DNA sequences. *Am J Botany* 2002;89:132–44.

48. Levin RA, Wagner WL, Hoch PC et al. Family-level relationships of Onagraceae based on chloroplast *rbcL* and *ndhF* data. *Am J Botany* 2003;**90**:107–15.
49. Kress WJ, Erickson DL, Shiu S. 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.
50. Fazekas AJ, Burgess KS, Kesanakurti PR et al. Multiple multilocus DNA barcodes from the plastid genome discriminate plant species equally well. *PLoS One* 2008;**3**:e2802.
51. Taberlet P, Coissac E, Pompanon F et al. Power and limitations of the chloroplast *trnL* (UAA) intron for plant DNA barcoding. *Nucl Acids Res* 2007;**35**:e14.
52. Chen S, Yao H, Han J et al. Validation of the ITS2 region as a novel DNA barcode for identifying medicinal plant species. *PLoS One* 2010;**5**:e8613.
53. Sang T, Crawford DJ, Stuessy TF. Chloroplast DNA phylogeny, reticulate evolution, and biogeography of *Paeonia* (Paeoniaceae). *Am J Botany* 1997;**84**:1120–36.
54. Tate JA, Simpson BB. Paraphyly of *Tarasa* (Malvaceae) and diverse origins of the polyploid species. *Syst Botany* 2003;**28**:723–37.
55. Manning J, Boatwright JS, Daru BH et al. A molecular phylogeny and generic classification of Asphodelaceae subfamily Alooideae: a final resolution of the prickly issue of polyphyly in the alooids? *Syst Botany* 2014;**39**:55–74.