

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

Alfred J. Arulandhu <sup>1,2‡</sup>, Martijn Staats <sup>1‡</sup>, Rico Hagelaar <sup>1</sup>, Marleen M. Voorhuijzen <sup>1</sup>, Theo W. Prins <sup>1</sup>, Ingrid Scholtens <sup>1</sup>, Adalberto Costessi <sup>3</sup>, Danny Duijsings <sup>3</sup>, François Rechenmann <sup>4</sup>, Frédéric B. Gaspar <sup>5</sup>, Maria Teresa Barreto Crespo <sup>5</sup>, Arne Holst-Jensen <sup>6</sup>, Matthew Birck <sup>7</sup>, Malcolm Burns <sup>8</sup>, [Edward Haynes](#)~~Hez Hird~~ <sup>9</sup>, Rupert Hochegger <sup>10</sup>, Alexander Klingl <sup>11</sup>, Lisa Lundberg <sup>12</sup>, Chiara Natale <sup>13</sup>, Hauke Niekamp <sup>14</sup>, Elena Perri <sup>15</sup>, Alessandra Barbante <sup>15</sup>, Jean-Philippe Rosec <sup>16</sup>, Ralf Seyfarth <sup>17</sup>, Tereza Sovová <sup>18</sup>, Christoff Van Moorlegem <sup>19</sup>, Saskia van Ruth <sup>1,2</sup>, Tamara Peelen <sup>20</sup> and Esther Kok <sup>1\*</sup>

11 Alfred J. Arulandhu <sup>1</sup> - RIKILT Wageningen University & Research, P.O. Box 230, 6700 AE Wageningen, The  
12 Netherlands – [alfred.arulandhu@wur.nl](mailto:alfred.arulandhu@wur.nl)

13 Alfred J. Arulandhu <sup>2</sup> – Food Quality and Design Group, Wageningen University and Research, P.O. Box 8129,  
14 6700 EV Wageningen, The Netherlands – [alfred.arulandhu@wur.nl](mailto:alfred.arulandhu@wur.nl)

15 Martijn Staats <sup>1</sup> - RIKILT Wageningen University & Research, P.O. Box 230, 6700 AE Wageningen, The  
16 Netherlands – [martijn.staats@wur.nl](mailto:martijn.staats@wur.nl)

17 Rico Hagelaar <sup>1</sup> - RIKILT Wageningen University & Research, P.O. Box 230, 6700 AE Wageningen, The  
18 Netherlands – [rico.hagelaar@wur.nl](mailto:rico.hagelaar@wur.nl)

19 Marleen M. Voorhuijzen <sup>1</sup> - RIKILT Wageningen University & Research, P.O. Box 230, 6700 AE Wageningen,  
20 The Netherlands - [marleen.voorhuijzen@wur.nl](mailto:marleen.voorhuijzen@wur.nl)

21 Theo W. Prins <sup>1</sup> - RIKILT Wageningen University & Research, P.O. Box 230, 6700 AE Wageningen, The  
22 Netherlands - [theo.prins@wur.nl](mailto:theo.prins@wur.nl)

23 Ingrid M.J. Scholtens <sup>1</sup> - RIKILT Wageningen University & Research, P.O. Box 230, 6700 AE Wageningen, The  
24 Netherlands - [ingrid.scholtens@wur.nl](mailto:ingrid.scholtens@wur.nl)

25 Adalberto Costessi <sup>3</sup> - Baseclear B. V, Einsteinweg 5, 2333 CC Leiden, The Netherlands -  
26 [Adalberto.Costessi@baseclear.nl](mailto:Adalberto.Costessi@baseclear.nl)

27 Danny Duijsings <sup>3</sup> - Baseclear B. V, Einsteinweg 5, 2333 CC Leiden, The Netherlands -  
28 [Danny.Duijsings@baseclear.nl](mailto:Danny.Duijsings@baseclear.nl)

29 François Rechenmann <sup>4</sup> - GenoStar Bioinformatics Solutions, 60 rue Lavoisier, 38330 Montbonnot Saint Martin,  
30 France - [rechenmann@genostar.com](mailto:rechenmann@genostar.com)

31 Frédéric B. Gaspar <sup>5</sup> – iBET, Instituto de Biologia Experimental e Tecnológica, Apartado 12, 2780-901 Oeiras,  
32 Portugal - [fgaspar@ibet.pt](mailto:fgaspar@ibet.pt)

33 Maria Teresa Barreto Crespo <sup>5</sup> - iBET, Instituto de Biologia Experimental e Tecnológica, Apartado 12, 2780-901  
34 Oeiras, Portugal - [tcrespo@ibet.pt](mailto:tcrespo@ibet.pt)

35 Arne Holst-Jensen <sup>6</sup> - Norwegian Veterinary Institute, Ullevaalsveien 68, P.O.Box 750 Sentrum, 0106 Oslo,  
36 Norway - [arne.holst-jensen@vetinst.no](mailto:arne.holst-jensen@vetinst.no)

37 Matthew Birck <sup>7</sup> - U.S. Customs and Border Protection Laboratory, 1100 Raymond Blvd Newark, NJ 07102 USA  
38 - [MATTHEW.BIRCK@cbp.dhs.gov](mailto:MATTHEW.BIRCK@cbp.dhs.gov)

39 Malcolm Burns <sup>8</sup> - LGC, Queens Road, Teddington, Middlesex, TW11 0LY, United kingdom -  
40 [Malcolm.Burns@lgcgroup.com](mailto:Malcolm.Burns@lgcgroup.com)

41 [Edward Haynes](mailto:Edward.Haynes@fera.co.uk) ~~Hez Hird~~<sup>9</sup> – Fera, Sand Hutton, York, YO41 1LZ, United Kingdom -  
42 [Edward.Haynes@fera.co.uk](mailto:Edward.Haynes@fera.co.uk) ~~Hez.Hird@fera.co.uk~~

43 Rupert Hochegger <sup>10</sup> - Austrian Agency for Health and Food Safety, Spargelfeldstrasse 191, 1220 Vienna,  
44 Austria - [rupert.hochegger@ages.at](mailto:rupert.hochegger@ages.at)

45 Alexander Klingl <sup>11</sup> – Generalzollverwaltung, Direktion IX, Bildungs- und Wissenschaftszentrum der  
46 Bundesfinanzverwaltung, Dienstort Hamburg, Baumacker 3, D-22523 Hamburg, Germany -  
47 [Alexander.Klingl@bwz.bund.de](mailto:Alexander.Klingl@bwz.bund.de)  
48 Lisa Lundberg <sup>12</sup> - Livsmedelsverket, Att. Lisa Lundberg, Strandbodgatan 4, SE 75323 Uppsala, Sweden -  
49 [lisa.lundberg@slv.se](mailto:lisa.lundberg@slv.se)  
50 Chiara Natale <sup>13</sup> - AGENZIA DELLE DOGANE E DEI MONOPOLI, Laboratori e servizi chimici – Laboratorio  
51 Chimico di Genova, 16126 Genova, Via Rubattino n.6, Italy - [chiara.natale@agenziadogane.it](mailto:chiara.natale@agenziadogane.it)  
52 Hauke Niekamp <sup>14</sup> - Eurofins GeneScan GmbH, Engesserstrasse 4 79108 Freiburg, Germany -  
53 [HaukeNiekamp@eurofins.de](mailto:HaukeNiekamp@eurofins.de)  
54 Elena Perri <sup>15</sup> - CREA-SCS sede di Tavazzano - Laboratorio via Emilia, Km 307, 26838 Tavazzano, Italy -  
55 [elena.perri@crea.gov.it](mailto:elena.perri@crea.gov.it)  
56 Alessandra Barbante <sup>15</sup> - CREA-SCS sede di Tavazzano - Laboratorio via Emilia, Km 307, 26838 Tavazzano,  
57 Italy - [alessandra.barbante@crea.gov.it](mailto:alessandra.barbante@crea.gov.it)  
58 Jean-Philippe Rosec <sup>16</sup> - Service Commun des Laboratoires, Laboratoire de Montpellier, Parc Euromédecine,  
59 205 rue de la Croix Verte, 34196 Montpellier Cedex 5, France - [Jean-Philippe.ROSEC@scl.finances.gouv.fr](mailto:Jean-Philippe.ROSEC@scl.finances.gouv.fr)  
60 Ralf Seyfarth <sup>17</sup> - Biolytix AG, Benkenstrasse 254, 4108 Witterswil, Switzerland - [Ralf.seyfarth@biolytix.ch](mailto:Ralf.seyfarth@biolytix.ch)  
61 Tereza Sovová <sup>18</sup> – Crop Research Institute, Department of Molecular Genetics, Drnovská 507, 161 06 Prague,  
62 Czech Republic - [mail@terezasovova.cz](mailto:mail@terezasovova.cz)  
63 Christoff Van Moorleghem <sup>19</sup> - Laboratory of Customs & Excises, Blijde Inkomststraat 20, B-3000 Leuven,  
64 Belgium - [christoff.vanmoorleghem@minfin.fed.be](mailto:christoff.vanmoorleghem@minfin.fed.be)  
65 Saskia van Ruth <sup>1</sup> - RIKILT Wageningen University & Research, P.O. Box 230, 6700 AE Wageningen, The  
66 Netherlands - [saskia.vanruth@wur.nl](mailto:saskia.vanruth@wur.nl)  
67 Saskia van Ruth <sup>2</sup> - Food Quality and Design Group, Wageningen University and Research, P.O. Box 8129, 6700  
68 EV Wageningen, The Netherlands - [saskia.vanruth@wur.nl](mailto:saskia.vanruth@wur.nl)  
69 Tamara Peelen <sup>20</sup> - Dutch Customs Laboratory, Kingsfordweg 1, 1043 GN, Amsterdam, The Netherlands -  
70 [t.peelen@belastingdienst.nl](mailto:t.peelen@belastingdienst.nl)  
71 Esther Kok <sup>1\*</sup> - RIKILT Wageningen University & Research, P.O. Box 230, 6700 AE Wageningen, The  
72 Netherlands - [esther.kok@wur.nl](mailto:esther.kok@wur.nl)

73  
74  
75 ‡ Alfred J. Arulandhu and Martijn Staats contributed equally to this work.

76  
77  
78  
79 Corresponding author: Esther Kok  
80 e-mail: [esther.kok@wur.nl](mailto:esther.kok@wur.nl)

86 **Abstract (max. 250 words)**

87  
88 **Background:** DNA metabarcoding provides great potential for species identification in complex samples such  
89 as food supplements and traditional medicines. Such a method would aid CITES (the Convention on  
90 International Trade in Endangered Species of Wild Fauna and Flora) enforcement officers to combat wildlife  
91 crime by preventing illegal trade of endangered plant and animal species. The objective of this research was to  
92 develop a multi-locus DNA metabarcoding method for forensic wildlife species identification and to evaluate the  
93 applicability and reproducibility of this approach across different laboratories.

94  
95 **Results:** A DNA metabarcoding method was developed that makes use of 12 DNA barcode markers that have  
96 demonstrated universal applicability across a wide range of plant and animal taxa, and that facilitate the  
97 identification of species in samples containing degraded DNA. The DNA metabarcoding method was developed  
98 based on Illumina MiSeq amplicon sequencing of well-defined experimental mixtures, for which a  
99 bioinformatics pipeline with user-friendly web interface was developed. The performance of the DNA  
100 metabarcoding method was assessed in an international validation trial by 16 laboratories, in which the method  
101 was found to be highly reproducible and sensitive enough to identify species present in a mixture at 1% dry  
102 weight content.

103  
104 **Conclusion:** The advanced multi-locus DNA metabarcoding method assessed in this study provides reliable and  
105 detailed data on the composition of complex food products, including information on the presence of CITES-  
106 listed species. The method can provides improved resolution for species identification, while verifying species  
107 with multiple DNA barcodes contributes to an enhanced quality assurance.

108  
109 **Keywords:** Endangered species, CITES, Traditional medicines, DNA metabarcoding, Customs agencies, COI,  
110 *matK*, *rbcL*, *cyt b*, mini-barcodes.

## 121 **Background**

122  
123 The demand for endangered species as ingredients in traditional medicines (TMs) has become one of the major  
124 threats to the survival of a range of endangered species such as seahorse (*Hippocampus* sp.), agarwood  
125 (*Aquilaria* sp.), and Saiga antelope (*Saiga tatarica*) [1-3]. The Convention on the International Trade in  
126 Endangered Species of Wild Fauna and Flora (CITES) is one of the best supported conservation agreements to  
127 regulate trading of animal and plant species ([www.cites.org](http://www.cites.org)) and thereby conserve biodiversity. Currently,  
128 ~35,000 species are classified and listed by CITES in three categories based on their extinction level (CITES  
129 Appendix I, II and III) by which the trade in endangered species is regulated. The success of CITES is dependent  
130 upon the ability of customs inspectors to recognize and identify components and ingredients derived from  
131 endangered species, for which a wide range of morphological, chromatographic and DNA-based identification  
132 techniques can be applied [4,5].

133         Recent studies have shown the potential of DNA metabarcoding for identifying endangered species in  
134 TMs and other wildlife forensic samples [4-7]. DNA metabarcoding is an approach that combines DNA  
135 barcoding with next-generation sequencing (NGS), which enables sensitive high-throughput multispecies  
136 identification on the basis of DNA extracted from complex samples [8]. DNA metabarcoding uses more or less  
137 universal PCR primers to mass-amplify informative DNA barcode sequences [9, 10]. Subsequently, the obtained  
138 DNA barcodes are sequenced and compared to a DNA sequence reference database from well-characterized  
139 species for taxonomic assignment [8, 10]. The main advantage of DNA metabarcoding over other identification  
140 techniques is that it permits the identification of all animal and plant species within samples that are composed of  
141 multiple ingredients, which would not be possible through morphological means ~~and~~ time-consuming with  
142 traditional DNA barcoding [4-6]. Furthermore, the use of mini-barcode markers in DNA metabarcoding facilitate  
143 the identification of species in highly processed samples containing heavily degraded DNA [5, 6]. Such a  
144 molecular approach could aid the Customs Authorities to identify materials derived from endangered species in a  
145 wide variety of complex samples, such as food supplements and TMs [11].

146         Before routine DNA metabarcoding can be applied, there are some key issues that need to be taken  
147 into account. First, complex products seized by Customs, such as TM products, may contain plant and animal  
148 components that are highly processed, and from which the isolation of good quality DNA is challenging. Second,  
149 the universal DNA barcodes employed may not result in amplification of the related barcode for each species  
150 contained in a complex sample, due to DNA degradation or the lack of PCR primer sequence universality. For  
151 plants, for example, different sets of DNA barcodes have been suggested for different fields of application (i.e.

152 general taxonomic identification of land plants, identification of medicinal plants, etc.), and none of them meet  
1 the true requirements of universal barcodes [12]. Also, whilst PCR primers can be designed to accommodate  
2 153 shorter DNA barcode regions for degraded DNA samples, such mini-barcodes contain less information and their  
3 154 primers are more restrictive, often making them unsuitable for universal species barcoding [4, 13]. The third  
4 155 challenge is the reference sequence database quality and integrity, which is particularly problematic for law  
5 156 enforcement issues, where high quality and reliability are essential. The current underrepresentation of DNA  
6 157 barcodes from species protected under CITES and closely related species critically hampers their identification.  
7 158 The fourth challenge is that a dedicated bioinformatics pipeline is necessary to process raw NGS data for  
8 159 accurate and sensitive identification of CITES-listed species [9]. Finally, studies using the DNA metabarcoding  
9 160 approach are scarce and none of these methods have been truly validated [9, 14]. Therefore, before implementing  
10 161 DNA metabarcoding by Customs and other enforcement agencies, the above-mentioned challenges need to be  
11 162 thoroughly assessed to ensure accurate taxonomic identifications.  
12 163

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

179 The DNA metabarcoding method was developed and tested based on data generated for 15 well-  
180 defined complex mixtures. The use of well-characterised mixtures allowed for optimising the bioinformatics  
181 procedure and subsequent robustness testing of multiple parameter settings and thresholds. The practical

182 performance and reproducibility of the DNA metabarcoding strategy was assessed in an international validation  
183 trial by 16 laboratories from 11 countries, on the basis of eight other newly composed complex mixtures and two  
184 seized TMs, which were suspected to contain ingredients derived from CITES species. In this study, the multi-  
185 locus DNA metabarcoding method is presented and it is assessed whether the method can improve the  
186 compositional analysis of complex and real-life samples by enabling the sensitive and reproducible identification  
187 of CITES-listed taxa by enforcement agencies and other laboratories.

### 189 **Data description**

190 To constitute well-defined complex mixtures, 46 reference specimens were commercially purchased  
191 from shops or were provided by the Dutch Custom Laboratory. In addition, two TMs that were suspected to  
192 comprise endangered species material were also obtained from Dutch Customs Laboratory. Each reference  
193 specimen was identified morphologically. Genomic DNA was extracted from 29 animal and 17 plant reference  
194 species for DNA barcoding. Standard cytochrome c oxidase I (COI) barcodes for all animal specimens were  
195 generated and individually sequenced using the Sanger method, and compared against the Barcode of Life Data  
196 Systems and NCBI database for taxonomic confirmation. For plant species, the DNA barcodes *rbcL* and *matK*  
197 were sequenced to confirm species identity. For a number of plant and animal species the generated barcode  
198 sequence information was deposited in the European Nucleotide Archive (ENA) under accession numbers  
199 LT009695 to LT009705, and LT718651 (Additional file 1; Table S1).

200 The complex mixtures for the pilot study and interlaboratory validation trial were prepared with 2 to 11  
201 taxonomically well-characterised species present in relative concentrations (dry mass: dry mass) from 1% to  
202 47%. For all experimental mixtures in the interlaboratory trial, internal control species were used to verify the  
203 efficiency of homogenization and to check for possible sample cross-contamination using species specific qPCR  
204 assays. DNA was isolated from the complex mixtures and the concentration and purity of extracted DNA was  
205 determined using spectrophotometer (NanoDrop 1000, Thermo Fisher Scientific Inc.). Subsequently, PCR  
206 amplifications using 12 DNA barcode primer sets were performed. The pooled and purified amplicons of each  
207 sample were sequenced using an Illumina MiSeq paired-end 300 technology, following the manufacturer's  
208 instructions (Illumina, Inc.). The NGS datasets were analysed using the CITESpeciesDetect pipeline. ~~that~~  
209 ~~consists of three steps: 1) pre-processing of paired-end Illumina data involving quality trimming and filtering of~~  
210 ~~reads, followed by reads sorting per DNA barcode, 2) Operational Taxonomic Unit (OTU) clustering by DNA~~  
211 ~~barcode, and 3) taxonomy prediction and CITES identification.~~ All raw NGS datasets from both analyses were

12 deposited in ENA under accession numbers ERS1545972 to ERS1545988, ERS1546502 to ERS1546533,  
13 ERS1546540 to ERS1546619, ERS1546624 to ERS1546639, ERS1546742 to ERS1546757, ERS1546759 to  
14 ERS1546774, and study number PRJEB18620 (Additional file 4; Table S1). A web interface was developed for  
15 the CITESpeciesDetect pipeline to allow stakeholders to perform the NGS data analysis of their own samples.  
16 The web interface can be globally accessed via the SURFsara high-performance computing and data  
17 infrastructure (<http://decathlon-fp7.citespipe-wur.surf-hosted.nl:8080/>).

## Analyses

### Establishing a laboratory procedure for multi-locus DNA barcode amplification

18 Based on the previous studies on DNA isolation for TMs [4, 15] and from the comparison between modified  
19 Qiagen DNeasy plant mini kit [16] and CTAB isolation [17] (unpublished results), we identified that the CTAB  
20 isolation method in general yields better DNA purity and provides better PCR amplification success. Therefore,  
21 the CTAB DNA isolation method was selected for successive experiments.

22 The DNA barcode markers included in this study were selected based on Staats et al. [9] supplemented  
23 with additional primers from literature [13] (Table 1). DNA barcode markers were selected based on the  
24 availability of universal primer sets and DNA sequence information in public repositories [9]. Important  
25 considerations in selecting suitable primer sets were that, preferably, they are used in DNA barcoding campaigns  
26 and studies, and as such have demonstrated universal applicability across a wide range of taxa. Furthermore,  
27 primer sets for both the amplification of full-length barcodes and their respective mini-barcodes (i.e. short  
28 barcode regions < 300 nt within existing ones) were selected when available. This was done to facilitate PCR  
29 amplification from a range of wildlife forensic samples containing relatively intact DNA (using full-length  
30 barcodes) and/or degraded DNA (mini-barcodes). Based on these criteria, PCR primer sets for the following  
31 animal DNA barcodes were selected: regions of the mitochondrial genes encoding 16S rRNA gene (16S),  
32 cytochrome c oxidase I (COI) and cytochrome *b* (*cyt b*). For plant species identification, primer sets for the  
33 following DNA barcodes were selected: regions of the plastidial genes encoding maturase K (*matK*), ribulose-  
34 1,5-bisphosphate carboxylase (*rbcL*), tRNA<sup>Leu</sup> (UAA) intron sequence (*trnL* (UAA)), *psbA-trnH* intergenic  
35 spacer region (*psbA-trnH*), and the nuclear internal transcribed spacer 2 (ITS2) region (Table 1). The selected  
36 primer sets were modified to include the Illumina adapter sequence at the 5' end of the locus-specific sequence  
37 to facilitate efficient NGS library preparation. A gradient PCR experiment was performed to identify the optimal  
38 PCR annealing temperature. While the selected PCR primer sets had previously been published with their own  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65



243 annealing temperatures and conditions, the identification of a single optimal annealing temperature for all PCR  
244 primer sets would allow for increased efficiency of analysis. Initially, a thermal gradient of 49.0 °C to 55.0 °C  
245 was tested on the *Bos taurus* reference material with the primer sets for COI-2, 16S, mini-16S, and *cyt b*. The  
246 amplification efficiency across the PCR primers sets was determined by comparing the intensity of the  
247 amplicons across the thermal gradient. An optimal annealing temperature of 49.5 °C was identified, but  
248 additional non-specific amplicons were observed with some primers (not shown). To reduce the amounts of non-  
249 specific amplification products, the PCR program was modified to increase the annealing temperature after five  
250 cycles from 49.5 °C to 54.0 °C [18], and tested on all 15 PCR primer sets (Table 1). It was observed that certain  
251 PCR primer combinations still produced non-specific products (for *psbA-trnH* gene) or less intense PCR  
252 products (for *rbcL* gene with primers *rbcLa-F* and *rbcLajf634R*, and *matK* gene with primers *matK-390f* and  
253 *matK-1326r*). Consequently, these PCR primer sets were excluded from subsequent experiments.

254           Next, the selected PCR thermocycling protocol was evaluated with the remaining 12 PCR primer sets  
255 on a panel of 29 animal and 17 plant species, representing a phylogenetically wide range of taxa (Mammalia,  
256 Actinopterygii, Malacostraca, Bivalvia, Aves, Reptilia, Amphibia, Insecta, Angiospermae, and Cycadopsida;  
257 Additional file 1; Table S2 and S3). The overall PCR amplification success rates varied across reference species  
258 and across DNA barcode markers (Additional file 1; Table S2). For instance, no PCR amplification was  
259 observed with *cyt b* for the CITES-listed species *Balaenoptera physalus*, whereas intense amplification was seen  
260 for the same species with 16S, COI-2, mini-16S and mini-COI (Additional file 1; Table S2). Overall, at least one  
261 DNA barcode marker could successfully be amplified for each of the 46 plant and animal species (Additional file  
262 1; Table S2 and S3). For a number of plant and animal species the generated barcode sequence information was  
263 deposited in the European Nucleotide Archive (ENA) under accession numbers LT009695 to LT009705, and  
264 LT718651 (Additional file 1; Table S1).

265

### 266 **Development and pre-validation of the CITESspeciesDetect bioinformatics pipeline**

267 A dedicated bioinformatics pipeline, named CITESspeciesDetect, was developed for the purpose of rapid  
268 identification of CITES-listed species using Illumina paired-end sequencing technology. Illumina technology  
269 was selected because it produces NGS data with very low error rates, compared to other technologies [2, 19].  
270 Furthermore, the Illumina MiSeq platform enables paired-end read lengths of up to 300 nt, allowing relatively  
271 long DNA barcode regions of up to ~550 nt to be assembled. Also, the multiplexing capabilities of Illumina  
272 technology are well developed, allowing for simultaneous sequencing of multiple samples in one run, thereby

273 enabling more cost-efficient NGS. While NGS data analysis pipelines exist that allow processing of Illumina  
274 DNA metabarcoding datasets (e.g. CLOTU, QIIME, Mothur), the majority have been developed for specifically  
275 studying microbial communities using the 16S rRNA gene region. CITESpeciesDetect, developed in this study,  
276 extends on the frequently-used software tools developed within the USEARCH [19] and BLAST+ packages  
277 [20], and additionally includes dedicated steps for quality filtering, sorting of reads per barcode, and CITES  
278 species identification (Figure 1). The CITESpeciesDetect is composed of five linked tools and data analysis  
279 passes through three phases: 1) pre-processing of paired-end Illumina data involving quality trimming and  
280 filtering of reads, followed by sorting by DNA barcode, 2) Operational Taxonomic Unit (OTU) clustering by  
281 barcode, and 3) taxonomy prediction and CITES identification.

~~In establishing the pipeline, it~~ was found that with the current setup of the pipeline, reads generated for  
282 *cyt b* and mini-*cyt b* could not be separated based on the forward PCR primer, as the forward primers are  
283 identical. It was therefore decided to combine (pool) the overlapping reads of *cyt b* and mini-*cyt b* during pre-  
284 processing (primer selection) of reads to prevent reads from being double selected. This means that the results of  
285 *cyt b* and mini-*cyt b* are presented by the CITESpeciesDetect pipeline as *cyt b*. The same issue was found for  
286 COI barcode and mini-barcode markers, for which the results are presented as COI-2.

A parameter scan was performed in order to assess the effect of software settings on the ability to  
288 identify species. The evaluation allowed for the identification of important parameters and their effect on the  
289 sensitivity, specificity and robustness of the procedure. Changing the base quality score has a major impact on  
290 the number of reads per barcode (Additional file 1; Table S4). Increasing the strictness of the base quality score  
291 resulted in decreasing numbers of reads per barcode. Quality score values other than the default values (Q20 for  
292 95% of bases) did not yield better identifications. When applying strict quality filtering settings (Q20 for 100%  
293 of bases, or Q30 for 99% of bases) the species *Pieris brassicae* and *Anguilla anguilla* could not be detected with  
294 *cyt b* and/or mini-COI, indicating these settings were too strict (Additional file 1; Table S5). This is likely due to  
295 the resulting overall low read numbers for *cyt b* and mini-COI when applying these strict quality filtering  
296 settings (Additional file 1; Table S4).

~~The effect of error tolerance on Illumina adapter trimming and primer selection was assessed by varying  
298 the maximum number of errors allowed in assigning reads to DNA barcodes. Setting higher error tolerances  
299 resulted in slightly higher number of reads being selected per DNA barcode marker (not shown). With 0% error  
300 tolerance, however, reads were observed that still contained untrimmed Illumina adapters or primer residues.  
301 These untrimmed residues were not observed when applying a 0.2% error tolerance (not shown).~~

303 An OTU abundance threshold is generally applied to make DNA metabarcoding less sensitive to  
1  
2 304 (potential) false-positive identifications. False-positives may occur e.g. as contaminants during pre-processing of  
3  
4 305 samples (DNA extraction, PCR) or as cross-contamination during Illumina sequencing. Applying an OTU  
5  
6 306 abundance threshold higher than zero generally results in loss of sensitivity. We have found, however, that  
7  
8 307 applying an OTU abundance threshold of higher than zero may help in reducing noisy identifications and  
9  
10 308 potential false-positive identifications (results not shown). It should be noted that applying filtering thresholds  
11  
12 309 may always lead to false negative or false positive identifications. In this study, an OTU abundance threshold of  
13  
14 310 0.2% was set as default, however, the OTU abundance threshold may need re-evaluation for samples with  
15  
16 311 expected very low species abundances (< 1% dry weight).

18 312 The effect of applying a minimum DNA barcode length revealed that allowing DNA barcodes of  $\geq 10$   
19  
20 313 nt did not lead to additional identification of species, compared with default settings (e.g.  $\geq 200$  nt). Increasing  
21  
22 314 the minimal DNA barcode length to 250 nt, however, resulted in a failure to identify most plant species with  
23  
24 315 mini-*rbcL* and *rbcL*. ~~We recommend~~ implemented using a minimum DNA barcode length of 200 nt, except for  
25  
26 316 DNA barcodes with a basic length shorter than 200 nt, in which case the minimum expected DNA barcode  
27  
28 317 length is set to 100 nt for ITS2, 140 nt for mini-*rbcL*, 10 nt, e.g. in case of and 10 nt for the *trnL* (P6 loop)  
29  
30 318 marker.

32 319 The results of the parameter scan resulted in specifying recommended parameter values (default setting)  
33  
34 320 for analysing DNA metabarcoding datasets using the CITESspeciesDetect pipeline (see Methods section  
35  
36 321 “Bioinformatics analysis”). An online version of the CITESspeciesDetect pipeline with a user-friendly web-  
37  
38 322 interface was developed for skilled analysts with basic, but no expert level knowledge in bioinformatics and is  
39  
40 323 made available via <http://decathlon-fp7.citespipe-wur.surf-hosted.nl:8080/>.

42 324

#### 44 325 **Pilot study to assess the performance of the DNA metabarcoding procedure using experimental mixtures**

46 326 The DNA metabarcoding procedure was assessed in a pilot study, for which 15 complex mixtures (EM1 to  
47  
48 327 EM15) were prepared containing from 2 to 10 taxonomically well-characterised species with DNA barcode  
49  
50 328 reference sequences available in the NCBI reference database (Table 2). The experimental mixtures 10 and 11  
51  
52 329 (EM10 and EM11) were independently analysed twice to verify repeatability of the method (DNA isolation,  
53  
54 330 barcode panel analysis and pooling). Only mixtures were used with well-characterised species (DNA Sanger  
55  
56 331 barcoded and taxonomically verified) ingredients, at known dry weight concentrations, and with high quality  
57

332 DNA that would allow for an assessment of the performance of the DNA metabarcoding method under optimal  
1 conditions.  
2 333

3  
4 334 A total of 2.37 Gb of Illumina MiSeq sequencing data was generated for the 17 complex samples (15  
5 complex mixtures along with the two replicates). On average, 464,648 raw forward and reverse Illumina reads  
6 335 were generated per sample, with minimum and maximum read numbers ranging between 273,104 (mixture  
7 336 EM4) and 723,130 (mixture EM10R; Table 3). During raw data pre-processing with the default settings of the  
8 337 CITESspeciesDetect pipeline, the reads were first quality filtered and overlapping paired-end Illumina reads  
9 338 were merged into pseudo-reads (Figure 1). The samples contained on average 269,099 quality controlled (QC)  
10 339 unmerged (forward and reverse) reads and merged pseudo-reads, collectively named (pseudo)reads. On average  
11 340 88.27% (min = 77.38%, max = 96.26%) of raw reads passed the quality filtering and pre-processing steps,  
12 341 indicating that the overall quality of the Illumina data was high (not shown).  
13  
14 342

15  
16 343 Next, the (pseudo)reads were assigned to DNA barcodes based on PCR primer sequences. On average,  
17 344 96.44% (min = 88.78%, max = 98.21%) of QC pre-processed reads were assigned to DNA barcodes, indicating a  
18 345 high percentage of reads containing the locus-specific DNA barcode primers (Table 3). After this, the  
19 346 (pseudo)reads were clustered by 98% sequence similarity into OTUs. On average, 82.26% (min = 75.11%, max  
20 347 = 90.63%) of the DNA barcodes assigned reads were clustered into OTUs (Table 3). It was assumed that the  
21 348 small fraction of reads that was not assigned to OTUs contained non-informative (e.g. non-specific fragments,  
22 349 chimeras) sequences that may have been generated during PCR amplification, and were filtered out during  
23 350 clustering.  
24 351

25  
26 352 For taxonomy prediction, OTUs were assigned to dataset sequences using BLAST when aligning with  
27 353 at least 98% sequence identity, a minimum of 90% query coverage, and an E-value of at least 0.001. Generally,  
28 354 the best match (“top hit”) is used as best estimate of species identity. However, species identification using  
29 355 BLAST requires careful weighting of the evidence. To minimize erroneous taxonomic identifications a more  
30 356 conservative guideline was used that allowed a species to be assigned only when the best three matches  
31 357 identified the species. If the bit scores do not decrease after the top three hits, or if other species have identical  
32 358 bit scores, then identification was considered inconclusive. In such cases, OTUs were assigned to higher  
33 359 taxonomic levels (genus, family or order). All animal ingredients, except *Parapenaepsis* sp. could be identified  
34 360 at the species-level with one or more DNA barcode marker using the default settings of the CITESspeciesDetect  
35 361 pipeline (Table 4 and 5). For plants, *Lactuca sativa* could be identified at the species-level using the *trnL* (P6  
36 loop). All other plant taxa were identified at the genus or higher level (Table 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 the](#) (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 (Figure 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 percentage

392 of QC reads assigned to *matK* were 6.11% (14081 reads) and 0.02% (97 reads) in EM10 and EM10R  
1  
2 393 respectively (Figure 2). The low number of reads assigned to *matK* limited its use for taxonomy identification in  
3  
4 394 EM10R (Table 4). The multi-locus approach, however, allowed for the repeatable identification of taxa in EM10  
5  
6 395 and EM11, though not in all cases with all DNA barcode markers (Table 4 and 5).

7  
8 396 Based on the results obtained from the pilot study, precautions were taken when grinding the freeze-  
9  
10 397 dried materials and subsequent mixing to avoid cross-contamination during the laboratory handling of samples,  
11  
12 398 which were used to improve the SOP for the interlaboratory trial (Additional file 3). Also, control species were  
13  
14 399 added to experimental mixtures that were prepared for the inter-laboratory trial to allow better confirmation of  
15  
16 400 sample homogeneity and to verify that no cross-contamination had occurred during sample preparation.

#### 18 401 19 20 402 **Assessment of interlaboratory reproducibility of the DNA metabarcoding procedure**

21  
22 403 Altogether 16 laboratories from 11 countries (all experienced, well-equipped and proficient in advanced  
23  
24 404 molecular analysis work), including two of the method developers, participated in the inter-laboratory trial  
25  
26 405 (Table 6). The laboratories received ten anonymously labelled samples, each consisting of 250 mg powdered  
27  
28 406 material. Two of the samples, labelled S3 and S8, were authentic TM products seized by the Dutch Customs  
29  
30 407 Laboratory while the other eight samples were well-characterized mixtures of specimens from carefully  
31  
32 408 identified taxa in relative dry weight concentrations from 1% to 47% (Table 7). In all experimental mixtures, 1%  
33  
34 409 of *Zea mays* was added as quality control for homogeneity, which was confirmed with maize-specific *hmg* (high-  
35  
36 410 mobility group gene) qPCR [16]. Also, tests performed with species-specific qPCR assays indicated that cross-  
37  
38 411 contamination did not occur during sample preparation (Additional file 1; Table S6). The qPCR assay for the  
39  
40 412 detection of *Brassica napus*, however, also gave a positive signal for other *Brassica* sp. in the mixtures.

41  
42 413 Together with the sample materials, reagents for DNA extraction, and the complete set of barcode  
43  
44 414 primers, the participants received an obligatory (SOP). Any deviations from the SOP had to be reported. The  
45  
46 415 participants were instructed to extract DNA, perform PCR using the barcode primers, purify the amplified DNA  
47  
48 416 by removal of unincorporated primers and primer dimers, and assess the quality and quantity of the amplification  
49  
50 417 products by gel electrophoresis and UV-spectrophotometry. The purified PCR products were then collected by  
51  
52 418 the coordinator of the trial (RIKILT Wageningen University & Research, the Netherlands) and shipped to a  
53  
54 419 sequencing laboratory (BaseClear, the Netherlands) for Illumina sequencing using MiSeq PE300 technology.  
55  
56 420 The sequencing laboratory performed Index PCR and Illumina library preparation prior to MiSeq sequencing as  
57  
58  
59  
60  
61  
62

421 specified in the Illumina 16S metagenomics sequencing library preparation guide. The altogether 160 PCR  
1  
2 422 samples were sequenced using two Illumina flow cells with MiSeq reagent kit v3.

3  
4 423 The interlaboratory trial should ideally have included the use of the online version of the pipeline, but  
5  
6 424 unfortunately this was not possible due to shortage of time. Therefore, a single (developer) laboratory performed  
7  
8 425 these bioinformatics analyses. The 160 individual samples contained on average 269,057 raw reads, and more  
9  
10 426 than 150,000 reads per sample in 95% of the samples (Additional file 1; Table S7). One sample contained less  
11  
12 427 than 100,000 reads (51,750), which was considered more than sufficient for reliable species identification. After  
13  
14 428 pre-processing, the samples contained on average 142,938 (pseudo)reads. On average 94.66% of the reads (min  
15  
16 429 = 88.12%, max = 98.02%) passed the quality filtering indicating that the overall quality of the sequence data was  
17  
18 430 consistently high across the 160 datasets.

19  
20 431 OTU-clustering at 98% sequence similarity on average assigned 78.14% of the pre-processed and DNA barcode  
21  
22 432 assigned reads into OTUs (Additional file 1; Table S7). Only two samples, both from the same laboratory, had a  
23  
24 433 slightly lower percentage of the (pseudo-)reads assigned to OTUs (66.02% and 66.05%). This indicates that the  
25  
26 434 pipeline correctly removed PCR artefacts in the clustering phase.

27  
28 435 For taxonomy prediction, an OTU would be assigned to a database hit if they aligned with  $\geq 98\%$   
29  
30 436 sequence identity and  $\geq 90\%$  query coverage, and yielded an expect value (E-value) of at least 0.001. The  
31  
32 437 BLAST output of the NGS data was interpreted by participants according to the guidelines in the SOP. Variation  
33  
34 438 was observed among laboratories in interpreting the BLAST output: some laboratories consistently scored the  
35  
36 439 top hits, irrespective of bitscore, while other labs selected all hits belonging to the top three bitscores, or  
37  
38 440 interpreted only the first OTU of each DNA barcode, leading to large differences in identified taxa. Because of  
39  
40 441 these inconsistencies, the BLAST results were re-interpreted by RIKILT Wageningen University & Research  
41  
42 442 following the established guideline as mentioned in the SOP. These re-interpreted data are the data referred to in  
43  
44 443 the following sections.

45  
46 444 With one exception, all taxa mixed in at  $\geq 1\%$  (dry mass: dry mass) were reproducibly identified by at  
47  
48 445 least 13 (81%) laboratories (Table 7). *Beta vulgaris* in sample S6 could only be identified by 4 out of 16 (25%)  
49  
50 446 laboratories. *Beta vulgaris* specific sequences were present in all remaining datasets, but at very low read counts.  
51  
52 447 So these clusters did not fulfil the 0.2% OTU abundance threshold (results not shown). All six animal species  
53  
54 448 could be identified to species level with at least one barcode marker (COI), while only four of the 12 plant  
55  
56 449 species (*Brassica oleracea*, *Carica papaya*, *Gossypium hirsutum*, and *L. sativa*) could be identified to species  
57  
58  
59  
60  
61  
62

450 level (Additional file 2; Table S2). All other plant species were identified at the genus or higher level. For plants,  
1  
2 451 no single barcode marker was best, and the most reliable data were obtained by combining the plant barcodes.

3  
4 452 Three taxa that were misidentified or not intentionally included in the mixtures were reproducibly  
5  
6 453 identified across all laboratories. *Acipenser schrenckii* co-occurred in all samples containing *Huso dauricus*. We  
7  
8 454 have confirmed with DNA metabarcoding that the caviar used for preparing the experimental mixtures contains  
9  
10 455 both *H. dauricus* and *A. schrenckii* (results not shown). Furthermore, *Brassica rapa* was identified by ITS2 in  
11  
12 456 sample S4 by all 16 (100%) laboratories, instead of *Brassica napus*. We confirmed by Sanger sequencing *rbcL*  
13  
14 457 and *matK* that our reference specimen is indeed *Brassica napus*, but that its ITS2 sequence is identical to  
15  
16 458 *Brassica rapa* (LT718651). Finally, a taxon of the plant family Phellinaceae was reproducibly identified (by all  
17  
18 459 laboratories) using the mini-*rbcL* marker in all samples containing *L. sativa* (S6, S7, S9, S10). Species of the  
19  
20 460 family Phellinaceae and *L. sativa* both belong to the order Asterales. The evidence for Phellinaceae was not  
21  
22 461 strong, i.e. the family-level identification was based on a single NCBI reference sequence only (GenBank:  
23  
24 462 X69748). We therefore suspect a misidentification during the interpretation of the BLAST results.

25  
26 463 Taxa that were identified to be the result of possible contaminations were scarcely observed, i.e. these  
27  
28 464 were found in isolated cases and could possibly be explained by cross-sample contamination that may have  
29  
30 465 occurred during any step of sample processing (DNA isolation, PCR, NGS library preparation or NGS). For  
31  
32 466 example, a contamination with *Gossypium* sp. was observed using *trnL* (P6 loop) in sample S1 of one of the  
33  
34 467 participating labs. A total of 6 of such suspected cases of incidental cross-contaminations were observed (not  
35  
36 468 shown).

37  
38 469 For the authentic TMs S3 and S8, it was observed that only few labelled ingredients could reproducibly  
39  
40 470 be identified (Table 8 and 9). For sample S3 (Ma pak leung sea-dog), only the listed ingredients *Cuscuta* sp.  
41  
42 471 (Chinese dodder seed), and *Astragalus danicus* (*Astragalus* root) could be identified. For sample S8 (Cobra  
43  
44 472 performance enhancer), only the listed ingredients *Epimedium* sp. (Horny goat weed; Berberidaceae), *Panax*  
45  
46 473 *ginseng* (Korean ginseng; Araliaceae), and species of the plant families Arecaceae (*Serenoa repens*) and  
47  
48 474 Rubiaceae (*Pausinystalia johimbe*) could be identified. While most declared taxa were not identified, many non-  
49  
50 475 declared taxa were identified. For sample S3, the animal species *B. taurus*, and the plants *Cullen* sp. (Fabaceae),  
51  
52 476 *Melilotus officinalis* (Fabaceae), *Medicago* sp. (Fabaceae), *Bupleurum* sp. (Apiaceae), and *Rubus* sp. (Rosaceae)  
53  
54 477 were identified by at least 14 (88%) laboratories (Table 8). Furthermore, the fungi *Aspergillus fumigatus*  
55  
56 478 (*Aspergillaceae*) and *Fusarium* sp. (*Nectriaceae*) were reproducibly identified, of which the former is also a  
57  
58 479 known human pathogenic fungus. For sample S8, the animal species *B. taurus* and *Homo sapiens*, the plant  
59  
60  
61  
62  
63  
64  
65



480 species *Sanguisorba officinalis* and *Eleutherococcus sessiliflorus*, and members of the plant genera *Croton* and  
481 *Erythroxylum*, and families Meliaceae and Asteraceae, were reproducibly identified (Table 9).

## 482 **Discussion**

483  
484  
485 In this study, a DNA metabarcoding method was developed using a multi-locus panel of DNA barcodes for the  
486 identification of CITES protected species in highly complex products such as TMs. As a first step, ~~we selected~~  
487 ~~an optimal DNA isolation method for complex mixtures consisting of both animal and plant tissues. A~~ a CTAB  
488 ~~DNA~~ isolation method ~~was selected for~~ ~~was found to be the most efficient in obtaining~~ ~~efficiently extracting~~ high  
489 quality DNA from pure plant and animal reference materials as well as from complex mixtures. DNA isolation  
490 can be very difficult to standardise and optimise because of the complexity and diversity of wild life forensic  
491 samples, and a more systematic comparison of different DNA extraction methods is required. Secondly, a single  
492 PCR protocol, suitable for all the barcodes included, i.e. multiple universal plant and animal barcode and mini-  
493 barcode markers, was identified. This facilitated the design of a multi-locus panel of DNA barcodes. ~~With this~~  
494 ~~panel, the presence of a species was confirmed with a multiplex marker approach, which improves the resolution~~  
495 ~~for identification and quality assurance.~~ Furthermore, the developed DNA metabarcoding method includes a  
496 dedicated bioinformatics workflow, named CITESspeciesDetect, that was specifically developed for the analysis  
497 of Illumina paired-end reads. The developed pipeline requires skilled experts in bioinformatics, and applies  
498 scripts for command-line processing. NGS data analysis pipelines may provide a lot of flexibility to the user, as  
499 modifications are easily implemented by expert users. The design of the pipeline prevented *cyt b* and COI full-  
500 length barcodes to be separated from their corresponding mini-barcodes, as they have identical forward primers.  
501 Since, the 300 PE reads can read through the *cyt b* and COI mini-barcodes, and therefore contain both 5' primer  
502 and 3' primer information, separation should be feasible.  
503 To simplify the inter-laboratory validation of the pipeline, a user-friendly and intuitive web-interface with  
504 associated “Help” functions and “FAQs” was developed for the CITESspeciesDetect pipeline. The web interface  
505 was, however, not available in the course of the interlaboratory trial. Therefore, the sequence data generated in  
506 the interlaboratory study could not be analysed by the individual laboratories using the CITESspeciesDetect  
507 pipeline. A single (developer) laboratory therefore performed these analyses. Upon the availability of the online  
508 web-interface, individual participants were later given the opportunity to reanalyse their DNA metabarcoding  
509 data. Observations made in this part demonstrated concordance of results with those obtained by the developing  
510 laboratory, reinforcing the perception of CITESspeciesDetect as a user-friendly and reliable pipeline that may  
511 readily be used by enforcement agencies and other laboratories.

512 The performance of the DNA metabarcoding method was assessed in an interlaboratory trial in which  
1  
2 513 the method was found to be highly reproducible across laboratories, and sensitive enough to identify species  
3  
4 514 present at 1% dry weight content in experimental samples containing up to 11 different species as ingredients.  
5  
6 515 However, not all laboratories could identify all [specified ingredients \(species\) in the analysed experimental](#)  
7  
8 516 [samples. From the current study, we demonstrate that diverse animal taxa from phylogenetically unrelated orders](#)  
9  
10 517 could be identified at the species level, [which highlights the object of the in line with the objective that the](#)  
11  
12 518 method ~~to should~~ target [a wide range of all](#) animal species. COI (full-length COI-2 and mini-COI) was found to  
13  
14 519 be the most effective DNA barcode marker for animal species identification. This is not surprising considering  
15  
16 520 that COI is the standard barcode for almost all animal groups [21]. Nearly all animal species identifications were  
17  
18 521 supported by multiple DNA barcodes, thereby giving strong confidence to the correctness of the animal species  
19  
20 522 identifications. In contrast, plants could mainly be identified at the family level, and no single DNA barcode  
21  
22 523 marker was found to provide best resolution for identifying plant taxa. Ideally, adequate plant species  
23  
24 524 discrimination would require the combined use of multiple DNA barcode markers, e.g. *rbcL* + *matK* [22], but  
25  
26 525 this is technically not possible due to the nature of the target samples (heavily processed) and with the current  
27  
28 526 Illumina Miseq technology. For the identification of plant taxa listed by CITES, the use of DNA barcodes with  
29  
30 527 relatively modest discriminatory power at the genus or higher taxonomic level can still be useful, as it is often an  
31  
32 528 entire plant genus or family that is listed by CITES, rather than individual plant species. This was the case for e.g.  
33  
34 529 Orchidaceae and Cactaceae in this study. Yet, for some plant species (e.g. *Aloe variegata*) the resolution  
35  
36 530 provided by the used plant DNA barcodes may still be too low for unambiguous CITES identification. It is  
37  
38 531 important to note that the maximum achievable Illumina NGS read length limits the taxonomic resolution of  
39  
40 532 DNA barcodes that are longer than ~550 nt. This particularly limited the discriminatory power of the full-length  
41  
42 533 plant barcodes *matK* and *rbcL*. The DNA metabarcoding method may therefore benefit from (currently  
43  
44 534 unavailable) Illumina read lengths longer than 300 nt, or other long-read sequencing technologies. [Alternatively,](#)  
45  
46 535 [full-length barcodes may be resolved using an advanced bioinformatics strategy \(SOAPBarcode\) to assemble](#)  
47  
48 536 [Illumina shotgun sequences of PCR amplicons](#) [23]. Single barcodes in several cases failed to amplify or provide  
49  
50 537 resolution. The latter is likely to be caused mainly by database incompleteness, lack of genetic variability within  
51  
52 538 some loci/target sequences, and sample composition. However, combining multiple barcodes into a multi-locus  
53  
54 539 metabarcoding method mitigated the problems observed for individual barcodes. A high degree of confidence in  
55  
56 540 the taxonomic assignments based on the combined barcodes were therefore observed, providing for enhanced  
57  
58 541 quality assurance compared to the use of single barcodes.

542 While the use of well-characterised experimental mixtures allowed for an assessment of the  
1 performance of the DNA metabarcoding method under ideal conditions, the amplifiable DNA content of real-life  
2 543 samples encountered in routine diagnostic work are often of an unpredictable and variable quality. An analysis of  
3 544 two authentic TM products seized by the Dutch Customs Laboratory demonstrated that only few ingredients  
4 545 listed on the labels could be reproducibly identified. This does not mean that the undetected species were not  
5 546 used as ingredients. Ingredients may have been processed in such a way that the DNA is either degraded or  
6 547 effectively removed. This is e.g. the case with refined oils or cooked ingredients [24]. A PCR-free targeted DNA  
7 548 capturing approach coupled with shotgun sequencing was recently proposed for biodiversity assessments which  
8 549 may potentially also be suitable for enhancing species identification in difficult wildlife forensic samples [23,  
9 550 25]. The quality of the sequence reference database also strongly affects the ability to correctly identify species.  
10 551 Without correct references that also exhibit the necessary intraspecific variation, it is not possible to match and  
11 552 discriminate sequence reads correctly. It is well-known that accurate DNA barcoding depends on the use of a  
12 553 reference database that provides good taxonomic coverage [5, 9]. The current underrepresentation of DNA  
13 554 barcodes from species protected by CITES and closely related species critically hampers their identification. We  
14 555 estimate that only 18.8% of species on the CITES list contain one or more DNA barcodes (COI for animals, and  
15 556 matK or rbcL for plants). This will improve as DNA barcoding campaigns continue, in particular through  
16 557 initiatives such as the Barcode of Wildlife Project (BWP; [www.barcodeofwildlife.org](http://www.barcodeofwildlife.org)). Only by expansion of the  
17 558 sequence reference database of endangered and illegally-traded species can DNA barcoding provide the  
18 559 definitiveness required in a court of law.

561 A noteworthy observation was that most species that were reproducibly identified did not appear on the  
562 ingredients lists on the labels of the analysed TMs. This is possibly due to mislabelling. If the identifications are  
563 correct this also indicates that consumption may pose health risks. These findings corroborate earlier reports that  
564 DNA metabarcoding may provide valuable information about the quality and safety of TMs [5, 6].

565

## 566 **Potential implications**

567

568 Overall, our findings demonstrate that the multi-locus DNA metabarcoding method assessed in this study can  
569 provide reliable and detailed data on the composition of highly complex food products and supplements. This  
570 study highlights the necessity of a multi-locus DNA metabarcoding strategy for species identification in complex  
571 samples, since the use of multiple barcode markers can ~~enables~~ an increased resolution and quality assurance,  
572 even in heavily processed samples. The developed robust bioinformatics pipeline for Illumina data analysis with

573 user-friendly web interface allows the method to be directly applied in various fields such as: a) food  
1 574 mislabelling and fraud in the food industry [26], b) environmental monitoring of species [27], and c) wildlife  
2 575 forensics [28]. Furthermore, the pipeline can be readily used to analyse different types of Illumina paired-end  
3 576 datasets, even the future Illumina datasets (read length > 300 nt). Additionally, the web interface provides an  
4 577 opportunity for the global audience with limited expertise in bioinformatics, to analyse their own data. It also  
5 578 provides the liberty to select different primer sets and customise the settings for the selected purposes. As a result,  
6 579 the range of potential applications of the method to identify plant and animal species is diverse, the pipeline is  
7 580 versatile and adjustable to the user's needs, thus providing a powerful tool for research as well as enforcement  
8 581 purposes.  
9

## 582 **Methods**

### 583 **Reference materials and preparation of experimental mixtures**

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

593 For the initial pilot study, in which the SOP for the DNA metabarcoding approach was established and  
594 tested, 15 well-defined complex mixtures were artificially prepared (Table 2). These experimental mixtures were  
595 prepared with 2 to 10 taxonomically well-characterised species (Table 2). The ingredients were mixed based on  
596 dry weight ratio, for which individual materials were freeze-dried for 78 hours. The lyophilized ingredients were  
597 ground using an autoclaved mortar and pestle or blender in a cleaned fume hood, and subsequently stored at -  
598 20 °C °C. The individual ingredients of each complex mixture were weighted and mixed thoroughly using a  
599 tumbler (Heidolph Reax 2) for 20 hours and stored at -20 °C until further use.

600 For the interlaboratory validation trial, in which the applicability and reproducibility of the DNA  
601 metabarcoding method was assessed, eight additional well-characterised mixtures were artificially prepared  
602

605 using the above procedure. These complex mixtures were prepared with 8 to 11 taxonomically well-  
1 606 characterised species present at dry weight concentrations from 1% to 47% (Table 7). These complex mixtures  
2 607 were prepared in such a way that the efficiency of homogenization and possibility of sample cross-contamination  
3 608 could be verified using species-specific qPCR assays. In all samples, 1% of *Zea mays* was added as quality  
4 609 control for homogeneity. The presence of *Z. mays* was checked after sample mixing using maize-specific *hmg*  
5 610 qPCR along with a positive and negative control. A unique species was added at 1% dry weight to each mixture  
6 611 (*S1-Glycine max*, *S2-Gossypium sp.*, *S4-Brassica napus*, *S5-Triticum aestivum*, *S6-Beta vulgaris*, *S7-Meleagris*  
7 612 *gallopavo*, *S9-Carica papaya*, *S10-Solanum lycopersicum*) (Table 7). Species-specific qPCR was performed in  
8 613 duplex (together with positive and negative controls) in all samples, to check for possible cross-contamination  
9 614 between samples after sample preparation. Information about the qPCR primers and probes, and qPCR  
10 615 procedure can be found in the Additional file 1; Table S8-S10. In addition to the eight experimental mixtures,  
11 616 two TMs were included that were obtained from the Dutch Customs Laboratory: a) Ma pak leung sea-dog hard  
12 617 capsules (MA PAK LEUNG CO, LTD, Hong Kong), was labelled to contain among others rhizoma *Cibotii*  
13 618 (*Cibotium barometz*, CITES appendix II), and Herba *Cistanche* (*Cistanche sp.*, CITES appendix II) and b)  
14 619 Cobra performance enhancer hard capsules (Gold caps, USA), was labelled to contain among others Siberian  
15 620 ginseng (*Eleutherococcus senticosus*) and Korean ginseng (*Panax ginseng*). In both TMs, the medicine powder  
16 621 was encapsulated in a hard-capsule shell. All capsules were opened and the powder inside the capsules were  
17 622 stored in air-sealed and sterilized containers. The powdered medicines were thoroughly mixed using tumbler  
18 623 (Heidolph Reax 2) for 20 hours and stored at -20 °C until further use.  
19 624  
20 625  
21 626  
22 627

#### 627 **DNA isolation method**

628 A cetyltrimethylammonium bromide (CTAB) extraction method [17] was assessed for its ability to efficiently  
629 extract DNA from a range of plant and animal materials (Additional file 3). In brief, the CTAB method consists  
630 of an initial step to separate polysaccharides and organic soluble molecules using a CTAB extraction buffer (1X  
631 CTAB, 1.4M NaCl, 0.1 M Tris-HCl [pH 8.0], and 20mM NA<sub>2</sub>EDTA) and chloroform. Next, the DNA was  
632 precipitated with 96% ethanol, purified with 70% ethanol, and the obtained DNA was stored at 4 °C until further  
633 use. DNA was extracted from 100 mg reference materials (plant and animal), artificially made complex mixtures,  
634 and real-life samples (TMs) [along with an extraction control](#). The concentration and purity (OD<sub>260/280</sub> and  
635 OD<sub>260/230</sub> ratios) of the obtained DNA was determined by spectrophotometer (NanoDrop 1000 instrument,  
60  
61  
62  
63  
64  
65

636 Thermo Fisher Scientific Inc.). The OD<sub>260/280</sub> ratios between 1.7 and 2.0 were considered to indicate purity of the  
1  
2 | 637 obtained DNA. In case the extraction control contained DNA, the DNA isolation procedure was repeated.  
3

4 638

### 639 **Barcode markers**

7  
8 640 Candidate universal DNA barcode and mini-barcode markers and primer sets were identified using the  
9  
10 641 information provided in Staats et al. (2016) [9], supplemented with additional primer sets from literature (Table  
11  
12 642 1). The PCR primer sets were modified to have an additional Illumina tail sequence at 5' end of the primers  
13  
14 643 (Table 1).

15  
16 644

### 18 645 **PCR**

19  
20 646 A gradient PCR was performed with all PCR primer combinations using 10 ng of DNA. The tested PCR  
21  
22 647 conditions programme were according to the following protocol: 95 °C for 15 min, five cycles of 94 °C for 30 s,  
23  
24 648 annealing range (49-55 °C) for 40 s, and 72 °C for 60 s, followed by 35 cycles of 94 °C for 30 s, 54 °C for 40 s,  
25  
26 649 and 72 °C for 60 s, with a final extension at 72 °C for 10 min. The total volume of the PCR mixture was 25 µl,  
27  
28 650 which included 12.5 µl of HotStarTaq Master Mix (Qiagen), 0.5 µl of 10 µM each sense and antisense primer, 7  
29  
30 651 µl of RNase-free water (Qiagen) and 5 µl of 10 ng/µl of represented species DNA. PCR was performed in the  
31  
32 652 CFX96 thermal cycler (Bio-Rad) and the amplified products from all the analysed reference specimens,  
33  
34 | 653 artificially made complex mixtures, and real-life samples (TMs) together with the positive and negative control  
35  
36 | 654 reactions were visualised on 1% agarose gels. If amplification was observed in the negative control, the PCR  
37  
38 | 655 analysis was repeated. Prior to NGS library preparation, 8 µl of PCR product of each target (12 in total) per  
39  
40 656 sample was pooled and mixed. Next, the pooled PCR products were purified using the QIAquick PCR  
41  
42 657 purification kit (Qiagen) according to manufacturer's protocol, and the purified amplicons were visualized on 1%  
43  
44 | 658 agarose gels for all the artificially made complex mixtures, and real-life samples (TMs).  
45

46 659

47  
48 660

### 50 661 **Next Generation Sequencing**

51  
52 662 The pooled and purified PCR amplicons were sequenced using Illumina MiSeq paired-end 300 technology. Prior  
53  
54 663 to MiSeq sequencing, Index PCR and Illumina library preparation were performed as specified in the Illumina  
55  
56 664 16S metagenomics sequencing library preparation guide (Illumina document 15044223). All the DNA barcode  
57  
58 665 amplicons of each sample were treated as one sample during library preparation i.e. all DNA barcode amplicons  
59  
60  
61  
62

666 of each sample were tagged with the addition of the same, unique identifier, or index sequence, during library  
1  
2 667 preparation. The Index PCR was performed to add dual indices (multiplex identifiers) and Illumina sequencing  
3  
4 668 adapters using the Nextera XT Index Kit (Illumina, FC-131-1001). [The prepared Illumina libraries from each](#)  
5  
6 669 [sample](#) were quantified [using the Quant-iT dsDNA broad range assay \(Life Technologies\)](#). Furthermore, the  
7  
8 670 [normalised library pools were prepared and their concentration was quantified using KAPA library](#)  
9  
10 671 [quantification kit \(KAPA Biosystems\)](#) and pooled prior to MiSeq sequencing using MiSeq reagent kit v3.

### 672 13 14 673 **Bioinformatics analysis**

15 674  
16 675 The raw demultiplexed Illumina reads with Illumina 1.8+ encoding were processed using a bioinformatics  
17  
18 676 pipeline, called CITESspeciesDetect. The CITESspeciesDetect is composed of five linked tools with data  
19  
20 677 analysis passing through three phases: 1) pre-processing of paired-end Illumina data involving quality trimming  
21  
22 678 and filtering of reads, followed by sorting by DNA barcode, 2) OTU clustering by barcode, and 3) taxonomy  
23  
24 679 prediction and CITES identification (Figure 1).

25  
26 680 During preprocessing of reads, the 5' and 3' Illumina adapter sequences are trimmed using Cutadapt v1.9.1 [29]  
27  
28 681 using the respective substrings TGTGTATAAGAGACAG and CTGTCTCTTATACACA. After Illumina  
29  
30 682 adapter trimming, reads  $\leq 10$  bp are removed using Cutadapt. Then, the forward and reverse reads are merged to  
31  
32 683 convert a pair into a single pseudoread containing one sequence and one set of quality score using USEARCH  
33  
34 684 v8.1.1861 [19].

35  
36 685 Next, the merged pseudo-reads, unmerged forward reads and unmerged reverse reads are processed  
37  
38 686 separately during quality filtering using a sliding window method implemented in PRINSEQ [30]. During this  
39  
40 687 procedure, low quality bases with Phred scores lower than 20 are trimmed from 3'-end using a window size of  
41  
42 688 15 nt and a step size of 5 nt. After PRINSEQ, reads with a minimum of 95% per base quality  $\geq 20$  are kept,  
43  
44 689 while the remaining reads are removed using FASTX\_Toolkit v0.0.14 ([http://hannonlab.cshl.edu/fastx\\_toolkit/](http://hannonlab.cshl.edu/fastx_toolkit/)).  
45  
46 690 Then, reads are successively selected, trimmed and sorted per DNA barcode marker using Cutadapt [29]. The  
47  
48 691 following steps are followed for each DNA barcode marker separately during this procedure. First, reads  
49  
50 692 containing an anchored 5' forward primer or anchored 5' reverse primer (or their reverse complement) are  
51  
52 693 selected with a maximum error tolerance of 0.2 (=20%) and with the overlap parameter specified to 6 to ensure  
53  
54 694 specific selection of reads. Also, reads  $\leq 10$  nt are removed. The anchored 5' primer sequences are subsequently  
55  
56 695 trimmed. Second, primer sequences that are present at the 3' end of the selected reads are also removed. For each  
57  
58  
59  
60  
61  
62



696 DNA barcode, the primer-selected and unmerged reverse reads are reverse complemented and combined with  
1  
2 697 primer-selected merged and unmerged forward reads.

3  
4 698 The following procedure is used to cluster the quality trimmed reads of each DNA barcode into OTUs  
5  
6 699 using the UPARSE pipeline implemented in USEARCH [19] with the following modifications: reads are  
7  
8 700 dereplicated using the derep\_prefix command. Also, singleton reads and reads with minimum cluster size  
9  
10 701 smaller than 4 are discarded. Representative OTUs are generated using an OTU radius of 2 (98% identity  
11  
12 702 threshold) and 0.2% OTU abundance threshold with minimum barcode length per primer set. Filtering of  
13  
14 703 chimeric reads is performed using the default settings of the UPARSE-REF algorithm implemented in the  
15  
16 704 cluster\_otus command of USEARCH.

17  
18 705 To assign OTUs to taxonomy, standalone BLASTn megablast searches [20] of representative OTUs are  
19  
20 706 performed on the National Centre for Biotechnology Information (NCBI) GenBank nucleotide database using an  
21  
22 707 Expectation value (E-value) threshold of 0.001 and a maximum of 20 aligned sequences. OTUs are assigned to  
23  
24 708 the database sequence to which they align, based on bit score, and having at least 98% sequence identity and  
25  
26 709 minimum of 90% query coverage. To identify putative CITES-listed taxa, the taxon ID first was matched against  
27  
28 710 the NCBI taxonomy database using Entrez Direct (edirect) functions (available at  
29  
30 711 <ftp://ftp.ncbi.nlm.nih.gov/entrez/entrezdirect/>) to retrieve scientific name (species, genus, family, order and  
31  
32 712 synonym name). The scientific, synonym and/or family names are then matched against a local CITES database  
33  
34 713 that is retrieved from <https://speciesplus.net>. The final results are presented as a tab-separated values file (TSV)  
35  
36 714 containing the BLAST hit metadata (i.e. bit-score, e-value, accession numbers etc.), the scientific name,  
37  
38 715 synonym name, and in case a CITES-listed taxon was found, also the CITES appendix listing and taxonomic  
39  
40 716 group (i.e. species, genus, family or order name) under which the taxon is listed by CITES.

41  
42 717 The BLAST output was interpreted by following guidelines: first, to minimize the chance of erroneous  
43  
44 718 species identifications, the same species should have at least three top hits, i.e. highest bit scores. Secondly, if  
45  
46 719 multiple hits are obtained with identical quality results, but with different assigned species, or with less than  
47  
48 720 three top hits with same species designation, the OTU fragment was considered to lack the discriminatory power  
49  
50 721 to refer the hit to species level. In such cases, the OTU would then be downgraded to a genus-level identification.  
51  
52 722 Thirdly, if multiple hits are obtained with identical quality results, but with different assigned genera, the OTU  
53  
54 723 fragment lacks the discriminatory power to describe the hit to genus level. In such cases, the OTU would then be  
55  
56 724 downgraded to a family-level identification. An online web-interface based application for the  
57  
58 725 CITESspeciesDetect pipeline was developed which is available from <http://decathlon-fp7.citespipe-wur.surf->  
59  
60  
61  
62



726 [hosted.nl:8080/](https://hosted.nl:8080/). The web-interface facilitates intuitive BLAST identification of species listed by speciesplus.net  
1  
2 by highlighting species on CITES appendix I in red. Species listed on CITES appendix II and II are highlighted  
3  
4 in orange and yellow, respectively.  
5

6 729

### 7 730 **Pre-validation in-house of the CITESspeciesDetect pipeline**

9  
10 731 A parameter scan was performed in order to assess the effect of software settings on the ability to identify  
11  
12 732 species. This evaluation allowed for identification of important parameters and their effects on the sensitivity,  
13  
14 733 specificity and robustness of the procedure. This in turn resulted in specified, recommended (default) parameters  
15  
16 734 values for analysing DNA metabarcoding datasets using the CITESspeciesDetect pipeline. The effects of the  
17  
18 735 following parameters were assessed: base quality scores, error tolerance for primer selection, OTU radius, OTU  
19  
20 736 abundance threshold, expect E-value and query coverage threshold, percentage identity threshold, minimum  
21  
22 737 DNA barcode length and BLAST database. The parameters scan was performed on experimental mixture 11 of  
23  
24 738 the pilot study (Table 2). This mixture was selected because of its (relatively) high sample complexity, making it  
25  
26 739 the most challenging complex mixture to analyse. Furthermore, the parameter scan was limited to four barcode  
27  
28 740 primer sets: full-length cytochrome-B (*cyt b*), COI mini barcode (mini-COI), *rbcL* mini barcode (mini-*rbcL*) and  
29  
30 741 the full-length *rbcL* (*rbcL*) barcode.  
31

32 742

### 34 743 **Inter-laboratory validation trial: participants and method.**

35  
36 744 To assess the overall performance of the developed DNA metabarcoding approach, 16 laboratories from 11  
37  
38 745 countries participated in an international inter-laboratory validation. Only laboratories that regularly perform  
39  
40 746 molecular analyses and have well-equipped laboratory facilities were selected to participate (Table 6). The  
41  
42 747 majority are governmental or semi-official institutes and are considered highly authoritative within each  
43  
44 748 respective country. Participants were requested to follow the SOP (Additional file 3), and were asked to  
45  
46 749 document any deviations that were made. The chemicals and reagents that were provided to the laboratories were:  
47  
48 750 10 samples (eight experimental mixtures and two TMs), *B. taurus* and *L. sativa* positive control DNA, CTAB  
49  
50 751 extraction and precipitation buffer, 1.2 M NaCl solution, 12 universal plant and animal barcode and mini-  
51  
52 752 barcode primer sets (Table 1), Qiagen HotStarTaq master mix, and Qiagen PCR purification kits. All reagents  
53  
54 753 and samples were provided in quantities corresponding to 2.5× the amounts required for the planned experiments.  
55  
56 754 After following the SOP from DNA isolation to purification of the amplified products, all the purified samples  
57  
58 755 from all the laboratories (n=160) were collected and sequenced using Illumina MiSeq paired-end 300 technology  
59  
60  
61  
62

756 (at BaseClear, Leiden, NL). The Index PCR and Illumina library preparation was performed according to the  
1  
2 757 guideline and all 160 samples were sequenced on two Illumina flow cells. After Illumina MiSeq run, the raw  
3  
4 758 NGS data was processed using the default settings of the CITESspeciesDetect pipeline. BLAST outputs for the  
5  
6 759 samples were distributed back to the participating laboratories for interpretation of results. The laboratories  
7  
8 760 interpreted the BLAST output based on the guideline provided in the SOP.  
9

10 761

#### 11 762 **Availability of supporting data**

12 763 All the sequence data obtained from the pilot study and the international interlaboratory validation trial, the  
13  
14  
15 764 CITESspeciesDetect pipeline and access to web interface are freely available. The generated barcode sequence  
16  
17 765 information for some animal and plant species were deposited in GenBank under the accession numbers  
18  
19 766 LT009695 to LT009705, and LT718651 (Additional file 1; Table S1). The Illumina PE300 MiSeq data obtained  
20  
21 767 from the pilot study and the international interlaboratory validation trial (n=177) were deposited to ENA with  
22  
23 768 study ID PRJEB18620. The script for the CITESspeciesDetect pipeline is available at GitHub. The web interface  
24  
25 769 for CITESspeciesDetect pipeline can be accessed via the following link: [http://decathlon-fp7.citespipe-wur.surf-](http://decathlon-fp7.citespipe-wur.surf-hosted.nl:8080/)  
26  
27 770 [hosted.nl:8080/](http://decathlon-fp7.citespipe-wur.surf-hosted.nl:8080/). The access to analysis via the web interface will be provided on request.  
28

29 771

#### 30 772 **Availability and requirements**

31  
32  
33 773 Project name: CITESspeciesDetect

34  
35 774 Project home page: <https://github.com/RIKILT/CITESspeciesDetect>

36  
37  
38 775 Operating system(s): Linux

39  
40 776 Programming language: Python and Bash

41  
42 777 Other requirements: none

43  
44  
45 778 License: BSD 3-Clause License

46  
47 779 Any restrictions to use by non-academics: none  
48

49 780

50 781

51 782

52 783

53 784

54 785

55 786

56

57

58

59

787

## 788 **Additional files**

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

798

799 **Additional file 2: Table S1** Pilot study: Composition of the experimental mixtures, and taxa identified using the  
800 default settings of the CITESpeciesDetect pipeline. **Table S2** Interlaboratory trial: the taxonomic resolution  
801 provided by each DNA barcode marker for eight experimental mixtures (\*.xlsx).

802

803 **Additional file 3:** Standard operating procedure (SOP) for the multi-locus DNA metabarcoding method that was  
804 used in the inter-laboratory validation study (\*.pdf).

805

806 **Additional file 4: Table S1** ENA accession numbers of all raw NGS datasets obtained in this study (\*.xlsx).

807

## 808 **Abbreviations**

809 CITES: Convention on International trade in Endangered Species of Wild fauna and flora; TMs: Traditional  
810 Medicines; NGS: Next generation sequencing; CTAB: cetyltrimethylammonium bromide; COI: Cytochrome c  
811 oxidase subunit I; *cyt b*: Cytochrome *b* gene; 16S rDNA: 16S ribosomal DNA; *matK*: Maturase K gene; *rbcL*:  
812 ribulose-1,5-bisphosphate carboxylase large subunit gene; ITS2: Internal transcribed spacer region 2;; SOP:  
813 Standard operating procedure; OTU: Operational Taxonomic Unit; BLAST: Basic Local Alignment Search Tool.

814

## 815 **Competing interests**

816 The authors declare that they have no competing interest.

817

818

## 819 **Funding**

1  
2 820 The DECATHLON project has been funded with support from the European Commission in the context of the  
3  
4 821 Seventh Framework Programme (FP7). This publication and all its contents reflect the views only of the authors,  
5  
6 822 and the Commission cannot be held responsible for any use, which may be made of the information contained  
7  
8 823 therein.

## 10 824 **Authors' Contributions**

11 825  
12  
13 826 AJA and MS shared the first authorship. AJA, MS, MV, TP, AC, EK conceived and designed the experiments  
14  
15 827 for the pilot study. AJA performed the experiments for the pilot study. MS, RH, AJA developed the  
16  
17 828 CITESpeciesDetect pipeline. AJA, MS, RH analysed the NGS data obtained from the pilot study. AJA, MS,  
18  
19 829 MV, TP, TWP, IS, EK, FG, MTBC, AHJ involved in establishing the Standard Operation Procedure for the  
20  
21 830 validation trial. AJA, MS, MV, TP, EK conceived and designed the experiments for the validation trial. FG,  
22  
23 831 MTBC, AHJ, AJA, MS involved in coordinating the trial. AJA, MV prepared the samples and materials for the  
24  
25 832 validation trial and distributed to the participated laboratories. FR, MS, RH involved in developing the web-  
26  
27 833 interface. MS, TP, DD, MBI, MBU, [EHHH](#), RHO, AK, LL, CN, HN, EP, JPR, RS, TS, CVM took part in the  
28  
29 834 validation trial. AJA, MS, RH, MV analysed the NGS data obtained from the validation trial. AJA, MS, RH, MV,  
30  
31 835 SVR, EK contributed to the writing of the manuscript. All authors read and approved the final manuscript.

## 33 836 **Acknowledgements**

34 837  
35  
36 838 This work was supported by the DECATHLON project, which was funded by the European Commission under  
37  
38 839 Seventh Framework Programme (FP7).

## 40 840 **Reference:**

- 41 841  
42  
43 842  
44 843  
45 844 1. Chang C-H, Jang-Liaw N-H, Lin Y-S, Fang Y-C, Shao K-T: **Authenticating the use of dried**  
46 845 **seahorses in the traditional Chinese medicine market in Taiwan using molecular forensics.**  
47 846 *Journal of Food and Drug Analysis* 2013, **21**:310-316.  
48 847 2. Lee SY, Ng WL, Mahat MN, Nazre M, Mohamed R: **DNA Barcoding of the Endangered Aquilaria**  
49 848 **(Thymelaeaceae) and Its Application in Species Authentication of Agarwood Products Traded in**  
50 849 **the Market.** *PLOS One* 2016, **11**:e0154631.  
51 850 3. Milner-Gulland E, Bukreeva O, Coulson T, Lushchekina A, Kholodova M, Bekenov A, Grachev IA:  
52 851 **Conservation: Reproductive collapse in saiga antelope harems.** *Nature* 2003, **422**:135-135.  
53 852 4. Cheng X, Su X, Chen X, Zhao H, Bo C, Xu J, Bai H, Ning K: **Biological ingredient analysis of**  
54 853 **traditional Chinese medicine preparation based on high-throughput sequencing: the story for**  
55 854 **Liuwei Dihuang Wan.** *Scientific Reports* 2014, **4**: 5147.  
56 855 5. Coghlan ML, Haile J, Houston J, Murray DC, White NE, Moolhuijzen P, Bellgard MI, Bunce M: **Deep**  
57 856 **sequencing of plant and animal DNA contained within traditional Chinese medicines reveals**  
58 857 **legality issues and health safety concerns.** *PLOS Genetics* 2012, **8**:e1002657.  
59 858 6. Coghlan ML, Maker G, Crighton E, Haile J, Murray DC, White NE, Byard RW, Bellgard MI, Mullaney  
60 859 I, Trengove R: **Combined DNA, toxicological and heavy metal analyses provides an auditing**

- 860 **toolkit to improve pharmacovigilance of traditional Chinese medicine (TCM).** *Scientific Reports*  
 1 861 2015, **5**.
- 2 862 7. Ivanova NV, Kuzmina ML, Braukmann TWA, Borisenko AV, Zakharov EV: **Authentication of**  
 3 863 **Herbal Supplements Using Next-Generation Sequencing.** *PLOS One* 2016, **11**:e0156426.
- 4 864 8. Taberlet P, Coissac E, Pompanon F, Brochmann C, Willerslev E: **Towards next-generation**  
 5 865 **biodiversity assessment using DNA metabarcoding.** *Molecular Ecology* 2012, **21**:2045-2050.
- 6 866 9. Staats M, Arulandhu AJ, Gravendeel B, Holst-Jensen A, Scholtens I, Peelen T, Prins TW, Kok E:  
 7 867 **Advances in DNA metabarcoding for food and wildlife forensic species identification.** *Analytical*  
 8 868 *and Bioanalytical Chemistry* 2016:1-16.
- 9 869 10. Fahner NA, Shokralla S, Baird DJ, Hajibabaei M: **Large-scale monitoring of plants through**  
 10 870 **environmental DNA metabarcoding of soil: recovery, resolution, and annotation of four DNA**  
 11 871 **markers.** *PLOS One* 2016, **11**:e0157505.
- 12 872 11. Arulandhu AJ, Staats M, Peelen T, Kok E: **DNA metabarcoding of endangered plant and animal**  
 13 873 **species in seized forensic samples.** In *Genome*. 2015: 188-189.
- 14 874 12. Taylor H, Harris W: **An emergent science on the brink of irrelevance: a review of the past 8 years**  
 15 875 **of DNA barcoding.** *Molecular Ecology Resources* 2012, **12**:377-388.
- 16 876 13. Little DP: **A DNA mini-barcode for land plants.** *Molecular Ecology Resources* 2014, **14**:437-446.
- 17 877 14. Parveen I, Gafner S, Techen N, Murch SJ, Khan IA: **DNA Barcoding for the Identification of**  
 18 878 **Botanicals in Herbal Medicine and Dietary Supplements: Strengths and Limitations.** *Planta*  
 19 879 *Medica* 2016, **82**:1225-1235.
- 20 880 15. Chen R, Dong J, Cui X, Wang W, Yasmeen A, Deng Y, Zeng X, Tang Z: **DNA based identification of**  
 21 881 **medicinal materials in Chinese patent medicines.** *Scientific Reports* 2012, **2**:958.
- 22 882 16. Scholtens I, Laurensse E, Molenaar B, Zaaier S, Gaballo H, Boleij P, Bak A, Kok E: **Practical**  
 23 883 **experiences with an extended screening strategy for genetically modified organisms (GMOs) in**  
 24 884 **real-life samples.** *Journal of agricultural and food chemistry* 2013, **61**:9097-9109.
- 25 885 17. Murray M, Thompson WF: **Rapid isolation of high molecular weight plant DNA.** *Nucleic Acids*  
 26 886 *Research* 1980, **8**:4321-4326.
- 27 887 18. Ivanova NV, Zemlak TS, Hanner RH, Hebert PDN: **Universal primer cocktails for fish DNA**  
 28 888 **barcoding.** *Molecular Ecology Notes* 2007, **7**:544-548.
- 29 889 19. Edgar RC: **Search and clustering orders of magnitude faster than BLAST.** *Bioinformatics* 2010,  
 30 890 **26**:2460-2461.
- 31 891 20. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and**  
 32 892 **PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Research* 1997,  
 33 893 **25**:3389-3402.
- 34 894 21. Hebert PD, Cywinska A, Ball SL, deWaard JR: **Biological identifications through DNA barcodes.**  
 35 895 *Proceedings of the Royal Society of London B: Biological Sciences* 2003, **270**:313-321.
- 36 896 22. Group CPW, Hollingsworth PM, Forrest LL, Spouge JL, Hajibabaei M, Ratnasingham S, van der Bank  
 37 897 M, Chase MW, Cowan RS, Erickson DL: **A DNA barcode for land plants.** *Proceedings of the*  
 38 898 *National Academy of Sciences* 2009, **106**:12794-12797.
- 39 899 23. Liu S, Li Y, Lu J, Su X, Tang M, Zhang R, Zhou L, Zhou C, Yang Q, Ji Y: **SOAPBarcode: revealing**  
 40 900 **arthropod biodiversity through assembly of Illumina shotgun sequences of PCR amplicons.**  
 41 901 *Methods in Ecology and Evolution* 2013, **4**:1142-1150.
- 42 902 24. Gryson N: **Effect of food processing on plant DNA degradation and PCR-based GMO analysis: a**  
 43 903 **review.** *Analytical and Bioanalytical Chemistry* 2010, **396**:2003-2022.
- 44 904 25. Tang M, Hardman CJ, Ji Y, Meng G, Liu S, Tan M, Yang S, Moss ED, Wang J, Yang C: **High -**  
 45 905 **throughput monitoring of wild bee diversity and abundance via mitogenomics.** *Methods in Ecology*  
 46 906 *and Evolution* 2015, **6**:1034-1043.
- 47 907 26. Galimberti A, De Mattia F, Losa A, Bruni I, Federici S, Casiraghi M, Martellos S, Labra M: **DNA**  
 48 908 **barcoding as a new tool for food traceability.** *Food Research International* 2013, **50**:55-63.
- 49 909 27. Taberlet P, Coissac E, Hajibabaei M, Rieseberg LH: **Environmental DNA.** *Molecular Ecology* 2012,  
 50 910 **21**:1789-1793.
- 51 911 28. Iyengar A: **Forensic DNA analysis for animal protection and biodiversity conservation: a review.**  
 52 912 *Journal for Nature Conservation* 2014, **22**:195-205.
- 53 913 29. Martin M: **Cutadapt removes adapter sequences from high-throughput sequencing reads.** *EMBnet*  
 54 914 *journal* 2011, **17**:10-12.
- 55 915 30. Schmieder R, Edwards R: **Quality control and preprocessing of metagenomic datasets.**  
 56 916 *Bioinformatics* 2011, **27**:863-864.
- 57 917 31. Palumbi S, Martin A, Romano S, McMillan W, Stice L, Grabowski G: **The Simple Fool's Guide to**  
 58 918 **PCR, Version 2.0, privately published document compiled by S. Palumbi. Dept. Zoology, Univ**  
 59 919 **Hawaii, Honolulu, HI 1991, 96822.**

- 920 32. Sarri C, Stamatis C, Sarafidou T, Galara I, Godosopoulos V, Kolovos M, Liakou C, Tastsoglou S,  
 1 921 Mamuris Z: **A new set of 16S rRNA universal primers for identification of animal species.** *Food*  
 2 922 *Control* 2014, **43**:35-41.
- 3 923 33. Leray M, Yang JY, Meyer CP, Mills SC, Agudelo N, Ranwez V, Boehm JT, Machida RJ: **A new**  
 4 924 **versatile primer set targeting a short fragment of the mitochondrial COI region for**  
 5 925 **metabarcoding metazoan diversity: application for characterizing coral reef fish gut contents.**  
 6 926 *Front Zool* 2013, **10**:34.
- 7 927 34. Geller J, Meyer C, Parker M, Hawk H: **Redesign of PCR primers for mitochondrial cytochrome c**  
 8 928 **oxidase subunit I for marine invertebrates and application in all-taxa biotic surveys.** *Molecular*  
 9 929 *Ecology Resources* 2013, **13**:851-861.
- 10 930 35. Parson W, Pegoraro K, Niederstätter H, Föger M, Steinlechner M: **Species identification by means of**  
 11 931 **the cytochrome b gene.** *International Journal of Legal Medicine* 2000, **114**:23-28.
- 12 932 36. Fazekas AJ, Kuzmina ML, Newmaster SG, Hollingsworth PM: **DNA barcoding methods for land**  
 13 933 **plants.** *Methods in Molecular Biology* 2012, **858**:223-252.
- 14 934 37. Cuénoud P, Savolainen V, Chatrou LW, Powell M, Grayer RJ, Chase MW: **Molecular phylogenetics**  
 15 935 **of Caryophyllales based on nuclear 18S rDNA and plastid *rbcL*, *atpB*, and *matK* DNA sequences.**  
 16 936 *American Journal of Botany* 2002, **89**:132-144.
- 17 937 38. Levin RA, Wagner WL, Hoch PC, Nepokroeff M, Pires JC, Zimmer EA, Sytsma KJ: **Family-level**  
 18 938 **relationships of Onagraceae based on chloroplast *rbcL* and *ndhF* data.** *American Journal of Botany*  
 19 939 2003, **90**:107-115.
- 20 940 39. Kress WJ, Erickson DL: **A two-locus global DNA barcode for land plants: the coding *rbcL* gene**  
 21 941 **complements the non-coding *trnH-psbA* spacer region.** *PLOS One* 2007, **2**:e508.
- 22 942 40. Fazekas AJ, Burgess KS, Kesanakurti PR, Graham SW, Newmaster SG, Husband BC, Percy DM,  
 23 943 Hajibabaei M, Barrett SC: **Multiple multilocus DNA barcodes from the plastid genome**  
 24 944 **discriminate plant species equally well.** *PLOS One* 2008, **3**:e2802.
- 25 945 41. Taberlet P, Coissac E, Pompanon F, Gielly L, Miquel C, Valentini A, Vermet T, Corthier G,  
 26 946 Brochmann C, Willerslev E: **Power and limitations of the chloroplast *trnL* (UAA) intron for plant**  
 27 947 **DNA barcoding.** *Nucleic Acids Research* 2007, **35**:e14.
- 28 948 42. Chen S, Yao H, Han J, Liu C, Song J, Shi L, Zhu Y, Ma X, Gao T, Pang X: **Validation of the ITS2**  
 29 949 **region as a novel DNA barcode for identifying medicinal plant species.** *PLOS One* 2010, **5**:e8613.
- 30 950 43. Sang T, Crawford D, Stuessy T: **Chloroplast DNA phylogeny, reticulate evolution, and**  
 31 951 **biogeography of *Paeonia* (Paeoniaceae).** *American Journal of Botany* 1997, **84**:1120-1136.
- 32 952 44. Tate JA, Simpson BB: **Paraphyly of *Tarasa* (Malvaceae) and diverse origins of the polyploid**  
 33 953 **species.** *Systematic Botany* 2003, **28**:723-737.
- 34 954 45. Manning J, Boatwright JS, Daru BH, Maurin O, Bank Mvd: **A molecular phylogeny and generic**  
 35 955 **classification of Asphodelaceae subfamily Alooideae: a final resolution of the prickly issue of**  
 36 956 **polyphyly in the alooids?** *Systematic Botany* 2014, **39**:55-74.
- 37 957  
 38 958  
 39 959  
 40 960  
 41 961  
 42 962  
 43 963  
 44 964  
 45 965  
 46 966  
 47 967  
 48 968  
 49 969  
 50 970  
 51 971  
 52 972  
 53 973  
 54 974  
 55 975  
 56 976  
 57 977  
 58  
 59  
 60  
 61  
 62  
 63  
 64  
 65

## Table files:

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

Alfred J. Arulandhu <sup>1,2‡</sup>, Martijn Staats <sup>1‡</sup>, Rico Hagelaar <sup>1</sup>, Marleen M. Voorhuijzen <sup>1</sup>, Theo W. Prins <sup>1</sup>, Ingrid Scholtens <sup>1</sup>, Adalberto Costessi <sup>3</sup>, Danny Duijsings <sup>3</sup>, François Rechenmann <sup>4</sup>, Frédéric B. Gaspar <sup>5</sup>, Maria Teresa Barreto Crespo <sup>5</sup>, Arne Holst-Jensen <sup>6</sup>, Matthew Birck <sup>7</sup>, Malcolm Burns <sup>8</sup>, [Edward Haynes](#) <sup>9</sup>, Rupert Hohegger <sup>10</sup>, Alexander Klingl <sup>11</sup>, Lisa Lundberg <sup>12</sup>, Chiara Natale <sup>13</sup>, Hauke Niekamp <sup>14</sup>, Elena Perri <sup>15</sup>, Alessandra Barbante <sup>15</sup>, Jean-Philippe Rosec <sup>16</sup>, Ralf Seyfarth <sup>17</sup>, Tereza Sovová <sup>18</sup>, Christoff Van Moorlehem <sup>19</sup>, Saskia van Ruth <sup>1,2</sup>, Tamara Peelen <sup>20</sup> and Esther Kok <sup>1\*</sup>

**Table 1: Overview of the PCR primer sets used in this study for amplifying plant and animal DNA barcodes and mini-barcodes. (Line no: 22749)**

DNA Marker	Primer name	Primer sequence 5'-3'	Amplicon length (nt)	Reference
Universal animal DNA barcodes and mini-barcodes				
16S	16sar-L	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCGCCTGTTTATCAAAAACAT	500-600	Palumbi [29]
	16sar-H	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCCGGTCTGAACTCAGATCACGT		
mini-16S	16S-forward	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGAYAAGACGAGAAGACCC	250	Sarri <i>et al.</i> [30]
	16S-reverse	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGATTCGCGCTGTATTCC		
COI*	LepF1_t1	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGATTCAACCAATCATAAAAGATATTGG	648	Modified from Ivanova <i>et al.</i> [18]
	VF1_t1	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTTCTCAACCAACCACAAAGACATTGG		
	VF1d_t1	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTTCTCAACCAACCACAARGAYATYGG		
	VF1i_t1	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTTCTCAACCAACCAIAAIGAIATIGG		
	LepR1_t1	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGTAACTTCTGGATGTCCAAAAAATCA		
	VR1d_t1	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGTAGACTTCTGGGTGGCCRAARAAYCA		
	VR1_t1	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGTAGACTTCTGGGTGGCCAAAGAATCA		
	VR1i_t1	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGTAGACTTCTGGGTGICCIAAIAAICA		
mini-COI	mICOIntF	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGGWACWGGWTGAACWGTWTAYCCYCC	313	Leray <i>et al.</i> [31]., Geller <i>et al.</i> [32]
	jcHCO2198	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGTAIACYTCIGGRTGICRAARAAYCA		
cyt <i>b</i>	L14816	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCCATCCAACATCTCAGCATGATGAAA	743	Palumbi [29], Parson <i>et al.</i> [33]
	CB3-H	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGGCAAATAGGAARTATCATTCC		
mini-cyt <i>b</i>	L14816	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCCATCCAACATCTCAGCATGATGAAA	357	Parson <i>et al.</i> [33]
	H15173	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCCCTCAGAATGATATTTGTCCTCA		
Universal plant DNA barcodes and mini-barcodes				
<i>matK</i>	matK-KIM1R	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGACCCAGTCCATCTGGAATCTTGGTTC	656-889	Fazekas <i>et al.</i> [34]
	matK-KIM3F	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCGTACAGTACTTTTGTGTTTACGAG		
<i>matK</i> <sup>ex</sup>	matK-390f	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCGATCTATTTCATTCAATATTTCC	656-889	Cuénoud <i>et al.</i> [35]

	matK-1326r	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGTCTAGCACACGAAAGTCGAAGT		
<i>rbcL</i>	rbcLa-F	TCGTCCGGCAGCGTCAGATGTGTATAAGAGACAGATGTCACCACAAACAGAGACTAAAGC	654	Levin <i>et al.</i> [36]
	rbcLa-R	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGTAATAAATCAAGTCCACCRCG		Kress and Erickson[37]
<i>rbcL</i> <sup>&amp;</sup>	rbcL a-F	TCGTCCGGCAGCGTCAGATGTGTATAAGAGACAGATGTCACCACAAACAGAGACTAAAGC	607	Levin <i>et al.</i> [36]
	rbcLajf634R	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGAAACGGTCTCTCCAACGCAT		Fazekas <i>et al.</i> [38]
mini- <i>rbcL</i>	F52	TCGTCCGGCAGCGTCAGATGTGTATAAGAGACAGGTTGGATTCAAAGCTGGTGTTA	140	Little[13]
	R193	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCVGTCCAMACAGTWGTCCATGT		
<i>trnL</i> (UAA)	c	TCGTCCGGCAGCGTCAGATGTGTATAAGAGACAGCGAAATCGGTAGACGCTACG	767	Taberlet <i>et al.</i> [39]
	d	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGGGGATAGAGGGACTTGAAC		
<i>trnL</i> (P6 loop)	g	TCGTCCGGCAGCGTCAGATGTGTATAAGAGACAGGGGCAATCCTGAGCCAA	10-143	Taberlet <i>et al.</i> [39]
	h	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCCATTGAGTCTCTGCACCTATC		
ITS2	S2F	TCGTCCGGCAGCGTCAGATGTGTATAAGAGACAGATGCGATACTTGGTGTGAAT	160-320	Chen <i>et al.</i> [40]
	S3R	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGACGCTTCTCCAGACTACAAT		
<i>psbA-trnH</i> <sup>&amp;</sup>	psbAf	TCGTCCGGCAGCGTCAGATGTGTATAAGAGACAGGTTATGCATGAACGTAATGCTC	264-792	Sang <i>et al.</i> [41], Tate and Simpson [42]
	trnH2	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCGCGCATGGTGGATTACAATCC		

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

\*Modified COI cocktail primers without M13-tails were used [18].

& The primers were not included in the final panel of DNA barcodes.



**Table 2: Pilot study: Composition of the experimental mixtures, and taxa identified using the default setting of the CITESspeciesDetect pipeline. (Line no: 3284)**

		Experimental mixtures																
Species/Genus	Common name	EM1	EM2	EM3	EM4	EM5	EM6	EM7	EM8	EM9	EM10	EM10R	EM11	EM11R	EM12	EM13	EM14	EM15
<i>Bos taurus</i>	Cattle	99% (S)	90% (S)	1% (S)	0% (S)	99% (S)	95% (S)	85% (S)			10% (S)	10% (S)	46% (S)	46% (S)	95% (S)	85% (S)		
<i>Parapenaeopsis</i> sp.	Shrimp						1%	3%			10%	10%	1%	1%			1%	3%
<i>Anguilla anguilla</i> *	European eel						1%	3%			10% (S)	10% (S)	1% (S)	1% (S)			1% (S)	3% (S)
<i>Crocodylus niloticus</i> *	Nile crocodile						1% (S)	3% (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>Pieris 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. *	Barrel cactus								1% (F)	3% (F)	10% (F)	10% (F)	1% (F)	1% (F)	1% (F)	3% (F)		
<i>Euphorbia</i> sp. *	Spurge								1% (F)	3% (F)	10% (F)	10% (F)	1% (F)	1% (F)	1% (F)	3% (F)		
<i>Aloe variegata</i> *,&	Tiger aloe					1% (F)			1% (F)	3% (F)	10% (F)	10% (F)	1% (F)	1% (F)	1% (F)	3% (F)		
<i>Dendrobium</i> sp. *	Dendrobium (orchid)								1% (F)	3% (G)					1% (G)	3% (G)		
<i>Cycas revoluta</i> *	Sago palm								1%	3%	10% (G)	10% (G)	1%	1% (G)	1% (G)	3% (G)		
<i>Lactuca sativa</i>	Lettuce	1% (S)	10% (S)	99% (S)	90% (S)				95% (S)	85% (S)	10% (S)	10% (G)	46% (S)	46% (S)			95% (S)	85% (S)

Taxa were identified at the species-level unless otherwise indicated in brackets. Cells highlighted in grey indicate 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 and (S) – Species level

\* Species listed by CITES.

& *Aloe variegata* (synonym *Gonialoe variegata*) was recently assigned to the genus *Gonialoe* [43].

**Table 3: Pilot study: average number of Illumina MiSeq reads, the average number of (pseudo)reads that passed quality control (QC) and the percentage of QC (pseudo)reads that were assigned to DNA barcodes and Operational Taxonomic Units (OTUs) generated per sample. (Line no: 3373)**

Experimental mixture	Number of raw reads	Percentage of QC (pseudo)reads*	Percentage DNA barcode assigned (pseudo)reads*	Percentage OTU clustered (pseudo)reads*
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**	464,648	88.27	96.44	82.26

\* (pseudo)reads are the combined quality controlled (QC) pseudo-reads, and the QC processed unmerged forward and reverse reads.

\*\* Averaged across the 17 Illumina MiSeq datasets.

**Table 4: Taxonomic resolution provided by each DNA barcode marker for EM10 and EM10R. (Line no: 36054)**

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

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. (Line no: 36054)**

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

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

**Table 6: Laboratories participating in the interlaboratory trial. (Line no: 405396)**

<b>Laboratory</b>	<b>City and country</b>
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
U.S. Customs and Border Protection Laboratory	Newark, USA

**Table 7: Interlaboratory trial study: Composition of the complex mixtures, and taxa identified using the default setting of the CITESpeciesDetect pipeline. (Line no: 108399)**

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
<i>Glycine max</i>	Soy bean	1% (16) <i>Glycine</i> 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) Xanthorrhoeaceae	2% (16) Xanthorrhoeaceae	3% (16) Xanthorrhoeaceae	4% (16) Xanthorrhoeaceae	1% (16) Xanthorrhoeaceae	2% (16) Xanthorrhoeaceae	3% (16) Xanthorrhoeaceae	4% (16) Xanthorrhoeaceae
<i>Dendrobium</i> sp. *	Dendrobium orchid	<b>1% (16)</b> <b><i>Dendrobium</i> sp.</b>	<b>2% (16)</b> <b><i>Dendrobium</i> sp.</b>	<b>3% (16)</b> <b><i>Dendrobium</i> sp.</b>	<b>4% (16)</b> <b><i>Dendrobium</i> sp.</b>	<b>1% (16)</b> <b><i>Dendrobium</i> sp.</b>	<b>2% (16)</b> <b><i>Dendrobium</i> sp.</b>	<b>3% (16)</b> <b><i>Dendrobium</i> sp.</b>	<b>4% (16)</b> <b><i>Dendrobium</i> sp.</b>
<i>Huso dauricus</i> *	Sturgeon/Kaluga	<b>1% (16)</b>	<b>2% (16)</b>	<b>3% (16)</b>	<b>4% (16)</b>	<b>1% (14)</b>	<b>2% (16)</b>	<b>3% (16)</b>	<b>4% (16)</b>
<i>Crocodylus niloticus</i> *	Nile crocodile	<b>1% (14)</b>	<b>2% (14)</b>	<b>3% (15)</b>	<b>4% (16)</b>	<b>1% (9)</b>	<b>2% (15)</b>	<b>3% (15)</b>	<b>4% (15)</b>
<i>Lactuca sativa</i>	Lettuce					10% (16)	10% (16)	10% (16)	10% (16)
<i>Brassica oleracea</i>	White cabbage	47% (16)	45% (16)	43% (16)	41% (16)	32% (16)	30% (16)	28% (16)	26% (16)
<i>Sus scrofa</i>	Pig					10% (16)	10% (16)	10% (16)	10% (16)
<i>Bos taurus</i>	Cattle	47% (16)	45% (16)	43% (16)	41% (16)	32% (16)	30% (16)	28% (16)	26% (16)
<i>Pleuronectes platessa</i>	European plaice					10% (16)	10% (16)	10% (16)	10% (16)

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

\* Species listed by CITES

& *Aloe variegata* (synonym *Gonialoe variegata*) was recently assigned to the genus *Gonialoe* [43].

**Table 8: Sample S3 ingredients list and taxa (species, genus, family, order) identified. (Line no: 47064)**

Ingredients label:	Common name	Species/genus	Family	(Infra)Order
Herba Cistanches	Cistanche extract	<i>Cistanche</i> sp.	Orobanchaceae	Lamiales
Cauda cervi	Mature deer tail	<i>Cervus</i> sp.	Cervidae	Pecora
Radix Rehmanniae praeparata	Processed <i>Rehmannia</i> root	<i>Rehmanniae</i> 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 *		<i>Bos taurus</i> (16) <i>Cullen</i> 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 brackets. Species marked in grey are listed by CITES.

\* Species identified by at least 14 laboratories that were not mentioned on ingredients list

**Table 9: Sample S8 ingredients list and taxa (species, genus, family, order) identified. (Line no: 47064)**

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>	Areaceae (16)	Arecales
Yohimbe	Extract from the bark of <i>Pausinystalia johimbe</i>	<i>Pausinystalia johimbe</i>	Rubiaceae (16)	Gentianales
Magnesium stearate				
Undeclared identified taxa *		<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 brackets. Species marked in grey are listed by CITES.

\* Species identified by at least 14 laboratories that were not mentioned on ingredients list.

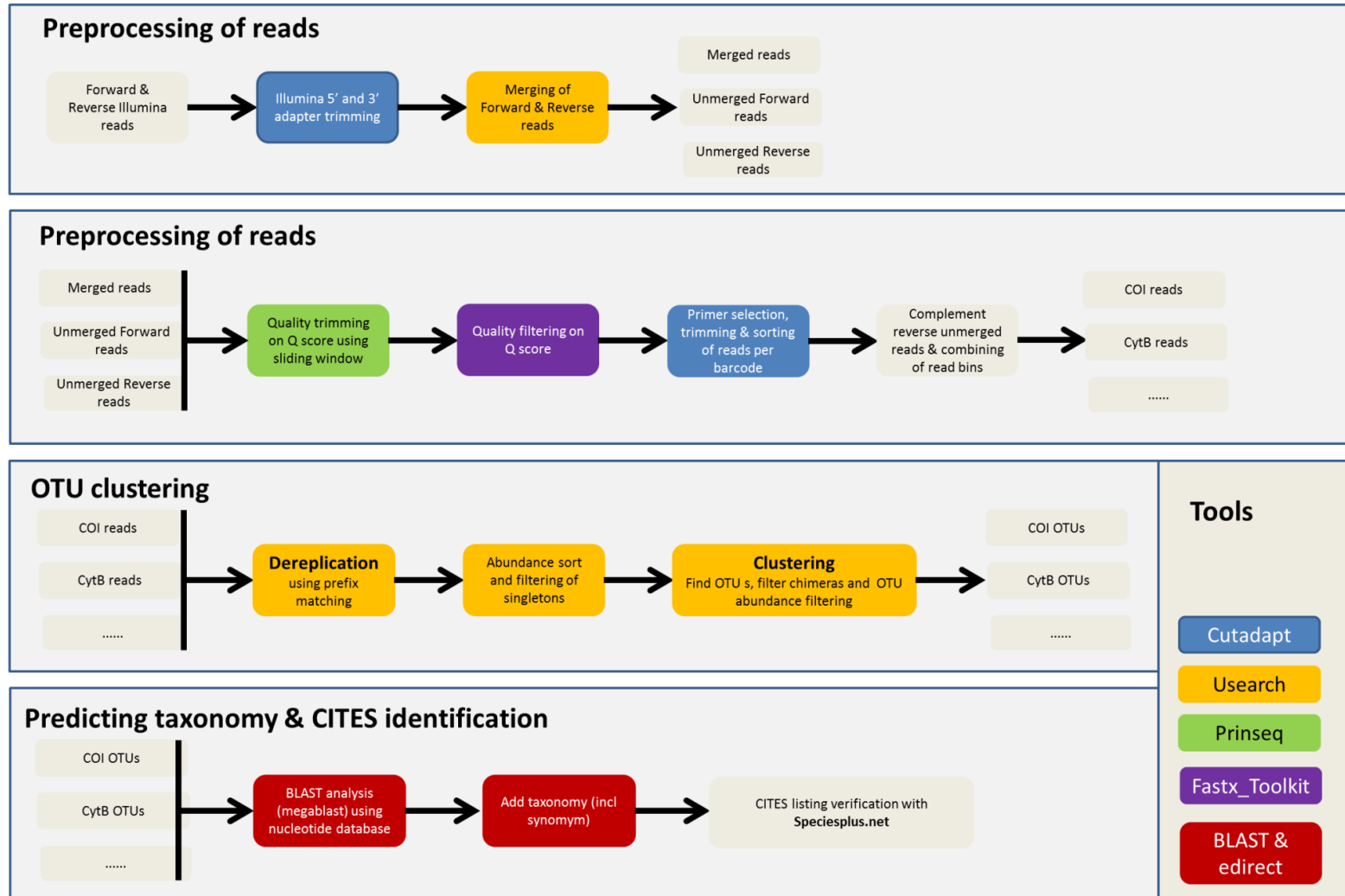


## Figure file:

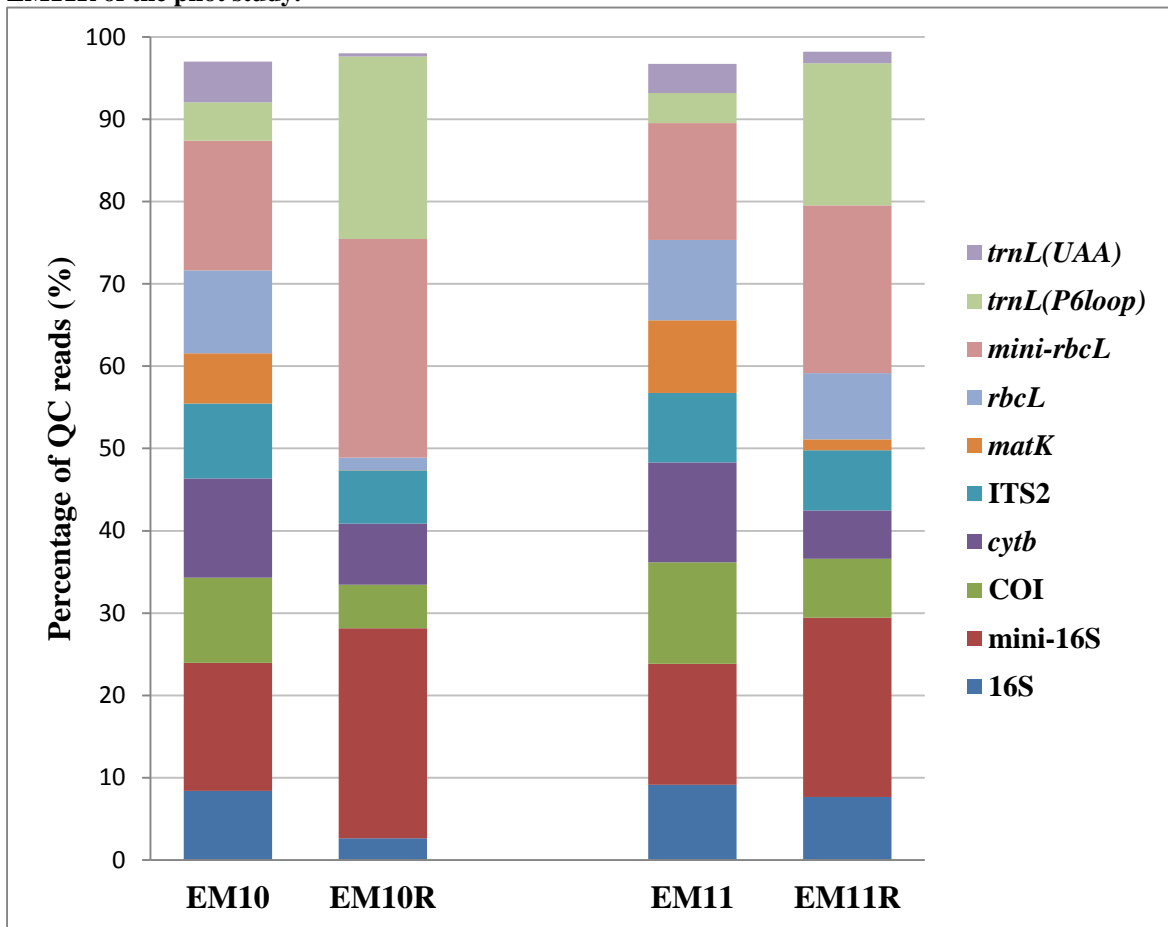
### **Development and validation trial of a multi-locus DNA metabarcoding method to identify endangered species in complex samples.**


Alfred J. Arulandhu, Martijn Staats, Rico Hagelaar, Marleen M. Voorhuijzen, Theo W. Prins, Ingrid Scholtens, Adalberto Costessi, Danny Duijsings, François Rechenmann, Frédéric B. Gaspar, Maria Teresa Barreto Crespo, Arne Holst-Jensen, Matthew Birck, Malcolm Burns, [Edward Haynes](#) ~~Hez-Hird~~, Rupert Hochegger, Alexander Klingl, Lisa Lundberg, Chiara Natale, Hauke Niekamp, Elena Perri, Alessandra Barbante, Jean-Philippe Rosec, Ralf Seyfarth, Tereza Sovová, Christoff Van Moorlegem, Saskia van Ruth, Tamara Peelen and Esther Kok

Figure 1: Schematic representation of the CITESpeciesDetect pipeline.

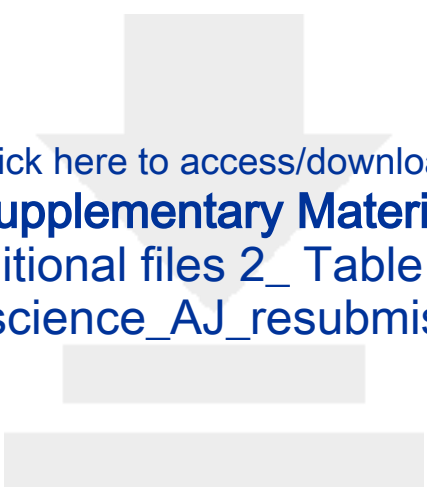


**Figure 2: The percentage of QC reads assigned to DNA barcodes for samples EM10, EM10R, EM11 and EM11R of the pilot study.**





Click here to access/download  
**Supplementary Material**  
Additional file 1\_Table S1-  
S10\_Gigascience\_AJ\_resubmission.docx



Click here to access/download  
**Supplementary Material**  
Additional files 2\_ Table S1-  
S2\_Gigascience\_AJ\_resubmission.xlsx



[Click here to access/download](#)

**Supplementary Material**

[Additional file 3\\_SOP\\_Gigascience\\_AJ\\_resubmission.pdf](#)

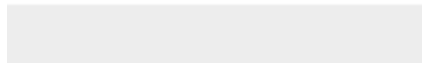




[Click here to access/download](#)

**Supplementary Material**

[Additional file 4\\_Gigascience\\_AJ\\_resubmission.xlsx](#)



Dear Scott Edmunds,

Hez Hird is the co-author from FERA, UK, after the submission the scientific officer from FERA required us to change the authorship to Edward Haynes. I already communicate about this issues with you before, one of your suggest was to change the co-author name in the resubmission process. According, I am changed the FERA co-author in the manuscript and additional file 1.

In the manuscript: line 6, from “Hez Hird” to “Edward Haynes”

In the author affiliation section: line 41, from “Hez Hird” to “Edward Haynes” and from “Hez.Hird@fera.co.uk” to “[Edward.Haynes@fera.co.uk](mailto:Edward.Haynes@fera.co.uk)”

In author contributions: line 822, from “HH” to EH”

In the separate file for figures and tables from “Hez Hird” to “Edward Haynes”

In the additional file 1, from “Hez Hird” to “Edward Haynes”

Warm regards,  
Alfred J Arulandhu