

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

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

11 Alfred J. Arulandhu ¹ - RIKILT Wageningen University & Research, P.O. Box 230, 6700 AE Wageningen, The
12 Netherlands – alfred.arulandhu@wur.nl

13 Alfred J. Arulandhu ² – Food Quality and Design Group, Wageningen University and Research, P.O. Box 8129,
14 6700 EV Wageningen, The Netherlands – alfred.arulandhu@wur.nl

15 Martijn Staats ¹ - RIKILT Wageningen University & Research, P.O. Box 230, 6700 AE Wageningen, The
16 Netherlands – martijn.staats@wur.nl

17 Rico Hagelaar ¹ - RIKILT Wageningen University & Research, P.O. Box 230, 6700 AE Wageningen, The
18 Netherlands – rico.hagelaar@wur.nl

19 Marleen M. Voorhuijzen ¹ - RIKILT Wageningen University & Research, P.O. Box 230, 6700 AE Wageningen,
20 The Netherlands - marleen.voorhuijzen@wur.nl

21 Theo W. Prins ¹ - RIKILT Wageningen University & Research, P.O. Box 230, 6700 AE Wageningen, The
22 Netherlands - theo.prins@wur.nl

23 Ingrid M.J. Scholtens ¹ - RIKILT Wageningen University & Research, P.O. Box 230, 6700 AE Wageningen, The
24 Netherlands - ingrid.scholtens@wur.nl

25 Adalberto Costessi ³ - Baseclear B. V, Einsteinweg 5, 2333 CC Leiden, The Netherlands -
26 Adalberto.Costessi@baseclear.nl

27 Danny Duijsings ³ - Baseclear B. V, Einsteinweg 5, 2333 CC Leiden, The Netherlands -
28 Danny.Duijsings@baseclear.nl

29 François Rechenmann ⁴ - GenoStar Bioinformatics Solutions, 60 rue Lavoisier, 38330 Montbonnot Saint Martin,
30 France - rechenmann@genostar.com

31 Frédéric B. Gaspar ⁵ – iBET, Instituto de Biologia Experimental e Tecnológica, Apartado 12, 2780-901 Oeiras,
32 Portugal - fgaspar@ibet.pt

33 Maria Teresa Barreto Crespo ⁵ - iBET, Instituto de Biologia Experimental e Tecnológica, Apartado 12, 2780-901
34 Oeiras, Portugal - tcrespo@ibet.pt

35 Arne Holst-Jensen ⁶ - Norwegian Veterinary Institute, Ullevaalsveien 68, P.O.Box 750 Sentrum, 0106 Oslo,
36 Norway - arne.holst-jensen@vetinst.no

37 Matthew Birck ⁷ - U.S. Customs and Border Protection Laboratory, 1100 Raymond Blvd Newark, NJ 07102 USA
38 - MATTHEW.BIRCK@cbp.dhs.gov

39 Malcolm Burns ⁸ - LGC, Queens Road, Teddington, Middlesex, TW11 0LY, United kingdom -
40 Malcolm.Burns@lgcgroup.com

41 Hez Hird ⁹ – Fera, Sand Hutton, York, YO41 1LZ, United Kingdom - Hez.Hird@fera.co.uk

42 Rupert Hochegger ¹⁰ - Austrian Agency for Health and Food Safety, Spargelfeldstrasse 191, 1220 Vienna,
43 Austria - rupert.hochegger@ages.at

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

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

Corresponding author: Esther Kok
e-mail: esther.kok@wur.nl

85

86

87

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

89

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

96

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

105

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

110

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

113

114

115

116

117

118

119

120 **Background**

121
122 The demand for endangered species as ingredients in traditional medicines (TMs) has become one of the major
123 threats to the survival of a range of endangered species such as seahorse (*Hippocampus* sp.), agarwood
124 (*Aquilaria* sp.), and Saiga antelope (*Saiga tatarica*) [1-3]. The Convention on the International Trade in
125 Endangered Species of Wild Fauna and Flora (CITES) is one of the best supported conservation agreements to
126 regulate trading of animal and plant species (www.cites.org) and thereby conserve biodiversity. Currently,
127 ~35,000 species are classified and listed by CITES in three categories based on their extinction level (CITES
128 Appendix I, II and III) by which the trade in endangered species is regulated. The success of CITES is dependent
129 upon the ability of customs inspectors to recognize and identify components and ingredients derived from
130 endangered species, for which a wide range of morphological, chromatographic and DNA-based identification
131 techniques can be applied [4,5].

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

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

151 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 shorter DNA barcode regions for degraded DNA samples, such mini-barcodes contain less information and their
3 primers are more restrictive, often making them unsuitable for universal species barcoding [4, 13]. The third
4 challenge is the reference sequence database quality and integrity, which is particularly problematic for law
5 enforcement issues, where high quality and reliability are essential. The current underrepresentation of DNA
6 barcodes from species protected under CITES and closely related species critically hampers their identification.
7 The fourth challenge is that a dedicated bioinformatics pipeline is necessary to process raw NGS data for
8 accurate and sensitive identification of CITES-listed species [9]. Finally, studies using the DNA metabarcoding
9 approach are scarce and none of these methods have been truly validated [9, 14]. Therefore, before implementing
10 DNA metabarcoding by Customs and other enforcement agencies, the above-mentioned challenges need to be
11 thoroughly assessed to ensure accurate taxonomic identifications.
12
13
14
15
16
17
18
19
20
21
22
23

24 The objective of this research was to develop a multi-locus DNA metabarcoding method for
25 (endangered) species identification and to evaluate the applicability and reproducibility of this approach in an
26 international interlaboratory study. The research was part of a larger programme on the development of
27 advanced DNA-based methods from the DECATHLON project (www.decathlon-project.eu), within the
28 European Union's Framework Programme 7. In the process of establishing the standard operating procedure
29 (SOP) for multi-locus DNA metabarcoding, all important aspects of the procedure (i.e. DNA isolation procedure,
30 DNA barcode marker, barcode primers, NGS strategy and bioinformatics) were evaluated. The challenges
31 concerning the quality and integrity of the DNA reference database(s) are discussed. The first step was aimed at
32 identifying an ideal DNA isolation method to extract DNA from complex mixtures consisting of both animal and
33 plant tissues. Secondly, animal and plant DNA barcode markers and corresponding primer sets were identified
34 from literature that allowed good resolution for identifying (endangered) species from a wide taxonomic range.
35 Thirdly, a panel of universal plant and animal DNA barcodes was selected and a single optimal PCR protocol
36 was identified for efficient amplification of a panel of DNA barcode markers. Finally, the suitability of the
37 Illumina MiSeq NGS technology was evaluated, and a bioinformatics pipeline with a user-friendly web interface
38 was established to allow stakeholders to perform the NGS data analysis without expert bioinformatics skills.
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53

54 The DNA metabarcoding method was developed and tested based on data generated for 15 well-
55 defined complex mixtures. The use of well-characterised mixtures allowed for optimising the bioinformatics
56 procedure and subsequent robustness testing of multiple parameter settings and thresholds. The practical
57
58
59
60
61
62

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

188 **Data description**

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

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

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

Analyses

Establishing a laboratory procedure for multi-locus DNA barcode amplification

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

225 The DNA barcode markers included in this study were selected based on Staats et al. [9] supplemented
226 with additional primers from literature [13] (Table 1). DNA barcode markers were selected based on the
227 availability of universal primer sets and DNA sequence information in public repositories [9]. Important
228 considerations in selecting suitable primer sets were that, preferably, they are used in DNA barcoding campaigns
229 and studies, and as such have demonstrated universal applicability across a wide range of taxa. Furthermore,
230 primer sets for both the amplification of full-length barcodes and their respective mini-barcodes (i.e. short
231 barcode regions < 300 nt within existing ones) were selected when available. This was done to facilitate PCR
232 amplification from a range of wildlife forensic samples containing relatively intact DNA (using full-length
233 barcodes) and/or degraded DNA (mini-barcodes). Based on these criteria, PCR primer sets for the following
234 animal DNA barcodes were selected: regions of the mitochondrial genes encoding 16S rRNA gene (16S),
235 cytochrome c oxidase I (COI) and cytochrome *b* (*cyt b*). For plant species identification, primer sets for the
236 following DNA barcodes were selected: regions of the plastidial genes encoding maturase K (*matK*), ribulose-
237 1,5-bisphosphate carboxylase (*rbcL*), tRNA^{Leu} (UAA) intron sequence (*trnL* (UAA)), *psbA-trnH* intergenic
238 spacer region (*psbA-trnH*), and the nuclear internal transcribed spacer 2 (ITS2) region (Table 1). The selected
239 primers sets were modified to include the Illumina adapter sequence at the 5' end of the locus-specific sequence
240 to facilitate efficient NGS library preparation. A gradient PCR experiment was performed to identify the optimal
241 PCR annealing temperature. While the selected PCR primer sets had previously been published with their own

242 annealing temperatures and conditions, the identification of a single optimal annealing temperature for all PCR
1
2 243 primer sets would allow for increased efficiency of analysis. Initially, a thermal gradient of 49.0 °C to 55.0 °C
3
4 244 was tested on the *Bos taurus* reference material with the primer sets for COI-2, 16S, mini-16S, and *cyt b*. The
5
6 245 amplification efficiency across the PCR primers sets was determined by comparing the intensity of the
7
8 246 amplicons across the thermal gradient. An optimal annealing temperature of 49.5 °C was identified, but
9
10 247 additional non-specific amplicons were observed with some primers (not shown). To reduce the amounts of non-
11
12 248 specific amplification products, the PCR program was modified to increase the annealing temperature after five
13
14 249 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
15
16 250 PCR primer combinations still produced non-specific products (for *psbA-trnH* gene) or less intense PCR
17
18 251 products (for *rbcL* gene with primers *rbcLa-F* and *rbcLajf634R*, and *matK* gene with primers *matK-390f* and
19
20 252 *matK-1326r*). Consequently, these PCR primer sets were excluded from subsequent experiments.
21

22 253 Next, the selected PCR thermocycling protocol was evaluated with the remaining 12 PCR primer sets
23
24 254 on a panel of 29 animal and 17 plant species, representing a phylogenetically wide range of taxa (Mammalia,
25
26 255 Actinopterygii, Malacostraca, Bivalvia, Aves, Reptilia, Amphibia, Insecta, Angiospermae, and Cycadopsida;
27
28 256 Additional file 1; Table S2 and S3). The overall PCR amplification success rates varied across reference species
29
30 257 and across DNA barcode markers (Additional file 1; Table S2). For instance, no PCR amplification was
31
32 258 observed with *cyt b* for the CITES-listed species *Balaenoptera physalus*, whereas intense amplification was seen
33
34 259 for the same species with 16S, COI-2, mini-16S and mini-COI (Additional file 1; Table S2). Overall, at least one
35
36 260 DNA barcode marker could successfully be amplified for each of the 46 plant and animal species (Additional file
37
38 261 1; Table S2 and S3). For a number of plant and animal species the generated barcode sequence information was
39
40 262 deposited in the European Nucleotide Archive (ENA) under accession numbers LT009695 to LT009705, and
41
42 263 LT718651 (Additional file 1; Table S1).
43

44 264

45

46 265 **Development and pre-validation of the CITESspeciesDetect bioinformatics pipeline**

47

48 266 A dedicated bioinformatics pipeline, named CITESspeciesDetect, was developed for the purpose of rapid
49
50 267 identification of CITES-listed species using Illumina paired-end sequencing technology. Illumina technology
51
52 268 was selected because it produces NGS data with very low error rates, compared to other technologies [2, 19].
53
54 269 Furthermore, the Illumina MiSeq platform enables paired-end read lengths of up to 300 nt, allowing relatively
55
56 270 long DNA barcode regions of up to ~550 nt to be assembled. Also, the multiplexing capabilities of Illumina
57
58 271 technology are well developed, allowing for simultaneous sequencing of multiple samples in one run, thereby
59
60
61
62

272 enabling more cost-efficient NGS. While NGS data analysis pipelines exist that allow processing of Illumina
1 DNA metabarcoding datasets (e.g. CLOTU, QIIME, Mothur), the majority have been developed for specifically
2 273 studying microbial communities using the 16S rRNA gene region. CITESpeciesDetect, developed in this study,
3 274 extends on the frequently-used software tools developed within the USEARCH [19] and BLAST+ packages
4 275 [20], and additionally includes dedicated steps for quality filtering, sorting of reads per barcode, and CITES
5 276 species identification (Figure 1). The CITESpeciesDetect is composed of five linked tools and data analysis
6 277 passes through three phases: 1) pre-processing of paired-end Illumina data involving quality trimming and
7 278 filtering of reads, followed by sorting by DNA barcode, 2) Operational Taxonomic Unit (OTU) clustering by
8 279 barcode, and 3) taxonomy prediction and CITES identification.
9 280

10 281 In establishing the pipeline, it was found that reads generated for *cyt b* and mini-*cyt b* could not be
11 282 separated based on the forward PCR primer, as the forward primers are identical. It was therefore decided to
12 283 combine (pool) the overlapping reads of *cyt b* and mini-*cyt b* during pre-processing (primer selection) of reads to
13 284 prevent reads from being double selected. This means that the results of *cyt b* and mini-*cyt b* are presented by the
14 285 CITESpeciesDetect pipeline as *cyt b*. The same issue was found for COI, for which the results are presented as
15 286 COI-2.
16 287

17 288 A parameter scan was performed in order to assess the effect of software settings on the ability to
18 289 identify species. The evaluation allowed for the identification of important parameters and their effect on the
19 290 sensitivity, specificity and robustness of the procedure. Changing the base quality score has a major impact on
20 291 the number of reads per barcode (Additional file 1; Table S4). Increasing the strictness of the base quality score
21 292 resulted in decreasing numbers of reads per barcode. Quality score values other than the default values (Q20 for
22 293 95% of bases) did not yield better identifications. When applying strict quality filtering settings (Q20 for 100%
23 294 of bases, or Q30 for 99% of bases) the species *Pieris brassicae* and *Anguilla anguilla* could not be detected with
24 295 *cyt b* and/or mini-COI, indicating these settings were too strict (Additional file 1; Table S5). This is likely due to
25 296 the resulting overall low read numbers for *cyt b* and mini-COI when applying these strict quality filtering
26 297 settings (Additional file 1; Table S4).
27 298

28 299 The effect of error tolerance on Illumina adapter trimming and primer selection was assessed by varying
29 300 the maximum number of errors allowed in assigning reads to DNA barcodes. Setting higher error tolerances
30 301 resulted in slightly higher number of reads being selected per DNA barcode marker (not shown). With 0% error
31 302 tolerance, however, reads were observed that still contained untrimmed Illumina adapters or primer residues.
32 303 These untrimmed residues were not observed when applying a 0.2% error tolerance (not shown).
33 304
34 305
35 306
36 307
37 308
38 309
39 310
40 311
41 312
42 313
43 314
44 315
45 316
46 317
47 318
48 319
49 320
50 321
51 322
52 323
53 324
54 325
55 326
56 327
57 328
58 329
59 330
60 331
61 332
62 333
63 334
64 335
65 336

302 An OTU abundance threshold is generally applied to make DNA metabarcoding less sensitive to
1 (potential) false-positive identifications. False-positives may occur e.g. as contaminants during pre-processing of
2 303 samples (DNA extraction, PCR) or as cross-contamination during Illumina sequencing. Applying an OTU
3 304 abundance threshold higher than zero generally results in loss of sensitivity. We have found, however, that
4 305 applying an OTU abundance threshold of higher than zero may help in reducing noisy identifications and
5 306 potential false-positive identifications (results not shown). In this study, an OTU abundance threshold of 0.2%
6 307 was set as default.
7 308

14 309 The effect of applying a minimum DNA barcode length revealed that allowing DNA barcodes of ≥ 10
15 310 nt did not lead to additional identification of species, compared with default settings (e.g. ≥ 200 nt). Increasing
16 311 the minimal DNA barcode length to 250 nt, however, resulted in a failure to identify most plant species with
17 312 mini-*rbcL* and *rbcL*. We recommend using a minimum DNA barcode length of 200 nt, except for DNA barcodes
18 313 with a basic length shorter than 200 nt, in which case the minimum expected DNA barcode length is set to 10 nt,
19 314 e.g. in case of the *trnL* (P6 loop) marker.
20 315

26 316 The results of the parameter scan resulted in specifying recommended parameter values (default setting)
27 317 for analysing DNA metabarcoding datasets using the CITESspeciesDetect pipeline (see Methods section
28 318 “Bioinformatics analysis”). An online version of the CITESspeciesDetect pipeline with a user-friendly web-
29 319 interface was developed for skilled analysts with basic, but no expert level knowledge in bioinformatics and is
30 320 made available via <http://decathlon-fp7.citespipe-wur.surf-hosted.nl:8080/>.
31 321

38 321 **Pilot study to assess the performance of the DNA metabarcoding procedure using experimental mixtures**

40 322 The DNA metabarcoding procedure was assessed in a pilot study, for which 15 complex mixtures (EM1 to
41 323 EM15) were prepared containing from 2 to 10 taxonomically well-characterised species with DNA barcode
42 324 reference sequences available in the NCBI reference database (Table 2). The experimental mixtures 10 and 11
43 325 (EM10 and EM11) were independently analysed twice to verify repeatability of the method (DNA isolation,
44 326 barcode panel analysis and pooling). Only mixtures were used with well-characterised species (DNA Sanger
45 327 barcoded and taxonomically verified) ingredients, at known dry weight concentrations, and with high quality
46 328 DNA that would allow for an assessment of the performance of the DNA metabarcoding method under optimal
47 329 conditions.
48 330

56 330 A total of 2.37 Gb of Illumina MiSeq sequencing data was generated for the 17 complex samples (15
57 331 complex mixtures along with the two replicates). On average, 464,648 raw forward and reverse Illumina reads
58 332

332 were generated per sample, with minimum and maximum read numbers ranging between 273,104 (mixture
1 EM4) and 723,130 (mixture EM10R; Table 3). During raw data pre-processing with the default settings of the
2 333 CITESpeciesDetect pipeline, the reads were first quality filtered and overlapping paired-end Illumina reads
3 334 were merged into pseudo-reads (Figure 1). The samples contained on average 269,099 quality controlled (QC)
4 335 unmerged (forward and reverse) reads and merged pseudo-reads, collectively named (pseudo)reads. On average
5 336 88.27% (min = 77.38%, max = 96.26%) of raw reads passed the quality filtering and pre-processing steps,
6 337 indicating that the overall quality of the Illumina data was high (not shown).
7 338

339 Next, the (pseudo)reads were assigned to DNA barcodes based on PCR primer sequences. On average,
14 339 96.44% (min = 88.78%, max = 98.21%) of QC pre-processed reads were assigned to DNA barcodes, indicating a
15 340 high percentage of reads containing the locus-specific DNA barcode primers (Table 3). After this, the
16 341 (pseudo)reads were clustered by 98% sequence similarity into OTUs. On average, 82.26% (min = 75.11%, max
17 342 = 90.63%) of the DNA barcodes assigned reads were clustered into OTUs (Table 3). It was assumed that the
18 343 small fraction of reads that was not assigned to OTUs contained non-informative (e.g. non-specific fragments,
19 344 chimeras) sequences that may have been generated during PCR amplification, and were filtered out during
20 345 clustering.
21 346

347 For taxonomy prediction, OTUs were assigned to dataset sequences using BLAST when aligning with
31 347 at least 98% sequence identity, a minimum of 90% query coverage, and an E-value of at least 0.001. Generally,
32 348 the best match (“top hit”) is used as best estimate of species identity. However, species identification using
33 349 BLAST requires careful weighting of the evidence. To minimize erroneous taxonomic identifications a more
34 350 conservative guideline was used that allowed a species to be assigned only when the best three matches
35 351 identified the species. If the bit scores do not decrease after the top three hits, or if other species have identical
36 352 bit scores, then identification was considered inconclusive. In such cases, OTUs were assigned to higher
37 353 taxonomic levels (genus, family or order). All animal ingredients, except *Parapenaepsis* sp. could be identified
38 354 at the species-level with one or more DNA barcode marker using the default settings of the CITESpeciesDetect
39 355 pipeline (Table 4 and 5). For plants, *Lactuca sativa* could be identified at the species-level using the *trnL* (P6
40 356 loop). All other plant taxa were identified at the genus or higher level (Table 4 and 5).
41 357

358 Putative contaminating species were observed in most of the experimental mixtures (Additional file 2;
52 358 Table S1). Even with the default OTU abundance threshold in place, the species *L. sativa*, *B. taurus* and *Gallus*
53 359 *gallus* were identified in mixtures that were not supposed to contain these species. To verify whether these
54 360 putative contaminations occurred during DNA isolation or Illumina sequencing, qPCR assays for the specific
55 361

362 detection of *B. taurus* and *G. gallus* were performed on selected DNA extracts. These results indicated that for
1
2 363 some experimental mixtures (EM8, EM9 and EM14) cross-contamination had occurred during sample
3
4 364 preparation or DNA isolation, while for other experimental mixtures (EM15) cross-contamination may have
5
6 365 occurred during PCR, Illumina library preparation or sequencing. In addition to these contaminants, a species of
7
8 366 *Brassica* was identified in experimental mixtures containing *P. brassica*. This result is most likely not a false-
9
10 367 positive, because the caterpillars used for this study had been fed on cabbage.

11
12 368 The DNA metabarcoding method was found to be sensitive enough to identify most plant and animal taxa at 1%
13
14 369 (dry mass: dry mass) in mixtures of both low (EM1, EM3 and EM5; Table 2) and relatively high complexity
15
16 370 (EM6, EM8, EM11, EM12, and EM14; Table 2). The exception being *Parapenaepsis* sp. (all mixtures), *A.*
17
18 371 *anguilla* in EM6, and *Cycas revoluta* in EM8 and EM11. Careful inspection of the NGS data revealed that in
19
20 372 nearly all cases OTUs related to *Parapenaepsis* sp., *A. anguilla*, and *C. revoluta* were present, but that these
21
22 373 sequences had been filtered out by the CITESspeciesDetect pipeline because their cluster sizes did not fulfil the
23
24 374 0.2% OTU abundance threshold. Lowering the OTU abundance threshold, however, would lead to (more) false-
25
26 375 positive identifications, and this was therefore not implemented.

27
28 376 The repeatability of the laboratory procedure (excluding NGS) was assessed by analysing the
29
30 377 experimental mixtures 10 and 11 (EM10R and EM11R; Table 2), which was independently performed twice, i.e.
31
32 378 DNA isolation and PCR barcode amplification, but NGS was performed on the same MiSeq flow cell as the
33
34 379 other samples of the pilot study. From the comparison, it was observed that the percentage of QC reads was
35
36 380 nearly twice as high in the replicate analyses (Table 3). Also, the percentage of QC reads assigned to DNA
37
38 381 barcodes varied among replicate analyses (Figure 2). Most notable were the observed differences among
39
40 382 replicate analyses in the percentage reads assigned to *matK* and the *trnL* (P6 loop). For example, the percentage
41
42 383 of QC reads assigned to *matK* were 6.11% (14081 reads) and 0.02% (97 reads) in EM10 and EM10R
43
44 384 respectively (Figure 2). The low number of reads assigned to *matK* limited its use for taxonomy identification in
45
46 385 EM10R (Table 4). The multi-locus approach, however, allowed for the repeatable identification of taxa in EM10
47
48 386 and EM11, though not in all cases with all DNA barcode markers (Table 4 and 5).

49
50 387 Based on the results obtained from the pilot study, precautions were taken when grinding the freeze-
51
52 388 dried materials and subsequent mixing to avoid cross-contamination during the laboratory handling of samples,
53
54 389 which were used to improve the SOP for the interlaboratory trial (Additional file 3). Also, control species were
55
56 390 added to experimental mixtures that were prepared for the inter-laboratory trial to allow better confirmation of
57
58 391 sample homogeneity and to verify that no cross-contamination had occurred during sample preparation.

1
2 **393 Assessment of interlaboratory reproducibility of the DNA metabarcoding procedure**

3
4 394 Altogether 16 laboratories from 11 countries (all experienced, well-equipped and proficient in advanced
5
6 395 molecular analysis work), including two of the method developers, participated in the inter-laboratory trial
7
8 396 (Table 6). The laboratories received ten anonymously labelled samples, each consisting of 250 mg powdered
9
10 397 material. Two of the samples, labelled S3 and S8, were authentic TM products seized by the Dutch Customs
11
12 398 Laboratory while the other eight samples were well-characterized mixtures of specimens from carefully
13
14 399 identified taxa in relative dry weight concentrations from 1% to 47% (Table 7). In all experimental mixtures, 1%
15
16 400 of *Zea mays* was added as quality control for homogeneity, which was confirmed with maize-specific *hmg* (high-
17
18 401 mobility group gene) qPCR [16]. Also, tests performed with species-specific qPCR assays indicated that cross-
19
20 402 contamination did not occur during sample preparation (Additional file 1; Table S6). The qPCR assay for the
21
22 403 detection of *Brassica napus*, however, also gave a positive signal for other *Brassica* sp. in the mixtures.

23
24 404 Together with the sample materials, reagents for DNA extraction, and the complete set of barcode
25
26 405 primers, the participants received an obligatory (SOP). Any deviations from the SOP had to be reported. The
27
28 406 participants were instructed to extract DNA, perform PCR using the barcode primers, purify the amplified DNA
29
30 407 by removal of unincorporated primers and primer dimers, and assess the quality and quantity of the amplification
31
32 408 products by gel electrophoresis and UV-spectrophotometry. The purified PCR products were then collected by
33
34 409 the coordinator of the trial (RIKILT Wageningen University & Research, the Netherlands) and shipped to a
35
36 410 sequencing laboratory (BaseClear, the Netherlands) for Illumina sequencing using MiSeq PE300 technology.
37
38 411 The sequencing laboratory performed Index PCR and Illumina library preparation prior to MiSeq sequencing as
39
40 412 specified in the Illumina 16S metagenomics sequencing library preparation guide. The altogether 160 PCR
41
42 413 samples were sequenced using two Illumina flow cells with MiSeq reagent kit v3.

44 414 The interlaboratory trial should ideally have included the use of the online version of the pipeline, but
45
46 415 unfortunately this was not possible due to shortage of time. Therefore, a single (developer) laboratory performed
47
48 416 these bioinformatics analyses. The 160 individual samples contained on average 269,057 raw reads, and more
49
50 417 than 150,000 reads per sample in 95% of the samples (Additional file 1; Table S7). One sample contained less
51
52 418 than 100,000 reads (51,750), which was considered more than sufficient for reliable species identification. After
53
54 419 pre-processing, the samples contained on average 142,938 (pseudo)reads. On average 94.66% of the reads (min
55
56 420 = 88.12%, max = 98.02%) passed the quality filtering indicating that the overall quality of the sequence data was
57
58 421 consistently high across the 160 datasets.

422 OTU-clustering at 98% sequence similarity on average assigned 78.14% of the pre-processed and DNA barcode
1
2 423 assigned reads into OTUs (Additional file 1; Table S7). Only two samples, both from the same laboratory, had a
3
4 424 slightly lower percentage of the (pseudo-)reads assigned to OTUs (66.02% and 66.05%). This indicates that the
5
6 425 pipeline correctly removed PCR artefacts in the clustering phase.

7
8 426 For taxonomy prediction, an OTU would be assigned to a database hit if they aligned with $\geq 98\%$
9
10 427 sequence identity and $\geq 90\%$ query coverage, and yielded an expect value (E-value) of at least 0.001. The
11
12 428 BLAST output of the NGS data was interpreted by participants according to the guidelines in the SOP. Variation
13
14 429 was observed among laboratories in interpreting the BLAST output: some laboratories consistently scored the
15
16 430 top hits, irrespective of bitscore, while other labs selected all hits belonging to the top three bitscores, or
17
18 431 interpreted only the first OTU of each DNA barcode, leading to large differences in identified taxa. Because of
19
20 432 these inconsistencies, the BLAST results were re-interpreted by RIKILT Wageningen University & Research
21
22 433 following the established guideline as mentioned in the SOP. These re-interpreted data are the data referred to in
23
24 434 the following sections.

25
26 435 With one exception, all taxa mixed in at $\geq 1\%$ (dry mass: dry mass) were reproducibly identified by at
27
28 436 least 13 (81%) laboratories (Table 7). *Beta vulgaris* in sample S6 could only be identified by 4 out of 16 (25%)
29
30 437 laboratories. *Beta vulgaris* specific sequences were present in all remaining datasets, but at very low read counts.
31
32 438 So these clusters did not fulfil the 0.2% OTU abundance threshold (results not shown). All six animal species
33
34 439 could be identified to species level with at least one barcode marker (COI), while only four of the 12 plant
35
36 440 species (*Brassica oleracea*, *Carica papaya*, *Gossypium hirsutum*, and *L. sativa*) could be identified to species
37
38 441 level (Additional file 2; Table S2). All other plant species were identified at the genus or higher level. For plants,
39
40 442 no single barcode marker was best, and the most reliable data were obtained by combining the plant barcodes.

41
42 443 Three taxa that were misidentified or not intentionally included in the mixtures were reproducibly
43
44 444 identified across all laboratories. *Acipenser schrenckii* co-occurred in all samples containing *Huso dauricus*. We
45
46 445 have confirmed with DNA metabarcoding that the caviar used for preparing the experimental mixtures contains
47
48 446 both *H. dauricus* and *A. schrenckii* (results not shown). Furthermore, *Brassica rapa* was identified by ITS2 in
49
50 447 sample S4 by all 16 (100%) laboratories, instead of *Brassica napus*. We confirmed by Sanger sequencing *rbcL*
51
52 448 and *matK* that our reference specimen is indeed *Brassica napus*, but that its ITS2 sequence is identical to
53
54 449 *Brassica rapa* (LT718651). Finally, a taxon of the plant family Phellinaceae was reproducibly identified (by all
55
56 450 laboratories) using the mini-*rbcL* marker in all samples containing *L. sativa* (S6, S7, S9, S10). Species of the
57
58 451 family Phellinaceae and *L. sativa* both belong to the order Asterales. The evidence for Phellinaceae was not

452 strong, i.e. the family-level identification was based on a single NCBI reference sequence only (GenBank:
1 X69748). We therefore suspect a misidentification during the interpretation of the BLAST results.
2

3
4 454 Taxa that were identified to be the result of possible contaminations were scarcely observed, i.e. these
5
6 455 were found in isolated cases and could possibly be explained by cross-sample contamination that may have
7
8 456 occurred during any step of sample processing (DNA isolation, PCR, NGS library preparation or NGS). For
9
10 457 example, a contamination with *Gossypium* sp. was observed using *tmL* (P6 loop) in sample S1 of one of the
11
12 458 participating labs. A total of 6 of such suspected cases of incidental cross-contaminations were observed (not
13
14 459 shown).

15
16 460 For the authentic TMs S3 and S8, it was observed that only few labelled ingredients could reproducibly
17
18 461 be identified (Table 8 and 9). For sample S3 (Ma pak leung sea-dog), only the listed ingredients *Cuscuta* sp.
19
20 462 (Chinese dodder seed), and *Astragalus danicus* (Astragalus root) could be identified. For sample S8 (Cobra
21
22 463 performance enhancer), only the listed ingredients *Epimedium* sp. (Horny goat weed; Berberidaceae), *Panax*
23
24 464 *ginseng* (Korean ginseng; Araliaceae), and species of the plant families *Arecaceae* (*Serenoa repens*) and
25
26 465 *Rubiaceae* (*Pausinystalia johimbe*) could be identified. While most declared taxa were not identified, many non-
27
28 466 declared taxa were identified. For sample S3, the animal species *B. taurus*, and the plants *Cullen* sp. (Fabaceae),
29
30 467 *Melilotus officinalis* (Fabaceae), *Medicago* sp. (Fabaceae), *Bupleurum* sp. (Apiaceae), and *Rubus* sp. (Rosaceae)
31
32 468 were identified by at least 14 (88%) laboratories (Table 8). Furthermore, the fungi *Aspergillus fumigatus*
33
34 469 (*Aspergillaceae*) and *Fusarium* sp. (*Nectriaceae*) were reproducibly identified, of which the former is also a
35
36 470 known human pathogenic fungus. For sample S8, the animal species *B. taurus* and *Homo sapiens*, the plant
37
38 471 species *Sanguisorba officinalis* and *Eleutherococcus sessiliflorus*, and members of the plant genera *Croton* and
39
40 472 *Erythroxylum*, and families *Meliaceae* and *Asteraceae*, were reproducibly identified (Table 9).
41

42 473 43 474 **Discussion**

44 475
45 476 In this study, a DNA metabarcoding method was developed using a multi-locus panel of DNA barcodes for the
46
47 477 identification of CITES protected species in highly complex products such as TMs. As a first step, we selected
48
49 478 an optimal DNA isolation method for complex mixtures consisting of both animal and plant tissues. A CTAB
50
51 479 isolation method was found to be the most efficient in obtaining high quality DNA from pure plant and animal
52
53 480 reference materials as well as from complex mixtures. Secondly, a single PCR protocol, suitable for all the
54
55 481 barcodes included, i.e. multiple universal plant and animal barcode and mini-barcode markers, was identified.
56
57 482 This facilitated the design of a multi-locus panel of DNA barcodes. With this panel, the presence of a species
58
59 483 was confirmed with a multiplex marker approach, which improves the resolution for identification and quality
60
61

484 assurance. Furthermore, the developed DNA metabarcoding method includes a dedicated bioinformatics
1 workflow, named CITESspeciesDetect, that was specifically developed for the analysis of Illumina paired-end
2 reads. The developed pipeline requires skilled experts in bioinformatics, and applies scripts for command-line
3
4 486 processing. NGS data analysis pipelines may provide a lot of flexibility to the user, as modifications are easily
5
6 487 implemented by expert users. To simplify the inter-laboratory validation of the pipeline, a user-friendly and
7
8 488 intuitive web-interface with associated “Help” functions and “FAQs” was developed for the CITESspeciesDetect
9
10 489 pipeline. The web interface was, however, not available in the course of the interlaboratory trial. Therefore, the
11
12 490 sequence data generated in the interlaboratory study could not be analysed by the individual laboratories using
13
14 491 the CITESspeciesDetect pipeline. A single (developer) laboratory therefore performed these analyses. Upon the
15
16 492 availability of the online web-interface, individual participants were later given the opportunity to reanalyse their
17
18 493 DNA metabarcoding data. Observations made in this part demonstrated concordance of results with those
19
20 494 obtained by the developing laboratory, reinforcing the perception of CITESspeciesDetect as a user-friendly and
21
22 495 reliable pipeline that may readily be used by enforcement agencies and other laboratories.
23
24 496

26 497 The performance of the DNA metabarcoding method was assessed in an interlaboratory trial in which
27
28 498 the method was found to be highly reproducible across laboratories, and sensitive enough to identify species
29
30 499 present at 1% dry weight content in experimental samples containing up to 11 different species as ingredients.
31
32 500 However, not all laboratories could identify all taxa in all cases. All animal taxa from phylogenetically unrelated
33
34 501 orders could be identified at the species level, in line with the objective that the method should target all animal
35
36 502 species. COI (full-length COI-2 and mini-COI) was found to be the most effective DNA barcode marker for
37
38 503 animal species identification. This is not surprising considering that COI is the standard barcode for almost all
39
40 504 animal groups [21]. Nearly all animal species identifications were supported by multiple DNA barcodes, thereby
41
42 505 giving strong confidence to the correctness of the animal species identifications. In contrast, plants could mainly
43
44 506 be identified at the family level, and no single DNA barcode marker was found to provide best resolution for
45
46 507 identifying plant taxa. Ideally, adequate plant species discrimination would require the combined use of multiple
47
48 508 DNA barcode markers, e.g. *rbcL* + *matK* [22], but this is technically not possible due to the nature of the target
49
50 509 samples (heavily processed) and with the current Illumina Miseq technology. For the identification of plant taxa
51
52 510 listed by CITES, the use of DNA barcodes with relatively modest discriminatory power at the genus or higher
53
54 511 taxonomic level can still be useful, as it is often an entire plant genus or family that is listed by CITES, rather
55
56 512 than individual plant species. This was the case for e.g. Orchidaceae and Cactaceae in this study. Yet, for some
57
58 513 plant species (e.g. *Aloe variegata*) the resolution provided by the used plant DNA barcodes may still be too low
59
60
61
62
63
64
65

514 for unambiguous CITES identification. It is important to note that the maximum achievable Illumina NGS read
1
2 515 length limits the taxonomic resolution of DNA barcodes that are longer than ~550 nt. This particularly limited
3
4 516 the discriminatory power of the full-length plant barcodes *matK* and *rbcL*. The DNA metabarcoding method may
5
6 517 therefore benefit from (currently unavailable) Illumina read lengths longer than 300 nt, or other long-read
7
8 518 sequencing technologies. Single barcodes in several cases failed to amplify or provide resolution. The latter is
9
10 519 likely to be caused mainly by database incompleteness, lack of genetic variability within some loci/target
11
12 520 sequences, and sample composition. However, combining multiple barcodes into a multi-locus metabarcoding
13
14 521 method mitigated the problems observed for individual barcodes. A high degree of confidence in the taxonomic
15
16 522 assignments based on the combined barcodes were therefore observed, providing for enhanced quality assurance
17
18 523 compared to the use of single barcodes.

20 524 While the use of well-characterised experimental mixtures allowed for an assessment of the
21
22 525 performance of the DNA metabarcoding method under ideal conditions, the amplifiable DNA content of real-life
23
24 526 samples encountered in routine diagnostic work are often of an unpredictable and variable quality. An analysis of
25
26 527 two authentic TM products seized by the Dutch Customs Laboratory demonstrated that only few ingredients
27
28 528 listed on the labels could be reproducibly identified. This does not mean that the undetected species were not
29
30 529 used as ingredients. Ingredients may have been processed in such a way that the DNA is either degraded or
31
32 530 effectively removed. This is e.g. the case with refined oils or cooked ingredients [23]. The quality of the
33
34 531 sequence reference database also strongly affects the ability to correctly identify species. Without correct
35
36 532 references that also exhibit the necessary intraspecific variation, it is not possible to match and discriminate
37
38 533 sequence reads correctly. It is well-known that accurate DNA barcoding depends on the use of a reference
39
40 534 database that provides good taxonomic coverage [5, 9]. The current underrepresentation of DNA barcodes from
41
42 535 species protected by CITES and closely related species critically hampers their identification. This will improve
43
44 536 as DNA barcoding campaigns continue, in particular through initiatives such as the Barcode of Wildlife Project
45
46 537 (BWP; www.barcodeofwildlife.org). Only by expansion of the sequence reference database of endangered and
47
48 538 illegally-traded species can DNA barcoding provide the definitiveness required in a court of law.

50 539 A noteworthy observation was that most species that were reproducibly identified did not appear on the
51
52 540 ingredients lists on the labels of the analysed TMs. This is possibly due to mislabelling. If the identifications are
53
54 541 correct this also indicates that consumption may pose health risks. These findings corroborate earlier reports that
55
56 542 DNA metabarcoding may provide valuable information about the quality and safety of TMs [5, 6].

58
59 543

Potential implications

Overall, our findings demonstrate that the multi-locus DNA metabarcoding method assessed in this study can provide reliable and detailed data on the composition of highly complex food products and supplements. This study highlights the necessity of a multi-locus DNA metabarcoding strategy for species identification in complex samples, since the use of multiple barcode markers enables an increased resolution and quality assurance, even in heavily processed samples. The developed robust bioinformatics pipeline for Illumina data analysis with user-friendly web interface allows the method to be directly applied in various fields such as: a) food mislabelling and fraud in the food industry [24], b) environmental monitoring of species [25], and c) wildlife forensics [26]. Furthermore, the pipeline can be readily used to analyse different types of Illumina paired-end datasets, even the future Illumina datasets (read length > 300 nt). Additionally, the web interface provides an opportunity for the global audience with limited expertise in bioinformatics, to analyse their own data. It also provides the liberty to select different primer sets and customise the settings for the selected purposes. As a result, the range of potential applications of the method to identify plant and animal species is diverse, the pipeline is versatile and adjustable to the user's needs, thus providing a powerful tool for research as well as enforcement purposes.

Methods

Reference materials and preparation of experimental mixtures

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

For the initial pilot study, in which the SOP for the DNA metabarcoding approach was established and tested, 15 well-defined complex mixtures were artificially prepared (Table 2). These experimental mixtures were prepared with 2 to 10 taxonomically well-characterised species (Table 2). The ingredients were mixed based on dry weight ratio, for which individual materials were freeze-dried for 78 hours. The lyophilized ingredients were

577 ground using an autoclaved mortar and pestle or blender in a cleaned fume hood, and subsequently stored at -
1 20 °C °C. The individual ingredients of each complex mixture were weighted and mixed thoroughly using a
2 578 tumbler (Heidolph Reax 2) for 20 hours and stored at -20 °C until further use.
3
4 579

5 580 For the interlaboratory validation trial, in which the applicability and reproducibility of the DNA
6 metabarcoding method was assessed, eight additional well-characterised mixtures were artificially prepared
7 581 using the above procedure. These complex mixtures were prepared with 8 to 11 taxonomically well-
8 582 characterised species present at dry weight concentrations from 1% to 47% (Table 7). These complex mixtures
9 583 were prepared in such a way that the efficiency of homogenization and possibility of sample cross-contamination
10 584 could be verified using species-specific qPCR assays. In all samples, 1% of *Zea mays* was added as quality
11 585 control for homogeneity. The presence of *Z. mays* was checked after sample mixing using maize-specific *hmg*
12 586 qPCR along with a positive and negative control. A unique species was added at 1% dry weight to each mixture
13 587 (*S1-Glycine max*, *S2-Gossypium sp.*, *S4-Brassica napus*, *S5-Triticum aestivum*, *S6-Beta vulgaris*, *S7-Meleagris*
14 588 *gallopavo*, *S9-Carica papaya*, *S10-Solanum lycopersicum*) (Table 7). Species-specific qPCR was performed in
15 589 duplex (together with positive and negative controls) in all samples, to check for possible cross-contamination
16 590 between samples after sample preparation. Information about the qPCR primers and probes, and qPCR
17 591 procedure can be found in the Additional file 1; Table S8-S10. In addition to the eight experimental mixtures,
18 592 two TMs were included that were obtained from the Dutch Customs Laboratory: a) Ma pak leung sea-dog hard
19 593 capsules (MA PAK LEUNG CO, LTD, Hong Kong), was labelled to contain among others rhizoma Cibotii
20 594 (*Cibotium barometz*, CITES appendix II), and Herba Cistanches (*Cistanche sp.*, CITES appendix II) and b)
21 595 Cobra performance enhancer hard capsules (Gold caps, USA), was labelled to contain among others Siberian
22 596 ginseng (*Eleutherococcus senticosus*) and Korean ginseng (*Panax ginseng*). In both TMs, the medicine powder
23 597 was encapsulated in a hard-capsule shell. All capsules were opened and the powder inside the capsules were
24 598 stored in air-sealed and sterilized containers. The powdered medicines were thoroughly mixed using tumbler
25 599 (Heidolph Reax 2) for 20 hours and stored at -20 °C until further use.
26 600
27
28
29
30
31
32
33
34
35
36
37
38
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

604 **DNA isolation method**

605 A cetyltrimethylammonium bromide (CTAB) extraction method [17] was assessed for its ability to efficiently
606 extract DNA from a range of plant and animal materials (Additional file 3). In brief, the CTAB method consists
607 of an initial step to separate polysaccharides and organic soluble molecules using a CTAB extraction buffer (1X

608 CTAB, 1.4M NaCl, 0.1 M Tris-HCl [pH 8.0], and 20mM NA₂EDTA) and chloroform. Next, the DNA was
609 precipitated with 96% ethanol, purified with 70% ethanol, and the obtained DNA was stored at 4 °C until further
610 use. DNA was extracted from 100 mg reference materials (plant and animal), artificially made complex mixtures,
611 and real-life samples (TMs). The concentration and purity (OD_{260/280} and OD_{260/230} ratios) of the obtained DNA
612 was determined by spectrophotometer (NanoDrop 1000 instrument, Thermo Fisher Scientific Inc.). The OD_{260/280}
613 ratios between 1.7 and 2.0 were considered to indicate purity of the obtained DNA.

615 **Barcode markers**

616 Candidate universal DNA barcode and mini-barcode markers and primer sets were identified using the
617 information provided in Staats et al. (2016) [9], supplemented with additional primer sets from literature (Table
618 1). The PCR primer sets were modified to have an additional Illumina tail sequence at 5' end of the primers
619 (Table 1).

621 **PCR**

622 A gradient PCR was performed with all PCR primer combinations using 10 ng of DNA. The tested PCR
623 conditions programme were according to the following protocol: 95 °C for 15 min, five cycles of 94 °C for 30 s,
624 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,
625 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,
626 which included 12.5 µl of HotStarTaq Master Mix (Qiagen), 0.5 µl of 10 µM each sense and antisense primer, 7
627 µl of RNase-free water (Qiagen) and 5 µl of 10 ng/µl of represented species DNA. PCR was performed in the
628 CFX96 thermal cycler (Bio-Rad) and the amplified products from all the analysed reference specimens,
629 artificially made complex mixtures, and real-life samples (TMs) were visualised on 1% agarose gels. Prior to
630 NGS library preparation, 8 µl of PCR product of each target (12 in total) per sample was pooled and mixed. Next,
631 the pooled PCR products were purified using the QIAquick PCR purification kit (Qiagen) according to
632 manufacturer's protocol, and the purified amplicons were visualized on 1% agarose gels.

635 **Next Generation Sequencing**

636 The pooled and purified PCR amplicons were sequenced using Illumina MiSeq paired-end 300 technology. Prior
637 to MiSeq sequencing, Index PCR and Illumina library preparation were performed as specified in the Illumina

638 16S metagenomics sequencing library preparation guide (Illumina document 15044223). All the DNA barcode
1 amplicons of each sample were treated as one sample during library preparation i.e. all DNA barcode amplicons
2 639
3 of each sample were tagged with the addition of the same, unique identifier, or index sequence, during library
4 640
5 preparation. The Index PCR was performed to add dual indices (multiplex identifiers) and Illumina sequencing
6 641
7 adapters using the Nextera XT Index Kit (Illumina, FC-131-1001). Illumina libraries were quantified and pooled
8 642
9 prior to MiSeq sequencing using MiSeq reagent kit v3.
10 643

11 644

14 645 **Bioinformatics analysis**

15 646
16 647 The raw demultiplexed Illumina reads with Illumina 1.8+ encoding were processed using a bioinformatics
17
18 648 pipeline, called CITESspeciesDetect. The CITESspeciesDetect is composed of five linked tools with data
19
20 649 analysis passing through three phases: 1) pre-processing of paired-end Illumina data involving quality trimming
21
22 650 and filtering of reads, followed by sorting by DNA barcode, 2) OTU clustering by barcode, and 3) taxonomy
23
24 651 prediction and CITES identification (Figure 1).

25
26 652 During preprocessing of reads, the 5' and 3' Illumina adapter sequences are trimmed using Cutadapt v1.9.1 [27]
27
28 653 using the respective substrings TGTGTATAAGAGACAG and CTGTCTCTTATACACA. After Illumina
29
30 654 adapter trimming, reads ≤ 10 bp are removed using Cutadapt. Then, the forward and reverse reads are merged to
31
32 655 convert a pair into a single pseudoread containing one sequence and one set of quality score using USEARCH
33
34 656 v8.1.1861 [19].

35
36 657 Next, the merged pseudo-reads, unmerged forward reads and unmerged reverse reads are processed
37
38 658 separately during quality filtering using a sliding window method implemented in PRINSEQ [28]. During this
39
40 659 procedure, low quality bases with Phred scores lower than 20 are trimmed from 3'-end using a window size of
41
42 660 15 nt and a step size of 5 nt. After PRINSEQ, reads with a minimum of 95% per base quality ≥ 20 are kept,
43
44 661 while the remaining reads are removed using FASTX_Toolkit v0.0.14 (http://hannonlab.cshl.edu/fastx_toolkit/).
45
46 662 Then, reads are successively selected, trimmed and sorted per DNA barcode marker using Cutadapt [27]. The
47
48 663 following steps are followed for each DNA barcode marker separately during this procedure. First, reads
49
50 664 containing an anchored 5' forward primer or anchored 5' reverse primer (or their reverse complement) are
51
52 665 selected with a maximum error tolerance of 0.2 (=20%) and with the overlap parameter specified to 6 to ensure
53
54 666 specific selection of reads. Also, reads ≤ 10 nt are removed. The anchored 5' primer sequences are subsequently
55
56 667 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

668 DNA barcode, the primer-selected and unmerged reverse reads are reverse complemented and combined with
1
2 669 primer-selected merged and unmerged forward reads.

3
4 670 The following procedure is used to cluster the quality trimmed reads of each DNA barcode into OTUs
5
6 671 using the UPARSE pipeline implemented in USEARCH [19] with the following modifications: reads are
7
8 672 dereplicated using the derep_prefix command. Also, singleton reads and reads with minimum cluster size
9
10 673 smaller than 4 are discarded. Representative OTUs are generated using an OTU radius of 2 (98% identity
11
12 674 threshold) and 0.2% OTU abundance threshold with minimum barcode length per primer set. Filtering of
13
14 675 chimeric reads is performed using the default settings of the UPARSE-REF algorithm implemented in the
15
16 676 cluster_otus command of USEARCH.

17
18 677 To assign OTUs to taxonomy, standalone BLASTn megablast searches [20] of representative OTUs are
19
20 678 performed on the National Centre for Biotechnology Information (NCBI) GenBank nucleotide database using an
21
22 679 Expectation value (E-value) threshold of 0.001 and a maximum of 20 aligned sequences. OTUs are assigned to
23
24 680 the database sequence to which they align, based on bit score, and having at least 98% sequence identity and
25
26 681 minimum of 90% query coverage. To identify putative CITES-listed taxa, the taxon ID first was matched against
27
28 682 the NCBI taxonomy database using Entrez Direct (edirect) functions (available at
29
30 683 <ftp://ftp.ncbi.nlm.nih.gov/entrez/entrezdirect/>) to retrieve scientific name (species, genus, family, order and
31
32 684 synonym name). The scientific, synonym and/or family names are then matched against a local CITES database
33
34 685 that is retrieved from <https://speciesplus.net>. The final results are presented as a tab-separated values file (TSV)
35
36 686 containing the BLAST hit metadata (i.e. bit-score, e-value, accession numbers etc.), the scientific name,
37
38 687 synonym name, and in case a CITES-listed taxon was found, also the CITES appendix listing and taxonomic
39
40 688 group (i.e. species, genus, family or order name) under which the taxon is listed by CITES.

41
42 689 The BLAST output was interpreted by following guidelines: first, to minimize the chance of erroneous
43
44 690 species identifications, the same species should have at least three top hits, i.e. highest bit scores. Secondly, if
45
46 691 multiple hits are obtained with identical quality results, but with different assigned species, or with less than
47
48 692 three top hits with same species designation, the OTU fragment was considered to lack the discriminatory power
49
50 693 to refer the hit to species level. In such cases, the OTU would then be downgraded to a genus-level identification.
51
52 694 Thirdly, if multiple hits are obtained with identical quality results, but with different assigned genera, the OTU
53
54 695 fragment lacks the discriminatory power to describe the hit to genus level. In such cases, the OTU would then be
55
56 696 downgraded to a family-level identification. An online web-interface based application for the
57
58
59
60
61
62

697 CITESpeciesDetect pipeline was developed which is available from <http://decathlon-fp7.citespipe-wur.surf-hosted.nl:8080/>.

699

700 **Pre-validation in-house of the CITESpeciesDetect pipeline**

701 A parameter scan was performed in order to assess the effect of software settings on the ability to identify
702 species. This evaluation allowed for identification of important parameters and their effects on the sensitivity,
703 specificity and robustness of the procedure. This in turn resulted in specified, recommended (default) parameters
704 values for analysing DNA metabarcoding datasets using the CITESpeciesDetect pipeline. The effects of the
705 following parameters were assessed: base quality scores, error tolerance for primer selection, OTU radius, OTU
706 abundance threshold, expect E-value and query coverage threshold, percentage identity threshold, minimum
707 DNA barcode length and BLAST database. The parameters scan was performed on experimental mixture 11 of
708 the pilot study (Table 2). This mixture was selected because of its (relatively) high sample complexity, making it
709 the most challenging complex mixture to analyse. Furthermore, the parameter scan was limited to four barcode
710 primer sets: full-length cytochrome-B (*cyt b*), COI mini barcode (mini-COI), *rbcL* mini barcode (mini-*rbcL*) and
711 the full-length *rbcL* (*rbcL*) barcode.

713 **Inter-laboratory validation trial: participants and method.**

714 To assess the overall performance of the developed DNA metabarcoding approach, 16 laboratories from 11
715 countries participated in an international inter-laboratory validation. Only laboratories that regularly perform
716 molecular analyses and have well-equipped laboratory facilities were selected to participate (Table 6). The
717 majority are governmental or semi-official institutes and are considered highly authoritative within each
718 respective country. Participants were requested to follow the SOP (Additional file 3), and were asked to
719 document any deviations that were made. The chemicals and reagents that were provided to the laboratories were:
720 10 samples (eight experimental mixtures and two TMs), *B. taurus* and *L. sativa* positive control DNA, CTAB
721 extraction and precipitation buffer, 1.2 M NaCl solution, 12 universal plant and animal barcode and mini-
722 barcode primer sets (Table 1), Qiagen HotStarTaq master mix, and Qiagen PCR purification kits. All reagents
723 and samples were provided in quantities corresponding to 2.5× the amounts required for the planned experiments.
724 After following the SOP from DNA isolation to purification of the amplified products, all the purified samples
725 from all the laboratories (n=160) were collected and sequenced using Illumina MiSeq paired-end 300 technology
726 (at BaseClear, Leiden, NL). The Index PCR and Illumina library preparation was performed according to the

727 guideline and all 160 samples were sequenced on two Illumina flow cells. After Illumina MiSeq run, the raw
1 NGS data was processed using the default settings of the CITESspeciesDetect pipeline. BLAST outputs for the
2
3 samples were distributed back to the participating laboratories for interpretation of results. The laboratories
4
5 interpreted the BLAST output based on the guideline provided in the SOP.
6
7

8 731

9 732 **Availability of supporting data**

10 733 All the sequence data obtained from the pilot study and the international interlaboratory validation trial, the
11
12 CITESspeciesDetect pipeline and access to web interface are freely available. The generated barcode sequence
13
14 information for some animal and plant species were deposited in GenBank under the accession numbers
15
16 LT009695 to LT009705, and LT718651 (Additional file 1; Table S1). The Illumina PE300 MiSeq data obtained
17
18 from the pilot study and the international interlaboratory validation trial (n=177) were deposited to ENA with
19
20 study ID PRJEB18620. The script for the CITESspeciesDetect pipeline is available at GitHub. The web interface
21
22 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/)
23
24 [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.
25
26

27 741

28 742 **Availability and requirements**

29
30
31 743 Project name: CITESspeciesDetect

32
33 744 Project home page: <https://github.com/RIKILT/CITESspeciesDetect>

34
35 745 Operating system(s): Linux

36
37 746 Programming language: Python and Bash

38
39 747 Other requirements: none

40
41 748 License: BSD 3-Clause License

42
43 749 Any restrictions to use by non-academics: none
44
45
46

47 750

48 751

49 752

50 753

51 754

52 755

53 756

54 757

55

56

57

58

59

60

61

758 **Additional files**

1
2 759 **Additional file 1: Table S1** Accession numbers of DNA barcode sequences of plant and animal species. **Table**
3
4 760 **S2** PCR success rate for animal reference species. **Table S3** PCR success rate for plant reference species. **Table**
5
6 761 **S4** Statistics of different quality filtering settings for four DNA barcodes. **Table S5** BLAST identification of
7
8 762 species with different quality filtering settings for four DNA barcodes. **Table S6** Results of species-specific
9
10 763 qPCR performed on the experimental mixtures prepared for the inter-laboratory validation trial. **Table S7**
11
12 764 Interlaboratory trial study: average number of Illumina reads per sample, the average number of (pseudo)reads
13
14 765 that passed quality control (QC) and the percentage of QC (pseudo)reads that were assigned to DNA barcodes
15
16 766 and Operational Taxonomic Units (OTUs). **Table S8** qPCR primer and probe information. **Table S9** qPCR
17
18 767 reagent composition. **Table S10** qPCR thermocycling program. (*.docx).

19
20 768
21
22 769 **Additional file 2: Table S1** Pilot study: Composition of the experimental mixtures, and taxa identified using the
23
24 770 default settings of the CITESpeciesDetect pipeline. **Table S2** Interlaboratory trial: the taxonomic resolution
25
26 771 provided by each DNA barcode marker for eight experimental mixtures (*.xlsx).

27
28 772
29 773 **Additional file 3:** Standard operating procedure (SOP) for the multi-locus DNA metabarcoding method that was
30
31 774 used in the inter-laboratory validation study (*.pdf).

32
33 775
34
35 776 **Additional file 4: Table S1** ENA accession numbers of all raw NGS datasets obtained in this study (*.xlsx).

36 777
37
38 778 **Abbreviations**
39
40 779 CITES: Convention on International trade in Endangered Species of Wild fauna and flora; TMs: Traditional
41
42 780 Medicines; NGS: Next generation sequencing; CTAB: cetyltrimethylammonium bromide; COI: Cytochrome c
43
44 781 oxidase subunit I; *cyt b*: Cytochrome *b* gene; 16S rDNA: 16S ribosomal DNA; *matK*: Maturase K gene; *rbcL*:
45
46 782 ribulose-1,5-bisphosphate carboxylase large subunit gene; ITS2: Internal transcribed spacer region 2;; SOP:
47
48 783 Standard operating procedure; OTU: Operational Taxonomic Unit; BLAST: Basic Local Alignment Search Tool.

49
50 784
51
52 785 **Competing interests**
53
54 786 The authors declare that they have no competing interest.

55
56 787
57 788
58
59 789
60 790

791

1 792 **Funding**

2
3 793 The DECATHLON project has been funded with support from the European Commission in the context of the
4
5 794 Seventh Framework Programme (FP7). This publication and all its contents reflect the views only of the authors,
6
7 795 and the Commission cannot be held responsible for any use, which may be made of the information contained
8
9 796 therein.

10
11 797
12 798 **Authors' Contributions**

13
14 799 AJA and MS shared the first authorship. AJA, MS, MV, TP, AC, EK conceived and designed the experiments
15
16 800 for the pilot study. AJA performed the experiments for the pilot study. MS, RH, AJA developed the
17
18 801 CITESspeciesDetect pipeline. AJA, MS, RH analysed the NGS data obtained from the pilot study. AJA, MS,
19
20 802 MV, TP, TWP, IS, EK, FG, MTBC, AHJ involved in establishing the Standard Operation Procedure for the
21
22 803 validation trial. AJA, MS, MV, TP, EK conceived and designed the experiments for the validation trial. FG,
23
24 804 MTBC, AHJ, AJA, MS involved in coordinating the trial. AJA, MV prepared the samples and materials for the
25
26 805 validation trial and distributed to the participated laboratories. FR, MS, RH involved in developing the web-
27
28 806 interface. MS, TP, DD, MBI, MBU, HH, RHO, AK, LL, CN, HN, EP, JPR, RS, TS, CVM took part in the
29
30 807 validation trial. AJA, MS, RH, MV analysed the NGS data obtained from the validation trial. AJA, MS, RH, MV,
31
32 808 SVR, EK contributed to the writing of the manuscript. All authors read and approved the final manuscript.

33
34 809
35 810 **Acknowledgements**

36
37
38 811 This work was supported by the DECATHLON project, which was funded by the European Commission under
39
40 812 Seventh Framework Programme (FP7).

41
42 813
43 814

44 815 **Reference:**

- 45 816
46 817 1. Chang C-H, Jang-Liaw N-H, Lin Y-S, Fang Y-C, Shao K-T: **Authenticating the use of dried**
47 818 **seahorses in the traditional Chinese medicine market in Taiwan using molecular forensics.**
48 819 *Journal of Food and Drug Analysis* 2013, **21**:310-316.
49 820 2. Lee SY, Ng WL, Mahat MN, Nazre M, Mohamed R: **DNA Barcoding of the Endangered Aquilaria**
50 821 **(Thymelaeaceae) and Its Application in Species Authentication of Agarwood Products Traded in**
51 822 **the Market.** *PLOS One* 2016, **11**:e0154631.
52 823 3. Milner-Gulland E, Bukreeva O, Coulson T, Lushchekina A, Kholodova M, Bekenov A, Grachev IA:
53 824 **Conservation: Reproductive collapse in saiga antelope harems.** *Nature* 2003, **422**:135-135.
54 825 4. Cheng X, Su X, Chen X, Zhao H, Bo C, Xu J, Bai H, Ning K: **Biological ingredient analysis of**
55 826 **traditional Chinese medicine preparation based on high-throughput sequencing: the story for**
56 827 **Liuwei Dihuang Wan.** *Scientific Reports* 2014, **4**: 5147.
57 828 5. Coghlan ML, Haile J, Houston J, Murray DC, White NE, Moolhuijzen P, Bellgard MI, Bunce M: **Deep**
58 829 **sequencing of plant and animal DNA contained within traditional Chinese medicines reveals**
59 830 **legality issues and health safety concerns.** *PLOS Genetics* 2012, **8**:e1002657.

- 831 6. Coghlan ML, Maker G, Crighton E, Haile J, Murray DC, White NE, Byard RW, Bellgard MI, Mullaney
1 832 I, Trengove R: **Combined DNA, toxicological and heavy metal analyses provides an auditing**
2 833 **toolkit to improve pharmacovigilance of traditional Chinese medicine (TCM).** *Scientific Reports*
3 834 2015, **5**.
- 4 835 7. Ivanova NV, Kuzmina ML, Braukmann TWA, Borisenko AV, Zakharov EV: **Authentication of**
5 836 **Herbal Supplements Using Next-Generation Sequencing.** *PLOS One* 2016, **11**:e0156426.
- 6 837 8. Taberlet P, Coissac E, Pompanon F, Brochmann C, Willerslev E: **Towards next-generation**
7 838 **biodiversity assessment using DNA metabarcoding.** *Molecular Ecology* 2012, **21**:2045-2050.
- 8 839 9. Staats M, Arulandhu AJ, Gravendeel B, Holst-Jensen A, Scholtens I, Peelen T, Prins TW, Kok E:
9 840 **Advances in DNA metabarcoding for food and wildlife forensic species identification.** *Analytical*
10 841 *and Bioanalytical Chemistry* 2016:1-16.
- 11 842 10. Fahner NA, Shokralla S, Baird DJ, Hajibabaei M: **Large-scale monitoring of plants through**
12 843 **environmental DNA metabarcoding of soil: recovery, resolution, and annotation of four DNA**
13 844 **markers.** *PLOS One* 2016, **11**:e0157505.
- 14 845 11. Arulandhu AJ, Staats M, Peelen T, Kok E: **DNA metabarcoding of endangered plant and animal**
15 846 **species in seized forensic samples.** In *Genome*. 2015: 188-189.
- 16 847 12. Taylor H, Harris W: **An emergent science on the brink of irrelevance: a review of the past 8 years**
17 848 **of DNA barcoding.** *Molecular Ecology Resources* 2012, **12**:377-388.
- 18 849 13. Little DP: **A DNA mini-barcode for land plants.** *Molecular Ecology Resources* 2014, **14**:437-446.
- 19 850 14. Parveen I, Gafner S, Techen N, Murch SJ, Khan IA: **DNA Barcoding for the Identification of**
20 851 **Botanicals in Herbal Medicine and Dietary Supplements: Strengths and Limitations.** *Planta*
21 852 *Medica* 2016, **82**:1225-1235.
- 22 853 15. Chen R, Dong J, Cui X, Wang W, Yasmeen A, Deng Y, Zeng X, Tang Z: **DNA based identification of**
23 854 **medicinal materials in Chinese patent medicines.** *Scientific Reports* 2012, **2**:958.
- 24 855 16. Scholtens I, Laurensse E, Molenaar B, Zaaier S, Gaballo H, Boleij P, Bak A, Kok E: **Practical**
25 856 **experiences with an extended screening strategy for genetically modified organisms (GMOs) in**
26 857 **real-life samples.** *Journal of agricultural and food chemistry* 2013, **61**:9097-9109.
- 27 858 17. Murray M, Thompson WF: **Rapid isolation of high molecular weight plant DNA.** *Nucleic Acids*
28 859 *Research* 1980, **8**:4321-4326.
- 29 860 18. Ivanova NV, Zemlak TS, Hanner RH, Hebert PDN: **Universal primer cocktails for fish DNA**
30 861 **barcoding.** *Molecular Ecology Notes* 2007, **7**:544-548.
- 31 862 19. Edgar RC: **Search and clustering orders of magnitude faster than BLAST.** *Bioinformatics* 2010,
32 863 **26**:2460-2461.
- 33 864 20. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and**
34 865 **PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Research* 1997,
35 866 **25**:3389-3402.
- 36 867 21. Hebert PD, Cywinska A, Ball SL, deWaard JR: **Biological identifications through DNA barcodes.**
37 868 *Proceedings of the Royal Society of London B: Biological Sciences* 2003, **270**:313-321.
- 38 869 22. Group CPW, Hollingsworth PM, Forrest LL, Spouge JL, Hajibabaei M, Ratnasingham S, van der Bank
39 870 M, Chase MW, Cowan RS, Erickson DL: **A DNA barcode for land plants.** *Proceedings of the*
40 871 *National Academy of Sciences* 2009, **106**:12794-12797.
- 41 872 23. Gryson N: **Effect of food processing on plant DNA degradation and PCR-based GMO analysis: a**
42 873 **review.** *Analytical and Bioanalytical Chemistry* 2010, **396**:2003-2022.
- 43 874 24. Galimberti A, De Mattia F, Losa A, Bruni I, Federici S, Casiraghi M, Martellos S, Labra M: **DNA**
44 875 **barcoding as a new tool for food traceability.** *Food Research International* 2013, **50**:55-63.
- 45 876 25. Taberlet P, Coissac E, Hajibabaei M, Rieseberg LH: **Environmental DNA.** *Molecular Ecology* 2012,
46 877 **21**:1789-1793.
- 47 878 26. Iyengar A: **Forensic DNA analysis for animal protection and biodiversity conservation: a review.**
48 879 *Journal for Nature Conservation* 2014, **22**:195-205.
- 49 880 27. Martin M: **Cutadapt removes adapter sequences from high-throughput sequencing reads.** *EMBnet*
50 881 *journal* 2011, **17**:10-12.
- 51 882 28. Schmieder R, Edwards R: **Quality control and preprocessing of metagenomic datasets.**
52 883 *Bioinformatics* 2011, **27**:863-864.
- 53 884 29. Palumbi S, Martin A, Romano S, McMillan W, Stice L, Grabowski G: **The Simple Fool's Guide to**
54 885 **PCR, Version 2.0, privately published document compiled by S. Palumbi.** Dept. Zoology, Univ
55 886 *Hawaii, Honolulu, HI* 1991, **96822**.
- 56 887 30. Sarri C, Stamatis C, Sarafidou T, Galara I, Godosopoulos V, Kolovos M, Liakou C, Tastsoglou S,
57 888 Mamuris Z: **A new set of 16S rRNA universal primers for identification of animal species.** *Food*
58 889 *Control* 2014, **43**:35-41.

- 890 31. Leray M, Yang JY, Meyer CP, Mills SC, Agudelo N, Ranwez V, Boehm JT, Machida RJ: **A new**
1 891 **versatile primer set targeting a short fragment of the mitochondrial COI region for**
2 892 **metabarcoding metazoan diversity: application for characterizing coral reef fish gut contents.**
3 893 *Front Zool* 2013, **10**:34.
- 4 894 32. Geller J, Meyer C, Parker M, Hawk H: **Redesign of PCR primers for mitochondrial cytochrome c**
5 895 **oxidase subunit I for marine invertebrates and application in all-taxa biotic surveys.** *Molecular*
6 896 *Ecology Resources* 2013, **13**:851-861.
- 7 897 33. Parson W, Pegoraro K, Niederstätter H, Föger M, Steinlechner M: **Species identification by means of**
8 898 **the cytochrome b gene.** *International Journal of Legal Medicine* 2000, **114**:23-28.
- 9 899 34. Fazekas AJ, Kuzmina ML, Newmaster SG, Hollingsworth PM: **DNA barcoding methods for land**
10 900 **plants.** *Methods in Molecular Biology* 2012, **858**:223-252.
- 11 901 35. Cuénoud P, Savolainen V, Chatrou LW, Powell M, Grayer RJ, Chase MW: **Molecular phylogenetics**
12 902 **of Caryophyllales based on nuclear 18S rDNA and plastid *rbcL*, *atpB*, and *matK* DNA sequences.**
13 903 *American Journal of Botany* 2002, **89**:132-144.
- 14 904 36. Levin RA, Wagner WL, Hoch PC, Nepokroeff M, Pires JC, Zimmer EA, Sytsma KJ: **Family-level**
15 905 **relationships of Onagraceae based on chloroplast *rbcL* and *ndhF* data.** *American Journal of Botany*
16 906 2003, **90**:107-115.
- 17 907 37. Kress WJ, Erickson DL: **A two-locus global DNA barcode for land plants: the coding *rbcL* gene**
18 908 **complements the non-coding *trnH-psbA* spacer region.** *PLOS One* 2007, **2**:e508.
- 19 909 38. Fazekas AJ, Burgess KS, Kesanakurti PR, Graham SW, Newmaster SG, Husband BC, Percy DM,
20 910 Hajibabaei M, Barrett SC: **Multiple multilocus DNA barcodes from the plastid genome**
21 911 **discriminate plant species equally well.** *PLOS One* 2008, **3**:e2802.
- 22 912 39. Taberlet P, Coissac E, Pompanon F, Gielly L, Miquel C, Valentini A, Vermet T, Corthier G,
23 913 Brochmann C, Willerslev E: **Power and limitations of the chloroplast *trnL* (UAA) intron for plant**
24 914 **DNA barcoding.** *Nucleic Acids Research* 2007, **35**:e14.
- 25 915 40. Chen S, Yao H, Han J, Liu C, Song J, Shi L, Zhu Y, Ma X, Gao T, Pang X: **Validation of the ITS2**
26 916 **region as a novel DNA barcode for identifying medicinal plant species.** *PLOS One* 2010, **5**:e8613.
- 27 917 41. Sang T, Crawford D, Stuessy T: **Chloroplast DNA phylogeny, reticulate evolution, and**
28 918 **biogeography of *Paeonia* (Paeoniaceae).** *American Journal of Botany* 1997, **84**:1120-1136.
- 29 919 42. Tate JA, Simpson BB: **Paraphyly of *Tarasa* (Malvaceae) and diverse origins of the polyploid**
30 920 **species.** *Systematic Botany* 2003, **28**:723-737.
- 31 921 43. Manning J, Boatwright JS, Daru BH, Maurin O, Bank Mvd: **A molecular phylogeny and generic**
32 922 **classification of Asphodelaceae subfamily Alooideae: a final resolution of the prickly issue of**
33 923 **polyphyly in the alooids?** *Systematic Botany* 2014, **39**:55-74.
34 924
35
36
37
38
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

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, 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.

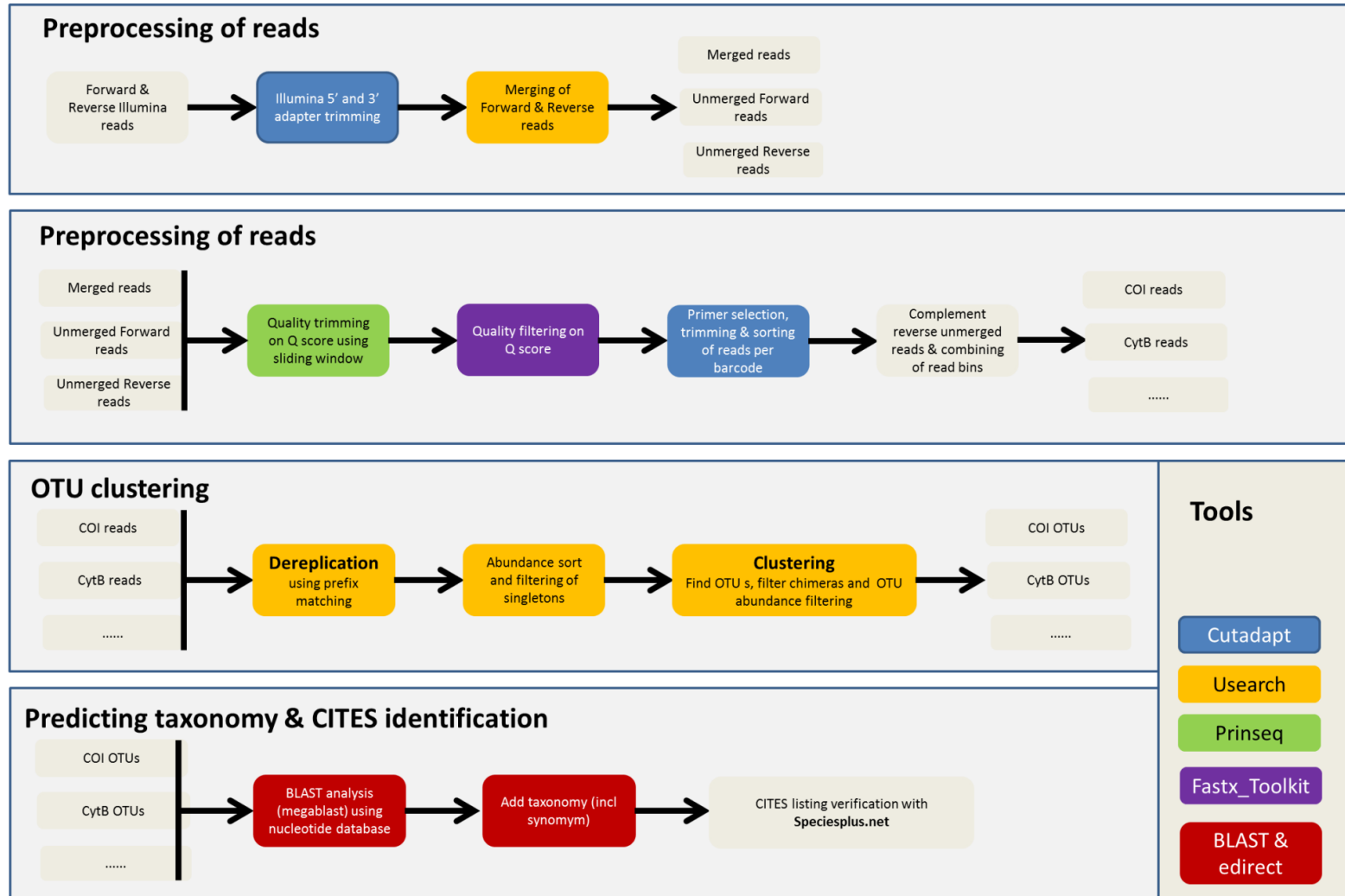
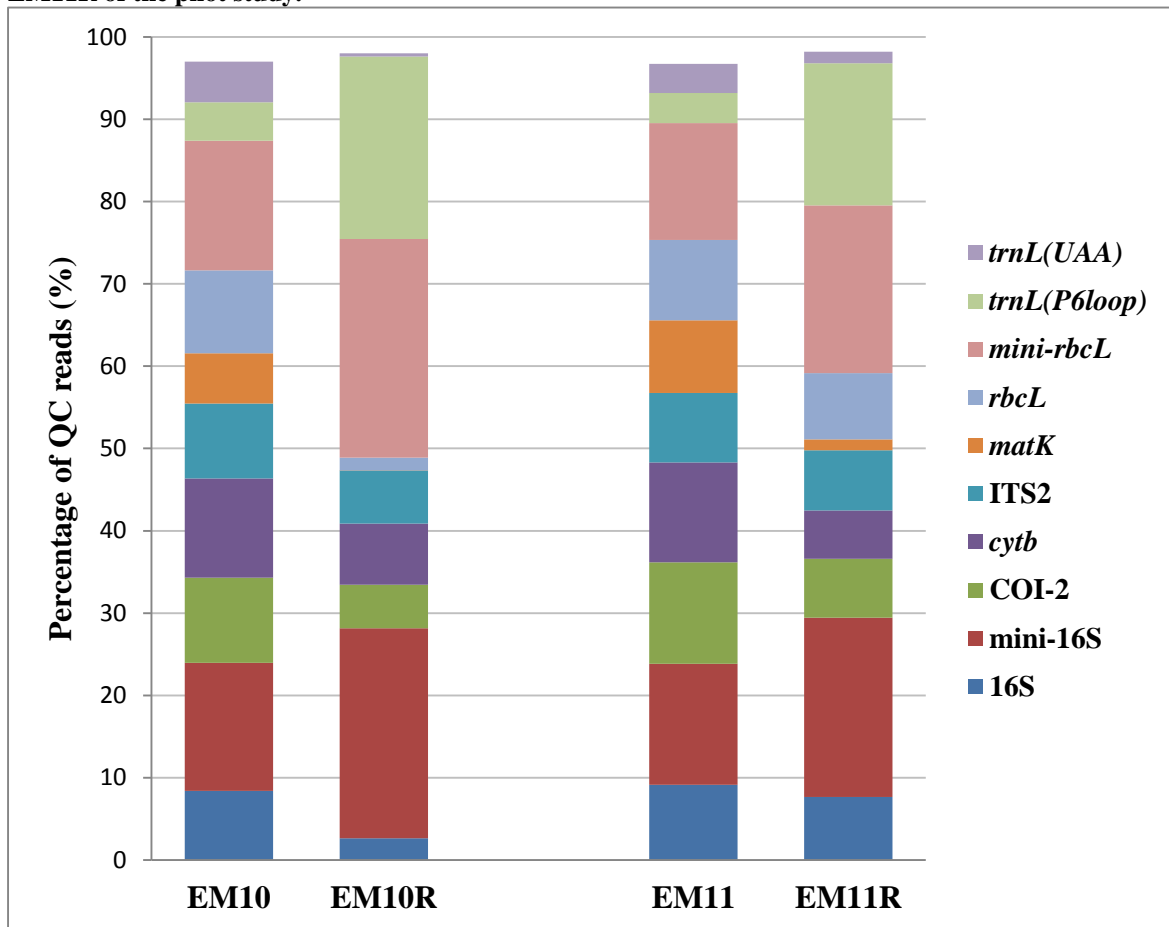


Figure 2: The percentage of QC reads assigned to DNA barcodes for samples EM10, EM10R, EM11 and EM11R of the pilot study.



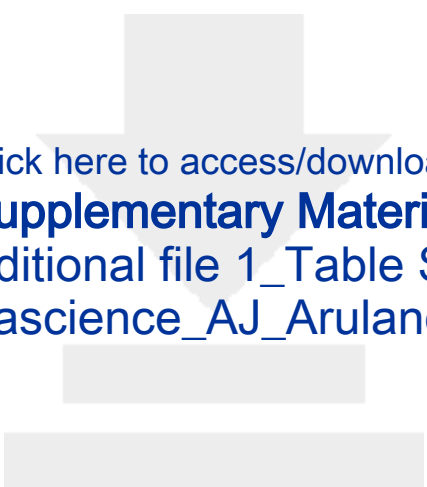


Click here to access/download

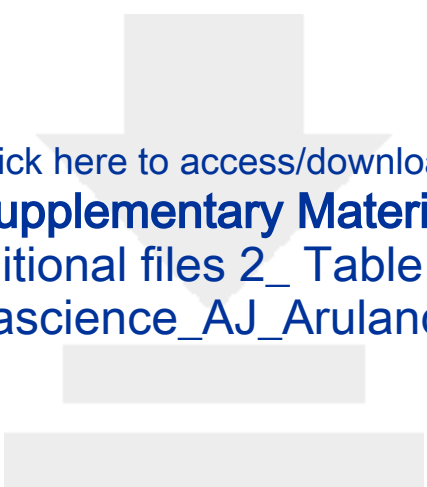
Supplementary Material

A multi-locus DNA metabarcoding
method_Tables_AJ_Arulandhu .docx





Click here to access/download
Supplementary Material
Additional file 1_Table S1-
S10_Gigascience_AJ_Arulandhu.docx



Click here to access/download
Supplementary Material
Additional files 2_ Table S1-
S2_Gigascience_AJ_Arulandhu.xlsx



[Click here to access/download](#)

Supplementary Material

[Additional file 3_SOP_Gigascience_AJ_Arulandhu.pdf](#)





[Click here to access/download](#)

Supplementary Material

[Additional file 4_Gigascience_AJ_Arulandhu.xlsx](#)



Dear editor,

The tables for the manuscript was uploaded under the supplementary file with the title: A multi-locus DNA metabarcoding method_Tables_AJ_Arulandhu.docx. Please include this file while reviewing the manuscript.

Warm regards,
Alfred J Arulandhu