

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, Edward Haynes ⁹, Rupert Hochegger 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 Moorlegem 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 Edward Haynes ⁹ – Fera, Sand Hutton, York, YO41 1LZ, United Kingdom - Edward.Haynes@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, ORCID: 0000-0003-1472-6710

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

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

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

86
87 **Background:** DNA metabarcoding provides great potential for species identification in complex samples such
88 as food supplements and traditional medicines. Such a method would aid CITES (the Convention on
89 International Trade in Endangered Species of Wild Fauna and Flora) enforcement officers to combat wildlife
90 crime by preventing illegal trade of endangered plant and animal species. The objective of this research was to
91 develop a multi-locus DNA metabarcoding method for forensic wildlife species identification and to evaluate the
92 applicability and reproducibility of this approach across different laboratories.

93
94 **Results:** A DNA metabarcoding method was developed that makes use of 12 DNA barcode markers that have
95 demonstrated universal applicability across a wide range of plant and animal taxa, and that facilitate the
96 identification of species in samples containing degraded DNA. The DNA metabarcoding method was developed
97 based on Illumina MiSeq amplicon sequencing of well-defined experimental mixtures, for which a
98 bioinformatics pipeline with user-friendly web interface was developed. The performance of the DNA
99 metabarcoding method was assessed in an international validation trial by 16 laboratories, in which the method
100 was found to be highly reproducible and sensitive enough to identify species present in a mixture at 1% dry
101 weight content.

102
103 **Conclusion:** The advanced multi-locus DNA metabarcoding method assessed in this study provides reliable and
104 detailed data on the composition of complex food products, including information on the presence of CITES-
105 listed species. The method can provide improved resolution for species identification, while verifying species
106 with multiple DNA barcodes contributes to an enhanced quality assurance.

107
108 **Keywords:** Endangered species, CITES, Traditional medicines, DNA metabarcoding, Customs agencies, COI,
109 *matK*, *rbcL*, *cyt b*, mini-barcodes.

120 **Background**

1 121
2 122 The demand for endangered species as ingredients in traditional medicines (TMs) has become one of the major
3
4 123 threats to the survival of a range of endangered species such as seahorse (*Hippocampus* sp.), agarwood
5
6 124 (*Aquilaria* sp.), and Saiga antelope (*Saiga tatarica*) [1-3]. The Convention on the International Trade in
7
8 125 Endangered Species of Wild Fauna and Flora (CITES) is one of the best supported conservation agreements to
9
10 126 regulate trading of animal and plant species (www.cites.org) and thereby conserve biodiversity. Currently,
11
12 127 ~35,000 species are classified and listed by CITES in three categories based on their extinction level (CITES
13
14 128 Appendix I, II and III) by which the trade in endangered species is regulated. The success of CITES is dependent
15
16 129 upon the ability of customs inspectors to recognize and identify components and ingredients derived from
17
18 130 endangered species, for which a wide range of morphological, chromatographic and DNA-based identification
19
20 131 techniques can be applied [4,5].

21
22 132 Recent studies have shown the potential of DNA metabarcoding for identifying endangered species in
23
24 133 TMs and other wildlife forensic samples [4-7]. DNA metabarcoding is an approach that combines DNA
25
26 134 barcoding with next-generation sequencing (NGS), which enables sensitive high-throughput multispecies
27
28 135 identification on the basis of DNA extracted from complex samples [8]. DNA metabarcoding uses more or less
29
30 136 universal PCR primers to mass-amplify informative DNA barcode sequences [9, 10]. Subsequently, the obtained
31
32 137 DNA barcodes are sequenced and compared to a DNA sequence reference database from well-characterized
33
34 138 species for taxonomic assignment [8, 10]. The main advantage of DNA metabarcoding over other identification
35
36 139 techniques is that it permits the identification of all animal and plant species within samples that are composed of
37
38 140 multiple ingredients, which would not be possible through morphological means and time-consuming with
39
40 141 traditional DNA barcoding [4-6]. Furthermore, the use of mini-barcode markers in DNA metabarcoding facilitate
41
42 142 the identification of species in highly processed samples containing heavily degraded DNA [5, 6]. Such a
43
44 143 molecular approach could aid the Customs Authorities to identify materials derived from endangered species in a
45
46 144 wide variety of complex samples, such as food supplements and TMs [11].

47
48 145 Before routine DNA metabarcoding can be applied, there are some key issues that need to be taken
49
50 146 into account. First, complex products seized by Customs, such as TM products, may contain plant and animal
51
52 147 components that are highly processed, and from which the isolation of good quality DNA is challenging. Second,
53
54 148 the universal DNA barcodes employed may not result in amplification of the related barcode for each species
55
56 149 contained in a complex sample, due to DNA degradation or the lack of PCR primer sequence universality. For
57
58 150 plants, for example, different sets of DNA barcodes have been suggested for different fields of application (i.e.
59
60
61
62

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 CITESpeciesDetect pipeline. All raw
208 NGS datasets from both analyses were deposited in ENA under accession numbers ERS1545972 to
209 ERS1545988, ERS1546502 to ERS1546533, ERS1546540 to ERS1546619, ERS1546624 to ERS1546639,
210 ERS1546742 to ERS1546757, ERS1546759 to ERS1546774, and study number PRJEB18620 (Additional file 3;

211 Table S1). A web interface was developed for the CITESpeciesDetect pipeline to allow stakeholders to perform
1 the NGS data analysis of their own samples. The web interface can be globally accessed via the SURFsara high-
2 performance computing and data infrastructure (<http://decathlon-fp7.citespipe-wur.surf-hosted.nl:8080/>).
3
4

214 215 **Analyses** 216

217 **Establishing a laboratory procedure for multi-locus DNA barcode amplification**

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

222 The DNA barcode markers included in this study were selected based on Staats et al. [9] supplemented
223 with additional primers from literature [13] (Table 1). DNA barcode markers were selected based on the
224 availability of universal primer sets and DNA sequence information in public repositories [9]. Important
225 considerations in selecting suitable primer sets were that, preferably, they are used in DNA barcoding campaigns
226 and studies, and as such have demonstrated universal applicability across a wide range of taxa. Furthermore,
227 primer sets for both the amplification of full-length barcodes and their respective mini-barcodes (i.e. short
228 barcode regions < 300 nt within existing ones) were selected when available. This was done to facilitate PCR
229 amplification from a range of wildlife forensic samples containing relatively intact DNA (using full-length
230 barcodes) and/or degraded DNA (mini-barcodes). Based on these criteria, PCR primer sets for the following
231 animal DNA barcodes were selected: regions of the mitochondrial genes encoding 16S rRNA gene (16S),
232 cytochrome c oxidase I (COI) and cytochrome *b* (*cyt b*). For plant species identification, primer sets for the
233 following DNA barcodes were selected: regions of the plastidial genes encoding maturase K (*matK*), ribulose-
234 1,5-bisphosphate carboxylase (*rbcL*), tRNA^{Leu} (UAA) intron sequence (*trnL* (UAA)), *psbA-trnH* intergenic
235 spacer region (*psbA-trnH*), and the nuclear internal transcribed spacer 2 (ITS2) region (Table 1). The selected
236 primers sets were modified to include the Illumina adapter sequence at the 5' end of the locus-specific sequence
237 to facilitate efficient NGS library preparation. A gradient PCR experiment was performed to identify the optimal
238 PCR annealing temperature. While the selected PCR primer sets had previously been published with their own
239 annealing temperatures and conditions, the identification of a single optimal annealing temperature for all PCR
240 primer sets would allow for increased efficiency of analysis. Initially, a thermal gradient of 49.0 °C to 55.0 °C
241 was tested on the *Bos taurus* reference material with the primer sets for COI, 16S, mini-16S, and *cyt b*. The

242 amplification efficiency across the PCR primers sets was determined by comparing the intensity of the
1 amplicons across the thermal gradient. An optimal annealing temperature of 49.5 °C was identified, but
2 243 additional non-specific amplicons were observed with some primers (not shown). To reduce the amounts of non-
3 244 specific amplification products, the PCR program was modified to increase the annealing temperature after five
4 245 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
5 246 PCR primer combinations still produced non-specific products (for *psbA-trnH* gene) or less intense PCR
6 247 products (for *rbcL* gene with primers *rbcLa-F* and *rbcLajf634R*, and *matK* gene with primers *matK-390f* and
7 248 *matK-1326r*). Consequently, these PCR primer sets were excluded from subsequent experiments.
8 249

15 250 Next, the selected PCR thermocycling protocol was evaluated with the remaining 12 PCR primer sets
16 251 on a panel of 29 animal and 17 plant species, representing a phylogenetically wide range of taxa (Mammalia,
17 252 Actinopterygii, Malacostraca, Bivalvia, Aves, Reptilia, Amphibia, Insecta, Angiospermae, and Cycadopsida;
18 253 Additional file 1; Table S2 and S3). The overall PCR amplification success rates varied across reference species
19 254 and across DNA barcode markers (Additional file 1; Table S2). For instance, no PCR amplification was
20 255 observed with *cyt b* for the CITES-listed species *Balaenoptera physalus*, whereas intense amplification was seen
21 256 for the same species with 16S, COI, mini-16S and mini-COI (Additional file 1; Table S2). Overall, at least one
22 257 DNA barcode marker could successfully be amplified for each of the 46 plant and animal species (Additional file
23 258 1; Table S2 and S3). For a number of plant and animal species the generated barcode sequence information was
24 259 deposited in the European Nucleotide Archive (ENA) under accession numbers LT009695 to LT009705, and
25 260 LT718651 (Additional file 1; Table S1).
26 261

262 **Development and pre-validation of the CITESspeciesDetect bioinformatics pipeline**

263 A dedicated bioinformatics pipeline, named CITESspeciesDetect, was developed for the purpose of rapid
264 identification of CITES-listed species using Illumina paired-end sequencing technology. Illumina technology
265 was selected because it produces NGS data with very low error rates, compared to other technologies [2, 19].
266 Furthermore, the Illumina MiSeq platform enables paired-end read lengths of up to 300 nt, allowing relatively
267 long DNA barcode regions of up to ~550 nt to be assembled. Also, the multiplexing capabilities of Illumina
268 technology are well developed, allowing for simultaneous sequencing of multiple samples in one run, thereby
269 enabling more cost-efficient NGS. While NGS data analysis pipelines exist that allow processing of Illumina
270 DNA metabarcoding datasets (e.g. CLOTU, QIIME, Mothur), the majority have been developed for specifically
271 studying microbial communities using the 16S rRNA gene region. CITESspeciesDetect, developed in this study,

272 extends on the frequently-used software tools developed within the USEARCH [19] and BLAST+ packages
1 [20], and additionally includes dedicated steps for quality filtering, sorting of reads per barcode, and CITES
2
3 species identification (Figure 1). The CITESspeciesDetect is composed of five linked tools and data analysis
4
5 passes through three phases: 1) pre-processing of paired-end Illumina data involving quality trimming and
6
7 filtering of reads, followed by sorting by DNA barcode, 2) Operational Taxonomic Unit (OTU) clustering by
8
9 barcode, and 3) taxonomy prediction and CITES identification.
10

11
12 It was found that with the current setup of the pipeline, reads generated for *cyt b* and mini-*cyt b* could
13
14 not be separated based on the forward PCR primer, as the forward primers are identical. It was therefore decided
15
16 to combine (pool) the overlapping reads of *cyt b* and mini-*cyt b* during pre-processing (primer selection) of reads
17
18 to prevent reads from being double selected. This means that the results of *cyt b* and mini-*cyt b* are presented by
19
20 the CITESspeciesDetect pipeline as *cyt b*. The same issue was found for COI barcode and mini-barcode markers,
21
22 for which the results are presented as COI.
23

24 A parameter scan was performed in order to assess the effect of software settings on the ability to
25
26 identify species. The evaluation allowed for the identification of important parameters and their effect on the
27
28 sensitivity, specificity and robustness of the procedure. Changing the base quality score has a major impact on
29
30 the number of reads per barcode (Additional file 1; Table S4). Increasing the strictness of the base quality score
31
32 resulted in decreasing numbers of reads per barcode. Quality score values other than the default values (Q20 for
33
34 95% of bases) did not yield better identifications. When applying strict quality filtering settings (Q20 for 100%
35
36 of bases, or Q30 for 99% of bases) the species *Pieris brassicae* and *Anguilla anguilla* could not be detected with
37
38 *cyt b* and/or mini-COI, indicating these settings were too strict (Additional file 1; Table S5). This is likely due to
39
40 the resulting overall low read numbers for *cyt b* and mini-COI when applying these strict quality filtering
41
42 settings (Additional file 1; Table S4).
43

44 An OTU abundance threshold is generally applied to make DNA metabarcoding less sensitive to
45
46 (potential) false-positive identifications. False-positives may occur e.g. as contaminants during pre-processing of
47
48 samples (DNA extraction, PCR) or as cross-contamination during Illumina sequencing. Applying an OTU
49
50 abundance threshold higher than zero generally results in loss of sensitivity. We have found, however, that
51
52 applying an OTU abundance threshold of higher than zero may help in reducing noisy identifications and
53
54 potential false-positive identifications (results not shown). It should be noted that applying filtering thresholds
55
56 may always lead to false negative or false positive identifications. In this study, an OTU abundance threshold of
57
58
59
60
61
62

301 0.2% was set as default, however, the OTU abundance threshold may need re-evaluation for samples with
1
2 302 expected very low species abundances (< 1% dry weight).

3
4 303 The effect of applying a minimum DNA barcode length revealed that allowing DNA barcodes of ≥ 10
5
6 304 nt did not lead to additional identification of species, compared with default settings (e.g. ≥ 200 nt). Increasing
7
8 305 the minimal DNA barcode length to 250 nt, however, resulted in a failure to identify most plant species with
9
10 306 mini-*rbcL* and *rbcL*. We implemented a minimum DNA barcode length of 200 nt, except for DNA barcodes with
11
12 307 a basic length shorter than 200 nt, in which case the minimum expected DNA barcode length is set to 100 nt for
13
14 308 ITS2, 140 nt for mini-*rbcL*, and 10 nt for the *trnL* (P6 loop) marker.

15
16 309 The results of the parameter scan resulted in specifying recommended parameter values (default setting)
17
18 310 for analysing DNA metabarcoding datasets using the CITESspeciesDetect pipeline (see Methods section
19
20 311 “Bioinformatics analysis”). An online version of the CITESspeciesDetect pipeline with a user-friendly web-
21
22 312 interface was developed for skilled analysts with basic, but no expert level knowledge in bioinformatics and is
23
24 313 made available via <http://decathlon-fp7.citespipe-wur.surf-hosted.nl:8080/>.

25
26 314

27 28 315 **Pilot study to assess the performance of the DNA metabarcoding procedure using experimental mixtures**

29
30 316 The DNA metabarcoding procedure was assessed in a pilot study, for which 15 complex mixtures (EM1 to
31
32 317 EM15) were prepared containing from 2 to 10 taxonomically well-characterised species with DNA barcode
33
34 318 reference sequences available in the NCBI reference database (Table 2). The experimental mixtures 10 and 11
35
36 319 (EM10 and EM11) were independently analysed twice to verify repeatability of the method (DNA isolation,
37
38 320 barcode panel analysis and pooling). Only mixtures were used with well-characterised species (DNA Sanger
39
40 321 barcoded and taxonomically verified) ingredients, at known dry weight concentrations, and with high quality
41
42 322 DNA that would allow for an assessment of the performance of the DNA metabarcoding method under optimal
43
44 323 conditions.

45
46 324 A total of 2.37 Gb of Illumina MiSeq sequencing data was generated for the 17 complex samples (15
47
48 325 complex mixtures along with the two replicates). On average, 464,648 raw forward and reverse Illumina reads
49
50 326 were generated per sample, with minimum and maximum read numbers ranging between 273,104 (mixture
51
52 327 EM4) and 723,130 (mixture EM10R; Table 3). During raw data pre-processing with the default settings of the
53
54 328 CITESspeciesDetect pipeline, the reads were first quality filtered and overlapping paired-end Illumina reads
55
56 329 were merged into pseudo-reads (Figure 1). The samples contained on average 269,099 quality controlled (QC)
57
58 330 unmerged (forward and reverse) reads and merged pseudo-reads, collectively named (pseudo)reads. On average

331 88.27% (min = 77.38%, max = 96.26%) of raw reads passed the quality filtering and pre-processing steps,
1
2 332 indicating that the overall quality of the Illumina data was high (not shown).

3
4 333 Next, the (pseudo)reads were assigned to DNA barcodes based on PCR primer sequences. On average,
5
6 334 96.44% (min = 88.78%, max = 98.21%) of QC pre-processed reads were assigned to DNA barcodes, indicating a
7
8 335 high percentage of reads containing the locus-specific DNA barcode primers (Table 3). After this, the
9
10 336 (pseudo)reads were clustered by 98% sequence similarity into OTUs. On average, 82.26% (min = 75.11%, max
11
12 337 = 90.63%) of the DNA barcodes assigned reads were clustered into OTUs (Table 3). It was assumed that the
13
14 338 small fraction of reads that was not assigned to OTUs contained non-informative (e.g. non-specific fragments,
15
16 339 chimeras) sequences that may have been generated during PCR amplification, and were filtered out during
17
18 340 clustering.

19
20 341 For taxonomy prediction, OTUs were assigned to dataset sequences using BLAST when aligning with
21
22 342 at least 98% sequence identity, a minimum of 90% query coverage, and an E-value of at least 0.001. Generally,
23
24 343 the best match (“top hit”) is used as best estimate of species identity. However, species identification using
25
26 344 BLAST requires careful weighting of the evidence. To minimize erroneous taxonomic identifications a more
27
28 345 conservative guideline was used that allowed a species to be assigned only when the best three matches
29
30 346 identified the species. If the bit scores do not decrease after the top three hits, or if other species have identical
31
32 347 bit scores, then identification was considered inconclusive. In such cases, OTUs were assigned to higher
33
34 348 taxonomic levels (genus, family or order). All animal ingredients, except *Parapenaepsis* sp. could be identified
35
36 349 at the species-level with one or more DNA barcode marker using the default settings of the CITESpeciesDetect
37
38 350 pipeline (Table 4 and 5). For plants, *Lactuca sativa* could be identified at the species-level using the *trnL* (P6
39
40 351 loop). All other plant taxa were identified at the genus or higher level (Table 4 and 5).

42 352 Putative contaminating species were observed in most of the experimental mixtures from multiple
43
44 353 markers, detailed information about the identified cross-contained species in a sample and the related markers
45
46 354 are specified in the Additional file 2; Table S1. Even with the default OTU abundance threshold in place, the
47
48 355 species *L. sativa*, *B. taurus* and *Gallus gallus* were identified in mixtures that were not supposed to contain these
49
50 356 species. To verify whether these putative contaminations occurred during DNA isolation or Illumina sequencing,
51
52 357 qPCR assays for the specific detection of *B. taurus* and *G. gallus* were performed on selected DNA extracts. The
53
54 358 high Cq values above 39 indicated the presence of these species, however, in low copy number, which suggests
55
56 359 that for some experimental mixtures (EM8, EM9 and EM14) cross-contamination had occurred during sample
57
58 360 preparation or DNA isolation, while for other experimental mixtures (EM15) cross-contamination may have
59
60
61
62

361 occurred during PCR, Illumina library preparation or sequencing. In addition to these contaminants, a species of
1
2 362 *Brassica* was identified in experimental mixtures containing *P. brassica*. This result is most likely not a false-
3
4 363 positive, because the caterpillars used for this study had been fed on cabbage.

5
6 364 The DNA metabarcoding method was found to be sensitive enough to identify most plant and animal taxa at 1%
7
8 365 (dry mass: dry mass) in mixtures of both low (EM1, EM3 and EM5; Table 2) and relatively high complexity
9
10 366 (EM6, EM8, EM11, EM12, and EM14; Table 2). The exception being *Parapenaeopsis* sp. (all mixtures), *A.*
11
12 367 *anguilla* in EM6, and *Cycas revoluta* in EM8 and EM11. Careful inspection of the NGS data revealed that in
13
14 368 nearly all cases OTUs related to *Parapenaeopsis* sp., *A. anguilla*, and *C. revoluta* were present, but that these
15
16 369 sequences had been filtered out by the CITESpeciesDetect pipeline because their cluster sizes did not fulfil the
17
18 370 0.2% OTU abundance threshold. There appeared to be no trend as to the type and length of DNA barcode marker
19
20 371 that had been filtered out by the CITESpeciesDetect pipeline. For instance, *Parapenaeopsis* sp. was detected
21
22 372 below the OTU threshold with *cyt b*, mini-16S, COI, and 16S markers (not shown). Lowering the OTU
23
24 373 abundance threshold, however, would lead to (more) false-positive identifications, and this was therefore not
25
26 374 implemented.

27
28 375 The repeatability of the laboratory procedure (excluding NGS) was assessed by analysing the
29
30 376 experimental mixtures 10 and 11 (EM10R and EM11R; Table 2), which was independently performed twice, i.e.
31
32 377 DNA isolation and PCR barcode amplification, but NGS was performed on the same MiSeq flow cell as the
33
34 378 other samples of the pilot study. From the comparison, it was observed that the percentage of QC reads was
35
36 379 nearly twice as high in the replicate analyses (Table 3). Also, the percentage of QC reads assigned to DNA
37
38 380 barcodes varied among replicate analyses (Figure 2). Most notable were the observed differences among
39
40 381 replicate analyses in the percentage reads assigned to *matK* and the *trnL* (P6 loop). For example, the percentage
41
42 382 of QC reads assigned to *matK* were 6.11% (14081 reads) and 0.02% (97 reads) in EM10 and EM10R
43
44 383 respectively (Figure 2). The low number of reads assigned to *matK* limited its use for taxonomy identification in
45
46 384 EM10R (Table 4). The multi-locus approach, however, allowed for the repeatable identification of taxa in EM10
47
48 385 and EM11, though not in all cases with all DNA barcode markers (Table 4 and 5).

49
50 386 Based on the results obtained from the pilot study, precautions were taken when grinding the freeze-dried
51
52 387 materials and subsequent mixing to avoid cross-contamination during the laboratory handling of samples, which
53
54 388 were used to improve the SOP for the interlaboratory trial (see protocols in
55
56 389 [21])([dx.doi.org/10.17504/protocols.io.ixbcfin](https://doi.org/10.17504/protocols.io.ixbcfin)). Also, control species were added to experimental mixtures that

390 were prepared for the inter-laboratory trial to allow better confirmation of sample homogeneity and to verify that
1
2 391 no cross-contamination had occurred during sample preparation.

3
4 392

5 6 393 **Assessment of interlaboratory reproducibility of the DNA metabarcoding procedure**

7
8 394 Altogether 16 laboratories from 11 countries (all experienced, well-equipped and proficient in advanced
9
10 395 molecular analysis work), including two of the method developers, participated in the inter-laboratory trial
11
12 396 (Table 6). The laboratories received ten anonymously labelled samples, each consisting of 250 mg powdered
13
14 397 material. Two of the samples, labelled S3 and S8, were authentic TM products seized by the Dutch Customs
15
16 398 Laboratory while the other eight samples were well-characterized mixtures of specimens from carefully
17
18 399 identified taxa in relative dry weight concentrations from 1% to 47% (Table 7). In all experimental mixtures, 1%
19
20 400 of *Zea mays* was added as quality control for homogeneity, which was confirmed with maize-specific *hmg* (high-
21
22 401 mobility group gene) qPCR [16]. Also, tests performed with species-specific qPCR assays indicated that cross-
23
24 402 contamination did not occur during sample preparation (Additional file 1; Table S6). The qPCR assay for the
25
26 403 detection of *Brassica napus*, however, also gave a positive signal for other *Brassica* sp. in the mixtures.

27
28 404 Together with the sample materials, reagents for DNA extraction, and the complete set of barcode
29
30 405 primers, the participants received an obligatory SOP. Any deviations from the SOP had to be reported. The
31
32 406 participants were instructed to extract DNA, perform PCR using the barcode primers, purify the amplified DNA
33
34 407 by removal of unincorporated primers and primer dimers, and assess the quality and quantity of the amplification
35
36 408 products by gel electrophoresis and UV-spectrophotometry. The purified PCR products were then collected by
37
38 409 the coordinator of the trial (RIKILT Wageningen University & Research, the Netherlands) and shipped to a
39
40 410 sequencing laboratory (BaseClear, the Netherlands) for Illumina sequencing using MiSeq PE300 technology.
41
42 411 The sequencing laboratory performed Index PCR and Illumina library preparation prior to MiSeq sequencing as
43
44 412 specified in the Illumina 16S metagenomics sequencing library preparation guide. The altogether 160 PCR
45
46 413 samples were sequenced using two Illumina flow cells with MiSeq reagent kit v3.

47
48 414 The interlaboratory trial should ideally have included the use of the online version of the pipeline, but
49
50 415 unfortunately this was not possible due to shortage of time. Therefore, a single (developer) laboratory performed
51
52 416 these bioinformatics analyses. The 160 individual samples contained on average 269,057 raw reads, and more
53
54 417 than 150,000 reads per sample in 95% of the samples (Additional file 1; Table S7). One sample contained less
55
56 418 than 100,000 reads (51,750), which was considered more than sufficient for reliable species identification. After
57
58 419 pre-processing, the samples contained on average 142,938 (pseudo)reads. On average 94.66% of the reads (min
59
60
61
62

420 = 88.12%, max = 98.02%) passed the quality filtering indicating that the overall quality of the sequence data was
1
2 421 consistently high across the 160 datasets.
3
4 422 OTU-clustering at 98% sequence similarity on average assigned 78.14% of the pre-processed and DNA barcode
5
6 423 assigned reads into OTUs (Additional file 1; Table S7). Only two samples, both from the same laboratory, had a
7
8 424 slightly lower percentage of the (pseudo-)reads assigned to OTUs (66.02% and 66.05%). This indicates that the
9
10 425 pipeline correctly removed PCR artefacts in the clustering phase.

11
12 426 For taxonomy prediction, an OTU would be assigned to a database hit if they aligned with $\geq 98\%$
13
14 427 sequence identity and $\geq 90\%$ query coverage, and yielded an expect value (E-value) of at least 0.001. The
15
16 428 BLAST output of the NGS data was interpreted by participants according to the guidelines in the SOP. Variation
17
18 429 was observed among laboratories in interpreting the BLAST output: some laboratories consistently scored the
19
20 430 top hits, irrespective of bitscore, while other labs selected all hits belonging to the top three bitscores, or
21
22 431 interpreted only the first OTU of each DNA barcode, leading to large differences in identified taxa. Because of
23
24 432 these inconsistencies, the BLAST results were re-interpreted by RIKILT Wageningen University & Research
25
26 433 following the established guideline as mentioned in the SOP. These re-interpreted data are the data referred to in
27
28 434 the following sections.

29
30 435 With one exception, all taxa mixed in at $\geq 1\%$ (dry mass: dry mass) were reproducibly identified by at
31
32 436 least 13 (81%) laboratories (Table 7). *Beta vulgaris* in sample S6 could only be identified by 4 out of 16 (25%)
33
34 437 laboratories. *Beta vulgaris* specific sequences were present in all remaining datasets, but at very low read counts.
35
36 438 So these clusters did not fulfil the 0.2% OTU abundance threshold (Additional file 2; Table S2) . In order to
37
38 439 provide insight into what alternative setting of the CITESspeciesDetect pipeline may have been better suited for
39
40 440 identifying *Beta vulgaris*, three data sets with relatively low (S6 – laboratory 13), medium (S6 – laboratory 14)
41
42 441 and high (S6 – laboratory 6) data volumes were reanalysed using a range of different settings for the OTU
43
44 442 minimum cluster size and OTU abundance threshold (Additional file 2: Table S3-S5). Setting the OTU
45
46 443 minimum cluster size to 2, 4, or 6 has no effect on taxon identification, and *Beta vulgaris* is not identified at the
47
48 444 species or higher taxonomic level in the data sets of laboratories 6 and 13. Setting the OTU abundance threshold
49
50 445 to zero allows identifying *Beta vulgaris* in all three samples, but at the expense of many false positive
51
52 446 identifications. Applying an OTU abundance threshold of 0.1% (default is 0.2%) allows identifying *Beta*
53
54 447 *vulgaris* at the species or genus level irrespective of any differences in data volume between the three samples.

55
56 448 All six animal species could be identified to species level with at least one barcode marker (COI), while
57
58 449 only four of the 12 plant species (*Brassica oleracea*, *Carica papaya*, *Gossypium hirsutum*, and *L. sativa*) could
59
60
61
62

450 be identified to species level (Additional file 2; Table S6). All other plant species were identified at the genus or
1
2 451 higher level. For plants, no single barcode marker was best, and the most reliable data were obtained by
3
4 452 combining the plant barcodes.

5
6 453 Three taxa that were misidentified or not intentionally included in the mixtures were reproducibly
7
8 454 identified across all laboratories. *Acipenser schrenckii* co-occurred in all samples containing *Huso dauricus*. We
9
10 455 have confirmed with DNA metabarcoding that the caviar used for preparing the experimental mixtures contains
11
12 456 both *H. dauricus* and *A. schrenckii* (results not shown). Furthermore, *Brassica rapa* was identified by ITS2 in
13
14 457 sample S4 by all 16 (100%) laboratories, instead of *Brassica napus*. We confirmed by Sanger sequencing *rbcL*
15
16 458 and *matK* that our reference specimen is indeed *Brassica napus*, but that its ITS2 sequence is identical to
17
18 459 *Brassica rapa* (LT718651). Finally, a taxon of the plant family Phellinaceae was reproducibly identified (by all
19
20 460 laboratories) using the mini-*rbcL* marker in all samples containing *L. sativa* (S6, S7, S9, S10). Species of the
21
22 461 family Phellinaceae and *L. sativa* both belong to the order Asterales. The evidence for Phellinaceae was not
23
24 462 strong, i.e. the family-level identification was based on a single NCBI reference sequence only (GenBank:
25
26 463 X69748). We therefore suspect a misidentification during the interpretation of the BLAST results.

27
28 464 Taxa that were identified to be the result of possible contaminations were scarcely observed, i.e. these
29
30 465 were found in isolated cases and could possibly be explained by cross-sample contamination that may have
31
32 466 occurred during any step of sample processing (DNA isolation, PCR, NGS library preparation or NGS). For
33
34 467 example, a contamination with *Gossypium* sp. was observed using *trnL* (P6 loop) in sample S1 of one of the
35
36 468 participating labs. A total of 6 of such suspected cases of incidental cross-contaminations were observed (not
37
38 469 shown).

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

480 known human pathogenic fungus. For sample S8, the animal species *B. taurus* and *Homo sapiens*, the plant
1
2 481 species *Sanguisorba officinalis* and *Eleutherococcus sessiliflorus*, and members of the plant genera *Croton* and
3
4 482 *Erythroxylum*, and families Meliaceae and Asteraceae, were reproducibly identified (Table 9).
5

6 483 **Discussion**

7 484
8 485
9 486 In this study, a DNA metabarcoding method was developed using a multi-locus panel of DNA barcodes for the
10
11 487 identification of CITES protected species in highly complex products such as TMs. As a first step, a CTAB
12
13 488 DNA isolation method was selected for efficiently extracting high quality DNA from pure plant and animal
14
15 489 reference materials as well as from complex mixtures. DNA isolation can be very difficult to standardise and
16
17 490 optimise because of the complexity and diversity of wild life forensic samples, and a more systematic
18
19 491 comparison of different DNA extraction methods is required. Secondly, a single PCR protocol, suitable for all
20
21 492 the barcodes included, i.e. multiple universal plant and animal barcode and mini-barcode markers, was identified.
22
23 493 This facilitated the design of a multi-locus panel of DNA barcodes. Furthermore, the developed DNA
24
25 494 metabarcoding method includes a dedicated bioinformatics workflow, named CITESpeciesDetect, that was
26
27 495 specifically developed for the analysis of Illumina paired-end reads. The developed pipeline requires skilled
28
29 496 experts in bioinformatics, and applies scripts for command-line processing. NGS data analysis pipelines may
30
31 497 provide a lot of flexibility to the user, as modifications are easily implemented by expert users. The design of the
32
33 498 pipeline prevented *cyt b* and COI full-length barcodes to be separated from their corresponding mini-barcodes,
34
35 499 as they have identical forward primers. Since, the 300 PE reads can read through the *cyt b* and COI mini-
36
37 500 barcodes, and therefore contain both 5' primer and 3' primer information, separation should be feasible.
38

39
40 501 To simplify the inter-laboratory validation of the pipeline, a user-friendly and intuitive web-interface
41
42 502 with associated “Help” functions and “FAQs” was developed for the CITESpeciesDetect pipeline. The web
43
44 503 interface was, however, not available in the course of the interlaboratory trial. Therefore, the sequence data
45
46 504 generated in the interlaboratory study could not be analysed by the individual laboratories using the
47
48 505 CITESpeciesDetect pipeline. A single (developer) laboratory therefore performed these analyses. Upon the
49
50 506 availability of the online web-interface, individual participants were later given the opportunity to reanalyse their
51
52 507 DNA metabarcoding data. Observations made in this part demonstrated concordance of results with those
53
54 508 obtained by the developing laboratory, reinforcing the perception of CITESpeciesDetect as a user-friendly and
55
56 509 reliable pipeline that may readily be used by enforcement agencies and other laboratories.
57

58 510 The performance of the DNA metabarcoding method was assessed in an interlaboratory trial in which
59
60 511 the method was found to be highly reproducible across laboratories, and sensitive enough to identify species
61

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

539 While the use of well-characterised experimental mixtures allowed for an assessment of the
540 performance of the DNA metabarcoding method under ideal conditions, the amplifiable DNA content of real-life
541 samples encountered in routine diagnostic work are often of an unpredictable and variable quality. An analysis of

542 two authentic TM products seized by the Dutch Customs Laboratory demonstrated that only few ingredients
1
2 543 listed on the labels could be reproducibly identified. This does not mean that the undetected species were not
3
4 544 used as ingredients. Ingredients may have been processed in such a way that the DNA is either degraded or
5
6 545 effectively removed. This is e.g. the case with refined oils or cooked ingredients [25]. A PCR-free targeted DNA
7
8 546 capturing approach coupled with shotgun sequencing was recently proposed for biodiversity assessments which
9
10 547 may potentially also be suitable for enhancing species identification in difficult wildlife forensic samples [24,
11
12 548 26]. The quality of the sequence reference database also strongly affects the ability to correctly identify species.
13
14 549 Without correct references that also exhibit the necessary intraspecific variation, it is not possible to match and
15
16 550 discriminate sequence reads correctly. It is well-known that accurate DNA barcoding depends on the use of a
17
18 551 reference database that provides good taxonomic coverage [5, 9]. The current underrepresentation of DNA
19
20 552 barcodes from species protected by CITES and closely related species critically hampers their identification. We
21
22 553 estimate that only 18.8% of species on the CITES list contain one or more DNA barcodes (COI for animals, and
23
24 554 *matK* or *rbcL* for plants). This will improve as DNA barcoding campaigns continue, in particular through
25
26 555 initiatives such as the Barcode of Wildlife Project (BWP; www.barcodeofwildlife.org). Only by expansion of the
27
28 556 sequence reference database of endangered and illegally-traded species can DNA barcoding provide the
29
30 557 definitiveness required in a court of law.

31
32 558 A noteworthy observation was that most species that were reproducibly identified did not appear on the
33
34 559 ingredients lists on the labels of the analysed TMs. This is possibly due to mislabelling. If the identifications are
35
36 560 correct this also indicates that consumption may pose health risks. These findings corroborate earlier reports that
37
38 561 DNA metabarcoding may provide valuable information about the quality and safety of TMs [5, 6].
39

40 562

42 563 **Potential implications**

43 564
44 565 Overall, our findings demonstrate that the multi-locus DNA metabarcoding method assessed in this study can
45
46 566 provide reliable and detailed data on the composition of highly complex food products and supplements. This
47
48 567 study highlights the necessity of a multi-locus DNA metabarcoding strategy for species identification in complex
49
50 568 samples, since the use of multiple barcode markers can enable an increased resolution and quality assurance,
51
52 569 even in heavily processed samples. The developed robust bioinformatics pipeline for Illumina data analysis with
53
54 570 user-friendly web interface allows the method to be directly applied in various fields such as: a) food
55
56 571 mislabelling and fraud in the food industry [27], b) environmental monitoring of species [28], and c) wildlife
57
58 572 forensics [29]. Furthermore, the pipeline can be readily used to analyse different types of Illumina paired-end
59
60
61
62

573 datasets, even the future Illumina datasets (read length > 300 nt). Additionally, the web interface provides an
1 opportunity for the global audience with limited expertise in bioinformatics, to analyse their own data. It also
2 574 provides the liberty to select different primer sets and customise the settings for the selected purposes. As a result,
3
4 575 the range of potential applications of the method to identify plant and animal species is diverse, the pipeline is
5
6 576 versatile and adjustable to the user's needs, thus providing a powerful tool for research as well as enforcement
7
8 577 purposes.
9
10 578

11 579 12 13 14 580 **Methods**

15 581 16 582 17 583 **Reference materials and preparation of experimental mixtures**

18
19 584 All reference specimens were obtained from a local shop in the Netherlands or provided by the Dutch Customs
20
21 585 Laboratory (Additional file 1; Table S2 and Table S3). The reference specimens were taxonomically
22
23 586 characterised to the finest possible taxonomic level. For each species, it was checked whether reference
24
25 587 sequences were present in NCBI GenBank. For taxonomic confirmation, standard COI barcodes for all animal
26
27 588 specimens were generated and individually Sanger sequenced, and compared against the NCBI and BOLD
28
29 589 nucleotide database. For plant species, the DNA barcodes *rbcL* and *matK* were Sanger sequenced to confirm
30
31 590 species identity. For a number of plant and animal species the generated barcode sequence information was
32
33 591 deposited in the European Nucleotide Archive (ENA) under accession numbers LT009695 to LT009705, and
34
35 592 LT718651 (Additional file 1; Table S1).

36
37 593 For the initial pilot study, in which the SOP for the DNA metabarcoding approach was established and
38
39 594 tested, 15 well-defined complex mixtures were artificially prepared (Table 2). These experimental mixtures were
40
41 595 prepared with 2 to 10 taxonomically well-characterised species (Table 2). The ingredients were mixed based on
42
43 596 dry weight ratio, for which individual materials were freeze-dried for 78 hours. The lyophilized ingredients were
44
45 597 ground using an autoclaved mortar and pestle or blender in a cleaned fume hood, and subsequently stored at -
46
47 598 20 °C °C. The individual ingredients of each complex mixture were weighted and mixed thoroughly using a
48
49 599 tumbler (Heidolph Reax 2) for 20 hours and stored at -20 °C until further use.

50
51 600 For the interlaboratory validation trial, in which the applicability and reproducibility of the DNA
52
53 601 metabarcoding method was assessed, eight additional well-characterised mixtures were artificially prepared
54
55 602 using the above procedure. These complex mixtures were prepared with 8 to 11 taxonomically well-
56
57 603 characterised species present at dry weight concentrations from 1% to 47% (Table 7). These complex mixtures
58
59 604 were prepared in such a way that the efficiency of homogenization and possibility of sample cross-contamination
60
61

605 could be verified using species-specific qPCR assays. In all samples, 1% of *Zea mays* was added as quality
1 control for homogeneity. The presence of *Z. mays* was checked after sample mixing using maize-specific *hmg*
2 606 qPCR along with a positive and negative control. A unique species was added at 1% dry weight to each mixture
3
4 607 (S1-*Glycine max*, S2-*Gossypium* sp., S4-*Brassica napus*, S5-*Triticum aestivum*, S6-*Beta vulgaris*, S7-*Meleagris*
5 608 *gallopavo*, S9-*Carica papaya*, S10-*Solanum lycopersicum*) (Table 7). Species-specific qPCR was performed in
6
7 609 duplex (together with positive and negative controls) in all samples, to check for possible cross-contamination
8
9 610 between samples after sample preparation. Information about the qPCR primers and probes, and qPCR
10
11 611 procedure can be found in the Additional file 1; Table S8-S10. In addition to the eight experimental mixtures,
12
13 612 two TMs were included that were obtained from the Dutch Customs Laboratory: a) Ma pak leung sea-dog hard
14
15 613 capsules (MA PAK LEUNG CO, LTD, Hong Kong), was labelled to contain among others rhizoma *Cibotii*
16
17 614 (*Cibotium barometz*, CITES appendix II), and Herba *Cistanche* (*Cistanche* sp., CITES appendix II) and b)
18
19 615 Cobra performance enhancer hard capsules (Gold caps, USA), was labelled to contain among others Siberian
20
21 616 ginseng (*Eleutherococcus senticosus*) and Korean ginseng (*Panax ginseng*). In both TMs, the medicine powder
22
23 617 was encapsulated in a hard-capsule shell. All capsules were opened and the powder inside the capsules were
24
25 618 stored in air-sealed and sterilized containers. The powdered medicines were thoroughly mixed using tumbler
26
27 619 (Heidolph Reax 2) for 20 hours and stored at -20 °C until further use.
28
29 620
30
31 621
32
33 622
34
35 623
36

37 624 **DNA isolation method**

38 625 A cetyltrimethylammonium bromide (CTAB) extraction method [17] was assessed for its ability to efficiently
39
40 626 extract DNA from a range of plant and animal materials (SOP). In brief, the CTAB method consists of an initial
41
42 627 step to separate polysaccharides and organic soluble molecules using a CTAB extraction buffer (1X CTAB,
43
44 628 1.4M NaCl, 0.1 M Tris-HCl [pH 8.0], and 20mM NA₂EDTA) and chloroform. Next, the DNA was precipitated
45
46 629 with 96% ethanol, purified with 70% ethanol, and the obtained DNA was stored at 4 °C until further use. DNA
47
48 630 was extracted from 100 mg reference materials (plant and animal), artificially made complex mixtures, and real-
49
50 631 life samples (TMs) along with an extraction control. The concentration and purity (OD_{260/280} and OD_{260/230} ratios)
51
52 632 of the obtained DNA was determined by spectrophotometer (NanoDrop 1000 instrument, Thermo Fisher
53
54 633 Scientific Inc.). The OD_{260/280} ratios between 1.7 and 2.0 were considered to indicate purity of the obtained DNA.
55
56 634 In case the extraction control contained DNA, the DNA isolation procedure was repeated.
57
58 635
59
60
61
62

636 **Barcode markers**

1
2 637 Candidate universal DNA barcode and mini-barcode markers and primer sets were identified using the
3
4 638 information provided in Staats et al. (2016) [9], supplemented with additional primer sets from literature (Table
5
6 639 1). The PCR primer sets were modified to have an additional Illumina tail sequence at 5' end of the primers
7
8 640 (Table 1).

9
10 641
11
12 642 **PCR**

13
14 643 A gradient PCR was performed with all PCR primer combinations using 10 ng of DNA. The tested PCR
15
16 644 conditions programme were according to the following protocol: 95 °C for 15 min, five cycles of 94 °C for 30 s,
17
18 645 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,
19
20 646 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,
21
22 647 which included 12.5 µl of HotStarTaq Master Mix (Qiagen), 0.5 µl of 10 µM each sense and antisense primer, 7
23
24 648 µl of RNase-free water (Qiagen) and 5 µl of 10 ng/µl of represented species DNA. PCR was performed in the
25
26 649 CFX96 thermal cycler (Bio-Rad) and the amplified products from all the analysed reference specimens,
27
28 650 artificially made complex mixtures, and real-life samples (TMs) together with the positive and negative control
29
30 651 reactions were visualised on 1% agarose gels. If amplification was observed in the negative control, the PCR
31
32 652 analysis was repeated. Prior to NGS library preparation, 8 µl of PCR product of each target (12 in total) per
33
34 653 sample was pooled and mixed. Next, the pooled PCR products were purified using the QIAquick PCR
35
36 654 purification kit (Qiagen) according to manufacturer's protocol, and the purified amplicons were visualized on 1%
37
38 655 agarose gels for all the artificially made complex mixtures, and real-life samples (TMs).

39
40 656
41
42 657
43
44 658 **Next Generation Sequencing**

45
46 659 The pooled and purified PCR amplicons were sequenced using Illumina MiSeq paired-end 300 technology. Prior
47
48 660 to MiSeq sequencing, Index PCR and Illumina library preparation were performed as specified in the Illumina
49
50 661 16S metagenomics sequencing library preparation guide [30]. All the DNA barcode amplicons of each sample
51
52 662 were treated as one sample during library preparation i.e. all DNA barcode amplicons of each sample were
53
54 663 tagged with the addition of the same, unique identifier, or index sequence, during library preparation. The Index
55
56 664 PCR was performed to add dual indices (multiplex identifiers) and Illumina sequencing adapters using the
57
58 665 Nextera XT Index Kit (Illumina, FC-131-1001). The prepared Illumina libraries from each sample were
59
60
61
62

666 quantified using the Quant-iT dsDNA broad range assay (Life Technologies). Furthermore, the normalised
1
2 667 library pools were prepared and their concentration was quantified using KAPA library quantification kit (KAPA
3
4 668 Biosystems) and pooled prior to MiSeq sequencing using MiSeq reagent kit v3.
5
6 669

7 8 670 **Bioinformatics analysis**

9 671
10 672 The raw demultiplexed Illumina reads with Illumina 1.8+ encoding were processed using a bioinformatics
11
12 673 pipeline, called CITESspeciesDetect. The CITESspeciesDetect is composed of five linked tools with data
13
14 674 analysis passing through three phases: 1) pre-processing of paired-end Illumina data involving quality trimming
15
16 675 and filtering of reads, followed by sorting by DNA barcode, 2) OTU clustering by barcode, and 3) taxonomy
17
18 676 prediction and CITES identification (Figure 1).

19
20 677 During preprocessing of reads, the 5' and 3' Illumina adapter sequences are trimmed using Cutadapt v1.9.1
21
22 678 (cutadapt, RRID:SCR_011841) [31] using the respective substrings TGTGTATAAGAGACAG and
23
24 679 CTGTCTCTTATACACA. After Illumina adapter trimming, reads ≤ 10 bp are removed using Cutadapt. Then,
25
26 680 the forward and reverse reads are merged to convert a pair into a single pseudoread containing one sequence and
27
28 681 one set of quality score using USEARCH v8.1.1861 [19].
29

30 682 Next, the merged pseudo-reads, unmerged forward reads and unmerged reverse reads are processed
31
32 683 separately during quality filtering using a sliding window method implemented in PRINSEQ (PRINSEQ,
33
34 684 RRID:SCR_005454) [32]. During this procedure, low quality bases with Phred scores lower than 20 are trimmed
35
36 685 from 3'-end using a window size of 15 nt and a step size of 5 nt. After PRINSEQ, reads with a minimum of 95%
37
38 686 per base quality ≥ 20 are kept, while the remaining reads are removed using FASTX_Toolkit v0.0.14 [32]. Then,
39
40 687 reads are successively selected, trimmed and sorted per DNA barcode marker using Cutadapt [31]. The
41
42 688 following steps are followed for each DNA barcode marker separately during this procedure. First, reads
43
44 689 containing an anchored 5' forward primer or anchored 5' reverse primer (or their reverse complement) are
45
46 690 selected with a maximum error tolerance of 0.2 (=20%) and with the overlap parameter specified to 6 to ensure
47
48 691 specific selection of reads. Also, reads ≤ 10 nt are removed. The anchored 5' primer sequences are subsequently
49
50 692 trimmed. Second, primer sequences that are present at the 3' end of the selected reads are also removed. For each
51
52 693 DNA barcode, the primer-selected and unmerged reverse reads are reverse complemented and combined with
53
54 694 primer-selected merged and unmerged forward reads.
55

56 695 The following procedure is used to cluster the quality trimmed reads of each DNA barcode into OTUs
57
58 696 using the UPARSE pipeline implemented in USEARCH [19] with the following modifications: reads are
59
60
61
62

697 dereplicated using the derep_prefix command. Also, singleton reads and reads with minimum cluster size
1
2 698 smaller than 4 are discarded. Representative OTUs are generated using an OTU radius of 2 (98% identity
3
4 699 threshold) and 0.2% OTU abundance threshold with minimum barcode length per primer set. Filtering of
5
6 700 chimeric reads is performed using the default settings of the UPARSE-REF algorithm implemented in the
7
8 701 cluster_otus command of USEARCH.

9
10 702 To assign OTUs to taxonomy, standalone BLASTn megablast searches (BLASTN, RRID:SCR_001598)
11
12 703 [20] of representative OTUs are performed on the National Centre for Biotechnology Information (NCBI)
13
14 704 GenBank nucleotide database using an Expectation value (E-value) threshold of 0.001 and a maximum of 20
15
16 705 aligned sequences. OTUs are assigned to the database sequence to which they align, based on bit score, and
17
18 706 having at least 98% sequence identity and minimum of 90% query coverage. To identify putative CITES-listed
19
20 707 taxa, the taxon ID first was matched against the NCBI taxonomy database using Entrez Direct (edirect) functions
21
22 708 (available at <ftp://ftp.ncbi.nlm.nih.gov/entrez/entrezdirect/>) to retrieve scientific name (species, genus, family,
23
24 709 order and synonym name). The scientific, synonym and/or family names are then matched against a local CITES
25
26 710 database that is retrieved from <https://speciesplus.net>. The final results are presented as a tab-separated values
27
28 711 file (TSV) containing the BLAST hit metadata (i.e. bit-score, e-value, accession numbers etc.), the scientific
29
30 712 name, synonym name, and in case a CITES-listed taxon was found, also the CITES appendix listing and
31
32 713 taxonomic group (i.e. species, genus, family or order name) under which the taxon is listed by CITES.

33
34 714 The BLAST output was interpreted by following guidelines: first, to minimize the chance of erroneous
35
36 715 species identifications, the same species should have at least three top hits, i.e. highest bit scores. Secondly, if
37
38 716 multiple hits are obtained with identical quality results, but with different assigned species, or with less than
39
40 717 three top hits with same species designation, the OTU fragment was considered to lack the discriminatory power
41
42 718 to refer the hit to species level. In such cases, the OTU would then be downgraded to a genus-level identification.
43
44 719 Thirdly, if multiple hits are obtained with identical quality results, but with different assigned genera, the OTU
45
46 720 fragment lacks the discriminatory power to describe the hit to genus level. In such cases, the OTU would then be
47
48 721 downgraded to a family-level identification. An online web-interface based application for the
49
50 722 CITESpeciesDetect pipeline was developed which is available from [http://decathlon-fp7.citespipe-wur.surf-](http://decathlon-fp7.citespipe-wur.surf-hosted.nl:8080/)
51
52 723 [hosted.nl:8080/](http://decathlon-fp7.citespipe-wur.surf-hosted.nl:8080/). The web-interface facilitates intuitive BLAST identification of species listed by speciesplus.net
53
54 724 by highlighting species on CITES appendix I in red. Species listed on CITES appendix II and II are highlighted
55
56 725 in orange and yellow, respectively.

57
58
59 726
60
61
62

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

1
2 728 A parameter scan was performed in order to assess the effect of software settings on the ability to identify
3
4 729 species. This evaluation allowed for identification of important parameters and their effects on the sensitivity,
5
6 730 specificity and robustness of the procedure. This in turn resulted in specified, recommended (default) parameters
7
8 731 values for analysing DNA metabarcoding datasets using the CITESpeciesDetect pipeline. The effects of the
9
10 732 following parameters were assessed: base quality scores, error tolerance for primer selection, OTU radius, OTU
11
12 733 abundance threshold, expect E-value and query coverage threshold, percentage identity threshold, minimum
13
14 734 DNA barcode length and BLAST database. The parameters scan was performed on experimental mixture 11 of
15
16 735 the pilot study (Table 2). This mixture was selected because of its (relatively) high sample complexity, making it
17
18 736 the most challenging complex mixture to analyse. Furthermore, the parameter scan was limited to four barcode
19
20 737 primer sets: full-length cytochrome-B (*cyt b*), COI mini barcode (mini-COI), *rbcL* mini barcode (mini-*rbcL*) and
21
22 738 the full-length *rbcL* (*rbcL*) barcode.
23

24 739

26 740 **Inter-laboratory validation trial: participants and method.**

27
28 741 To assess the overall performance of the developed DNA metabarcoding approach, 16 laboratories from 11
29
30 742 countries participated in an international inter-laboratory validation. Only laboratories that regularly perform
31
32 743 molecular analyses and have well-equipped laboratory facilities were selected to participate (Table 6). The
33
34 744 majority are governmental or semi-official institutes and are considered highly authoritative within each
35
36 745 respective country. Participants were requested to follow the SOP [21], and were asked to document any
37
38 746 deviations that were made. The chemicals and reagents that were provided to the laboratories were: 10 samples
39
40 747 (eight experimental mixtures and two TMs), *B. taurus* and *L. sativa* positive control DNA, CTAB extraction and
41
42 748 precipitation buffer, 1.2 M NaCl solution, 12 universal plant and animal barcode and mini-barcode primer sets
43
44 749 (Table 1), Qiagen HotStarTaq master mix, and Qiagen PCR purification kits. All reagents and samples were
45
46 750 provided in quantities corresponding to 2.5× the amounts required for the planned experiments. After following
47
48 751 the SOP from DNA isolation to purification of the amplified products, all the purified samples from all the
49
50 752 laboratories (n=160) were collected and sequenced using Illumina MiSeq paired-end 300 technology (at
51
52 753 BaseClear, Leiden, NL). The Index PCR and Illumina library preparation was performed according to the
53
54 754 guideline and all 160 samples were sequenced on two Illumina flow cells. After Illumina MiSeq run, the raw
55
56 755 NGS data was processed using the default settings of the CITESpeciesDetect pipeline. BLAST outputs for the
57
58
59
60
61
62

756 samples were distributed back to the participating laboratories for interpretation of results. The laboratories
1
2 757 interpreted the BLAST output based on the guideline provided in the SOP.

3
4 758

5 759 **Availability of supporting data**

6
7 760 All the sequence data obtained from the pilot study and the international interlaboratory validation trial, the
8
9 761 CITESspeciesDetect pipeline and access to web interface are freely available. The generated barcode sequence
10
11 762 information for some animal and plant species were deposited in GenBank under the accession numbers
12
13 763 LT009695 to LT009705, and LT718651 (Additional file 1; Table S1). The Illumina PE300 MiSeq data obtained
14
15 764 from the pilot study and the international interlaboratory validation trial (n=177) were deposited to ENA with
16
17 765 study ID PRJEB18620. The script for the CITESspeciesDetect pipeline is available at GitHub. The web interface
18
19 766 for CITESspeciesDetect pipeline can be accessed via the following link: [26
27 770](http://decathlon-fp7.citespipe-wur.surf-
20
21 767 <u>hosted.nl:8080/</u>. The access to analysis via the web interface will be provided on request. SOP protocols are
22
23 768 available from protocols.io [21] and snapshots of the code and example results are available from the
24
25 769 <i>GigaScience</i> database[34].</p></div><div data-bbox=)

28 771 **Availability and requirements**

29
30
31 772 Project name: CITESspeciesDetect

32
33
34 773 Project home page: <https://github.com/RIKILT/CITESspeciesDetect>

35
36 774 Operating system(s): Linux

37
38 775 Programming language: Python and Bash

39
40 776 Other requirements: none

41
42 777 License: BSD 3-Clause License

43
44
45 778 Any restrictions to use by non-academics: none

46
47 779

48
49 780

50
51 781

52
53 782

54
55 783

56
57 784

58
59 785

60
61 786

787 **Additional files**

1 788 **Additional file 1: Table S1** Accession numbers of DNA barcode sequences of plant and animal species. **Table**
2
3 789 **S2** PCR success rate for animal reference species. **Table S3** PCR success rate for plant reference species. **Table**
4
5 790 **S4** Statistics of different quality filtering settings for four DNA barcodes. **Table S5** BLAST identification of
6
7 791 species with different quality filtering settings for four DNA barcodes. **Table S6** Results of species-specific
8
9 792 qPCR performed on the experimental mixtures prepared for the inter-laboratory validation trial. **Table S7**
10
11 793 Interlaboratory trial study: average number of Illumina reads per sample, the average number of (pseudo)reads
12
13 794 that passed quality control (QC) and the percentage of QC (pseudo)reads that were assigned to DNA barcodes
14
15 795 and Operational Taxonomic Units (OTUs). **Table S8** qPCR primer and probe information. **Table S9** qPCR
16
17 796 reagent composition. **Table S10** qPCR thermocycling program. (*.docx).
18
19
20 797

21
22 798 **Additional file 2: Table S1** Pilot study: Composition of the experimental mixtures, and taxa identified using the
23
24 799 default settings of the CITESpeciesDetect pipeline. **Table S2** Interlaboratory trial: *Beta vulgaris* observed in the
25
26 800 sample S6 data sets generated by the 16 laboratories. **Table S3-S5** Interlaboratory trial: Assessment of the effect
27
28 801 of different settings (OTU clusters size, OTU abundance threshold) of the CITESpeciesDetect pipeline on the
29
30 802 identification of taxa using different data volume (low, medium and high) generated by three laboratory for S6.
31
32 803 **Table S6** Interlaboratory trial: the taxonomic resolution provided by each DNA barcode marker for eight
33
34 804 experimental mixtures (*.xlsx).
35

36 805
37 806
38
39 807 **Additional file 3: Table S1** ENA accession numbers of all raw NGS datasets obtained in this study (*.xlsx).
40
41 808

42 809 **Abbreviations**

43
44 810 CITES: Convention on International trade in Endangered Species of Wild fauna and flora; TMs: Traditional
45
46 811 Medicines; NGS: Next generation sequencing; CTAB: cetyltrimethylammonium bromide; COI: Cytochrome c
47
48 812 oxidase subunit I; *cyt b*: Cytochrome *b* gene; 16S rDNA: 16S ribosomal DNA; *matK*: Maturase K gene; *rbcL*:
49
50 813 ribulose-1,5-bisphosphate carboxylase large subunit gene; ITS2: Internal transcribed spacer region 2;; SOP:
51
52 814 Standard operating procedure; OTU: Operational Taxonomic Unit; BLAST: Basic Local Alignment Search Tool.
53

54 815

56 816 **Competing interests**

57
58 817 The authors declare that they have no competing interest.
59
60 818
61
62

819

1 820 **Funding**

2
3 821 The DECATHLON project has been funded with support from the European Commission in the context of the
4
5 822 Seventh Framework Programme (FP7). This publication and all its contents reflect the views only of the authors,
6
7 823 and the Commission cannot be held responsible for any use, which may be made of the information contained
8
9 824 therein.

10
11 825
12 826 **Authors' Contributions**

13
14 827 AJA and MS shared the first authorship. AJA, MS, MV, TP, AC, EK conceived and designed the experiments
15
16 828 for the pilot study. AJA performed the experiments for the pilot study. MS, RH, AJA developed the
17
18 829 CITESpeciesDetect pipeline. AJA, MS, RH analysed the NGS data obtained from the pilot study. AJA, MS,
19
20 830 MV, TP, TWP, IS, EK, FG, MTBC, AHJ involved in establishing the Standard Operation Procedure for the
21
22 831 validation trial. AJA, MS, MV, TP, EK conceived and designed the experiments for the validation trial. FG,
23
24 832 MTBC, AHJ, AJA, MS involved in coordinating the trial. AJA, MV prepared the samples and materials for the
25
26 833 validation trial and distributed to the participated laboratories. FR, MS, RH involved in developing the web-
27
28 834 interface. MS, TP, DD, MBI, MBU, EH, RHO, AK, LL, CN, HN, EP, JPR, RS, TS, CVM took part in the
29
30 835 validation trial. AJA, MS, RH, MV analysed the NGS data obtained from the validation trial. AJA, MS, RH, MV,
31
32 836 SVR, EK contributed to the writing of the manuscript. All authors read and approved the final manuscript.

33
34 837
35 838 **Acknowledgements**

36
37
38 839 This work was supported by the DECATHLON project, which was funded by the European Commission under
39
40 840 Seventh Framework Programme (FP7).

41
42 841
43 842

44 843 **Reference:**

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

- 859 6. Coghlan ML, Maker G, Crighton E, Haile J, Murray DC, White NE, Byard RW, Bellgard MI, Mullaney
1 860 I, Trengove R: **Combined DNA, toxicological and heavy metal analyses provides an auditing**
2 861 **toolkit to improve pharmacovigilance of traditional Chinese medicine (TCM).** *Scientific Reports*
3 862 2015, **5**.
- 4 863 7. Ivanova NV, Kuzmina ML, Braukmann TWA, Borisenko AV, Zakharov EV: **Authentication of**
5 864 **Herbal Supplements Using Next-Generation Sequencing.** *PLOS One* 2016, **11**:e0156426.
- 6 865 8. Taberlet P, Coissac E, Pompanon F, Brochmann C, Willerslev E: **Towards next-generation**
7 866 **biodiversity assessment using DNA metabarcoding.** *Molecular Ecology* 2012, **21**:2045-2050.
- 8 867 9. Staats M, Arulandhu AJ, Gravendeel B, Holst-Jensen A, Scholtens I, Peelen T, Prins TW, Kok E:
9 868 **Advances in DNA metabarcoding for food and wildlife forensic species identification.** *Analytical*
10 869 *and Bioanalytical Chemistry* 2016:1-16.
- 11 870 10. Fahner NA, Shokralla S, Baird DJ, Hajibabaei M: **Large-scale monitoring of plants through**
12 871 **environmental DNA metabarcoding of soil: recovery, resolution, and annotation of four DNA**
13 872 **markers.** *PLOS One* 2016, **11**:e0157505.
- 14 873 11. Arulandhu AJ, Staats M, Peelen T, Kok E: **DNA metabarcoding of endangered plant and animal**
15 874 **species in seized forensic samples.** In *Genome*. 2015: 188-189.
- 16 875 12. Taylor H, Harris W: **An emergent science on the brink of irrelevance: a review of the past 8 years**
17 876 **of DNA barcoding.** *Molecular Ecology Resources* 2012, **12**:377-388.
- 18 877 13. Little DP: **A DNA mini-barcode for land plants.** *Molecular Ecology Resources* 2014, **14**:437-446.
- 19 878 14. Parveen I, Gafner S, Techen N, Murch SJ, Khan IA: **DNA Barcoding for the Identification of**
20 879 **Botanicals in Herbal Medicine and Dietary Supplements: Strengths and Limitations.** *Planta*
21 880 *Medica* 2016, **82**:1225-1235.
- 22 881 15. Chen R, Dong J, Cui X, Wang W, Yasmeen A, Deng Y, Zeng X, Tang Z: **DNA based identification of**
23 882 **medicinal materials in Chinese patent medicines.** *Scientific Reports* 2012, **2**:958.
- 24 883 16. Scholtens I, Laurensse E, Molenaar B, Zaaier S, Gaballo H, Boleij P, Bak A, Kok E: **Practical**
25 884 **experiences with an extended screening strategy for genetically modified organisms (GMOs) in**
26 885 **real-life samples.** *Journal of agricultural and food chemistry* 2013, **61**:9097-9109.
- 27 886 17. Murray M, Thompson WF: **Rapid isolation of high molecular weight plant DNA.** *Nucleic Acids*
28 887 *Research* 1980, **8**:4321-4326.
- 29 888 18. Ivanova NV, Zemlak TS, Hanner RH, Hebert PDN: **Universal primer cocktails for fish DNA**
30 889 **barcoding.** *Molecular Ecology Notes* 2007, **7**:544-548.
- 31 890 19. Edgar RC: **Search and clustering orders of magnitude faster than BLAST.** *Bioinformatics* 2010,
32 891 **26**:2460-2461.
- 33 892 20. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and**
34 893 **PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Research* 1997,
35 894 **25**:3389-3402.
- 36 895 21. Alfred J. Arulandhu, Martijn Staats, Rico Hagelaar, Marleen M. Voorhuijzen, Theo W. Prins, Ingrid
37 896 Scholtens, Tamara Peelen and Esther Kok (2017): Development and validation of a multi-locus DNA
38 897 metabarcoding method to identify endangered species in complex samples SOP. protocols.io.
39 898 <http://dx.doi.org/10.17504/protocols.io.ixbcfin>
- 40 899 22. Hebert PD, Cywinska A, Ball SL, deWaard JR: **Biological identifications through DNA barcodes.**
41 900 *Proceedings of the Royal Society of London B: Biological Sciences* 2003, **270**:313-321.
- 42 901 23. Group CPW, Hollingsworth PM, Forrest LL, Spouge JL, Hajibabaei M, Ratnasingham S, van der Bank
43 902 M, Chase MW, Cowan RS, Erickson DL: **A DNA barcode for land plants.** *Proceedings of the*
44 903 *National Academy of Sciences* 2009, **106**:12794-12797.
- 45 904 24. Liu S, Li Y, Lu J, Su X, Tang M, Zhang R, Zhou L, Zhou C, Yang Q, Ji Y: **SOAPBarcode: revealing**
46 905 **arthropod biodiversity through assembly of Illumina shotgun sequences of PCR amplicons.**
47 906 *Methods in Ecology and Evolution* 2013, **4**:1142-1150.
- 48 907 25. Gryson N: **Effect of food processing on plant DNA degradation and PCR-based GMO analysis: a**
49 908 **review.** *Analytical and Bioanalytical Chemistry* 2010, **396**:2003-2022.
- 50 909 26. Tang M, Hardman CJ, Ji Y, Meng G, Liu S, Tan M, Yang S, Moss ED, Wang J, Yang C: **High -**
51 910 **throughput monitoring of wild bee diversity and abundance via mitogenomics.** *Methods in Ecology*
52 911 *and Evolution* 2015, **6**:1034-1043.
- 53 912 27. Galimberti A, De Mattia F, Losa A, Bruni I, Federici S, Casiraghi M, Martellos S, Labra M: **DNA**
54 913 **barcoding as a new tool for food traceability.** *Food Research International* 2013, **50**:55-63.
- 55 914 28. Taberlet P, Coissac E, Hajibabaei M, Rieseberg LH: **Environmental DNA.** *Molecular Ecology* 2012,
56 915 **21**:1789-1793.
- 57 916 29. Iyengar A: **Forensic DNA analysis for animal protection and biodiversity conservation: a review.**
58 917 *Journal for Nature Conservation* 2014, **22**:195-205.

- 918 30. **16s Metagenomic Sequencing Library Preparation. Illumina document 15044223.**
 919 [https://support.illumina.com/content/dam/illumina-](https://support.illumina.com/content/dam/illumina-support/documents/documentation/chemistry_documentation/16s/16s-metagenomic-library-prep-guide-15044223-b.pdf)
 920 [support/documents/documentation/chemistry_documentation/16s/16s-metagenomic-library-prep-guide-](https://support.illumina.com/content/dam/illumina-support/documents/documentation/chemistry_documentation/16s/16s-metagenomic-library-prep-guide-15044223-b.pdf)
 921 [15044223-b.pdf](https://support.illumina.com/content/dam/illumina-support/documents/documentation/chemistry_documentation/16s/16s-metagenomic-library-prep-guide-15044223-b.pdf)
- 922 31. Martin M: **Cutadapt removes adapter sequences from high-throughput sequencing reads.** *EMBnet*
 923 *journal* 2011, **17**:10-12.
- 924 32. Schmieder R, Edwards R: **Quality control and preprocessing of metagenomic datasets.**
 925 *Bioinformatics* 2011, **27**:863-864.
- 926 33. **FASTX_Toolkit v0.0.14** http://hannonlab.cshl.edu/fastx_toolkit/
- 927 34. Arulandhu, A, J; Hagelaar, R; Staats, M; Voorhuijzen, M, M; Prins, T, W; Scholtens, I, M; Costessi, A;
 928 Duijsings, D; Rechenmann, F; Gaspar, F, B; Barreto Crespo, M, T; Holst-Jensen, A; Birck, M; Burns,
 929 M; Haynes, E; Hochegger, R; Klingl, A; Lundberg, L; Natale, C; Niekamp, H; Perri, E; Barbante, A;
 930 Rosec, J; Seyfarth, R; Sovova, T; Moorleggem, C, V; Ruth, S, V; Peelen, T; Kok, E (2017): Supporting
 931 data for "Development and validation of a multi-locus DNA metabarcoding method to identify
 932 endangered species in complex samples" GigaScience Database. <http://dx.doi.org/10.5524/100330>
- 933 35. Palumbi S, Martin A, Romano S, McMillan W, Stice L, Grabowski G: **The Simple Fool's Guide to**
 934 **PCR, Version 2.0, privately published document compiled by S. Palumbi. Dept. Zoology, Univ**
 935 **Hawaii, Honolulu, HI 1991, 96822.**
- 936 36. Sarri C, Stamatis C, Sarafidou T, Galara I, Godosopoulos V, Kolovos M, Liakou C, Tastsoglou S,
 937 Mamuris Z: **A new set of 16S rRNA universal primers for identification of animal species.** *Food*
 938 *Control* 2014, **43**:35-41.
- 939 37. Leray M, Yang JY, Meyer CP, Mills SC, Agudelo N, Ranwez V, Boehm JT, Machida RJ: **A new**
 940 **versatile primer set targeting a short fragment of the mitochondrial COI region for**
 941 **metabarcoding metazoan diversity: application for characterizing coral reef fish gut contents.**
 942 *Front Zool* 2013, **10**:34.
- 943 38. Geller J, Meyer C, Parker M, Hawk H: **Redesign of PCR primers for mitochondrial cytochrome c**
 944 **oxidase subunit I for marine invertebrates and application in all-taxa biotic surveys.** *Molecular*
 945 *Ecology Resources* 2013, **13**:851-861.
- 946 39. Parson W, Pegoraro K, Niederstätter H, Föger M, Steinlechner M: **Species identification by means of**
 947 **the cytochrome b gene.** *International Journal of Legal Medicine* 2000, **114**:23-28.
- 948 40. Fazekas AJ, Kuzmina ML, Newmaster SG, Hollingsworth PM: **DNA barcoding methods for land**
 949 **39.** *Methods in Molecular Biology* 2012, **858**:223-252.
- 950 41. Cuénoud P, Savolainen V, Chatrou LW, Powell M, Grayer RJ, Chase MW: **Molecular phylogenetics**
 951 **of Caryophyllales based on nuclear 18S rDNA and plastid rbcL, atpB, and matK DNA sequences.**
 952 *American Journal of Botany* 2002, **89**:132-144.
- 953 42. Levin RA, Wagner WL, Hoch PC, Nepokroeff M, Pires JC, Zimmer EA, Sytsma KJ: **Family-level**
 954 **relationships of Onagraceae based on chloroplast rbcL and ndhF data.** *American Journal of Botany*
 955 2003, **90**:107-115.
- 956 43. Kress WJ, Erickson DL: **A two-locus global DNA barcode for land plants: the coding rbcL gene**
 957 **complements the non-coding trnH-psbA spacer region.** *PLOS One* 2007, **2**:e508.
- 958 44. Fazekas AJ, Burgess KS, Kesanakurti PR, Graham SW, Newmaster SG, Husband BC, Percy DM,
 959 Hajibabaei M, Barrett SC: **Multiple multilocus DNA barcodes from the plastid genome**
 960 **discriminate plant species equally well.** *PLOS One* 2008, **3**:e2802.
- 961 45. Taberlet P, Coissac E, Pompanon F, Gielly L, Miquel C, Valentini A, Vermet T, Corthier G,
 962 Brochmann C, Willerslev E: **Power and limitations of the chloroplast trnL (UAA) intron for plant**
 963 **DNA barcoding.** *Nucleic Acids Research* 2007, **35**:e14.
- 964 46. Chen S, Yao H, Han J, Liu C, Song J, Shi L, Zhu Y, Ma X, Gao T, Pang X: **Validation of the ITS2**
 965 **region as a novel DNA barcode for identifying medicinal plant species.** *PLOS One* 2010, **5**:e8613.
- 966 47. Sang T, Crawford D, Stuessy T: **Chloroplast DNA phylogeny, reticulate evolution, and**
 967 **biogeography of Paeonia (Paeoniaceae).** *American Journal of Botany* 1997, **84**:1120-1136.
- 968 48. Tate JA, Simpson BB: **Paraphyly of Tarasa (Malvaceae) and diverse origins of the polyploid**
 969 **species.** *Systematic Botany* 2003, **28**:723-737.
- 970 49. Manning J, Boatwright JS, Daru BH, Maurin O, Bank Mvd: **A molecular phylogeny and generic**
 971 **classification of Asphodelaceae subfamily Alooideae: a final resolution of the prickly issue of**
 972 **polyphyly in the alooids?** *Systematic Botany* 2014, **39**:55-74.

978
1 979
2 980
3 981
4 982
5 983
6 984
7 985
8 986
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
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

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 , 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.

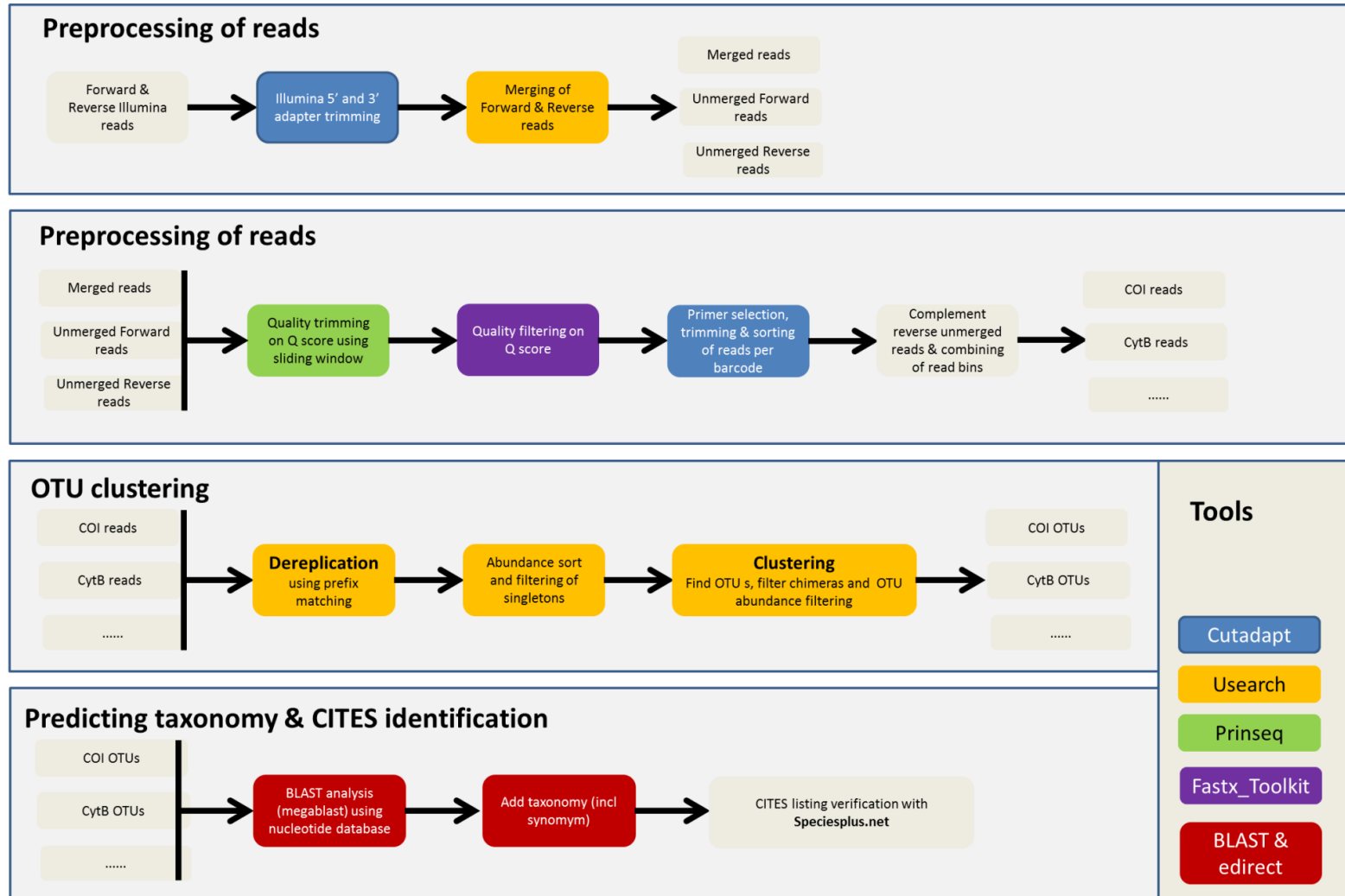
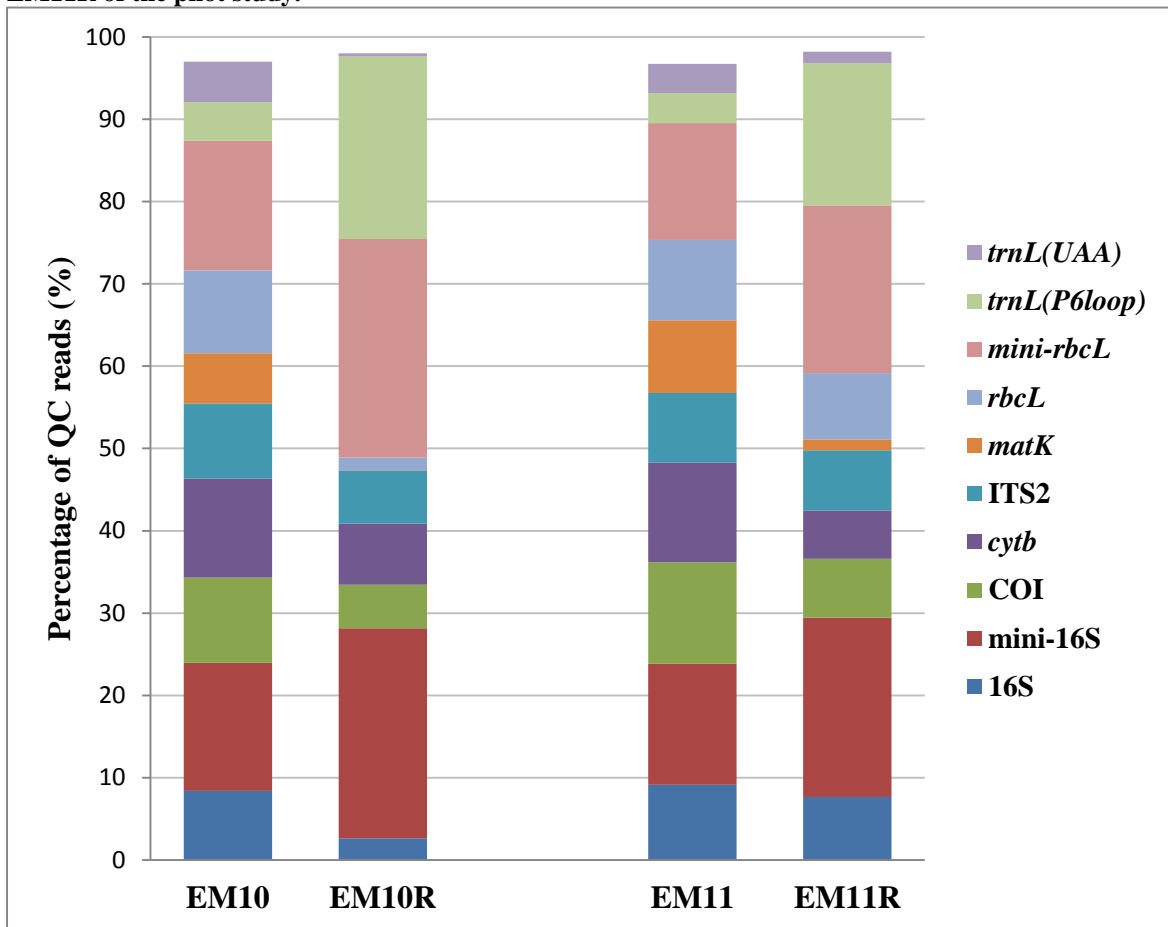
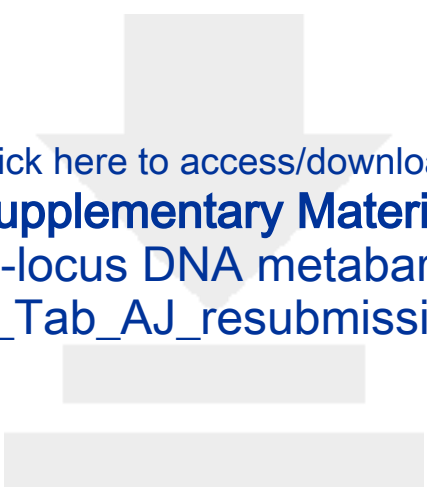



Figure 2: The percentage of QC reads assigned to DNA barcodes for samples EM10, EM10R, EM11 and EM11R of the pilot study.

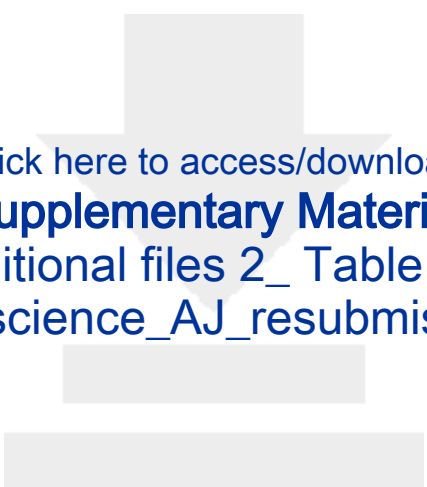




Click here to access/download
Supplementary Material
A multi-locus DNA metabarcoding
method_Tab_AJ_resubmission.docx



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_Gigascience_AJ_resubmission.xlsx](#)

