De Boer et al. DNA metabarcoding of orchid-derived products reveals widespread illegal orchid trade. Proceedings of the Royal Society B. DOI: 10.1098/rspb.2017-1182. Electronic Supplementary Material (ESM) Text S1

DNA extraction and quantification

Samples were processed in small batches (8 – 15; typically 15) with an extraction blank to monitor for potential cross-contamination in a laboratory designed to conduct analyses on historical and sensitive DNA analyses, including separate pre- and post-PCR rooms, with separate equipment for DNA extraction, PCR (including a laminar flow hood), and post-PCR processing. Total DNA was extracted using a modified CTAB protocol with STE [1,2]. In the initial step 1 ml of freshly made STE (0.25 M sucrose, 0.03 M Tris, 0.05 M EDTA) was added to the sample powder. This was vortexed, and then centrifuged at 2,000g for 10 min. The supernatant was discarded and the STE wash step repeated. Subsequently a standard CTAB extraction procedure was followed by addition of 600 µl CTAB solution (2% CTAB, 1% PVP, 1.4 M NaCl, 100 mM Tris–HCl pH 8.0, 20 mM EDTA pH 8.0) and incubation at 60 °C for 40 min with occasional shaking. The final elution volume of the total DNA was 50 µl. Extracted DNA was quantified using a Qubit 2.0 Fluorometer (Thermo Fisher Scientific, Waltham, USA), and measured for integrity and fragment size distribution on a Fragment Analyzer (Advanced Analytical Technologies, Heidelberg, Germany) using a DNF-488 High Sensitivity Genomic DNA Analysis Kit. Prior to amplicon library preparation all extracts including negative controls where tested for amplification of nrITS1 and nrITS2 using PCR. The following primers were used for nrITS1, 17SE_F (5'-ATGGTCCGGTGAAGTGTTC-3') and 5.8I-1_R (5'-GTTGCCGAGAGTCGT-3'), and for nrITS2, 5.8I-2_F (5'-GCCTGGGCGTCACGC-3') and 26SE_R (5'-CCCGGTTCGCTCGCCGTTAC-3') [3]. The 15 µl PCR reaction volume contained 0.08 µl of 5U Taq DNA polymerase, 1.5 µl of 10X PCR buffer, 0.3 µl of 10 mM dNTPs, 1.9 µl of 0.4 g/l bovine serum albumin (BSA), 1 µl of 25 mM $MgCl_{2,}$ 0.66 µl of 10 µM of each primer and 2 µl undiluted template DNA. Cycling conditions consisted of an initial denaturation step of 95 °C for 3 min, followed by 35 cycles of 20 s at 95 °C, 60 s at 45 °C and 45 s at 72 °C, and a final elongation step of 72 °C for 5 min.

DNA metabarcoding

Amplicon libraries of samples and controls were made using fusion PCR based on two nuclear ribosomal target sequences, internal transcribed spacers nrITS1 (~246 bp) and nrITS2 (~233 bp, [4]), using PGM fusion primers. The PGM fusion primers were based on the same primer as used in the amplification pilot, 17SE and 5.8I1, and 5.8I2 and 26SE respectively [3]. The forward primers were labeled with unique 10 bp multiplex identifier (MID) tags and the reverse primers with uniform truncated P1 (trP1) tags (85 bp forward and reserve primer combined). EcoPCR [5] was used to test these primers in-silico on a subset of the NCBI Nucleotide database containing only plant DNA sequences for nrITS with the following parameters: up to 3 bases mismatch allowed for annealing, min length of the product (excluding primers) 50bp, and max length of the product (excluding primers) 500bp. The in-silico PCR confirmed that the main orchid species in *Anacamptis*, *Dactylorhiza*, *Himantoglossum*, *Ophrys* and *Orchis* could be amplified using nrITS2. nrITS1 amplified many of the potential cereal adulterants but no salep orchids.

Thermal cycling was carried out in 25 µl reaction volumes, and each reaction contained 5 µl 5X Q5 reaction buffer, 0.5 µl 10 µM of each primer, 0.5 µl 10 mM dNTPs, 0.25 µl 20U/µl Q5 High-Fidelity DNA Polymerase, 10.25 µl of Milli-Q ultrapure water and 0.5 µl of template DNA. The following thermocycling protocols was used: 30 seconds of initial denaturation at 98 °C, followed by 35 cycles of denaturation at 98 °C for 10 seconds, annealing at 30 seconds, and elongation at 72 °C for 30 seconds, followed by a final elongation step at 72 °C for 2 minutes. The annealing temperature for nrITS1 was 56 °C, and 71 °C for nrITS2.

The size, purity and the molar concentration (in nmol/l) of each amplicon library were measured using a Fragment Analyzer and a DNF-910 dsDNA Reagent Kit (35 bp - 1,500 bp). An equimolar pool (1.5 ng/µl/library) was prepared from the amplicon libraries using the Biomek 4000 Laboratory Automation Workstation (Beckman Coulter, Brea, USA). Negative controls were prepared as undiluted libraries. Agencourt AMPure XP (Beckman Coulter) was used for removal of unincorporated primers and nucleotides using the manufacturer's instructions (Agencourt AMPure XP v. B37419AA). The total concentration of the purified pooled amplicon library stock and three serial dilutions (undiluted, 1/5, 1/10) were analyzed using the Fragment Analyzer (Advanced Analytical Technologies) and DNF-488 High Sensitivity Genomic DNA Analysis Kit in order to identify the optimum concentration range for the template preparation.

An Ion Chef (Life Technologies (LT), Thermo-Fisher Scientific) was used to prepare Pooled Ion AmpliSeq libraries (LT) for emulsion PCR and to load the sequencing chips. The input DNA template concentration was adjusted to the number of Ion Sphere Particles (ISPs) and added to the emulsion PCR master mix. The emulsion PCR was done using the Ion Chef, and template-positive ISPs were enriched and loaded onto two Ion 316 v2 Chips (LT) and sequenced on an Ion Torrent Personal Genome Machine (LT) using the Ion PGM Sequencing 400 kit (LT). Sequencing read data was analyzed and demultiplexed into FASTQ files per chip and sample using Torrent Suite version 5.0.4 (LT) and deposited in DRYAD doi:10.5061/dryad.5q447.

Bioinformatics analysis

FASTQ read files were processed using the HTS-barcode-checker pipeline [6] available as a Galaxy pipeline at the Naturalis Biodiversity Center (http://145.136.240.164:8080/). Using the HTS pipeline, nrITS1 and nrITS2 primer sequences were used to demultiplex the sequencing reads per sample and to filter out reads that did not match any of the primers. PRINSEQ [7] was used to inspect read lengths, Phred base qualities and mean quality scores [8]. Reads were selected with a minimal length of 300 bp in order to filter out short reads below the target amplicon length. Reads were trimmed to a maximum length of 360 bp as base quality scores dropped sharply beyond that point. This exceeds the length of the nrITS1 and nrITS2 amplicons and the 85 bp indexed forward and reserve primer combined, respectively ~331 bp and ~318 bp. Reads with mean Phred quality scores below 25 were filtered to avoid selecting reads with errors or poor base calling. CD-HIT-EST [9] was used to cluster reads into molecular operational taxonomic units (MOTUs) defined by a sequence similarity of >99 % and a minimum number of 2 reads. The DNA barcoding reference library of salep orchids made by Ghorbani et al. [10] showed that 97% similarity clusters more than 3% of taxa in certain genera, whereas 99% similarity reduces this to below 1%. The consensus sequences of non-singleton MOTUs were queried using BLAST [11] against a local copy of the NCBI/GenBank nucleotide database, with a maximum e-value of 0.05, a minimum hit length of 100 bp and sequence identity of >97 %. The number of reads per MOTU as well as the top BLAST hits per MOTU were compiled using custom scripts available online as part of the HTS-barcode-checker pipeline http://145.136.240.164:8080/ [6], and the species assignments per MOTU are deposited in DRYAD doi:10.5061/dryad.5q447.

ESM Text S1 References

1. Doyle JJ, Doyle JL. 1987 A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochem. Bull.* **19**, 11–15.

2. Shepherd LD, McLay TGB. 2011 Two micro-scale protocols for the isolation of DNA from polysaccharide-rich plant tissue. *J. Plant Res.* **124**, 311–314. (doi:10.1007/s10265-010-0379-5)

3. Sun Y, Skinner DZ, Liang GH, Hulbert SH. 1994 Phylogenetic analysis of Sorghum and related taxa using internal transcribed spacers of nuclear ribosomal DNA. *Theor. Appl. Genet.* **89**, 26–32.

4. Han J, Zhu Y, Chen X, Liao B, Yao H, Song J, Chen S, Meng F. 2013 The short ITS2 sequence serves as an efficient taxonomic sequence tag in comparison with the full-length ITS. *BioMed Res. Int.* **2013**.

5. Ficetola GF, Coissac E, Zundel S, Riaz T, Shehzad W, Bessière J, Taberlet P, Pompanon F. 2010 An In silico approach for the evaluation of DNA barcodes. *BMC Genomics* **11**, 434.

6. Lammers Y, Peelen T, Vos RA, Gravendeel B. 2014 The HTS barcode checker pipeline, a tool for automated detection of illegally traded species from high-throughput sequencing data. *BMC Bioinformatics* **15**, 44. (doi:10.1186/1471-2105-15-44)

7. Schmieder R, Edwards R. 2011 Quality control and preprocessing of metagenomic datasets. *Bioinformatics* **27**, 863–864.

8. Ewing B, Green P. 1998 Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.* **8**, 186–194.

9. Li W, Godzik A. 2006 Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658–1659.

10. Ghorbani A, Gravendeel B, Selliah S, Zarre S, de Boer HJ. 2016 DNA barcoding of tuberous Orchidoideae: A resource for identification of orchids used in Salep. *Mol Ecol Resour* (doi:10.1111/1755-0998.12615)

11. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990 Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410.