scientific data

OPEN

DATA DESCRIPTOR

Semi-automated sequence curation for reliable reference datasets in ITS2 vascular plant DNA (meta-) barcoding

Andreia Quaresma[1,2,3,4,5], Markus J. Ankenbrand[6], Carlos Ariel Yadró Garcia[1,2], José Rufino[2,7], Mónica Honrado[1,2], Joana Amaral[1,2], Robert Brodschneider[8], Valters Brusbardis[9], Kristina Gratzer[8], Fani Hatjina[10], Ole Kilpinen[11], Marco Pietropaoli[12], Ivo Roessink[13], Jozef van der Steen[14], Flemming Vejsnæs[11], M. Alice Pinto[1,2,16] & Alexander Keller[15,16] ✉

One of the most critical steps for accurate taxonomic identification in DNA (meta)-barcoding is to have an accurate DNA reference sequence dataset for the marker of choice. Therefore, developing such a dataset has been a long-term ambition, especially in the *Viridiplantae* kingdom. Typically, reference datasets are constructed with sequences downloaded from general public databases, which can carry taxonomic and other relevant errors. Herein, we constructed a curated (i) global dataset, (ii) European crop dataset, and (iii) 27 datasets for the EU countries for the ITS2 barcoding marker of vascular plants. To that end, we first developed a pipeline script that entails (i) an automated curation stage comprising five filters, (ii) manual taxonomic correction for misclassified taxa, and (iii) manual addition of newly sequenced species. The pipeline allows easy updating of the curated datasets. With this approach, 13% of the sequences, corresponding to 7% of species originally imported from GenBank, were discarded. Further, 259 sequences were manually added to the curated global dataset, which now comprises 307,977 sequences of 111,382 plant species.

[1]Centro de Investigação de Montanha (CIMO), Instituto Politécnico de Bragança, Campus de Santa Apolónia, 5300-253, Bragança, Portugal. [2]Laboratório Associado para a Sustentabilidade e Tecnologia em Regiões de Montanha (SusTEC), Instituto Politécnico de Bragança, Campus de Santa Apolónia, 5300-253, Bragança, Portugal. [3]Departamento de Biologia, Faculdade de Ciências da Universidade do Porto, Rua do Campo Alegre, S/N, Edifício FC4, 4169-007, Porto, Portugal. [4]CIBIO, Centro de Investigação em Biodiversidade e Recursos Genéticos, InBIO Laboratório Associado, Campus de Vairão, Universidade do Porto, 4485-661, Vairão, Vila do Conde, Portugal. [5]BIOPOLIS Program in Genomics, Biodiversity and Land Planning, CIBIO, Campus de Vairão, 4485-661, Vairão, Vila do Conde, Portugal. [6]Center for Computational and Theoretical Biology, Faculty of Biology, Julius-Maximilians-Universität Würzburg, Klara-Oppenheimer-Weg 32, 97074, Würzburg, Germany. [7]Research Centre in Digitalization and Intelligent Robotics (CeDRI), Instituto Politécnico de Bragança, Bragança, Portugal. [8]Institute of Biology, University of Graz, Universitätsplatz 2, 8010, Graz, Austria. [9]Latvian Beekeepers' Association (LBA), Rigas iela 22, LV-3004, Jelgava, Latvia. [10]Ellinikos Georgikos Organismos DIMITRA (ELGO- DIMITRA), Kourtidou 56-58, GR-11145, Athina, Greece. [11]Danish Beekeepers Association (DBF), Fulbyvej 15, DK-4180, Sorø, Denmark. [12]Istituto Zooprofilattico Sperimentale del Lazio e della Toscana "M. Aleandri" (IZSLT), Via Appia Nuova 1411, IT-00178, Roma, Italy. [13]Wageningen Environmental Research, WageningenUniversity&Research, Droevendaalsesteeg 3, 6700 AA, Wageningen, Netherlands. [14]Alveus AB Consultancy, Kerkstraat 96, 5061, Oisterwijk, EL, Netherlands. [15]Cellular and Organismic Interactions, Biocenter, Faculty of Biology, Ludwig-Maximilians-Universität München, Großhaderner Str. 2-4, 82152, Planegg-Martinsried, Germany. [16]These authors contributed equally: M. Alice Pinto, Alexander Keller. ✉e-mail: keller@biologie.uni-muenchen.de

Barcoding marker	Number of entries in 2015*	Number of entries in 2023**	Increase (%)
rbcLa	155,634	409,911	163
trnH-psbA	86,828	176,688	103
matK	127,990	270,486	111
ITS2	243,155	460,121	89

Table 1. Number of sequences available in GenBank for each of the Viridiplantae DNA barcoding marker in 2015 and 2023, and corresponding increase rate during this period. *Retrieved from Bell *et al.*[11]. Accessed on 4 November 2015 using the following search strings: "ITS2 OR internal transcribed spacer 2[All Fields] AND plants [filter]"; "rbcL OR rbc-L or Rubisco [All Fields] AND plants[filter]"; "trnH OR trn-H OR trnH-psbA OR psbA-trnH [All Fields] AND plants[filter]"; "matK OR mat-K OR maturase K [All Fields] AND plants[filter]". **Accessed on 22 February 2023 using the search string of Bell *et al.*[9].

Background & Summary

DNA barcoding, a concept put forward by Hebert *et al.*[1] in 2003, was developed to facilitate species identification using molecular methods. DNA barcoding standardizes the taxonomic identification of organisms based on well-established short genomic regions that have high interspecific and low intraspecific variability. By definition, a DNA barcoding marker must be universal, reliable, and show good discriminatory power at the species level[2]. For animals and fungi, the mitochondrial cytochrome c oxidase I gene (COI)[1] and the internal transcribed spacer (ITS) region[3], respectively, have been defined and accepted by the scientific community as the genomic regions that fulfil these criteria. However, in the *Viridiplantae* kingdom, there is no single barcoding marker that satisfies all of those criteria, and several markers in the mitochondrial, chloroplastidial, and nuclear genomes have been under dispute[4–6]. Finally, four DNA barcoding markers have been agreed upon for taxonomic identification of plants, including the chloroplastidial regions rbcL, matK, and trnH-psbA, as well as the nuclear internal transcribed spacer (ITS) region of the ribosome, particularly the ITS2 region[2,7,8].

The emergence of high-throughput sequencing (HTS) techniques is tightly linked to the recent burst of DNA metabarcoding studies[9,10], which have used one or more of the four markers for taxonomic identification. DNA metabarcoding is a powerful approach for resolving mixed-species samples or environmental DNA (eDNA)[11] at large spatial scales, with multiple applications in the fields of ecology, taxonomy, evolution, and conservation[11,12] for a wide array of organisms. In plants, DNA metabarcoding has been applied in the authentication of herbal teas[13–15], determining herbivore diets[16–19], unravelling plant-pollinator interactions[20–23], identifying botanical origin of honey[24–26], monitoring allergy-related airborne pollen sources[27,28], assessing biodiversity[29–32], or even in forensic analysis[33]. These studies have either employed single DNA marker or their combinations, with most relying on rbcL and/or ITS2[34–36]. ITS2 has been increasingly popular due to its better taxonomic discriminatory capabilities[2] and the higher number of sequences available in GenBank[37] (Table 1) as compared with the other three plant barcoding markers.

Botanical identification of mixed-species samples by DNA metabarcoding entails their laboratorial processing to obtain the sequence reads with HTS technologies. The millions of reads generated by the HTS are then classified against sequences of known taxonomic origin, which are typically *a priori* compiled in reference datasets, either constructed using own sequences or such retrieved from GenBank or other public databases. The quality of identification depends on the quality and completeness of the reference dataset built for the target barcoding marker, which in turn is determined by the breadth and size of the dataset (number of taxa and number of sequences per taxon) as well as by the taxonomic accuracy of the compiled sequences[20,38,39]. Many plant studies have relied on sequence data directly retrieved from GenBank for identifying unknown samples[24,40–43]. The problem with this approach is that sequences deposited in GenBank are not rigorously checked for taxonomic mistakes and other inconsistencies that might affect barcoding purposes. Erroneous records are common and can, for example, be due to fungi inhabiting the surface or tissue of plants that are sequenced instead of the targeted plant, or to plants that were morphologically misidentified[38]. This results in inaccurate classifications using direct hit methods (e.g., VSEARCH[44], USEARCH[45], BLAST[46]), and also in poor models for hierarchical classifications (e.g., RDPclassifier[47], SINTAX[48]).

Construction of high-quality reference datasets for plants has been sought over the years, and several attempts have been made, specifically for ITS2 and rbcL. The first ITS2 reference database was released in 2006[49] for different kingdoms. This database underwent several updates until 2015[50]. In the same year, Sickel *et al.*[51] built the first *Viridiplantae* specific ITS2 dataset from the original multi-kingdom ITS2 database, which has been used in several plant metabarcoding studies[20,34,52,53]. However, due to the ever-increasing number of sequences deposited in GenBank, this dataset soon became outdated (Table 1). In 2017, Bell *et al.*[52] developed an rbcL dataset, which was combined with the existing ITS2 *Viridiplantae*[51], for species-level identification in angiosperms. This rbcL dataset was last updated in 2021, at the same time that a new ITS2 dataset for *Magnoliopsida* was developed by the same group[54]. In 2019, Curd *et al.*[55] developed the ANACAPA toolkit, which comprises a module to generate custom reference datasets for any marker. In 2020, Banchi *et al.*[38] published an ITS dataset, named PLANiTS, that includes datasets for ITS, ITS1 and ITS2. In addition, these authors developed a script that performs a species identity check on the sequences downloaded from GenBank, although it is a QIIME2 based script. Also in 2020, Richardson *et al.*[56], developed the toolkit MetaCurator, which generates reference datasets dedicated to taxonomically informative genetic markers, while Keller *et al.*[39] developed BCdatabaser, a tool that allows generating generic datasets of any marker by linking sequences and taxonomic information retrieved from GenBank. In 2022, Dubois *et al.*[12] developed a workflow that allows the building of plant reference datasets dedicated to ITS2 and rbcL. However, this workflow can only be used on the QIIME2 platform.

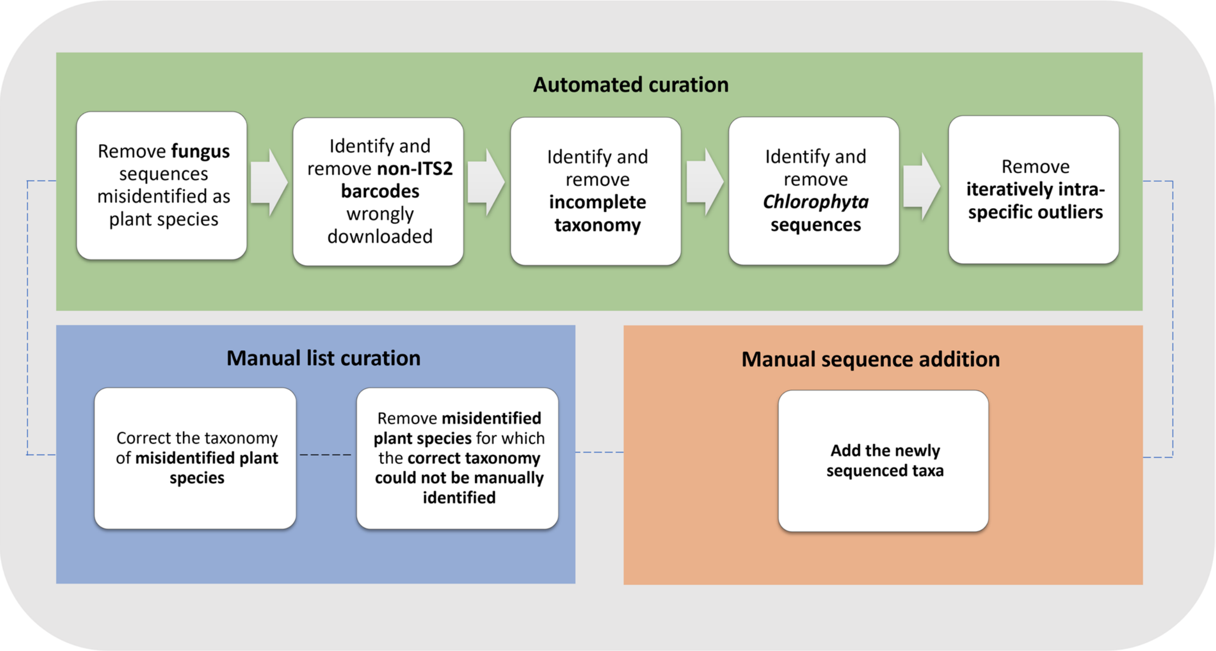
Automated curation
Remove **fungus** sequences misidentified as plant species
Identify and remove **non-ITS2 barcodes** wrongly downloaded
Identify and remove **incomplete taxonomy**
Identify and remove ***Chlorophyta* sequences**
Remove **iteratively intra-specific outliers**
Manual list curation
Correct the taxonomy of **misidentified plant species**
Remove misidentified plant species for which the **correct taxonomy could not be manually identified**
Manual sequence addition
Add the newly sequenced taxa

Fig. 1 Schematic representation of the curation pipeline. The components 'Automated curation', 'Manual list curation', and 'Manual sequence addition' can be used singly or in conjunction.

The first developed datasets were static and, therefore, easily outdated due to the rapid flow of new sequences being deposited in GenBank (Table 1). Moreover, most of the available datasets are global-scale, which may lead to taxa misidentifications because of sequence conflicts originating from misidentified GenBank sequences or even from polyphyletic species. Accordingly, it might be helpful to have a dataset tailored for the geographical area under analysis, as a way of reducing the identification error by including only the extant flora, therefore minimizing the detection of unlikely taxa[57]. Complementary to this, it is also important to have user-friendly tools that automatically perform the generation and curation of reliable and updatable reference datasets. Currently, most of the available tools require some level of user bioinformatics expertise or lack a good curation method for handling the problem of misidentified GenBank sequences. For instance, BCdatabaser[39] is a user-friendly tool as it entails a single command to produce a taxonomy-linked *fasta* file, which can be used by several taxonomic classifiers. However, it lacks a curation method, and it includes the download of non-target sequences incorrectly annotated in GenBank[12] (e.g., rbcl sequences that are labelled as ITS2).

In this context, the goal of this study was to provide curated datasets for ITS2 (meta)-barcoding, and a reproducible, public, and pipeline-based workflow that is compatible with other custom datasets. The script was designed to be applied after using BCdatabaser[39] or similar workflows that generate taxonomically linked *fasta* files. The workflow consists of three main stages: (i) automated curation of the downloaded sequences that accounts for five major problems detected in sequences deposited in GenBank (fungal sequences identified as vascular plants, *Chlorophyta* sequences, non-target sequences, incomplete taxonomies, and erroneous taxonomy annotation); (ii) a manual taxonomic correction option for misidentified taxa; and (iii) the addition of custom sequenced species to conform with the common syntax of the database. Using this workflow, we generated an ITS2 reference dataset that comprises worldwide vascular plant taxa, as well as individual subsets of this database for each of the 27 countries of the European Union and a reference dataset for European crops.

Methods

Curation pipeline-based workflow. The pipeline-based workflow comprises three independent stages for generating more accurate reference datasets: (i) automated curation, (ii) manual list curation, and (iii) manual sequence addition (Fig. 1). These can be performed singly or in conjunction, depending on the user's needs. The pipeline script is publicly available at GitHub (https://github.com/chiras/database-curation) and has as dependencies the also publicly available software tools R[58], SeqFilter v2.1.10[59] (https://github.com/BioInf-Wuerzburg/SeqFilter), and VSEARCH v2.18.0[44] (https://github.com/torognes/vsearch). It is designed to start after the point of pulling reference sequences from GenBank with BCdatabaser (or equivalent tools) or from other public sources that follow the same syntax needed for a variety of classifiers (https://molbiodiv.github.io/bcdatabaser/output.html). The pipeline was executed successfully on the bash command line of Ubuntu 20.04.6 and Mac OSX 12.3.

Automated curation. The automated curation is the most important stages of the pipeline-based workflow. Five major cleaning steps are implemented during curation (Fig. 1):

i. The first filter identifies fungal sequences and removes them. These are identified by using a hierarchical classification with the *sintax*[48] command from VSEARCH against the RDP curated fungal ITS dataset[60], with a cut-off of 0.90;

ii. The second filter performs the removal of non-ITS2 (non-target) sequences. For this, we manually created a preliminary ITS2 reference dataset of selected trustworthy sequences representing all vascular plant families from the ITS2 database[50]. In the automated curation, the command *usearch_global* by VSEARCH is used to identify only vascular plant sequences with an identity threshold of 70%;

iii. The third filter checks for incomplete taxonomy entries in the metadata and removes such entries as they are not suitable for barcoding purposes and might interfere with finding better resolved references;

iv. The fourth filter removes all the sequences that are classified as *Chlorophyta* as our intention was to create a reliable vascular plant dataset. Wrong annotations of *Chlorophyta* sequences can also interfere with vascular plant identification;

v. The fifth filter applies a deterministic assessment of intraspecific variability for the respective dataset on-the-fly. However, this filter is only applied to species that are represented by more than four sequences. The dataset is hereby split into subsets for each plant species, and, for each separate species datasets pairwise all-against-all global alignments are performed with *allpairs_global* from VSEARCH. An iterative R script increases a drop-out threshold for each species in steps of 50%, 75%, 80%, 85%, 90%, 92.5%, 95%, and 97%, removing sequences that have a lower median identity to all other sequences of the species than the threshold, but only while a threshold is given that removes less than 50% of the remaining sequences per species. This 50% threshold is a balanced trade-off between removing taxa with wrong GenBank taxonomic assignments and retaining sequences that are still within expected intraspecific variability (see ´Assessment of intraspecific variability´ section for further details).

Manual list curation. The manual list curation is intended to serve as a community-driven approach. Scientists that spot erroneous GenBank entries that are not identified by the automated curation are invited to add a simple tabular text file to our GitHub repository. Based on these text files, researchers curating a dataset can choose to use or discard manual curations from different contributors. The text file format is kept as simple as possible, and examples are given in the code repository:

```
NCBIAccessionNumber;WrongScientificName;CorrectedScientificName;CuratorName
```

If the file specifies the *CorrectedScientificName*, the script will proceed to correct the taxonomy in the reference dataset. The field can be left empty as well, indicating that the curator is sure that this is a wrong taxonomic metadata and yet unsure about the correct identification, which will result in the sequences being removed from the dataset.

Manual sequence addition. The manual addition allows users to add own generated sequences to the reference dataset, and automating the gathering of taxonomic metadata and formatting. This is a tedious step, especially when many sequences are added. The requirement for this step is the provision of common *fasta* files with the species name as the header. Examples are provided in the GitHub repository.

Global dataset subdivision. The subdivision of the global dataset allows the user to reduce the number of species from the global reference dataset to a local reference dataset that contains a geographically delimited number of species. For this step, it is required to provide a list of the intended local flora in a csv file format.

Application of the pipeline for curation of ITS2 datasets. A *Viridiplantae* ITS2 reference dataset, hereafter called "global dataset" was created on 17 January of 2023 using the following command of BCdatabaser[39]:

```
docker run -u $UID:$GID -v $PWD:/data \
        --rm iimog/bcdatabaser \
        --outdir its2.global.$today \
        --marker-search-string "(ITS2 OR internal transcribed spacer 2)" \
        --taxonomic-range Viridiplantae \
        --sequences-per-taxon=25 \
        --sequence-length-filter 100:2000 \
        --names-dmp-path /NCBI-Taxonomy/names.dmp \
        --warn-failed-tax-names
```

This dataset comprises a maximum of 25 sequences per species of the *Viridiplantae* kingdom, within a length range of 100–2,000 bp. Across the study, we found that crop species represent a special case for barcoding purposes because they show a high intraspecific variability of the ITS2 region, often due to hybridizations or other genomic interventions (e.g.: *Brassica* and *Malus*). Therefore, we considered that there was an additional need for developing a reference dataset only for European crops, which is further referred to as the "crop dataset". This dataset was generated in the same way, but instead of 25 sequences, a maximum of 100 ITS2 sequences per species was downloaded from GenBank to account for a higher representation of intraspecific variability.

Enrichment with new sequences. In addition, 536 leaf samples representing 322 species, selected from expert knowledge as important pollen sources for the honey bee (*Apis mellifera*), collected from nine European

Filters/steps	Global dataset Sequences removed	 Sequences/taxa retained	Crop dataset Sequences removed	 Sequences/ taxa retained
Start	—	354,690/119,830	—	4,206/81
Automated curation				
Fungal sequences	127	354,563	3	4,203
Non-ITS2 sequences	29,341	325,222	249	3,954
Incomplete taxonomy	6	325,216	0	3,954
Chlorophyta sequences	781	324,435	0	3,954
High intraspecific variability	16,711	307,724	611	3,343
Total	**46,966/8,453**	**307,724/111,377**	**863/0**	**3,343/81**
Manual list curation				
Misidentified sequences	6	307,718	0	3,343
Taxonomy corrected	5*	307,718	0	3,343
Total	**6/3**	**307,718/111,374**	**0/0**	**3,343/81**
Manual sequence addition				
New sequences	—	259/182		
Gran-total	**—**	**307,977/111,382**	**0/0**	**3,343/81**

Table 2. Number of sequences/corresponding taxa that were removed/retained/added by the curation pipeline (automated curation, manual list curation, and manual sequence addition) from/in/to the ITS2 global and crop datasets. *Number of detected sequences with incorrect taxonomic classification; These were not removed but instead corrected and retained in the datasets.

countries (Austria, Denmark, France, Greece, Italy, Latvia, The Netherlands, Norway, and Portugal) were further sequenced for the ITS2 region, aiming for manual addition into the dataset (Table 2). These species were missing or underrepresented in the initial global dataset. The leaves were cut into small pieces and transferred to a 2.0 ml screwcap tube with two 3 mm zirconia beads. After being grounded in a Precellys 24 tissue homogeniser (Bertin Instruments), the DNA was extracted with the Macherey-Nagel NucleoSpin Plant II Kit, according to the manufacturer's instructions. DNA extracts were amplified targeting the ITS2 region using the primers ITS-S2F[61] and ITS-S4R[62]. PCR was carried out in a 25 µL total volume using 12.5 µL of Q5 High-Fidelity 2X Master Mix (New England Biolabs), 1.25 µL of each primer (10 µM), and 1 µL of DNA (10 ng/µL). Reactions were performed in a T100 Thermal Cycler (BioRad^{TM}) using the temperature profile consisting of an initial denaturation of 98 °C for 3 min, followed by 35 cycles of 98 °C for 10 s, 52 °C for 30 s, and 72 °C for 40 s, and a final extension of 72 °C for 2 min. The amplicons were Sanger sequenced at STABVIDA Inc. (Portugal) and then analysed using Mega v10.1.7[63].

From the 536 samples submitted to DNA sequencing, 259 clean and sufficiently long, high-quality sequences were generated, representing 182 species (Table 2). The new sequences were collected in a *fasta* format file and then added to the global dataset using the manual sequence addition script, as described above. These sequences are also available in the GitHub repository.

Country-level datasets. After curation, the global ITS2 dataset was subdivided into two local ITS2 dataset for each of the 27 EU countries, according to the local flora retrieved from two online flora databases: Euro + Med PlantBase (https://www.emplantbase.org/home.html) and GBIF (https://www.gbif.org/). These databases complement each other, enabling a more comprehensive representation of the local flora across the 27 EU countries. A more extensive list of plant taxa was retrieved from GBIF than from the Euro + Med PlantBase for the 27 EU countries. Still, there were taxa in the Euro + Med PlantBase list that were missing in the GBIF list.

Data Records

All final ITS2 datasets are publicly available as fasta files on Zenodo:

(i) global dataset: https://doi.org/10.5281/zenodo.7968519[64];
(ii) crop dataset: https://doi.org/10.5281/zenodo.7969940[65], and
(iii) country-level datasets for the 27 EU countries: https://doi.org/10.5281/zenodo.7970046[66].

New ITS2 sequences were publicly deposited in GenBank (https://www.ncbi.nlm.nih.gov/nuccore) under the BioProject PRJNA1033169.

The curation scripts are publicly available as bash and R code at https://github.com/chiras/database-curation.

A web interface has been developed that allows search for accessions and taxonomic names to assess which sequences were kept or removed during the curation (global dataset). The web interface also allows selection of sequences and refer to the corresponding NCBI records for further investigative purposes. The web interface is available at https://its2curation.molecular.eco.

Country	Sequences A/B	Species with sequences A/B	Species in the flora A/B	Species coverage (%) A/B
Austria	25,209/40,297	2,747/5,141	3,572/11,316	77/45
Belgium	18,083/57,279	1,810/9,070	2,182/17,359	83/52
Bulgaria	23,460/26,591	2,812/3,137	3,839/5,599	73/56
Croatia	20,640/28,011	2,400/3,385	3,053/5,852	79/58
Cyprus	10,777/13,524	1,296/1,624	1,710/2,600	76/62
Czechia	18,620/32,640	1,929/3,707	2,411/6,650	80/56
Denmark	16,804/29,735	1,583/3,218	1,846/5,307	86/61
Estonia	14,709/29,816	1,352/3,363	1,534/5,586	88/60
Finland	16,214/27,716	1,527/2,958	1,747/5,351	87/55
France	32,207/59,623	4,134/9,426	5,799/30,227	71/31
Germany	26,709/57,740	2,979/8,516	4,107/20,101	73/42
Greece	24,909/34,231	3,550/5,033	5,382/10,966	66/46
Hungary	20,670/30,557	2,089/3,584	2,535/6,925	82/52
Ireland	14,341/18,944	1,328/1,884	1,582/2,682	84/70
Italy	32,955/47,441	4,310/6,808	5,948/16,427	72/41
Latvia	13,405/21,771	1,210/2,178	1,367/3,148	89/69
Lithuania	14,504/16,702	1,316/1,486	1,497/1,736	88/86
Luxembourg	1,627/22,033	137/2,147	178/3,085	77/70
Malta	9,874/6,672	1,108/677	1,371/929	81/73
The Netherlands	16,693/43,884	1,530/5,770	1,881/11,115	81/52
Poland	22,099/33,094	2,256/3,711	2,785/6,740	81/55
Portugal	19,593/37,654	2,372/5,103	3,031/10,347	78/49
Romania	25,349/28,296	2,820/3,221	3,673/5,858	77/55
Slovakia	18,344/26,330	1,925/2,795	2,448/4,616	79/61
Slovenia	18,329/25,063	1,968/2,727	2,434/4,129	81/66
Spain	32,265/58,332	4,384/9,628	6,380/27,128	69/35
Sweden	20,468/48,305	2,048/6,736	2,446/14,648	84/46

Table 3. Sizes of the country-level ITS2 datasets in relation to the vascular plant species inventories extracted from (A) Euro + Med PlantBase (https://www.emplantbase.org/home.html) and (B) GBIF platforms (https://www.gbif.org/). Number of ITS2 sequences, number of species with ITS2 sequences, number of species extracted from Euro + Med PlantBase and GBIF, and proportion of species with ITS2 sequences in the dataset.

Global dataset. The global dataset downloaded from GenBank originally held a total of 354,690 sequences, representing 119,830 unique species (Table 2). However, many sequences were identified as problematic and were thus removed (see Table S1 for the full list of the removed accession numbers) after the automated implementation of the five sequential curation filters, as follows: (i) 127 fungal sequences; (ii) 29,341 non-ITS2 sequences; (iii) six sequences with incomplete taxonomies; (iv) 781 *Chlorophyta* sequences; and (v) 16,711 sequences with unexpectedly high intraspecific variability for the respective species. After this automated curation, 307,724 sequences (13% loss) were retained in the global dataset representing 111,377 species (7% loss). The manual list curation detected 11 misidentified sequences in the global dataset, of which six were removed due to incorrect taxonomic classification, which was not possible to edit, and five were replaced by their correct taxonomic classification. After this additional step, a total of 307,718 sequences, representing 111,374 species, were retained in the global dataset. With the addition of our own ITS2 sequences, the final global dataset contains 307,977 sequences, representing 534 families, 11,034 genera, and 111,382 species of vascular plants.

Crop dataset. A list of European crop species, containing for each entry an accurate taxonomic classification string, was carefully assembled and then used to retrieve the matching sequences from GenBank. A total of 4,206 sequences, representing 81 taxa, were downloaded from GenBank. The automated curation workflow identified and removed (Table S1) from this dataset the following number of sequences: (i) three fungal sequences; (ii) 249 non-ITS2 sequences; and (iii) 611 sequences with high intraspecific variability for the respective species (Table 2). As expected from the nature of the assembled list, no 'Incomplete taxonomy' or '*Chlorophyta*' problems were detected. Furthermore, no sequences were removed or added by the 'Manual list curation' and 'Manual sequence addition' components of the pipeline. Accordingly, the final crop dataset comprises 3,343 sequences (21% loss), representing 25 families, 50 genera, and 81 species (0% loss).

Country-level datasets. Table 3 compiles the sizes of the two ITS2 datasets generated for each of the 27 EU countries, taking into account the local flora extracted from Euro + Med PlantBase and GBIF. The 27 ITS2 datasets generated using the Euro + Med PlantBase lists cover between 66% and 89% of the vascular plant species listed for each country (Fig. 2). The ITS2 datasets of the Mediterranean countries show the lowest coverage of the local flora, with Greece having 66%, Spain 69%, France 71%, and Italy 72%. In contrast, the ITS2 datasets obtained for the

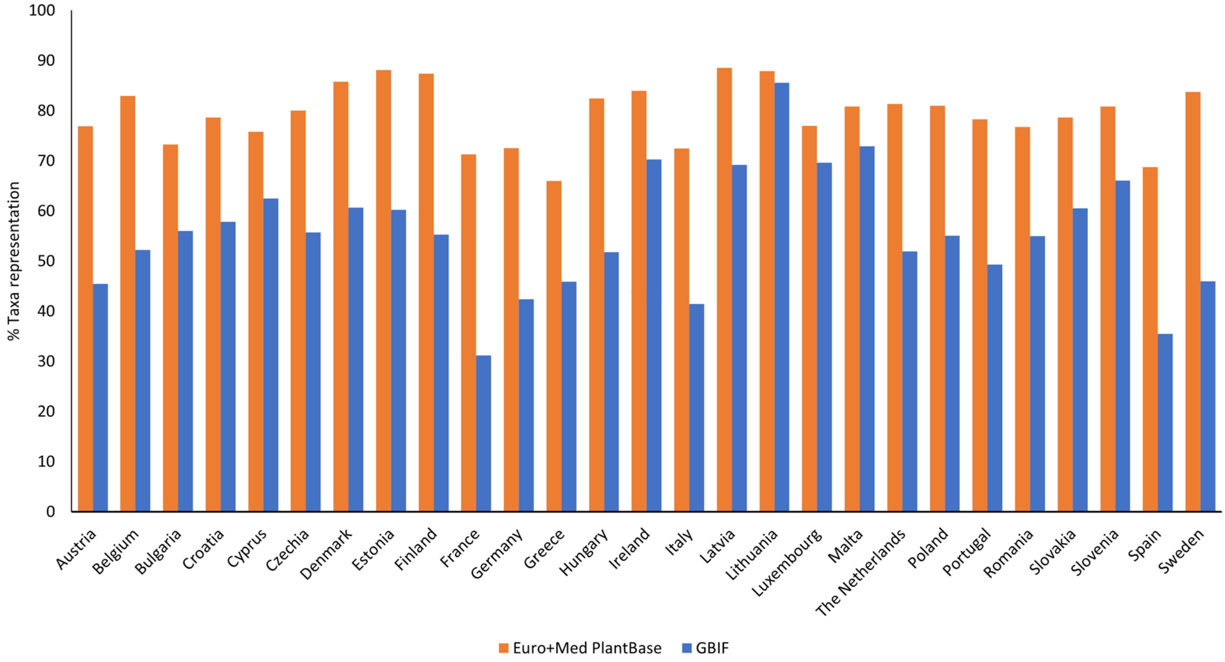

Fig. 2 Taxa representation of the two reference ITS2 datasets generated for each of the 27 EU countries, using the flora information extracted from the Euro + Med PlantBase (https://www.emplantbase.org/home.html) and GBIF platforms (https://www.gbif.org/).

Baltic countries contain sequences representing a high proportion of their plant diversity, with Latvia (89%) at the top of the ranking, followed by Lithuania and Estonia (88%), and Finland (87%). The findings for Mediterranean countries were expected due to their higher species richness, thereby requiring a higher sequencing effort to achieve the levels of the Baltic countries. Apart from Malta, the lists extracted from GBIF are species-richer than those extracted from Euro + Med PlantBase, explaining the lower coverage of the corresponding ITS2 datasets. Hence, the coverage of the ITS2 datasets generated using the GBIF lists is lower than that generated using the Euro + Med PlantBase lists, varying between 31% for France and 86% for Lithuania (Fig. 2).

Technical Validation

Fungal sequences identified as plants in GenBank. A total of 127 fungal sequences were detected among the sequences identified as plants in GenBank. Of these, 55 (43%) belonged to the phylum *Ascomycota*, and the most common genera were *Erysiphe* (15%), *Aspergillus* (14%), *Davidiella* (11%), *Gibberella*, and *Mycosphaerella (8%)*, and *Eurotium* (6%). These fungi are either pathogens or endophytes commonly detected in plant tissues (e.g., *Erysiphe* causes powdery mildew, and *Mycosphaerella* causes leaf blight). Fungal PCR-amplifications from infected plant tissues are well documented for ITS2 primers designed for plants[67], explaining the misidentified sequences deposited in GenBank. One such example comes from the single ITS2 sequence available in GenBank for *Rumex stenophyllus* (accession number MG235257). During the automated curation, this sequence was identified as belonging to the genus *Alternaria*, leading to its removal from the global dataset.

Plant sequences assigned an incorrect taxonomic classification. The automated curation allowed the identification and removal of sequences that were deposited in GenBank with incorrect taxonomic classification. For instance, the sequences with accession numbers KF454376 and KF454377, originally identified in GenBank as *Typha angustifolia* (*Typhaceae*), turned out to belong to the genus *Taraxacum* (*Asteraceae*) after manual verification. With the intraspecific analysis implemented by the fifth filter of the automated curation, these two sequences were automatically removed from the global dataset.

Assessment of intraspecific variability. The accuracy of the taxonomic classification depends on the power of the chosen marker in discriminating between interspecific and intraspecific variation, i.e., the overlap of the genetic variation between species should be small or ideally non-existent. Hybridization is a common natural or human-mediated phenomenon in many wild plant species as well as in many crops, such as *Brassica napus* and *Brassica rapa*, or *Malus domestica* and *Pyrus communis*. This erodes species delimitations and increases intraspecific variability, making automated curation a more challenging endeavour.

The last step of the automated curation (the fifth filter) applies a deterministic assessment of intraspecific variability for the respective species. In the initial configuration of the pipeline, the sequences that had a median identity lower than 97% in pairwise all-against-all global alignments were removed from the dataset in a single iteration. This revealed itself to be very stringent for taxa suffering from high intraspecific variability, leading to the removal of all the sequences from the curated dataset. Hence, this direct approach (approach A) was replaced by the iterative increment of the drop-out threshold (approach B), as explained in the section 'Automated curation'. While an

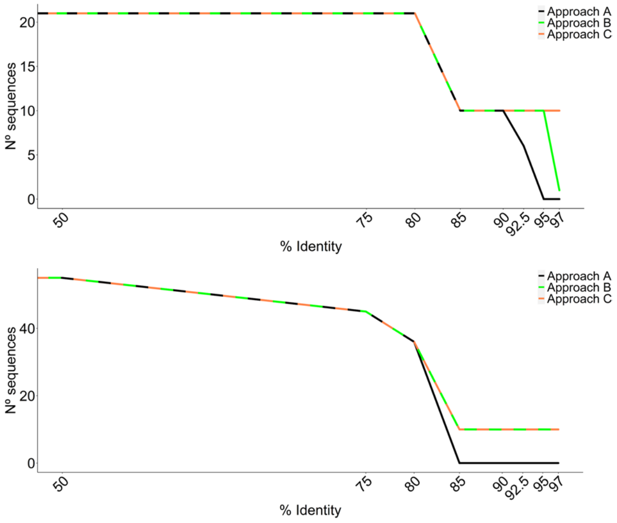

Fig. 3 Number of sequences retained in the ITS2 dataset for *Malus pumila* (top chart) *and Pyrus communis* (bottom chart) by the automated curation workflow. Approach A: sequences with a median identity <97% in pairwise all-against-all global alignments are removed in a single iteration; Approach B: sequences are removed iteratively using an incremental drop-out identity threshold of 50%, 75%, 80%, 85%, 90%, 92.5%, 95%, and 97%; Approach C: sequences are removed using the incremental threshold of ´Approach B´ while ensuring that 50% of the initial sequences are retained in the dataset.

improvement in the pipeline's performance was noted, there was still a low number of retained sequences in the curated dataset (e.g., *Malus domestica* was represented by a single sequence). Lastly, in the final configuration of the automated curation (see the 'Automated curation' section), the introduced threshold that retains 50% of the initial sequences (approach C) seems to represent a good trade-off between removing taxa with wrong GenBank taxonomic assignments and retaining the sequences that are still within expected intraspecific variability.

The outcomes of these three approaches are illustrated in Fig. 3 for *Malus pumila* and *Pyrus communis*. No sequences or a single sequence were retained in the curated dataset for *Malus pumila* with approaches A and B, respectively. In contrast, 10 of the initial 20 sequences were retained in the curated dataset at 85% identity when approach C was applied. In the case of *Pyrus communis*, approaches B and C performed equally well, retaining 10 of the initial 55 sequences, whereas all the sequences were removed from the dataset when applying approach A.

Comparison with other datasets. The global ITS2 dataset generated in this study contains sequences from 111,377 species, representing an increase of over 62% when compared to the datasets of Sickel *et al.*[51] (72,325 species) and Dubois *et al.*[12] (~70,000 species). The implementation of the automated curation script developed herein is able to resolve troublesome sequences downloaded from GenBank while still retaining a good representation of worldwide species in the curated dataset. Moreover, the manual list curation step prevents reliable sequences from being removed at the same time that the manual sequence addition step facilitates dataset enrichment.

Code availability

All code used in this study is freely available in https://github.com/chiras/database-curation. The developed global and country-level datasets are also provided in the same repository as well as in Zenodo. A web interface with a list of sequences that were kept or removed during curation is available at https://its2curation.molecular.eco.

Received: 10 July 2023; Accepted: 12 January 2024;
Published online: 25 January 2024

References

1. Hebert, P. D. N., Cywinska, A., Ball, S. L. & deWaard, J. R. Biological identifications through DNA barcodes. *Proceedings of the Royal Society of London. Series B: Biological Sciences* **270**, 313–321, https://doi.org/10.1098/rspb.2002.2218 (2003).
2. Li, D.-Z. *et al.* Comparative analysis of a large dataset indicates that internal transcribed spacer (ITS) should be incorporated into the core barcode for seed plants. *Proc. Natl. Acad. Sci. (PNAS)* **108**, 19641–19646, https://doi.org/10.1073/pnas.1104551108 (2011).
3. Schoch, C. L. *et al.* Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. *Proc. Natl. Acad. Sci. (PNAS)* **109**, 6241–6246, https://doi.org/10.1073/pnas.1117018109 (2012).
4. Kress, W. J., Wurdack, K. J., Zimmer, E. A., Weigt, L. A. & Janzen, D. H. Use of DNA barcodes to identify flowering plants. *Proc. Natl. Acad. Sci. (PNAS)* **102**, 8369–8374, https://doi.org/10.1073/pnas.0503123102 (2005).
5. Newmaster, S. G., Fazekas, A. J., Steeves, R. A. D. & Janovec, J. Testing candidate plant barcode regions in the Myristicaceae. *Mol. Ecol. Resour.* **8**, 480–490, https://doi.org/10.1111/j.1471-8286.2007.02002.x (2008).
6. Lahaye, R. *et al.* DNA barcoding the floras of biodiversity hotspots. *Proc. Natl. Acad. Sci. USA (PNAS)* **105**, 2923–2928, https://doi.org/10.1073/pnas.0709936105 (2008).
7. Hollingsworth, P. M. *et al.* A DNA barcode for land plants. *Proc. Natl. Acad. Sci. (PNAS)* **106**, 12794–12797, https://doi.org/10.1073/pnas.0905845106 (2009).
8. Li, X. *et al.* Plant DNA barcoding: from gene to genome. *Biol. Rev.* **90**, 157–166, https://doi.org/10.1111/brv.12104 (2015).

9. Ruppert, K. M., Kline, R. J. & Rahman, M. S. Past, present, and future perspectives of environmental DNA (eDNA) metabarcoding: A systematic review in methods, monitoring, and applications of global eDNA. *Glob. Ecol. Conserv*. **17**, https://doi.org/10.1016/j.gecco.2019.e00547 (2019).
10. Bell, K. L. *et al*. Plants, pollinators and their interactions under global ecological change: The role of pollen DNA metabarcoding. *Mol. Ecol.* https://doi.org/10.1111/mec.16689 (2022).
11. Bell, K. L. *et al*. Pollen DNA barcoding: current applications and future prospects. *Genome* **59**, 629–640, https://doi.org/10.1139/gen-2015-0200 (2016).
12. Dubois, B. *et al*. A detailed workflow to develop QIIME2-formatted reference databases for taxonomic analysis of DNA metabarcoding data. *BMC Genom. Data* **23**, 53, https://doi.org/10.1186/s12863-022-01067-5 (2022).
13. Frigerio, J. *et al*. DNA-Based Herbal Teas' Authentication: An ITS2 and psbA-trnH Multi-Marker DNA Metabarcoding Approach. *Plants* **10**, https://doi.org/10.3390/plants10102120 (2021).
14. Zhang, G. X. *et al*. Tracing the Edible and Medicinal Plant Pueraria montana and Its Products in the Marketplace Yields Subspecies Level Distinction Using DNA Barcoding and DNA Metabarcoding. *Front. Pharmacol.* **11**, https://doi.org/10.3389/fphar.2020.00336 (2020).
15. Anthoons, B. *et al*. Metabarcoding reveals low fidelity and presence of toxic species in short chain-of-commercialization of herbal products. *J Food Compost Anal.* **97**, https://doi.org/10.1016/j.jfca.2020.103767 (2021).
16. Moorhouse-Gann, R. J. *et al*. New universal ITS2 primers for high-resolution herbivory analyses using DNA metabarcoding in both tropical and temperate zones. *Sci. Rep.* **8**, 8542, https://doi.org/10.1038/s41598-018-26648-2 (2018).
17. Wang, B. *et al*. Seasonal variations in the plant diet of the Chinese Monal revealed by fecal DNA metabarcoding analysis. *Avian Res.* **13**, https://doi.org/10.1016/j.avrs.2022.100034 (2022).
18. Fujii, T., Ueno, K., Shirako, T., Nakamura, M. & Minami, M. Identification of *Lagopus muta japonica* food plant resources in the Northern Japan Alps using DNA metabarcoding. *PLoS One* **17**, https://doi.org/10.1371/journal.pone.0252632 (2022).
19. König, S., Krauss, J., Keller, A., Bofinger, L. & Steffan-Dewenter, I. Phylogenetic relatedness of food plants reveals highest insect herbivore specialization at intermediate temperatures along a broad climatic gradient. *Glob. Change Biol.* **28**, 4027–4040, https://doi.org/10.1111/gcb.16199 (2022).
20. Bell, K. L. *et al*. Applying pollen DNA metabarcoding to the study of plant–pollinator interactions. *Appl. Plant Sci.* **5**, 1600124, https://doi.org/10.3732/apps.1600124 (2017).
21. Arstingstall, K. A. *et al*. Capabilities and limitations of using DNA metabarcoding to study plant-pollinator interactions. *Mol. Ecol.* **30**, 5266–5297, https://doi.org/10.1111/mec.16112 (2021).
22. Encinas-Viso, F. *et al*. Pollen DNA metabarcoding reveals cryptic diversity and high spatial turnover in alpine plant-pollinator networks. *Mol. Ecol.* https://doi.org/10.1111/mec.16682 (2022).
23. Bell, K. L. *et al*. Plants, pollinators and their interactions under global ecological change: The role of pollen DNA metabarcoding. *Mol. Ecol.*, 1–18, https://doi.org/10.1111/mec.16689 (2022).
24. Hawkins, J. *et al*. Using DNA Metabarcoding to Identify the Floral Composition of Honey: A New Tool for Investigating Honey Bee Foraging Preferences. *PLoS One* **10**, e0134735, https://doi.org/10.1371/journal.pone.0134735 (2015).
25. Milla, L., Schmidt-Lebuhn, A., Bovill, J. & Encinas-Viso, F. Monitoring of honey bee floral resources with pollen DNA metabarcoding as a complementary tool to vegetation surveys. *Ecol. Solut. Evid*. **3**, https://doi.org/10.1002/2688-8319.12120 (2022).
26. Khansaritoreh, E. *et al*. Employing DNA metabarcoding to determine the geographical origin of honey. *Heliyon* **6**, https://doi.org/10.1016/j.heliyon.2020.e05596 (2020).
27. Korpelainen, H. & Pietilainen, M. Biodiversity of pollen in indoor air samples as revealed by DNA metabarcoding. *Nord. J. Bot.* **35**, 602–608, https://doi.org/10.1111/njb.01623 (2017).
28. Omelchenko, D. O. *et al*. Assessment of ITS1, ITS2, 5 '-ETS, and trnL-F DNA Barcodes for Metabarcoding of Poaceae Pollen. *Diversity* **14**, https://doi.org/10.3390/d14030191 (2022).
29. Fahner, N. A., Shokralla, S., Baird, D. J. & Hajibabaei, M. Large-Scale Monitoring of Plants through Environmental DNA Metabarcoding of Soil: Recovery, Resolution, and Annotation of Four DNA Markers. *PLoS One* **11**, https://doi.org/10.1371/journal.pone.0157505 (2016).
30. Vasconcelos, S. *et al*. Unraveling the plant diversity of the Amazonian canga through DNA barcoding. *Ecol. Evol.* **11**, 13348–13362, https://doi.org/10.1002/ece3.8057 (2021).
31. Timpano, E. K., Scheible, M. K. R. & Meiklejohn, K. A. Optimization of the second internal transcribed spacer (ITS2) for characterizing land plants from soil. *PLoS One* **15**, https://doi.org/10.1371/journal.pone.0231436 (2020).
32. Yau, S. *et al*. Mantoniella beaufortii and Mantoniella baffinensis sp. nov. (Mamiellales, Mamiellophyceae), two new green algal species from the high arctic(1). *J. Phycol.* **56**, 37–51, https://doi.org/10.1111/jpy.12932 (2020).
33. Liu, Y. L., Xu, C., Dong, W. P., Yang, X. Y. & Zhou, S. L. Determination of a criminal suspect using environmental plant DNA metabarcoding technology. *Forensic Sci. Int*. **324**, https://doi.org/10.1016/j.forsciint.2021.110828 (2021).
34. Higashi, Y., Hirota, S. K., Suyama, Y. & Yahara, T. Geographical and seasonal variation of plant taxa detected in faces of *Cervus nippon yakushimae* based on plant DNA analysis in Yakushima Island. *Ecol. Res.* **37**, 582–597, https://doi.org/10.1111/1440-1703.12319 (2022).
35. Fox, G. *et al*. Complex urban environments provide *Apis mellifera* with a richer plant forage than suburban and more rural landscapes. *Ecol. Evol.* **12**, https://doi.org/10.1002/ece3.9490 (2022).
36. Quaresma, A. *et al*. Preservation methods of honey bee-collected pollen are not a source of bias in ITS2 metabarcoding. *Environ. Monit. Assess*. **193**, https://doi.org/10.1007/s10661-021-09563-4 (2021).
37. Benson, D. A. *et al*. GenBank. *Nucleic Acids Res.* **45**, D37–D42, https://doi.org/10.1093/nar/gkw1070 (2017).
38. Banchi, E. *et al*. PLANiTS: a curated sequence reference dataset for plant ITS DNA metabarcoding. *Database* **2020**, https://doi.org/10.1093/database/baz155 (2020).
39. Keller, A. *et al*. BCdatabaser: on-the-fly reference database creation for (meta-)barcoding. *Bioinformatics* **36**, 2630–2631, https://doi.org/10.1093/bioinformatics/btz960 (2020).
40. Kraaijeveld, K. *et al*. Efficient and sensitive identification and quantification of airborne pollen using next-generation DNA sequencing. *Mol. Ecol. Resour.* **15**, 8–16, https://doi.org/10.1111/1755-0998.12288 (2015).
41. Keller, A. *et al*. Evaluating multiplexed next-generation sequencing as a method in palynology for mixed pollen samples. *Plant Biol.* **17**, 558–566, https://doi.org/10.1111/plb.12251 (2015).
42. Richardson, R. T. *et al*. Rank-based characterization of pollen assemblages collected by honey bees using a multi-locus metabarcoding approach. *Appl. Plant Sci.* **3**, 1500043, https://doi.org/10.3732/apps.1500043 (2015).
43. Edwards, C. E., Swift, J. F., Lance, R. F., Minckley, T. A. & Lindsay, D. L. Evaluating the efficacy of sample collection approaches and DNA metabarcoding for identifying the diversity of plants utilized by nectivorous bats. *Genome* **62**, 19–29, https://doi.org/10.1139/gen-2018-0102 (2019).
44. Rognes, T., Flouri, T., Nichols, B., Quince, C. & Mahé, F. VSEARCH: A versatile open source tool for metagenomics. *PeerJ* (2016).
45. Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**, 2460–2461, https://doi.org/10.1093/bioinformatics/btq461 (2010).
46. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410, https://doi.org/10.1016/S0022-2836(05)80360-2 (1990).
47. Wang, Q., Garrity, G. M., Tiedje, J. M. & Cole, J. R. Naïve Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy. *Appl Environ Microbiol* **73**, 5261–5267, https://doi.org/10.1128/AEM.00062-07 (2007).

48. Edgar, R. C. SINTAX: a simple non-Bayesian taxonomy classifier for 16S and ITS sequences. *bioRxiv*, 074161, https://doi.org/10.1101/074161 (2016).
49. Schultz, J. *et al*. The internal transcribed spacer 2 database—a web server for (not only) low level phylogenetic analyses. *Nucleic Acids Res*. **34**, W704–W707, https://doi.org/10.1093/nar/gkl129 (2006).
50. Ankenbrand, M. J., Keller, A., Wolf, M., Schultz, J. & Förster, F. ITS2 Database V: Twice as Much. *Mol. Biol. Evol.* **32**, 3030–3032, https://doi.org/10.1093/molbev/msv174 (2015).
51. Sickel, W. *et al*. Increased efficiency in identifying mixed pollen samples by meta-barcoding with a dual-indexing approach. *BMC Ecology* **15**, 1–9, https://doi.org/10.1186/s12898-015-0051-y (2015).
52. Bell, K. L., Loeffler, V. M. & Brosi, B. J. An rbcL reference library to aid in the identification of plant species mixtures by DNA metabarcoding. *Appl. Plant Sci*. **5**, https://doi.org/10.3732/apps.1600110 (2017).
53. Wirta, H., Abrego, N., Miller, K., Roslin, T. & Vesterinen, E. DNA traces the origin of honey by identifying plants, bacteria and fungi. *Sci. Rep*. **11**, https://doi.org/10.1038/s41598-021-84174-0 (2021).
54. Bell, K. L. *et al*. Comparing whole-genome shotgun sequencing and DNA metabarcoding approaches for species identification and quantification of pollen species mixtures. *Ecol. Evol.* **11**, 16082–16098, https://doi.org/10.1002/ece3.8281 (2021).
55. Curd, E. E. *et al*. Anacapa Toolkit: An environmental DNA toolkit for processing multilocus metabarcode datasets. *Methods Ecol. Evol.* **10**, 1469–1475, https://doi.org/10.1111/2041-210x.13214 (2019).
56. Richardson, R. T., Sponsler, D. B., McMinn-Sauder, H. & Johnson, R. M. MetaCurator: A hidden Markov model-based toolkit for extracting and curating sequences from taxonomically-informative genetic markers. *Methods Ecol. Evol.* **11**, 181–186, https://doi.org/10.1111/2041-210x.13314 (2020).
57. Keck, F., Couton, M. & Altermatt, F. Navigating the seven challenges of taxonomic reference databases in metabarcoding analyses. *Mol. Ecol. Resour*., https://doi.org/10.1111/1755-0998.13746.
58. R: A language and environment for statistical computing (Vienna, Austria 2013).
59. Hackl, T., Hedrich, R., Schultz, J. & Förster, F. proovread: large-scale high-accuracy PacBio correction through iterative short read consensus. *Bioinformatics* **30**, 3004–3011, https://doi.org/10.1093/bioinformatics/btu392 (2014).
60. Deshpande, V. *et al*. Fungal identification using a Bayesian classifier and the Warcup training set of internal transcribed spacer sequences. *Mycologia* **108**, 1–5, https://doi.org/10.3852/14-293 (2016).
61. Chen, S. *et al*. Validation of the ITS2 region as a novel DNA barcode for identifying medicinal plant species. *PLoS One* **5**, e8613, https://doi.org/10.1371/journal.pone.0008613 (2010).
62. White, T. J., Bruns, T., Lee, S. J. W. T. & Taylor, J. in *PCR protocols: a guide to methods applications* (ed Gelfand, D. H. Innis, M. A., Sninsky, J. J., White, T. J.) 315-322 (Academic Press, 1990).
63. Kumar, S., Stecher, G., Li, M., Knyaz, C. & Tamura, K. MEGA X: molecular evolutionary genetics analysis across computing platforms. *Mol. Biol. Evol.* **35**, 1547–1549, https://doi.org/10.1093/molbev/msy096 (2018).
64. Quaresma, A. *et al*. ITS2 Global database. *Zenodo* https://doi.org/10.5281/zenodo.7968519 (2023).
65. Quaresma, A. *et al*. ITS2 Crop database. *Zenodo* https://doi.org/10.5281/zenodo.7969940 (2023).
66. Quaresma, A. *et al*. ITS2 European countries. *Zenodo* https://doi.org/10.5281/zenodo.7970046 (2023).
67. Cheng, T. *et al*. Barcoding the kingdom Plantae: new PCR primers for ITS regions of plants with improved universality and specificity. *Mol. Ecol. Resour.* **16**, 138–149, https://doi.org/10.1111/1755-0998.12438 (2016).

Acknowledgements

AQ acknowledges the PhD scholarship (2020.05155.BD), funded by the Portuguese Foundation for Science and Technology (FCT). This work was developed in the framework of INSIGNIA – Environmental monitoring of pesticide use through honeybees (SANTE/E4/SI2.788418-SI2.788452-INSIGINIA-PP-1-1-2018) and INSIGNIA-EU - Preparatory action for monitoring of environmental pollution using honey bees (Procurement procedure ENV/2021/OP/0014 of 28-09-2021). FCT provided financial support by national funds (FCT/MCTES) to CIMO (UIDB/00690/2020 and UIDP/00690/2020) and SusTEC (LA/P/0007/2021).

Author contributions

A.Q., A.K. and M.A.P. conceived the ideas and designed the methodology. A.K. and A.Q. developed the scripts and the datasets. M.A. and A.K. developed the web interface. J.R. extracted the list of European flora from Euro + Med PlantBase and assisted with the computational resources. Plant leaves for the manual sequence addition were provided by M.A.P., J.A., R.B., V.B., K.G., F.H., O.K., M.P., I.R. and F.V. A.Q., C.A.Y.G. and M.H. performed the DNA extractions of the plant leaves. A.Q., M.A.P. and A.K., wrote the manuscript. All the authors critically reviewed the manuscript for important intellectual content. JvdS acquired INSIGNIA's funding.

Funding

Open Access funding enabled and organized by Projekt DEAL.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41597-024-02962-5.

Correspondence and requests for materials should be addressed to A.K.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.