Pollen identification by ITS2 metabarcoding: curation of the sequences retrieved from GenBank to build a reference database

Quaresma, Andreia[1,2,3]; Keller, Alexander[4]; Rufino, José[5]; van der Steen, Jozef[6]; Pinto, M. Alice[1,2]

[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]Organismic and Cellular Interactions, Biocenter, Ludwig-Maximilians-Universität München, Planegg, Germany; [5]Research Centre in Digitalization and Intelligent Robotics (CeDRI), Campus de Santa Apolónia, 5300-253 Bragança, Portugal; [6]Alveus AB Consultancy, Oisterwijk, Netherlands

Framework

A powerful way of studying the quality of the environment is by examining pollen collected by honey bees as it contains information on the availability of plant sources, and spatial and temporal floral diversity. The botanical origin of pollen is typically addressed by classical palynology, a costly approach that provides low taxonomic resolution, is time-consuming and labour intensive, and requires plant taxonomy expertise. However, as high-throughput sequencing is becoming increasingly affordable, pollen metabarcoding is gaining momentum and it is a promising alternative to classical palynology. Given that one of the main drawbacks of pollen metabarcoding is the lack of good quality reference databases for the barcode of choice, developing tools to help curating reference databases is of uttermost importance.

BCdatabaser

BCdatabaser was used to automatically generate the standardized database, for arbitrary barcodes and taxonomic groups, from GenBank (Keller et al. 2020)

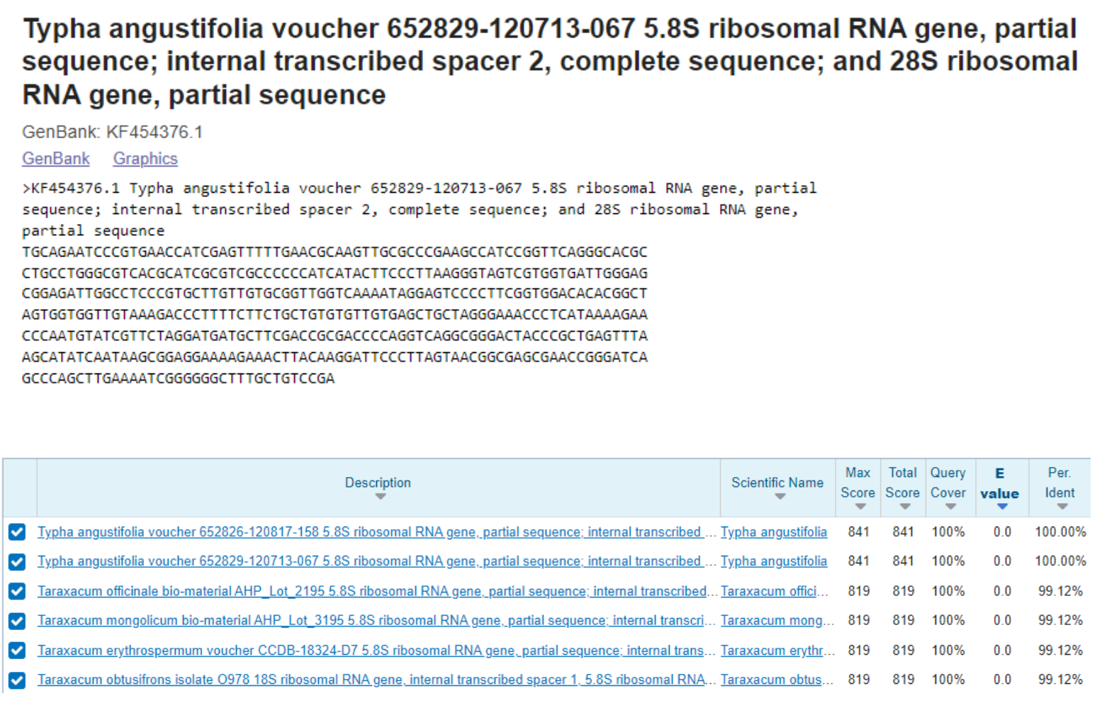
GenBank sequences

Viridiplantae ITS2 sequences were downloaded from GenBank

Problems

Misidentified sequences were detected in GenBank downloads

Plant sequences assigned to the wrong taxa



Typha angustifolia voucher 652829-120713-067 5.8S ribosomal RNA gene, partial sequence; internal transcribed spacer 2, complete sequence; and 28S ribosomal RNA gene, partial sequence

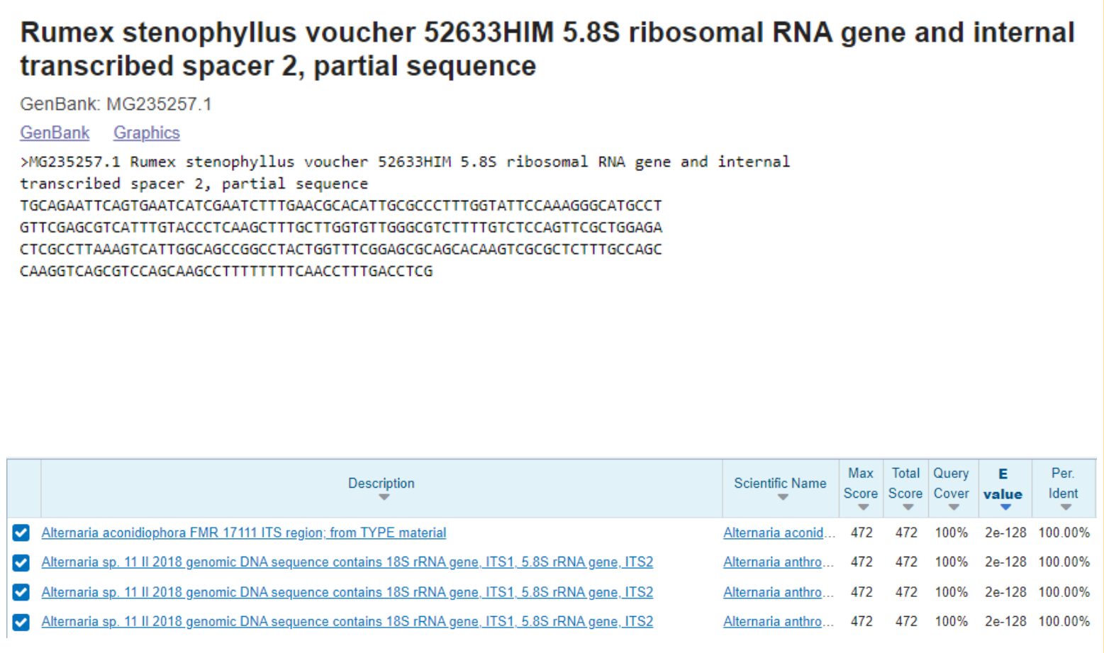
GenBank: KF454376.1

GenBank Graphics

```
>KF454376.1 Typha angustifolia voucher 652829-120713-067 5.8S ribosomal RNA gene, partial
sequence; internal transcribed spacer 2, complete sequence; and 28S ribosomal RNA gene,
partial sequence
TGCAGAATCCCGTGAACCATCGAGTTTTTGAACGCAAGTTGCGCCCGAAGCCATCCGGTTCAGGGCACGC
CTGCCTGGGCGTCACGCATCGCGTCGCCCCCCATCATACTTCCCTTAAGGGTAGTCGTGGTGATTGGGAG
CGGAGATTGGCCTCCCGTGCTTGTTGTGCGGTTGGTCAAAATAGGAGTCCCCTTCGGTGGACACACGGCT
AGTGGTGGTTGTAAAGACCCTTTTCTTCTGCTGTGTGTTGTGAGCTGCTAGGGAAACCCTCATAAAAGAA
CCCAATGTATCGTTCTAGGATGATGCTTCGACCGCGACCCCAGGTCAGGCGGGACTACCCGCTGAGTTTA
AGCATATCAATAAGCGGAGGAAAAGAAACTTACAAGGATTCCCTTAGTAACGGCGAGCGAACCGGGATCA
GCCCAGCTTGAAAATCGGGGGGCTTTGCTGTCCGA
```

	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident
☑	Typha angustifolia voucher 652826-120817-158 5.8S ribosomal RNA gene, partial sequence; internal transcribed ...	Typha angustifolia	841	841	100%	0.0	100.00%
☑	Typha angustifolia voucher 652829-120713-067 5.8S ribosomal RNA gene, partial sequence; internal transcribed ...	Typha angustifolia	841	841	100%	0.0	100.00%
☑	Taraxacum officinale bio-material AHP_Lot_2195 5.8S ribosomal RNA gene, partial sequence; internal transcribed...	Taraxacum offici...	819	819	100%	0.0	99.12%
☑	Taraxacum mongolicum bio-material AHP_Lot_3195 5.8S ribosomal RNA gene, partial sequence; internal transcri...	Taraxacum mong...	819	819	100%	0.0	99.12%
☑	Taraxacum erythrospermum voucher CCDB-18324-D7 5.8S ribosomal RNA gene, partial sequence; internal trans...	Taraxacum erythr...	819	819	100%	0.0	99.12%
☑	Taraxacum obtusifrons isolate O978 18S ribosomal RNA gene, internal transcribed spacer 1, 5.8S ribosomal RNA...	Taraxacum obtus...	819	819	100%	0.0	99.12%

Fungi sequences identified as plants species



Rumex stenophyllus voucher 52633HIM 5.8S ribosomal RNA gene and internal transcribed spacer 2, partial sequence

GenBank: MG235257.1

GenBank Graphics

```
>MG235257.1 Rumex stenophyllus voucher 52633HIM 5.8S ribosomal RNA gene and internal
transcribed spacer 2, partial sequence
TGCAGAATTCAGTGAATCATCGAATCTTTGAACGCACATTGCGCCCTTTGGTATTCCAAAGGGCATGCCT
GTTCGAGCGTCATTTGTACCCTCAAGCTTTGCTTGGTGTTGGGCGTCTTTTGTCTCCAGTTCGCTGGAGA
CTCGCCTTAAAGTCATTGGCAGCCGGCCTACTGGTTTCGGAGCGCAGCACAAGTCGCGCTCTTTGCCAGC
CAAGGTCAGCGTCCAGCAAGCCTTTTTTTTCAACCTTTGACCTCG
```

	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident
☑	Alternaria aconidiophora FMR 17111 ITS region; from TYPE material	Alternaria aconid...	472	472	100%	2e-128	100.00%
☑	Alternaria sp. 11 II 2018 genomic DNA sequence contains 18S rRNA gene, ITS1, 5.8S rRNA gene, ITS2	Alternaria anthro...	472	472	100%	2e-128	100.00%
☑	Alternaria sp. 11 II 2018 genomic DNA sequence contains 18S rRNA gene, ITS1, 5.8S rRNA gene, ITS2	Alternaria anthro...	472	472	100%	2e-128	100.00%
☑	Alternaria sp. 11 II 2018 genomic DNA sequence contains 18S rRNA gene, ITS1, 5.8S rRNA gene, ITS2	Alternaria anthro...	472	472	100%	2e-128	100.00%

Script

A script in bash and R was developed to curate the ITS2 global reference database for the EU space

ITS2 database curation steps

1. Removal of Fungi sequences using the RDP Classifier Fungal ITS database
2. Pairwise alignment using vsearch v2.14.1 (Rognes et al. 2016)
3. Development of a R script for the removal of low identity alignments among plant species
4. Development of single databases for the 27 EU countries from the curated global database using Euro+Med PlantBase (https://www.emplantbase.org/home.html), and GBIF (https://www.gbif.org/) list

Some crop taxa were undetected (e.g. *Malus sp.* and *Pyrus sp.*), as revealed by comparing metabarcoding against palynological data

Crops reference database

A reference database was developed for the European Union crops by adding more sequences/species from GenBank to overcome the identity percentage removal after the pairwise alignment

Developed pipeline

- ✓ Allows easy and fast development of reference databases
- ✓ Facilitates regular update of reference databases
- ✓ Can be applied to other barcodes and organisms

References

Keller, A., Hohlfeld, S., Kolter, A., Schultz, J., Gemeinholzer, B., Ankenbrand, M. J. (2020). BCdatabaser: on-the-fly reference database creation for (meta-) barcoding. Bioinformatics, 36(8), 2630-2631.

Rognes, T., Flouri, T., Nichols, B., Quince, C., Mahé, F. (2016). VSEARCH: a versatile open source tool for metagenomics. PeerJ, 4, e2584.

Acknowledgements:

AQ acknowledges the PhD scholarship (DFA/BD/5155/2020) funded by FCT. This study was funded by DG Environment through the INSIGNIA-EU preparatory action for monitoring of environmental pollution using honey bees. Procurement procedure ENV/2021/OP/0014 of 28-09-2021.