

Aberystwyth University

Barcode UK

Jones, Laura E.; Twyford, Alex D.; Ford, Col R.; Rich, Tim C. G.; Davies, Helena; Forrest, Laura L.; Hart, Michelle L.; McHaffie, Heather; Brown, Max R.; Hollingsworth, Peter M.; De Vere, Natasha

Published in:
Molecular Ecology Resources

DOI:
[10.1111/1755-0998.13388](https://doi.org/10.1111/1755-0998.13388)

Publication date:
2021

Citation for published version (APA):

Jones, L. E., Twyford, A. D., Ford, C. R., Rich, T. C. G., Davies, H., Forrest, L. L., Hart, M. L., McHaffie, H., Brown, M. R., Hollingsworth, P. M., & De Vere, N. (2021). Barcode UK: A complete DNA barcoding resource for the flowering plants and conifers of the United Kingdom. *Molecular Ecology Resources*.
<https://doi.org/10.1111/1755-0998.13388>

Document License CC BY

General rights

Copyright and moral rights for the publications made accessible in the Aberystwyth Research Portal (the Institutional Repository) are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the Aberystwyth Research Portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the Aberystwyth Research Portal


Take down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

tel: +44 1970 62 2400
email: is@aber.ac.uk

RESOURCE ARTICLE

Barcode UK: A complete DNA barcoding resource for the flowering plants and conifers of the United Kingdom

Laura Jones¹ | Alex D. Twyford^{2,3}  | Col R. Ford¹  | Tim C. G. Rich⁴ |
Helena Davies¹ | Laura L. Forrest² | Michelle L. Hart² | Heather McHaffie² |
Max R. Brown³ | Peter M. Hollingsworth² | Natasha de Vere^{1,5} 

¹National Botanic Garden of Wales, Llanarthne, UK

²Royal Botanic Garden Edinburgh, Edinburgh, UK

³School of Biological Sciences, Institute of Evolutionary Biology, Edinburgh, UK

⁴Freelance Botanist, Cardiff, UK

⁵Institute of Biological, Environmental and Rural Sciences, Aberystwyth University, Aberystwyth, UK

Correspondence

Natasha de Vere, National Botanic Garden of Wales, Llanarthne, Carmarthenshire, SA32 8HG, UK.

Email: natasha.devere@gardenofwales.org.uk

Funding information

Welsh Government's Enabling Natural Resources and Well-being in Wales Grants; Welsh Government Rural Communities – Rural Development Programme 2014-2020, which is funded by the European Agricultural Fund for Rural Development and the Welsh Government

Abstract

DNA barcoding and metabarcoding provide new avenues for investigating biological systems. These techniques require well-curated reference libraries with extensive coverage. Generating an exhaustive national DNA barcode reference library can open up new avenues of research in ecology, evolution and conservation, yet few studies to date have created such a resource. In plant DNA barcoding, herbarium collections provide taxonomically robust material but also pose challenges in lab processing. Here, we present a national DNA barcoding resource covering all of the native flowering plants and conifers of the United Kingdom. This represents 1,482 plant species, with the majority of specimens (81%) sourced from herbaria. Using Sanger sequencing of the plant DNA barcode markers, *rbcl*, *matK*, and ITS2, at least one DNA barcode was retrieved from 98% of the UK flora. We sampled from multiple individuals, resulting in a species coverage for *rbcl* of 96% (4,477 sequences), 90% for *matK* (3,259 sequences) and 75% for ITS2 (2,585 sequences). Sequence recovery was lower for herbarium material compared to fresh collections, with the age of the specimen having a significant effect on the success of sequence recovery. Species level discrimination was highest with ITS2, however, the ability to successfully retrieve a sequence was lowest for this region. Analyses of the genetic distinctiveness of species across a complete flora showed DNA barcoding to be informative for all but the most taxonomically complex groups. The UK flora DNA barcode reference library provides an important resource for many applications that require plant identification from DNA.

KEYWORDS

conifers, DNA barcoding, flowering plants, UK

1 | INTRODUCTION

The identification of plant species is vitally important for the monitoring, conservation and utilisation of biodiversity but is limited by the availability of taxonomic expertise. DNA barcoding, the method of characterising

species using one or a few standardised regions of DNA (Hebert et al., 2003), has been used to both characterise existing biodiversity and identify new or cryptic species. Species identification is possible even where morphological identification was previously limited, as in juvenile, sterile, mixed or degraded plant material (Hollingsworth et al., 2016).

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2021 The Authors. *Molecular Ecology Resources* published by John Wiley & Sons Ltd.

The applications of DNA barcoding and DNA metabarcoding techniques (when DNA barcoding is used in combination with high throughput sequencing from multispecies samples) cover a wide range of purposes, with the potential for rapid identification of species composition from many different sources of DNA (Deiner et al., 2017; Hollingsworth et al., 2016). In environmental monitoring, DNA metabarcoding has been used to detect the presence of rare species (Harper et al., 2018), to reveal pollinator communities by retrieving insect DNA from the flowers they visit (Thomsen & Sigsgaard, 2019), monitoring biodiversity change by reconstructing historical plant communities (Edwards et al., 2017) and to identify pollen from the air (Brennan et al., 2019) or from the bodies of insects (Lucas et al., 2018). DNA barcoding has been widely employed for diet analysis, as when examining food partitioning in herbivores (Kartzinel et al., 2015) or the trophic specialisation of bats (Arrizabalaga-Escudero et al., 2019).

Potential commercial applications involve using DNA barcoding to assess plant based products for sale, such as confirming the claimed identity of herbal medicines (Raclariu et al., 2018), verifying the geographic origin of honey (Prosser & Hebert, 2017; Saravanan et al., 2019), or checking the composition of plant species present in commercial tea preparations (de Boer et al., 2017; Stoeckle et al., 2011).

Underpinning this diversity of applications are well-curated reference libraries, with associated voucher specimens and sample metadata (Hebert et al., 2003). By using a high-quality reference library, the accuracy achieved within the many applications that require DNA-based species identification can be improved (Bell et al., 2016; Landi et al., 2014). Sequences with these metadata provide the opportunity to appropriately evaluate any identification results (Pentinsaari et al., 2020). The gold standard is a complete reference database of sequences for multiple, well-verified individuals, for all species from a country. However, this goal is challenging to achieve in most countries and for most organismal groups, with sample availability, high species richness and taxonomic expertise often being limiting factors. Large-scale DNA barcode reference data sets combining exhaustive taxonomic coverage and multispecies sampling have been completed for numerous animal groups (Hebert et al., 2016) including Canadian spiders (1,460 species) (Blagoev et al., 2016) and New Zealand's birds (236 species) (Tizard et al., 2019). In animals, the use of a single DNA region for most taxa, cytochrome c oxidase I (COI), for which there are reliable primers and standardised protocols, makes the process of generating large reference databases routine. In plants, however, the logistics are more complicated, as no single DNA region provides suitable levels of species discrimination, and instead two core plastid DNA barcode regions, parts of the genes *rbcl* (c. 600 bp) and *matK* (c. 800 bp) are commonly used (CBOL Plant Working Group, 2009), often alongside a portion of the more variable internal transcribed spacers of nuclear ribosomal DNA (ITS2) (Chen et al., 2010; Kuzmina et al., 2017; Li et al., 2011; Yao et al., 2010), or sometimes the plastid intergenic spacer *trnH-psbA* (Kress & Erickson, 2007).

The UK has a long history of botanical recording, with an extremely well-studied flora (Preston et al., 2002; Walker & Preston,

2006). There are extensive data sets on alien taxa (Stace & Crawley, 2015), conservation status, current distributions and distribution change over time (Preston et al., 2002), cytotype variation, ecological traits (Hill et al., 2004), genetic diversity (Ruhsam et al., 2018), genome sizes (Pellicer & Leitch, 2020), and hybridisation (Stace et al., 2015). This information provides a rich contextual background for UK DNA barcoding data that can then be used in evolutionary and ecological studies. There are also extensive sample resources, with over half the native flora held in the living collections of botanic gardens, and ~1.5 million specimens representing >99% UK native species in major UK herbaria (Clubbe et al., 2020). By using national herbarium collections, a comprehensive collection of UK plant species can be accessed for DNA barcoding (de Vere et al., 2012), providing accurate taxonomic identification, source material for DNA extraction, and voucher specimens that can be linked to the sequencing data. In addition, the cost and time associated with establishing a reference database can be significantly reduced by using herbarium collections (Kuzmina et al., 2017). In connecting well-annotated and curated collections of specimens, housed in open-access repositories, to their DNA sequences, the voucher specimens can be easily revisited for any further research purposes.

Here we present the complete reference library for the UK native flora for *rbcl*, *matK* and ITS2, providing a three-locus barcode library representing 1,482 British native plant species. We provide coverage for all of the native and archaeophyte (naturalised prior to 1500 CE) flowering plants and conifers of the UK, with this representing a major taxonomic and geographic extension of the reference library previously created for Wales (de Vere et al., 2012). For the Welsh flora DNA barcode database, 1,143 plant species were sequenced for plastid regions only (*rbcl* and *matK*). In this study, an additional 339 plant species were targeted to gain representation for plant species not found within Wales, but present within England, Scotland and Northern Ireland. We also, for the first time, target the ITS2 marker for the 1,482 UK species. Our primary aim is to provide a resource that can be used in integrated ecological and evolutionary analyses at the scale of a whole flora. Given our extensive use of herbarium material, we also assess how the success of Sanger-based sequence recovery is affected by specimen age and by higher-level taxonomy; how the three markers *rbcl*, *matK*, and ITS2 vary in their ability to successfully recover a sequence, how the three markers vary in their species level discrimination and discuss their applicability for DNA metabarcoding studies.

2 | MATERIALS AND METHODS

2.1 | Sample collection

The UK native flora targeted here represents 1,482 flowering plants and conifers, representing 550 genera, 104 families, and 35 orders (Preston et al., 2002; Stace, 2019). Taxonomic classifications use Stace (2019). The apomictic microspecies complexes of *Hieracium*, *Rubus* and *Taraxacum*, that are difficult to distinguish morphologically

as species (Ellstrand et al., 1996), were represented using aggregate species groupings. Leaf material was obtained for 1,473 (99%) of the targeted species. In total, 6,100 individuals were sampled, 4,965 from herbarium specimens and 1,135 from recent collections of leaf material into silica-gel desiccant, from throughout the UK. Of the 6,100 specimens, 4,272 were sampled and extracted during previous work on DNA barcoding the Welsh flora (de Vere et al., 2012), while 1,828 specimens represent newly sequenced herbarium and recent field collections, to gain coverage for those UK plant species not present in Wales. At least three individuals of each species were targeted for collection.

For samples from herbarium specimens, approximately 2 cm² of leaf material was removed from a part of the specimen that would not detract from its scientific value. Further criteria for specimen selection in order of importance included: being a morphologically typical representative of the species, being as recently collected as possible within the specimens available; being collected from geographically distinct locations; and having additional taxonomic verification present for the specimen. The majority of herbarium samples came from specimens housed in the National Museum Wales (NMW) collections (89%), with additional samples from the Royal Botanic Garden Edinburgh (E), National Botanic Garden of Wales (NBGW), National Museums Liverpool (LIV), Dublin Botanic Garden (DBN), Bangor University (UCNW) and Aberystwyth University (ABS).

For freshly collected material, approximately 2 cm² of undamaged leaf or flower material was sampled in the field and placed into bags of silica gel desiccant to dry. Regional floras, online databases from the Botanical Society of Britain and Ireland, and knowledge from local recorders were used to locate species for sampling. Herbarium vouchers were created for all freshly collected material, with the exception of threatened species, and were placed in the herbaria at the National Botanic Garden of Wales, National Museum Wales and the Royal Botanic Garden, Edinburgh. Where collection of a voucher specimen was prohibited, a photographic voucher was taken instead. Full collection recommendations are detailed in Data S1.

2.2 | DNA extraction and amplification

For freshly collected leaf material, Plant DNeasy 96 kits (Qiagen) were used, following the manufacturer's protocol, using leaves dried in silica gel. For herbarium samples, the Plant DNeasy protocol was modified to improve success, following de Vere et al., (2012). This used a lysis buffer containing 400 µl AP1 buffer from the Qiagen kit, 80 µl DTT (0.75 mg/ml) (Melford Laboratories) and 20 µl proteinase K (1 mg/ml) (Sigma). Each sample had 400 µl of this lysis buffer added to the leaf material before disruption with a TissueLyser II (Qiagen) with 3 mm tungsten carbide beads. The incubation phase using the modified AP1 buffer was then extended to 1 h at 65°C. The final elution stage with the AE buffer was extended to 15 min for the herbarium samples.

Amplification for the *rbcl* marker was carried out using primers *rbcl*A-F and *rbcl*Lr590 (Table S1, de Vere et al., 2012). For *matK*, multiple primer combinations were used, beginning with universal primer combinations, *matK*-390F with *matK*1326R, and *matK*-2.1a with *matK*-3Fkim-R. If these failed, order specific primers were then attempted (Table S2, de Vere et al., 2012). For the ITS2 region, ITS2F and ITS3R primers were used (Table S3, Chen et al., 2010). Only one primer pair was used for ITS2 amplification, to reduce the risk of generating sequences from different paralogous copies of ITS / ribosomal arrays (Yao et al., 2010). Failed amplifications were attempted a further two times for each sample and primer pair.

The PCR amplifications were carried out in 20 µl reactions, using 10 µl of Biomix (Bioline), 0.4 µl of each primer (10 µM), 0.8 µl of BSA (1 mg/ml), 6.4 µl of molecular biology grade H₂O, and 2 µl of template DNA. PCR conditions for *rbcl* and *matK* were 95°C for 2 min, followed by 95°C for 30 s, 50°C for 90 s, and 72°C for 40 s, for 45 cycles, followed by 72°C for 5 min and 30°C for 10 sec. The ITS2 PCR cycle was as follows: 94°C for 5 min, followed by 94°C for 30 s, 56°C for 30 s and 72°C for 45 s for 40 cycles, and then 72°C for 10 min. Samples were visualised on 1% agarose gels and successfully amplified samples were sent for forward and reverse Sanger sequencing to Macrogen Europe or to Edinburgh Genomics, where they were run on an ABI3730XL sequencer (Applied Biosystems).

2.3 | Sequencing

For each returned sequence, the sequences were quality trimmed (with 25 bp window segments where more than 2 bp showing a quality value of <20 were removed), the primer sequences were removed, and the contigs then assembled. Each contig was manually checked for base call disagreements and manually edited as needed. Low quality sequences were removed, and the two coding regions (*rbcl* and *matK*), were also checked for stop codons. All sequence assembly and editing was completed using Sequencher v 5.0 (GeneCodes Corp). Summary quality statistics were calculated for each marker, including the mean sequence length, the mean sequence QV, the mean sequence overlap between forward and reverse reads, the mean percentage of high-quality bases (QV>30) per sequence and the mean percentage of low-quality bases (QV<20) per sequence.

Sequence identification quality control included comparing sequences from multiple individuals of species and creating neighbour-joining trees. Any species that were misplaced within the tree (i.e. not with other accessions of the same species or not close to related taxa) were investigated to verify their identity. Sequences were also checked against available records on the NCBI database GenBank using BLAST.

The reference library, for all three loci, was deposited in the Barcode of Life Database (BOLD) (accessions: FPUK001-14 to FPUK1362-20; POWNA001-10 to POWNA3220-13; POWNB001-10 to POWNB237-10) and GenBank (accessions: JN890545–JN896265; KX165423–KX167996; MK924423–MK926404). Each sequence in the BOLD database is available with quality statistics for the

TABLE 1 Summary statistics for the DNA barcode database for the UK flora of 1,482 species.

| | <i>rbcL</i> | <i>matK</i> | ITS2 | All markers |
|---|-------------|-------------|------------|-------------|
| Number of species successfully DNA barcoded out of 1,482 | 1418 (96%) | 1334 (90%) | 1105 (75%) | 1035 (70%) |
| Species with more than one individual DNA barcoded out of 1,482 | 1312 (89%) | 1042 (70%) | 786 (53%) | 662 (45%) |
| Mean (SD) number of DNA barcodes per species | 3.0 (1.5) | 2.2 (1.5) | 1.7 (1.6) | 7 (3.8) |
| Total number of DNA barcodes from 6,100 specimens sampled | 4477 | 3259 | 2585 | 10,321 |

sequences in addition to the collection information, including location, collector, date collected, and a scan of the herbarium voucher.

2.4 | Species recoverability

The ability to successfully retrieve a sequence from a species was summarised overall and by plant order. To assess the effect of both year of sample collection and plant order on the successful recovery of a sequence, a binomial generalised linear model was fit with the proportion of successfully recovered sequences as the response variable. This analysis was restricted to plant orders with ten or more species sampled. The effect of year of sample collection, plant order and the interaction between the two were included as explanatory variables. Each marker was fitted as a separate model. Model selection was based on the lowest Akaike information criterion (AIC) score.

The success of sequence recoverability was also examined for the herbarium material separately. Year of collection was divided into nine classes for specimens from 1912 to 2010. Specimens from either side of the range were excluded, due to the small sample size in these age classes. The relationship between year of collection and sequence recovery was assessed with Spearman's rank correlation for each marker. All analysis was completed in R v. 3.5.2 (R Development Core Team, 2011).

2.5 | Species discrimination

The ability of the DNA barcode markers to discriminate to different levels of taxonomic identification was evaluated using BLAST searches that queried each sequence in turn against the database. For the Welsh plant flora, similar results in the discrimination ability were found between using BLAST, barcode gaps or monophyletic groups within neighbour-joining trees (de Vere et al., 2012). As BLAST identification is a common method in the application of DNA barcode reference libraries when assigning taxonomic information to unknown sequences in DNA metabarcoding (Deiner et al., 2017), the discrimination ability of the UK native reference library was assessed using BLAST. To allow comparison between the three markers, the BLAST database was restricted to the plant species that had multiple sequences for all three markers, giving 634 species. Each sequence was matched against a database that excluded the query sequence, and the level of discrimination assessed for a species, genus, family or order level match.

2.6 | Intra- and interspecific genetic distances

Alignments were created for the three markers using the R package DECIPHER (Wright, 2016). The function *alignseqs* was used for *rbcL* and ITS2. For ITS2, alignments were completed separately for each family with short sequences excluded, estimated as sequences which were below 400 bp in length. For *matK*, *aligntranslation* was used to align the amino acid translation of the DNA sequences and back-translate to base pairs. The uncorrected intra- and interspecific genetic distances were calculated using the function *distancematrix* for each pairwise comparison. For ITS2, pairwise comparisons were completed within the family alignment. A barcode gap, where the minimum interspecific genetic distance is greater than the maximum intraspecific genetic distance, was assessed for sequences that had more than one specimen per species. The percentage of species within each genus that showed a barcode gap was then calculated.

3 | RESULTS

3.1 | Recoverability within the UK native species

For the 1,482 native and archaeophyte flowering plants and conifers of the UK, a total of 10,321 barcode sequences were recovered with the *rbcL*, *matK* and ITS2 primers (Table 1). This represented 4,477 sequences for *rbcL* covering 96% of species, 97% of genera and 99% of families (Figure 1a). For *matK*, 3,259 sequences were recovered across 90% of species, 93% of genera and 93% of families, while 2,585 ITS2 sequences, representing 75% of species, 79% of genera and 82% of families were recovered (Figure 1a). All three markers were obtained for 1,035 species (representing 70% of the total UK flora), while 98% of plant species, 97% of genera and 100% of families were represented with at least one marker (Table 1).

When looking at sequence success at the family level, a sequence recovery of above 50% of specimens was seen for 85% of plant families for *rbcL*, 54% of families for *matK*, and 32% of families for ITS2 (Figure 1a). The family success level for *matK* was still low despite the use of multiple primer combinations. Malvaceae was the most consistently successful family with over 80% sequence recovery across all three markers. Looking at the families with ten or more specimens attempted, the *rbcL* locus was recovered from all families. ITS2 was not recovered from four plant families (Araceae, Zosteraceae, Iridaceae and Valerianaceae), three of which are monocots. The *matK* locus was not recovered from three of the plant families with more than ten specimens (Elatinaceae, Hypericaceae and Polygalaceae).

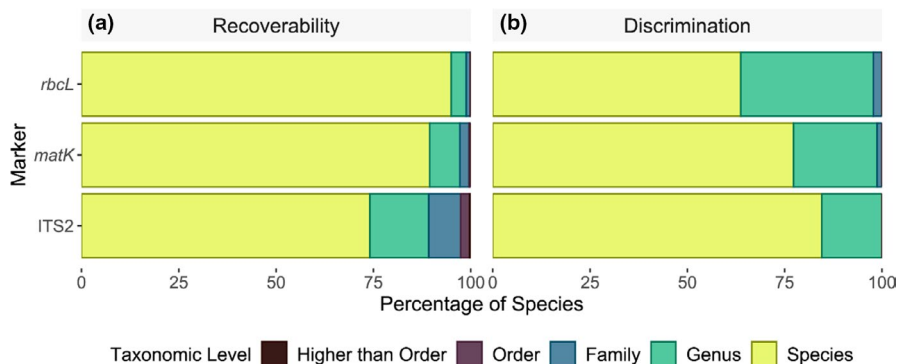
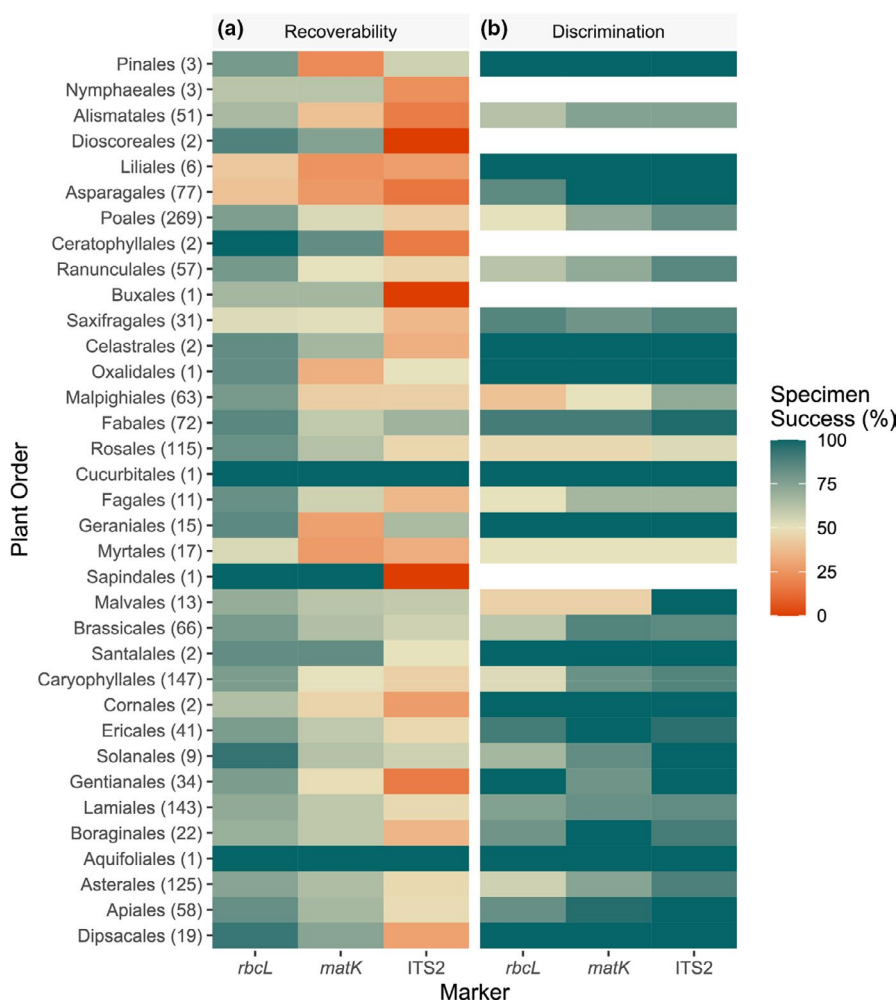


FIGURE 1 The overall level of taxonomic representation in the database for the native species of the UK flora. Recoverability (a) shows the level of representation for the native species of the UK flora ($n = 1,482$) in the reference library. Discrimination (b) shows the taxonomic resolution achieved using BLAST for those plant species in the reference library that were represented by all three markers more than once ($n = 634$)

FIGURE 2 Species level recoverability (a) and discrimination (b) by plant order for each marker. Recoverability shows the percentage of specimens in each plant order that were successfully sequenced. The discrimination level was assessed for plant species that were represented by all three markers more than once, showing the percentage of species in each plant order that were identified to species level. Number of species represented by an order is shown in brackets. The mode number of specimens per species was three



Recoverability was tested at the order level for plant orders with 10 or more species (Figure 2a, Figure S1). For *rbcL* both the year of sample collection (Likelihood ratio; $LR_{1,20} = 838.90, p < .001$) and plant order ($LR_{1,20} = 330.22, p < .001$) were found to significantly predict the success of sequence recovery, with a significant interaction also found between plant order and year ($LR_{1,20} = 80.92, p < .001$), suggesting that plant orders have different patterns of

DNA loss or degradation in herbarium specimens (Figure S1). The same pattern was seen for both *matK*, with year ($LR_{1,20} = 1075.63, p < .001$), plant order ($LR_{1,20} = 296.23, p < .001$) and their interaction ($LR_{1,20} = 96.67, p < .001$), and ITS2, with year ($LR_{1,20} = 577.24, p < .001$), plant order ($LR_{1,20} = 381.82, p < .001$) and their interaction ($LR_{1,20} = 82.79, p < .001$). For *rbcL* three plant orders showed a significant correlation between sequence recovery and year class after

Bonferroni correction for multiple testing; six orders with *matK* and eight orders for ITS2 (Figure S1, full correlation results detailed in Table S4).

3.2 | Discrimination within the UK native species

Discrimination ability was assessed for 634 of the sequenced species that had multiple individuals sequenced for all markers (Figure 1b). For *rbcl*, 64% of species were returned as a species level match, 98% were a genus level match and 100% were a family level match. For *matK* there was greater species discrimination with 77% of species returned at species level, 99% at genus and 100% at family. ITS2 had the highest level of species discrimination with 85% to species and 100% returned to genus.

The increased discrimination ability of ITS2 was also reflected in the relationship between the minimum interspecific genetic distance and the maximum intraspecific genetic distance (Figure 3). A barcode gap, where the minimum interspecific distance is greater than the maximum intraspecific distance, was found for 50% of *rbcl* sequences, 63% of *matK* sequences and 76% of ITS2 sequences.

There were some similar taxonomic patterns across the three markers for discrimination ability, with reduced species level discrimination in the species-rich Rosales: *rbcl* and *matK* distinguish 47% of species to species level, and ITS2 distinguishes 52% (Figure 2b). The large apomictic genus *Sorbus* (36 species) within Rosales, showed limited species discrimination across all three markers (Figure 4). Discrimination was also poor across all three markers within the Myrtales order in distinguishing *Epilobium* species. ITS2 showed increased discrimination ability compared to *matK* and *rbcl* in Malpighiales and Malvales. With *rbcl* the lowest species level discrimination achieved in a plant order was 39% (Malpighiales), while in *matK* the lowest was 44% (Malvales), and for ITS2 it was 50% (Myrtales).

At the genus level, the proportion of species showing a barcode gap differed (Figure 4). For *matK* and ITS2 a negative correlation between genus size and the proportion of species with a barcode gap was found (Spearman's rank correlation coefficient, *matK*: $r_s = -0.200$, $p = .002$; ITS2: $r_s = -0.33$, $p < .001$). No correlation was found for *rbcl* ($r_s = -0.05$, $p = .405$). While ITS2 showed an increased number of barcode gaps for some genera (e.g., *Carex*) compared with *rbcl* and ITS2, certain genera remain problematic in identification across all three markers, e.g., *Euphrasia* and *Sorbus*.

3.3 | Comparison between recoverability from herbarium and fresh material

In total, 6,100 accessions were sampled, 4,965 from herbarium specimens and 1,135 from recent silica-gel dried tissue samples (Table 2). For all three markers, freshly collected leaf material was significantly more likely to yield a successful DNA barcode than herbarium

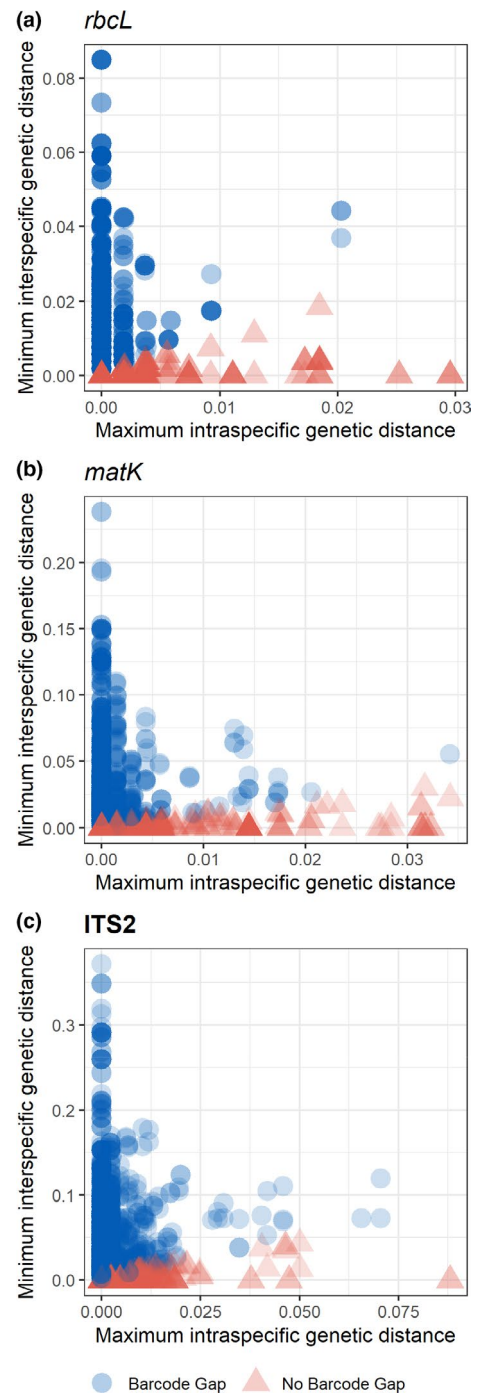
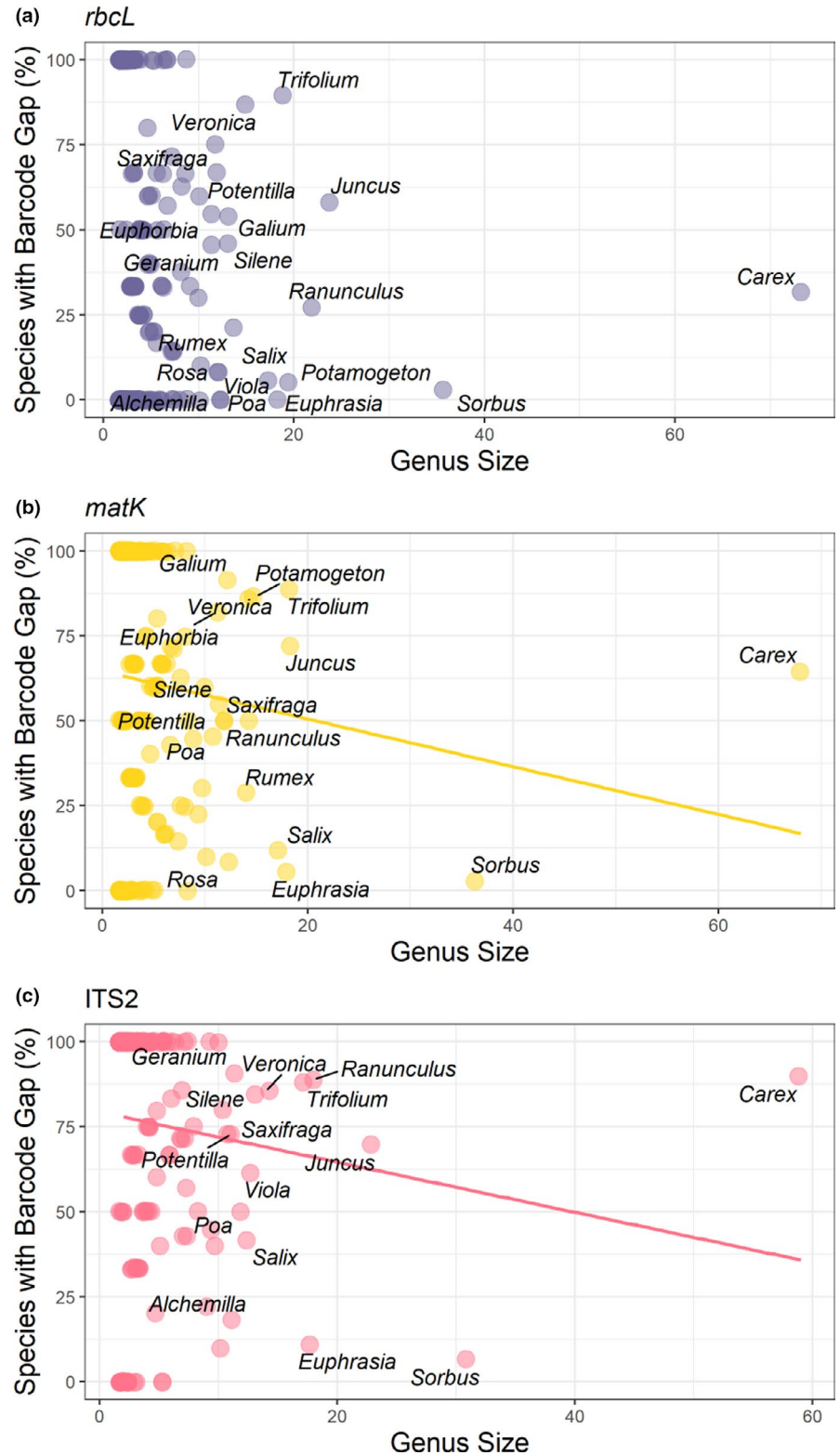


FIGURE 3 The relationship between the maximum intraspecific genetic distance and minimum interspecific distance for each specimen and locus in the UK flora reference library, using pairwise comparisons calculated from the multiple alignment of each locus

material. For *rbcl*, overall 73% of specimens yielded a sequence, with 88% success from fresh samples and 70% from herbarium samples (chi-squared test, with Yates correction; $\chi^2 = 145.45$, $d.f = 1$, $p < .001$). Lower sequence recoverability was found for *matK*, with 53% of specimens working overall, 74% for fresh material and 49% for herbarium ($\chi^2 = 236.26$, $d.f = 1$, $p < .001$). ITS2 showed the lowest overall recoverability at 42% of samples, 65% for fresh material and

FIGURE 4 The percentage of species with a barcode gap per genus compared with genus size. Genera with more than one species are plotted. Genera with more than 10 species are labelled. A significant correlation was found for *matK* (Spearman's: $r_s = -0.200$, $p = .002$) and ITS2 ($r_s = -0.33$, $p < .001$). No correlation was found for *rbcL* ($r_s = -0.05$, $p = .405$)



37% for herbarium specimens ($\chi^2 = 288.34$, $d.f = 1$, $p < .001$). Looking across the sequence quality statistics, overall the quality and consistency of returned sequences was improved within fresh samples compared to herbarium (Table 2, Table S5).

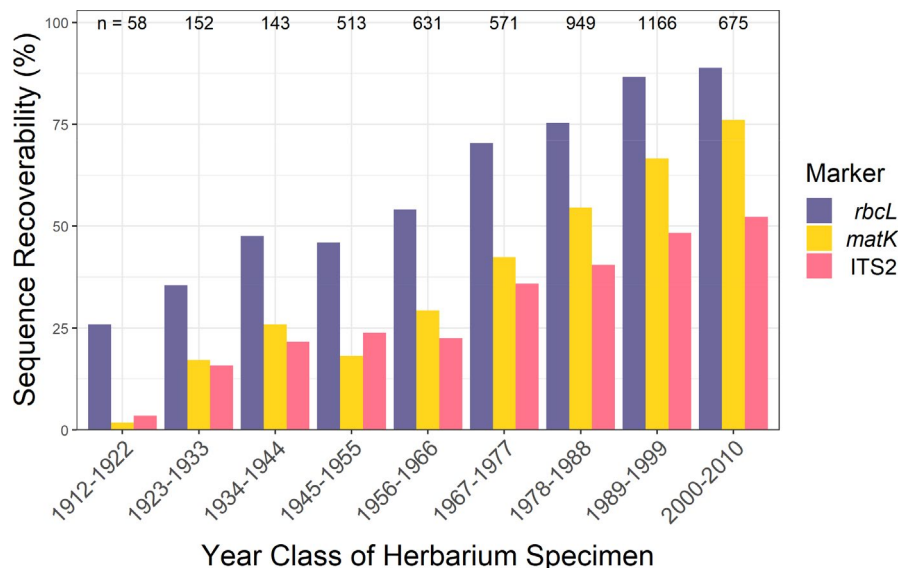
Examining the success of herbarium specimens alone, a significant correlation between the success of sequence recovery and

the age of the herbarium specimen was found, with greater success of sequencing for more recently collected specimens (Figure 5; Spearman's rank correlation coefficient for *rbcL* $r_s = 0.982$, $p < .001$; for *matK*, $r_s = 0.927$, $p < .001$; for ITS2, $r_s = 0.890$, $p < .001$). The collection year of successfully sequenced herbarium specimens ranged from 1868 to 2011.

TABLE 2 The sequence quality for the DNA barcodes retrieved from fresh and herbarium material for *rbcl*, *matK* and ITS2. Overall, freshly collected leaf material performed better than herbarium material, in terms of returned sequence quality metrics (Table S5).

| | <i>rbcl</i> | | | <i>matK</i> | | | ITS2 | | |
|--|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|
| | Overall | Fresh | Herbarium | Overall | Fresh | Herbarium | Overall | Fresh | Herbarium |
| Number of samples collected | 6100 | 1135 | 4965 | 6100 | 1135 | 4965 | 6100 | 1135 | 4965 |
| Number successfully barcoded | 4477 (73%) | 994 (88%) | 3483 (70%) | 3259 (53%) | 845 (74%) | 2414 (49%) | 2585 (42%) | 738 (65%) | 1847 (37%) |
| Mean sequence length (SD) | 573 bp (25.04) | 578 bp (41.17) | 572 bp (17.66) | 843 bp (62.83) | 833 bp (43.79) | 847 bp (68.35) | 513 bp (97.91) | 515 bp (92.06) | 512 bp (99.94) |
| Mean sequence QV (SD) | 56 (4.48) | 57 (3.06) | 56 (4.78) | 50 (4.86) | 51 (4.29) | 50 (5.02) | 53 (5.23) | 54 (4.77) | 52 (5.32) |
| Mean (SD) sequence overlap | 472 (29.76) | 481 (18.95) | 469 (31.64) | 719 (123.49) | 742 (93.7) | 710 (132.13) | 388 (32.55) | 393 (25.97) | 386 (34.42) |
| Mean (SD) of high-quality bases: QV>30 (%) | 97% (3.32) | 97% (2.64) | 97% (3.47) | 91% (6.53) | 92% (5.77) | 91% (6.77) | 94% (5.39) | 95% (4.91) | 94% (5.55) |
| Mean (SD) of low-quality bases: QV<20 (%) | 1% (1.01) | 1% (0.95) | 1% (1.02) | 2% (2.03) | 2% (1.67) | 2% (2.14) | 2% (2.39) | 2% (2.11) | 2% (2.48) |

FIGURE 5 Effect of herbarium specimen age on sequence recoverability. The success of 4,858 herbarium samples was assessed over nine-year classes, between 1912–2010. Older specimens were not included due to the lower sample sizes. The sample size of each age class is annotated above the bar. There was a significant correlation for all three markers, between success of sequence recoverability and age of the herbarium specimen (Spearman's rank correlation coefficient for *rbcl*, $r_s = 0.982$, $p < .001$; for *matK*, $r_s = 0.927$, $p < .001$; for ITS2, $r_s = 0.890$, $p < .001$)



4 | DISCUSSION

DNA barcoding is increasingly deployed on large sample sets and to develop large-scale reference libraries. However, there are still very few national DNA barcode databases for plants. Here, we present a comprehensive and high-quality DNA barcode reference library for the native UK flora, for three loci, complete with metadata and herbarium voucher images for each specimen. This provides a foundational resource for species identification and will enable future large-scale analyses of evolutionary and ecological processes. We discuss the development of this resource, and the results of our sequence analyses, in terms of producing DNA barcode data sets for other floras, the use of herbarium specimens in large-scale DNA barcoding efforts, and the utility of DNA barcoding for species identification within the flora.

4.1 | Developing national DNA barcoding resources

Complete DNA barcoding data sets for regions or countries are scarce—with examples to date mostly limited taxonomically, to confined geographic areas or small countries (e.g., research centres in East African savanna (Gill et al., 2019), and the arctic (Wirta et al., 2016)). The most comparable effort to ours is the DNA barcode data set for the vascular plants of Canada, where 5,190 plant species were targeted with *rbcl*, *matK* and ITS2 (Braukmann et al., 2017; Kuzmina et al., 2017). We achieved similar species coverage across the markers to the Canadian flora for *rbcl* and ITS2, while our species coverage of *matK* is higher (89% compared with 41%), although more primer combinations were attempted. The PhyloAlps and PhyloNorway projects represent another large scale floristic barcoding effort, using genome skimming to assemble whole chloroplast data and retrieve barcode regions (Alsos et al., 2020). The successful retrieval of *rbcl*, *matK* and ITS2 was shown to be higher on a per specimen basis with genome skimming; however, the increased cost needs to be evaluated by a project's requirements.

Our complete DNA barcoding resource for the British flora enables the large community of researchers studying the British flora to place their research in a genetic context. For example, ecologists interested in community structure can incorporate phylogenetic relatedness from DNA sequencing information in their models (Gill et al., 2019; Heckenhauer et al., 2017; Lim et al., 2014). Complete country-level DNA barcode databases will also facilitate sample verification in large-scale genome sequencing projects. For example, the Darwin Tree of Life (DTOL) project aims to sequence the complete genomes of all British native eukaryotic organisms, including all flowering plant species. DTOL samples can now have their DNA barcode sequence checked against our reference database to ensure correct sample identification before the considerable cost of genome sequencing is incurred.

4.2 | Use of herbarium material for generating reference DNA barcoding libraries

Herbarium material represents a readily available and well-curved resource for plant DNA barcoding efforts. However, degraded museum DNA comes with its own set of challenges for DNA sequencing. In our study we found herbarium material performed worse in terms of successful sequence recovery, across all three markers. When looking at the success of herbarium material alone, the sequence recovery was strongly related to the age of the specimen, with more recently collected specimens having increased success. Similar patterns of effect of specimen age were seen during the Canadian plant barcoding campaign from herbarium samples (Kuzmina et al., 2017); however, they found ITS2 was less affected by specimen age, which they theorised was due to its shorter length (~350 bp); similar results were obtained by Särkinen et al., (2012). That pattern was not observed here, with ITS2 performing poorly with herbarium material compared to freshly collected material.

While the use of herbarium material can increase the lab and processing time involved in gaining a DNA barcode, this increase

is mitigated by the relative effort involved in collecting, identifying and processing new plant specimens to create high-quality DNA barcodes. However, certain considerations are required when using herbarium material to maximise success, given that it is destructive sampling that requires the permission of the holding institution. As evidenced here, sample collection from herbaria should focus on younger specimens. In addition, quality control checks should monitor for potential contamination at the specimen sampling stage for both herbarium and recent samples, including the presence of algae on the plants, which can be difficult to detect when collecting (de Vere et al., 2012). One possible way to minimise algal contamination for aquatic plants would be to selectively sample aerial flowering parts of the plants that will have reduced algae.

Herbaria around the world have been recognised as a potential source of efficiently capturing the accumulated taxonomic expertise they house (Dormontt et al., 2018; Kuzmina et al., 2017; Xu et al., 2015), as well as answering a broad range of questions beyond core species identification and taxonomy (James et al., 2018). With the decreasing costs of genome skimming approaches, which by-pass the need for amplicon based methods that are reliant on relatively undegraded DNA, the use of herbarium collections for DNA barcoding is likely to continue to grow (Alsos et al., 2020).

By including voucher information with the DNA sequence, taxonomic updates to the sampled species can continue to be incorporated into the reference library. Going forward, collections can incorporate data that will benefit research, with samples stored with future DNA analysis in mind (Heberling & Isaac, 2017). This involves the creation of a "secondary voucher", one intended for destructive analysis (Kageyama et al., 2007). For DNA analysis this often consists of leaf material stored in silica gel, vouchered by the primary traditional herbarium specimen; this material can potentially be used for other destructive analyses, e.g., analysis of secondary plant compounds.

4.3 | Utility of DNA barcoding across the UK flora

The relative levels of recovery and discrimination seen in *rbcl*, *matK* and ITS2 show the need for a balance between taxonomic universality and the level of species discrimination gained. For the UK flora, ITS2 performed poorly compared to *rbcl* and *matK* in its species recovery, suggesting that the primers used are less universal than those for the other markers (Chen et al., 2010; Li et al., 2011). This needs to be balanced however by the fact that *matK* was attempted with multiple primer pairs, whilst just one primer pair was used for ITS2. For ITS2, attempting additional primer combinations may have improved recoverability, but could also have increased the risk of amplifying different copies within this multicopy marker. In contrast, previous work by de Vere et al., (2012) showed that one pair of primers for *rbcl* was highly effective for species recovery across a broad taxonomic range and that the addition of further primer combinations did not increase overall recoverability. Overall, all three of the markers performed well with the UK flora when looking at the ability

to discriminate to at least genus level (98%–100%), reflecting similar patterns with the Canadian flora (>90% across *rbcl*, *matK*, and ITS2) (Braukmann et al., 2017).

Analysing DNA barcoding data within and between species across the British flora reveals a gradient of genetic divergence, from those genera where species have a barcoding gap and are easy to tell apart, to those where no single marker or multimer combination provides species-level resolution. Examples of taxonomically complex genera where species level discrimination remained low across markers include *Euphrasia* and *Sorbus*. DNA barcoding has been shown to have limited utility in species identification within these and other taxonomically complex genera, arising from processes including apomixis, recent species divergence and past hybridisation events (Twyford, 2014; Wang et al., 2018). Groups which have poor overall sequence recovery can also appear to have higher levels of discrimination because of missing conspecifics. While plant species are, in general, inherently more difficult to distinguish with DNA barcoding data than animals (Fazekas et al., 2009), at least in the relatively species depauperate British flora, DNA barcoding provides a useful tool for identification in most flowering plant groups.

With DNA metabarcoding applications, the mixed source samples used will often provide challenges with low amounts of DNA, shorter fragment lengths, and poorer quality template, such as extractions from soil, honey (Hawkins et al., 2015; Jones et al., 2021), or faecal samples (Moorhouse-Gann et al., 2018). The longer length of *matK* combined with the number of primer combinations required to gain taxonomic coverage makes it generally unsuitable for amplicon-based metabarcoding, where the sequencing length is often limited by both the sequencing platform and the degraded quality of the source DNA. Both *rbcl* and ITS2 can be used for DNA metabarcoding applications, using primer pairs for *rbcl* that produce a shorter amplicon (Table S6, de Vere et al., 2017). However, care needs to be taken when using ITS2 as the sole marker in DNA metabarcoding studies that are surveying a potentially wide range of species. While ITS2 can give increased species discrimination, it may not be representing the target community fully due its lower universality. The higher length variability with ITS2 compared with comparatively fixed-length amplicons is another source of potential sequencing bias on length restricted sequencing platforms.

Overall, the UK flora as a national reference library may allow for higher levels of taxonomic discrimination due to species poor groups. For DNA metabarcoding studies at a finer geographic scale, the level of species discrimination achieved can be improved further by reducing the set of locally occurring species and thereby reducing the complexity of the identification challenge (de Vere et al., 2012).

4.4 | Conclusions

The DNA barcode reference library presented here represents a high-quality database that is publicly available and able to facilitate wide-ranging applications that require plant identification as well as

providing a robust resource for continuing phylogenetic analyses. We have demonstrated the effective use of herbarium collections for retrieving their “stored” taxonomic expertise to rapidly build a robust DNA barcode library.

ACKNOWLEDGEMENTS

We thank J. Moughan, A. Griffith, A. Lowe, E. Chapman, C. Long, S. Trinder, C.W. Moore, E. Brittain, D. Satterthwaite, and A. Sweeney for their help collecting and processing specimens. Thank you to the vice-county recorders of the Botanical Society of Britain and Ireland for plant record data.

Natasha de Vere has received funding through the Welsh Government Rural Communities – Rural Development Programme 2014–2020, which is funded by the European Agricultural Fund for Rural Development and the Welsh Government. Laura Jones and Natasha de Vere have received funding through the Welsh Government’s Enabling Natural Resources and Well-being in Wales Grants (ENRaW).

AUTHOR CONTRIBUTIONS

Natasha de Vere, Laura Jones, Tim C.G. Rich and Peter M. Hollingsworth conceived and designed the experiments. Natasha de Vere, Laura Jones, Helena Davies, Tim C.G. Rich, Laura L. Forrest, Michelle L. Hart and Heather McHaffie completed the fieldwork and labwork. Laura Jones, Natasha de Vere, Col R. Ford, Helena Davies, Laura L. Forrest, Michelle L. Hart analysed the data. Laura Jones, Natasha de Vere, Alex D. Twyford, Max R. Brown, Laura L. Forrest, Peter M. Hollingsworth wrote the manuscript. All authors contributed to the final submitted manuscript.

DATA AVAILABILITY STATEMENT

Specimen data and DNA barcodes: BOLD and Genbank accessions are listed with specimen metadata in Supporting Information. BOLD accessions: FPUK001-14 to FPUK1362-20; POWNA001-10 to POWNA3220-13; POWNB001-10 to POWNB237-10. GenBank accessions: JN890545–JN896265; KX165423–KX167996; MK924423–MK926404.

ORCID

Alex D. Twyford  <https://orcid.org/0000-0002-8746-6617>

Col R. Ford  <https://orcid.org/0000-0002-5881-4707>

Natasha de Vere  <https://orcid.org/0000-0001-9593-6925>

REFERENCES

Alsos, I. G., Lavergne, S., Merkel, M. K. F., Boleda, M., Lammers, Y., Alberti, A., Pouchon, C., Denoed, F., Pitelkova, I., Puşcaş, M., Roquet, C., Hurdu, B.-I., Thuiller, W., Zimmermann, N. E., Hollingsworth, P. M., & Coissac, E. (2020). The treasure vault can be opened: Large-scale genome skimming works well using herbarium and silica gel dried material. *Plants*, 9(4), 432. <https://doi.org/10.3390/plants9040432>

Arrizabalaga-Escudero, A., Merckx, T., Garcia-Baqueró, G., Wahlberg, N., Aizpurua, O., Garin, I., Goiti, U., & Aihartza, J. (2019). Trait-based functional dietary analysis provides a better insight into the foraging ecology of bats. *Journal of Animal Ecology*, 88(10), 1587–1600. <https://doi.org/10.1111/1365-2656.13055>

Bell, K. L., de Vere, N., Keller, A., Richardson, R. T., Gous, A., Burgess, K. S., & Brosi, B. J. (2016). Pollen DNA barcoding: current applications and future prospects. *Genome*, 59(9), 629–640. <https://doi.org/10.1139/gen-2015-0200>

Blagoev, G. A., deWaard, J. R., Ratnasingham, S., deWaard, S. L., Lu, L., Robertson, J., & Hebert, P. D. N. (2016). Untangling taxonomy: A DNA barcode reference library for Canadian spiders. *Molecular Ecology Resources*, 16(1), 325–341. <https://doi.org/10.1111/1755-0998.12444>

Braukmann, T. W. A., Kuzmina, M. L., Sills, J., Zakharov, E. V., & Hebert, P. D. N. (2017). Testing the efficacy of DNA barcodes for identifying the vascular plants of Canada. *PLoS One*, 12(1), 1–19. <https://doi.org/10.1371/journal.pone.0169515>

Brennan, G. L., Potter, C., de Vere, N., Griffith, G. W., Skjøth, C. A., Osborne, N. J., Wheeler, B. W., McInnes, R. N., Clewlow, Y., Barber, A., Hanlon, H. M., Hegarty, M., Jones, L., Kurganskiy, A., Rowney, F. M., Armitage, C., Adams-Groom, B., Ford, C. R., Petch, G. M., & Creer, S. (2019). Temperate airborne grass pollen defined by spatio-temporal shifts in community composition. *Nature Ecology & Evolution*, 3(5), 750–754. <https://doi.org/10.1038/s41559-019-0849-7>

CBOL Plant Working Group (2009). A DNA barcode for land plants. *Proceedings of the National Academy of Sciences of the United States of America*, 106(31), 12794–12797. <https://doi.org/10.1073/pnas.0905845106>

Chen, S., Yao, H., Han, J., Liu, C., Song, J., Shi, L., Zhu, Y., Ma, X., Gao, T., Pang, X., Luo, K., Li, Y., Li, X., Jia, X., Lin, Y., & Leon, C. (2010). Validation of the ITS2 region as a novel DNA barcode for identifying medicinal plant species. *PLoS One*, 5(1), 1–8. <https://doi.org/10.1371/journal.pone.0008613>

Clubbe, C., Ainsworth, A. M., Bárrios, S., Bensusan, K., Brodie, J., Cannon, P., Chapman, T., Copeland, A. I., Corcoran, M., Dani Sanchez, M., David, J. C., Dines, T., Gardiner, L. M., Hamilton, M. A., Heller, T., Hollingsworth, P. M., Hutchinson, N., Llewelyn, T., Lowe Forrest, L., ... Fay, M. F. (2020). Current knowledge, status, and future for plant and fungal diversity in Great Britain and the UK Overseas Territories. *Plants, People, Planet*, 2(5), 557–579. <https://doi.org/10.1002/ppp3.10142>

de Boer, H. J., Ghorbani, A., Manzanilla, V., Raclariu, A.-C., Kreziou, A., Ounjai, S., Osathanunkul, M., & Gravendeel, B. (2017). DNA metabarcoding of orchid-derived products reveals widespread illegal orchid trade. *Proceedings of the Royal Society B: Biological Sciences*, 284(1863), 20171182. <https://doi.org/10.1098/rspb.2017.1182>

de Vere, N., Jones, L. E., Gilmore, T., Moscrop, J., Lowe, A., Smith, D., Hegarty, M. J., Creer, S., & Ford, C. R. (2017). Using DNA metabarcoding to investigate honey bee foraging reveals limited flower use despite high floral availability. *Scientific Reports*, 7(1), 42838. <https://doi.org/10.1038/srep42838>

de Vere, N., Rich, T. C. G., Ford, C. R., Trinder, S. A., Long, C., Moore, C. W., Satterthwaite, D., Davies, H., Allainguillaume, J., Ronca, S., Tatarinova, T., Garbett, H., Walker, K., & Wilkinson, M. J. (2012). DNA Barcoding the Native Flowering Plants and Conifers of Wales. *PLoS One*, 7(6), e37945. <https://doi.org/10.1371/journal.pone.0037945>

Deiner, K., Bik, H. M., Mächler, E., Seymour, M., Lacoursière-Roussel, A., Altermatt, F., & Bernatchez, L. (2017). Environmental DNA metabarcoding: Transforming how we survey animal and plant communities. *Molecular Ecology*, 26(21), 5872–5895. <https://doi.org/10.1111/mec.14350>

Dormontt, E. E., van Dijk, K.-J., Bell, K. L., Biffin, E. D., Breed, M. F., Byrne, M., Caddy-Retalic, S., Encinas-Viso, F., Nevill, P. G., Shapcott, A., Young, J. M., Waycott, M., & Lowe, A. J. (2018). Advancing DNA Barcoding and Metabarcoding Applications for Plants Requires Systematic Analysis of Herbarium Collections—An Australian Perspective. *Frontiers in Ecology and Evolution*, 6, 1–12. <https://doi.org/10.3389/fevo.2018.00134>

- Edwards, M. E., Gielly, L., Langdon, C. T., Croudace, I. W., Kristine, M., Merkel, F., & Alsos, I. G. (2017). Lake sedimentary DNA accurately records 20th Century introductions of exotic conifers in Scotland. *New Phytologist*, 213, 929–941. <https://doi.org/10.1111/nph.14199>.
- Ellstrand, N. C., Whitkus, R., & Rieseberg, L. H. (1996). Distribution of spontaneous plant hybrids. *Proceedings of the National Academy of Sciences of the United States of America*, 93(10), 5090–5093.
- Fazekas, A. J., Kesanakurti, P. R., Burgess, K. S., Percy, D. M., Graham, S. W., Barrett, S. C. H., Newmaster, S. G., Hajibabaei, M., & Husband, B. C. (2009). Are plant species inherently harder to discriminate than animal species using DNA barcoding markers? *Molecular Ecology Resources*, 9(Suppl. 1), 130–139. <https://doi.org/10.1111/j.1755-0998.2009.02652.x>.
- Gill, B. A., Musili, P. M., Kurukura, S., Hassan, A. A., Goheen, J. R., Kress, W. J., Kuzmina, M., Pringle, R. M., & Kartzinel, T. R. (2019). Plant DNA-barcode library and community phylogeny for a semi-arid East African savanna. *Molecular Ecology Resources*, 19(4), 838–846. <https://doi.org/10.1111/1755-0998.13001>.
- Harper, L. R., Lawson Handley, L., Hahn, C., Boonham, N., Rees, H. C., Gough, K. C., & Hänfling, B. (2018). Needle in a haystack? A comparison of eDNA metabarcoding and targeted qPCR for detection of the great crested newt (*Triturus cristatus*). *Ecology and Evolution*, 8(12), 6330–6341. <https://doi.org/10.1002/ece3.4013>.
- Hawkins, J., de Vere, N., Griffith, A., Ford, C. R., Allainguillaume, J., Hegarty, M. J., Baillie, L., & Adams-Groom, B. (2015). Using DNA Metabarcoding to Identify the Floral Composition of Honey: A New Tool for Investigating Honey Bee Foraging Preferences. *PLoS One*, 10(8), e0134735. <https://doi.org/10.1371/journal.pone.0134735>.
- Heberling, J. M., & Isaac, B. L. (2017). Herbarium specimens as exaptations: New uses for old collections. *American Journal of Botany*, 104(7), 963–965. <https://doi.org/10.3732/ajb.1700125>.
- Hebert, P. D. N., Cywinska, A., Ball, S. L., & DeWaard, J. R. (2003). Biological identifications through DNA barcodes. *Proceedings of the Royal Society B: Biological Sciences*, 270(1512), 313–321. <https://doi.org/10.1098/rspb.2002.2218>.
- Hebert, P. D. N., Hollingsworth, P. M., & Hajibabaei, M. (2016). From writing to reading the encyclopedia of life. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 371(1702), 20150321. <https://doi.org/10.1098/rstb.2015.0321>.
- Heckenhauer, J., Abu Salim, K., Chase, M. W., Dexter, K. G., Pennington, R. T., Tan, S., Kaye, M. E., & Samuel, R. (2017). Plant DNA barcodes and assessment of phylogenetic community structure of a tropical mixed dipterocarp forest in Brunei Darussalam (Borneo). *PLoS One*, 12(10), 1–24. <https://doi.org/10.1371/journal.pone.0185861>.
- Hill, M. O., Preston, C. D., & Roy, D. B. (2004). PLANTATT - Attributes of British and Irish Plants: Status, Size, Life history, Geography and Habitats. Centre for Ecology and Hydrology, 80.
- Hollingsworth, P. M., Li, D., van der Bank, M., & Twyford, A. D. (2016). Telling plant species apart with DNA: from barcodes to genomes. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 371(1702), 20150338. <https://doi.org/10.1098/rstb.2015.0338>.
- James, S. A., Soltis, P. S., Belbin, L., Chapman, A. D., Nelson, G., Paul, D. L., & Collins, M. (2018). Herbarium data: Global biodiversity and societal botanical needs for novel research: Global. *Applications in Plant Sciences*, 6(2), 1–8. <https://doi.org/10.1002/aps3.1024>.
- Jones, L., Brennan, G. L., Lowe, A., Creer, S., Ford, C. R., & de Vere, N. (2021). Shifts in honeybee foraging reveal historical changes in floral resources. *Communications Biology*, 4(1), 37. <https://doi.org/10.1038/s42003-020-01562-4>.
- Kageyama, M., Monk, R. R., Bradley, R. D., Edson, G. F., & Baker, R. J. (2007). The changing significance and definition of the biological voucher. In *Museum Studies: Perspectives and Innovations* (pp. 257–264). Society for the Preservation of Natural History Collections.
- Kartzinel, T. R., Chen, P. A., Coverdale, T. C., Erickson, D. L., Kress, W. J., Kuzmina, M. L., Rubenstein, D. I., Wang, W., & Pringle, R. M. (2015). DNA metabarcoding illuminates dietary niche partitioning by African large herbivores. *Proceedings of the National Academy of Sciences*, 112(26), 8019–8024. <https://doi.org/10.1073/pnas.1503283112>.
- Kress, W. J., & Erickson, D. L. (2007). A Two-Locus Global DNA Barcode for Land Plants: The Coding rbcL Gene Complements the Non-Coding trnH-psbA Spacer Region. *PLoS One*, 2(6), e508. <https://doi.org/10.1371/journal.pone.0000508>.
- Kuzmina, M. L., Braukmann, T. W. A., Fazekas, A. J., Graham, S. W., Dewaard, S. L., Rodrigues, A., Bennett, B. A., Dickinson, T. A., Saarela, J. M., Catling, P. M., Newmaster, S. G., Percy, D. M., Fenneman, E., Lauron-Moreau, A., Ford, B., Gillespie, L., Subramanyam, R., Whittton, J., Jennings, L., ... Hebert, P. D. N. (2017). Using Herbarium-Derived DNAs to Assemble a Large-Scale DNA Barcode Library for the Vascular Plants of Canada. *Applications in Plant Sciences*, 5(12), 1700079. <https://doi.org/10.3732/apps.1700079>.
- Landi, M., Dimech, M., Arculeo, M., Biondo, G., Martins, R., Carneiro, M., Carvalho, G. R., Brutto, S. L., & Costa, F. O. (2014). DNA barcoding for species assignment: The case of Mediterranean marine fishes. *PLoS One*, 9(9), e106135. <https://doi.org/10.1371/journal.pone.0106135>.
- Li, D.-Z., Gao, L.-M., Li, H.-T., Wang, H., Ge, X.-J., Liu, J.-Q., Chen, Z.-D., Zhou, S.-L., Chen, S.-L., Yang, J.-B., Fu, C.-X., Zeng, C.-X., Yan, H.-F., Zhu, Y.-J., Sun, Y.-S., Chen, S.-Y., Zhao, L., Wang, K., Yang, T., & Duan, G.-W. (2011). Comparative analysis of a large dataset indicates that internal transcribed spacer (ITS) should be incorporated into the core barcode for seed plants. *Proceedings of the National Academy of Sciences*, 108(49), 19641–19646. <https://doi.org/10.1073/pnas.1104551108>.
- Lim, J., Crawley, M. J., de Vere, N., Rich, T., & Savolainen, V. (2014). A phylogenetic analysis of the British flora sheds light on the evolutionary and ecological factors driving plant invasions. *Ecology and Evolution*, 4(22), 4258–4269. <https://doi.org/10.1002/ece3.1274>.
- Lucas, A., Bodger, O., Brosi, B. J., Ford, C. R., Forman, D. W., Greig, C., Hegarty, M., Jones, L., Neyland, P. J., & de Vere, N. (2018). Floral resource partitioning by individuals within generalised hoverfly pollination networks revealed by DNA metabarcoding. *Scientific Reports*, 8(1), 5133. <https://doi.org/10.1038/s41598-018-23103-0>
- Moorhouse-Gann, R. J., Dunn, J. C., de Vere, N., Goder, M., Cole, N., Hipperson, H., & Symondson, W. O. C. (2018). New universal ITS2 primers for high-resolution herbivory analyses using DNA metabarcoding in both tropical and temperate zones. *Scientific Reports*, 8(1), 1–15. <https://doi.org/10.1038/s41598-018-26648-2>.
- Pellicer, J., & Leitch, I. J. (2020). The Plant DNA C-values database (release 7.1): an updated online repository of plant genome size data for comparative studies. *New Phytologist*, 226(2), 301–305. <https://doi.org/10.1111/nph.16261>.
- Pentinsaari, M., Ratnasingham, S., Miller, S. E., & Hebert, P. D. N. (2020). BOLD and GenBank revisited - Do identification errors arise in the lab or in the sequence libraries? *PLoS One*, 15(4), 1–10. <https://doi.org/10.1371/journal.pone.0231814>.
- Preston, C. D., Pearman, D. A., & Dines, T. D. (2002). *New Atlas of the British and Irish Flora: An Atlas of the Vascular Plants of Britain, Ireland, The Isle of Man and the Channel Islands*. Oxford University Press.
- Prosser, S. W. J., & Hebert, P. D. N. (2017). Rapid identification of the botanical and entomological sources of honey using DNA metabarcoding. *Food Chemistry*, 214, 183–191. <https://doi.org/10.1016/j.foodchem.2016.07.077>.
- R Development Core Team (2011). *R: A Language and Environment for Statistical Computing*. In R Foundation for Statistical Computing. doi: <https://doi.org/10.1007/978-3-540-74686-7>
- Raclariu, A. C., Țebrencu, C. E., Ichim, M. C., Ciupercă, O. T., Brysting, A. K., & de Boer, H. (2018). What's in the box? Authentication of Echinacea herbal products using DNA metabarcoding and HPTLC. *Phytomedicine*, 44, 32–38. <https://doi.org/10.1016/j.phymed.2018.03.058>.

- Ruhsam, M., Squirrell, J., Gornall, R. J., French, G. C., Pullan, M., & Hollingsworth, P. M. (2018). *Genetic flora of the British Isles database*. Retrieved from <http://elmer.rbge.org.uk/geneticflora/gflora.php%0A%0A>
- Saravanan, M., Mohanapriya, G., Laha, R., & Sathishkumar, R. (2019). DNA barcoding detects floral origin of Indian honey samples. *Genome*, *62*(5), 341–348. <https://doi.org/10.1139/gen-2018-0058>.
- Särkinen, T., Staats, M., Richardson, J. E., Cowan, R. S., & Bakker, F. T. (2012). How to open the treasure chest? Optimising DNA extraction from herbarium specimens. *PLoS One*, *7*(8), e43808. <https://doi.org/10.1371/journal.pone.0043808>.
- Stace, C. A. (2019). *New Flora of the British Isles* (4th ed.). C&M Floristics.
- Stace, C. A., & Crawley, M. J. (2015). *Alien plants*. William Collins.
- Stace, C. A., Preston, C. D., & Pearman, D. A. (2015). *Hybrid flora of the British Isles*. Botanical Society of Britain and Ireland.
- Stoeckle, M. Y., Gamble, C. C., Kirpekar, R., Young, G., Ahmed, S., & Little, D. P. (2011). Commercial teas highlight plant DNA barcode identification successes and obstacles. *Scientific Reports*, *1*, 1–7. <https://doi.org/10.1038/srep00042>.
- Thomsen, P. F., & Sigsgaard, E. E. (2019). Environmental DNA metabarcoding of wild flowers reveals diverse communities of terrestrial arthropods. *Ecology and Evolution*, *9*(4), 1665–1679. <https://doi.org/10.1002/ece3.4809>.
- Tizard, J., Patel, S., Waugh, J., Tavares, E., Bergmann, T., Gill, B., Norman, J., Christidis, L., Scofield, P., Haddrath, O., Baker, A., Lambert, D., & Millar, C. (2019). DNA barcoding a unique avifauna: An important tool for evolution, systematics and conservation. *BMC Evolutionary Biology*, *19*(1), 1–13. <https://doi.org/10.1186/s12862-019-1346-y>.
- Twyford, A. D. (2014). Testing evolutionary hypotheses for DNA barcoding failure in willows. *Molecular Ecology*, *23*(19), 4674–4676. <https://doi.org/10.1111/mec.12892>.
- Walker, K. J., & Preston, C. D. (2006). Ecological predictors of extinction risk in the flora of lowland England. *UK Biodiversity and Conservation*, *15*(6), 1913–1942. <https://doi.org/10.1007/s10531-005-4313-4>.
- Wang, X., Gussarova, G., Ruhsam, M., de Vere, N., Metherell, C., Hollingsworth, P. M., & Twyford, A. D. (2018). DNA barcoding a taxonomically complex hemiparasitic genus reveals deep divergence between ploidy levels but lack of species-level resolution. *AoB PLANTS*, *10*(3), 1–13. <https://doi.org/10.1093/aobpla/ply026>.
- Wirta, H., Várkonyi, G., Rasmussen, C., Kaartinen, R., Schmidt, N. M., Hebert, P. D. N., Barták, M., Blagoev, G., Disney, H., Ertl, S., Gjelstrup, P., Gwiazdowicz, D. J., Huldén, L., Ilmonen, J., Jakovlev, J., Jaschhof, M., Kahanpää, J., Kankaanpää, T., Krogh, P. H., ... Roslin, T. (2016). Establishing a community-wide DNA barcode library as a new tool for arctic research. *Molecular Ecology Resources*, *16*(3), 809–822. <https://doi.org/10.1111/1755-0998.12489>.
- Wright, E. S. (2016). Using DECIPHER v2.0 to Analyze Big Biological Sequence Data in R. *The R Journal*, *8*(1), 352–359. <https://doi.org/10.32614/RJ-2016-025>.
- Xu, C., Dong, W., Shi, S., Cheng, T., Li, C., Liu, Y., Wu, P., Wu, H., Gao, P., & Zhou, S. (2015). Accelerating plant DNA barcode reference library construction using herbarium specimens: Improved experimental techniques. *Molecular Ecology Resources*, *15*(6), 1366–1374. <https://doi.org/10.1111/1755-0998.12413>.
- Yao, H., Song, J., Liu, C., Luo, K., Han, J., Li, Y., Pang, X., Xu, H., Zhu, Y., Xiao, P., & Chen, S. (2010). Use of ITS2 Region as the Universal DNA Barcode for Plants and Animals. *PLoS One*, *5*(10), e13102. <https://doi.org/10.1371/journal.pone.0013102>.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Jones L, Twyford AD, Ford CR, et al. Barcode UK: A complete DNA barcoding resource for the flowering plants and conifers of the United Kingdom. *Mol Ecol Resour*. 2021;00:1–13. <https://doi.org/10.1111/1755-0998.13388>