Natural England Joint Publication JP035

Evaluation of DNA barcode libraries used in the UK and developing an action plan to fill priority gaps

Funded by the Defra DNA Centre of Excellence

First published October 2020

www.gov.uk/natural-england

Foreword

Natural England is part of the Defra DNA Centre of Excellence, which champions the uptake of DNA based tools for monitoring the environment to inform its management and regulation. Natural England commissioned this report on behalf of the DNA Centre of Excellence. Natural England commission a range of reports from external contractors to provide evidence and advice to assist us in delivering our duties. The views in this report are those of the authors and do not necessarily represent those of Natural England.

Background

DNA – based methods offer a significant opportunity to monitor individual species and species assemblages where appropriate, for example those that may be difficult to monitor using traditional methods. However, with the exception of some individual species such as the great crested newt, there is still much development of these techniques required before they can be used in routine monitoring.

Natural England has been developing the use of DNA-based methods for monitoring for several years and is a founding member of the Defra DNA Centre of Excellence, which was set up to encourage collaboration across the Defra group to progress the use of DNA based methods by tackling cross-cutting barriers.

Gaps in DNA reference libraries of UK species were identified by the Defra DNA Centre of Excellence Working Group as one of the main barriers preventing the further uptake of DNA based methods for environmental species monitoring.

This report is the first step towards rectifying this by providing an assessment of the current state of reference libraries available for all known UK taxa; and prioritising key taxa where obtaining DNA barcode references should be a priority.

This report should be cited as:

Price, Briscoe, Misra and Broad (2020) DEFRA Centre of Excellence for DNA Methods: Evaluation of DNA barcode libraries used in the UK and developing an action plan to fill priority gaps. *Natural England Joint Publication JP035*.

Natural England Project Manager – Andy Nisbet, Principal Adviser, Natural England
andy.nisbet@naturalengland.org.uk

Contractor – The Natural History Museum

Keywords – Defra DNA Centre of Excellence, DNA barcode libraries, Tree of Life

Further information
This report can be downloaded from the Natural England Access to Evidence Catalogue: **http://publications.naturalengland.org.uk/** . For information on Natural England publications contact the Natural England Enquiry Service on 0300 060 3900 or e-mail **enquiries@naturalengland.org.uk**.

ISBN 978-1-78354-671-8

DEFRA Centre of Excellence for DNA Methods:

Evaluation of DNA barcode libraries used in the UK and developing an action plan to fill priority gaps

Benjamin Price, Andrew Briscoe, Raju Misra and Gavin Broad

Natural History Museum, London

For further information please contact:

The Natural History Museum
Cromwell Road
London, SW7 5BD
https://nhm.ac.uk/

Acknowledgements:

We would like to acknowledge the contribution of those that responded to the survey, attended the workshop and contributed in the follow up consultation. This project was led and funded by the Defra DNA Centre of Excellence, which aims to transform Defra Group science to support policy and delivery by sharing and demonstrating the wider potential of existing methods; enabling the rapid development and trialling of new DNA based method applications; providing a focal point for leadership in Defra, UK and across other Government Departments; and building shared capability including skills and facilities.

Contents

Executive Summary

- There are approximately 76,000 eukaryote species recognised in the UK, and while we know some of them in great detail, the majority of these species are poorly known, and hundreds of new species are discovered each year.
- DNA barcoding uses a short, standardised segment of an organism's genome for identification by comparison to a reference library; however, the UK lags behind several countries in Europe and North America in that we lack trusted, reliable and openly accessible reference sequences for key UK taxa.
- This report is the first step in rectifying this by engaging diverse stakeholders to facilitate collaboration and coordination; providing robust stakeholder-based and independent assessment of the current state of reference libraries available for all known UK taxa; and prioritising key taxa.
- A survey was developed and shared with the UK research and end user community, receiving 80 responses from a wide range of stakeholders and covering the focal taxa / assemblages and habitats; the DNA reference libraries in use, their quality assurance and perceived coverage.
- A formal gap analysis of the public DNA data in major DNA reference libraries highlighted that an estimated 52% of UK species have publicly available DNA data of some sort; however, coverage in gene specific reference libraries varies greatly (eg 2 – 52%), as does the associated quality assurance.
- Priority taxa highlighted by end users had coverage in reference libraries ranging from almost complete, in the case of known invasive non-native species, to significant coverage (71%) for taxa with conservation designations. However, these data also vary by kingdom and reference library, as does the associated quality assurance.
- If taking a strict requirement of DNA data provided by UK specimens and held in UK repositories, for robust QC and QA, then the proportion of UK species with public DNA data in reference libraries falls to less than 4% in the largest reference library assessed (BOLD).
- While standard genes for DNA-based identification have essentially been established, more work is required to establish the priority taxa required for regulatory delivery in contrast to taxa that are surveyed in a non-regulatory framework.
- Several barriers to the development of barcode libraries were highlighted, the most relevant being sustained large scale funding, expertise, capacity, laboratory skills and equipment, quality control and assurance, collecting logistics (eg permits and access) and communication.
- Significant opportunities identified include a large network of interested experts, several organisations with significant delivery capabilities, current large-scale projects and funding opportunities, emerging technologies and the economy of scale for DNA sequencing.
- Following a stakeholder workshop, we have outlined a concise action plan to provide reliable, open access reference sequences, linked to open access vouchers, identified by known experts, to facilitate UK academic and regulatory aims.

Appendices

Appendix 1: Survey questions (PDF)
Appendix 2: Supplementary tables S1, S2, S3 (PDF)
Appendix 3: Priority single species (MS Excel). Available at: http://doi.org/10.5281/zenodo.3965809
Appendix 4: Raw data (MS Excel). Available at: http://doi.org/10.5281/zenodo.3965809
Appendix 5: Workshop agenda (PDF)

List of tables

List of figures

Chapter 1: Introduction

Background

There are approximately 76,000 species of eukaryote recognised in the UK, ranging from single celled diatoms to baleen whales. We know in great detail the species composition, distribution and some ecology of much of the plant fauna; many vertebrates, especially birds; some invertebrates, notably butterflies and moths; and a few other small groups of organisms. However, many of these species are relatively poorly known; there are hundreds of species of fungi and animals found new to the UK every year, comprising species not previously recorded in the UK and newly described species. Our knowledge of several kingdoms of life, which are traditionally combined under ‘Chromista’ and ‘Protozoa’, is largely rudimentary. For those few species with sufficient data, their abundance and distribution has, on average, declined since 1970 and of the 8,431 species that have been formally assessed, 15% are threatened with extinction and 133 species are already extinct in the UK (Hayhow and others, 2019).

Table 1: Species totals for major groups of eukaryotes, derived from the UK Species Inventory (UKSI), maintained at the Natural History Museum, together with the numbers of non-native species, species with a conservation designation or legal protection.

Kingdom	UKSI species	Non-native	JNCC designated*	Legally protected**
Animalia	42,780	1,325	7,506	517
Chromista	3,828	12	10	0
Fungi	18,547	1	2,460	35
Plantae	10,429	1,543	2,301	198
Protozoa	665	1	20	0
Total	**76,249**	**2,883**	**12,297**	**750**

** 1,051 taxa on the JNCC list are not at the species level and are not reported here; ** 19 legally protected taxa on the JNCC list are not at the species level and are not reported here. See Chapter 3 for details on conventions / criteria used to denote legal protection.*

A relatively small number of UK organisms are protected by any legislation (Table 1) and few are routinely monitored to track the health of ecosystems. However, there is a huge diversity of life that could be more routinely monitored and identified to provide a more complete picture of our biodiversity and the health of ecosystems.

Taxonomic capacity

For various groups of organisms, including some of our most species-rich groups, we lack the national capacity to identify species. The national shortfall in taxonomic expertise prompted the House of Lords Science and Technology Committee to launch an inquiry into the state of systematics research and taxonomy: 'What on Earth? The threat to the science underpinning conservation' (House of Lords, 2001-02), and a follow-up inquiry: 'Systematics and Taxonomy: Follow-up' (House of Lords, 2007-08).

The report made numerous suggestions that were agreed upon by the government: 'The Government's Response and the Committee's Commentary' (House of Lords, 2002-03), concerning important areas such as increasing training of taxonomists, digitising collections and increasing capacity for high throughput molecular identifications. Key organisations for delivering these goals, particularly Biotechnology and Biological Sciences Research Council (BBSRC), Natural Environment Research Council (NERC) and the Natural History Museum (NHM), agreed that they were committed to delivering an increase in systematics output and that their strategies reflected these goals. In practice, this has not translated into stable funding streams and there has been no discernible increase in professional taxonomists, although successful initiatives such as the HLF-funded 'Identification Trainers of the Future', and the Field Studies Council's identification courses have ensured that identification skills are being maintained in some of the more difficult groups of organisms. In other words, our lack of capacity to identify many groups of organisms remains at a similar level that was concerning enough for the House of Lords to launch its inquiries almost two decades ago.

DNA barcoding

DNA "barcoding" uses a short, standardized segment of an organism's genome for identification, much like the barcodes found on commercial products (Hebert and others, 2003). These DNA-based identifications require comparison to reference libraries of DNA "barcodes" sequenced from identified individuals.

The International Barcode of Life (iBoL) launched in 2008 and has since grown to be the largest and most established DNA barcode sequencing consortium in the world. iBoL leads a network of national hubs, across over 30 countries and more than 1,000 researchers, to help progress the goal to barcode sequence all life; and align with national strategic goals to create a robust framework for biodiversity surveillance and diagnostic applications. iBoL has successfully delivered on its first phase "Barcode 500k", creating a global network of national government driven barcoding programmes, informatics and sequencing infrastructure to barcode sequence 500,000 species. The second phase "BIOSCAN" launched in June 2019, leverages this current network to barcode 2 million species by 2026.

iBoL as a hub for this global network has initiated several national programmes. Examples of European barcoding programmes include, Norway (NorBol), Germany (GBol), Switzerland (SwissBol) and Finland (FinBol). Each is a key member of the iBol network, supported by national government, aligning directly with government strategies to benefit from biodiversity surveillance, monitoring and diagnostics. The UK currently lacks a coordinated national barcoding campaign, which was recognized as a limiting factor by the House of Lords review (2007-08):

"*The Committee is concerned about lack of co-ordination of barcoding effort nationally and about the potential for duplication of effort. The efficiency of barcoding as a diagnostic technique increases in proportion to the number of different species barcodes available for comparison.*"

However, since this report there has been an increase in DNA barcoding in the UK, most notably the almost complete coverage of flowering plants and conifers of Wales (de Vere and others, 2012). Furthermore the UK has already begun to integrate DNA based monitoring into the existing regulatory framework, for example complimenting manual surveys for Great Crested Newts with an environmental DNA (eDNA) tool (Biggs and others, 2014). While DNA based methods for other protected species are being developed, they are yet to be incorporated into regulation.

Gaps in reference libraries

The principal limiting factor in the successful implementation of DNA-based identification is the inadequate coverage of DNA reference libraries for focal taxa. The largest global reference library is BOLD (Ratnasingham and Hebert, 2007), which currently comprises 8 million barcodes of 675,000 putative species, however this is still a small proportion of the approximately 2 million known and 10 million estimated eukaryote species (Mora and others, 2011; Costello and others, 2012; Larsen and others, 2017).

In 2018, the DNAqua-Net network (Leese and others, 2016) examined the 28,000 aquatic species surveyed in Europe under the EU Water Framework Directive (WFD) and Marine Strategy Framework Directive (MSFD) for DNA data in public reference libraries (Weigand and others, 2019). Their analysis of BOLD showed that coverage varies strongly among taxonomic groups, and among geographic regions. Furthermore, a large proportion of species (up to 50%) in several taxonomic groups are only represented by private data (Weigand and others, 2019).

More recently, in a UK perspective, a gap analysis of BOLD for the 13,773 taxa in the Pantheon database (Webb and others, 2018) revealed that 168 species (1.2%) have sufficient data to enable identification with "high" confidence, whereas 3,025 species (22.2 %) have sufficient data to be identified with "moderate" confidence (Macadam and others, 2020). To date no analysis of the whole UK biota has been undertaken.

Aims and objectives

This project aims to provide (1) robust stakeholder-based and independent assessments of the current state of single / multi gene and genomic reference libraries available for all known UK taxa; (2) prioritise key taxa for reference library development in line with regulatory requirements; and (3) develop a concise action plan to rapidly provide reliable, open access reference sequences, linked to open access vouchers, identified by known experts, in order to facilitate academic and regulatory aims.

In order to achieve the aims the project was organized into four main tasks:

1. Catalogue and evaluate DNA barcode libraries used in the UK (see Chapter 2)
2. Describe end user needs and prioritise gaps (see Chapter 3)
3. Identify opportunities and barriers to developing barcode libraries (see Chapter 4)
4. Develop proposals for improving DNA barcode libraries for the UK (see Chapter 5)

Methods

Survey and Consultation: A survey was developed and provided online through Cognito Forms (survey Appendix 1). Consultation included members of the UK DNA working group, regulatory agencies, academic organisations, commercial entities, non-governmental organizations and non-departmental public bodies. The survey was run publicly from 22 January to 14 February 2020. Responses to key questions are summarized below with additional comments where appropriate.

Gap Analysis: A formal gap analysis of reference libraries identified in the survey was undertaken using the UK species inventory, a compilation of species and higher taxa from several component checklists, used as the UK standard list in biological recording. The JNCC list of conservation designations and the non-native invasive species list, were also included to assess barcoding gaps against key taxa.

Seven of the DNA sequence databases most cited by survey respondents were then assessed against the UKSI list. Taxon synonyms and misspellings were not collated or assessed, thus these analyses are a conservative estimate of the current state of database completeness relative to UK taxa. Several smaller databases identified from the survey results were not assessed due to an inability to access the dataset, or to avoid redundancy by excluding databases containing sequence information mined directly from larger repositories which were assessed.

Workshop: A workshop was held at the Natural History Museum (12th March 2020) to summarise the results of the survey and consultation, elicit feedback on the draft report and outline an action plan.

Chapter 2: DNA barcode libraries and their coverage of UK species

Several different DNA sequence libraries are used in the UK, each with differing focal taxa, levels of completeness, accessibility and quality assurance. This fragmented landscape prevents end users from routinely validating taxonomic assignments and developing robust DNA-based methods for biodiversity monitoring. This chapter details the survey responses and the formal gap analysis.

Survey Results

Overview of respondents

The survey resulted in 80 responses with most responses coming from academics (31), followed by regulators/Government organisations (29), NGOs (10), Commercial (6) and joint Commercial / Government ventures (3). While the respondents were self-selecting the number and range of respondents suggests this is representative of the UK community. Some individual responses were from multiple people who collated the views of their organisation before responding - typically in the Government sector (Appendix 2 - Table S1).

Most respondents work across the UK, while some are restricted to one or more of the countries making up the UK (Table 2). Several respondents work internationally in addition to the UK, working either in South East Asia, Germany, Scandinavia, Canada or undefined international locations. Two respondents worked exclusively outside of the UK but in UK overseas territories.

Table 2: Summary of focal region studied.

Region	**Count**	**% respondents**
UK wide	46	58
England	17	21
Northern Ireland	2	2
Scotland	7	9
Wales	6	8
Overseas territories	2	2
Global	8	10

Most respondents undertake multiple types of surveys, primarily academic research and / or biodiversity / conservation assessment and monitoring, followed by conservation / environmental management (Table 3).

Table 3: Summary of the types of surveys undertaken.

Type of survey	Count	% respondents
Academic	56	70
Biodiversity / Conservation Assessment and Monitoring	57	70
Conservation / Environmental Management	39	49
Recording / Citizen Science	25	31
Regulatory / Statutory	25	31
Commercial	17	21

The level of involvement in DNA-based identification was mixed and several respondents provided multiple answers (eg thirteen respondents both carry out research / surveys and commission them). Approximately 75% of respondents carry out DNA-based identification themselves, while 25% commission DNA-based monitoring surveys and use the resulting identifications (Table 4). Understandably responses from academics were skewed towards carrying out research / surveys, whereas responses from other end users (eg Government, commercial organisations, NGOs) were more balanced between the three categories. Of the 11 respondents who answered “none”, three expect their organisation to use DNA-based identification within the next five years, while three were unsure and five did not expect this to occur.

Table 4: Summary of the level of involvement in DNA-based identification by organisation type.

Organisation type	Involvement	Responses
Academic	Carry out research / surveys (including molecular lab and bioinformatics)	26
	Carry out research / surveys (including molecular lab and bioinformatics) Commission work and use resulting identifications	3
	None	4
Commercial	Carry out research / surveys (including molecular lab and bioinformatics)	4
	Commission work and use resulting identifications Carry out research / surveys (including molecular lab and bioinformatics)	1
	None	1
Government (including non-departmental public bodies)	Carry out research / surveys (including molecular lab and bioinformatics)	13
	Carry out research / surveys (including molecular lab and bioinformatics) Commission work and use resulting identifications	6
	Commission work and use resulting identifications	6
	None	4
Joint Venture (commercial + govt / private)	Carry out research / surveys (including molecular lab and bioinformatics)	1
	Commission work and use resulting identifications Carry out research / surveys (including molecular lab and bioinformatics)	1
NGO / trust / non-profit	Carry out research / surveys (including molecular lab and bioinformatics)	3
	Commission work and use resulting identifications	3
	Commission work and use resulting identifications Carry out research / surveys (including molecular lab and bioinformatics)	1
	None	2

The majority of respondents' survey for all native species (Table 5), likely reflecting the large number of academic surveys focusing on multi-species assemblages or those without a specific taxon focus (ie all species). Over half of the respondents' survey native (protected) and / or invasive non-native species. The need to generate data on taxon abundance was listed as a requirement by 51 respondents (64%); however, further work is required to assess which groups / species require abundance data in a regulatory framework.

Table 5: Summary of the status of focal species surveyed with either morphological or DNA methods.

Species status	**Count**	**% respondents**
Native (not protected species)	64	80
Native (protected species)	49	61
Invasive non-native species	42	53
Potentially invasive non-native species	32	40
Horticultural	1	1
Migratory species	1	1
Pests and pathogens	1	1

The majority of respondents (61%) survey multiple habitat types, with freshwater and / or terrestrial habitats being the most surveyed habitats overall (Table 6).

Table 6: Summary of the habitats surveyed.

Habitat	**Count**	**% respondents**
Freshwater	54	68
Terrestrial	49	61
Marine	33	41
Soil	26	33
Estuarine	3	4
Bark	1	1
Faecal	1	1
Host Associated	1	1

Approximately half of the respondents surveyed listed only a single taxonomic scope (ie one of the options shown in Table 7), whereas the other half surveyed multiple groups with differing taxonomic scope. The majority of respondents focus on targeted multi-species assemblages.

Table 7: Summary of the taxonomic scope.

Taxonomic scope	Count	% respondents
Targeted multi-species assemblages	56	70
All species (ie no specific targets)	39	49
Single species	37	46

Reference libraries

The survey responses highlighted that the vast majority of respondents use the International Nucleotide Sequence Database Collaboration (INSDC), comprising Genbank at the National Centre for Biotechnology Information (NCBI), the European Nucleotide Archive (ENA), and the DNA DataBank of Japan (DDBJ), followed by BOLD and bespoke databases reflecting their taxonomic scope (Table 8).

Table 8: DNA sequence databases used for identification in the UK.

Database	Count	% respondents (59 total)
INSDC (NCBI / ENA / DDBJ) https://www.insdc.org	52	89
BOLD http://www.boldsystems.org/	34	58
Bespoke (see public links below)	28	47
EUKREF http://eukref.org/	16	27
UNITE https://unite.ut.ee/	11	19
MIDORI http://reference-midori.info/index.html	4	7
ArthemisDB@se http://arthemisdb.supagro.inra.fr/	1	2
Diat.barcode https://www6.inrae.fr/carrtel-collection/Barcoding-database	3	6
EPPO Q-Bank https://qbank.eppo.int/	1	2

Of the 28 respondents who use bespoke reference libraries only seven make use of four databases that are publicly available:

1. Data processing workflow for Handley and others, 2019
2. ScreenForBio
3. A generalised, dynamic DNA reference library for UK fishes
4. THAPBI Phytophthora ITS1 Classifier Tool (PICT)

A further public curated ITS database for plants (PLANiTS) was recently published (Banchi and others, 2020), and while not highlighted by respondents, this library has been included in the formal gap analysis.

Quality assurance: QA measures that were highlighted by respondents as present in the reference libraries they use, are summarised below (Table 9). As multiple reference libraries are used by most respondents some multiple values (eg those including “none”) were disambiguated.

Table 9: Quality assurance measures in various reference libraries used by respondents.

Quality assurance measure	Count	% respondents (60 total)
Specimen collection data provided	30	50
None	25	42
Vouchers deposited in recognized public repositories	24	30
Multiple vouchers & sequences available per species	23	29
Database curated by taxon experts	23	29

Notable “other” comments included that quality assurance is very mixed across databases and is variable within a single reference library. Single responses included QA through “historical data and statistical power of the study”; “manual curation”; “proprietary modelling of taxonomic performance”; and “manual cross-checking of taxon assignments”. The “taxonomic muddle” [the occurrence of reference sequence(s) for a single taxon under different names in different databases] was highlighted as causing issues when using some databases.

Target genes: A total of 15 gene regions were identified by respondents (Table 10). Most respondents survey multiple different taxon groups and therefore use different genes for identification, depending on the taxonomic group. The most widely used gene is mitochondrial Cytochrome c oxidase subunit 1 (COI), which is understandable given it is the basis of metazoan barcoding and comprises the majority of records included in the BOLD database.

Other notable gene regions include several hypervariable regions of the 18S ribosomal gene, which is primarily used for whole community assessment of eukaryotes, and mitochondrially encoded 12S ribosomal gene, which is increasingly used for fish identification from environmental samples. The standard DNA barcode for Fungi is the internal transcribed spacer (ITS1/2) situated between the nuclear 18S, 5.8S and 28S ribosomal genes, while plant identifications typically utilise ITS as well as the chloroplast RuBisCo (rbcL) and Maturase K (matK) genes, and bacterial identification uses 16S ribosomal RNA.

Table 10: Summary of the DNA regions used for taxonomic identification.

Region	**Taxonomic scope**	**Count**	**% respondents (57 total)**
COI	Invertebrates Vertebrates (Mammals)	41	72
18S	All	28	49
ITS	Plants Fungi	26	46
12S	Fish Lichen / Cyanobacteria	24	42
16S (Bacteria)	Bacteria	23	40
rbcL	Plants	23	40
16S (Eukaryote)	Invertebrates	15	26
matK	Plants	10	18
28S	Eukaryotes	6	11
trnL	Plants	3	5
psbA-trnF	Plants	1	2
cytb	Animals	1	2
MCM7	Lichen / Cyanobacteria	1	2
RPB1/2	Lichen / Cyanobacteria	1	2
rbcLX	Lichen / Cyanobacteria	1	2

Library coverage: The majority of respondents, 44/69 (64%), agree that the current reference libraries do not currently have sufficient taxonomic coverage for their work, while a further 12 (17%) do not know. When asked to estimate the coverage of the reference libraries for their target taxa most respondents did not know, while three considered coverage to be 100% (Table 11). Several respondents highlighted that coverage varies widely amongst taxonomic groups and this could not be reflected in the structure of the survey.

The majority of respondents, 42/60 (70%), contribute sequences to public repositories (notably NCBI / ENA / DDBJ), though only half of these respondents, and a third of all respondents (20/60; 33%), associate these sequences with “publicly accessible voucher specimens”.

Table 11: Summary of estimated coverage of reference libraries for focal taxa.

Approximate coverage	Count	% respondents (69 total)
100%	3	0.4
75%	10	14
50%	10	14
25%	12	17
Don't know	34	49

Formal gap analysis

Several taxon lists were combined for analysis (Table 12) using the UK species inventory (UKSI), held and managed by the Natural History Museum, as a backbone. As the UKSI is a compilation of species from several component checklists, it includes taxa that would not normally be found in the UK but have needed to be recorded at some point in the past (eg vagrant species).

In addition to the UKSI, the JNCC list of conservation designations of UK taxa and the non-native invasive species list, the latter compiled by the GB non-native species secretariat, were also included to assess barcoding gaps against key taxa / taxonomic groups within the wider UKSI list. In order to ensure that information in databases were comparable, all taxa were matched based on either their UKSI taxon version key (TVK), organism key, or NCBI TaxID where appropriate. The higher taxonomy follows that of the National Biodiversity Network (NBN) which is originally sourced from the UKSI.

As these databases contain records at different taxonomic levels (ie order, family, genus), all non-species level records were removed to limit spurious matches at higher taxonomic ranks and avoid difficulties in matching subspecies (see Table 12 for numbers of records excluded).

Table 12: Summary of taxon data sources for gap analysis.

Taxon list	Scope and resource link	Date accessed
UK Species Inventory (UKSI)	All taxa recorded in the UK. Assessed at the species level (76,249 eukaryote species) due to complexity of matching subspecies ranks to external resources. UK species	9 October 2019
JNCC conservation designations for UK taxa	UK taxa with conservation designations (13,353 taxa). Matched against UKSI using Taxon_Version_Key and/or UKSI Organism_Key. Note: 1,051 non species level taxa were removed due to complexity of matching to external resources. Conservation designations for UK taxa	17 January 2020
NBN Biota	UKSI mapped to a standardized higher taxonomy (ie missing sub-ranks). Includes NBN native / non-native status. NBN Biota list	03 February 2020
Non-native invasive species list	Invasive non-native species listed by GB non-native species secretariat. Species identification sheets	13 February 2020

Seven of the DNA sequence databases most cited by survey respondents were then assessed against the UKSI list. Taxon synonyms and misspellings were not collated or assessed, thus the gap analysis relied on exact matches to taxonomy, either through NCBI taxon ID codes or taxon names (outlined in Table 13). As a result, these analyses are a conservative estimate of the current state of database completeness relative to UK taxa. The data from BOLD was not assessed for private vs public data. Several databases identified from the survey results were not assessed due to an inability to access the dataset (ArthemisDB@se & EPPO-Q-bank), or to avoid redundancy by excluding databases containing sequence information mined directly from NCBI’s GenBank. In addition to these omissions, MIDORI was also left out of the analyses as it has not been updated since February 2018.

Table 13: Summary of DNA databases and how assessed.

Database	How assessed
INSDC (ENA/NCBI / DDBJ) https://www.insdc.org	All UKSI species using a simple taxonomy match via NCBI taxonomy match service. Relevant genes were not assessed at this stage.
BOLD http://www.boldsystems.org/	Three checklists loaded into BOLD and publicly available within BOLD: CL-UKSI1 Animals CL-UKSI2 Plants CL-UKSI3 Fungi Chromista, Bacteria and Protozoa not assessed
EUKREF - PR2 (Protist Ribosomal Reference Database) https://github.com/pr2database/pr2database	Exact match to NCBI taxonomy IDs
EUKREF - SILVA https://www.arb-silva.de/	Exact match to NCBI taxonomy IDs
UNITE https://unite.ut.ee/	Exact match to NCBI taxonomy IDs
PLANiTS https://github.com/apallavicini/PLANiTS	Exact match to combined list of taxon names for ITS1 and ITS2
Bespoke (public)	Not assessed
MIDORI http://reference-midori.info/index.html	Not assessed
ArthemisDB@se http://arthemisdb.supagro.inra.fr/	Not assessed
Diat.barcode https://www6.inrae.fr/carrtel-collection/Barcoding-database	Exact match to list of taxon names
EPPO-Q-Bank https://qbank.eppo.int/	Not assessed

The UKSI contains 76,249 eukaryote species, broken down into five kingdoms as grouped by the NBN (Figure 1, Table 1) comprising Animals (56%), Fungi (24%) and Plants (14%), with Chromista and Protozoa making up the remaining diversity (6%). These latter groups are poorly known at the species level and harbour much more diversity in the UK.

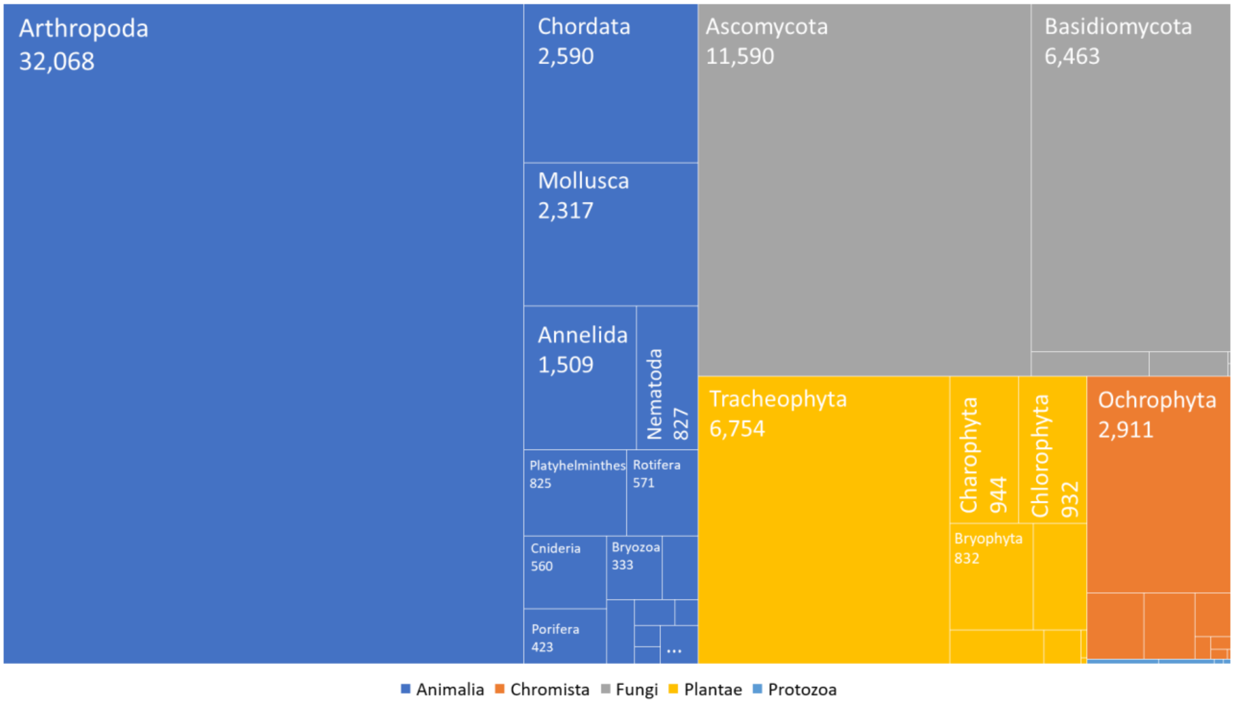

Figure 1: Valid eukaryote species in the UKSI grouped by phylum (boxes) with kingdom (colour). Note only major phyla are labelled due to space constraints.

Overall just over half (52%) of all known UK species have DNA data in the global INSDC public archive, however the coverage in curated gene specific reference libraries is highly variable between taxon groups, with reference libraries for plants (52%) and animals (50%) being the most well represented in BOLD (Figure 2, Table 14).

Table 14: Known eukaryote species diversity and sequence data publicly available by kingdom.

		Species with public DNA data							
Kingdom	**UKSI species**	**ENA**	**BOLD**	**PR2 (16S)**	**PR2 (18S)**	**SILVA**	**UNITE**	**PLANiTS**	**Diat. barcode**
Animalia	42,780	21,037	21,446		3,784	689			
Chromista	3,828	1,455		105	805	387			539
Fungi	18,547	9,249	5,537		2,456	580	4,222		
Plantae	10,429	7,865	5,423	373	1,203	666	4	4,207	19
Protozoa	665	339		60	206	50	1		
Total	**76,249**	**39,946**	**32,406**	**538**	**8,454**	**2,373**	**4,227**	**4,207**	**558**

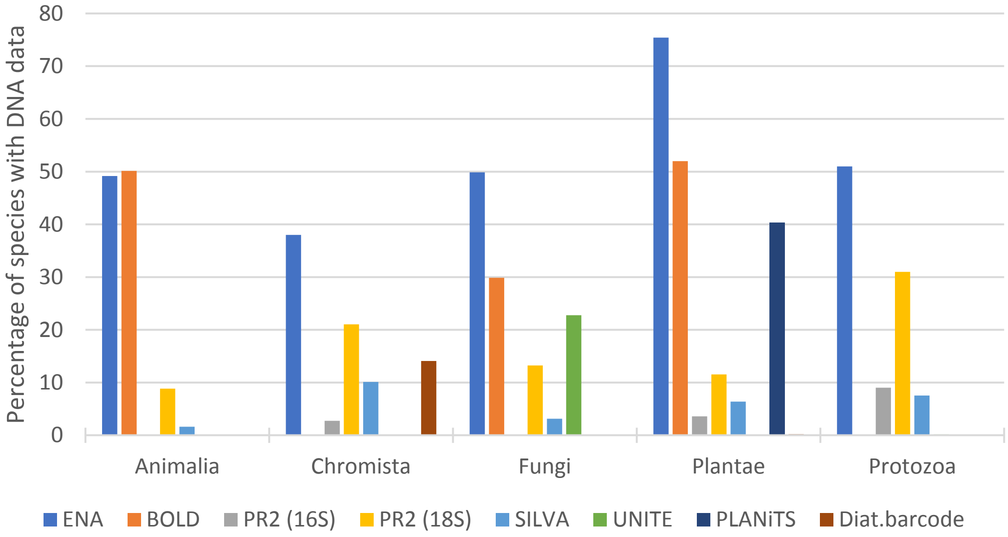

Figure 2: Proportion of UK species with publicly available DNA data grouped by kingdom.

Approximately 70% of all species included on the JNCC list of taxa with conservation designations, have DNA data in the global public archive (ENA/NCBI/DDBJ). The coverage in curated gene specific reference libraries follows the general trend for all UK taxa, and is highly variable between taxon groups and reference libraries with Plants (78%) and Animals (72%) being the most well represented in BOLD (Figure 3, Table 15).

Table 15: Overview of species (not taxa) on the JNCC conservation designation list and the percentage with public sequence data grouped by kingdom.

		% species with conservation designations and DNA data							
Kingdom	**JNCC species**	**ENA**	**BOLD**	**PR2 (16S)**	**PR2 (18S)**	**SILVA**	**UNITE**	**PLANiTS**	**Diat. barcode**
Animalia	7,506	69.4	71.8	-	9.5	2.1	-	-	-
Chromista	10	80.0	-	-	20.0	10.0	-	-	-
Fungi	2,460	62.5	46.5	-	15.0	1.6	30.5	-	-
Plantae	2,301	85.6	78.1	3.2	10.0	5.5	-	51.9	-
Protozoa	20	50.0	-	-	25.0	-	-	-	-
Total	**12,297**	**71.0**	**67.7**	**-**	**10.7**	**2.7**	**-**	**-**	**-**

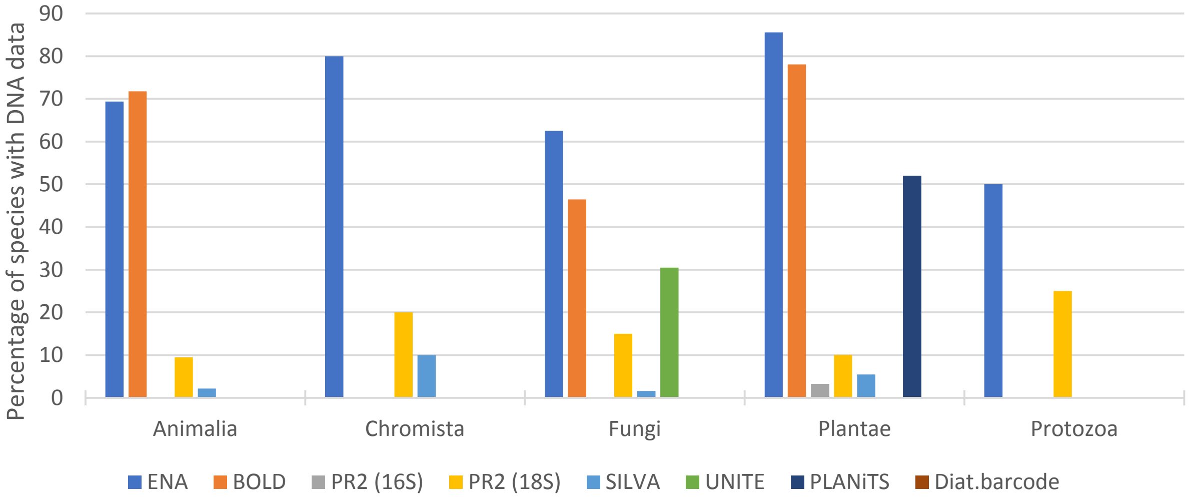

Figure 3: Proportion of UK species with conservation designations and publicly available DNA data grouped by kingdom.

UK specimens and vouchers

The recent analysis of species listed in the Pantheon database and the BOLD reference library (Macadam and others, 2020) highlighted that only a small proportion of UK arthropod species are represented with UK specimens and / or specimens vouchered in UK institutions. In order to assess this against the entire UK species inventory, an initial analysis examined BOLD as this contains the most structured voucher metadata. The specimen records within BOLD, designated as collected in the UK, were downloaded and investigated. A total of 24,548 specimen records from the UK are present in BOLD (see Appendix 4), comprising 5,479 species and 3,092 BINs[1]. This equates to 4% of the UK species having at least one reference sequence on BOLD generated from UK specimens. Of these UK based specimens 11,763 specimens are likely to also be vouchered in the UK based on their institutional metadata, corresponding to 2,724 species and 1,443 BINs, equating to 1.9% of the UK species with a UK specimen vouchered in a UK institution (see Appendix 4).

As anticipated by the previous analysis (Macadam and others, 2020) the low proportion of barcodes based on UK specimens extends to the entire UK species list within BOLD, and it is likely this pattern would be observed in the other reference libraries if examined in detail. However, the lack of metadata held in many of the most widely used reference libraries, coupled with a lack of consistency and standardisation in the minimum amount of specimen metadata required for sequence submission / deposition (Table 16), limit the opportunities to perform this analysis across other sequence databases.

[1] The Barcode Index Number (BIN) system clusters sequences using well established algorithms to produce operational taxonomic units that closely correspond to species.

Reference library QA/QC

The minimum requirements for data submission, to the reference databases identified through the survey, are summarized in Table 16, along with additional quality assurance / quality control measures summarized below. It is important to note that many of the databases highlighted by end-users do not allow direct submission of new data but are curated by consortia of taxonomic experts from public data. Methods of quality control vary across all databases with most not factoring in quality of the sequencing data itself but relying on clustering techniques or phylogenetic analysis to remove incorrect / poorly identified records.

The International Nucleotide Sequence Database Collaboration (INSDC) is comprised of three databases; DNA Data Bank of Japan (DDBJ), European Nucleotide Archive (ENA), and National Center for Biotechnology Information (NCBI) that house annotated sequence data, as well as associated sequencing reads, with the aim of providing free and unrestricted access to the records contained in their databases. The databases that comprise the INSDC are not gene or taxon specific and while some quality checks are in place (eg translation of coding regions), the quality and accuracy of the data is the responsibility of the submitting author.

Barcode of Life Data System (BOLD) is a cloud-based data storage system and analysis platform developed by the Centre for Biodiversity Genomics, Guelph, Canada (Ratnasingham and Hebert, 2007). Data on BOLD is primarily focussed on four main barcoding genes; COI-5P (metazoa), ITS (fungi), matK & rbcL (plants) but do accept data from over 150 other markers commonly used for DNA barcoding. Both DNA sequences and associated metadata are quality checked before approval on BOLD. Records with sequences meeting specific criteria (eg COI sequence longer than 500bp and containing less than 1% ambiguous bases) are assigned a BIN (Ratnasingham and Hebert, 2013). All submissions to BOLD, or edits made to existing records, will be periodically submitted to NCBI's GenBank database.

PR2: Protist Ribosomal Reference Database is a reference database for small sub-unit rRNA (18S) sequences (Guillou and others, 2013). The database mainly consists of nuclear-encoded protistan sequences. However, metazoans, land plants, macrosporic fungi and eukaryotic organelles (mitochondrion, plastid and others) are also included. Sequence annotation is performed by experts for each taxonomic group and the database does not accept public submissions. Along with SILVA, PR2 is part of the EUKREF 18S RNA Collaborative Annotation Initiative (http://eukref.org/).

SILVA is a comprehensive on-line resource providing quality-checked and aligned small sub-unit (16S/18S) and large sub-unit (23S/28S) ribosomal RNA sequences across all three domains of life: Bacteria, Archaea, and Eukarya (Quast and others, 2013; Yilmaz and others, 2014). Along with PR2 it is part of the EUKREF 18S RNA Collaborative Annotation Initiative (http://eukref.org/).

UNITE is a web-based database focussing on the molecular identification of fungi, targeting the formal fungal barcode - the nuclear ribosomal internal transcribed spacer region (ITS). UNITE contains all eukaryotic ITS sequences available from the INSDC clustered to a standard species level (97-100% identity) creating a species hypothesis. Each species hypothesis is given its own Digital Object Identifier (DOI) to facilitate unambiguous scientific communication (Nilsson and others, 2018).

PLANiTS is a curated reference dataset for plant ITS sequences (Banchi and others, 2020). It contains all the available sequences from NCBI's GenBank dataset of Viridiplantae ITS1, ITS2 and entire ITS sequences including both Chlorophyta and Streptophyta. The sequences are retrieved from NCBI, and the ITS region is extracted. The sequences undergo an identity check to remove misidentified records and are clustered at 99% identity to reduce redundancy and computational effort.

MIDORI is a web platform that uses a curated reference dataset for taxonomic classification of metazoan mitochondrial-encoded gene sequences (Machida and others, 2017). The dataset is comprised of quality filtered mitochondrial protein coding and ribosomal gene sequences taken from NCBI's BLAST nucleotide database (nt).

ArthemisDB@se is a database containing COI-5P and ITS2 barcode sequences of arthropod species sequenced in INRAE, CIRAD and SupAgro laboratories (France). The database also hosts information about species distribution, biology and ecology with a focus on pest species, and their predators. The main groups represented in the database belong to Coleoptera, Diptera, Hemiptera, Hymenoptera and Lepidoptera as well as some mite families.

Diat.barcode is a sequence database dedicated to chloropoast rbcL sequences of diatoms (Rimet and others, 2019). DNA sequences come from two sources: the NCBI nucleotide database and unpublished sequencing data of culture collections.

EPPO-Q-BANK is a database to support diagnostic activities on plant pests, which comprises data of properly documented species and strains present in collections from which items can be used as controls in identification and detection tests. The entries in EPPO-Q-bank are updated by a team of curators from organisations with connections to phytosanitary collections. In addition to housing reference sequence data EPPO-Q-bank also contains protocols for DNA barcoding.

Table 16: Minimum requirements for record submissions to selected DNA sequence databases.

Database	Minimum submission requirements	Additional notes / features
INSDC GenBank, NCBI[1]	Specimen voucher, mitochondrial genetic code, only metazoan COI sequences – no flanking data	Can submit raw HTS[2] data via SRA[3] to allow external validation.
BOLD Specimen submission	Sample ID, Field ID, Voucher ID, Institution storing, Phylum, Country	
BOLD Sequence Submission	Sample ID, Marker, Institution	Sanger trace files only needed if sample has “barcode” status. Uses contamination library to flag suspect submissions.
EUKREF PR2	No direct submission	Curated by taxonomic experts.
EUKREF SILVA	No direct submission	Curated by taxonomic experts.
UNITE	No direct submission	Curated by UNITE community.
PLANiTS	No direct submission – data mined from NCBI	All records identity checked, and sequences clustered at 99% similarity.
MIDORI	No direct submission – data mined from NCBI	Records quality filtered by removing non-species names and sub-species ranks.
ArthemisDB@se	No direct submission	Only incorporates data from insect pests generated from approved research groups.
Diat.barcode	No direct submission – data mined from NCBI or UK diatom barcoding project	Quality checked using phylogenetic methods.
EPPO Q-BANK	No direct submission	Provides approved methodology and workflows for barcoding pest species.

[1] requirements listed are for submission of metazoan COI barcode data only; [2] HTS = High Throughput Sequencing; [3] SRA = Sequence Read Archive (https://www.ncbi.nlm.nih.gov/sra).

Chapter 3: End user needs and prioritising gaps

Identification of taxa using DNA sequencing is not uniformly applied across end users and regulators, in part due to a lack of quality assured end-to-end workflows and reference libraries for some groups. This chapter summarises the end user needs as identified in the survey and follow-up consultation; the gaps in the priority list are then briefly discussed.

The majority of survey respondents required species level identification (Table 17), followed by mixed taxon level identification (eg macroinvertebrate assemblages), highlighting that DNA methods will need to have species level ID for maximum uptake, thus requiring comprehensive reference libraries of UK species to be developed.

Table 17: Resolution required for taxon identification.

Resolution	Count	% respondents
Phylum	1	1
Order	10	14
Family	17	23
Genus	29	40
Species	56	77
Strain	2	3
Mixed taxon levels	42	56
Taxonomy free (mOTUs)	24	33

Key taxa

End users survey taxa either individually or as part of multi-species assemblages, each is dealt with below.

Individual species surveyed

The individual species were collated, deduplicated and are listed in Appendix 2 (Table S2) and the supplementary spreadsheet (Appendix 4). For Natural England the focal species were collated from the JNCC list of conservation designations, comprising only species with legal protection (Bern Convention; Bonn Convention; Birds Directive Annex 1; Habitats and Species Directive Annex 2; CITES Annex A; Wildlife and Countryside Act 1981; The Wildlife (Northern Ireland) Order 1985; The Conservation Regulations 1994 including Northern Ireland; and the Protection of Badgers Act 1992), which were supplemented with the NERC section 41 species.

A gap analysis of the priority species identified (Table 18) shows that almost all species have some public DNA data (89.3%), and that this wide coverage extends to the BOLD reference library (81.3%). One caveat in this analysis is that 38 taxa do not reliably match UKSI taxa, of which 18 have public sequence data in ENA, suggesting they would not negatively impact our estimated coverage.

Table 18: Overview of individual species listed in Appendix 3 as priority taxa and the percentage with public sequence data grouped by kingdom.

		% priority single species with public DNA data							
Kingdom	**Priority species**	**ENA**	**BOLD**	**PR2 (16S)**	**PR2 (18S)**	**SILVA**	**UNITE**	**PLANiTS**	**Diat. barcode**
Animalia	1,076	90.0	86.3	-	9.0	2.3	-	-	-
Chromista	16	100.0	-	6.3	37.5	37.5	-	-	-
Fungi	172	75.6	57.0	-	19.8	3.5	47.1	-	-
Plantae	416	92.8	81.3	3.4	10.6	3.6	-	54.3	-
Total	**1,680**	**89.3**	**81.3**	**0.9**	**10.8**	**3.1**	**-**	**-**	**-**

Multiple respondents highlighted invasive non-native taxa as targets for either single species or assemblage surveys. Almost all invasive non-native species, as listed by the non-native species secretariat (NNSS), have public DNA data (95%) and a significant proportion are represented in reference libraries, notably BOLD with 78% coverage (Table 19).

Table 19: Overview of invasive non-native species listed by the NNSS, and the percentage with public sequence data grouped by kingdom.

		% invasive non-native species with public DNA data						
Kingdom	**NNSS species**	**ENA**	**BOLD**	**PR2 (16S)**	**PR2 (18S)**	**PLANiTS**	**SILVA**	**Diat. barcode**
Animalia	52	88.4	80.4	-	29.4	-	7.8	-
Chromista	2	100	-	-	50	-	-	-
Plantae	91	100	78.9	4.4	15.6	63.3	8.9	-
Total	**146**	**95.2**	**78.3**	**2.8**	**21.0**	**39.9**	**8.4**	**-**

Assemblages monitored

Respondents to the survey listed almost all possible assemblages, highlighting the diversity of respondents and their research within the UK. We have summarised those examined in a regulatory framework (Appendix 2 - Table S3); however, as multiple options could be selected in the survey not all listed may be surveyed as part of statutory regulation. A gap analysis of each assemblage is precluded by the UKSI lacking complete annotation of associated habitat data.

Priority taxa and genes

The gap analysis has shown that invasive non-native species and individual species reported in the survey as priorities are already well represented in current reference libraries, for example 78% and 81% coverage respectively on BOLD (Tables 18 and 19). Furthermore, extending the list of focal species to all those on the JNCC list of taxa with conservation designations still results in 67% coverage on BOLD (Table 15). However, the overall coverage differs between taxonomic groups and this is compounded by the variability in quality assurance provided by current public reference libraries. While we did not examine the BOLD data for all UK species to the same level of detail as Macadam and others (2020), it is likely that their results on the species listed in the Pantheon database can be broadly extrapolated to the entire UK species list: a small percentage can be considered as having a high confidence in identification.

While the priorities need further investigation, especially around the assessment of assemblages, where they would require (a) defined need and (b) proven feasibility; the priorities would likely include (1) invasive species; (2) species monitored by end-users within a regulatory framework (eg legally protected species); and (3) component species within assemblages, which are also monitored within a regulatory framework and are wholly / partly identified to the species level with current methods.

Of the 1,680 priority species identified in the survey 140 do not have any public DNA sequence data and could be considered the highest priority (Table 20). There are 129 additional species with public DNA data but lacking a standard DNA barcode (defined as at least one sequence in a suitable reference library: Animalia = BOLD; Chromista = SILVA; Fungi = UNITE; Plantae = BOLD or PLANiTS), these species could be considered the second highest priority (Table 20). The third highest priority would need to be further analysed based on the confidence associated with each of the species for which there is already "barcode" data (Table 20). However, based on the analysis of Macadam and others (2020), it is likely the vast majority of these species will require new specimens and sequence data to have high confidence in their identification.

Finally if three strict requirements of (a) quality assured identifications; (b) based on UK specimens; and (c) vouchered in UK public institutions are to be met to enable uptake of DNA based identifications within a regulatory framework the reference library for UK taxa will need to be developed from the ground up, excluding vascular plants which already have data that broadly fit these criteria (see Chapter 2 gap analysis).

Table 20: Priority species identified by respondents to the survey, split into three priorities for gap filling: (1) those without any data in ENA; (2) those with data but no “barcode”; or (3) those with a “barcode” but the confidence has not been assessed.

			Priority		
Kingdom	**Phylum**	**Total**	**1** **no data**	**2** **no barcode**	**3** **confidence?**
Animalia		**1,076**	**76**	**71**	**929**
	Annelida	13	3		10
	Arthropoda	394	45	21	328
	Bryozoa	4		2	2
	Chordata	589	3	37	549
	Cnidaria	15	4	4	7
	Echinodermata	6		1	5
	Echiura	1		1	
	Mollusca	53	21	5	27
	Platyhelminthes	1			1
Chromista		**16**		**10**	**6**
	Ochrophyta	11		8	3
	Oomycota	5		2	3
Fungi		**172**	**38**	**20**	**114**
	Ascomycota	117	32	15	70
	Basidiomycota	54	6	5	43
	Chytridiomycota	1			1
Plantae		**416**	**26**	**28**	**362**
	Bryophyta	116	11	15	90
	Charophyta	10	1		9
	Hepaticae	3	2	1	
	Marchantiophyta	21	8	1	12
	Pteridophyta	11		1	10
	Rhodophyta	5		5	
	Tracheophyta	250	4	5	241
Total		**1,680**	**140**	**129**	**1,411**

The genes targeted for barcoding are relatively well established (Table 10) and while COI is the *de facto* barcode for animals, other genes such as 12S are increasingly being used by the community (eg fish identification). The development of whole genome reference libraries (ie the Darwin Tree of Life project) will mean that, once completed, any DNA fragment that is suitably variable could be used for the identification of UK eukaryotes. However, this ambitious project will likely take a decade or more and requires significant additional funding to deliver these 76,000 genomes. Furthermore, shorter standardised DNA barcodes are likely to continue as the mainstay of DNA-based identification due to the reduced costs associated with barcoding vs genome skimming.

Summary of barcode libraries

There were 80 survey responses covering a wide range of end user groups. Respondents were primarily from academic or government organisations, working across the UK. Surveys are primarily academic and have a focus on biodiversity assessment / conservation. Most respondents survey native species, while half survey invasive species. Freshwater and terrestrial habitats are primarily targeted, with most respondents surveying multi-species assemblages rather than singe species. Most end users require species level identifications.

A total of twelve public reference libraries were reported, including four bespoke databases. In addition, several private databases were reported. Reference libraries have mixed Quality Assurance, with only 50% of respondents' reporting the provision of voucher information. A total of 15 gene regions were reported, with most end users using COI. The perceived coverage of reference libraries was low, with only 14% of end users considering coverage to be better than 75% of their focal taxa.

A formal gap analysis using the UK Species Inventory (76,249 eukaryote species) found that half of all species have public DNA data, furthermore 50% of animals and 52% of plants have barcode data in BOLD, the most comprehensive reference library. When considering species with conservation designations, 67% have barcode data in BOLD, while 78% of currently recognised invasive non-native species have barcodes. There are 1,680 species identified as a priority group by end users, of which only 269 lack barcode data.

A separate analysis of BOLD data by Macadam and others (2020) shows that only 1.2% of species in the Pantheon database can be identified with "high confidence" using current BOLD data. Our analysis of all UK species in BOLD shows that only 4% of UK species are represented by specimens collected in the UK, and only 1.9% of species have UK specimens vouchered in UK institutions in order to facilitate easier verification of identifications.

Chapter 4: Opportunities and barriers to developing barcode libraries

The survey asked respondents to list the opportunities and barriers to developing barcode libraries. These were then refined in a workshop held at the Natural History Museum (12th March 2020) and are summarised below.

Summary of workshop and survey responses

Respondents and workshop participants agreed that the most important opportunities for developing barcode libraries can be broadly summarised as, in decreasing importance:

1. capabilities of people and organisations;
2. timeliness of this endeavour;
3. the emergence of new technologies and associated economies of scale.

Funding was identified as a more nuanced opportunity, but a significant barrier.

We have an amazing national capacity to identify organisms, especially in the amateur community. This capacity goes hand in hand with a global appetite to understand how the world is changing, and the technology to enable us to investigate this at scale.

Barriers to the development of barcode libraries fell into the following categories, again of decreasing perceived importance:

1. lack of funding;
2. no agreed standards on data quality, nor regulatory standards beyond the great crested newt;
3. small numbers of professionals with the necessary taxonomic expertise.

Additional funding is needed to take advantage of significant opportunities to build a national barcode library infrastructure and develop a world class set of data standards.

Opportunities and barriers are explored in more depth below. We then present some examples of relevant projects and partnerships that we should learn from, including existing projects that we can capitalise on.

Barriers identified from survey

The following barriers to the development of barcode libraries were highlighted as the most significant and are explored in turn, in decreasing order of perceived significance:

1. Funding
2. Expertise & capacity
3. Laboratory skills & equipment
4. Quality Control & Quality Assurance
5. Permits / access / legality / ownership
6. Network communication

Funding

The primary barrier to DNA barcoding of the UK fauna that was identified in the survey was the lack of specific funding for DNA barcoding. Concerted barcoding efforts require significant resourcing maintained over a number of years, typically provided by governments, as exemplified by the summary of European national barcoding initiatives (Table 21). Analysis of the funding provided to national campaigns in Europe and iBOL shows that the funders of DNA barcoding, in decreasing order of significance are:

1. Government direct or through independent government bodies
2. Non-profit foundations / charities
3. In house / local university funding
4. Private / commercial sector

Table 21: Summary of primary funding for national barcoding campaigns in Europe and iBOL.

Project	Primary funders
iBOL	● Canadian equivalent of UKRI ● Foundations: ○ Walder Foundation ○ Gordon and Betty Moore Foundation ○ Richard Lounsbery Foundation ● Government funding - including non-Canadian governments. ○ Environment Canada ○ Federal Ministry of Education and Research (Germany) ○ National Natural Science Foundation of China (China) ○ National Science Foundation (USA) ● Private / Commercial ○ Not listed, but includes discounted hardware and consumable costs
NorBOL	● Research councils of Norway ● Norwegian Biodiversity Information Centre ● Norway Taxonomy Initiative ● Partner institutions ● iBoL ● Universities and Heritage sites
GBOL	● Government ○ Federal Ministry for Education and Research (BMBF Germany)
SwissBOL	● Government ○ Federal Office for the Environment

Project	Primary funders
FinBOL	● Government ○ Ministry of the Environment ● Academy of Finland (similar to UKRI) ● Foundations (non-profit) ○ Finnish Cultural Foundation ○ Kone Foundation ● In house / university funding ○ University of Oulu

Expertise and capacity

Taxonomy

There are not enough people skilled in identification of many groups of organisms. There are also too few taxonomists working on many groups, which results in a very uncertain taxonomic backbone to underpin knowledge of UK flora and fauna. This has been emphasised in House of Lords enquiries into the state of taxonomy in the UK (House of Lords, 2001-02; 2002-03; 2007-08). While recent checklists are available for many taxa, thus far incorporation of these into the UK Species Inventory (UKSI) is piecemeal and results in a lack of a standardised UK checklist, which in turn makes recording more difficult for some groups and reduces the likelihood of correctly associating species concepts (eg between species distribution and DNA barcode data sources). An additional problem is the lack of standardised global checklists for many groups and poor integration between standard nomenclatures, although Fauna Europaea was an attempt to unify European animal checklists (Jong and others, 2014).

For many groups of organisms, we have little idea of the impact that cryptic species will have on overall species numbers. Museum collections of some groups of organisms contain misidentified specimens, or identifications that cannot be verified with current expertise. Integrated curation and critical identifications are needed to make the most of these collections, including as voucher material for DNA barcode identifications. Basically, for particular groups of organisms there is a lack of benchmark specimens to assure the quality of identifications.

Limited understanding of ecology

Few species are well-mapped, which limits our ability to target their collection for barcoding. A recent study by Outhwaite and others (2020) was able to use species occupancy trends for 5,000 species of insects, lichens and bryophytes, but that leaves c. 30,000 species within these groups for which there is still insufficient data.

Basic ecological traits for many species, such as phenology, food plants, host, etc., are lacking for most organisms. A high proportion of species are essentially known only from a small number of specimens in museum collections with limited associated data.

One questionnaire respondent put the case for the marine barcoding gap very succinctly: “for marine and brackish marine benthic habitats (as opposed to plankton), knowledge of biodiversity and availability of reference sequences is so limited (eg I estimate much less than 1% of UK species sequenced) that DNA-based identification is currently impossible except at the genus or family level, at which very little useful or interesting information is gained from metabarcoding."

Projects, such as Darwin Tree of Life (DToL), present opportunities to locate rare species for barcoding; however, collecting will be dependent on project funding, licensing, field craft and serendipity. Some species will need a significant input of time to arrange licenses, permissions etc.

Phylogenetic bias in expertise and interest (particularly from the amateur community) means that we have limited representation in reference libraries of some important functional groups, such as soil fauna, benthic invertebrates, and parasitoids. This hampers our ability to respond to global challenges such as food security, soil health, ecosystem complexity, etc. Taxon-specific grants are needed to develop identification tools for particular groups and to create these associated reference libraries.

Sample storage

There is a lack of clarity on the different preservatives used in the UK, such as Industrial Denatured Alcohol (IDA/IMS), Isopropanol and Ethanol, and their short- or longer-term effects on DNA preservation for DNA barcoding. In some cases, there are good data in research papers, but the results have not yet reached the necessary audiences. Similarly, how do preparation techniques and historical storage conditions of museum specimens affect DNA preservation? Dedicated molecular sample storage facilities are in short supply, although the CryoArks biobank initiative (https://www.cryoarks.org/) aims to bring together the diverse collections of animal frozen material found in museums, zoos, research institutes and universities across the UK to make them accessible to the UK’s research and conservation community. For physical vouchers, museums for the most part, are not in a position to voucher all samples that should be vouchered as there is a lack of infrastructure and staff.

Laboratory skills and equipment

Morphological identification can be carried out with relatively little investment (ie a hand lens, binoculars or a microscope), whereas barcoding requires a laboratory; however, newer portable sequencers, and the reducing costs of consumables, will likely make these technologies more accessible in the medium term.

DNA extraction and sequencing skill sets are lacking in most of the potential user community. Barcoding is still the preserve of wet lab people, therefore there has been limited engagement from the morphological community. However, if knowledge is shared and, importantly, samples are used for barcoding projects rapidly (ie don’t sit around gathering dust) then this is less of a barrier to engagement.

Bioinformatics support: staff time is needed to manually edit sequence data and upload these to public databases. There is a lack of automated workflows, in both the wet lab and in data management. It was pointed out in the workshop, though, that user friendly bioinformatics tools (eg MBRAVE) are dismantling these barriers.

Quality Control (QC) and Quality Assurance (QA)

Many issues here boil down to a lack of curation and minimum data standards in DNA databases. Multiple public repositories lack sufficient data standards, both for voucher specimens and associated sequences, for adequate quality assurance. Many small-scale projects are providing data to public repositories without adequate quality control and assurance, for example by using reverse taxonomy rather than independent identifications by experts. Public databases can be poorly curated and it is difficult to get errors fixed, although it is important to note that some databases do insist on strict quality control, eg RDP and SILVA are resources for quality checked and aligned ribosomal RNA data.

Due to the lack of UK based specimens in reference libraries there is an over-reliance on non-UK specimens that are not readily available for examination, often using reverse taxonomy for identification. There is a need for accreditation and validation, which should be led by statutory agencies.

Permits / access / legality / ownership

To be accessioned in museums, voucher specimens need to be accompanied by proof of legality or ownership. Sanger and other institutions involved in releasing large quantities of genomic data are restricted to processing material that meets a high threshold of due diligence in terms of legality and ethics. Statutory agencies are limited in their ability to issue wide-ranging permits for collecting organisms across their range, or wide ranges of organisms. Sampling a diversity of birds and mammals, for example, requires many permit applications. Similarly, a national collecting campaign would require many individual permits for collecting on different sites.

Network / Communication

Respondents highlighted that the UKDNA working group would have more impact with improved structure and resources. There is clearly a need for a coordinated national barcoding campaign, as highlighted by the House of Lords review (2007-08).

Opportunities identified from the survey

The opportunities are considerable and can be broken down into essentially:

1. People and Organisational capabilities
2. Funding opportunities
3. Existing networks
4. Availability of identified specimens for barcoding
5. Emerging technologies and economy of scale
6. Timeliness

People / Organisations

The UK has the museum and herbarium infrastructure to support large scale barcoding and vouchering, with some additional investment. Statutory Nature Conservation Bodies (SNCB) could support the collection of specimens on Sites of Special Scientific Interest (SSSI) and National Nature Reserves, including staff on site and staff with taxonomic expertise. However, in order to do this to a significant degree funding would be required to cover staff time. The SNCBs also have the ability to influence recording societies and get their assistance and expertise in contributing well identified vouchers and possibly testing methods. Although there is a shortage of professional taxonomists and curators, the UK is perhaps particularly blessed with expertise in the amateur community and in the ecological surveying sector. Much of the expertise and many publications on identification for groups such as bees, some groups of flies, some plants and fungi, lies in this sector. The UK has arguably the most comprehensive community of amateur naturalists and recording schemes globally; for example, 10 million records are added per year to the NBN atlas and UK recording schemes cover a considerable range of organisms.

Funding opportunities

From a regulatory perspective funding for barcoding development is possible if (1) benefits outweigh the costs for existing monitoring, and / or (2) a strong case can be made that key evidence gaps can be filled (eg soil biota). There is a diverse array of likely funding for DNA barcoding, including:

- The DEFRA Centre of Excellence
- Programme funding within Statutory Nature Conservation Bodies
- Funding from the Darwin Tree of Life: several of the consortium partners have barcoding costs as part of their project plans, to cover a variety of animal, plant and fungal groups.
- For certain use cases or groups of organisms, industry funding (eg from the food industry) is available for projects.
- Government interest in regulation, pathogens and risk can release funding for barcoding tools. For example, there is considerable interest in the regulatory and fish-farming sectors to develop alternatives methods for assessing / monitoring marine benthos.
- Internal funding from organisations (universities, agencies, herbaria, museums, etc.) is available for pilot studies.
- Additional funding agencies (eg UKRI) are listed in the summary of potential funders section.

Existing networks

The following networks are just some of the major existing or potential contributors to a comprehensive UK barcoding reference library, however coordination of these networks is key:

- The UKDNA working group (in particular the reference libraries technical group).
- National Biodiversity Network (NBN)
- Biological Records Centre (BRC)
- The State of Nature partnership

- The JNCC UK Terrestrial Evidence Partnership of Partnerships
- The Chartered Institute of Ecology and Environmental Management (CIEEM)
- Teams of local naturalists who are interested in applying the techniques in habitat restoration and monitoring, identification of fauna and flora, etc.
- Major consortia now need this data and will rely on barcoding (eg DToL). These networks can leverage funds and enthusiasm, with major goals as clear destinations.
- The development of a SNH DNA framework and the Sottish DNA Hub present opportunities away from the usual South-East of England bias.

Availability of identified specimens for barcoding

National collections and herbaria hold vast collections of verified material that can be used to produce reference barcode collections. While older material often requires different techniques utilising shorter DNA fragments, there can be greater opportunities for recently collected, authoritatively identified material of some taxonomic groups at these institutions. For example the FreshBase project is currently building a genomic resource of expertly identified freshwater macroinvertebrates, assembled through a cross-disciplinary initiative and vouchered at NHM. Current PhD projects are sampling various groups and habitats, such as sediment DNA and aquatic macrophytes at sites in the UK Upland Waters Monitoring Network (http://awmn.defra.gov.uk/). DNA BioBlitzes, such as the Ainsdale 2019 event, provide the opportunity to engage multiple stakeholders and collect high quality voucher material and DNA barcodes over very short timescales.

Emerging technologies and economy of scale

New hardware, such as the Oxford Nanopore MinION, have the potential to democratise sequencing by reducing start-up costs. Furthermore the latest sequencing technology in established sequencing facilities enables many more samples to be sequenced at a time, significantly reducing per sample costs. Advances in the analysis of DNA from mixed samples can have higher sensitivity than traditional methods, including population / haplotype identification, should be more objective than morphological identification, and can be standardised more easily.

Emerging research and conservation priorities

The development of a national barcoding project would be very timely, given the emergence of global and local initiatives to better understand and conserve biodiversity in the face of a biodiversity crisis (IPBES, 2019). Increased attention on the biodiversity crisis has highlighted the lack of data on trends in species abundance. Furthermore climate change and shifting patterns of global trade are both contributing to increases in invasive species and problems with pest management. UK legislation means that DNA-based monitoring is already in use, for example newt surveys, paving the way for further development of the field.

Summary of current / past funding

A summary of the funding sources for barcoding highlighted in the survey responses is below (Table 22).

Table 22: Current funding sources highlighted in the survey.

Region	Source	Funding
UK	Government	Direct from Government Government agencies / regulators Research councils (ie UKRI, NERC, BBSRC) Defra DNA Centre of Excellence Defra EU Exit innovation funding Ministry of Housing, Communities & Local Government NERC National Capability Innovate UK
	Societies	British Ecological Society
	Public sector grants	No specific examples given
	Private	Water companies End users (ie ecologists and citizen scientists) Private contracts
	Trusts	Wellcome Trust Leverhulme Trust
	Institutional	Internal (institutional) funding PhD project funds (including DTPs)
Europe	EU funding	European Agricultural Fund for Rural Development EU Synthesys awards

Summary of relevant projects and partnerships

Funding for mass barcoding projects in the UK has been ad hoc, mostly grant-funded. However, some projects have been ambitious in scope. The examples below are funded by a mix of government agencies, charitable trusts and individual donors.

Examples of funded large-scale barcoding projects in the UK

The Welsh angiosperms and gymnosperms (native and archaeophytes) were barcoded through funding from a mixture of sources: National Botanic Garden of Wales, National Museum Wales, Welsh Government, Countryside Council for Wales, and from donations from the public. This project had a well-defined taxon scope, an ambitious but feasible number of species (1,143), and all specimens were vouchered in herbaria, summarised by de Vere and others (2012).

Darwin Tree of Life has received £9.4 million of funding from the Wellcome trust to sequence and assemble the genomes of 2,000 UK eukaryotes (2019 – 2022), with 10 organisations collaborating closely to achieve this. DNA barcoding is an essential tool for identification verification within the workflow and the NHM plans to barcode sequence at least 10,000 metazoan individuals within the pilot project. Specimens are being vouchered in the NHM collections. Plant barcoding is being led by RBGE.

Defra funded a taxonomic fellowship to support the National Pollinator Strategy. This project, based at the Natural History Museum, DNA-barcoded 60.1% of the UK bee species and 19.6% of the hoverflies. Additionally, CO1 data were assembled from BOLD to produce a barcode library for 92.4% (255 species) of UK bees. Reference specimens are vouchered at NHM (DEFRA, 2016).

Smaller scale UK barcoding projects

Some UK-focused projects fall within the remit of the International Barcode of Life and are funded by iBoL, for example Charles Godfray's work (Oxford University) on parasitoids of leaf-mining Diptera. Other similar projects are not funded by iBoL but are paid for on a plate by plate basis by the researchers, eg Mark Shaw's (National Museums of Scotland) studies of Lepidoptera parasitoids. Both projects add to our barcode voucher libraries and contributing primary taxonomic knowledge of the UK fauna, but at the relatively small scale of hundreds of specimens and emphasise the opportunistic and unreliable nature of funding for these activities.

Two ambitious ecological projects have received funding recently - Brilliant Butterflies and Urban Nature - that include DNA metabarcoding as a means of assessing invertebrate communities and their response to changes in land usage. The generation of reference libraries is a part of this, though the scale is not clear as they have just begun. These initiatives have been funded by the People's Postcode Lottery (Brilliant Butterflies) and National Lottery Heritage Fund, together with a variety of charitable donors (Urban Nature).

International projects of relevance to a UK barcoding campaign

The International Barcode of Life (iBOL) consortium was established in 2008 and has overseen the completion of one major program to barcode 500,000 species, at a cost of $150 million. The current iBOL consortium consists of 32 member and 8 associate member nations, including the UK through RBGE and NHM, and has recently launched a second phase project (BIOSCAN), which will extend barcode coverage to 2.5 million species by 2026 at an estimated cost of $180 million including $50 million from in-kind specimen collection and identification. BIOSCAN will provide a platform for the third phase of iBOL, known as the Planetary Biodiversity Mission, a research initiative that aims to deliver a comprehensive understanding of multicellular life by 2045. Funding for iBOL is provided by research organizations in partner countries, the Canadian government (Environment Canada) and Canadian UKRI equivalent, several other national government funders: Federal Ministry of Education and Research (Germany), National Natural Science Foundation of China (China) and the National Science Foundation (USA). Furthermore, several large foundations support iBOL, including the Walder Foundation, Gordon and Betty Moore Foundation and the Richard Lounsbery Foundation.

There are also several independent national barcoding campaigns across Europe including Norway (NorBOL), Germany (GBOL and BFB), Switzerland (SwissBOL), Finland (FinBOL), Austria (ABOL) and Croatia (CroBOL). There are incipient barcoding consortia developing in several European countries including Romania, Turkey, Poland and Belarus. A key commonality in these national barcoding campaigns is that their funding is primarily from government sources and foundations.

In addition to national barcoding campaigns there are thematic and habitat specific networks, the most relevant being the EU funded DNAqua-Net COST action (https://dnaqua.net/). Running since 2017, with a focus on aquatic habitats (freshwater and marine) and the implementation of the Water Framework Directive and Marine Strategy Framework Directive across Europe, DNAqua-Net consists of 400 members in 49 countries, including the UK, and is subdivided into five working groups:

- WG1 – DNA Barcode References
- WG2 – Biotic Indices & Metrics
- WG3 – Field & Lab Protocols
- WG4 – Data Analysis & Storage
- WG5 – Implementation Strategy & Legal Issues

Summary of potential funders

As highlighted previously the primary barrier to DNA barcoding of the UK fauna is the lack of large scale coordinated funding, maintained over a number of years which is typically provided by governments (see Table 19 for summary of funding for other national barcoding initiatives). Below several relevant large-scale funding sources are summarised:

BBSRC BBR

Link: https://bbsrc.ukri.org/funding/filter/2019-bioinformatics-biological-resources-fund/

The Bioinformatics and Biological Resources (BBR) Fund aims to facilitate the establishment, maintenance and enhancement of high-quality bioinformatics and biological resources to support the UK bioscience research community. The indicative budget for the call is up to £6 million, subject to the quality of applications received.

Wellcome Trust

The funding of Darwin Tree of Life was a significant departure from Wellcome's usual foci. Whether that funding continues will depend on the success of DToL over the next 30 months and the demonstration of this programme's societal relevance.

NERC National Capability

Link: https://nerc.ukri.org/funding/available/nc-funding/

National capability (NC) funding describes the element of NERC-funded activity directly procured by NERC due to a combination of its scale and complexity. These features result in a need for NC provision with a critical mass of size and budget that makes direct procurement the only practical option.

NC comprises:

- NC-science, which integrates over at least national and decadal timescales.
- NC-large-scale research infrastructure.
- smaller-scale NC-services, facilities and data that provide a service to the environmental science research community.
- delivering NC-national and public good activities, which comprise advice to government departments and wider information to the public at large.

Budgets for these four NC categories are determined by council as part of its business planning to deliver NERC's strategy.

NERC Strategic research

Link: https://nerc.ukri.org/funding/available/programmes/

NERC's strategic research funding supports research into environmental areas of major economic and societal importance. It aims to address key science challenges and priorities for the 21st century. NERC plans strategic research funding opportunities via its Science Committee, which uses ideas from the community on where strategic research should be targeted. Once strategic research funding opportunities have been agreed, the community are asked to respond with grant proposals.

There are three types of strategic research funding:

- Highlight topics: funding opportunities once a year, requesting proposals for large-size grants to address one of a defined list of strategic topics.
- NERC strategic programme areas: there will be one or more funding opportunity per programme, requesting proposals for grants to address specific aspects of the programme's objectives.
- Partnerships and Opportunities: when NERC contributes to strategic activities led by other funders, our partners may publish and manage the funding opportunity.

One of the first highlight topics funded by NERC, in 2015, was ‘eDNA: a tool for 21st century ecology’. This, and some other highlight topics, rely on the existence of reliable barcode libraries, but funding for generation of this infrastructure has not been implicit in the funding calls.

UKRI Strategic Priorities Fund

Link: https://www.ukri.org/research/themes-and-programmes/strategic-priorities-fund/

The Strategic Priorities Fund (SPF) is being led by UKRI to:

- drive an increase in high quality multi and interdisciplinary research and innovation.
- ensure that UKRI’s investment links up effectively with government research priorities and opportunities.
- and ensure the system responds to strategic priorities and opportunities.

Relevant themes from wave 1 (the current call) include Landscape Decisions: Developing a new understanding to help individuals, communities and country make the best choices regarding land use in the UK.

The second wave (to be announced) includes Sustainable Management of Marine Resources: This programme will ensure that the UK realises sustainable societal and economic benefits through better management of the UK's marine resources.

Charitable trusts / Lotteries

The People's Postcode Lottery and the National Lottery Heritage Fund will potentially be significant funders of barcode reference libraries through their funding of landscape conservation projects. Smaller charitable trusts continue to contribute through particular projects.

Chapter 5: Action plan for developing barcode libraries and filling priority gaps

Workshop overview

A workshop was held at the Natural History Museum (12th March 2020) to summarise the results of the survey and consultation, elicit feedback on the draft report and outline an action plan (see Agenda: Appendix 5). Following the workshop, a draft action plan is outlined below.

Recommendations

In the workshop it was agreed that the UK should establish a Barcode of Life project to develop a standardised, open access, vouchered UK reference library to facilitate reliable species identification using DNA data. Several additional recommendations are outlined below to enable this project, divided into (1) project initiation and governance and (2) interim work which can be tackled before large scale funding is secured.

Project initiation and governance

1. Setup a project steering group in consultation with:
    a. National iBoL member organisations (NHM and RGBE);
    b. End user groups (Defra CoE, UKDNA steering group, Scottish DNA Hub, NBN, UKCEH);
    c. National repositories (Museums and Herbaria);
    d. National sequencing facilities (eg Wellcome Sanger Institute).

2. Steering group to oversee development of:
    a. 5-year plan;
    b. Business case for large-scale funding;
    c. Engage with potential funders to secure funding
    d. Identify and establish “Task & Deliver” groups, including:
        i. Communication and engagement (including a register of taxon experts);
        ii. Data standards;
        iii. Priority taxa;
        iv. Field standard operation procedures (SOPs) / workflows;
        v. Lab SOPs / workflows;
        vi. Website and data portal (eg GBOL, NorBOL).

Interim work

In the interim, before large scale funding is secured, the following tasks should be initiated:

1. Refine and publish the list of priority taxa and gaps across taxonomic groups in consultation with end users.
2. Develop communication/engagement plans for working with the wider taxonomic and recording communities.
3. Develop SOPs for field and lab work.
4. Begin to tackle the highest priority taxa.

Milestones

The steering committee would agree the project milestones; however, we have suggested several below that we would anticipate:

1. Formation of project steering group with regular meetings each year
2. Business case submitted to government/funders
3. Funding secured
4. Agreed framework for prioritising taxa and initial priority taxon list published as a report
5. Agreed annual collecting / barcoding targets
6. Agreed data standard(s) published as a report
7. Website made live
8. Relevant SOPs agreed and published as reports
9. Data portal added to website and linked to appropriate international repositories

Resources

Creating a coordinated network of organisations to deliver a national barcoding programme will require resourcing, including, but not limited to:

1. Coordination
   a. Staff time / additional staff
   b. Funds for regular meetings
   c. Funds for website development
2. Sample collection
   a. Register of taxon experts
   b. Support from permitting authorities
   c. Standardized consumables (vials, preservatives, labels etc)
   d. Support for sample collection and identification where this is not provided in-kind
3. Repositories
   a. Staff time / additional staff
   b. Consumables / curatorial supplies
   c. Imaging equipment
4. Sequencing Facilities
   a. Staff time / additional staff
   b. Consumables
   c. High-throughput sample processing equipment
5. Bioinformatics
   a. Staff time / additional staff

Data attribution

The gap analysis used the JNCC "Conservation Designations for UK taxa", which contains JNCC/NE/NRW/SNH/NIEA data © copyright and database right 2019.

References

BANCHI, E., AMETRANO, C.G., GRECO, S. STANKOVIĆ, D., MUGGIA, L., PALLAVICINI, A. (2020) PLANiTS: a curated sequence reference dataset for plant ITS DNA metabarcoding. Database 2020.

BIGGS J., EWALD N., VALENTINI A., GABORIAUD C., GRIFFITHS R.A., FOSTER J., WILKINSON J., ARNETT A., WILLIAMS P., DUNN F. (2014) Analytical and methodological development for improved surveillance of the Great Crested Newt Defra Project WC1067. Freshwater Habitats Trust, Oxford.

COSTELLO M.J., WILSON S., HOULDING B. (2012) Predicting Total Global Species Richness Using Rates of Species Description and Estimates of Taxonomic Effort. Systematic Biology, 61: 871.

DE VERE N., RICH T.C.G., FORD C.R., TRINDER S.A., LONG C., MOORE C.W., and others. (2012) DNA Barcoding the Native Flowering Plants and Conifers of Wales. PLoS ONE, 7(6): e37945.

DEFRA (2016) Taxonomic fellowship to support the National Pollinator Strategy. EVID4 Evidence Project Final Report (Rev. 10/14) 27pp.

GUILLOU, L., BACHAR, D., AUDIC, S., BASS, D., BERNEY, C., BITTNER, L., BOUTTE, C. and others. 2013. The Protist Ribosomal Reference database (PR2): a catalog of unicellular eukaryote Small Sub-Unit rRNA sequences with curated taxonomy. Nucleic Acids Research, 41: D597–604.

HAYHOW D.B., EATON M.A., STANBURY A.J., BURNS F, KIRBY W.B., BAILEY, N (2019) The State of Nature 2019. The State of Nature partnership.

HEBERT, P. D., CYWINSKA, A., BALL, S. L., & DEWAARD, J. R. (2003) Biological identifications through DNA barcodes. Proceedings. Biological sciences, 270 (1512): 313–321.

House of Lords Science and Technology Committee, Fourth Report, Session 2001-02, *What on Earth? The threat to the science underpinning conservation*. HL Paper 118(i). https://publications.parliament.uk/pa/ld200102/ldselect/ldsctech/118/11801.htm

House of Lords Science and Technology Committee, Third Report, Session 2002-03, *What on Earth? The threat to the science underpinning conservation: the government's response and the committee's commentary.* HL Paper 130(i). https://publications.parliament.uk/pa/ld200203/ldselect/ldsctech/130/13001.htm

House of Lords Science and Technology Committee, Fifth Report, Session 2007-08, *Systematics and Taxonomy: Follow-up.* HL Paper 162. https://publications.parliament.uk/pa/ld200708/ldselect/ldsctech/162/162.pdf

IPBES (2019) Global assessment report on biodiversity and ecosystem services of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services. E. S. Brondizio, J. Settele, S. Díaz, and H. T. Ngo (editors). IPBES secretariat, Bonn, Germany.

JONG Y. DE, VERBEEK M., MICHELSEN V., BJØRN P. DE P., LOS W., STEEMAN F., and others. (2014) Fauna Europaea – all European animal species on the web. Biodiversity Data Journal, 2: e4034.

JNCC (2020) Conservation Designations for UK taxa. Available from: https://jncc.gov.uk/our-work/conservation-designations-for-uk-taxa/ (Downloaded: 5 February 2020)

LARSEN B.B., MILLER E.C., RHODES M.K., WIENS J.J. (2017) Inordinate Fondness Multiplied and Redistributed: The Number of Species on Earth and the New Pie of Life. The Quarterly Review of Biology, 92: 229–265.

LEESE F., ALTERMATT F., BOUCHEZ A., EKREM T., HERING D., MEISSNER K., and others. (2016) DNAqua-Net: Developing new genetic tools for bioassessment and monitoring of aquatic ecosystems in Europe. Research Ideas and Outcomes, 2: e11321.

MACADAM C., ROBINS J., THOMPSON T. (2020) 2020 Gap Analysis of the BOLD Database for Key English Invertebrates. Buglife Report, Peterborough, UK.

MACHIDA, R., LERAY, M., HO, S-L., KNOWLTON, N (2017) Metazoan mitochondrial gene sequence reference datasets for taxonomic assignment of environmental samples. Scientific Data, 4: 170027.

MORA C., TITTENSOR D.P., ADL S., SIMPSON A.G.B., WORM B. (2011) How Many Species Are There on Earth and in the Ocean? PLoS Biology, 9: e1001127.

NILSSON R.H., LARSSON K-H., TAYLOR A.F.S., BENGTSSON-PALME J., JEPPESEN T.S., SCHIGEL D., KENNEDY P., PICARD K., GLÖCKNER F.O., TEDERSOO L., SAAR I., KÕLJALG U., ABARENKOV K. (2018) The UNITE database for molecular identification of fungi: handling dark taxa and parallel taxonomic classifications. Nucleic Acids Research, 47 (D1): D259–D264.

OUTHWAITE, C.L., GREGORY, R.D., CHANDLER, R.E. and others. (2020) Complex long-term biodiversity change among invertebrates, bryophytes and lichens. Nature Ecology and Evolution, 4: 384–392.

QUAST C., PRUESSE E., YILMAZ P., GERKEN J., SCHWEER T., YARZA P., PEPLIES J., GLÖCKNER F.O. (2013) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Research 41 (D1): D590-D596.

RATNASINGHAM S., HEBERT P.D.N. (2007) BOLD: The Barcode of Life Data System (http://www.barcodinglife.org). Molecular Ecology Notes, 7: 355–364.

RATNASINGHAM S., HEBERT P.D.N. (2013) A DNA-Based Registry for All Animal Species: The Barcode Index Number (BIN) System. PLoS ONE 8 (7): e66213.

RIMET F., GUSEV E., KAHLERT M., KELLY M., KULIKOVSKIY M., MALTSEV Y., MANN D., PFANNKUCHEN M., TROBAJO R., VASSELON V., ZIMMERMANN J., BOUCHEZ A. (2019) Diat.barcode, an open-access curated barcode library for diatoms. Scientific Reports, 9: 15116.

WEBB J., HEAVER D., LOTT D., DEAN H.J., VAN BREDA J., CURSON J., HARVEY M.C., GURNEY M., ROY D.B., VAN BREDA A., DRAKE M., ALEXANDER K.N.A., FOSTER, G. (2018) Pantheon - database version 3.7.6 (https://www.brc.ac.uk/pantheon/)

YILMAZ P., PARFREY L.W., YARZA P., GERKEN J., PRUESSE E., QUAST C., SCHWEER T., PEPLIES J., LUDWIG W., GLÖCKNER F.O. (2014) The SILVA and "All-species Living Tree Project (LTP)" taxonomic frameworks. Nucleic Acids Research 42: D643-D648