Project:	Synthesis of systematic resources
Project acronym:	SYNTHESYS3
Grant Agreement number:	312253
Workpackage:	2: Improving collections management and enhancing accessibility
Deliverable number:	2.6: Strategic priorities for barcoding
Deliverable title:	Strategic Priorities for DNA Barcoding Natural History Collections
Deliverable author(s):	Peter M. Hollingsworth; David Harris; Michelle Hart; Gabi Dröge; Isabel Rey; Inés Álvarez Fernández; Javier Fuertes Aguilar; Beatriz Alvarez Dorda; Tim Fulcher; Stefanie Krause; Thomas von Rintelen; Michel Guiraud; Carole Paleco; Thierry Backeljau; Patrick Grootaert; Barbara Gravendeel, René Dekker; Elisabeth Haring; Karin Wiltschke-Schrotta; Nikolaus Szucsich; Luise Kruckenhauser; Vacek František; Jiří Kvaček; Jackie Mackenzie-Dodds; Stephen Russell; Garin Cael; Danny Meirte; Sean Prosser; Paul Hebert
Date:	October 2016

Strategic Priorities for DNA Barcoding Natural History Collections

Authors: Peter M. Hollingsworth[1*]; David Harris[1]; Michelle Hart[1]; Gabi Dröge[2]; Isabel Rey[3]; Inés Álvarez Fernández[4]; Javier Fuertes Aguilar[4]; Beatriz Alvarez Dorda[3]; Tim Fulcher[5]; Stefanie Krause[6]; Thomas von Rintelen[6]; Michel Guiraud[7]; Carole Paleco[8]; Thierry Backeljau[8]; Patrick Grootaert[8]; Barbara Gravendeel[9], René Dekker[9]; Elisabeth Haring[10]; Karin Wiltschke-Schrotta[10]; Nikolaus Szucsich[10]; Luise Kruckenhauser[10]; Vacek František[11]; Jiří Kvaček[11]; Jackie Mackenzie-Dodds[12]; Stephen Russell[12]; Garin Cael[13]; Danny Meirte[13]; Sean Prosser[14]; Paul Hebert[14]

[1]Royal Botanic Garden Edinburgh, 20A Inverleith Row, Edinburgh, EH3 5LR, UK

[2]Botanischer Garten und Botanisches Museum Berlin-Dahlem, Freie Universität Berlin, Königin-Luise-Straße 6-8, 14195, Berlin, Deutschland

[3]Museo Nacional de Ciencias Naturales, CSIC, José Gutiérrez Abascal 2, 28006 Madrid, España

[4]Real Jardín Botánico de Madrid, Real Jardín Botánico, CSIC, Plaza de Murillo 2, 28014 Madrid, España

[5]Royal Botanic Gardens Kew, Richmond, Surrey, TW9 3AE, UK

[6]Museum für Naturkunde, Leibniz-Institut für Evolutions- und Biodiversitätsforschung Invalidenstraße 43, 10115 Berlin, Deutschland

[7]Museum National d'Histoire Naturelle 57, rue Cuvier75005 Paris, France

[8]Royal Belgian Institute of Natural Sciences**.** Museum of Natural Sciences, Vautier Street 29, 1000 Brussels, Belgium

[9]Naturalis Biodiversity Center, Vondellaan 55, Postbus 9517, 2300 RA Leiden, Netherlands

[10]Naturhistorisches Museum Wien, Burgring 7, 1010 Vienna, Austria

[11]National Museum Prague, Václavské nám. 68, 115 79 Praha 1, Czech Republic

[12]Life Sciences, Natural History Museum, Cromwell Road, London DE7 5BD, UK

[13]Royal Museum for Central Africa, Leuvensesteenweg 13, 3080 Tervuren, Belgium

[14]Biodiversity Institute of Ontario, University of Guelph, 50 Stone Road, E Guelph, Ontario, Canada N1G 2W1

*Correspondence: p.hollingsworth@rbge.org.uk; orcid.org/0000-0003-0602-0654

Contents

Summary and key recommendations 4
DNA barcoding overview 5
The role of natural history collections in DNA barcoding 6
Large scale barcode studies of natural history collections 7
Factors influencing recoverability of Sanger sequenced barcodes from natural history collections 7
Specimen age 7
Body size 8
Preservation method 9
Target locus length 9
DNA barcoding and next generation sequencing technologies 9
Strategic priorities for DNA barcoding natural history collections 11
Scientific priorities 11
Societal needs 12
Practical considerations influencing the use of natural history specimens in DNA barcoding projects 13
Supporting access to preserved specimens for DNA barcoding 14
Ongoing major initiatives among SYNTHESYS3 partners 14
References 14
Annexe 1. Case studies 17
A.1.1. Mollusc DNA in museum collections 17
A.1.2. *Delias* butterfly radiation on New Guinea 18
A.1.3. Sequencing nuclear genes from herbarium specimens 19
A.1.4. DNA barcoding of CITES protected species 20

Summary and key recommendations

DNA barcoding involves the standardised use of DNA sequences to tell species apart. This report explores the feasibility and strategic use of natural history collections for DNA barcoding and wider genomic approaches for discriminating among species. The report is based on discussions at the *International Barcode of Life Conference* in Guelph, Canada (August 2015) and a workshop at the Royal Botanic Garden Edinburgh, UK (September 2015), supported by further dialogue and literature surveys.

Recovery of DNA sequence data from preserved natural history collections has traditionally been hampered by DNA degradation. However, large-scale studies have recently demonstrated effective recovery of DNA barcode data from natural history collections even using traditional Sanger sequencing approaches. Such effective access to expertly verified material in natural history collections represents a strategically important contribution to constructing a DNA barcode reference library for life on earth.

The development of new technologies focused on massively parallel short-read sequencing is enabling a further step change in recovery of nucleotide sequences from preserved specimens. Both target-capture methods and shotgun sequencing are resulting in rapid production of large genetic datasets from preserved museum and herbarium specimens suitable for species discrimination.

The authors of this report identified the following (non-exhaustive) strategic priorities for DNA barcoding of natural history collections:

- Sequencing of type specimens to formalise the link between scientific names and sequences
- Targeted sequencing to fill gaps in phylogenetic coverage (e.g. genera with no sequenced specimens)
- DNA barcode surveys of large genera to facilitate taxonomic revisions of difficult groups
- DNA barcoding of endangered species to support enforcement of wildlife crime
- DNA barcoding of key pollinator species to enhance understanding of ecosystem service provision / food security
- Sequencing invasive non-natives, pest and pathogens to support control and management
- DNA barcoding species impacting human health to support effective diagnostics and control

The selection of samples and the design of barcoding projects for natural history collections also involves important practical considerations: (a) Recently collected material typically has high resolution spatial coordinates, rich meta-data and close links to ongoing projects. There is clear efficiency benefit to targeting efforts toward such recently collected and actively used material. (b) Sequencing type-specimens can be extremely useful in linking names to genetic data, but this value is only realised if the regions being sequenced are diagnostic at the species level (an assumption not satisfied in e.g. many plant species). Careful attention is required to promote genetic access to type material, whilst avoiding tissue sacrifice for uninformative assays. (c) Projects using natural history collections have often focused on particular taxonomic groups or geographical areas. Addressing some of the pressing applications of DNA barcoding (e.g. diagnostics for invasive species, pests and pathogens) will require new large-scale inter-institutional collaborative projects drawing on expertise and samples sets with very wide phylogenetic coverage and geographical spread.

DNA barcoding overview

DNA barcoding involves the standardised use of one or a few DNA regions to tell the world's eukaryotic species apart (Box 1). The need for the approach is driven by the large uncertainty as to the total number of species on earth, the small proportion of those that have yet been described, and the widespread and routine difficulties encountered in identifying unknown specimens to already known species. There is thus a strong scientific imperative and societal need for improved mechanisms to tell species apart.

DNA barcoding was first proposed in 2003 [1] and has since gained widespread uptake across the globe (Fig.1; Box 1). As of June 2016, there are barcode records from more than 500,000 species from 5,002,218 specimens, and more than one thousand barcoding papers were published in 2015 alone [2]. Key working principles of DNA barcoding include (i) minimalism: using a minimal set of loci to maximise efficiency given the large scale of the task; (ii) standardisation: community adoption of an agreed set of barcode markers to enable individual projects to contribute towards a shared global resource; and (iii) quality control: establishment of data standards and links between barcode sequences and voucher specimens to maximise the reliability of the resource.

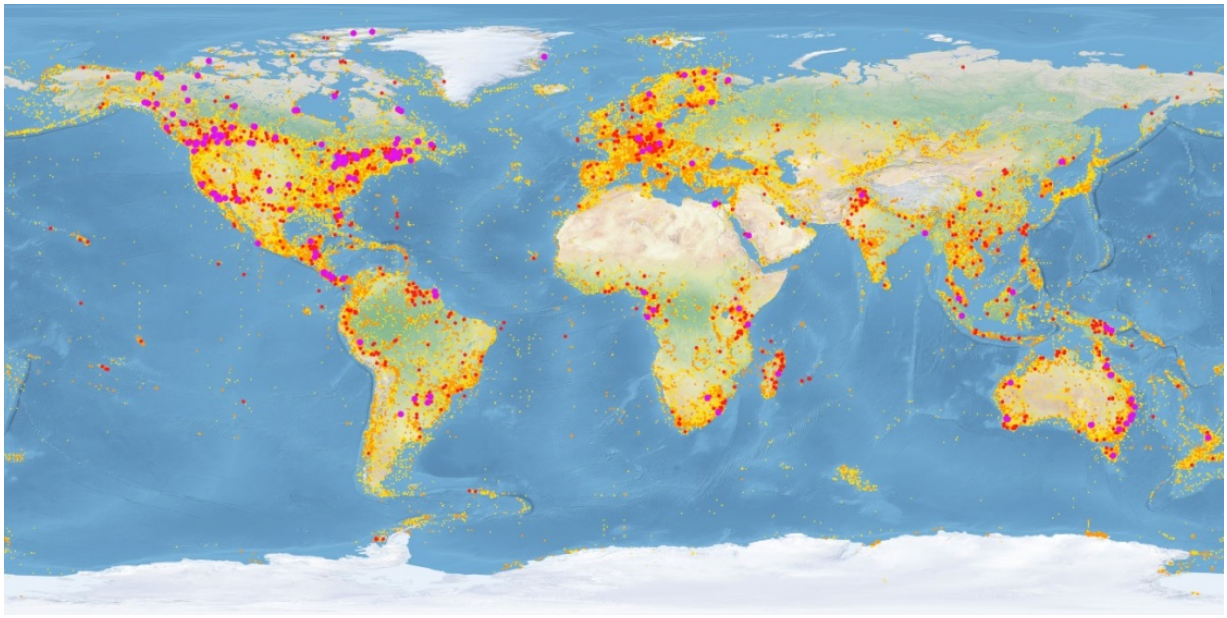


Fig. 1. Heat map summarising the global distribution of 5 million DNA barcode records in the Barcode of Life Datasystem (BOLD) as of June 2016.

Box 1 DNA barcode markers for different major groups

Animals: The standard barcode for animals is a 648 bp fragment of the mitochondrial CO1 region [1]. The CO1 barcode region shows high levels of discriminatory power in many animal groups, and the sequence clustering algorithms based around Barcode Identification Numbers show extremely high concordance between sequence clusters and existing species taxonomies [3, 4].

Plants: Researchers focusing on specimen-based barcoding typically use 2-3 plastid regions (*rbcL, matK, trnH-psbA*) along with the internal transcribed spacers of nuclear ribosomal DNA (nrDNA ITS) [5]. In contrast, researchers focusing on meta-barcoding have favoured the plastid *trnL* intron which contains a short mini-barcode flanked by extremely conserved primer sites which make it particularly well suited to recovery from degraded samples [6]. The resolving power of these loci, singularly or in combination is typically lower than that of CO1, with plant barcodes often shared among closely related species [7].

Fungi: The coordinated use of ITS sequence data to discriminate between fungal species preceded formal proposals for global DNA barcoding efforts. The ITS region was adopted as the fungal barcode in 2012 with the expectation that it would be supplemented with other loci on a clade-by-clade basis where ITS does not provide species level resolution [8, 9].

Protists: A segment of the 18S rDNA has been adopted as a 'backbone barcode' for protists to be augmented with clade-specific markers to provide species-level resolution in different protist groups [10].

The role of natural history collections in DNA barcoding

Species names provide the link between an individual organism, and the accumulated knowledge of the biological attributes of the species to which it belongs. A key step in building the global DNA barcode reference library involves sequencing expertly verified material. Three major tissue sources have been used for constructing this reference library:

- Natural history museums and herbaria represent a resource of billions of samples and are the repository for type and other expertly verified specimens. Species determinations are frequently updated and individual specimens benefit from annotations spanning decades or centuries. However, despite the intrinsic richness of this resource, obtaining sequence data from museum/herbarium specimens can be challenging due to DNA degradation.
- Freshly collected field samples are generally well suited for DNA sequencing, but require further work to obtain expert identifications. This can be straightforward for samples that are the subject of active research programmes, or small sample sets from relatively well known groups or areas, but it becomes a substantial challenge for general collections of large sample sets and/or those from poorly known areas.
- Living collections (e.g. botanic gardens and zoos) and well curated dedicated DNA and tissue collections (e.g. ggbn.org) represent a mid-point between these two extremes. They contain accessible and expertly verified material that is well suited to sequence recovery. However despite rapid expansion in tissue databasing and increased accessibility of material in these collections they still only contain a fraction of the species (and their types) compared to the historically accumulate set of preserved specimens in natural history museums and herbaria.

Thus despite the technical difficulties of recovering barcode samples from preserved specimens, natural history collections represent a critically important sample set for constructing a DNA barcode reference library [2].

Large scale barcode studies of natural history collections

Many studies have used museum specimens as a tissue source for DNA, although these studies typically focus on sample sets consisting of 10s or a few hundred samples. A landmark large-scale DNA barcoding project using natural history collections was conducted by Hebert *et al.* [11]. This study involved processing 41,650 Lepidoptera specimens from 12,699 species from the Australian National Insect collection, with an average age of 30 years old. Up to 4 individuals per species were assayed, either via a full length 658bp amplicon or amplifying the CO1 region in two parts. Sequences were recovered from 31,585 specimens (76%) and of these 59% were classed as barcode compliant (at least 487 bp) in total representing 10,931 (86%) of the species.

The first national barcode inventory for plants (flora of Wales) was based on 4272 specimens of 1143 species, with 3637 of these specimens derived from herbarium samples (85%)[12]. In total, 74% of herbarium specimens yielded an *rbcL* barcode sequence, compared to 94% of fresh samples. Likewise 53% of herbarium specimens yielded a *matK* barcode compared to 79% for fresh samples. The authors stressed that the increased costs of processing preserved material were more than offset by the efficiency gains of avoiding costs of new field collections/voucher preparation/and expert verifications [12].

An investigation into the recovery of ITS barcode sequences from fungal herbarium samples was undertaken using material from the Museum of Natural History in Vienna [13]. From a start point of about 5000 samples, some 1107 specimens produced bi-directional sequences of sufficient quality for submitting to GenBank. When the failed samples were reanalysed targeting just ITS1 (as opposed to the full length ITS), substantial gains in recovery were realised [13].

In total about 25K plant specimens from herbaria, and 250K animal specimens from museum collections have DNA barcode sequences lodged in the Barcode of Life Identification System (BOLD) database (P. Hebert, pers comm, 25/6/2016).

Factors influencing recoverability of Sanger sequenced barcodes from natural history collections

Various studies have assessed factors impacting on the recovery of DNA sequences from natural history collections. A common limitation of this knowledge base is the relatively small sample sizes of many of the studies, and the number of candidate variables - and the interactions among those variables. The list of issues to consider is extensive and includes intrinsic attributes such as organism size, shape, physiology and morphology, and age and health at the time of collection. Extrinsic factors include environmental conditions of the collection event, sampling method, preservation method, storage conditions of the sample, and subsequent tissue type and volume used for DNA extraction, the length of target DNA region, the sequencing protocol and chemistry, and the quality of laboratory facilities and expertise [14-20].

Specimen age

There is a general tendency for recovery success of barcode sequences to be correlated with the age of the specimen, with older specimens proving more recalcitrant. The barcode survey of the Lepidoptera of the Australian National Insect Collection (ANIC) [11] provided a large scale quantification of this issue (Figure 2). Recovery rates were high for the first decade or so after collection, and then steadily decline with increasing age. Other factors will of course influence recovery of DNA sequences, and the linear deterioration noted here does not discount difficulties with recently collected samples [21].

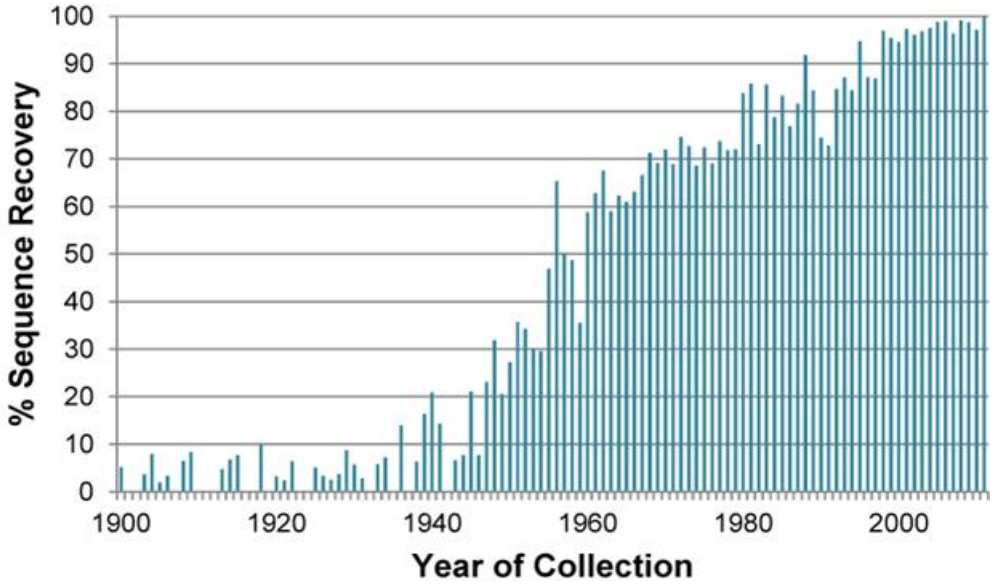


Fig. 2 Recovery of DNA barcode sequences with respect to specimen age from the Australian National Insect Collection. The plot illustrates the recovery of at least 1 barcode compliant record per species via a 658 bp amplicon or a pair of amplicons (307 bp, 408 bp), based on 4 individuals per species from the Australian National Insect Collection (taken from [11]).

Body size

Body size and shape may influence sequence recovery in multiple ways including efficacy of tissue stabilisation post-sampling (slower where surface area is small with respect to tissue volume) and also the total quantities of DNA available in a given sample. In the ANIC survey the Lepidoptera specimens with the very smallest bodies had poor sequence recovery [11], and the authors recommended targeting of younger specimens or increased tissue amounts (e.g. targeting abdomen, rather than legs as the tissue source) for very small insects. A similar association of lower sequence recovery from smaller older specimens was detected in time-series comparison of spider specimens stored in alcohol [22].

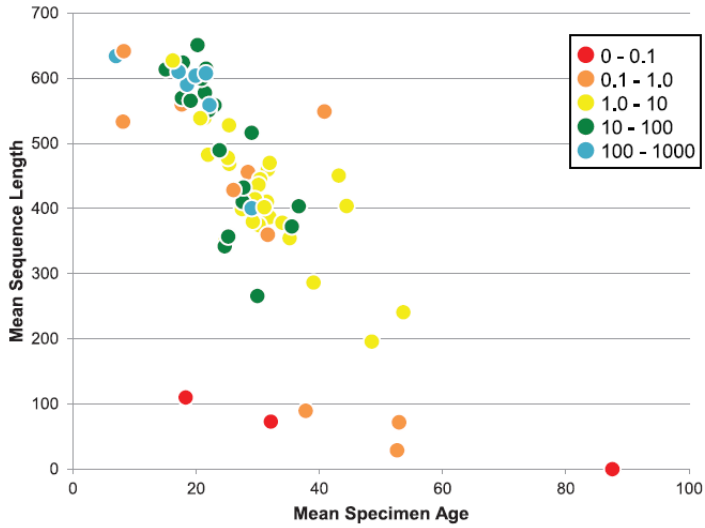


Figure 3: Taken from [11]. Impact of variation in body size (mg) and specimen age on the mean length of barcode sequences recovered from specimens in 66 families of Lepidoptera sampled from

ANIC. Each circle represents a different family while the mean specimen age is the average for all species analysed in a family. The mean body mass (mg) for species in a family are shown by colouration.

Preservation method

Tissue storage conditions and preservation method can have striking impacts on the recovery of DNA sequences [17, 23]. The negative impacts of formalin preservation on DNA recovery are well documented [20], as are the deleterious impacts of alcohol or use of excessive heat for sample drying on collections in the wet-tropics [15]. Within individual taxonomic groups, there can be a further marked influence of individual collectors, reflecting the combined effects of practices from the field to deposition in a collection. This was shown clearly in the ANIC survey, with differences in the recovery of barcode sequences from specimens originating from different collectors [11].

Target locus length

A well-established generalisation is that the longer the DNA target region the more difficult it is to recover from older specimens. This has been shown in various plant [16, 24], animal [25] and fungal studies [13]. Recovery of barcode regions in multiple parts [18, 26, 27], or use of a subset of the barcode region as a minibarcode [13] represent work-around solutions, but with additional labour costs or sacrifice of information content. The problem is particularly acute for some barcode regions such as *matK* which is >800bp long, and lacks robust internal priming sites [5].

DNA barcoding and next generation sequencing technologies

Fig 4. summarises the landscape of issues for DNA barcoding in light of developments in DNA sequencing platforms.

Mainstream DNA barcoding workflows are focused on Sanger sequencing protocols of standard barcode markers. This workflow has been optimised to deliver high-throughput sequencing in large institutional facilities for CO1 barcodes at costs as low as $3 USD, whilst at the same time being compatible with data produced from small scale operations in less well-resourced laboratories. The benefits of this approach are the accessibility of the technology and well-established protocols, practices, and data standards ([28]; Fig 4). A further important benefit is the focus on a small number of loci with limited bio-functional information which limits concerns associated with Access and Benefit Sharing [29]. In terms of limitations - for some taxonomic groups (such as plants), the levels of resolution with standard barcodes can be sub-optimal (Box 1) which has triggered a search for improved barcoding protocols [7, 29]. In addition, the size of the barcode regions, all > 600bp presents challenges for Sanger sequencing of degraded tissues.

New sequencing platforms have resulted in substantial improvements in the recovery of barcode sequences from degraded DNAs. Massively parallel short-read sequencing, followed by assembly into longer regions enables efficient recovery of DNA barcodes [27, 30]. These approaches are intrinsically well suited to use with herbarium or museum specimens [27, 30], and represent a step-change in preserved specimen use. An initial constraint was the cost of ID tags for tracking and assembling sequences from individual samples, but various combinatorial indexing strategies are addressing this problem (e.g. [31]). An emergent pressing general issue for DNA barcoding is the establishment of new data standards for new sequencing platforms.

In addition to effective recovery of standard DNA barcodes, machines such as the Illumina HiSeq 4000 (which produces up to 750 gb per run) offers the potential for greatly increased depth of

sequence coverage for individual samples. This means that additional barcode loci can be recovered from species groups where resolution levels are suboptimal, or where different DNA regions are preferred for different applications [7, 29]. Two main areas under exploration for DNA barcoding are shotgun-based genome-skimming and hybrid capture approaches. The issues involved in both of these approaches are summarised by Hollingsworth *et al.* and Coissac *et al.* [7, 29]. In summary, shotgun sequencing genome skims recover organelle genomes, ribosomal repeats and a random fraction of the nuclear genome. They are an effective method of recovering standard barcode loci and other useful gene sequence data. The approach is intrinsically scaleable as the depth of the skim can increase as sequencing costs fall, thus future proofing the methodology. The approaches has been effectively utilised with herbarium and museum specimens [14, 32, 33]. The big challenges are the informatics and storage needs from the large quantities of data that the approach produces and effective ways of utilising the partially overlapping nuclear fraction of the data. A more targeted approach is to establish a set of informative loci followed by a capture-based sequencing approach. This is intrinsically well suited to recovery from degraded DNAs. Capture probe sets with wide phylogenetic utility are available for organelle DNAs (e.g. [34]), and for nuclear markers for more restricted assemblages [35, 36]. A key challenge in this area is development of bait sets for the nuclear genome that are of broad phylogenetic utility and which still show species-level discrimination [7].

In light of these new approaches to DNA barcoding is important to note that any extension of the standard barcodes to add more loci involves a trade-off (Fig. 4). Even modest gains in the costs and amounts of data have a big impact when scaled over millions of samples. Furthermore, a key element of building the reference library of life on earth is accessibility to laboratories and facilities around the world with varying levels financial resources and technical expertise. If the approach to build the library outstrips the capacity of these facilities or other specimen holders, then sample access will become a rate limiting step.

In summary, DNA barcoding of standard barcode markers is evolving into a combination of classic Sanger-based and parallel sequencing approaches. This provides increased opportunities for sequencing of museum and herbarium specimens. In parallel, efforts to extend (rather than replace) standard barcoding approaches are underway [7, 29].

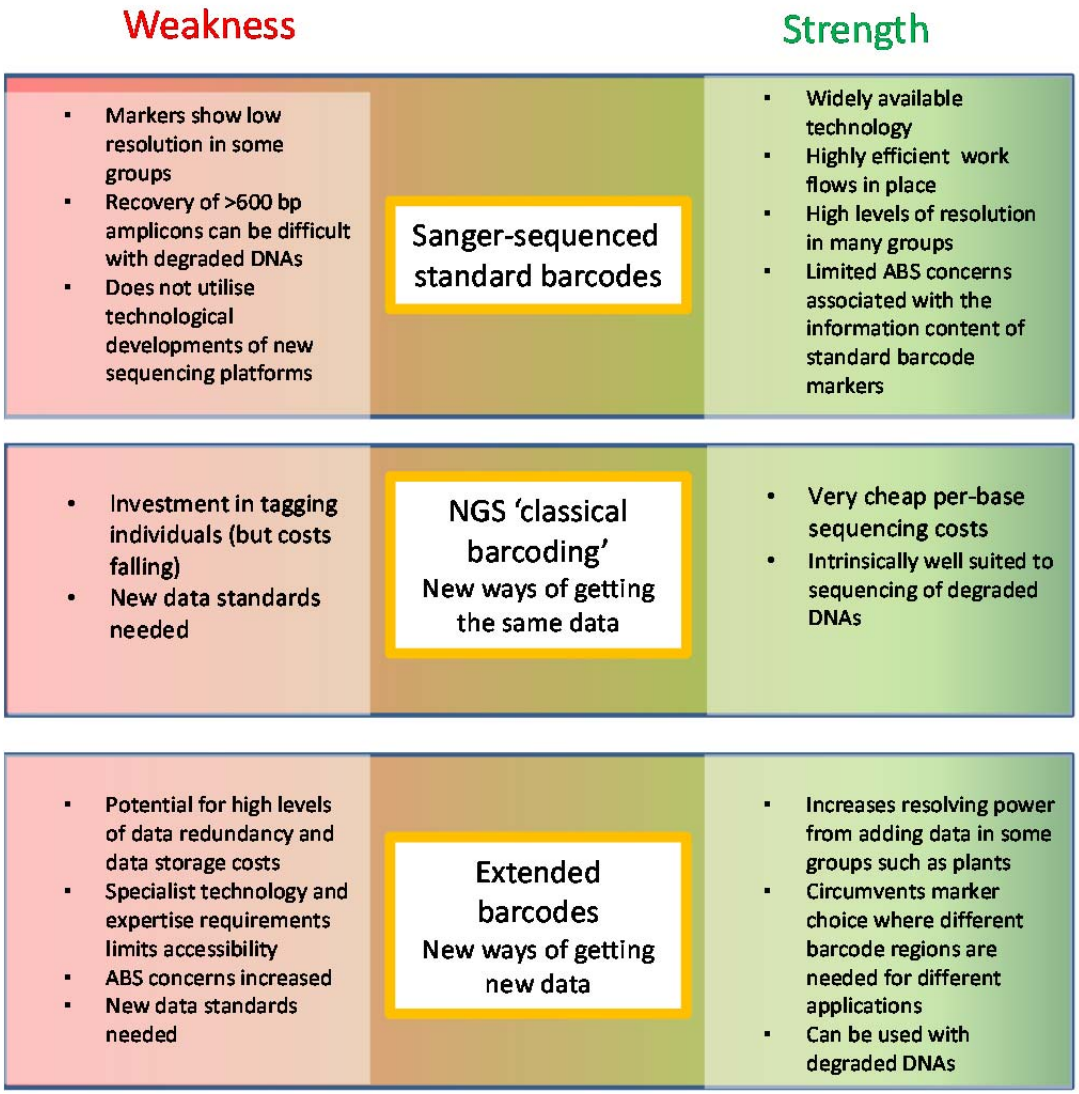


Figure 4. Strengths and weaknesses of different approaches for DNA barcoding: Sanger sequencing of standard barcode markers (top panel), use of NGS for standard barcode markers (middle panel), and NGS approaches to add additional loci (bottom panel)

Strategic priorities for DNA barcoding natural history collections

Future priorities for DNA barcoding natural history collections reflect the interplay of scientific and societal needs, with the overlay of practical constrains and considerations.

Scientific priorities

Type specimens: Sequencing type material provides a clear link between a species name and the intrinsic indexing capacity of barcode sequences. Improvements in sequence recovery have lowered the risk of tissue sacrifice for no returns, although protocol tests on non-type material are always recommended. Where sequence data is not recoverable from type specimens for technical reasons or because permission to sample is not granted, the production of DNA barcode sequence from an alternative representative specimen for each species is beneficial. A key consideration for deciding whether to DNA barcode type material, is to ensure that the target loci provide discrimination at the species-level. For many animal groups this is the default, whereas for plants there is no *a priori*

assumption that standard barcoding markers will show species-specific clusters, which inevitably limits the potential information gains from sequencing type material for those loci for barcoding.

Filling phylogenetic gaps: Even in relatively well-documented groups of organisms like land plants, many genera lack any nucleotide sequence data in public databases. Targeted sampling to fill these gaps is a high priority. In megadiverse groups like invertebrates or fungi, massive uncertainty in underlying species and phylogenetic diversity patterns represents a substantial knowledge gap. Filling the major gaps in the tree of life with even single barcode sequences would act as an effective set of anchor points for further study, and would assist in the wider interpretation of eDNA and metagenomics studies. Plugging these gaps will inevitably involve substantial de novo sampling of under-studied geographic regions and habitats. However, a useful exercise would be targeted cross comparison of known families or genera represented by material in natural history collections against sequence records in public sequence repositories.

Big genera: Species-rich groups are a particularly well suited target for DNA barcoding. Many large genera lack recent robust monographic treatments, with a key contributing factor simply being the large scale of the task. Barcode sequencing across large genera can provide a sequence-based framework of specimens, species and related sets of species – effectively compartmentalising the task into more manageable units. Integrating these studies with named material in natural history collections is an integral part of effective taxonomic revisions.

All taxon biodiversity inventories: Comparators on the amount and distribution of species diversity are typically based around occurrence records of species names at locations. There is an obvious benefit to comparing diversity patterns based on DNA sequences (intrinsically well suited to quantitative comparisons) rather than solely using species names. Comprehensive sampling of the species in a given place offers this potential, as well as establishing a dataset for the study of inter-specific interactions [37-39]. Major projects are underway barcoding the tree species in forest plots [40], local insect populations via the Global Malaise Trap project (globalmalaise.org) and various other location-based studies (e.g. Area de Conservación Guanacaste in Costa Rica [41] and barcoding all species of Moorea; mooreabiocode.org). As the place-based studies are well suited to field sampling, there is a wider issue in some cases in establishing voucher specimens in the first place, rather than using them for the barcoding itself. Nevertheless – given the extensive investment of efforts into these studies, integrating sequencing of field samples with that of preserved specimens has the potential to enrich the study.

Societal needs

Endangered species: Effective identification of endangered species (and the species they interact with) is useful for species management as well as assessing and regulating material in trade [40] [42, 43]. DNA barcodes obtained from preserved specimens can be particularly useful for very rare species as it avoids the potential harm to living individuals from sampling, and by-passes the often complicated and extensive permit application procedures for endangered species. DNA barcoding pre-existing expertly verified preserved specimens makes best use of that material, particularly for regulatory enforcement where a verifiable link between the sequence and expertly identified material is important.

Pollinators: Global concerns about food security include decreases in production and/or increase in costs due to pollinator declines. Yet identification challenges, cryptic species, and high levels of diversity in insect pollinators hampers understanding of the biology and status of individual species. Key priorities for DNA barcoding pollinators include establishing the reference library for key pollinator species of bees, Diptera, Coleoptera and Lepidoptera [44].

Invasive non-native species: Invasive non-native species including pathogens, exotic pests and introduced weeds can have major economic impacts. Identification of these organisms at an early

stage is critical to their effective control (e.g. border control, quarantine, early post establishment eradication) [45-48]. Barcode datasets and effective assays are available for many species of concern, but there is no systematically compiled register of the barcode availability (and genetic diagnosibility) of the major pests, pathogens and other invasive species. Development of this gap analysis is a pre-requisite to targeting natural history collections to fill the gap.

Human health: Comprehensive barcode datasets to support the identification of species which impact on human health address a clear societal need. Biological identifications relevant to human health include vectors of human diseases, parasites, pathogens and allergens. It also includes species which are poisonous when consumed [49], or conversely species with beneficial medicinal properties [50]. As with invasive non-native species, there are numerous barcode and other types of diagnostic genetic assays, but no systematically compiled register of species for mapping barcode availability to societal identification needs. Given the open-ended nature of this issue, a comprehensive register is not feasible, but a pragmatic start point is identifying high priority groups of organisms lacking barcodes but which are well represented in natural history collections (e.g. poisonous fungi similar in appearance to edible fungi) [9].

Practical considerations influencing the use of natural history specimens in DNA barcoding projects

Cost-effectiveness and tractability: Obtaining DNA barcodes from previously collected preserved specimens can represent a substantial saving over the costs of planning and executing new field work, securing permits, and preparing and curating newly collected voucher specimens. The 'full-cost' of new field collections are substantial. On the other hand, failure rates in DNA barcoding preserved samples can be high. An important consideration in use of natural history specimens in DNA barcoding projects is the cost-benefit analysis of the savings on new field collections traded off against the increased costs of laboratory processing preserved rather than fresh materials.

Complementarity: In 2015 there were 14,000 users of BOLD from >1000 institutes from 94 countries (Sujeevan Ratnasingham pers comm. 2/9/2015). Factoring in the wider pool of researchers who use different informatics platforms, the number of individual barcoding projects is very high. Given an overall limitation in most projects is the number of specimens that can be handled, there is a strategic need for greater cross-project coordination. This is particularly relevant to Europe, where various countries have national barcoding campaigns (e.g. Germany www.bolgermany.de; Norway www.norbol.org; Austria www.abol.ac.at; Switzerland www.swissbol.ch). Given a typical project sample of n=3-5 individuals per species, a natural outcome is an excess of samples for many species (the sum of multiple national barcode campaigns) while other species remain un-sampled.

Increased emphasis on applied barcoding: Natural history collections and their associated research teams have a rich history of working on taxa and places. This provides data and knowledge which can be repurposed to meet stakeholder needs. However, given greater urgency and visibility of major societal challenges (e.g. UN Sustainable Development Goals; sustainabledevelopment.un.org), a further shift in emphasis is required to application-based, rather than taxon-based or place-based studies. This requires changes in working practices and new collaborative networks as many users of applied barcode datasets will need very diverse sample sets and broad geographical coverage. Likewise where barcode datasets are targeted to support regulatory enforcement, the working flows need to match the relevant standards of procedure [51].

Maximising added value: Generating barcode data from natural history collections represents an investment of time and money. Given a choice of which specimen to sequence, one criterion is 'added value'. There is a clear scientific benefit to sequencing type specimens or specimens that fill a gap in unsequenced phylogenetic space as outlined above. In addition, there is a benefit to barcoding specimens with particularly rich meta-data (e.g. high resolution spatial coordinates, high

resolution images) or which have high likelihood of incorporation into other scientific studies. Furthermore, the barcoding of recently collected museum or herbarium specimens creates an opportunity to feedback to the original collectors information on specimen identities which can serve to steer / harness further collecting efforts.

Prioritising actively used specimens: Specimens which are in active use indicate some level of user interest and demand. Given this involves specimen handling and databasing, this is another relevant criterion for consideration in choosing material for DNA barcoding of collections. As these samples are being manually removed from the collections, and are more likely to have their identifications and database entries up-to-date, there are relatively low 'add on costs' to barcode these specimens.

Supporting access to preserved specimens for DNA barcoding

Institutional policies on tissue access for DNA barcoding are often idiosyncratic and not strategic, with access decisions being made at various levels in organisational management structures. There are clear benefits to the production of clear guidelines on when to agree to (semi-) destructive sampling. Establishment of a generic framework document that can be adapted for different institutional needs is a high priority. As part of this, a clear need is a decision-support document for curators to help guide the decision on when to permit tissue sampling. The key steps are to assess:

(a) Whether a pilot project has been conducted demonstrating feasibility, or whether the specimens targeted for sampling are suitable for protocol development because excess tissue is available or many duplicates exist.
(b) Will the information being generated be informative for the question in hand (e.g. has the case been made that the genetic markers targeted by the investigator are informative for question at hand?). The level of stringency of this assessment is obviously dependent on the uniqueness of the individual specimen, but is a particularly important step when type specimens are targeted.

Ongoing major initiatives among SYNTHESYS3 partners

Annexe 2 (associated spreadsheet) summarises major DNA barcoding initiatives underway among member organisations of the SYNTHESYS3 project. Additional information on Barcoding projects can be recovered from BOLD (boldsystems.org).

References

1 Hebert, P.D.N.*, et al.* (2003) Biological identifications through DNA barcodes. *Proceedings of the Royal Society of London, Series B* 270, 313-321
2 Hebert, P.D.N.*, et al.* (2016) From writing to reading the encyclopedia of life. *Philosophical Transactions of the Royal Society B: Biological Sciences* 371, 20150321
3 Ratnasingham, S. and Hebert, P.D.N. (2013) A DNA-based registry for all animal species: the Barcode Index Number (BIN) system. *PLoS ONE* 8, e66213
4 Hebert, P.D.N.*, et al.* (2016) Counting species with DNA barcodes: Canadian insects. *Philosophical Transactions of the Royal Society, Series B.* 371, 20150321
5 Hollingsworth, P.M.*, et al.* (2011) Choosing and using a plant DNA barcode. *PLoS ONE* 6, e19254
6 Taberlet, P.*, et al.* (2007) Power and limitations of the chloroplast trnL (UAA) intron for plant DNA barcoding. *Nucleic Acids Research* 35, e14
7 Hollingsworth, P.M.*, et al.* (2016) Telling plant species apart with DNA: from barcodes to genomes. *Philosophical Transactions of the Royal Society B: Biological Sciences* 371, 20150338

8 Schoch, C.L., *et al.* (2012) Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. *Proceedings of the National Academy of Sciences* 109, 6241-6246
9 Yahr, R., *et al.* (2016) Scaling up discovery of hidden diversity in fungi: impacts of barcoding approaches. *Philosophical Transactions of the Royal Society B: Biological Sciences* 371, 20150336
10 Pawlowski, J., *et al.* (2012) CBOL Protist Working Group: Barcoding Eukaryotic Richness beyond the Animal, Plant, and Fungal Kingdoms. *PLoS Biol* 10, e1001419
11 Hebert, P.D.N., *et al.* (2013) A DNA 'Barcode Blitz': Rapid digitization and sequencing of a natural history collection. *PLoS ONE* 8, e68535
12 de Vere, N., *et al.* (2012) DNA barcoding the native flowering plants and conifers of Wales. *PLoS ONE* 7, e37945
13 Osmundson, T.W., *et al.* (2013) Filling Gaps in Biodiversity Knowledge for Macrofungi: Contributions and Assessment of an Herbarium Collection DNA Barcode Sequencing Project. *PLoS ONE* 8, e62419
14 Bakker, F.T., *et al.* (2016) Herbarium genomics: plastome sequence assembly from a range of herbarium specimens using an Iterative Organelle Genome Assembly pipeline. *Biological Journal of the Linnean Society* 117, 33-43
15 Bakker, F.T. (2015) DNA sequences from plant herbarium tissue. In *Next-Generation Sequencing in Plant Systematics* (Hörandl, E. and Appelhans, M.S., eds), Koeltz Scientific Books
16 Särkinen, T., *et al.* (2012) How to open the treasure chest? Optimising DNA extraction from herbarium specimens. *PLoS ONE* 7, e43808
17 Staats, M., *et al.* (2011) DNA damage in plant herbarium tissue. *PLoS ONE* 6, e28448
18 Mitchell, A. (2015) Collecting in collections: a PCR strategy and primer set for DNA barcoding of decades-old dried museum specimens. *Molecular Ecology Resources* 15, 1102-1111
19 Roth, S. (2012) Pouring new wine into old bottles: DNA studies on collection material. *Heteropteron* 36, 5-8
20 Zimmermann, J., *et al.* (2008) DNA damage in preserved specimens and tissue samples: a molecular assessment. *Frontiers in Zoology* 5, 18-18
21 Burrell, A.S., *et al.* (2015) The use of museum specimens with high-throughput DNA sequencers. *Journal of Human Evolution* 79, 35-44
22 Miller, J.A., *et al.* (2013) Which specimens from a museum collection will yield DNA barcodes? A time series study of spiders in alcohol. *ZooKeys*, 365, 245-261
23 Choi, J., *et al.* (2015) All that is gold does not glitter? Age, taxonomy, and ancient plant DNA quality. *PeerJ* 3, e1087
24 Little, D.P. (2014) A DNA mini-barcode for land plants. *Molecular Ecology Resources* 14, 437-446
25 Hajibabaei, M. and McKenna, C. (2012) DNA mini-barcodes. *Methods in molecular biology (Clifton, N.J.)* 858, 339-353
26 Hernández-Triana, L.M., *et al.* (2014) Recovery of DNA barcodes from blackfly museum specimens (Diptera: Simuliidae) using primer sets that target a variety of sequence lengths. *Molecular Ecology Resources* 14, 508-518
27 Prosser, S.W.J., *et al.* (2016) DNA barcodes from century-old type specimens using next-generation sequencing. *Molecular Ecology Resources* 16, 487-497
28 Hanner, R. (2005 (updated 2009)) DNA standards for BARCODE records in INSDC (BRIs). *http://studentdnabarcoding.org/pdf/Barcode%20Data%20Standards.pdf*
29 Coissac, E., *et al.* (2016) From barcodes to genomes: extending the concept of DNA barcoding. *Molecular Ecology* 25, 1423-1428
30 Shokralla, S., *et al.* (2015) Massively parallel multiplex DNA sequencing for specimen identification using an Illumina MiSeq platform. *Scientific Reports* 5, 9687
31 Peterson, B.K., *et al.* (2012) Double digest RADseq: An inexpensive method for *de novo* SNP dscovery and genotyping in model and non-model species. *PLoS ONE* 7, e37135
32 Crampton-Platt, A., *et al.* (2016) Mitochondrial metagenomics: letting the genes out of the bottle. *GigaScience* 5, 1-11

33 Besnard, G.*, et al.* (2016) Valuing museum specimens: high-throughput DNA sequencing on historical collections of New Guinea crowned pigeons (*Goura*). *Biological Journal of the Linnean Society* 117, 71-82
34 Stull, G.W.*, et al.* (2013) A targeted enrichment strategy for massively parallel sequencing of angiosperm plastid genomes. *Applications in Plant Sciences* 1, apps.1200497
35 de Sousa, F.*, et al.* (2014) Phylogenetic properties of 50 nuclear loci in *Medicago* (Leguminosae) generated using multiplexed sequence capture and next-generation sequencing. *PLoS ONE* 9, e109704
36 Hart, M.L.*, et al.* (2016) Retrieval of hundreds of nuclear loci from herbarium specimens. *Taxon* 65, 1081–1092
37 Taberlet, P.*, et al.* (2012) Towards next-generation biodiversity assessment using DNA metabarcoding. *Molecular Ecology* 21, 2045-2050
38 Valentini, A.*, et al.* (2009) New perspectives in diet analysis based on DNA barcoding and parallel pyrosequencing: the *trnL* approach. *Molecular Ecology Resources* 9, 51-60
39 Kartzinel, T.R.*, et al.* (2015) DNA metabarcoding illuminates dietary niche partitioning by African large herbivores. *Proceedings of the National Academy of Sciences* 112, 8019-8024
40 Kress, W.J.*, et al.* (2012) Generating plant DNA barcodes for trees in long-term forest dynamics plots. *Methods in molecular biology (Clifton, N.J.)* 858, 441-458
41 Janzen, D.H. and Hallwachs, W. (2011) Joining Inventory by Parataxonomists with DNA Barcoding of a Large Complex Tropical Conserved Wildland in Northwestern Costa Rica. *Plos One* 6 e18123
42 Shapcott, A.*, et al.* (2015) Mapping biodiversity and setting conservation priorities for SE Queensland's rainforests using DNA barcoding. *PLoS One* 10 e0122164
43 Yesson, C.*, et al.* (2011) DNA barcodes for Mexican Cactaceae, plants under pressure from wild collecting. *Molecular ecology resources* 11, 775-783
44 http://ibol.org/wg-1-6-pollinators/
45 Cross, H.B.*, et al.* (2011) DNA barcoding of invasive species. In *Fifty Years of Invasion Ecology: The Legacy of Charles Elton* (Richardson, D.M., ed), pp. 289-299, Wiley-Blackwell
46 Armstrong, K. (2010) DNA barcoding: a new module in New Zealand's plant biosecurity diagnostic toolbox. *Bulletin OEPP* 40, 91-100
47 Pelletier, Y.*, et al.* (2012) A New approach for the identification of aphid vectors (Hemiptera: Aphididae) of potato virus Y. *Journal of Economic Entomology* 105, 1909-1914
48 Smith, K.F.*, et al.* (2012) Barcoding of the cytochrome oxidase I (COI) indicates a recent introduction of *Ciona savignyi* into New Zealand and provides a rapid method for *Ciona* species discrimination. *Aquatic Invasions* 7, 305-313
49 Xie, L.*, et al.* (2014) Prospects and problems for identification of poisonous plants in China using DNA barcodes. *Biomedical and Environmental Sciences* 27, 794-806
50 Wallace, L.J.*, et al.* (2012) DNA barcodes for everyday life: Routine authentication of Natural Health Products. *Food Research International* 49, 446-452
51 Ogden, R. (2010) Forensic science, genetics and wildlife biology: getting the right mix for a wildlife DNA forensics lab. *Forensic Science, Medicine, and Pathology* 6, 172-179

Annexe 1. Case studies

A.1.1. Mollusc DNA in museum collections

Nikolaus Szucsich, Naturhistorisches Museum Wien, Austria

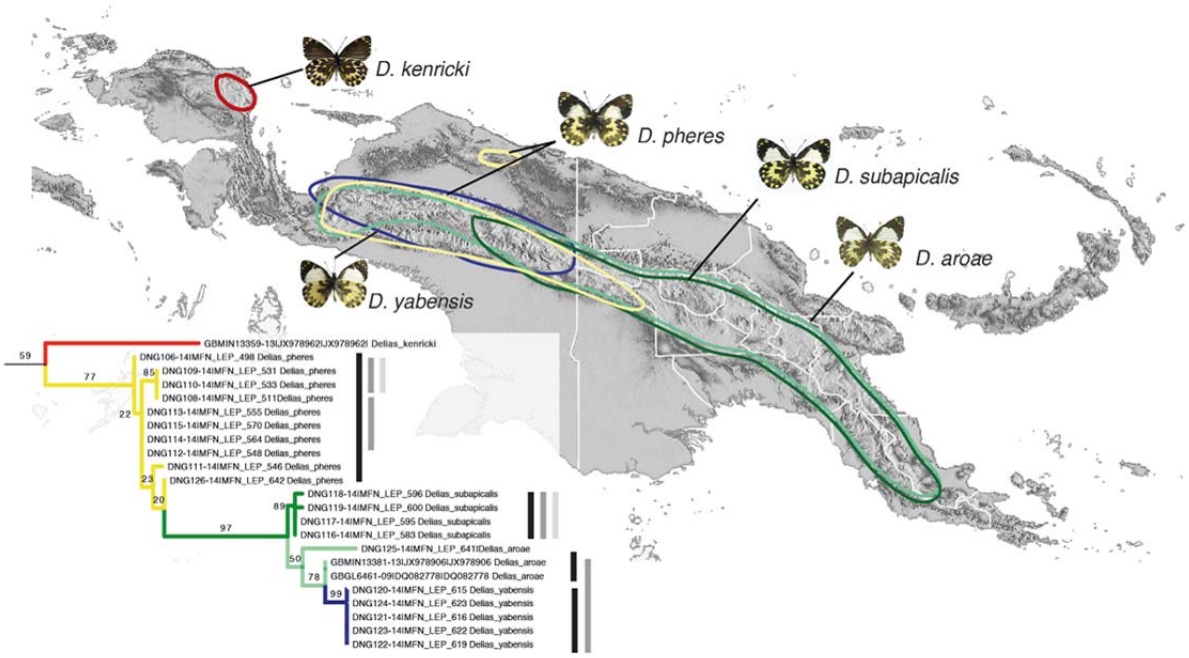
Museums harbour huge treasuries of new or old, regular or odd, common or rare animals. Along with conserving all the morphological variation for future generations, scientific collections represent a hidden source for studying genetic diversity. Vouchers of molluscs are assessed to be of limited value in this respect, since apart of the DNA-fragmentation over time, DNA isolation and PCR amplification are known to be distracted by polysaccharides in tissue and mucus. Funded by the European SYNTHESYS2 Program a project dealt with the tricky task to extract DNA from old specimens of molluscs. In total 72 glasses, comprising 20 different taxa of 4 classes of molluscs were analyzed, with collecting dates ranging from 1877 to 2002. Different extraction protocols and PCR primers were compared for their ability to amplify two short sections of the mitochondrial genome (COI and 16S). Results are published in Jaksch et al. (2016) and will be applied and further developed in the ongoing DNA barcoding project of Austria molluscs, conducted in the framework of ABOL (Austrian Barcode of Life). Already, a good working protocol was established to extract DNA from mummies. i.e. remains of tissue in empty shells of dry collections (Jaksch et al. in prep.).

Jaksch, K., Eschner, A., Rintelen, T. V. & Haring, E. (2016) DNA analysis of molluscs from a museum wet collection: a comparison of different extraction methods. BMC Research Notes 9 (1): 348 – DOI: 10.1186/s13104-016-2147-7

A.1.2. *Delias* butterfly radiation on New Guinea

Dr. Thomas von Rintelen, Museum für Naturkunde, Leibniz Institute for Evolution and Biodiversity Science, Germany

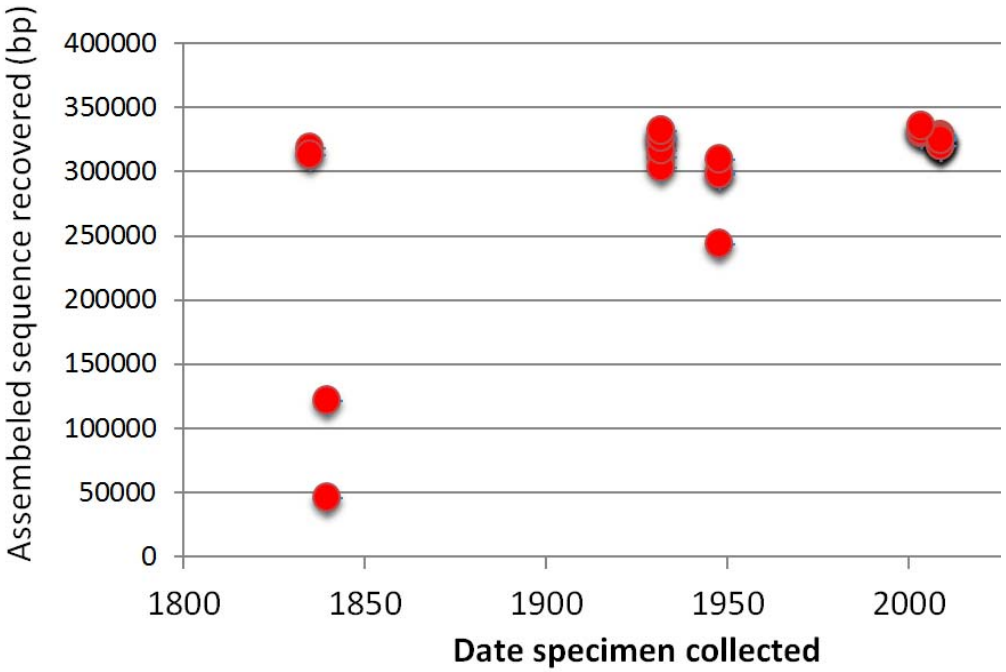
The pierid butterfly genus *Delias* has radiated extensively on the island of New Guinea, where more than 54 % of the c. 250 species occur. The variability of wing patterns and coloration has led to the description of many forms and consequently considerable taxonomic uncertainty. Getting a grip on species boundaries within *Delias* is essential for understanding the diversity of species and their distribution on New Guinea. This has considerable implications for assessing the impact of anthropogenic threats including climate change on this predominantly high-altitude taxon. While a DNA barcoding approach has proven extremely successful in other butterfly groups, no fresh samples of *Delias* from all species and forms across their wide reported distribution ranges were available. The *Delias* barcoding project (BOLD data project ID: DS-DNG02) is thus entirely based on dried specimens from museum collections, particularly the van Mastright collection in Jayapura, Indonesia, and the Berlin Museum für Naturkunde, with specimen age ranging from 8 to 110 years (mostly > 20 years old). In an initial study using a traditional PCR approach targeting up to four short fragments of the standard COI barcoding fragment (658 bp), a >500 bp fragment could be obtained for 92% of the 161 samples used. Even with this limited dataset conflicts of morphology based species delimitations and BINs was observed for 21 out of 40 species represented by multiple sequences, providing evidence for both cryptic species (N=7) and oversplitting (N=14). Building on this success, more than 1,000 additional specimens of *Delias* will be barcoded within the next six months using a hybrid capture approach that has also been successfully tested in the pilot study, creating an extensive barcode library as a resource for studying species and speciation in a highly diverse tropical taxon.



A.1.3. Sequencing nuclear genes from herbarium specimens

Michelle Hart, Royal Botanic Garden Edinburgh, UK

DNA barcoding in plants with standard plastid and ribosomal DNA markers often provides resolution to species-group, rather than unique species identification. This is a particular challenge in plant groups which have shown recent radiations or have other forms of taxonomic complexity (e.g. hybridisation, polyploidy, narrowly circumscribed taxa). To improve resolution in these cases, multiple unlinked nuclear markers are required. Hart *et al.* (2016) addressed this, using a hybrid capture approach to recover 276 nuclear loci from the legume genus *Inga.* This approach was then tested on herbarium specimens up to 180 years old. The hybrid capture approach recovered high quality, high coverage sequence data from specimens collected as early as 1835, and from as little as 16ng of starting DNA. This study illustrates the potential of new sequencing technologies to substantially improve the recovery of DNA barcode type data from preserved natural history collections.



Hart, M. L., Forrest, L. L, Nicholls, J., Kidner, C. 2016. Retrieval of hundreds of nuclear loci from herbarium specimens. *Taxon* 65: 1081-1092.

A.1.4. DNA barcoding of CITES protected species

Barbara Gravendeel, Naturalis Biodiversity Center, Leiden, The Netherlands

The Convention on International Trade in Endangered Species of Wild Fauna and Flora (CITES) entered into force in 1975. It aims to control and regulate trade in endangered species. The convention produces appendices that list taxa in which trade is controlled or prohibited. Monitoring of trade in CITES taxa is a challenge to customs authorities worldwide when parts or organisms are used in mixtures such as in traditional medicines.

High-throughput sequencing (HTS) techniques yield increasingly large volumes of barcode sequences. This leads to a greater identifying potential for complex species samples at low cost (Staats *et al.* 2016). The process of going through a set of identified sequences and manually comparing them to the CITES appendices is labour intensive and error prone. CITES appendices often also list higher taxa (e.g. genera, families) whereas reference sequences are annotated to species level. False positive hits can occur for DNA barcodes deposited in NCBI GenBank due to incorrect taxonomic name annotations.

Open-source, freely available pipelines were recently developed that automate the identification and CITES listing verification steps to enable efficient scanning of large sample sequence datasets for quick detection of presence of DNA barcodes derived from protected species Lammers *et al.* 2014). The number of CITES protected species was 30713 in 2014, of which roughly 55% were present in NCBI GenBank with DNA barcodes. A total of 13883 species (45%) remained to be sequenced. DNA barcoding of CITES listed species therefore needs to be prioritized.

Staats, M., Arulandhu, A.J., Gravendeel, B., Holst-Jensen, A., Scholtens, I., Peelen, T., Prins, T.W., Kok, E. 2016. Advances in DNA metabarcoding for food and wildlife forensic species identification. *Analytical and Bioanalytical Chemistry* 408(17): 4615-4630.

Lammers, Y., Peelen, T., Vos, R.A., Gravendeel, B. 2014. The CITES checker pipeline, a tool for automated detection of illegally traded species from high-throughput sequencing data. *BMC Bioinformatics* 15: 44.