MOLECULAR ECOLOGY
RESOURCES
Molecular Ecology Resources (2016) **16**, 487–497 doi: 10.1111/1755-0998.12474

DNA barcodes from century-old type specimens using next-generation sequencing

SEAN W. J. PROSSER,* JEREMY R. DEWAARD,* SCOTT E. MILLER† and PAUL D. N. HEBERT*
Biodiversity Institute of Ontario, University of Guelph, Guelph, ON, Canada, †National Museum of Natural History, Smithsonian Institution, Washington, DC, USA

Abstract

Type specimens have high scientific importance because they provide the only certain connection between the application of a Linnean name and a physical specimen. Many other individuals may have been identified as a particular species, but their linkage to the taxon concept is inferential. Because type specimens are often more than a century old and have experienced conditions unfavourable for DNA preservation, success in sequence recovery has been uncertain. This study addresses this challenge by employing next-generation sequencing (NGS) to recover sequences for the barcode region of the cytochrome *c* oxidase 1 gene from small amounts of template DNA. DNA quality was first screened in more than 1800 century-old type specimens of Lepidoptera by attempting to recover 164-bp and 94-bp reads via Sanger sequencing. This analysis permitted the assignment of each specimen to one of three DNA quality categories – high (164-bp sequence), medium (94-bp sequence) or low (no sequence). Ten specimens from each category were subsequently analysed via a PCR-based NGS protocol requiring very little template DNA. It recovered sequence information from all specimens with average read lengths ranging from 458 bp to 610 bp for the three DNA categories. By sequencing ten specimens in each NGS run, costs were similar to Sanger analysis. Future increases in the number of specimens processed in each run promise substantial reductions in cost, making it possible to anticipate a future where barcode sequences are available from most type specimens.

Keywords: degraded DNA, DNA barcoding, DNA sequencing, next-generation sequencing, type specimens

Received 20 July 2015; revision received 14 September 2015; accepted 25 September 2015

Introduction

The immense repositories of identified specimens in the world's natural history museums provide the opportunity to construct a DNA barcode reference library that can subsequently be used to identify newly collected specimens (Hebert *et al.* 2003, 2013). However, the scientific value of this library would be greatly enhanced if each species was represented by sequences from its type material, particularly the holotype (Kvist *et al.* 2010). Without such information, there are many cases in which the correct application of taxon names is uncertain. For example, the analysis of type(s) is critical when the study of modern specimens suggests synonymy (e.g. Liimatainen *et al.* 2014; Mutanen *et al.* 2014; Antoni *et al.* 2015) or when it indicates that a long-known species is actually a complex of two or more morphologically similar taxa (e.g. Hausmann *et al.* 2009a; Porco *et al.* 2012; Liimatainen *et al.* 2014; Johnson & Wilmer 2015). The recovery of a barcode sequence from type material is also essential when it represents the only known record (s) for a taxon – a situation that is surprisingly common (e.g. Kirchman *et al.* 2010).

Prior studies have often encountered difficulty in recovering sequence information from old museum specimens because of DNA degradation (Zimmermann *et al.* 2008; Allentoft *et al.* 2012; Dabney *et al.* 2013). This barrier has been reduced as protocols have improved, but there are still important constraints (Hausmann *et al.* 2009a,b; Lees *et al.* 2010, 2011; Strutzenberger *et al.* 2012). Past studies have generally employed several PCRs to generate a set of short amplicons that were Sanger sequenced and assembled into a barcode record. When many amplification reactions are required, as in cases where difficulties in primer binding are encountered, template can be depleted before sequence is recovered. There is no easy solution because DNA extracts are small (<50 μL) and concentrations are low (typically <0.5 pg/μL) so dilution is rarely feasible (Hausmann *et al.* 2009a; Rougerie *et al.* 2012). As a consequence, sequence recovery from many type specimens is not currently possible.

Correspondence: Sean W. J. Prosser, Fax: +1-519-824-5703;
E-mail: sprosser@uoguelph.ca

Next-generation sequencing (NGS) is increasingly used for studies on both freshly collected and museum specimens (Rowe *et al.* 2011; Hughey *et al.* 2013; Tin *et al.* 2014). Work on fresh specimens has shown that the barcode region can be recovered from hundreds of individuals at a time using multiplex identifier (MID) tags to associate the sequence records to each specimen (Shokralla *et al.* 2014, 2015). This study extends earlier work by recovering full-length barcodes from type specimens with heavily degraded DNA by employing multiplex PCR to generate short amplicons covering the barcode region and then using NGS for their characterization. In practice, there are complexities in multiplex analysis that must be resolved to ensure efficient, unbiased of fragments spanning the barcode region, but this study has created a NGS protocol that overcomes these problems. Its effectiveness was validated by recovering sequences from century-old specimens of Lepidoptera, including those where Sanger analysis completely failed. Equally important, this NGS approach escapes problems that often confront Sanger analysis, such as uncertain primer binding, amplification bias and the need for large amounts of template DNA.

Materials and methods

Type specimens

Tissue samples were obtained from 1820 specimens (mostly primary types but some were equally important nontypes) of Geometridae (Lepidoptera) from the Natural History Museum (London) as part of a project to develop a strongly validated taxonomic system to support species inventories and studies of host plant use in Papua New Guinea (Holloway *et al.* 2009; Miller 2014). Genitalic dissections of these specimens generated residual tissue that was held frozen until its use in this study. All 1820 specimens were analysed using a traditional Sanger-based approach and, based on those results, three subsets of 10 specimens (n = 30) were chosen for further analysis via NGS.

DNA extraction

All tissue samples were processed in an isolated 'clean' laboratory at the Canadian Centre for DNA Barcoding (CCDB; www.ccdb.ca) with dedicated reagents, supplies and protective clothing. Each sample was incubated overnight in lysis buffer, following a modified protocol of Knolke *et al.* (2005) (tissue samples was incubated overnight in lysis buffer + ProK prior to dissection), before DNA was extracted using a silica membrane-based method in either single columns or 96-well plate format (Ivanova *et al.* 2006). To maximize the concentration of extracted DNA, elution from each silica membrane was performed with 30 μL of prewarmed (to 56 °C) 10 mM Tris-HCl. Total DNA concentrations were too low to quantify but were expected to be <0.05 pg/μL on average.

Sanger sequencing

As DNA quality varies greatly, even among specimens of similar age (Dean & Ballard 2001; Hebert *et al.* 2013), each DNA extract was initially assessed by Sanger analysis. This involved an attempt to amplify both 164-bp (C_microLepF1_t1 + C_TypeR1) and 94-bp (C_TypeF1 + C_TypeR1) regions of the COI barcode (Hebert *et al.* 2013; Hernández-Triana *et al.* 2013). PCR amplification and cycle sequencing employed standard CCDB protocols (Ivanova *et al.* 2006; deWaard *et al.* 2008; Hebert *et al.* 2013) with amplicons bidirectionally sequenced on an ABI 3730XL (Applied Biosystems). All traces were edited using CodonCode v. 4.2.7 (CodonCode Corporation) and the resulting 164-bp and 94-bp sequences were validated by comparison with sequences from conspecific individuals or, when they were unavailable, by neighbour-joining (NJ) analysis to ensure that each sequence branched as one would expect relative to their most closely related taxa. These tests for sequence recovery permitted the assignment of DNA from each specimen to one of three categories: (i) high quality (HQ) – those that generated a 164-bp sequence, (ii) medium quality (MQ) – those that generated a 94-bp sequence and (iii) low quality (LQ) – those that failed to generate any sequence. This study examined ten specimens from each category with the goal of developing a NGS protocol effective across varying levels of DNA degradation. Preliminary experiments (not shown) involving sequencing members of the same genus across all quality classes showed that phylogenetic disturbances were not affecting our analysis. Therefore, in this study, we chose to maximize taxonomic coverage, so the specimens selected for analysis (Table S1, Supporting information) included a single representative from 30 different genera of the family Geometridae, all more than a century old (mean age = 111 years). Sequences, electropherograms and primer details for the specimens are on BOLD (dx.doi.org/10.5883/DS-NGSTYPES).

Next-generation sequencing

DNA degradation often limits PCR amplicons to <200 bp in specimens that are more than 50 years old (Allentoft *et al.* 2012), precluding efforts to recover the entire barcode region with one or two primer sets. As a consequence, primer sets were designed to amplify fragments ranging in length from 120 to 148 bp with enough

overlap to permit recovery of the 658-bp barcode region. These primers needed to be tailed with adapter sequences for analysis on an Ion Torrent PGM (Life Technologies) and with multiplex identifier (MID) tags to distinguish sequence reads from each specimen. Ten sets of MID-tagged primers, each consisting of six forward and six reverse primers, were employed to analyse ten type specimens per NGS run (Table S2, Supporting information).

Optimization of NGS protocols

Optimization studies tested the impact of varied primer combinations, number of PCR cycles, differential concentrations of primers and nesting of PCRs. Efforts to multiplex all six forward and reverse primers in a single reaction were unsuccessful because the small regions of overlap were preferentially amplified over the six target fragments. Splitting the PCR into two reactions, each targeting nonadjacent fragments (e.g. PCR1 = F1 + R1, F3 + R3, F5 + R5; PCR2 = F2 + R2, F4 + R4, F6 + R6), solved this issue, but revealed another problem: the dominance of certain amplicons. This problem was overcome by mixing the forward primers with the full complement of reverse primers (e.g. PCR1 = F1 + F3 + F5 + six reverse primers; PCR2 = F2 + F4 + F6 + 5 reverse primers). This allowed each forward primer to potentially pair with several downstream reverse primers, creating redundancy that maximized sequence recovery while reducing the dominance of any particular amplicon. For example, depending upon the quality of the template DNA, the barcode segment amplified by primers F4 + R4 could be amplified by any of the twelve combinations of F1, F2, F3 or F4 paired with R4, R5 or R6. This redundancy aided the recovery of full-length barcodes from specimens with varied degrees of DNA degradation or with particular primer mismatches (as evidenced by the lack of a certain product in the final sequence array). When DNA quality is poor, primer binding becomes increasingly important to 'kick start' amplification (Van Houdt *et al.* 2010). Perfect primer binding is impossible when diverse taxa are analysed, but the prospects for recovery of desired amplicons can be improved by raising the number of PCR cycles and by increasing the primer degeneracy (Hebert *et al.* 2013). Both tactics were employed in the present NGS protocol. Two rounds of PCR were employed, with 60 cycles in the first and 40 cycles in the second. All forward and reverse primers included degeneracy at the sites most important for primer binding (i.e. 3′ terminus). Considering this degeneracy, the 12 forward and reverse primers were actually a compilation of 2010 different oligonucleotides. Other factors were found to have important impacts on final outcomes. For example, initial tests revealed that primers with the 33- to 40-bp adapter/MID tails required for NGS were much less effective in generating product than the same primers without tails, a difference that was particularly strong for LQ extracts. This difference was probably due to interference with primer binding caused by the formation of secondary structures in the primers with tails. Although primers without tails produced the highest amplification success, their use allowed short, nontarget amplicons to act as longer primers on the strand opposite and immediately upstream of the original primer, generating chimeric amplicons that combined sequence information from primers and the specimen. To overcome this problem, 10-bp tails lacking complementarity to any region in the target genomes were added to the 5′ terminus of all primers. Their presence inhibited polymerase elongation when short amplicons or primer dimers attempted to act as primers, preventing the formation of chimeric amplicons while avoiding the secondary structure issues inherent with longer tails. Although the first round of PCR was effective in generating amplicons, a second round of PCR was required to introduce the adapter-tailed primers required for sequence analysis. It likely had the additional benefit of reducing amplification bias because it involved six separate reactions, one for each forward primer, dampening amplification bias by limiting primer competition (Berry *et al.* 2011).

Final NGS protocol

These background studies led to the development of a two-stage, nested, multiplex PCR protocol which produced sequence records spanning the barcode region. The first round of PCR included two reactions for each specimen (PCR 1.1 and PCR 1.2 in Fig. 1a), each consuming 2 μL of genomic DNA as template. Each reaction included three forward primers (F1 + F3 + F5 or F2 + F4 + F6) with six and five reverse primers, respectively, allowing each forward primer to generate from 1 to 6 amplicons, depending on the quality of DNA and its binding position in relation to the reverse primers. Detailed reaction components (final volume = 12.5 μL) are provided in Table S3 (Supporting information). Thermocycling consisted of 94 °C for 2 min, 60 cycles of 94 °C for 40 s, 48 °C for 40 s and 72 °C for 30 s and a final extension of 72 °C for 5 min.

The second round of PCR used product from the first PCRs as template and included six reactions per specimen (PCR 2.1–2.6 in Fig. 1a), each coupling a single forward primer with one to three reverse primers and using 2 μL of the appropriate primary PCR product as template. It boosted amplicon yields while also adding the required sequencing adapters. Each secondary PCR generated 1–3 amplicons which collectively spanned the COI barcode

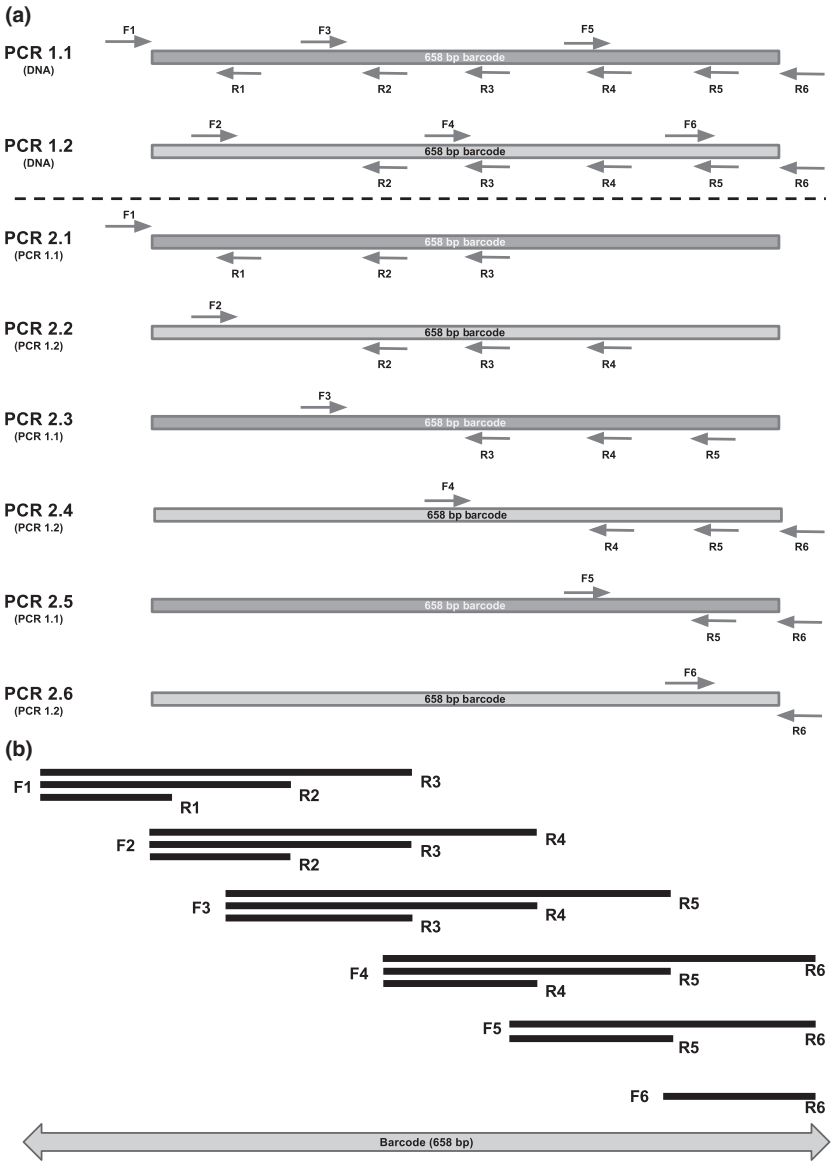

Fig. 1 Primer positions for the first and second rounds of PCR (a) and all possible final amplicons (b). The initial round of PCR includes two separate reactions (a – above broken line) using 10-bp tailed primers and genomic DNA as template (shown in parentheses below reaction names). The second round of PCR includes six separate reactions (a – below broken line) using adapter-tailed primers and the products from the first PCRs as template (shown in parentheses below reaction names). The second PCR can generate up to 15 amplicons – spanning the entire COI barcode region – which range in size from 119 bp to 366 bp (b). To assign each amplicon to a particular type specimen, each forward PCR2 primer is tailed with MID tags unique to that specimen. To assign each amplicon to a particular reaction (i.e. 2.1, 2.2, 2.3), each reverse PCR2 primer is tailed with a MID tag unique for each reaction in the second round of PCR.

region (Fig. 1b). The first four PCRs (2.1–2.4 in Fig. 1a) contained forward primers F1–F4, each combined with the three immediately downstream reverse primers (e.g. F1 + R1 + R2 + R3). The fifth PCR (2.5 in Fig. 1a) combined F5 with R5 and R6, while the sixth PCR (2.6 in Fig. 1a) combined F6 with R6. All of these reactions employed primers with adapter tails and MID tags to enable NGS to discriminate fragments and/or individuals in postprocessing. Detailed reaction components (final volume = 12.5 μL) are provided in Table S3 (Supporting information). Thermocycling consisted of 94 °C for 2 min, 40 cycles of 94 °C for 40 s, 48 °C for 40 s and 72 °C for 30 s, and a final extension of 72 °C for 5 min.

The secondary PCR products from each specimen (six reactions) were pooled, and a double size selection protocol (PCRClean DX kit – Aline Biosciences) was employed to remove genomic DNA, primer dimers and residual primers. The first cleanup step was designed to remove any high molecular weight genomic DNA (>800 bp) that might reflect recent contamination (e.g. human DNA from researchers working with the specimens). Briefly, the PCR product and magnetic beads were incubated in a 2:1 ratio (volume PCR product: volume beads) for 8 min at room temperature followed by 2 min on a magnet, to bind molecular weights >700 bp to the beads. The pellet of beads was discarded, while the supernatant (containing DNA <700 bp) was retained for the second cleanup step, designed to bind molecular weights >250 bp (i.e. the PCR products) to the beads, while lower molecular weight DNA (primer dimers, residual primers) remained in solution. This step was carried out by mixing enough beads and sterile water to

generate a 5:4 ratio (PCR product: beads) and incubated for 8 min followed by two minutes on a magnet. The supernatant was discarded and the pellet of beads was washed three times with 80% ethanol before the PCR products were eluted from the beads with 36 μL of sterile water. Following cleanup, the concentration of each purified PCR product was measured on a Qubit 2.0 spectrophotometer using the Qubit dsDNA HS Assay Kit (Life Technologies), and all 10 samples were normalized to 1 ng/μL and mixed in equal proportions. From this mixture, the final sequencing template library was created by making a 1/300 dilution. An Ion PGM Template OT2 400 kit (Life Technologies) was used for template preparation and sequencing was carried out on an Ion Torrent PGM following the manufacturer's instructions. Sequencing was performed using three 316 v2 chips – each loaded with libraries created from ten MID-tagged samples – using an Ion PGM Sequencing 400 Kit (Life Technologies).

Data analysis

Raw data from each Ion Torrent PGM run were uploaded to the Galaxy platform for analysis (https://usegalaxy.org/) (Blankenberg *et al.* 2010). Several filters (e.g. end trimming of reads, size selection, stringent alignment of reads to closely related taxa) were applied to remove low-quality, short and nontarget reads before an alignment was constructed to assemble the full barcode contig. Representative examples of the sequence reads recovered from HQ, and LQ extracts are shown in Fig S1 (Supporting information). The resultant FASTA file was then exported to permit comparisons with Sanger-generated sequences in BOLD. The authenticity of each NGS-generated sequence was subsequently validated by querying the sequence against the BOLD Identification Engine (www.boldsystems.org) to check for contamination or nontarget amplification. Further validation was performed via neighbour-joining (NJ) analysis that included the NGS-generated sequences as well as sequences from recently collected specimens of the same species or close relatives. The compiled reads from each run were deposited in the Sequence Read Archive (SRA; http://www.ncbi.nlm.nih.gov/sra) under study Accession SRP055961 (see Table S1, Supporting information for individual sample Accession nos), while the barcode contig for each specimen was deposited in the BOLD data set (dx.doi.org/10.5883/DS-NGSTYPES) and in GenBank (see Table S1, Supporting information for Accession nos). Additional sequence alignment and tree files have been archived on the Dryad digital repository (http://datadryad.org/) under the DOI: doi:10.5061/dryad.n1cg7.

Results

Because the NGS protocol allowed the simultaneous processing of ten specimens, just three runs were required to analyse the 30 specimens. The average number of sequence reads per specimen showed fivefold variation (182K, 59K, 36K), while the average depth of coverage per base showed sixfold variation (36K, 12K, 6K) across the three DNA categories (Fig. 2a,b). The number of reads per specimen averaged 90K, resulting in an average coverage depth of 18K per base. Sequences were recovered from every specimen with reads averaging 610 bp, 578 bp and 458 bp for the HQ, MQ and LQ extracts, respectively (Fig. 2c). Barcode compliant sequences (>487 bp) were recovered from eight HQ,eight MQ and four LQ specimens (Table S1, Supporting information), while sequence records >400 bp were recovered from 25 of 30 specimens (83%). In fact, more than 200 bp of sequence data was recovered from all 30 specimens (Table S1, Supporting information).

The sequences generated by NGS samples from the HQ and MQ specimens were perfectly matched in their zones of overlap to the shorter sequences generated by Sanger analysis (Fig. 3). Further confirmation of their validity was provided by the fact that they grouped with sequences from closely allied taxa (Fig. 4). It was more difficult to verify the sequences obtained via NGS from the LQ specimens because they had no Sanger counterparts for comparison. In six cases, the NGS sequences clearly derived from a single species, but reads from the other four specimens appeared to originate from two or more species. Obvious contaminants (e.g. fungi, bacteria) were easily removed during postprocessing, but some sequences in these four records appeared to derive from closely allied species or pseudogenes. In principle, the contaminants and authentic sequences could be discriminated if reference sequences were available from modern specimens of these species, but they were not. Because the four specimens showing these admixtures generated the fewest sequence reads and the lowest depth of coverage (Table S1, Supporting information), it is likely that their DNA was heavily degraded. Once contemporary sequences for these species become available, it should be possible to recognize the authentic sequences.

Discussion

Many earlier studies have recovered sequence information from museum specimens, including beetles (Gilbert *et al.* 2007; Thomsen *et al.* 2009), flies (Dean & Ballard 2001; Van Houdt *et al.* 2010; Hernández-Triana *et al.* 2013), true bugs (Bluemel *et al.* 2011) and moths (Hebert *et al.* 2013; : Hausmann *et al.* 2009a,b; Lees *et al.* 2010).

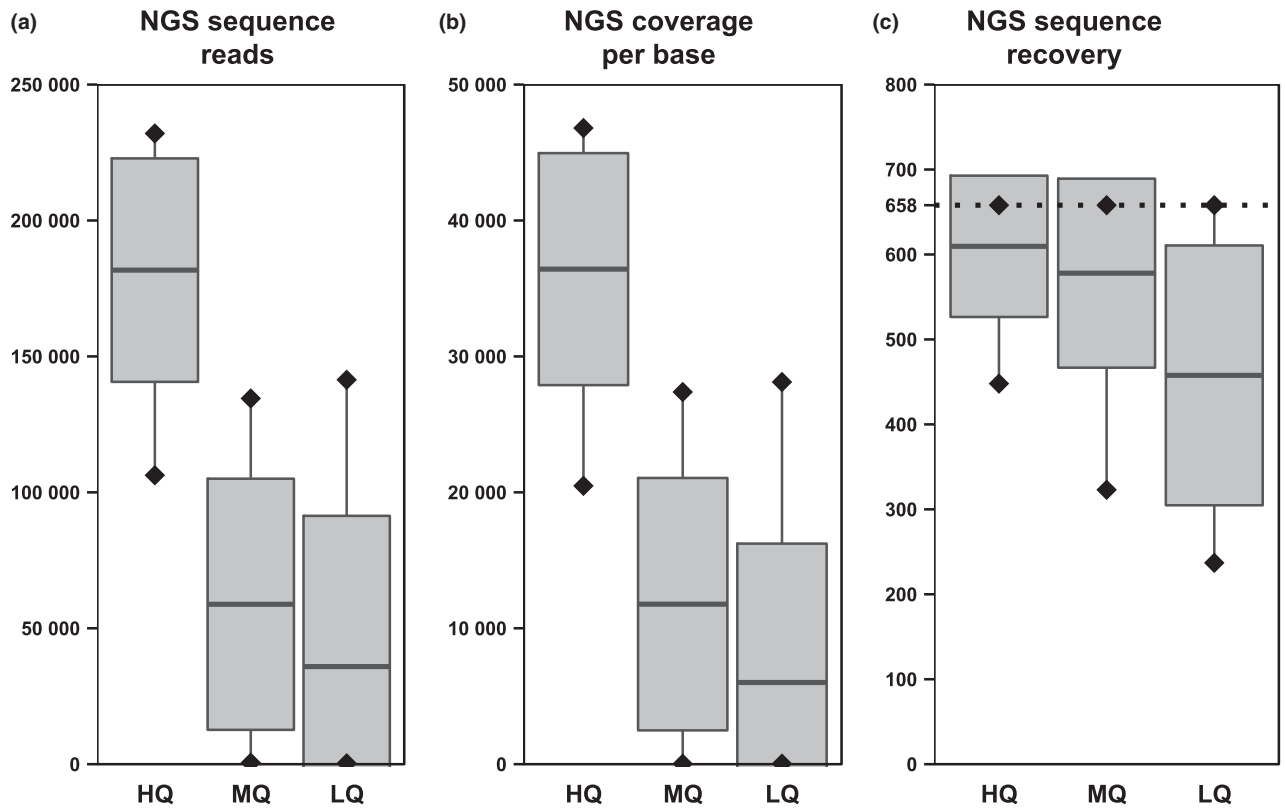

Fig. 2 Recovery of sequences from ten type specimens in each of three DNA categories. (a) Number of reads, (b) per base coverage and (c) number of base pairs (bp) recovered via NGS. HQ – high quality; MQ – medium quality, LQ – low quality. Mean (horizontal black line), standard deviation (edges of box), min and max (whiskers, ◆) are shown. The horizontal broken line in (c) represents a full-length (658 bp) barcode.

Some of these investigations analysed specimens that were relatively young (<50 years), while others extracted DNA from whole specimens. However, Hausmann *et al.* (2009a,b) and Rougerie *et al.* (2012) recovered barcode sequences from a single leg of type specimens more than 100 years old with a protocol that required six PCRs and twelve sequencing reactions (see details in Lees *et al.* 2010). Strutzenberger *et al.* (2012) reduced costs by processing specimens in batches of 95, but the basic protocol was unchanged, requiring substantial template DNA and careful inspection of data to ensure that contamination among wells had not produced chimeric sequences. As well, the failure of any single reaction led to an incomplete sequence for the barcode region. The present NGS protocol has the advantage of requiring very little template DNA and providing protection against the failure of any particular amplification reaction because of the initial multiplex PCR. The protocol does involve 100 cycles of PCR amplification, but there was no evidence of artefacts when the NGS sequences were compared to their Sanger counterparts (Fig. 3).

DNA quality is generally highest in young museum specimens, but exposure to certain killing or preservation agents can lead to its almost immediate degradation (Dean & Ballard 2001; Allentoft *et al.* 2012; Hebert *et al.* 2013). Because exposure histories are usually unknown, the assessment of DNA quality in museum specimens requires a screening procedure, such as the Sanger method used in this study. The present analysis revealed a rough correspondence between DNA quality as revealed by Sanger analysis and by sequence recovery through NGS. While DNA samples designated 'low quality' could in reality be high quality if their Sanger analysis failed due to primer-template mismatches – as opposed to DNA degradation – we find such a scenario unlikely because experiments using fresh DNA (not shown) demonstrated that the 164-bp and 94-bp primers function over a very broad taxonomic range and are unlikely to fail to amplify Lepidoptera DNA in the presence of a high quality template. The NGS protocol did recover more than 200 bp of sequence data from every specimen, but the records from four specimens with the lowest quality DNA included enough contamination to compromise data interpretation. However, future expansion of the barcode reference library will likely make it possible to disentangle sequences

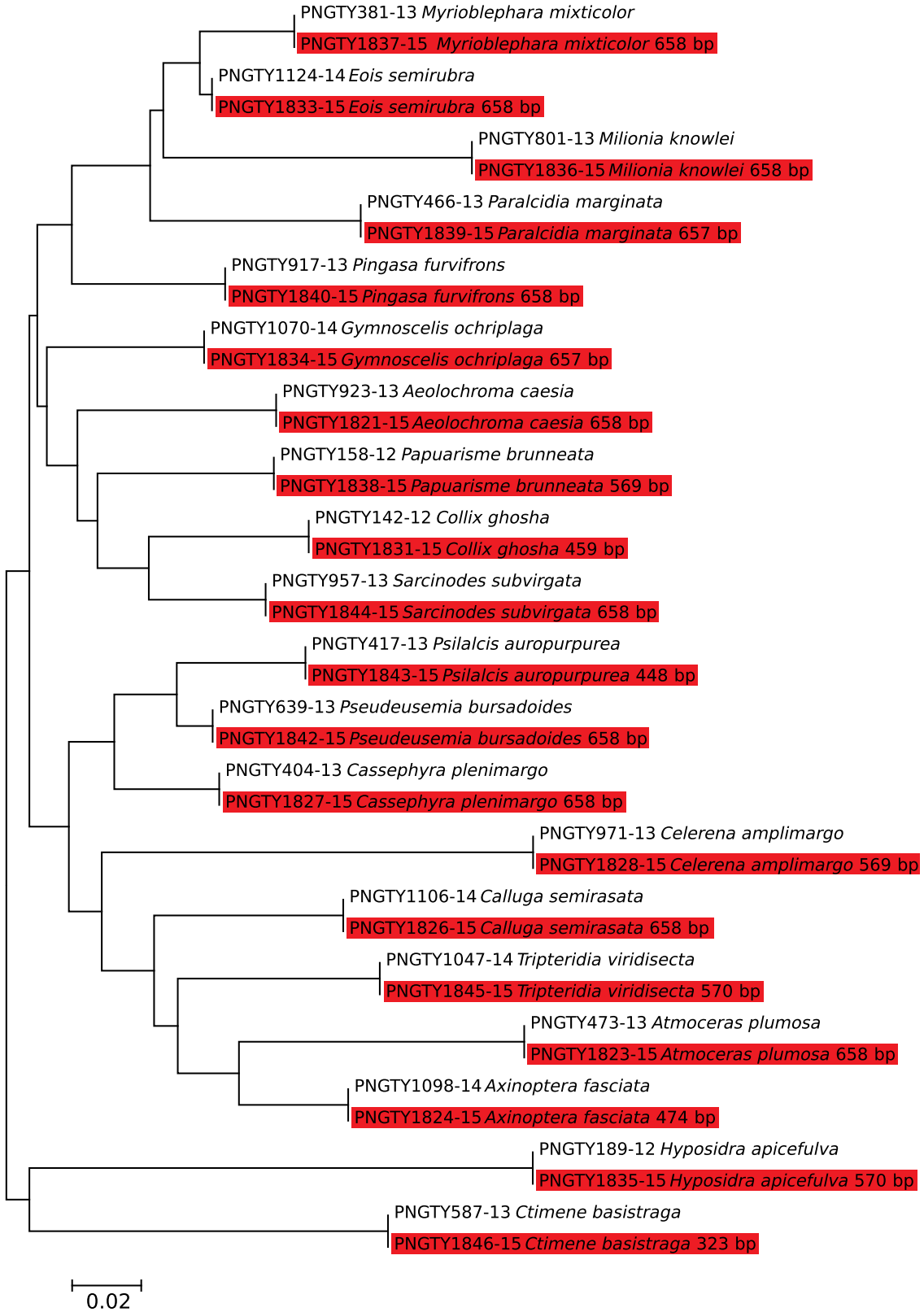

Fig. 3 Neighbour-joining tree showing 100% concordance between sequences generated from type specimens using NGS and Sanger sequencing. Taxonomy, BOLD Process IDs and/or total recovered base pairs are shown for the Sanger- (nonhighlighted) and NGS-generated (highlighted) sequences.

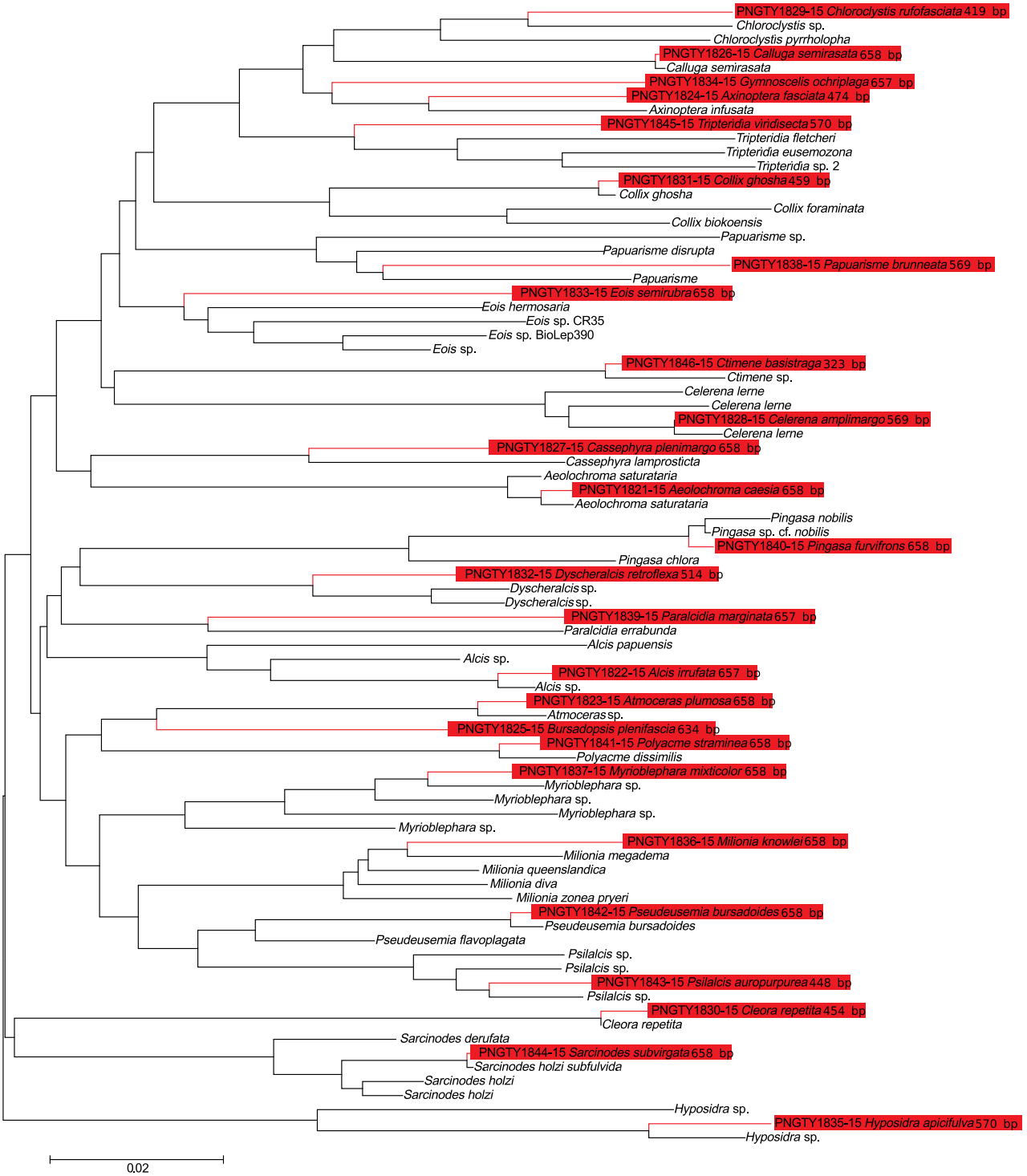

Fig. 4 Neighbour-joining tree of barcodes generated century-old type specimens and contemporary congeneric taxa (where available). Barcodes from the 26 century-old specimens (highlighted) were generated via NGS. BOLD Process IDs and total recovered base pairs are shown for each NGS-generated sequence. Four cases involve confirmed or suspected synonymy: *Celerna amplimargo* and *C. lerne*, *Aeolochroma caesia* and *A. saturataria*, *Sarcinodes subvirgata* and *S. holzi*, *Pingasa furvifrons* and *P. nobilis*.

derived from the type specimen in these cases from those arising through contamination. Alternatively, re-analysing another tissue source (either from the same specimen or from another type in the same series if possible) could provide more interpretable sequence data.

The superior performance of NGS in sequence recovery is likely due, in part, to the multiplex nature of the PCRs which allowed high primer redundancy. As a result, each DNA extract processed with the current protocol was exposed to amplification by 2010 primers vs. the 20 primers used in the Sanger analysis. The high diversity of primers undoubtedly meant that there was a greater chance of achieving the primer-template homology necessary for successful amplification (Pierce & Wangh 2015). The higher success of the NGS protocol is likely also a consequence of the greater sensitivity of these sequencing platforms. This difference was evidenced by the fact that 16 of 20 specimens, which failed to generate a 164-bp sequence via Sanger analysis, generated sequence reads for the same region with NGS. Although our results show that it was sometimes possible to recover a full-length COI barcode with NGS from specimens that failed with Sanger analysis, it is premature to declare that Sanger analysis is obsolete. In practice, it often generates short reads from type specimens that are taxonomically informative (Hajibabaei *et al.* 2006; Shokralla *et al.* 2011; Hebert *et al.* 2013; Hernández-Triana *et al.* 2013) at very low cost (< $1 per specimen).

Shokralla *et al.* (2014, 2015) used NGS to recover full-length barcodes from freshly collected specimens of Lepidoptera with a single primer pair. The present study extends their work by demonstrating that NGS can regularly recover complete or near-complete barcodes from century-old specimens with heavily degraded DNA. Moreover, because it requires little template DNA, much of each DNA extract remains for future analysis. Although analytical costs were approximately $50 a specimen, a 10-fold increase in the number of specimens processed in each run is feasible as the average coverage depth exceeded 20K per base in this study. With this minor shift in protocols or a move to a NGS platform generating more reads, the cost of sequence analysis could drop by an order of magnitude.

This study only analysed specimens of Geometridae, but the same primer sets have recovered sequences from type specimens in other families of Lepidoptera (e.g. Spiedel *et al.* 2015). These cases of success, combined with the extensive primer degeneracy of the multiplex PCRs, suggest the current protocol may be generally effective for insects. If failures are encountered, they should be easily overcome by developing new primer sets, a task facilitated by the well-parameterized barcode reference library for the animal kingdom. NGS has the power to sequence genomes, but this study has demonstrated its value in probing sequence diversity in single gene regions. A large-scale programme to sequence type specimens would represent a major advance in stabilizing and validating the application of scientific names. As well, because many type specimens derive from developing nations, it would represent an important step in the repatriation of knowledge that will aid these nations in managing their biodiversity by enabling DNA-powered identification systems, a major advance in settings where the scientific workforce is small and biodiversity is high.

Acknowledgements

This research was primarily funded by a grant to PDNH and SEM from the Gordon and Betty Moore Foundation. Funding from the government of Canada through Genome Canada and the Ontario Genomics Institute to the International Barcode of Life Project also aided the work. Support from the Ontario Ministry of Research and Innovation enabled the development of BOLD, while the Canada Foundation for Innovation provided key research infrastructure. Imaging and specimen lysis was supported by the US National Institutes of Health through ICBG 5UO1TW006671 granted to SEM. Building the Papua New Guinea background library has been supported by the US National Science Foundation (via grants DEB-0211591, 0515678 and others to SEM). We thank the Natural History Museum and its staff (particularly Jacqueline Mackenzie-Dodds, Geoff Martin and John Chainey) for their support and for permitting access to type specimens. We are grateful to David Pollock, Margaret Rosati and Jeremy Holloway who aided tissue sampling and/or taxonomic validation. We also thank Nataly Ivanova, Evgeny Zakharov and Sujeevan Ratnasingham who contributed to protocol development and sequence analyses. Thermo Fisher Scientific provided advice on protocols for the Ion Torrent platform, but did not influence experimental design, data analysis or data interpretation.

References

Allentoft ME, Collins M, Harker D *et al.* (2012) The half-life of DNA in bone: measuring decay kinetics in 158 dated fossils. *Proceedings of Biological Sciences*, **279**, 4724–4733.

Antoni MY, Delpiani SM, Stewart AL, González-Castro M, Astarloa JM (2015) Merluccius tasmanicus Matallanas & Lloris 2006 is a junior synonym of M. australis (Hutton 1872) (Gadiformes: Merluciidae) based on morphological and molecular data. *Zootaxa*, **3956**, 29–55.

Berry D, Mahfoudh KB, Wagner M, Loy A (2011) Barcoded primers used in multiplex amplicon pyrosequencing bias amplification. *Applied and Environment Microbiology*, **77**, 7846–7849.

Blankenberg D, Von KG, Coraor N *et al.* (2010) Galaxy: a web-based genome analysis tool for experimentalists. *Current Protocols in Molecular Biology*, Chapter 19, Unit 19, 10, 11–21.

Bluemel JK, King RA, Virant-Doberlet M, Symondson WOC (2011) Primers for identification of type and other archived specimens of *Aphrodes* leafhoppers (Hemiptera, Cicadellidae). *Molecular Ecology Resources*, **11**, 770–774.

Dabney J, Meyer M, Paabo S (2013) Ancient DNA damage. *Cold Spring Harbor Perspectives in Biology*, **5**, a012567.

Dean MD, Ballard JWO (2001) Factors affecting mitochondrial DNA quality from museum preserved *Drosophila simulans*. *Entomologia Experimentalis et Applicata*, **98**, 279–283.

Gilbert MTP, Moore W, Melchior L, Worobey M (2007) DNA extraction from dry museum beetles without conferring external morphological damage. *PLoS ONE*, **2**, e272.

Hajibabaei M, Smith MA, Janzen DH, Rodriguez JJ, Whitfield JB, Hebert PDN (2006) A minimalist barcode can identify a specimen whose DNA is degraded. *Molecular Ecology Notes*, **6**, 959–964.

Hausmann A, Hebert PDN, Mitchell A, Rougerie A, Sommerer M, Edwards T (2009a) Revision of the Australian *Oenochroma vinaria* Guenée, 1858 species-complex (Lepidoptera: Geometridae, Oenochrominae): DNA barcoding reveals cryptic diversity and assesses status of type specimen without dissection. *Zootaxa*, **2239**, 1–21.

Hausmann A, Sommerer M, Rougerie R, Hebert P (2009b) *Hypobapta tachyhalotaria n. sp.* from Tasmania – an example of a new species revealed by DNA barcoding (Lepidoptera, Geometridae). *Spixiana*, **32**, 161–166.

Hebert PDN, Cywinska A, Ball S, deWaard JR (2003) Biological identifications through DNA barcodes. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, **270**, 313–321.

Hebert PDN, deWaard JR, Zakharov EV *et al.* (2013) A DNA 'Barcode Blitz': rapid digitization and sequencing of a natural history collection. *PLoS ONE*, **8**, e68535.

Hernández-Triana LM, Prosser SW, Rodríguez-Perez MA, Chaverri LG, Hebert PDN, Gregory TR (2013) 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.

Holloway JD, Miller SE, Pollock DM, Helgen L, Darrow K (2009) GONGED (Geometridae of New Guinea Electronic Database): a progress report on development of an online facility of images. *Spixiana*, **32**, 122–123.

Hughey JR, Gabrielson PW, Rohmer L *et al.* (2013) Minimally destructive sampling of type specimens of *Pyropia* (Bangiales, Rhodophyta) recovers complete plastid and mitochondrial genomes. *Scientific Reports*, **4**, 5113.

Ivanova NV, deWaard JR, Hebert PDN (2006) An inexpensive, automation-friendly protocol for recovering high-quality DNA. *Molecular Ecology Notes*, **6**, 998–1002.

Johnson JW, Wilmer JW (2015) Plectorhinchus caeruleonothus, a new species of sweetlips (Perciformes: Haemulidae) from northern Australia and the resurrection of P. unicolor (Macleay, 1883), species previously confused with P. schotaf (Forsskål, 1775). *Zootaxa*, **13985**, 491–522.

Kirchman JJ, Witt CC, McGuire JA, Graves GR (2010) DNA from a 100-year-old holotype confirms the validity of a potentially extinct hummingbird species. *Biology Letters*, **6**, 112–115.

Knolke S, Erlacher S, Hausmann A, Miller MA, Segerer AH (2005) A procedure for combined genitalia dissection and DNA extraction in Lepidoptera. *Insect Systematics and Evolution*, **35**, 401–409.

Kvist S, Oceguera-Figueroa A, Siddall ME, Erséus C (2010) Barcoding, types and the Hirudo files: using information content to critically evaluate the identity of DNA barcodes. *Mitochondrial DNA*, **21**, 198–205.

Lees DC, Rougerie R, Zeller-Lukashort C, Kristensen NP (2010) DNA mini-barcodes in taxonomic assignment: a morphologically unique new homoneurous moth clade from the Indian Himalayas described in *Micropterix* (Lepidoptera, Micropterigidae). *Zoologica Scripta*, **39**, 642–661.

Lees DC, Lack HW, Rougerie R *et al.* (2011) Tracking origins of invasive herbivores using herbaria and archival DNA: the case of the horse-chestnut leaf miner. *Frontiers in Ecology and the Environment*, **9**, 322–328.

Liimatainen K, Niskanen T, Dima B, Kytövuori I, Ammirati JF, Frøslev TG (2014) The largest type study of Agaricales species to date: bringing identification and nomenclature of Phlegmacium (Cortinarius) into the DNA era. *Persoonia*, **33**, 98–140.

Miller SE (2014) DNA barcode enabled ecological research on Geometridae in Papua New Guinea. *Spixiana*, **37**, 245–246.

Mutanen M, Kekkonen M, Prosser SW, Hebert PDN, Kaila L (2014) One species in eight: DNA barcodes from type specimens resolve a taxonomic quagmire. *Molecular Ecology Resources*, **15**, 967–984.

Pierce KE, Wangh LJ (2015) Low-concentration initiator primers improve the amplification of gene targets with high sequence variability. *Methods in Molecular Biology*, **1275**, 73–89.

Porco D, Potapov M, Bedos A *et al.* (2012) Cryptic diversity in the ubiquist species Parisotoma notabilis (Collembola, Isotomidae): a long-used chimeric species? *PLoS ONE*, **7**, e46056.

Rougerie R, Naumann S, Nassig WA (2012) Morphology and molecules reveal unexpected cryptic diversity in the enigmatic genus *Sinobirma* Bryk, 1944 (Lepidoptera: Saturniidae). *PLoS ONE*, **7**, e43920.

Rowe KC, Singhal S, Macmanes MD *et al.* (2011) Museum genomics: low-cost and high-accuracy genetic data from historical specimens. *Molecular Ecology Resources*, **11**, 1082–1092.

Shokralla S, Zhou X, Janzen DH *et al.* (2011) Pyrosequencing for mini-barcoding of fresh and old museum specimens. *PLoS ONE*, **6**, e21252.

Shokralla S, Gibson JF, Nikbakht H, Janzen DH, Hallwachs W, Hajibabaei M (2014) Next-generation DNA barcoding: using next-generation sequencing to enhance and accelerate DNA barcode capture from single specimens. *Molecular Ecology Resources*, **14**, 892–901.

Shokralla S, Porter T, Gibson J *et al.* (2015) Massively parallel multiplex DNA sequencing for specimen identification using an Illumina MiSeq platform. *Scientific Reports*, **5**, 9687.

Spiedel W, Hausmann A, Muller GC *et al.* (2015) Taxonomy 2.0: next generation sequencing of old type specimens supports the description of two new species of the *Lasiocampa decolorata* group from Morocco (Lepidoptera: Lasiocampidae). *Spixiana*, **3**, 401–412.

Strutzenberger P, Brehm G, Fiedler K (2012) DNA barcode sequencing from old type specimens as a tool in taxonomy: a case study in the diverse genus *Eois* (Lepidoptera: Geometridae). *PLoS ONE*, **7**, e49710.

Thomsen PF, Elias S, Gilbert MTP *et al.* (2009) Non-destructive sampling of ancient insect DNA. *PLoS ONE*, **4**, e5048.

Tin MM-Y, Economo EP, Mikheyev AS (2014) Sequencing degraded DNA from non-destructively sampled museum specimens for RAD-tagging and low-coverage shotgun phylogenetics. *PLoS ONE*, **9**, e96793.

Van Houdt JKJ, Breman FC, Virgilio M, De Meyer M (2010) Recovering full DNA barcodes from natural history collections of Tephritid fruit-flies (Tephritidae, Diptera) using mini barcodes. *Molecular Ecology Resources*, **10**, 459–465.

deWaard JR, Ivanova NV, Hajibabaei M, Hebert PDN (2008) Assembling DNA barcodes. *Methods in Molecular Biology*, **410**, 275–293.

Zimmermann J, Hajibabaei M, Blackburn DC *et al.* (2008) DNA damage in preserved specimens and tissue samples: a molecular assessment. *Frontiers in Zoology*, **5**, 18.

S.W.J.P., J.R.D. and P.D.N.H. planned and designed the experiments. S.E.M. provided the specimens, which were sampled by J.R.D. S.W.J.P. developed and executed the laboratory work. S.W.J.P. and J.R.D. analysed the experimental data. S.W.J.P., J.R.D., S.E.M. and P.D.N.H. wrote and edited the manuscript.

Data accessibility

Sanger sequences, electropherograms and specimen details: dx.doi.org/10.5883/DS-NGSTYPES.

NGS contig sequences: GenBank Accessions KR070762 - KR070787.

NGS raw reads: SRA Accessions SRR1867808, SRR1867811 - SRR1867819, SRR1867935 - SRR1867944, SRR1945335, SRR1945382 - SRR1945389, SRR1946575.

Sanger and NGS sequence alignments, and tree files: doi:10.5061/dryad.n1cg7.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Figure S1. Alignments of sequence records derived from two type specimens of Geometridae, one with high quality DNA (a) and one with low quality DNA (b).

Table S1. Type specimens analyzed, including sequencing results and accession numbers

Table S2. Primers used in the first (PCR1) and second (PCR2) reactions to allow the analysis of 10 specimens in an Ion Torrent PGM run

Table S3. Components of PCR reactions in the NGS protocol