DOI: https://doi.org/10.3897/arphapreprints.e84305

Enhancing DNA barcode reference libraries by harvesting terrestrial arthropods at the National Museum of Natural History

Bernardo Santos, Meredith Miller, Margarita Miklasevskaja, Jaclyn McKeown, Niamh Redmond, Jonathan Coddington, Jessica Bird, Scott Miller, Ashton Smith, Seán Brady, Matthew Buffington, M. Lourdes Chamorro, Torsten Dikow, Michael Gates, Paul Goldstein, Alexander Konstantinov, Robert Kula, Nicholas Silverson, M. Alma Solis, Stephanie deWaard, Suresh Naik, Nadya Nikolova, Mikko Pentinsaari, Sean Prosser, Jayme Sones, Evgeny Zakharov, Jeremy deWaard

ARPHA Preprints *Author-formatted, not peer-reviewed document posted on 28/03/2022*. DOI: https://doi.org/10.3897/arphapreprints.e84305

Enhancing DNA barcode reference libraries by harvesting terrestrial arthropods at the National Museum of Natural History

Bernardo Santos[‡,§], Meredith E. Miller[|], Margarita Miklasevskaja[|], Jaclyn T.A. McKeown[|], Niamh E. Redmond[§], Jonathan A. Coddington[§], Jessica Bird[§], Scott E. Miller[§], Ashton Smith[§], Seán G. Brady[§], Matthew L. Buffington[¶], M. Lourdes Chamorro[¶], Torsten Dikow[§], Michael W. Gates[¶], Paul Goldstein[¶], Alexander Konstantinov[¶], Robert Kula[¶], Nicholas D. Silverson[§], M. Alma Solis[¶], Stephanie L. deWaard[|], Suresh Naik[|,#], Nadya Nikolova[|], Mikko Pentinsaari[|], Sean W.J. Prosser[|], Jayme E. Sones[|], Evgeny V. Zakharov[|,#], Jeremy R. deWaard[|,§,¤]

‡ Institut de Systématique, Evolution, Biodiversité (ISYEB), Muséum National d'Histoire naturelle, CNRS, SU, EPHE, UA, Paris, France
§ National Museum of Natural History, Smithsonian Institution, Washington, United States of America
| Centre for Biodiversity Genomics, University of Guelph, Guelph, Canada
¶ Systematic Entomology Laboratory, Beltsville Agricultural Research Center, Agricultural Research Service, U.S. Department of Agriculture, Washington, United States of America
# Department of Integrative Biology, University of Guelph, Guelph, Canada
¤ School of Environmental Sciences, University of Guelph, Guelph, Canada

Corresponding author: Bernardo Santos (bernardofsantos@gmail.com)

Abstract

The use of DNA barcoding has revolutionized biodiversity science, but its application depends on the existence of comprehensive and reliable reference libraries. For many poorly known taxa, such reference sequences are missing even at higher-level taxonomic scales. We harvested the collections of the Smithsonian's National Museum of Natural History (USNM) to generate DNA barcoding sequences for genera of terrestrial arthropods previously not recorded in one or more major public sequence databases. Our workflow used a mix of Sanger and Next-Generation Sequencing (NGS) approaches to maximize sequence recovery while ensuring affordable cost. In total, COI sequences were obtained for 5,686 specimens belonging to 3,888 genera and 202 families. Success rates varied widely according to collection data and focal taxon. NGS helped recover sequences of specimens that failed a previous run of Sanger sequencing. Success rates and the optimal balance between Sanger and NGS are the most important drivers to maximize output and minimize cost in future projects. The corresponding sequence and taxonomic data can be accessed through the Barcode of Life Data System, GenBank, the Global Biodiversity Information Facility, the Global Genome Biodiversity Network Data Portal and the NMNH data portal.

Keywords

COI, cox1, dark taxa, OTUs, BINs, natural history collection, museum harvesting

Introduction

The use of DNA barcoding has revolutionized how biodiversity can be surveyed and identified, with applications in fields as broad as biodiversity assessment, invasive species monitoring, agricultural pest control, identification of disease vectors, integrative taxonomy and evolutionary studies (Hubert and Hanner 2015). However, the accuracy of DNA barcoding identifications depends to a large degree on the availability of comprehensive reference libraries, which allow the assignment of scientific names to operational taxonomic units (OTUs) delimited by analysis of barcoding sequences. The construction of reliable reference libraries, often region- or taxon-specific, has received a lot of attention in recent years (e.g., Hawlitschek et al. (2015), Raupach et al. (2014), Morinière et al. (2017), Porco et al. (2018)). In spite of these advances, assembling reference libraries that can support robust identifications at a broad scale is still challenging for poorly known taxa such as insects and other terrestrial arthropods with extremely high species number. Identification tools applicable to physical vouchers are often lacking, and many taxa (including genera) are known only from a few specimens, often collected decades or even over a century ago.

In the face of these challenges, one of the most promising avenues for building comprehensive reference libraries is directly harvesting museum specimens that are authoritatively determined (Puillandre et al. 2012, Hebert et al. 2013, Mitchell 2015, Chambers and Hebert 2016, Sire et al. 2019, Rinkert et al. 2021). Major natural history museums often harbor specimens from several thousands of determined species and can support a considerable increase in the availability of reliable entries for barcode reference libraries. The use of such collections, however, is not free of challenges; the sheer scale of collections, diversity of storing and preserving techniques across taxa, and the old age of many specimens poses the need to develop optimized, logistic protocols and molecular techniques to amplify and sequence barcoding fragments from often degraded material.

The Smithsonian Institution’s National Museum of Natural History (USNM) comprises the largest natural history collection in the world, with a large portion of its holdings represented by terrestrial invertebrates. For many taxa, the USNM holds the most complete inventory of species of any collection in the world, and the vast majority of invertebrate orders have a complete inventory of the holdings at species level. These qualities make it ideally suited to contribute to the general effort of building a global reference library for DNA barcodes, especially for “dark taxa” not otherwise represented in repositories such as GenBank (Benson et al. 2012; https://www.ncbi.nlm.nih.gov/genbank/), the Barcode of Life Data System (BOLD; Ratnasingham and Hebert 2007; http://

www.boldsystems.org), or Global Genome Biodiversity Network (GGBN; Droege et al. 2014).

Herein we report results of the project “Barcoding NMNH terrestrial invertebrate genera”, which aims to generate DNA barcoding sequences for genera not previously represented on GenBank, BOLD or GGBN, and to initiate the long-term preservation of publicly-accessible genomic DNA extracts and high-resolution images to accompany the physical USNM vouchers. We describe the operational protocol employed, provide statistics and metrics for the results of the project to date, and discuss these in the context of the general utility of museum collections in the generation of reference libraries and supporting resources.

Material and methods

Specimen Selection and USNM Loan Organization

In 2018 and 2019, staff from the Centre for Biodiversity Genomics (CBG) completed six visits (46 days total) to the Smithsonian Institution’s National Museum of Natural History, Department of Entomology (USNM). Prior to each visit, a number of target taxa, such as families or superfamilies, were defined based on number of available specimens, level of curation and physical localization in the museum. Available species inventories for target taxa were compared with the holdings of GenBank and BOLD using a custom application, the GGI Gap Analysis Tool (Global Genome Initiative 2019) to define target genera for sampling. Over the six visits, 8,549 specimens were selected and loaned. Two representatives of different species for each target genus (whenever possible) were selected. Curator specifications, specimen age, collection method, preservation method, number of specimens per genus within the collection, and taxonomy were used to determine the appropriate extraction and sequencing protocols for each specimen. Overall, 7,589 specimens were selected for analysis using the CBG’s Sanger-based sequencing protocol (Ivanova et al. 2006), and 950 specimens were selected for a protocol involving Next-Generation Sequencing (NGS; also referred to as High-Throughput Sequencing or HTS) (Hebert et al. 2013, Prosser et al. 2015). Of the 7,589 specimens selected for Sanger sequencing, 380 specimens were processed using whole voucher specimens, and 7,219 specimens were processed using a tissue sample (leg). Of the 950 specimens selected for NGS, 184 specimens were processed using whole voucher specimens (usually minute Hymenoptera specimens), and 766 specimens were processed using a tissue sample (typically a leg). Specimens were loaned to CBG for processing and sequencing following the 'museum harvesting' protocol outlined in Levesque-Beaudin et al. (In press). Specimen data including taxonomy, country of collection, sample ID, and specimen cabinet/drawer locations within the USNM collection were recorded by CBG staff at the time of loan organization.

Imaging, Digitization, Tissue Sampling and Sequencing

At the end of each visit, specimens were transferred to CBG for processing. Each specimen was assigned a sample ID, accession number and labelled with a Barcode of Life Data Systems (BOLD) (Ratnasingham and Hebert 2007) specimen label, as well as an unique specimen identifier (USNM ENT) label. Digitization, imaging, and sub-sampling were completed following the protocol outlined in Levesque-Beaudin et al. (submitted), following predetermined specifications by USNM museum curators for each taxonomic group. After digitization, imaging, and sub-sampling were complete, data and images were uploaded to BOLD in projects organized by project year and visit (Table A in Suppl. material 1). DNA samples were extracted using the silica-based protocol outlined in Ivanova et al. (2006), PCR amplification and sequencing following protocols detailed in Hebert et al. (2013), Prosser et al. (2015), and D'Ercole et al. (2021) that target overlapping fragments of the cytochrome c oxidase I (COI) gene. Following sequence editing, sequences were uploaded to BOLD in the appropriate project. Following BOLD upload, DNA extracts were spilt (20µl each) with one half stored in the CBG DNA archive and the other sent to the USNM Biorepository. All voucher specimens from the six visits and loans were returned to their original locations within the USNM collection following the protocol outlined in Levesque-Beaudin et al. (In press).

NGS Failure Tracking

A list of genera sampled in Year 1 (Fig. 1) that failed to yield sequences (0 bp) using the Sanger protocol was compiled, and 475 specimens were selected for Next Generation Sequencing (NGS) processing and sequencing (NGSFT Round 1). An additional list of genera sampled in Year 1 and Year 2 that failed to yield sequences (0 to 300 bp) using both the Sanger and NGS protocol was compiled, including 143 specimens that failed to yield sequences after the initial round of NGS failure tracking. In NGSFT Round 2, 1013 specimens were selected for NGS processing and sequencing (Fig. 1). Specimen selection was based on genera that would generate the maximum number of unique new GenBank records. These DNA samples (which were stored in CBG's DNA Archive) underwent an NGS pipeline that targets overlapping fragments (as outlined in Prosser et al. (2015)) and sequences the amplicons on the PacBio Sequel II (Quicke et al. 2020, D'Ercole et al. 2021). After sequencing and validation was complete, sequencing data were uploaded and stored on BOLD associated with each individual sample ID.

Data and Other Resources

All successfully sequenced records from all BOLD projects (>200 bp) were made public and submitted to GenBank. USNM voucher information is listed in the "specimen voucher" field of all GenBank records, ensuring the correct linkage with records in the USNM EMu Collection Management System (https://collections.nmnh.si.edu/search/ento). CBG provided the USNM Entomology Data Manager all GenBank Accession numbers, DNA bank data (following the GGBN Data Standard; Droege et al. (2016)) and specimen images which were submitted to the USNM EMu collection management system. After sequence validation was complete, the successfully sequenced records were added to the BOLD dataset DS-NMNHSEQ titled 'Barcoding NMNH Terrestrial Arthropod Genera' (http://dx.doi.org/10.5883/DS-NMNHSEQ).

Data resources

The specimen data, images and sequencing data for all 8,549 specimen records are available on BOLD in the public dataset DS-NMNHALL (http://dx.doi.org/10.5883/DS-NMNHALL) and searchable in the Public Data Portal on BOLD (www.boldsystems.org/index.php/Public_BINSearch) or downloadable by utilizing BOLD's API (www.boldsystems.org/index.php/resources/api).

Specimen records include taxonomy, collection date and location, USNM ENT identifiers, EZID reference numbers (corresponding to EMu-minted records that have globally-unique identifier status), BINs, and any additional voucher specimen details. All specimen images are publicly available under the Creative Commons No Rights Reserved (CC0 1.0) license. All data was submitted and stored in the USNM EMu collection management system and individual records are accessible at https://collections.nmnh.si.edu/search/ento/. Specimen data and DNA storage information were submitted to the Global Genome Biodiversity Network (GGBN) Data Portal (Droege et al. 2014; https://www.ggbn.org/ggbn_portal/search/result?voucherCol=NMNH%2C+Washington).

All sequences have been submitted to GenBank; the dataset can be accessed through NCBI's BioProject PRJNA81359 (https://www.ncbi.nlm.nih.gov/bioproject/81359). All specimen data have also been uploaded to the Global Biodiversity Information Facility (GBIF; http://www.gbif.org) in the 'NMNH Extant Specimen Records (USNM, US)' occurrence dataset (https://doi.org/10.15468/hnhrg3). DNA extracts derived from sequenced specimens are held in the CBG DNA Archive (as specified in deWaard et al. 2019) and in the NMNH Biorepository (https://naturalhistory.si.edu/research/biorepository).

Results

A complete list of the 8,549 specimens (including USNM ENT IDs, Process IDs, BOLD ID's, COI sequence length, country of origin, collection date and taxonomy) is provided in Suppl. material 1. Specimens represent 13 orders, 210 families, 4,510 genera and 4,864 identified species collected from 148 countries. In total, 8,549 label images and 12,096 specimen images (TIF format) were completed by CBG imaging technicians.

Of the 4,510 selected genera, 879 genera were represented by 1 specimen, 3,428 genera were represented by 2 specimens, 102 genera were represented by 3 specimens, 75 genera were represented by 4 specimens and the remaining 26 genera were represented by 5 or more specimens. At the time of specimen selection (Table A in Suppl. material 1)

4,415 of the 4,510 selected genera were new to GGBN, 4,117 were new to GenBank and 2,696 were new to BOLD.

NGS-based failure-tracking was conducted in two stages (Fig. 1). In round 1,475 specimens that failed to gain a sequence (0 bp) using the Sanger method (Table 1) were sequenced using next generation sequencing, resulting in 310 recovered sequences (>0 bp). Of the 310 specimens that gained a sequence, 300 were of acceptable length (or 'acceptable bacodes', here defined as >300 bp), resulting in a success rate of 63.2% (Table 2). In round 2 of NGS failure tracking, 1,013 specimens with sequences between 0 and 300 bp were selected, these included 145 specimens that failed to gain a sequence (0 bp) in round 1 of NGS FT (Fig. 1). Round 2 of NGSFT resulted in 674 recovered sequences (>0 bp). Of the 674 recovered sequences, 501 were acceptable barcodes (>300 bp), and a success rate of 49.5% (Table 2).

After NGS-based failure tracking, overall sequence recovery by specimen was 66.5% (5,686 of 8,549 records gained a sequence (>0 bp) (Table 3). Of the 5,686 records that gained a sequence, 5,220 (61.1%) were acceptable barcodes (>300 bp) with 3,278 records with sequences 500 bp or greater. Specimen collection dates (by decade) and corresponding sequencing success rates are plotted in Fig. 2.

After NGS-based failure tracking, overall sequence recovery by specimen was 66.5% (5,686 of 8,549 records gained a sequence (>0 bp) (Table 3). Of the 5,686 records that gained a sequence, 5,220 (61.1%) were acceptable barcodes (>300 bp) with 3,278 records with sequences 500 bp or greater. Specimen collection dates (by decade) and corresponding sequencing success rates are plotted in .

Of the 4,510 selected genera, 3,888 gained a sequence >0 bp (86.2%), with 3,641 genera gaining a sequence that was an acceptable barcode (>300 bp), resulting in a success rate of 80.7% (Table 4). In total, COI sequences (>0 bp) were obtained for 5686 specimens belonging to 3743 species, 3888 genera and 202 families. The sequences of acceptable barcodes (>300 bp) constitute 2,433 BINs on BOLD (Ratnasingham and Hebert 2007), with 1,925 unique BINs (79.1%) added to BOLD from this project.

Sequence recovery (>0 bp) for all selected taxonomic groups (Orders) was between 81.48% and 93.98% (Fig. 3, Table 4). Sequence success by genus for each taxonomic group (>300 bp) was between 64.81% and 91.97%. The group defined as “Other Orders”, (consisting of specimens of Araneae, Mecoptera, Megaloptera, Neuroptera, Odonata, Plecoptera, Raphidioptera and Trichoptera), has the greatest genus sequencing success of all taxonomic groups with 91.97%, followed by Lepidoptera (86.27%), Coleoptera (83.18%), Diptera (83.02%), Hemiptera (80.68%), Hymenoptera (73.25%) and Araneae (64.81%), respectively (Table 4).

Hymenoptera specimens were sequenced using a sample of leg tissue (1,542/2,017 specimens, representing 823 Hymenoptera genera) or using the whole voucher (475/2,017 total specimens, representing 253 Hymenoptera genera) (Table 5). Prior to NGS failure tracking, for specimens sequenced using a leg tissue sample, sequence recovery using the

Sanger protocol was 48.40% (865 specimens with sequences >0 bp), and specimens sequenced with NGS was 65.13% (127 specimens with sequences >0 bp). For specimens sequenced using the whole voucher, sequence recovery using the Sanger protocol was 47.37% (180 specimens with sequences >0 bp), and specimens sequenced with NGS was 63.16% (60 specimens with sequences >0 bp). Prior to NGS failure tracking, genus sequence recovery for leg tissue (using Sanger and NGS protocols combined) was 51.26% (428 of 835 genera >300 bp), and genus sequence recovery for whole voucher was 46.51% (120 of 258 genera >300 bp). After NGS failure tracking was complete, genus sequence recovery for leg tissue increased to 77.13 % (644 of 835 genera >300 bp), and genus sequence recovery for whole voucher increased to 60.47% (156 of 258 genera >300 bp; Table 6).

Discussion

The persistent scarcity of reliable reference libraries for many poorly known invertebrate taxa has been a growing concern, reflected in the recent emergence of specific projects and initiatives aimed specifically at such groups, such as “GBOL III: Dark Taxa” by the German Barcode of Life Initiative (Rduch and Peters 2020). Our study intentionally targeted genera that were not represented in existing public databases of barcode sequences, keeping in line with the Global Genome Initiative’s objective of increasing barcode representation along the major branches of the Tree of Life.

Using authoritatively identified material from one of the most prominent natural history collections in the world, we were able to provide novel DNA barcoding data for thousands of genera which had not yet been sequenced, and for 3,743 determined species of terrestrial arthropods. This data release represents not only an important advance in the availability of species-level reference barcodes for several taxa but also has the potential to assist genus-level identifications for groups in which reference sequences are sorely lacking. These results were attained by using a workflow that combines on-site sampling with off-site processing of specimens and DNA extracts (Levesque-Beaudin et al. In press), with the use of the high-throughput infrastructure at the CBG allowing for protocol standardization and gains of scale in terms of cost and output.

The laboratory protocol used for this study was primarily based on Sanger sequencing, with an NGS pipeline used as an alternative method to recover sequences for very old or small taxa, or to specifically target samples that had failed to sequence using the Sanger-based methodology. In our case, this increased overall success. As costs associated with NGS processing continue to decline (National Human Genome Research Institute 2019), we envision a point where our hybrid approach will no longer be cost-effective compared to NGS alone.

In strict terms, matching cost levels are achieved when the difference in cost (C) per specimen between NGS and Sanger approaches matches the difference in success rate, or efficiency (E) between the two approaches (i.e., when $C_{Sanger}/E_{Sanger} = C_{NGS}/E_{NGS}$). Monitoring this 'tipping point' is essential for the efficiency of studies aiming to produce reference libraries, but calculating this specific point of inflection is not always straightforward. While the difference in cost per specimen is easily calculable, the difference in efficiency between Sanger and NGS depends on specimen age, size, preservation method and other factors. Many of these variables are often opaque – while specimen age is usually preserved in the labels, means of preservation prior to mounting is usually unknown for each given specimen. In some cases, indirect evidence can be inferred based on collector name or collection method, as well as specific historic aspects of the material being harvested for DNA. As experience accumulates with particular collections, it may become clear that certain collectors used methods that are compatible with Sanger sequencing (Hebert et al. 2013). For example, in moths, different practices include either killing and mounting individual specimens versus holding specimens in humid 'relaxing boxes' for extended periods before mounting, the latter of which is more prone to deteriorate DNA.

In our case, NGS was only attempted for specimens that were either unlikely to be successfully sequenced with Sanger approaches (i.e. very small or old) or as part of failure tracking; hence, our success rates for NGS cannot be used as baseline for overall success if the whole project was conducted under this approach. Overall, our data and those of Levesque-Beaudin et al. (In press) suggest that NGS success rates are less correlated with specimen age than those of Sanger, meaning that an entirely NGS-based approach may be preferable for studies harvesting largely decades-old material, especially considering the potential evolution of DNA barcoding towards genome skimming (Dodsworth 2015, Coissac et al. 2016, Bohmann et al. 2020). Large-scale studies should consider running pilot projects to investigate differences in efficiency rates among different approaches in order to choose an optimal balance.

Acknowledgements

We wish to thank the numerous USNM curators and other staff members who contributed directly or indirectly with this work: Thomas Henry, Stuart McKamey, Charyn J. Micheli, Allen Norrbom, Robert Robbins, Ted Schultz, Floyd W. Shockley, and Hannah M. Wood. Funds for this project, including a postdoctoral fellowship to BFS, were provided by the Smithsonian’s Global Genome Initiative. The CBG receives funding support from a number of sources, including the Canada Foundation for Innovation, Genome Canada through Ontario Genomics, the Natural Sciences and Engineering Research Council of Canada, the Ontario Ministry of Research, Innovation and Science, the Gordon and Betty Moore Foundation, Ann McCain Evans, and Chris Evans. This article also contributes to the University of Guelph’s Food from Thought research program supported by the Canada First Research Excellence Fund. We would also like to thank colleagues at the CBG for their contributions to this research, including Allison Brown, Gergin Blagoev, Tyler Elliott, Liuqiong Lu, Renee Miskie, Norm Monkhouse, Crystal Sobel, Angela Telfer, Connor Warne, and Paul Hebert. Mention of trade names or commercial products in this publication is solely for the purpose of providing specific information and does not imply recommendation

or endorsement by the USDA. USDA is an equal opportunity provider and employer. The authors have not detected any conflict of interest to declare.

References

- Benson D, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman D, Ostell J, Sayers E (2012) GenBank. Nucleic Acids Research 41 https://doi.org/10.1093/nar/gks1195
- Bohmann K, Mirarab S, Bafna V, Gilbert MTP (2020) Beyond DNA barcoding: The unrealized potential of genome skim data in sample identification. Molecular Ecology 29 (14): 2521-2534. https://doi.org/10.1111/mec.15507
- Chambers EA, Hebert PN (2016) Assessing DNA Barcodes for Species Identification in North American Reptiles and Amphibians in Natural History Collections. PLOS ONE 11 (4). https://doi.org/10.1371/journal.pone.0154363
- Coissac E, Hollingsworth P, Lavergne S, Taberlet P (2016) From barcodes to genomes: extending the concept of DNA barcoding. Molecular Ecology 25 (7): 1423-1428. https://doi.org/10.1111/mec.13549
- D'Ercole J, Prosser SW, Hebert PD (2021) A SMRT approach for targeted amplicon sequencing of museum specimens (Lepidoptera)-patterns of nucleotide misincorporation. PeerJ 9: 10420. https://doi.org/10.7717/peerj.10420
- deWaard J, Ratnasingham S, Zakharov E, Borisenko A, Steinke D, Telfer A, Perez KJ, Sones J, Young M, Levesque-Beaudin V, Sobel C, Abrahamyan A, Bessonov K, Blagoev G, deWaard S, Ho C, Ivanova N, Layton KS, Lu L, Manjunath R, McKeown JA, Milton M, Miskie R, Monkhouse N, Naik S, Nikolova N, Pentinsaari M, Prosser SJ, Radulovici A, Steinke C, Warne C, Hebert PN (2019) A reference library for Canadian invertebrates with 1.5 million barcodes, voucher specimens, and DNA samples. Scientific Data 6 (1). https://doi.org/10.1038/s41597-019-0320-2
- Dodsworth S (2015) Genome skimming for next-generation biodiversity analysis. Trends in Plant Science 20 (9): 525-527. https://doi.org/10.1016/j.tplants.2015.06.012
- Droege G, Barker K, Astrin J, Bartels P, Butler C, Cantrill D, Coddington J, Forest F, Gemeinholzer B, Hobern D, Mackenzie-Dodds J, Ó Tuama É, Petersen G, Sanjur O, Schindel D, Seberg O (2014) The Global Genome Biodiversity Network (GGBN) Data Portal. Nucleic Acids Research 42 https://doi.org/10.1093/nar/gkt928
- Droege G, Barker K, Seberg O, Coddington J, Benson E, Berendsohn WG, Bunk B, Butler C, Cawsey EM, Deck J, Döring M, Flemons P, Gemeinholzer B, Güntsch A, Hollowell T, Kelbert P, Kostadinov I, Kottmann R, Lawlor RT, Lyal C, Mackenzie-Dodds J, Meyer C, Mulcahy D, Nussbeck SY, O'Tuama É, Orrell T, Petersen G, Robertson T, Söhngen C, Whitacre J, Wieczorek J, Yilmaz P, Zetzsche H, Zhang Y, Zhou X (2016) The Global Genome Biodiversity Network (GGBN) Data Standard specification. Database 2016 https://doi.org/10.1093/database/baw125
- Global Genome Initiative (2019) GGI Biodiversity Data Tools - GGI Gap Analysis Tool. https://www.globalgeno.me
- Hawlitschek O, Morinière J, Dunz A, Franzen M, Rödder D, Glaw F, Haszprunar G (2015) Comprehensive DNA barcoding of the herpetofauna of Germany. Molecular Ecology Resources 16 (1): 242-253. https://doi.org/10.1111/1755-0998.12416

Author-formatted, not peer-reviewed document posted on 28/03/2022. DOI: https://doi.org/10.3897/arphapreprints.e84305

- Hebert PN, deWaard J, Zakharov E, Prosser SJ, Sones J, McKeown JA, Mantle B, La Salle J (2013) A DNA 'Barcode Blitz': Rapid Digitization and Sequencing of a Natural History Collection. PLoS ONE 8 (7). https://doi.org/10.1371/journal.pone.0068535
- Hubert N, Hanner R (2015) DNA Barcoding, species delineation and taxonomy: a historical perspective. DNA Barcodes 3 (1). https://doi.org/10.1515/dna-2015-0006
- Ivanova N, deWaard J, Hebert PN (2006) An inexpensive, automation-friendly protocol for recovering high-quality DNA. Molecular Ecology Notes 6 (4): 998-1002. https://doi.org/10.1111/j.1471-8286.2006.01428.x
- Levesque-Beaudin V, Miller ME, Dikow T, Miller SE, Prosser SW, Zakharow EV, McKeown JT, Sones JE, Redmond NE, Coddington JA, Santos BF, Bird J, deWaard JR (In press) A workflow for the expansion of a DNA barcode reference library through 'museum harvesting' of natural history collections. Biodiversity Data Journal.
- 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 (5): 1102-1111. https://doi.org/10.1111/1755-0998.12380
- Morinière J, Hendrich L, Balke M, Beermann A, König T, Hess M, Koch S, Müller R, Leese F, Hebert PN, Hausmann A, Schubart C, Haszprunar G (2017) A DNA barcode library for Germany's mayflies, stoneflies and caddisflies (Ephemeroptera, Plecoptera and Trichoptera). Molecular Ecology Resources 17 (6): 1293-1307. https://doi.org/10.1111/1755-0998.12683
- National Human Genome Research Institute (2019) DNA Sequencing Costs. www.genome.gov/sequencingcostsdata
- Porco D, Chang C, Dupont L, James S, Richard B, Decaëns T (2018) A reference library of DNA barcodes for the earthworms from Upper Normandy: Biodiversity assessment, new records, potential cases of cryptic diversity and ongoing speciation. Applied Soil Ecology 124: 362-371. https://doi.org/10.1016/j.apsoil.2017.11.001
- Prosser SJ, deWaard J, Miller S, Hebert PN (2015) DNAbarcodes from century-old type specimens using next-generation sequencing. Molecular Ecology Resources 16 (2): 487-497. https://doi.org/10.1111/1755-0998.12474
- Puillandre N, Bouchet P, Boisselier-Dubayle M-, Brisset J, Buge B, Castelin M, Chagnoux S, Christophe T, Corbari L, Lamboudière J, Lozouet P, Marani G, Rivasseau A, Silva N, Terryn Y, Tillier S, Utge J, Samadi S (2012) New taxonomy and old collections: integrating DNA barcoding into the collection curation process. Molecular Ecology Resources 12 (3): 396-402. https://doi.org/10.1111/j.1755-0998.2011.03105.x
- Quicke DLJ, Belokobylskij SA, Braet Y, van Achterberg C, Hebert PDN, Prosser SWJ, Austin AD, Fagan-Jeffries EP, Ward DF, Shaw MR, Butcher BA (2020) Phylogenetic reassignment of basal cyclostome braconid parasitoid wasps (Hymenoptera) with description of a new, enigmatic Afrotropical tribe with a highly anomalous 28S D2 secondary structure. Zoological Journal of the Linnean Society 190 (3): 1002-1019. https://doi.org/10.1093/zoolinnean/zlaa037
- Ratnasingham S, Hebert PN (2007) BARCODING: bold: The Barcode of Life Data System (http://www.barcodinglife.org). Molecular Ecology Notes 7 (3): 355-364. https://doi.org/10.1111/j.1471-8286.2007.01678.x
- Raupach M, Hendrich L, Küchler S, Deister F, Morinière J, Gossner M (2014) Building-Up of a DNA Barcode Library for True Bugs (Insecta: Hemiptera: Heteroptera) of Germany Reveals Taxonomic Uncertainties and Surprises. PLoS ONE 9 (9). https://doi.org/10.1371/journal.pone.0106940

- Rduch V, Peters RS (2020) GBOL III: Dark Taxa – die dritte Phase der German Barcode of Life Initiative hat begonnen. Koenigiana 14: 91-107.
- Rinkert A, Misiewicz T, Carter B, Salmaan A, Whittall J (2021) Bird nests as botanical time capsules: DNA barcoding identifies the contents of contemporary and historical nests. PLOS ONE 16 (10). https://doi.org/10.1371/journal.pone.0257624
- Sire L, Gey D, Debruyne R, Noblecourt T, Soldati F, Barnouin T, Parmain G, Bouget C, Lopez-Vaamonde C, Rougerie R (2019) The Challenge of DNA Barcoding Saproxylic Beetles in Natural History Collections—Exploring the Potential of Parallel Multiplex Sequencing With Illumina MiSeq. Frontiers in Ecology and Evolution 7 https://doi.org/10.3389/fevo.2019.00495

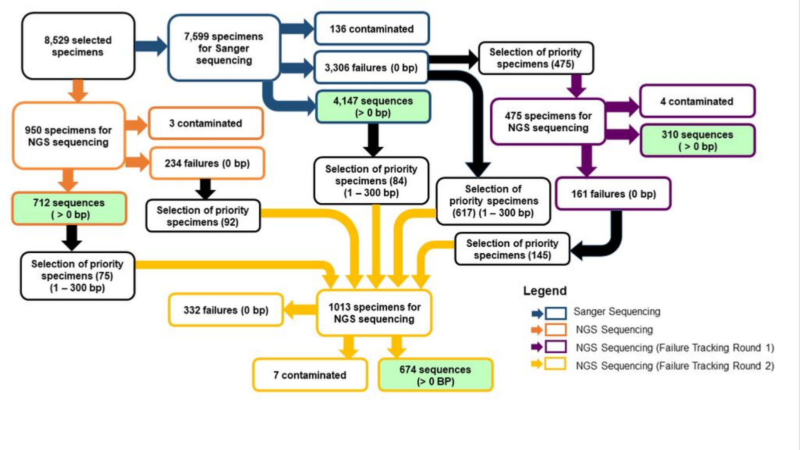
8,529 selected specimens
7,599 specimens for Sanger sequencing
136 contaminated
3,306 failures (0 bp)
4,147 sequences (> 0 bp)
Selection of priority specimens (475)
475 specimens for NGS sequencing
4 contaminated
310 sequences (> 0 bp)
161 failures (0 bp)
950 specimens for NGS sequencing
3 contaminated
234 failures (0 bp)
712 sequences (> 0 bp)
Selection of priority specimens (92)
Selection of priority specimens (84) (1 – 300 bp)
Selection of priority specimens (617) (1 – 300 bp)
Selection of priority specimens (145)
Selection of priority specimens (75) (1 – 300 bp)
332 failures (0 bp)
1013 specimens for NGS sequencing
7 contaminated
674 sequences (> 0 BP)
Legend
Sanger Sequencing
NGS Sequencing
NGS Sequencing (Failure Tracking Round 1)
NGS Sequencing (Failure Tracking Round 2)

Figure 1.

Sanger and NGS Sequencing Flowchart for 8,529 USNM specimens.

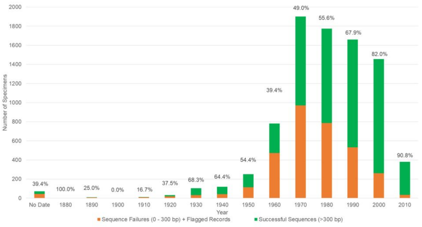

Figure 2.

C0I Sequence length by specimen collection date for the 8,549 USNM specimens selected in 2018 and 2019. The green bar represents the percent of specimens collected per decade with recovered sequences (>300 bp) and orange represents specimens with failed sequences (0 - 299 bp) or flagged sequences.

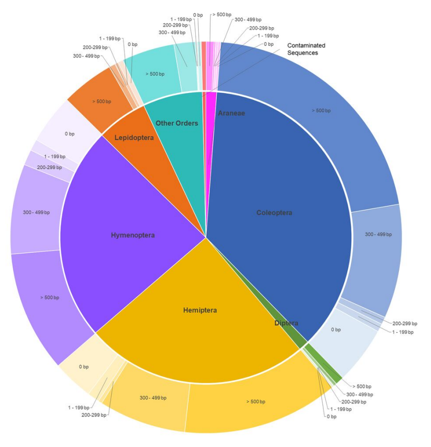

Figure 3.

Sequencing results by taxonomic group for 4,510 USNM genera.

Table 1.

Initial sequencing results by sequencing method for 8,549 USNM specimen records prior to NGS Failure Tracking. 675 genera gained at least 1 sequence using both the Sanger and NGS protocol during initial sequencing.

Initial Sequencing Method	Total Specimens	> 500 bp	300 - 499 bp	200 - 299 bp	0 - 199 bp	0 bp	Contaminated Sequences
Sanger Protocol	7,599	2,246	1,609	239	53	3316	136
NGS Protocol	950	445	120	64	84	234	3
TOTAL	**8,549**	**2,691**	**1,729**	**303**	**137**	**3,550**	**139**
(% of Total)		**31.48%**	**20.22%**	**3.54%**	**1.60%**	**41.53%**	**1.63%**

Table 2.

NGS Failure Tracking sequencing results for 8,549 USNM specimen records. 145 specimens failed (0 bp) on the first round of NGS failure tracking, and were therefore included in the second round.

Sequencing Method	**Total Specimens**	**> 500 bp**	**300 - 499 bp**	**200 - 299 bp**	**0 - 199 bp**	**0 bp**	**Contaminated Sequences**
NGSFT (Round 1)	475	231	69	3	7	161	4
(% of Total)		**48.63%**	**14.53%**	**0.63%**	**1.47%**	**33.89%**	**0.84%**
NGSFT (Round 2)	1,013	356	145	60	113	332	7
(% of Total)		**35.10%**	**14.30%**	**5.90%**	**11.20%**	**32.80%**	**0.70%**

Table 3.

Sequencing results by taxonomic group for 8,549 USNM specimens. **Other Orders**: Neuroptera, Odonata,Mecoptera, Megaloptera, Plecoptera,Trichoptera and Raphidoptera.

Order	**Total Specimens**	**> 500 bp**	**300 - 499 bp**	**200 - 299 bp**	**1 - 199 bp**	**0 bp**	**Contaminated Sequences**
Araneae	**95**	42	12	1	13	26	1
Coleoptera	**3,257**	1284	689	79	41	1095	69
Diptera	**103**	44	17	0	1	37	4
Hemiptera	**2,042**	776	542	30	58	596	40
Hymenoptera	**2,017**	563	493	133	80	736	12
Lepidoptera	**454**	281	46	4	13	104	6
Other Orders *	**581**	288	143	11	2	119	18
Total	**8,549**	**3,278**	**1,942**	**258**	**208**	**2,713**	**150**
(% of Total)		**38.30%**	**22.70%**	**3.00%**	**2.40 %**	**31.70%**	**1.80%**

Table 4.

Sequencing results by taxonomic group for 4510 USNM genera.

Order	**Total Genera**	**> 500 bp**	**300 - 499 bp**	**200 - 299 bp**	**1 - 199 bp**	**0 bp**	**Contaminated Sequences**
Araneae	**54**	29	6	1	8	10	0
Coleoptera	**1,653**	951	424	28	30	214	6
Diptera	**53**	32	12	0	1	7	1
Hemiptera	**1,123**	581	325	14	45	152	6
Hymenoptera	**1,073**	451	335	58	43	185	1
Lepidoptera	**255**	197	23	0	13	20	2
Other Orders	**299**	196	79	4	2	16	2
Total	**4,510**	**2,437**	**1,204**	**105**	**142**	**604**	**18**
(% of Total)		54.04%	26.70%	2.33%	3.15%	13.39%	0.40%

Table 5.

Tissue type and sequencing method by specimen for Hymenoptera prior to NGS Failure tracking.

Sequencing Method	**Total Specimens**	**> 500 bp**	**300 - 499 bp**	**200 - 299 bp**	**1 - 199 bp**	**0 bp**	**Contaminated Records**
Sanger (leg tissue)	1347	260	268	93	31	686	9
NGS (leg tissue)	195	68	24	10	25	68	0
Sanger (whole voucher)	380	57	91	32	0	197	3
NGS (whole voucher)	95	3	29	20	8	35	0
Total	**2,017**	563	493	133	80	736	12

Author-formatted, not peer-reviewed document posted on 28/03/2022. DOI: https://doi.org/10.3897/arphapreprints.e84305

Table 6.

Tissue type and sequencing method by specimen for Hymenoptera after NGS Failure tracking.

Sequencing Method	**Total Specimens**	**> 500 bp**	**300 - 499 bp**	**200 - 299 bp**	**1 - 199 bp**	**0 bp**	**Contaminated Records**
Sanger (leg tissue)	1347	415	326	77	47	473	9
NGS (leg tissue)	195	72	27	10	25	61	0
Sanger (whole voucher)	380	71	106	32	1	167	3
NGS (whole voucher)	95	5	34	14	7	35	0
Total	**2,017**	563	493	133	80	736	12

Supplementary material

Suppl. material 1: Appendix 1.

Authors: Santos B.F. et al.
Data type: Table
Brief description: Specimen selection visits by CBG staff to the Smithsonian Institution National Museum of Natural History, Department of Entomology (NMNH) and corresponding BOLD project on theBarcode of Life Data Systems (BOLD) (Ratnasingham & Hebert, 2007)
Download file (3.58 MB)