GGI-CBG 'Barcoding NMNH Genera' Project – Year 2 Final Report

May 2020

PI/Co-PIs and Project Team: Bernardo Santos, Niamh Redmond, Jonathan Coddington, Nicolas Silverson, Ashton Smith, Jessica Bird, Scott Miller, Meredith Miller, Margarita Miklasevskaja, Allison Brown, Jaclyn McKeown, and Jeremy deWaard

Associated NMNH Curators: Tom Henry, Robert Kula, Michael Gates, Ted Schultz, Sean Brady, Scott Miller, Charyn Micheli, Matt Buffington, Bob Robbins, and Hannah Wood

Methods:

1) Specimen Selection and Loan Organization

In 2019, staff from the Centre for Biodiversity Genomics (CBG) completed three visits to the Smithsonian Institution National Museum of Natural History, Department of Entomology (NMNH): Phase 1 (May 28th to June 6th), Phase 2 (September 10th to 19th) and Phase 3 (December 3rd to 13th). Forty-five arrays of 95 specimens each (4275 specimens total) were selected and loaned to CBG for processing. Thirty-five of these arrays were assembled for analysis by CBG's Sanger-based sequencing protocol, and ten for a protocol involving Next-Generation Sequencing (NGS). Four of the Sanger plates and one NGS plate were processed as whole vouchers. Two representatives from each genus (wherever possible) that were new to both GGBN and GenBank were selected, following the GGI-CBG project guidelines. Taxonomy, country of collection, sample ID, and specimen cabinet/drawer locations were carefully recorded by CBG staff at the time of loan organization. A report of the species names, sample IDs, and country of collection of all specimens was given to the GGI project manager. The loan was approved by museum curators prior to the transport of the specimens to CBG.

2) Imaging, Digitization, Subsampling and Sequencing

Once transferred to CBG, specimens were accessioned and labelled with Barcode of Life Datasystems (BOLD) and USNM ENT labels prior to digitization. All necessary precautions were taken to prevent cross-contamination of and/or damage to the specimens during imaging and subsampling. Digitization, imaging, and tissue sampling for all 45 arrays were completed following pre-determined specifications by museum curators and uploaded to BOLD. DNA samples were extracted using the silica-based protocol outlined in Ivanova, deWaard & Hebert (2006; DOI: 10.1111/j.1471-8286.2006.01428.x). DNA samples were PCR amplified and sequenced following protocols detailed in Hebert et al. (2013; DOI: 10.1371/journal.pone.0068535) and Prosser et al. (2016; DOI: 10.1111/1755-0998.12474) that target overlapping fragments of the cytochrome c oxidase I (COI) gene. The BOLD projects "Hemiptera, Coleoptera, Hymenoptera - USNM June 2019" (Phase 1 - SICOD), "Coleoptera, Hemiptera, Hymenoptera - USNM Sept 2019" (Phase 2 - SICOE) and "Coleoptera, Hemiptera, Hymenoptera, Araneae - USNM Dec 2019" (Phase 3 - SICOF) were created to store all specimen data, images, sequence data and associated files. Jonathan Coddington, Niamh Redmond, Michael Trizna, Bernardo Santos, Scott Miller and Ashton Smith were added to the BOLD projects as full-access users.

3) Data and Other Resources

Specimen data and images for SICOD and SICOE were provided to the GGI project manager in Dec 2019, data and images for SICOF will be provided in September 2020, or will be provided remotely. DNA bank data (following the GGBN Data Standard; Droege et al. 2016; DOI: 10.1093/database/baw125) were provided to the NMNH data manager for input into NMNH's EMu collection management system. Authorship of the specimen records will be completed by GGI staff. DNA extracts were split (20 ul each) between the DNA archives of CBG and NMNH. Extracts for the NMNH Biorepository will be given to the GGI project manager in September 2020. All voucher specimens from Phase 1 (SICOD) and Phase 2 (SICOE) were returned to their original locations within the collection; all voucher specimens from Phase 3 (SICOF) will be returned in September 2020. All successfully sequenced records from the BOLD projects (> 200 bp) will be submitted to GenBank (see Appendix 1) and will be moved to BioProject PRJNA81359 and made public by the GGI Project Manager. USNM voucher information will be listed in the "specimen voucher" field of all GenBank records, ensuring the correct linkage with records in the NMNH collections database (EMu).

Results:

In total, 4275 specimens (45 arrays) were borrowed from NMNH between June and December 2019. Appendix 1 shows a complete list of specimens and their associated data, downloaded from BOLD in May 2020 (Ratnasingham & Hebert, 2007). This represents 5 orders (Figure 1), 96 families, 2358 genera and 2342 identified species collected from 110 countries (Figure 2). As of May 2020, 2345 of the 2358 selected genera were new to GGBN, 2169 were new to GenBank and 1508 were new to BOLD. This constitutes 1033 Barcode Index Numbers (BINs) with 66.2% (684 BINs) being unique to this project. Specimen collection dates (by decade) are summarized in Table 1 and Figure 3. Overall sequencing success was 64.2% (2746/4275) and is summarized in Tables 2 and 3 and Figures 4 and 5. Sequencing success by taxonomic group is summarized in Tables 4 to 13 and Figures 6 and 7. After validation, the 2746 successfully sequenced records were added to the private dataset DS-NMNH2020 titled 'Barcoding NMNH Insect Genera 2019-20' (which will receive the following DOI when made public: http://dx.doi.org/10.5883/DS-NMNH2020. In total, 4275 label images and 4387 specimen images were completed by CBG imaging technicians. Specimen images were largely habitus and/or dorsal view. All Specimen and label images (in TIF format and labelled with USNM ENT numbers) will be provided to the GGI project manager by September 2020. Specimen images are also viewable on BOLD and in Appendix 2.

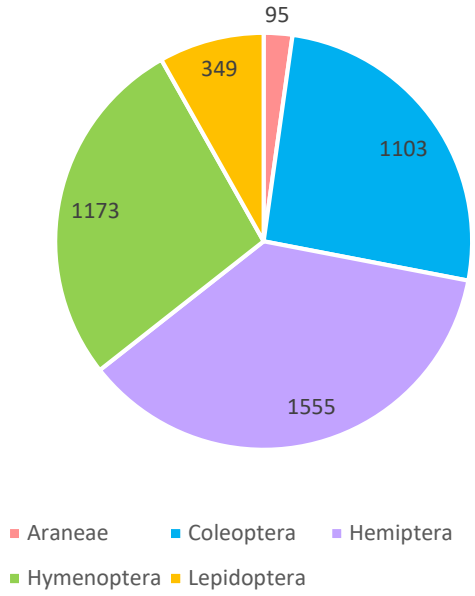

Figure 1: Taxonomic breakdown of specimens sampled

In Years 1 (2018) and 2 (2019) combined, 8549 specimens (90 arrays) were borrowed from the NMNH. This represents 13 orders, 212 families, 4521 genera and 4865 identified species collected from 143 countries. In total, 4415 of the 4521 selected genera were new to GGBN, 4117 were new to GenBank and 2696 were new to BOLD. This constitutes 2101 BINs with 61.3% (1287 BINs) unique to the 'Barcoding NMNH Genera' Project, and 5256 successfully sequenced records (2510 from Year 1 and 2746 from Year 2). In Year 1, NGS-based failure-tracking of 475 specimens resulted in 370 recovered sequences (77.9%). Of the 370 specimens that gained a sequence, 291 (61.3%) were 300 bp or greater and 131 (27.6%) were 500 bp or greater. Appendix 3 provides details on the protocol(s) used and the sequencing success for all records from Years 1 and 2. In addition, 8549 label images and 12,096 specimen images were completed by CBG imaging technicians.

Figure 2: Map of collection locations for specimens borrowed from NMNH (BOLD, 2020)

Table 1: Specimen Collection Dates

Collection Date Decade	Specimen Count	Percent of Total Samples
1880-1890	1	0.06%
1890-1899	5	0.29%
1900-1909	2	0.12%
1910-1919	12	0.70%
1920-1929	30	1.75%
1930-1939	103	6.02%
1940-1949	110	6.43%
1950-1959	197	11.52%
1960-1969	235	13.74%
1970-1979	653	38.19%
1980-1989	819	47.89%
1990-1999	928	54.27%
2000-2009	887	51.87%
2010-2018	238	13.92%
No Year	55	3.22%
Total	**4275**	

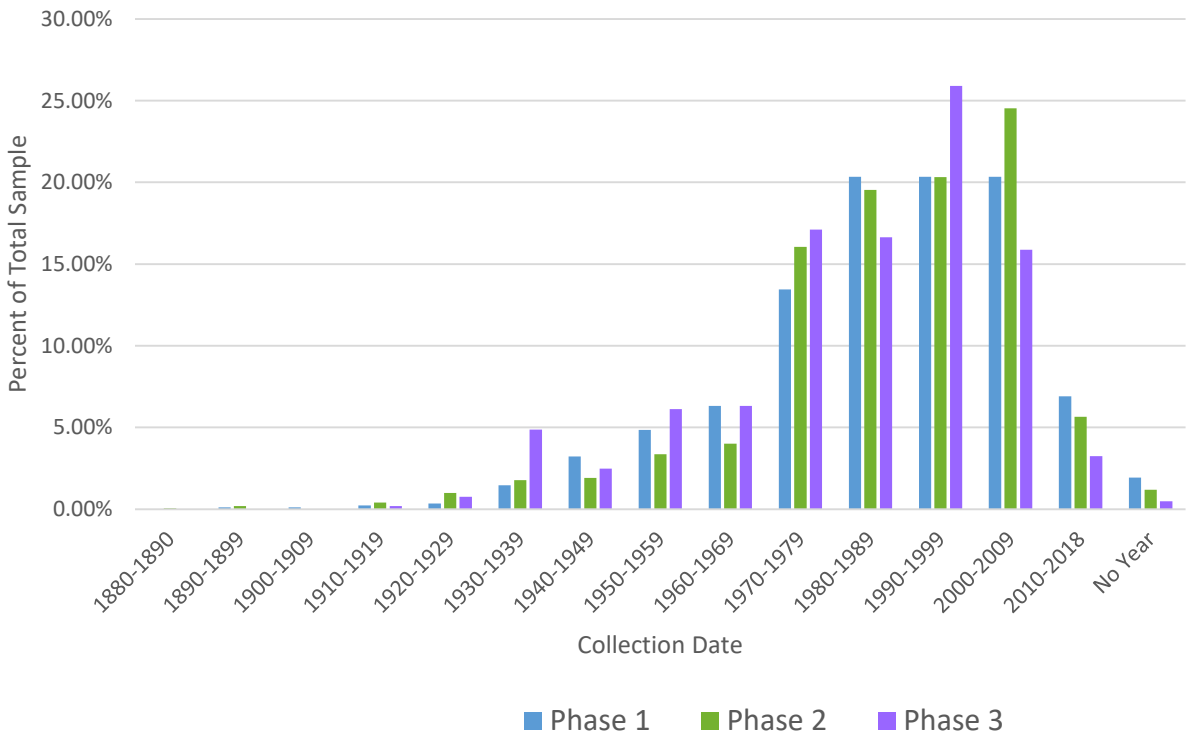

Figure 3: Collection Dates for Year 2 by Decade

Table 2: Overall Success

OVERALL SUCCESS – All Specimens		
Total Records	**4275**	
> 500 bp	1608	37.6%
300-499 bp	892	20.9%
100-299 bp	246	5.8%
0 bp	1529	35.8%

Table 3: Overall Genera Success

OVERALL SUCCESS – Genera		
Total Genera	**2358**	
> 500 bp	1115	47.3%
300-499 bp	505	21.4%
100-299 bp	145	6.1%
0 bp	593	25.1%

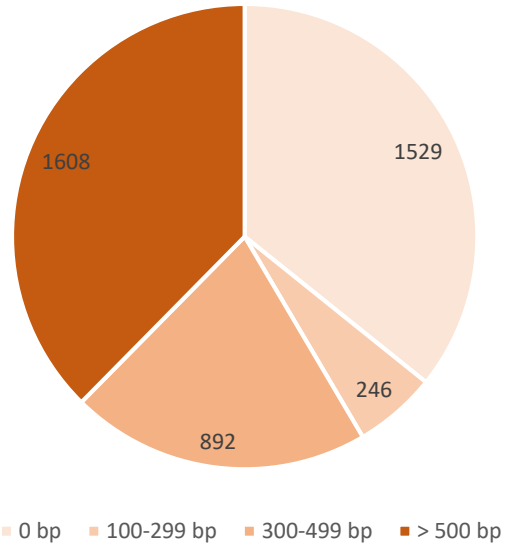

Figure 4: Overall Success

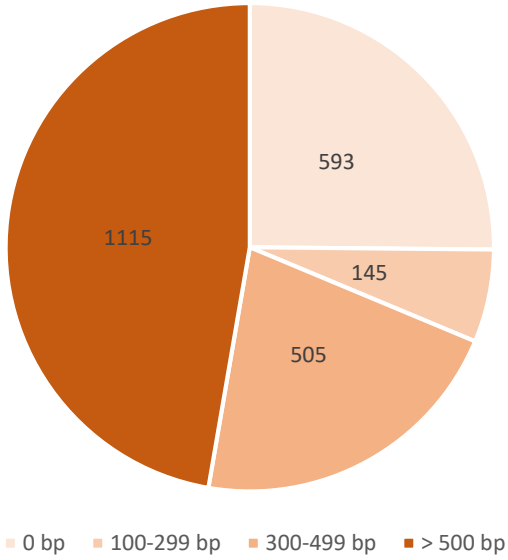

Figure 5: Overall Genera Success

RESULTS: Success by Taxonomic Group

Table 4: Araneae – Overall Success

OVERALL SUCCESS – All Specimens		
Total Records	**95**	
> 500 bp	42	44.2%
300-499 bp	13	13.7%
100-299 bp	12	12.6%
0 bp	28	29.5%

Table 5: Araneae – Genera Success

OVERALL SUCCESS – Genera		
Total Genera	**54**	
> 500 bp	29	53.7%
300-499 bp	6	11.1%
100-299 bp	8	14.8%
0 bp	11	20.4%

Table 6: Coleoptera – Overall Success

OVERALL SUCCESS – All Specimens		
Total Records	**1103**	
> 500 bp	565	51.2%
300-499 bp	200	18.1%
100-299 bp	40	3.6%
0 bp	298	27.0%

Table 7: Coleoptera – Genera Success

OVERALL SUCCESS – Genera		
Total Genera	**609**	
> 500 bp	383	62.9%
300-499 bp	112	18.4%
100-299 bp	19	3.1%
0 bp	95	15.6%

Table 8: Hemiptera – Overall Success

OVERALL SUCCESS – All Specimens		
Total Records	**1555**	
> 500 bp	552	35.5%
300-499 bp	355	22.8%
100-299 bp	66	4.2%
0 bp	582	37.4%

Table 9: Hemiptera – Genera Success

OVERALL SUCCESS – Genera		
Total Genera	**864**	
> 500 bp	397	45.9%
300-499 bp	196	22.7%
100-299 bp	42	4.9%
0 bp	229	26.5%

Table 10: Hymenoptera – Overall Success

OVERALL SUCCESS – All Specimens		
Total Records	**1173**	
> 500 bp	258	22.0%
300-499 bp	293	25.0%
100-299 bp	125	10.7%
0 bp	497	42.4%

Table 11: Hymenoptera – Genera Success

OVERALL SUCCESS – Genera		
Total Genera	**632**	
> 500 bp	172	27.2%
300-499 bp	174	27.5%
100-299 bp	74	11.7%
0 bp	212	33.5%

Table 12: Lepidoptera – Overall Success

OVERALL SUCCESS – All Specimens		
Total Records	**349**	
> 500 bp	191	54.7%
300-499 bp	31	8.9%
100-299 bp	3	0.9%
0 bp	124	35.5%

Table 13: Lepidoptera – Genera Success

OVERALL SUCCESS – Genera		
Total Genera	**199**	
> 500 bp	134	67.3%
300-499 bp	17	8.5%
100-299 bp	2	1.0%
0 bp	46	23.1%

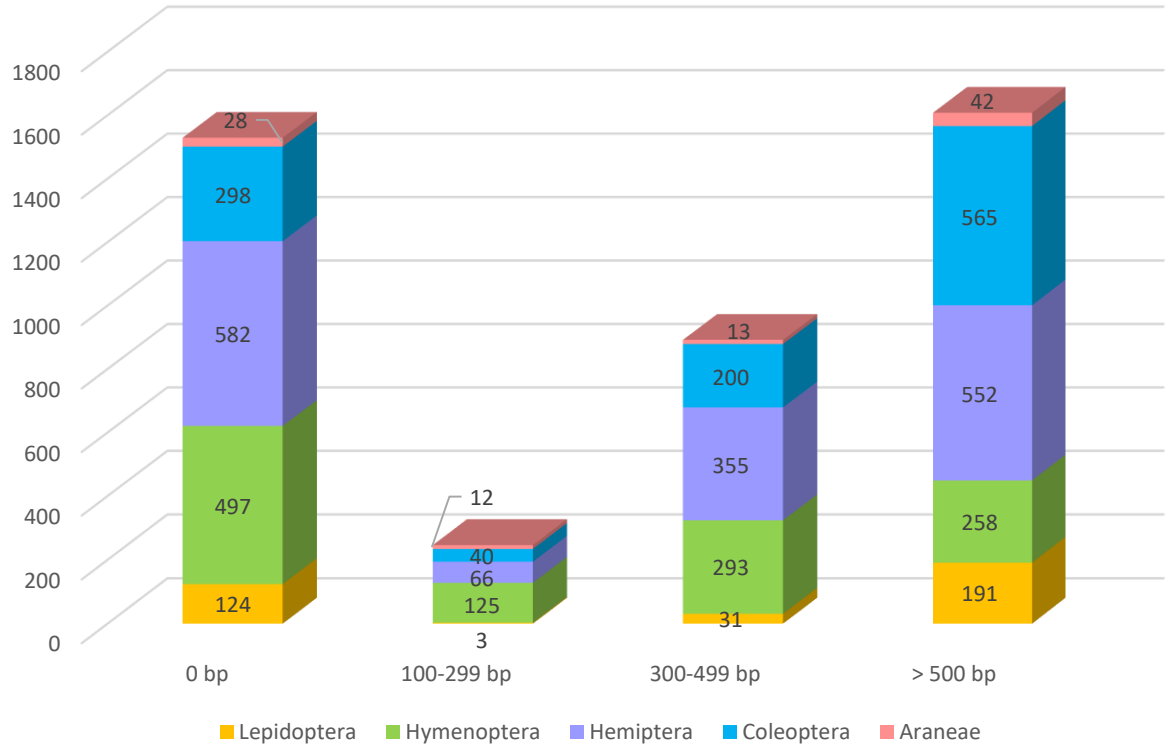

Figure 6: Overall Success by Taxonomic Group

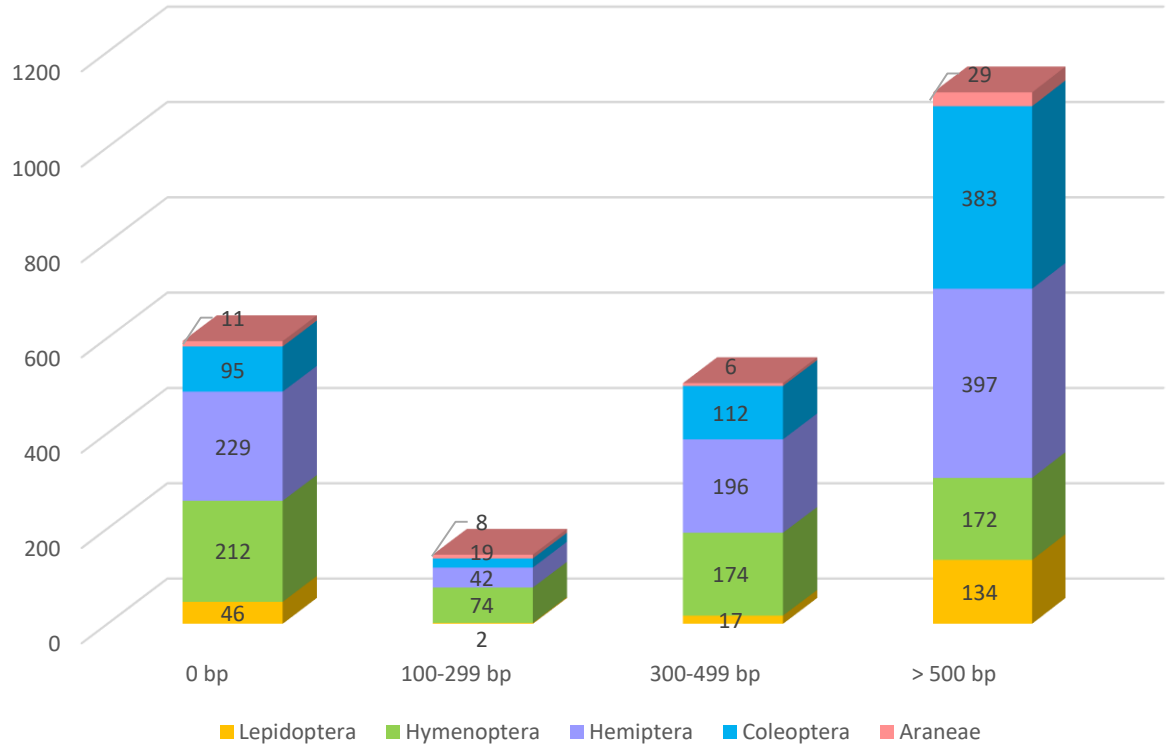

Figure 7: Genera Success by Taxonomic Group

Attached Appendices

Appendix 1: Sequencing Summary and BOLD data spreadsheet for the 4275 NMNH specimens analyzed in Year 2.

Appendix 2: Image library for the 4275 NMNH specimens analyzed in Year 2.

Appendix 3: Year 1 and Year 2 sequencing success.