PLOS | ONE

RESEARCH ARTICLE

The Hemiptera (Insecta) of Canada: Constructing a Reference Library of DNA Barcodes

Rodger A. Gwiazdowski[1]*, Robert G. Foottit[2], H. Eric L. Maw[2], Paul D. N. Hebert[1]

1 Biodiversity Institute of Ontario, University of Guelph, Guelph, Ontario, N1G 2W1, Canada, **2** Agriculture and Agri-Food Canada, Invertebrate Biodiversity—National Environmental Health Program, and Canadian National Collection of Insects, Arachnids and Nematodes, Ottawa, Ontario, K1A 0C6, Canada

* rodger.gwiazdowski@gmail.com

Citation: Gwiazdowski RA, Foottit RG, Maw HEL, Hebert PDN (2015) The Hemiptera (Insecta) of Canada: Constructing a Reference Library of DNA Barcodes. PLoS ONE 10(4): e0125635. doi:10.1371/journal.pone.0125635

Academic Editor: Bernd Schierwater, University of Veterinary Medicine Hanover, GERMANY

Received: November 14, 2014

Accepted: March 18, 2015

Published: April 29, 2015

Data Availability Statement: All relevant data are within the paper and its Supporting Information files. Additionally, all specimen records, and their sequences used in this study are available from the Barcode of Life Datasystems, under the following project names, and DOIs: dx.doi.org/10.5883/DS-HECALIB Hemiptera of Canada - DNA Barcode Library dx.doi.org/10.5883/DS-HECAMAIN Hemiptera of Canada - Main dataset dx.doi.org/10.5883/DS-HECAMN1 Hemiptera of Canada - Main dataset, part II dx.doi.org/10.5883/DS-HECANEW Hemiptera of Canada - New records for release dx.doi.org/10.5883/DS-HECANEW1 Hemiptera of Canada - New

Abstract

DNA barcode reference libraries linked to voucher specimens create new opportunities for high-throughput identification and taxonomic re-evaluations. This study provides a DNA barcode library for about 45% of the recognized species of Canadian Hemiptera, and the publically available R workflow used for its generation. The current library is based on the analysis of 20,851 specimens including 1849 species belonging to 628 genera and 64 families. These individuals were assigned to 1867 Barcode Index Numbers (BINs), sequence clusters that often coincide with species recognized through prior taxonomy. Museum collections were a key source for identified specimens, but we also employed high-throughput collection methods that generated large numbers of unidentified specimens. Many of these specimens represented novel BINs that were subsequently identified by taxonomists, adding barcode coverage for additional species. Our analyses based on both approaches includes 94 species not listed in the most recent Canadian checklist, representing a potential 3% increase in the fauna. We discuss the development of our workflow in the context of prior DNA barcode library construction projects, emphasizing the importance of delineating a set of reference specimens to aid investigations in cases of nomenclatural and DNA barcode discordance. The identification for each specimen in the reference set can be annotated on the Barcode of Life Data System (BOLD), allowing experts to highlight questionable identifications; annotations can be added by any registered user of BOLD, and instructions for this are provided.

Introduction

In this study, we present a DNA barcode library as a set of publicly available COI-5' sequences linked to voucher specimens on the Barcode of Life Data System (BOLD) [1] that meet DNA barcode data standards [1], are identified to species listed in a taxonomic catalogue, and specified using a Digital Object Identifier (DOI, http://www.doi.org/). DNA barcode libraries make it possible sequences with their source specimens, which have been collected across time,

records for release, part II dx.doi.org/10.5883/DS-HECATAX Hemiptera of Canada - Taxonomist ID'd specimens dx.doi.org/10.5883/DS-HECATEN Hemiptera of Canada - Tentative New Specimen/ Species Records The R code version used in this publication can be found at: dx.doi.org/10.5281/zenodo.12582 Hemiptera-of-Canada-DNA-Barcode-Library-workflow An active version of the R code is available at: https://github.com/RodgerG/Hemiptera-of-Canada-DNA-Barcode-Library-workflow.git.

Funding: This research was enabled by funding from NSERC, Genome Canada through the Ontario Genomics Institute, and by Agriculture and Agri-Food Canada. The funders had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

habitats, identifiers, and institutions. Queries that use a reference set to make identifications rely on the quality of the reference material [2]. We propose that the definition of a set of DNA-barcoded specimens as a library, improves it as a basis for identifications in two main ways: #1) the specimens are explicitly defined for community review [3], allowing the library to be collaboratively improved [4]—this may often be the first time many specimens, identified to species, have been explicitly compared with each other [5]; and #2) query results can be rapidly compared with the recognized diversity of a taxon or region, based on prior taxonomic work.

DNA barcode libraries as taxonomic tools

Because of its role as a repository for DNA barcode sequences and associated specimen data, BOLD, the Barcode of Life Data System [1] can serve as a workbench for constructing a reference library. Its effectiveness in supporting identifications [6,7] has been extended by introduction of the Barcode Index Number system (BIN) [8]. The BIN system employs a defined set of algorithms to group DNA barcode sequences into Operational Taxonomic Units (OTU)[9], which often correspond to species [8], and assigns each OTU with a unique identifier (a BIN number, linked to a DOI) [8]. BOLD [1] automatically generates a web page for each BIN that provides summary statistics on its member specimens, and enables the comparison and download of specimen records.

Aggregating specimens into presumptive species-level 'bins' has been practiced by entomologists for over a century. For example, C. V. Riley and H. H. Knight, leading hemipterists in the 19th and early 20th centuries, were among the first to advocate the storage of specimens in unit trays [10,11], a practice now the standard for separating insect specimens assigned to different species. Most unit trays in any collection contain specimens whose taxonomic provenance varies from certain (holotype) to ambiguous—all placed in a particular unit tray by curators or visiting specialists over time; the voucher specimens assembled on each BIN page are similarly accessed as a group. However, the use of a standardized algorithmic approach allows the merger of both identified and unidentified specimens in a way that makes it possible to visually compare DNA barcode diversity with nomenclatural diversity. The BIN page provides a clear report on the level of congruence (or discordance) among the DNA barcode records comprising a BIN and nomenclature [8] exposing taxonomic conflicts with the goal of aiding their resolution [6].

Assembling DNA barcode libraries

The development of a DNA barcode library, for any group of organisms, requires the integration of taxonomic expertise, technologies, and cooperation among institutions [12,13], and helps to create new collaborative opportunities for constructing libraries at national and regional scales [14,15,16]. Prior efforts to develop DNA barcode libraries have adopted diverse approaches (S1 Table). Here, we build on these methods to describe the workflow employed in the construction of a DNA barcode library for Canadian species of Hemiptera [17]. As the fifth largest order of insects (after Coleoptera, Hymenoptera, Lepidoptera, Diptera), the Canadian fauna includes at least 3900 species of Hemiptera [16]. These taxa occur in both aquatic and terrestrial environments across Canada, and include some of its most damaging pest species as well as species used for biocontrol. In this manuscript, we provide an overview of the construction of the library, indicate methods for its use, and report on taxon coverage, taxon diversity, and DNA barcode-based genetic diversity within this group.

Materials and Methods

Data Release

We analyzed records for 54,280 specimens, which included 2671 species of Hemiptera from Canada whose specimen and DNA barcode information are available on BOLD. All records analyzed in this study are included in two projects on BOLD: '*Hemiptera of Canada—Main dataset parts I and II*' that can be accessed via two DOIs: 1) dx.doi.org/10.5883/DS-HECAMAIN; 2) dx.doi.org/10.5883/DS-HECAMN1. A subset of the above dataset, that is only new records for public release, are included in the projects: '*Hemiptera of Canada—New records for release parts I and II*' that can be accessed via two DOIs: 1) dx.doi.org/10.5883/DS-HECANEW; 2) dx.doi.org/10.5883/DS-HECANEW1. We consider a subset of records from the complete dataset (described in **Data Analyses**, below) as the current draft DNA barcode library for the Hemiptera of Canada. This subset includes 20,851 records that provide coverage for 1849 species assigned to 1867 BINs, which are available on BOLD in the project '*Hemiptera of Canada—DNA barcode library*' via DOI dx.doi.org/10.5883/DS-HECALIB. All library specimen records, and their sequences are consistent with the DNA Barcode data standard [1] in terms of both sequence length (>500bp) and quality (less than 1% ns) as well as required specimen metadata. Many of the specimens in this data release are consistent, rather than strictly compliant, with the barcode data standard because they are based on high quality unidirectional reads (see **Specimen Collection and Processing**) rather than the conventional requirement for bidirectional reads.

Most of specimens in this study were processed using high-throughput protocols at the Canadian Center for DNA Barcoding (CCDB), available at: http://www.ccdb.ca/resources.php. These techniques are an integrated workflow of specimen preparation, data recording, photography, tissue sampling, sequencing, and data integration managed using a laboratory information system (LIMS)—which results in individual specimen pages, connected through BINs on BOLD [1,8,18,19,20,21]. The full laboratory history for each specimen can be accessed via its sequence page link on BOLD, with the exception of a few specimens that were processed in other facilities. Fig 1 details the primers used to generate sequences included in the data release, whereas S2 Table lists the sources/collectors of all data.

Choice of a Taxonomic Checklist/Catalogue

The "Checklist of the Hemiptera of Canada and Alaska" by Maw et al. [17] was used as the sole nomenclatural basis for this study, although several more recent studies have led to some shifts in generic and species revision [22,23,24] (also, G. G. E. Scudder (University of British Columbia, CA), C. H. Dietrich (University of Illinois Natural History Survey, USA), C. Bartlett (University of Delaware, USA), J. N. Zahniser (University of Illinois Natural History Survey, USA), personal communication). Generally, all but the most recent catalogues require revision, and the adoption of a particular list/catalogue (or set) during the initial library construction provides a nomenclatural baseline against which future updates can be compared. Here, all new taxonomic identifications for this study were made consistent with the names of Maw et al. [17].

Specimen Collection and Processing

Specimens in this study originate from museum collections, from recent contemporary mass-collecting efforts, and from colleagues contributing public records on BOLD. Museum specimens were borrowed under a formal Memorandum of Understanding with the source institution (see institutions listed in S3 Table). When borrowing museum specimens for DNA

Amplicon (bp)		Primer pairs	Sequence (5' -> 3')	Reference
658	1	C_LepFolF C_LepFolR	LepF1 + LCO1490 (sequences below) LepR1 + HCO2198 (sequences below)	BOLD Primer Database 2014
	2	C_tRWFt1 LepR1	tRWF1_t1 + tRWF2_t1 (sequences below) TAAACTTCTGGATGTCCAAAAAATCA	Park et al. 2010 Hebert et al. 2004
	3	LCO1490 HCO2198	GGTCAACAAATCATAAAGATATTGG TAAACTTCAGGGTGACCAAAAAATCA	Folmer et al. 1994
	4	LCO1490_t1 HCO2198_t1	TGTAAAACGACGGCCAGTGGTCAACAAATCATAAAGATATTGG CAGGAAACAGCTATGACTAAACTTCAGGGTGACCAAAAAATCA	Foottit et al. 2009
	5	LepF1 LepR1	ATTCAACCAATCATAAAGATATTGG as above	Hebert et al. 2004
	6	LepF2_t1 LepR1	TGTAAAACGACGGCCAGTAATCATAARGATATYGG as above	Park et al. 2011 Hebert et al. 2004
	7	PcoF LepR1	CCTTCAACTAATCATAAAAATATYAG as above	Park et al. 2010 Hebert et al. 2004
407	8	LCO1490 MLepR1	as above CCTGTTCCAGCTCCATTTTC	Folmer et al. 1994 published as "MH-MR1" Hajibabaei et al. 2006
	9	MHemF C_LepFolR	GCATTYCCACGAATAAATAAYATAAG as above	Park et al. 2011 BOLD Primer Database 2014
	10	MHemF LepR1	as above as above	Park et al. 2011 Hebert et al. 2004
	11	MLepF1 C_LepFolR	GCTTTCCCACGAATAAATAATA as above	Hajibabaei et al. 2006 BOLD Primer Database 2014
	12	MLepF1 HCO2198_t1	as above as above	Hajibabaei et al. 2006 Foottit et al. 2009
	13	MLepF1 LepR1	as above as above	Hajibabaei et al. 2006 Hebert et al. 2004
	14	MLepF1 LepR1_t1	as above CAGGAAACAGCTATGACTAAACTTCTGGATGTCCAAAAAATCA	Hajibabaei et al. 2006 Zhang & Hanner 2012
307	15	C_LepFolF MHemR	as above GGTGGATAAACTGTTCAWCC	BOLD Primer Database 2014 Park et al. 2011
	16	C_LepFolF MLepR2	as above GTTCAWCCWGTWCCWGCYCCATTTTC	BOLD Primer Database 2014 Hebert et al. 2013
	17	C_tRWFt1 MHemR	as above as above	Park et al. 2010
	18	LepF1 MLepR1	as above as above	Hebert et al. 2004 published as "MH-MR1" Hajibabaei et al. 2006
	19	LepF1 C_ANTMR1D	as above RonIIdeg_R + AMR1deg_R (sequences below)	Hebert et al. 2004 Smith & Fisher 2009
	20	LepF2_t1 MHemR	as above as above	Park et al. 2011
295	21	MLepF2_t1 microLepR2_t1	TGTAAAACGACGGCCAGTGCWTTCCCMCGWATAAATAATATAAG CAGGAAACAGCTATGACGTAATWGCWCCWGCTARWACWGG	Hebert et al. 2013
189	22	AncientLepF2 MLepR2	ATTRRWRATGATCAARTWTATAAT as above	Hebert et al. 2013
164	23	C_microLepF1_t1 C_TypeR1	microLepF2_t1 + microLepF3_t1 (sequences below) TypeR1:TypeR2:TypeR3 (sequences below)	Hebert et al. 2013
Compoments of Cocktail primers (as above)		AMR1deg_R	CAWCCWGTWCCKRMNCCWKCAT	Smith & Fisher 2009
		RonIIdeg_R	GGRGGRTARAYAGTTCATCCWGTWCC	Smith & Fisher 2009
		microLepF2_t1	TGTAAAACGACGGCCAGTCATGCWTTTATTATAATTTTYTTTATAG	Hebert et al. 2013
		microLepF3_t1	TGTAAAACGACGGCCAGTCATGCWTTTGTAATAATTTTYTTTATAG	Hebert et al. 2013
		tRWF1	TGTAAAACGACGGCCAGTAAACTAATARCCTTCAAAG	Park et al. 2010
		tRWF2	TGTAAAACGACGGCCAGTAAACTAATAATYTTCAAAATTA	Park et al. 2010
		TypeR1	GGAGGRTAAACWGTTCAWCC	Hebert et al. 2013
		TypeR2	GGAGGGTAAACTGTTCAWCC	Hebert et al. 2013
		TypeR3	GGTGGATAAACAGTTCAWCC	Hebert et al. 2013

Fig 1. Primers amplifying the COI barcode region of Hemiptera examined in this study. This table pairs with the primer-use heatmap in Fig 2, where the sequentially shaded numbers in the left columns connect primer sequences and citations with their amplification success, per hemipteran family, in this study. Primer sequences are provided with their first occurrence, and sequences for cocktail primer components are provided below. All cocktail primers are used in a 1:1 ratio.

doi:10.1371/journal.pone.0125635.g001

Barcoding [25,26], we used the workflows given in Hebert et al. [21, see for descriptive photos] described in brief here. Tissue sampling from museum specimens involved the removal of a mid, or hind leg; and for some recent collections or very small specimens, such as those from malaise traps, tissue sampling involved whole specimens preserved as vouchers. High-throughput laboratory protocols are based on a 96-well plate format, and specimens are collected into (and returned in) customized Schmitt boxes [27] arrayed into 96 plate positions. As specimens were removed from the collection, a unique CCDB label was swapped into the specimen's location, indicating the specimens' unique Schmitt box and CCDB accession number. Wherever possible, we selected 3–5 specimens of each species that was already represented in BOLD, targeting those with clear locality data, sex determination, their membership in a type series, and from a series where the specimen's labels indicated it was part of a prior study. Small insect specimens, especially those 20 years old or older have lower sequence recovery rates [28] but this can be compensated for by sampling multiple individuals [21]. In this study, the youngest specimens, approximating the above criteria, were preferred.

Because recently collected specimens can yield high-quality sequence results, we supplemented work on museum collections by analyzing hemipterans collected by a range of methods, including Malaise traps, sweep netting, and pitfall taps. These methods include the School Malaise trap program [29] (an array of Malaise traps deployed across southern Ontario) and the BioBus [30], a technician-run collection vehicle specializing in continent-scale specimen collection for DNA barcoding. No specific collection permissions were required for recently collected specimens in this study, as they do not involve endangered or protected species and were previously collected for other projects under blanket collecting permits and permissions issued to the Biodiversity Institute of Ontario. Specimen details, including the holding institution, and original accession number, are provided with each specimen's record on BOLD, and are accessible through the DOIs and project names mentioned above (see **Data Release**).

When initially compared against identified North American Hemiptera on BOLD, these recent collections (primarily collected in 2012–2013) yielded over 1000 BINs not identified to species. We prioritized the identification of 2–3 specimens from a selection of these BINs based on collection location, as well as those that most closely matched a species on the Canadian checklist, when queried on BOLD, using its ID Engine [1]. This subset of specimens was identified by taxonomic specialists, whose working time was greatly reduced by preliminary identification via BOLD [31] [also, C. Bartlett (University of Delaware, USA), M. D. Schwartz (Canadian National Collection of Insects, Arachnids and Nematodes, CA), J. N. Zahniser (University of Illinois Natural History Survey, USA), personal communication]. When specimens of species were unavailable for Canadian localities, we sought representatives from the US or Mexico.

Specimen Identification: how specimens on BOLD get their names

BOLD is a wiki-like environment [32], and all taxonomic names are supplied by data-submitters (i.e., those with edit-level access to particular records). Fields are also available to indicate the identifier and the identification method, and we encourage all data-submitters to clearly indicate their identification method. Records in BOLD obtain their taxonomic information from any of three sources, described below.

1) User submitted names: data-submitters indicate species names, and other taxonomic information during data submission; this includes identifiers indicated on museum specimen labels.

2) Institutional names: Insect specimens, in museum collections, are often assigned to a species' unit tray by curators, or visiting specialists. When these specimens do not have a

determination label, we listed the identifier as [Institutional] Curator with the name, or institutional abbreviation (e.g., Smithsonian Curator, or CNC curator), and the identification method is "Institutional ID".

3) BOLD ID Engine, or BIN-based match names: The BOLD ID engine first aligns a protein translation of the query COI sequence, using a Hidden Markov Model, to a 'query-optimized' alignment-set of specimens on BOLD [1]; the composition of this set corresponds to the search database chosen by the user, during the query. The engine then performs a linear search of this 'query-optimized' set to produce a match. The BIN algorithm (mentioned above) aggregates unidentified and identified specimens together in a BIN. In brief, the BIN algorithm does this by a staged clustering process using uncorrected pairwise distances. The process uses a threshold distance for an initial clustering step, and further refines groupings within, and between these clusters via Markov Clustering [8]. This is a dynamic process that depends on the data available, and for a detailed explanation of this process, please sees pages 2–6 by Ratnasingham & Hebert [8]. These algorithms provide a suggested identification, but the taxonomic names of specimens can only be changed by the data owner through submitting an update to BOLD.

Data Analyses

The main workflow for analyses, tables, and figures was written and performed in R [33] using a BOLD-formatted download (as a csv file) as input (this was all data from BOLD projects: *Hemiptera of Canada—Main dataset, parts I and II*, in DOIs dx.doi.org/10.5883/DS-HECAMAIN, and dx.doi.org/10.5883/DS-HECAMN1). The analyses can be reproduced using the annotated R file, and can be adapted to accommodate other taxa. R code and all supporting R workflow data files are available as S1 Code and described in the legend for S1 Code. R code in-development is deposited at https://github.com/RodgerG/Hemiptera-of-Canada-DNA-Barcode-Library-workflow.git, and the version used for this study (V1.0) is at the DOI: http://dx.doi.org/10.5281/zenodo.12582. For all analyses involving the library dataset, we use the set available as BOLD project: *Hemiptera of Canada—DNA barcode library* at DOI dx.doi.org/10.5883/DS-HECALIB.

Specimens collected in Canada, in BINs without any named specimens, represent possible new species records on BOLD. These records are part of the data release, but not included in the library dataset.

To investigate the utility of library DNA barcodes to differentiate species, we calculated the number of species per BIN, from the library. BINs containing a single species are considered to be concordant; those with more than one species are considered discordant. Concordance will vary between the library and on BOLD because the number of records in need of editing or revision for various taxonomic, or metadata-quality reasons, will be higher on BOLD; and these records tend to influence a species' status toward discordance.

Library species that could be successfully identified were scored as those only occurring in concordant library BINs (BINs containing only one species name). Species sharing barcodes (species that have specimens in discordant library BINs) were determined with the same analysis. This analysis determines species' concordance only within the library dataset. To determine a library species' concordance on BOLD, all library specimens were also analyzed via a BIN discordance report (an online tool) on BOLD. For all library species concordant on BOLD (a conservative estimate), intraspecific divergence was calculated with a pairwise distance analysis using the Kimura 2-parameter model [43] as implemented in BOLD, and the mean standard error, minimum and maximum intraspecific sequence divergence was calculated in R. Potential cryptic species, with mean sequence diversity >2%, were identified with the same analysis.

To provide an overview of the library's contents, we visually explored several aspects of the data. BIN discordance among specimens from the three largest taxonomic families (Aphididae, Cicadellidae, Miridae), identified by taxonomic specialists, was calculated using the same method for barcode sharing (above), and visualized in three dimensions: by family, by number of species in a BIN, and by number of specimens in a BIN. Also, primer usage and proportion of amplification success for all specimens in the data release was plotted as a proportional heatmap. The primer usage data were provided by the BOLD technical staff.

Results

Overall Diversity

The Checklist of Hemiptera of Canada [16] recognizes 78 hemipteran families, 870 genera, and 3944 species as native, or naturalized to Canada. The general data release of specimens matching this checklist or unidentified from Canadian localities, is in the BOLD projects: *Hemiptera of Canada—Main dataset*, *parts I and II*: available at the DOIs dx.doi.org/10.5883/DS-HECAMAIN, and dx.doi.org/10.5883/DS-HECAMN1 and contains 54,280 records, all identified to family (74 families), 39,493 identified to genus (769 genera), and 28,408 identified to species (2,671 species). All records comprise 2,714 BINs, of which 736 are unnamed; specimens from unnamed BINs are presented in S4 Table.

Among samples borrowed from museum collections, we find at least 94 species from Canadian localities not recognized by Maw et al., representing a potential ~ 3% increase of the known fauna. The workflow, after processing through the API of the Catalogue of Life (COL) [34], initially identified >280 species as tentatively new to the checklist. Further nomenclatural review revealed many 'new' species to be: previously published synonyms (101 species), species described or reported since Maw et al. 2000 (49 species), spelling or case agreement errors (29 species), or require further taxonomic clarification (>30 species). Brief status notes for the remaining species tentatively considered new to the checklist are presented in S5 Table, and the specimens are available in the BOLD project *Hemiptera of Canada—Tentative New Specimen/Species Records*, via DOI dx.doi.org/10.5883/DS-HECATEN.

We present a DNA barcode library for the Hemiptera of Canada containing 20,851 specimens classified in 64 families, 628 genera, and 1849 species, assigned to 1867 BINs that can be accessed at the DOI dx.doi.org/10.5883/DS-HECALIB; instructions for specimen-level access are provided as S1 Instructions. A summary table of the taxonomic contents at the ordinal level for the data-release and library are presented in Table 1. A summary of species-based library coverage at the family level is presented in Fig 3. The species-level diversity for library specimens is presented as an annotated version of the Maw et al. checklist in S6 Table. Annotation for each species in S6 Table includes nomenclature status on the Catalogue of Life [34], specimen numbers, divergence metrics, barcode-sharing, and BIN discordance both within the library, and on BOLD. We find 1,312 library species correspond to concordant BINS—within—the library (368 of these species are represented by singleton BINs); 510 Library species appear to share BINs (are discordant within the library), and 27 have not yet been placed in a BIN (as of this writing); species-specific results are presented in S6 Table. Of the 1,312 library species in concordant or singleton BINs, 919 can be successfully identified based on their concordance on BOLD (741 species are in concordant BINs on BOLD, and 178 are singleton BINs; S6 Table). Lastly, we find 27 library species with > 2% mean intra-specific pairwise divergence, which may represent cryptic species complexes, listed in S7 Table.

Table 1. Ordinal-level summary for all data-release and library specimens.

Suborder	Lib. Families	Families	Prop.Named. Family	Lib.Genera	Genera	Prop.Named. Genera
Not.assigned	1	3	0	1	3	0
Archaeorrhyncha	8	8	1	35	47	0.73
Clypeorrhyncha	5	5	1	146	186	0.97
Heteroptera	41	46	1	291	358	0.98
Sternorrhyncha	9	12	1	155	175	0.87
Totals	64	74	*	628	769	*
Suborder	**Lib.Species**	**Species**	**Lib.BINs**	**total.BINs**	**UnNamed.BINs**	
Not.assigned	2	3	5	157	129	
Archaeorrhyncha	87	177	83	119	20	
Clypeorrhyncha	604	875	594	867	227	
Heteroptera	703	1135	724	822	76	
Sternorrhyncha	453	481	461	749	284	
Totals	1849	2671	1867	2714	736	
Suborder	**Lib.>2%div**	**Lib. Specimens**	**Specimens**	**Prop.Named. Specimens**	**Lib.spp.sharing. barcodes**	
Not.assigned	1	8	10117	0	0	
Archaeorrhyncha	0	527	1222	0.63	23	
Clypeorrhyncha	11	7373	15806	0.6	225	
Heteroptera	12	4724	10823	0.85	166	
Sternorrhyncha	2	8219	14587	0.61	96	
Totals	26	20851	52555	*	510	

Shaded columns indicate library specimens, light columns indicate those from the data-release. This table provides a suborder level totals of specimens by taxon-based coverage, proportion of named specimens, BIN totals, species sharing barcodes, and those with >2% divergence. Specimens not assigned to family represent specimens with nomenclature-yet-to-be-confirmed, or specimens without DNA barcodes. The "Not Assigned" category is a workflow-filter that identifies specimens with species names consistent with the checklist, but whose higher taxonomy is not consistent with the checklist.

doi:10.1371/journal.pone.0125635.t001

Visual Data Summaries

Specimens in the data-release come from several sources (Fig 4), including sampling of species from Institutional collections and a range of fresh collection methods. The number of specimens representing species in the library is a function of taxon abundance and representation in collections, and the rate of successful DNA barcode sequence recovery.

Although most of the available sequences are from are recently collected specimens collected within the past few years (S1 Fig), the majority of specimens accessioned with names are older specimens from museums. Less than 30% of all data-release specimens come from museums, but museums account for ~ 50% of the library specimens (S2 Fig). This disparity highlights both the importance of museum specimens to provide a context for specimens from recent projects, and the utility of the latter for extending coverage of geographic range and COI sequence variability within and among taxa.

The number of species per family in the data-release (Table 1) and library (Table 1 and Fig 3) reflects the relative diversity of those taxa in the checklist (S6 Table). This representation is not only frequency-dependent, reflecting fresh captures from recent projects (Fig 4), but also corresponds to contributor interest, as the more species-rich groups contain members of economic or systematic importance (in particular: the Aphididae, Cicadellidae, and Miridae; Fig 2 and Fig 3). However some relatively diverse groups have limited representation, notably scale insects (Coccoidea) and jumping plant lice (Psylloidea). These groups not only require


Acanaloniidae
Acanthosomatidae
Achilidae
Adelgidae
Aleyrodidae
Alydidae
Anthocoridae
Aphididae
Aradidae
Arthropleidae
Belostomatidae
Beryliidae
Blissidae
Calliscelidae
Calophyidae
Ceratocombidae
Cercopidae
Cicadellidae
Cicadidae
Cimicidae
Cixiidae
Clastopteridae
Coccidae
Coreidae
Corixidae
Cydnidae
Cymidae
Delphacidae
Derbidae
Diaspididae
Enicocephalidae
Eriococcidae
Flatidae
Fulgoridae
Gelastocoridae
Geocoridae
Gerridae
Hebridae
Heterogastridae
Hydrometridae
Issidae
Kinnaridae
Lasiochilidae
Lygaeidae
Membracidae
Mesoveliidae
Miridae
Naboidae
Naucoridae
Nepidae
Notonectidae
Ochteridae
Orthezidae
Oxycarenidae
Pachygronthidae
Pentatomidae
Phylloxeridae
Phymatidae
Piesmatidae
Pleidae
Pseudococcidae
Psyllidae
Reduviidae
Rhopalidae
Rhyparochromidae
Saldidae
Scutelleridae
Thyreocoridae
Tingidae
Triozidae
Veliidae
C_LepFolF & C_LepFolR
C_tRWFt1 & LepR1
LCO1490 & HCO2198
LCO1490_t1 & HCO2198_t1
LepF1 & LepR1
LepF2_t1 & LepR1
PcoF & LepR1
LCO1490_t1 & MLepR1
MHemF & C_LepFolR
MHemF & LepR1
MLepF1 & C_LepFolR
MLepF1 & HCO2198_t1
MLepF1 & LepR1
MLepF1 & LepR1_t1
C_LepFolF & MHemR
C_LepFolF & MLepR2
C_tRWFt1 & MHemR
LepF1 & MLepR1
LepF1 & ANTMR1D
LepF2_t1 & MHemR
MLepF2_t1 & microLepR2_t1
AncientLepF2 & MLepR2
C_microLepF1_t1 & C_TypeR1
658
407
307
295
189
164
Proportion of amplification
1.00
0.75
0.50
0.25
0.00

Fig 2. Proportions of PCR amplifications by taxon, generated with particular primer pairs, for all Hemiptera specimens in this study. This heatmap pairs with Fig 1, where the sequentially shaded numbers in the left columns connect primer sequences and citations with their use, per hemipteran family. Primer pairs appear in descending order of amplicon lengths, at left. The dot area is proportional to all amplification attempts for that primer/family, the total number of specimens analyzed per primer combination is adjacent to each dot, and the shading indicates the proportion of successful amplifications (illustrated in the key at lower left). All cocktail primers are used in a 1:1 ratio.

doi:10.1371/journal.pone.0125635.g002

specialized collecting, preservation, and identification techniques, but also museums do not usually maintain specimens of some groups in a state usable for DNA Barcoding (e.g., specimens are slide-mounted or preserved in inappropriate fluids).

Amplification success by primer (Figs 1 & 2) tends to be highest with full-length (658bp) DNA barcode primers (Fig 2) applied to specimens collected within the past 30 years (S1 Fig). Particularly, shorter-fragment "mini" primers (Figs 1 & 2) have mixed success across many groups, and these primers are often attempted with older or museum specimens after full-length DNA barcode primers are unsuccessful. Amplification bias does not appear to be systematic across families, but in some cases may be taxon-dependent (e.g., primer pairs applied within the Aleyrodidae, Corixidae, or Miridae), and Fig 2 may assist targeting amplification by taxon.

BIN discordance among library specimens identified to species by taxonomists is presented in Fig 5. This figure displays a range of congruent and incongruent states that BINs for a taxon, or entire library can possess, based on the diversity of specimen information, and DNA barcode variation aggregated in an individual BIN. A table of these taxonomists and their number of specimens is provided in S8 Table. Of the 676 BINs in Fig 5, the majority (606 BINs) are concordant (contain one species), 45 BINs have two species, and 25 have three or more. An expanded view of BIN discordance for all families in the library is presented in S3 Fig, and the taxonomists involved are listed in S9 Table.

Discussion

We have established a DNA barcode library with reference to a taxonomic list or catalogue for an order of insects within a country, using a DOI to define library specimens on BOLD. Additionally, we have made the R-based workflow publically available, and reproducible through the S1 Code. This library presents a defined set of reference specimens of the Hemiptera of Canada, which is open to community review and comments (to do this, please see S1 Instructions). Using the data from a DOI in an R-based workflow, several new developments are possible. We found the ability to rapidly categorize a burgeoning stream of BOLD data against a taxonomic catalogue, and the existing library allowed us to dynamically identify new library specimens, as well as tentative new species records, and unnamed BINs both of which were subsequently identified by taxonomic review. Automatically querying taxonomic names of new records against the Catalogue of Life using 'taxize' [35] easily revealed nomenclatural differences between 'new' records, and our catalogue with the most comprehensive global index available [34]. These 'new' species records (initially >280 species) were obtained through general museum surveys, similar to the methods of Hebert et al. (2013) [21]—and we hypothesize this result may apply for most taxa. Manual research revealed that many of these 'new' records (S5 Table) actually occur in the checklist, because our auto-processing missed them due to: synonymy, changes in generic combination, misspelling/wrong gender agreement from both BOLD and/or the checklist, and differences in recognition/non-recognition of subspecies. Lastly, because the R-workflow is based on a BOLD-formatted spreadsheet (the standard data download from BOLD), it is adaptable for most BOLD users, and most datasets.

Suborder	Family	Checklist/Library	% in Library
Archaeorrhyncha	Flatidae	1/1	100
	Fulgoridae	8/7	88
	Caliscelidae	11/8	73
	Achilidae	19/10	53
	Delphacidae	125/47	38
	Derbidae	14/5	36
	Issidae	3/1	33
	Cixiidae	35/8	23
Clypeorrhyncha	Cercopidae	24/22	92
	Cicadidae	21/13	62
	Membracidae	105/60	57
	Cicadellidae	1088/506	47
	Clastopteridae	12/3	25
Sternorrhyncha	Adelgidae	18/10	56
	Aphididae	828/428	52
	Eriococcidae	3/1	33
	Phylloxeridae	8/2	25
	Diaspididae	29/4	14
	Aleyrodidae	11/1	9
	Coccidae	24/2	8
	Psyllidae	105/4	4
	Triozidae	25/1	4
	Pseudococcidae	23/0	0
	Margarodidae	5/0	0
	Ortheziidae	5/0	0
	Asterolecaniidae	2/0	0
	Cryptococcidae	2/0	0
	Aclerdidae	1/0	0
	Calophyidae	1/0	0
	Dactylopiidae	1/0	0
	Kermesidae	1/0	0
	Spondyliaspididae	1/0	0
Heteroptera	Blissidae	6/6	100
	Pachygronthidae	3/3	100
	Artheneidae	1/1	100
	Enicocephalidae	1/1	100
	Gelastocoridae	1/1	100
	Hydrometridae	1/1	100
	Ochteridae	1/1	100
	Pleidae	1/1	100
	Rhopalidae	16/15	94
	Coreidae	13/12	92
	Pentatomidae	69/55	80
	Berytidae	5/4	80
	Cymidae	5/4	80
	Alydidae	9/7	78
	Lygaeidae	25/19	76
	Notonectidae	12/9	75
	Acanthosomatidae	4/3	75
	Belostomatidae	4/3	75
	Nepidae	4/3	75
	Nabidae	19/13	68
	Phymatidae	3/2	67
	Thyreocoridae	11/7	64
	Reduviidae	26/16	62
	Cydnidae	10/6	60
	Geocoridae	10/6	60
	Rhyparochromidae	65/37	57
	Anthocoridae	38/21	55
	Scutelleridae	13/7	54
	Tingidae	47/24	51
	Oxycarenidae	4/2	50
	Mesoveliidae	2/1	50
	Miridae	691/341	49
	Saldidae	38/18	47
	Corixidae	77/31	40
	Gerridae	23/9	39
	Lyctocoridae	6/2	33
	Cimicidae	7/2	29
	Piesmatidae	4/1	25
	Hebridae	6/1	17
	Veliidae	7/1	14
	Aradidae	53/6	11
	Microphysidae	4/0	0
	Ceratocombidae	1/0	0
	Heterogastridae	1/0	0
	Lasiochilidae	1/0	0
	Naucoridae	1/0	0

Fig 3. Species-level summary by family, of library coverage given the Checklist. Families are grouped by suborder, and then by the proportion of species with barcode records. The numbers in the Checklist/library column indicate the number of species for each family in the Checklist, and the number with barcode coverage.

doi:10.1371/journal.pone.0125635.g003

The past several years have seen a rise in Hemiptera DNA barcode publications that span intra-specific [36,37] and inter-specific [38] population studies, regional libraries from Europe [39], India [40], Korea and the far east [41], and the start of a global focus at the family (Aradidae) [42], and suborder (heteroptera)[43] levels. Thus far all Hemiptera libraries, including this study, predictably find patterns of taxonomic agreement, cryptic diversity or barcode sharing based on their specimens examined. These preliminary observations are important, but isolated, in light of the comparative contributions each study can provide, if regional, or taxon-based libraries could be collaboratively organized on a global scale.

Prior to developing our workflow we considered 41 citations, spanning the past five years, which explicitly discuss creating a DNA barcode Reference library (presented in S1 Table, which includes search methods, and keywords). Approximately half of these 41 studies proceed with an a priori taxonomic catalogue, whereas the others generate a post-hoc list from the study results. Very few are continental (3 studies) or global (1 study) in scope, and there is a clear taxon bias toward insects (19 studies), toward fish among vertebrates (7/10 studies) and a pioneering representation for community assemblages. While data from most of these studies is publically available, we found none combined a prior taxonomic hypothesis (catalogue) with a publically available dataset designated by a DOI, and developed a scalable workflow relative to their data-sources. Development of an open-source workflow that is reproducible serves as a common-source tool for broad community use [44]. Such community 'conversation' is a frontier in cybertaxonomy [45], and we offer our workflow as a point for discussions toward other library construction efforts. BIN discordance exists in the library at various taxonomic levels (Fig 5 and S3 Fig), and one of the next developmental steps for the Hemiptera of Canada library is specimen-level annotation by taxonomic specialists. Similar to the curation of museum-specimens by external experts, such annotations could accumulate a community consensus of comments on particular specimens (or BINs) (please see S1 Instructions for a step-by-step guide to accessing and annotating library specimens, or BINs).

Currently, queries on BOLD are confined to the OpenIdEngine (http://www.boldsystems.org/index.php/IDS_OpenIdEngine), and direct queries via BOLD of the library we present here, are not yet available. However, library specimens can be added to any private project, and then used with all of the analysis tools available on BOLD. Furthermore, the library contents can be accessed directly (see S1 Instructions), and the ProcessIDs/SampleIDs from library specimens can be compared to search results from the OpenIdEngine, or to specimens in BINs.

The library contents range from species represented by a single specimen to concordant species spanning several BINs with hundreds of specimens per BIN; and although this first draft of the library varies in taxonomic coverage and agreement, several patterns useful for making identifications are becoming clear. Many species represented by singleton BINs (and thus, a single specimen), such as the mirid plant bug *Tropidosteptes cardinalis* (Uhler, 1878)[46] (http://www.boldsystems.org/index.php/Public_BarcodeCluster?clusteruri=BOLD:AAG8871), are important first-references that can be prioritized for further sampling in future library drafts. Some species with concordant BINs have very few specimens identified by a taxonomist, relative to, perhaps, hundreds of specimens collected from across the continent (for example, the cicadellid leafhopper *Macrosteles quadrilineatus* (Forbes, 1885)[47] (http://www.boldsystems.org/index.php/Public_BarcodeCluster?clusteruri=BOLD:AAA9422), and these BINs appear to reflect a clear biological grouping.

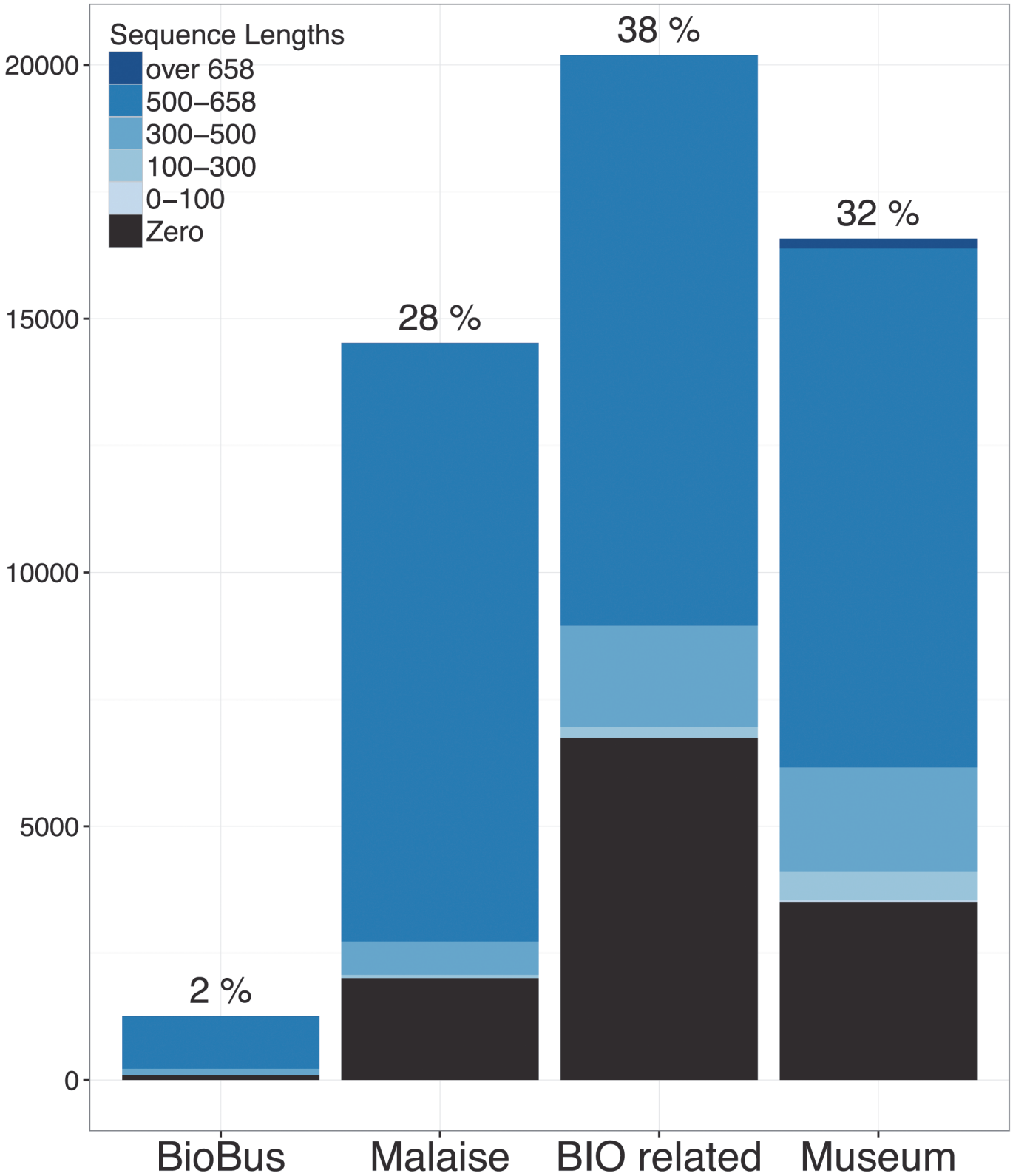

Fig 4. Proportion of data sources, and sequence length for all records in this study. These generalized-collection methods acquire a broad spectrum of taxa, and are valuable sources of diversity when combined with conventional museum collections, and research projects through BIO. Some specimens, especially those from museums, lack sequence data, but do have images, names, and associated metadata on BOLD.

doi:10.1371/journal.pone.0125635.g004

BIN discordances within the library (e.g. Fig 5) are likely to be biological realities, and may also highlight inconsistencies in identifications or application of species concepts by different workers. Some taxonomic groups vary widely in their concordance, for example the plant bug genus *Lygus* (family Miridae). A single BIN corresponds to *Lygus lineolaris* (Palisot, 1818)[48] that contains > 700 specimens collected from across North America, with identifications by multiple taxonomists (http://www.boldsystems.org/index.php/Public_BarcodeCluster?clusteruri=BOLD:AAA5803).

However, another *Lygus* (Hahn, 1833)[49] BIN includes ~ 600 specimens of 27 different taxonomist-identified morphospecies (http://www.boldsystems.org/index.php/Public_BarcodeCluster?clusteruri=BOLD:ACF4388), reflecting an intriguing taxonomic issue (G. G. E. Scudder & M. D. Schwartz, personal communication from identifiers of these same *Lygus* specimens). Lastly, even discordant multi-species BINs represent a useful level of identification when patterns of discordance are consistent, especially in genera containing one or more complexes of closely related species.

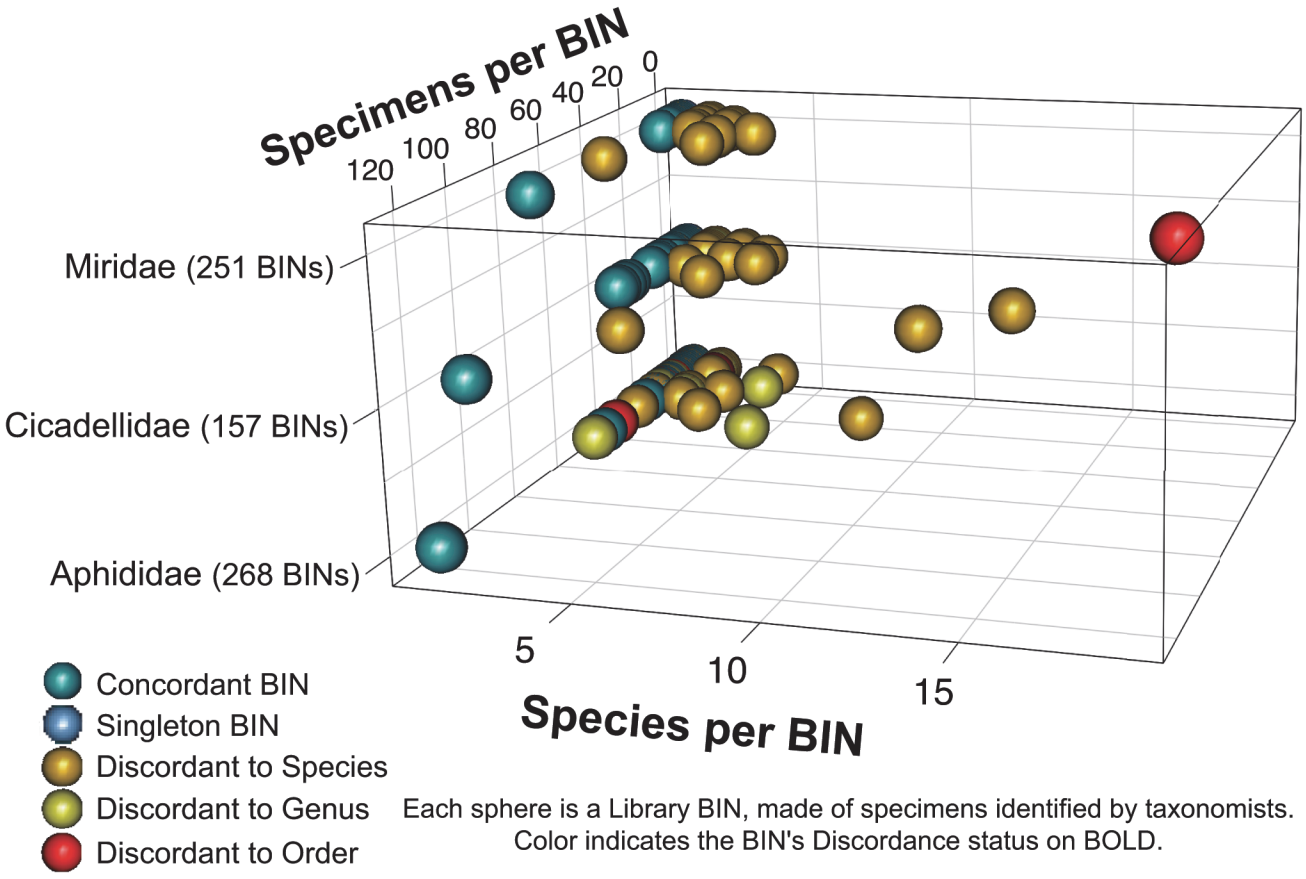

Fig 5. A three dimensional visualization of taxonomic congruence between BINs and species identifications, based on morphology, for three families of Canadian Hemiptera (Aphididae, Cicadellidae, and Miridae). The taxonomists who identified these specimens are listed in S7 Table. A robustly concordant BIN (one species per BIN, with many specimens in that BIN) occupies a forward position in the plane along the left wall, whereas discordant BINs (those with many specimens, and several species names) appear towards the upper right quadrant. The colors correspond to the concordance status of the same BIN on BOLD. In many cases, BINs that are concordant in this curated dataset are discordant on BOLD. This disparity highlights the utility of defining a reference set of specimens, as library specimens will be grouped by BOLD in BINs with misidentified specimens.

doi:10.1371/journal.pone.0125635.g005

Future Directions

Comparison of additional specimens with this first iteration of a DNA barcode library of the Hemiptera of Canada is a means of identifying candidate species that can contribute to the development of a more complete dataset [50,51,52]. As additional data become available on BOLD, and the Hemiptera of Canada library grows, targetable gaps in taxon and geographic coverage become clear—clarifying the contributions needed from taxon specialists, and highlighting species of interest that can be targeted by novel resources, such as DNA barcode Bioblitzes [53], or student projects via the Education Barcode of life's (eBol) student data portal [54].

Lastly, we recognize the multimodal nature of biodiversity inference can be unifying. DNA sequences without names, specimens without sequences, names that are synonymies—can all be reciprocally illuminating in a comparative framework. Conventionally, these frameworks have been important, but isolated works: catalogues, monographs, revisions, descriptions, sequences, and their epistemological foundation—specimens. As demonstrated here, an infrastructure that digitally aggregates specimens in a DNA-mediated reference library can unite these isolated resources into open-sourced 'virtual unit trays' of specimens, and sequences, from across the world's collections. Such an effort integrates isolated resources to create a shared understanding about taxon diversity.

Communicating about how we evolve such an infrastructure is a frontier in biodiversity science. And we share the ideas in this paper to help evolve inference methods that are at once public, repeatable, collaborative, and comparative.

Supporting Information

S1 Fig. Specimen number and sequence length recovery by year for specimens included in the data-release. Years are grouped by decade until 2000–2005 to reflect the recent rise in number of specimens added to BOLD in each year.
(EPS)

S2 Fig. Proportion of data sources, and sequence length for the records included in the library dataset, showing museums as an important source for library specimens.
(EPS)

S3 Fig. A three dimensional visualization of taxonomic congruence between BINs and species identifications based on morphology, for all families of Canadian Hemiptera; this presents a snapshot of the contents and structure of the entire 'library'. Such a view could be used to access macro-level progress during early phases of library construction. The taxonomists who identified these specimens are listed in S9 Table. A robustly concordant BIN (one species per BIN, with many specimens in that BIN) occupies a forward position in the plane along the left wall, whereas discordant BINs (those with many specimens, and several species names) appear towards the upper right quadrant. The colors correspond to the concordance status of the same BIN on BOLD. In many cases, BINs that are concordant in this curated dataset are discordant on BOLD. This disparity highlights the utility of defining a reference set of specimens, as library specimens will be grouped by BOLD in BINs with misidentified specimens.
(EPS)

S1 Instructions. Step-by-step instructions for accessing and commenting on individual records within the DNA barcode library for the Hemiptera of Canada. These same instructions

are applicable for adding comments to a particular BIN page.
(PDF)

S1 Code. Zipped file containing data, accessory files and R code to reproduce all analyses and tables presented in this paper. In R, once the working directory is set to a folder containing these files, all code can be executed and run at once. The R code calls for the installation and use of all necessary packages although mac users may need to update their installation of XQuartz (http://xquartz.macosforge.org), to render the three dimensional images presented in Fig 5, and S3 Fig. For further details, please see the annotation in the preamble, and throughout the R code.
(ZIP)

S1 Table. Library citations and brief characterization of the methods employed by publications describing the construction of DNA barcode libraries. Construction-method categories were arbitrarily designated after considering the diversity of methods used in all publications.
(PDF)

S2 Table. Contributors of specimens for all data released as part of this manuscript. These data can be found on BOLD, in projects: '*Hemiptera of Canada—Main dataset parts I and II*' that can be accessed via two DOIs: 1) dx.doi.org/10.5883/DS-HECAMAIN; 2) dx.doi.org/10.5883/DS-HECAMN1.
(CSV)

S3 Table. Institutions that provided specimens for processing at the Canadian Center for DNA Barcoding.
(CSV)

S4 Table. Specimens in BINs that lack specimen with a species assignment. These specimens are found, using R, by identifying all BINs containing species, and excluding all specimens in these BINs, through a subset of the data. These taxa represent tentative new species records for the DNA barcode library, but they await identification.
(CSV)

S5 Table. Tentative new species records for Canada. These specimens are found, using R, by identifying species that do not match the Maw et al. (2000) checklist, but that were collected within Canada. Duplicate entries have been removed. These specimens derive from thorough sampling of museum collections and from targeted searches by taxon. Many of these species have been recognized as native or naturalized after Maw et al (2000), and each species has been annotated with brief notes regarding its tentative taxonomic status. The full list of specimens is available on BOLD in the project *Hemiptera of Canada—Tentative New Specimen/Species Records*, available at the DOI dx.doi.org/10.5883/DS-HECATEN.
(CSV)

S6 Table. Species-specific information for taxa included in the Maw et al. (2000) checklist, as found in the reference library, and on BOLD. The curated dataset of the library contains fewer specimens, and fewer cases of BIN discordance than the full public data on BOLD, and this information is summarized here. For each species, this table also lists the taxonomic status of each on the Catalogue of Life, the number of specimens on BOLD, the number of specimens in the library, the mean, standard error, maximum and minimum intraspecific pairwise divergence (using the Kimura two-parameter model available on BOLD), whether a species shares barcodes and, if so, the number of species involved, the number of BINs per species, and lastly

the number of specimens in BINs on BOLD, relative to the BIN's concordance status.
(CSV)

S7 Table. Species in the library dataset that display greater than 2 percent uncorrected pairwise divergence. It is possible these species represent either particularly divergent, or cryptic groups.
(CSV)

S8 Table. Taxonomists whose identifications were used to generate the concordance reports in Fig 5.
(CSV)

S9 Table. Taxonomists whose identifications were used to generate concordance reports for S3 Fig.
(CSV)

Acknowledgments

We thank Jeremy Andersen, Andrew Frewin, Thomas Henry, Ian King, and two anonymous reviewers whose comments substantially improved earlier drafts of this manuscript. We also thank Charles Bartlett (University of Delaware, USA), Christopher Dietrich (University of Illinois Natural History Survey, USA), Thomas Henry (National Museum of Natural History, USA) Michael Schwartz (Canadian National Collection of Insects, Arachnids and Nematodes, CA), Geoffrey Scudder (University of British Columbia, CA) and James Zahniser (University of Illinois Natural History Survey, USA) who provided important advice on the identification of certain specimens. We also acknowledge the hemipterists whose taxonomic efforts over the past century provided the foundation for our work.

R.A.G is grateful for the keen assistance and conscientious effort from all staff at the Biodiversity Institute of Ontario—without which this work would not be possible; and to Robert Hanner and Sujeevan Ratnasingham for stimulating discussions about related ideas. R.A.G also thanks Jerri Larsson for sharing his research on the origin of Unit Trays, and also: Thomas Henry, Dennis Kopp, Stuart McKamey, and Scott Miller of the NMNH; Sydney Cannings, and Claudia Copley of the RBCM; along with Karen Needham and Geoffrey Scudder of the Beaty Biodiversity Museum for their hospitality, and help with sampling their collections.

Author Contributions

Conceived and designed the experiments: RAG PDNH RGF. Performed the experiments: RAG HELM. Analyzed the data: RAG HELM RGF. Contributed reagents/materials/analysis tools: PDNH RGF HELM RAG. Wrote the paper: RAG RGF HELM PDNH.

References

1. Ratnasingham S, Hebert PDN. BOLD: The Barcode of Life Data System. Molecular Ecology Notes. 2007; 7: 355–364. Available: http://www.barcodinglife.org. PMID: 18784790
2. Meyer CP, Paulay G. DNA Barcoding: error rates based on comprehensive sampling. PLOS Biology. 2005; 3: e422. PMID: 16336051
3. Ruedas LA, Salazar-Bravo J, Dragoo JW, Yates TL. The importance of being earnest: What, if anything, constitutes a "specimen examined?". Molecular Phylogenetics and Evolution. 2000; 17: 129–132. PMID: 11020311
4. Kvist S. Barcoding in the dark?: A critical view of the sufficiency of zoological DNA barcoding databases and a plea for broader integration of taxonomic knowledge. Molecular Phylogenetics and Evolution. 2013; 69: 39–45. doi: 10.1016/j.ympev.2013.05.012 PMID: 23721749

5. Lim GS, Balke M, Meier R. Determining species boundaries in a world full of rarity: singletons, species delimitation methods. Systematic Biology. 2011; 61: 165–169. doi: 10.1093/sysbio/syr030 PMID: 21482553
6. DeSalle R, Egan MG, Siddall M. The unholy trinity: taxonomy, species delimitation and DNA barcoding. Philosophical Transactions Of The Royal Society B–Biological Sciences. 2005; 360: 1905–1916.
7. Wheeler QD, Valdecasas AG. Cybertaxonomy and ecology. Nature Education Knowledge. 2010; 3: 6.
8. Ratnasingham S, Hebert PDN. A DNA–based registry for all animal species: the Barcode Index Number (BIN) System. PLOS ONE. 2013; 8: e66213. doi: 10.1371/journal.pone.0066213 PMID: 23861743
9. Blaxter M, Mann J, Chapman T, Thomas F, Whitton C, Floyd R, et al. Defining operational taxonomic units using DNA barcode data. Philosophical Transactions of the Royal Society of London B: Biological Sciences. 2005; 360: 1935–1943. PMID: 16214751
10. Smith CR. The tray system for insect collections. Transactions of the Kansas Academy of Science. 1926; 31: 77–80, + 81a.
11. Aldrich JM. The Division of Insects in the United States National Museum. Annual Report of the Smithsonian Institution for the Year 1919, Smithsonian Institution. 1919; 367–379 p.
12. Howe D, Costanzo M, Fey P, Gojobori T, Hannick L, Hide W, et al. Big data: The future of biocuration. Nature. 2008; 455: 47–50. doi: 10.1038/455047a PMID: 18769432
13. Burge S, Attwood TK, Bateman A, Berardini TZ, Cherry M, O'Donovan C, et al. Biocurators and Biocuration: surveying the 21st century challenges. Database. 2012; 2012: 1–7.
14. German Barcode of Life Project. 2014; 5:14. Available: https://www.bolgermany.de/
15. The Quarantine Barcode of Life. 2014; 5:14. Available: http://www.qbol.org/en/qbol.htm
16. Foottit RG, Maw E, Hebert PDN. DNA Barcodes for Nearctic Auchenorrhyncha (Insecta: Hemiptera). PLOS ONE. 2014; 9: e101385. doi: 10.1371/journal.pone.0101385 PMID: 25004106
17. Maw HEL, Foottit RG, Hamilton KGA, Scudder GGE. Checklist of the Hemiptera of Canada and Alaska. Ottawa: NRC Research Press; 2000
18. Hajibabaei M, deWaard JR, Ivanova NV, Ratnasingham S, Dooh RT, Kirk SL, et al. Critical factors for assembling a high volume of DNA barcodes. Philosophical Transactions of the Royal Society B: Biological Sciences. 2005; 360: 1959–1967. PMID: 16214753
19. Ivanova NV, Dewaard JR, Hebert PDN. An inexpensive, automation–friendly protocol for recovering high–quality DNA. Molecular Ecology Notes. 2006; 6: 998–1002.
20. deWaard JR, Ivanova NV, Hebert PDN. Assembling DNA Barcodes: Analytical Protocols. In: Martin C, editor. Methods in Molecular Biology: Environmental Genetics Totowa, N.J., USA: Humana Press Inc.; 2008. pp. 275–293.
21. Hebert PDN, deWaard JR, Zakharov EV, Prosser SWJ, Sones JE, McKeown JTA, et al. A DNA 'Barcode Blitz': Rapid digitization and sequencing of a natural history collection. PLOS ONE. 2013; 8: e68535. doi: 10.1371/journal.pone.0068535 PMID: 23874660
22. Zahniser J, Dietrich CH. A review of the tribes of Deltocephalinae (Hemiptera: Auchenorrhyncha: Cicadellidae). European Journal of Taxonomy, North America. 2013; 45: 1–211.
23. Bartlett CR. A new genus and species of delphacid planthoppers (Hemiptera: Fulgoroidea) from Canada. Entomological News. 2002; 113: 97–102.
24. Dietrich CH, Dmitriev DA. Review of the New World genera of the leafhopper tribe Erythroneurini (Hemiptera: Cicadellidae: Typhlocybinae). Bulletin of the Illinois Natural History Survey. 2006; 37: 119–190.
25. de Vere N, Rich TCG, Ford CR, Trinder SA, Long C, Moore CW, et al. DNA barcoding the native flowering plants and conifers of Wales. PLOS ONE. 2012; 7: e37945. doi: 10.1371/journal.pone.0037945 PMID: 22701588
26. Gibson CM, Kao RH, Blevins KK, Travers PD. Integrative taxonomy for continental–scale terrestrial insect observations. PLOS ONE. 2012; 7: e37528. doi: 10.1371/journal.pone.0037528 PMID: 22666362
27. Suter WR, Rupprecht J. The Father Of The Schmitt Box. Entomological News. 1974; 85: 298–300.
28. Miller J, Beentjes K, van Helsdingen P, Ijland S. Which specimens from a museum collection will yield DNA barcodes? A time series study of spiders in alcohol. Zookeys. 2013; 365: 245–261. doi: 10.3897/zookeys.365.5787 PMID: 24453561
29. Steinke D. The School Malaise Trap Program: Involving students in active science through the exploration of insect diversity. Barcode Bulletin: The Newsletter of the International Barcode of Life Project. http://www.ibol.org/news-and-events/newsletter/. International Barcode of Life; 2013. pp. 8–9.

30. Gleason J, Williams J. BIObus provides unique opportunities: Students describe their experiences as part of the crew. Barcode Bulletin: The Newsletter of the International Barcode of Life Project. http://www.ibol.org/news-and-events/newsletter/. International Barcode of Life; 2012. pp. 8–9.
31. Goldstein P, DeSalle R. Integrating DNA barcode data and taxonomic practice: Determination, discovery, and description. Bioessays. 2010; 33: 135–147.
32. Ratnasingham S, Hebert PDN. BOLD's role in barcode data management and analysis: a response. Molecular Ecology Resources. 2011; 11: 941–942.
33. Team RDC. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2013.
34. Roskov Y, Abucay L, Orrell T, Nicolson D, Kunze T, Culham A, et al., eds. Species 2000 & ITIS Catalogue of Life, 2014 Annual Checklist. Species 2000: Naturalis, Leiden, the Netherlands; 2014. Available: www.catalogueoflife.org/annual-checklist/
35. Chamberlain S, Szöcs E. taxize—taxonomic search and retrieval in R (v. 0.3.0). F1000Research. 2013; 2: 191. doi: 10.12688/f1000research.2-191.v2 PMID: 24555091
36. Zhou C, Kandemir I, Walsh DB, Zalom FG, Lavine LC. Identification of *Lygus hesperus* by DNA Barcoding Reveals Insignificant Levels of Genetic Structure among Distant and Habitat Diverse Populations. PLOS ONE. 2012; 7: e34528. doi: 10.1371/journal.pone.0034528 PMID: 22479640
37. Rebijith KB, Asokan R, Kumar NKK, Srikumar KK, Ramamurthy VV, Bhat PS, et al. DNA Barcoding and Development of Species-Specific Markers for the Identification of Tea Mosquito Bugs (Miridae: Heteroptera) in India. Environmental Entomology. 2012; 41: 1239–1245. doi: 10.1603/EN12096 PMID: 23068182
38. Li M, Liu Q, Xi L, Liu Y, Zhu G, et al. Testing the Potential of Proposed DNA Barcoding Markers in Nezara virudula and Nezara antennata When Geographic Variation and Closely Related Species Were Considered. Journal of Insect Science. 2014; 14: 1–11. doi: 10.1093/jis/14.1.1 PMID: 25373148
39. Raupach MJ, Hendrich L, Küchler SM, Deister F, Morinière J, Gossner MM. Building-Up of a DNA Barcode Library for True Bugs (Insecta: Hemiptera: Heteroptera) of Germany Reveals Taxonomic Uncertainties and Surprises. PLOS ONE. 2014; 9: e106940. doi: 10.1371/journal.pone.0106940 PMID: 25203616
40. Tembe S, Shouche Y, Ghate HV. DNA barcoding of Pentatomomorpha bugs (Hemiptera: Heteroptera) from Western Ghats of India. Meta Gene. 2014; 2: 737–745. doi: 10.1016/j.mgene.2014.09.006 PMID: 25606457
41. Jung S, Duwal RK, Lee S. COI barcoding of true bugs (Insecta, Heteroptera). Molecular Ecology Resources. 2011; 11: 266–270. doi: 10.1111/j.1755-0998.2010.02945.x PMID: 21429132
42. Grebennikov VV, Heiss E. DNA barcoding of flat bugs (Hemiptera: Aradidae) with phylogenetic implications. Arthropod Systematics & Phylogeny. 2014; 72 213–219.
43. Park D- S, Foottit R, Maw E, Hebert PDN. Barcoding bugs: DNA–based identification of the True Bugs (Insecta: Hemiptera: Heteroptera). PLOS ONE. 2011; 6: e18749. doi: 10.1371/journal.pone.0018749 PMID: 21526211
44. Vasilevsky NA, Brush MH, Paddock H, Ponting L, Tripathy SJ, LaRocca GM, et al. On the reproducibility of science: unique identification of research resources in the biomedical literature. PeerJ 2013; 1: e148. doi: 10.7717/peerj.148 PMID: 24032093
45. Heraty JM, Burks RA, Cruaud A, Gibson GAP, Liljeblad J, Munro J, et al. A phylogenetic analysis of the megadiverse Chalcidoidea (Hymenoptera). Cladistics. 2012; 1–77.
46. Uhler PR. Notices of the Hemiptera Heteroptera in the collection of the late T. W. Harris, M.D. Proceedings of the Boston Society of Natural History. 1878; 19: 365–446.
47. Forbes SA. Fourteenth report of the state entomologist on the noxious and beneficial insects of the state of Illinois. Springfield, Ill.;1885.
48. Palisot de Beauvois AMFJ. Insectes recueillis en Afrique et en Amérique, dans les royaumes d'Oware et de Benin, a Saint-Domingue et dans les États-Unis pendant les années 1786–1797. Imprimerie de Fain et Compagnie, Paris. 1818;Parts 11–12: 173–208.
49. Hahn CW. Die Wanzenartigen Insecten. Nürnberg: der C.H. Zeh'schen Buchhandlung. 1833;147–148 p.
50. Aylagas E, Borja Á, Rodríguez-Ezpeleta N. Environmental status assessment using DNA metabarcoding: towards a genetics based Marine Biotic Index (gAMBI). PLOS ONE. 2014; 9: e90529. doi: 10.1371/journal.pone.0090529 PMID: 24603433
51. Virgilio M, Jordaens K, Breman FC, Backeljau T, De Meyer M. Identifying insects with incomplete DNA barcode libraries, African Fruit Flies (Diptera: Tephritidae) as a test case. PLOS ONE. 2012; 7: e31581. doi: 10.1371/journal.pone.0031581 PMID: 22359600

52. Maddison DR, Guralnick R, Hill A, Reysenbach A- L, McDade LA. Ramping up biodiversity discovery via online quantum contributions. Trends in Ecology & Evolution. 2012; 27: 72–77.

53. Pigg S. BioBlitz crowd counts all creatures great and small. The Star, Sunday ed. Toronto: Toronto Star Newspapers Limited. Accessed 16 September 2013.

54. Santschi L, Hanner RH, Ratnasingham S, Riconscente M, Imondi R. Barcoding Life's Matrix: Translating biodiversity genomics into high school settings to enhance life science education. PLOS Biology. 2013; 11: e1001471. doi: 10.1371/journal.pbio.1001471 PMID: 23382648