A DNA barcode survey of insect biodiversity in Pakistan

Muhammad Ashfaq [Corresp., 1], **Arif M. Khan** [2], **Akhtar Rasool** [3], **Saleem Akhtar** [4], **Naila Nazir** [5], **Nazeer Ahmed** [6], **Farkhanda Manzoor** [7], **Jayme Sones** [8], **Kate Perez** [8], **Ghulam Sarwar** [9], **Azhar A. Khan** [10], **Muhammad Akhter** [11], **Shafqat Saeed** [12], **Riffat Sultana** [13], **Hafiz M. Tahir** [14], **Muhammad A. Rafi** [15], **Romana Iftikhar** [16], **Muhammad Tayyib Naseem** [17], **Mariyam Masood** [18], **Muhammad Tufail** [19], **Santosh Kumar** [20], **Sabila Afzal** [21], **Jaclyn McKeown** [8], **Ahmed Ali Samejo** [13], **Imran Khaliq** [19], **Michelle L. D'Souza** [8], **Shahid Mansoor** [22], **Paul D. N. Hebert** [1]

[1] Centre for Biodiversity Genomics & Department of Integrative Biology, University of Guelph, Guelph, Canada

[2] Department of Biotechnology, University of Sargodha, Sargodha, Pakistan

[3] Centre for Animal Sciences and Fisheries, University of Swat, Mingora, Pakistan

[4] Directorate of Entomology, Ayub Agricultural Research Institute, Faisalabad, Pakistan

[5] Department of Entomology, University of Poonch, Rawalakot, Azad Kashmir, Pakistan

[6] Faculty of Life Sciences and Informatics, Balochistan University of Information Technology, Engineering and Management Sciences, Quetta, Pakistan

[7] Department of Zoology, Lahore College for Women University, Lahore, Pakistan

[8] Centre for Biodiversity Genomics, University of Guelph, Guelph, Canada

[9] Institute of Zoology, University of the Punjab, Lahore, Pakistan

[10] College of Agriculture, Bahauddin Zakariya University Bahadur Campus, Layyah, Pakistan

[11] Pulses Research Institute, Ayub Agricultural Research Institute, Faisalabad, Pakistan

[12] Faculty of Agriculture and Environmental Sciences, MNS University of Agriculture, Multan, Pakistan

[13] Department of Zoology, University of Sindh, Jamshoro, Pakistan

[14] Department of Zoology, Government College University Lahore, Lahore, Pakistan

[15] National Insect Museum, National Agricultural Research Center, Islamabad, Pakistan

[16] Department of Plant Pathology, Washington State University, Pullman, WA, United States

[17] Department of Biology, University of Copenhagen, Copenhagen, Denmark

[18] Government College Women University Faisalabad, Faisalabad, Pakistan

[19] Ghazi University, Dera Ghazi Khan, Pakistan

[20] Department of Zoology, Cholistan University of Veterinary and Animal Sciences, Bahawalpur, Pakistan

[21] Department of Zoology, University of Narowal, Narowal, Pakistan

[22] National Institute for Biotechnology and Genetic Engineering, Faisalabad, Pakistan

Corresponding Author: Muhammad Ashfaq
Email address: mashfaq@uoguelph.ca

Although Pakistan has rich biodiversity, many groups are poorly known, particularly insects. To address this gap, we employed DNA barcoding to survey its insect diversity. Specimens obtained through diverse collecting methods at 1,858 sites across Pakistan from 2010–2019 were examined for sequence variation in the 658 bp barcode region of the cytochrome *c* oxidase 1 (COI) gene. Sequences from nearly 49,000 specimens were assigned to 6,590 Barcode Index Numbers (BINs), a proxy for species, most (88%) also possessing a representative image on the Barcode of Life Data System (BOLD). By coupling morphological inspections with barcode matches on BOLD, every BIN was

assigned to an order (19) and most (99.8%) were placed to a family (362). However, just 40% of the BINs were assigned to a genus (1,375) and 21% to a species (1,364). Five orders (Coleoptera, Diptera, Hemiptera, Hymenoptera, Lepidoptera) accounted for 92% of the specimens and BINs. More than half of the BINs (59%) are so far only known from Pakistan, but others have also been reported from Bangladesh (13%), India (12%), and China (8%). Representing the first DNA barcode survey of the insect fauna in any South Asian country, this study provides the foundation for a complete inventory of the insect fauna in Pakistan while also contributing to the global DNA barcode reference library.

A DNA barcode survey of insect biodiversity in Pakistan

Muhammad Ashfaq[1]*, Arif M. Khan[2], Akhtar Rasool[3], Saleem Akhtar[4], Naila Nazir[5], Nazeer Ahmed[6], Farkhanda Manzoor[7], Jayme Sones[8], Kate Perez[8], Ghulam Sarwar[9], Azhar A. Khan[10], Muhammad Akhter[11], Shafqat Saeed[12], Riffat Sultana[13], Hafiz M. Tahir[14], Muhammad A. Rafi[15], Romana Iftikhar[16], Muhammad T. Naseem[17], Mariyam Masood[18], Muhammad Tufail[19], Santosh Kumar[20], Sabila Afzal[21], Jaclyn McKeown[8], Ahmed Ali Samejo[13], Imran Khaliq[19], Michelle D'Souza[8], Shahid Mansoor[22], Paul D. N. Hebert[1]

[1] Centre for Biodiversity Genomics & Department of Integrative Biology, University of Guelph, Guelph, ON, Canada.

[2] Department of Biotechnology, University of Sargodha, Sargodha, Pakistan.

[3] Centre for Animal Sciences and Fisheries, University of Swat, Mingora, Pakistan.

[4] Directorate of Entomology, Ayub Agricultural Research Institute, Faisalabad, Pakistan.

[5] Department of Entomology, University of Poonch Rawalakot, Azad Kashmir, Pakistan.

[6] Faculty of Life Sciences and Informatics, Balochistan University of Information Technology, Engineering and Management Sciences, Quetta, Pakistan.

[7] Department of Zoology, Lahore College for Women University, Lahore, Pakistan.

[8] Centre for Biodiversity Genomics, University of Guelph, Guelph, Canada

[9] Institute of Zoology, University of the Punjab, Lahore, Pakistan.

[10] College of Agriculture, Bahauddin Zakariya University Bahadur Campus Layyah, Pakistan.

[11] Pulses Research Institute, Ayub Agricultural Research Institute, Faisalabad, Pakistan.

[12] Faculty of Agriculture and Environmental Sciences, MNS University of Agriculture, Multan, Pakistan.

[13] Department of Zoology, University of Sindh, Jamshoro, Pakistan.

[14] Department of Zoology, Government College University, Lahore, Pakistan.

[15] National Insect Museum, National Agricultural Research Center, Islamabad, Pakistan.

[16] Department of Plant Pathology, Washington State University, Pullman, WA, USA.

[17] Department of Biology, University of Copenhagen, Denmark.

[18] Department of Zoology, Government College Women University, Faisalabad.

[19] Ghazi University, Dera Ghazi Khan, Pakistan.

[20] Department of Zoology, Cholistan University of Veterinary and Animal Sciences, Bahawalpur, Pakistan.

[21] Department of Zoology, University of Narowal, Pakistan.

[22] National Institute for Biotechnology and Genetic Engineering (NIBGE), Faisalabad, Pakistan.

*Corresponding author:

Muhammad Ashfaq

Centre for Biodiversity Genomics, University of Guelph, Guelph, ON, N1G 2W1, Canada

Email: mashfaq@uoguelph.ca

Phone: (519) 824-4120 Ext. 56393

Key Words: DNA barcoding, cytochrome oxidase I, Barcode Index Number, biodiversity overlap, BOLD.

Abstract:

Although Pakistan has rich biodiversity, many groups are poorly known, particularly insects. To address this gap, we employed DNA barcoding to survey its insect diversity. Specimens obtained through diverse collecting methods at 1,858 sites across Pakistan from 2010–2019 were examined for sequence variation in the 658 bp barcode region of the cytochrome *c* oxidase 1 (COI) gene. Sequences from nearly 49,000 specimens were assigned to 6,590 Barcode Index Numbers (BINs), a proxy for species, most (88%) also possessing a representative image on the Barcode of Life Data System (BOLD). By coupling morphological inspections with barcode matches on BOLD, every BIN was assigned to an order (19) and most (99.8%) were placed to a family (362). However, just 40% of the BINs were assigned to a genus (1,375) and 21% to a species (1,364). Five orders (Coleoptera, Diptera, Hemiptera, Hymenoptera, Lepidoptera) accounted for 92% of the specimens and BINs. More than half of the BINs (59%) are so far only known from Pakistan, but others have also been reported from Bangladesh (13%), India (12%), and China (8%). Representing the first DNA barcode survey of the insect fauna in any South Asian country, this study provides the foundation for a complete inventory of the insect fauna in Pakistan while also contributing to the global DNA barcode reference library.

Introduction:

With an area of 882,000 km^2, Pakistan includes four biomes (desert, mountain, temperate grassland, tropical seasonal forest) and portions of the Afro-Tropical, Indo-Malayan, and

Palearctic biogeographic realms (Cox and Moore 1993; Baig and Al-Subaiee 2009). Because of this physiographic and climatic variation, its faunal diversity is high (Anwar et al. 2008). While its vertebrate fauna is well known (Khan 1976; Roberts 1997; Grimmett et al. 2008), prior studies on other animal lineages have been restricted to specific taxa or regions (Inayat et al. 2010; Iftikhar et al. 2016a; Manzoor et al. 2020). Although insects are undoubtedly the most diverse component of the fauna, no wide-ranging assessments have been undertaken because of the scarcity of taxonomists, and because many insect species are undescribed (Ward & Lariviere 2004). Because conventional morphological approaches (Pompeo et al. 2017) are difficult to implement at scale (Fattorini 2013), the species count for Pakistan remains uncertain as it does globally (Scheffers et al. 2012).

The effectiveness of DNA barcoding (Hebert et al. 2003) in both specimen identification and species discovery (Kress et al. 2015) has stimulated its rapid adoption (DeSalle & Goldstein 2019). This work has generated DNA barcode coverage for more than 760,000 animal species on the Barcode of Life Data System (BOLD) (www.boldsystems.org) (Ratnasingham and Hebert 2007). The effectiveness of the Barcode Index Number (BIN) system (Ratnasingham and Hebert 2013) as a species proxy (Hausmann et al. 2013) has made it possible to rapidly evaluate species diversity, enabling large-scale biotic inventories (Hebert et al. 2016; Wirta et al. 2016). Because BINs show close congruence with species boundaries established through morphological study (Ortiz et al. 2017; Huemer et al. 2019) they can be used to delineate newly encountered species (Mitchell et al. 2020), discern cryptic species (Zahiri et al. 2017; Zhou et al. 2019), to plot species distributions (Ren et al. 2018), to estimate species richness in bulk samples (Andújar et

al. 2018; Braukmann et al. 2019), to analyze museum collections (Pentinsaari et al. 2020) and to assess faunal similarity at regional and global scales (Ashfaq et al. 2017).

The effectiveness of DNA barcoding coupled with advances in sequencing technology allow it to support large-scale biodiversity analysis (Wilson et al. 2017). However, the intensity of study has varied among regions (Weigand et al. 2019). For example, the BIN count (84,000) for Canada is 8x that for Russia (11,000), although the latter nation is 1.7x larger (www.boldsystems.org, accessed 7 September 2021). In a similar fashion, the BIN count for Germany (23,000) is 4x that for India (5,800), although the latter nation is 9x larger. The current study extends DNA barcode coverage for Pakistan to both advance knowledge of the taxonomic composition of its insect fauna and to develop ~~the~~ barcode reference ~~inventory~~ needed to support eDNA and metabarcoding studies.

Material and Methods

Sample collection and preparation:

Insects were sampled at 1,858 sites across Pakistan (Fig. 1) from 2010–2019 using both active and passive collecting methods including sweep nets, hand collections, hostplant beating, light traps, Malaise traps, pitfall traps, and UV illuminated sheets. Plans for the specimen collections/sites were approved by the Director, National Institute for Biotechnology and Genetic Engineering, Faisalabad under the project HEC No. 20-1403/R& D/09. The specimens were identified to an order and, where possible, to lower taxonomic ranks. Large specimens

were either pinned and preserved dry or placed in Matrix tubes with 95% ethanol. Small specimens were individually placed in a well containing 30 µl of 95% ethanol in 96-well microplates. Specimen metadata and an image (except for Malaise samples where only representative specimens of each BIN were imaged) were submitted to BOLD where the information can be accessed on both the specimen page and corresponding BIN page. Voucher specimens are archived at the National Institute for Biotechnology and Genetic Engineering (NIBGE), Faisalabad, Pakistan (with sample ID prefix NIBGE) or at the Centre for Biodiversity Genomics (CBG), Guelph, Canada (with ID prefix BIOUG).

DNA barcoding:

60,273 insects were barcoded following standard protocols (deWaard et al. 2019a; 2019b). In brief, a leg was removed with sterile forceps from each large specimen and transferred to a well preloaded with 30 µl of 95% EtOH. As smaller specimens were already in plates, they were ready for analysis, but vouchers were recovered after DNA extraction (Porco et al. 2010). DNA extraction, PCR amplification, and sequencing were performed at the Canadian Centre for DNA Barcoding (CCDB) following established protocols (Ivanova et al. 2006; Hebert et al. 2018; deWaard et al. 2019b). PCR reactions were either 6 µl or 12 µl (Hebert et al. 2013). Three quarters (73%) of the specimens were Sanger sequenced while the rest were analyzed using SMRT sequencing on a Sequel platform (Pacific Biosciences). Sanger sequencing employed BigDye Terminator Cycle Sequencing Kit (v3.1) on an Applied Biosystems 3730XL DNA Analyzer. Sequences were assembled, aligned and edited using CodonCode Aligner before submission to BOLD. SMRT sequencing employed protocols described by Hebert et al. (2018). The resultant

sequences were uploaded to mBRAVE (Multiplex Barcoding Research and Visualization Environment; www.mbrave.net) for editing (sequence trimming, quality filtering, de-replication), identification, and generation of operational taxonomic units (OTUs). The edited sequences were subsequently exported to BOLD for BIN assignment and reference library development. The specimen records, sequence data, electropherograms, and primer details are available in the dataset "DS-INSCTPAK" (dx.doi.org/10.5883/DS-INSCTPAK). All DNA extracts are stored within the DNA archive facility at the CBG.

Data analysis:

The final dataset (~~N~~=50,592) included 50,094 new barcode records and 498 public records on BOLD from specimens collected in Pakistan (Table S1). All records were assigned taxonomy and BINs following the workflow outlined by DeWaard et al. (2019b). In brief, once the barcode data was on BOLD, each record went through a taxonomic assignment and verification workflow. Earlier studies (Ashfaq et al. 2013; Nazir et al. 2014; Iftikhar et al. 2016b; Akhtar et al. 2018; Naseem et al. 2019) on five ~~lineages~~ (antlions, aphids, butterflies, grasshoppers, thrips) coupled analysis of barcode results with detailed morphological study by taxonomic specialists. All sequences meeting the quality criteria were either assigned to an existing BIN or founded a new one (Ratnasingham & Hebert 2013). Sequences founding a new BIN had to possess >500 bp of the barcode region with <1% ambiguous bases and no stop codons. Shorter sequences (300–495) that met the latter two quality criteria and that were a close sequence match to an established BIN were assigned to it (deWaard et al. 2019a). The remaining short sequences (1230) that failed to gain a BIN assignment were run through the stand-alone version of the

147 RESL algorithm (using the function Cluster Sequences on BOLD) to estimate the number of
148 additional OTUs among them. One representative from each OTU was then queried against the
149 BOLD ID Engine to link them with known BINs (deWaard et al. 2019b). The BIN details with
150 specimen records and representative images (where available) are accessible on BOLD
151 (dx.doi.org/10.5883/DS-INSCTPAK).

152 Various statistical approaches were used to estimate the number of insect species in Pakistan
153 (Chao & Chiu 2016) including the parametric estimator Preston's log-normal as well as non-
154 parametric estimators *Chao1*, and the *first-order* and *second-order jackknife*. A bias-corrected
155 version of each non-parametric estimator, designed to improve performance under conditions
156 of low sampling effort, was also included (Lopez et al. 2012). All estimates were calculated using
157 R packages *vegan* and *BAT*. In addition, a species accumulation curve was drawn based on a
158 sample-size-based rarefaction and extrapolation to at most double the minimum observed
159 sample size, guided by an estimated asymptote using the R package *iNEXT* (Hsieh et al. 2016)

160

161 Results:

162 DNA barcodes were recovered from 50,094 (83%) of the 60,273 specimens analyzed. The other
163 17% either failed to amplify or generated problematic sequences (e.g., contamination, NUMTs,
164 stop codons, endosymbionts) that were excluded from subsequent analysis. Considering orders
165 with 100 or more specimens, sequence recovery ranged from a low of 63% for Blattodea to

95% for Lepidoptera. Sequence recovery for the other four major orders of insects showed considerable variation (Diptera–91%, Coleoptera–80%, Hymenoptera–78%, Hemiptera–69%).

All 50,592 insects with a barcode were assigned to one of 19 orders while 99.8% received an assignment to one of 362 families (Table 1, S1 table). Five orders represented 92% of the specimens: Diptera (40%), Hymenoptera (21%), Lepidoptera (12%), Hemiptera (11%), and Coleoptera (8%) (Fig. 2). Six orders (Mantodea, Neuroptera, Odonata, Orthoptera, Psocodea, Thysanoptera) were each represented by >100 specimens while the other eight possessed fewer representatives (Fig. 2, S1 Table). Most of these sequences (98%) received a BIN assignment, leading to a total of 6,590 BINs. The other 1,230 barcode sequences did not meet the criteria for BIN assignment, but included 629 OTUs when analyzed using "Cluster Sequences" function on BOLD. The BOLD ID Engine assigned 82 of these OTUs to known BINs, but the other 547 OTUs likely represent taxa new to BOLD. Many (57%) of the 6,590 BINs were represented by two or more sequences, but 43% were represented by just a single specimen. The ratio of these singletons was above 40% in all five major orders, but was highest in Hymenoptera (48%). Most BINs (88%; N = 5,754) possessed an image of at least one voucher.

The percentage of records in each of the five major orders with a BIN assignment ranged from 93% (Coleoptera) to 99% (Diptera, Lepidoptera) with Hemiptera and Hymenoptera intermediate (96%) (Table 1). These five orders also contributed most of the BINs (92%) and families (81%) (Table 1, Fig. 3A, B). Only 40% of BINs were placed to a genus and 21% to a species, but this still led to records for 1,375 genera and 1,364 species (Table 1, S1 Table). Among the five major orders, more BINs were identified to a genus (72%) and species (41%) in

Lepidoptera than in the other four orders (Table 1). For example, just 13.8% of Diptera BINs and 10.2% of Hymenoptera BINs were assigned to a species (Table 1).

Specimen counts for the 362 families were highly variable as 15 families were each represented by >1,000 specimens while 38 had just one (S1 Table). This pattern was also reflected in the number of BINs as 15 families had > 100 BINs while 86 had just one. The Chironomidae (N = 3,258) and Braconidae (N = 2,174) were represented by the most specimens while Cecidomyiidae (238 BINs) and Platygastridae (230 BINs) were most diverse. Fig. 4 shows the BIN diversity and BIN:specimen ratio for the 15 families with >100 BINs. The ratio was highest (0.33) for Geometridae (Lepidoptera) and lowest (0.05) for Chironomidae. The species accumulation curve did not reach an asymptote indicating more species await detection (Fig. 5). Species estimates for the country ranged from 9,253 to 12,246 species suggesting that, on average, 40% of species remain to be sampled (Table 2).

BOLD was searched to ascertain if the 6,590 insect BINs from Pakistan were known from other countries. This analysis showed that 2,684 BINs (41%) were shared with 199 other countries while the others (3,906) are so far only known from Pakistan. The percentage of shared species ranged from 0.02% to 13%. Fig. 5 shows the overlap values between Pakistan and countries with >1,000 BINs. BIN overlap was higher with nearby countries (Bangladesh–13%, India–12%, China–8%) than for other regions. For example, Pakistan shared just 5% of its BINs with Australia, South Africa, and Germany (Fig. 6). The overlap between Canada and Costa Rica, both countries with >50,000 insect BINs, was only 4% and 1% respectively (Fig. 6).

Discussion:

Current estimates of the number of insect species which occur in Pakistan range from 5,000 (Ministry of Climate Change, 2019) to 20,000 species (Hasnain 1998), but they are certainly too low (Baig & Al-Subaiee 2009). The current study aimed to refine estimates of species richness by coupling DNA barcoding with the BIN system. With over 50,000 specimens sequenced, this study represents, by far, the largest effort to assemble a DNA barcode registry for the insect fauna of any South Asian country. While success (83%) in barcode recovery was good, it varied considerably among orders – from 63% for Blattodea to 95% for Lepidoptera. Similar variation in barcode recovery among different insect taxa has been reported in other studies (Geiger et al. 2016; Pentinsaari et al. 2020). For example, a study on the insect fauna of French Polynesia reported 91% recovery for Diptera versus 63% for Coleoptera (Ramage et al. 2017). Similarly, a large-scale Canadian study revealed 95% recovery for Diptera versus 77% for Hemiptera and 74% for Hymenoptera (DeWaard et al. 2019a). Although DNA quantity and quality play an important role (Ballare et al. 2019; Velasco-Cuervo et al. 2019), failures in primer binding often underlie low sequence recovery (Hajibabaei et al. 2005, Hebert et al. 2016). Such failures can lead to the underestimation of species richness in insect groups where recovery is low (Hebert et al. 2016).

The coupling of morphological inspection with barcode matches on BOLD (DeWaard et al. 2019a; 2019b) was very effective at placing BINs to an order (100%) and family (>99%). However, just 40% of the BINs could be assigned to a genus and 21% to a species indicating the need for better parameterization of the barcode reference library. This was particularly true for the three most diverse orders where species assignments were less than 15% (Diptera–13.8%, Coleoptera–13.3%, Hymenoptera–10.2%). Considerably higher assignment success has been

reported for Malaise samples from Germany (34%) and Canada (38%) (Geiger et al. 2016, DeWaard et al. 2019a) reflecting the more comprehensive DNA barcode reference libraries available for these nations. Despite the limited reference database (Virgilio et al. 2010), the present analysis identified representatives from 1,375 genera and 1,364 species showing the value of the global reference library (BOLD) which far exceeds the results obtained by morphology alone (Marshall et al. 2009). The present analysis revealed 6,590 BINs with species richness estimates indicating that the fauna of Pakistan certainly includes more than 10,000 species. As these estimates are based on specimens collected with uneven sampling and limited geographic coverage, they are likely to increase with more comprehensive efforts.

Although 19 insect orders were detected, five (Coleoptera, Diptera, Hemiptera, Hymenoptera, Lepidoptera) were dominant (92%), reinforcing prior results from morphological (Stork 2018) and barcoding studies (Ritter et al. 2019; Pentinsaari et al. 2020). Malaise traps preferentially capture low-flying insects such as Diptera and Hymenoptera (Cooksey & Barton 1981; deWaard et al. 2019b), the two orders that made 61% of the collections. Other studies have reported a similar pattern (Brown 2005; Karlsson et al. 2020). For example, a Canadian study found that Diptera comprised 57% of the collections (DeWaard et al. 2019b).

Fifteen of the 362 families dominated with 1,000 or more specimens and this pattern was also reflected in the BIN diversity. The uneven detection of families in the survey is supported by the fact that 38 families were represented by just one specimen and 88 by one BIN. Interestingly, nine of the 15 families with most BINs were dipterans and hymenopterans with Cecidomyiidae and Platygastridae being most diverse. Although Coleoptera has long been viewed as the most

diverse insect order (Stork et al. 2018), the present study supports conclusions from prior DNA barcode studies that Diptera and Hymenoptera are considerably more diverse (Hebert et al. 2016; Karlsson et al. 2020).

Because BOLD now hosts DNA barcode records for more than 760,000 animal species, it provides a good basis for assessing faunal overlap using BINs. Only 41% of the 6,590 insect BINs from Pakistan are currently known from other countries. As expected, BIN overlap was highest with neighboring countries. This result reflects the endemism of biodiversity (Werneck et al. 2012), and underscores the need to develop local biodiversity inventories. The current survey represents a first step towards building an inventory for the insect fauna of Pakistan.

Acknowledgements:

This study was enabled by grant 106106-001 “Engaging Developing Nations in iBOL” from the International Development Research Centre (IDRC) in Canada and by grant HEC No. 20-1403/R&D/09 “Sequencing DNA Barcodes of Economically Important Insect Species from Pakistan” from the Higher Education Commission of Pakistan. Sequence analysis was made possible by a grant from the Government of Canada through Genome Canada and Ontario Genomics in support of the International Barcode of Life (iBOL) project. This is a contribution to the ‘Food from Thought’ project supported by the Canada First Research Excellence Fund.

References:

271 Akhtar S, Ashfaq M, Zia A, Ali S, Ali GM, Zafar Y, and Farhatullah. 2018. First report and
272 redescription of five species of genus *Myrmeleon* (Neuroptra: Myrmeleontidae) from Pakistan.
273 *Journal of Biodiversity and Environmental Sciences* 13: 180–190.

274 Andújar C, Arribas P, Gray C, Bruce C, Woodward G, and Yu DW et al. 2018. Metabarcoding of
275 freshwater invertebrates to detect the effects of a pesticide spill. *Molecular Ecology* 27: 146–
276 166.

277 Anwar M, Jasra AW, and Ahmad I. 2008. Biodiversity conservation status in Pakistan - a review.
278 *Pakistan Journal of Forestry* 58: 39–48.

279 Ashfaq M, Akhtar S, Khan AM, Adamowicz SJ, and Hebert PDN (2013) DNA barcode analysis of
280 butterfly species from Pakistan points towards regional endemism. *Molecular Ecology*
281 *Resources* 13: 832–843.

282 Ashfaq M, and Hebert PDN. 2016. DNA barcodes for bio-surveillance: Regulated and
283 economically important arthropod plant pests. *Genome* 59: 933–945.

284 Ashfaq M, Akhtar S, Rafi MA, Mansoor S, and Hebert PDN. 2017. Mapping global biodiversity
285 connections with DNA barcodes: Lepidoptera of Pakistan. *PLoS ONE* 12: e0174749.

286 Baig MB, and Al-Subaiee FS. 2009. Biodiversity in Pakistan: Key issues. *Biodiversity* 10: 20–29.

287 Ballare KM, Pope NS, Castilla AR, Cusser S, Metz RP, and Jha S. 2019. Utilizing field collected
288 insects for next generation sequencing: Effects of sampling, storage, and DNA extraction
289 methods. *Ecology and Evolution* 9: 13690–13705.

290 Braukmann TWA, Ivanova NV, Prosser SWJ, Elbrecht V, Steinke D, Ratnasingham S, et al. 2019.
291 Metabarcoding a diverse arthropod mock community. *Molecular Ecology Resources* 19: 711–
292 727.

293 Brown B. 2005. Malaise trap catches and the crisis in Neotropical Dipterology. *American*
294 *Entomologist* 51: 180–183.

295 Chao A, and Chiu C. 2016. Nonparametric estimation and comparison of species richness.

296 *eLS*. DOI: 10.1002/9780470015902.a002632.

297 Cooksey LM, and Barton HE. 1981. Flying insect populations as sampled by Malaise trap on
298 Crowley's Ridge in northeast Arkansas. *Journal of Arkansas Academy of Sciences* 35: 29–32.

299 Cox CB, and Moore PD. 1993. Biogeography: an ecological and evolutionary approach (5th Edn).
300 Oxford: Blackwell Scientific Publications. 326p.

301 DeSalle R, and Goldstein P. 2019. Review and interpretation of trends in DNA barcoding.
302 *Frontiers of Ecology and Evolution* 7: 302.

303 deWaard JR, Ratnasingham S, Zakharov EV, Borisenko AV, Steinke D, Telfer AC, et al. 2019a. A
304 reference library for the identification of Canadian invertebrates: 1.5 million DNA barcodes,
305 voucher specimens, and genomic samples. *Scientific Data* 6: 308.

306 deWaard JR, Levesque-Beaudin V, deWaard SL, Ivanova NV, McKeown JTA, Miskie R, et al.
307 2019b. Expedited assessment of terrestrial arthropod diversity by coupling Malaise traps with
308 DNA barcoding. *Genome* 62: 85–95.

309 Fattorini S. 2013. Regional Insect inventories require long time, extensive spatial sampling and
310 good will. *PLoS ONE* 8: e62118.

311 Geiger M, Moriniere J, Hausmann A, Haszprunar G, Wägele W, Hebert PDN, et al. 2016. Testing
312 the Global Malaise Trap Program – How well does the current barcode reference library identify
313 flying insects in Germany? *Biodiversity Data Journal* 4: e10671.

314 Grimmett R, Roberts T, and Inskipp T. 2008. Birds of Pakistan. A & C Black Publishers Ltd.
315 London. 255p.

316 Hajibabaei M, deWaard JR, Ivanova NV, Ratnasingham S, Dooh RT, Kirk SL, et al. 2005. Critical
317 factors for assembling a high volume of DNA barcodes. *Philosophical Transactions of the Royal*
318 *Society London B: Biological Sciences* 360: 1959–1967.

319 Hasnain T. 1998. Implementation of Convention on Biological Diversity in Pakistan: Policy brief
320 series # 2. Sustainable Development Policy Institute (SDPI), Islamabad, Pakistan.

321 Hausmann A, Godfray HCJ, Huemer P, Mutanen M, Rougerie R, van Nieukerken EJ, et al. 2013.
322 Genetic patterns in European geometrid moths revealed by the Barcode Index Number (BIN)
323 system. *PLoS ONE* 8: e84518.

324 Hebert PDN, Cywinska A, Ball SL, and deWaard JR. 2003. Biological identifications through DNA
325 barcodes. *Proceedings of the Royal Society B: Biological Sciences* 270: 313–321.

326 Hebert PDN, Penton EH, Burns JM, Janzen DH, and Hallwachs W. 2004. Ten species in one: DNA
327 barcoding reveals cryptic species in the Neotropical skipper butterfly *Astraptes fulgerator*.
328 *Proceedings of the National Academy of Sciences USA*. 101: 14812–14817.

329 Hebert PDN, deWaard JR, Zakharov EV, Prosser SWJ, Sones JE, McKeown JTA, et al. 2013. A DNA
330 “Barcode Blitz”: rapid digitization and sequencing of a natural history collection. *PLoS ONE* 8:
331 e68535.

332 Hebert PDN, Ratnasingham S, Zakharov EV, Telfer AC, Levesque-Beaudin V, Milton MA, et al.
333 2016. Counting animal species with DNA barcodes: Canadian insects. *Philosophical Transactions*
334 *of the Royal Society B*. 371: 20150333.

335 Hebert PDN, Braukmann TWA, Prosser SWJ, Ratnasingham S, deWaard JR, Ivanova NV, et al.
336 2018. A Sequel to Sanger: amplicon sequencing that scales. *BMC Genomics* 19: 219.
337 https://doi.org/10.1186/s12864-018-4611-3.

338 Hsieh T, Ma K, and Chao A. 2016. iNEXT: an R package for rarefaction and extrapolation of
339 species diversity (Hill numbers). *Methods in Ecology and Evolution* 7:1451–1456.

340 Huemer P, Mutanen M, Sefc KM, and Hebert PDN. 2014. Testing DNA barcode performance in
341 1000 species of European Lepidoptera: large geographic distances have small genetic impacts.
342 *PLoS ONE* 9: e115774.

343 Huemer P, Wieser C, Stark W, Hebert PDN, and Wiesmair B. 2019. DNA barcode library of
344 megadiverse Austrian Noctuoidea (Lepidoptera) - a nearly perfect match of Linnean taxonomy.
345 *Biodiversity Data Journal* 7: e37734.

346 Iftikhar R, Ullah I, Diffie S, and Ashfaq M. 2016a. Deciphering Thysanoptera: A comprehensive
347 study on the distribution and diversity of thrips fauna in Pakistan. *Pakistan Journal of Zoology*
348 48: 1233–1240.

Iftikhar R, Ashfaq M, Rasool A, and Hebert PDN. 2016b. DNA barcode analysis of thrips (Thysanoptera) diversity in Pakistan reveals cryptic species complexes. *PLoS ONE* 11: e0146014.

Inayat TP, Rana SA, Khan HA, and Rehman K. 2010. Diversity of insect fauna in croplands of district Faisalabad. *Pakistan Journal of Agricultural Sciences* 47: 245–250.

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

Karlsson D, Hartop E, Forshage M, Jaschhof M, Ronquist F. 2020. The Swedish Malaise Trap Project: A 15 year retrospective on a countrywide insect inventory. Biodiversity Data Journal 8: e47255.

Khan MS. 1976. An annotated checklist and key to the amphibians of Pakistan. *Biologia* 22: 201–210.

Kress WJ, García-Robledo C, Uriarte M, and Erickson DL. 2015. DNA barcodes for ecology, evolution, and conservation. *Trends in Ecology and Evolution* 30: 25–35.

Lopez LCS, de Fracasso MPA, Mesquita DO, Palma ART, and Riul P. 2012. The relationship between percentage of singletons and sampling effort: a new approach to reduce the bias of richness estimates. *Ecological Indicators* 12: 164–169.

Manzoor M, Khan SW, Shah SA. 2020. An annotated checklist of butterflies at elevated protected areas of Pakistan. *Journal of Bioresource Management* 7: 41–52.

Marshall S, Paiero S, Buck M. 2009. Point Pelee National Park species list. [Online.] Available from http://www.uoguelph.ca/debu/pelee_specieslist.htm [accessed 20 June 2020]

Ministry of Climate Change, Pakistan. 2019. Pakistan's sixth national report to the United Nations Convention on Biological Diversity. Convention on Biological Diversity, Montreal, Canada. https://www.cbd.int/doc/nr/nr-06/pk-nr-06-en.pdf.

Mitchell A, Moeseneder CH, and Hutchinson PM. 2020. Hiding in plain sight: DNA barcoding suggests cryptic species in all 'well-known' Australian flower beetles (Scarabaeidae: Cetoniinae). *PeerJ*: DOI 10.7717/peerj.9348.

Naseem MT, Ashfaq M, Khan AM, Rasool A, Asif M, and Hebert PDN. 2019. BIN overlap confirms transcontinental distribution of pest aphids (Hemiptera: Aphididae). *PLoS ONE* 14: e0220426.

Nazir N, Mehmood K, Ashfaq M, and Rahim J. 2014. Morphological and molecular identification of acridid grasshoppers (Acrididae: Orthoptera) from Poonch division, Azad Jammu Kashmir, Pakistan. *Journal of Threatened* Taxa 6: 5544–5552.

Ortiz AS, Rubio RM, Guerrero JJ, Garre MJ, Serrano J, Hebert PDN, et al. 2017. Close congruence between Barcode Index Numbers (BINs) and species boundaries in the Erebidae (Lepidoptera: Noctuoidea) of the Iberian Peninsula. *Biodiversity Data Journal* 5: e19840.

Pentinsaari M, Blagoev GA, Hogg ID, Levesque-Beaudin V, Perez K, Sobel CN, et al. 2020. DNA barcoding survey of an Arctic arthropod community: implications for future monitoring. *Insects* 11: 46.

Pompeo PN, de Oliveira FLCI, Santos MAB, Mafra AL, Filho OK, and Baretta D. 2017. Morphological diversity of Coleoptera (Arthropoda: Insecta) in agriculture and forest systems. *Revista Brasileira de Ciência do Solo* 41: e0160433.

Porco D, Rougerie R, Deharveng L, and Hebert P. 2010. Coupling non-destructive DNA extraction and voucher retrieval for small soft-bodied arthropods in a high-throughput context: The example of Collembola. *Molecular Ecology Resources* 10: 942–945.

Ramage T, Martins-Simoes P, Mialdea G, Allemand R, Duplouy A, Rousse P, et al. 2017. A DNA barcode-based survey of terrestrial arthropods in the Society Islands of French Polynesia: host diversity within the SymbioCode Project. *European Journal of Taxonomy* 272: 1–13

Ratnasingham S, and Hebert PDN. 2007. BOLD: the Barcode of Life Data system (www.barcodinglife.org). *Molecular Ecology Notes* 7: 355–364.

Ratnasingham S, and Hebert PDN. 2013. A DNA-based registry for all animal species: the Barcode Index Number (BIN) system. *PLoS ONE* 8: e66213.

Ren J, Ashfaq M, Hu X, Ma J, Liang F, Hebert PDN, et al. 2018. Barcode index numbers expedite quarantine inspections and aid the interception of nonindigenous mealybugs (Pseudococcidae). *Biological Invasions* 20: 449–460.

Ritter CD, Häggqvist S, Karlsson D, Sääksjärvi IE, Muasya AM, Nilsson RH, Antonelli A. 2019. Biodiversity assessments in the 21st century: the potential of insect traps to complement environmental samples for estimating eukaryotic and prokaryotic diversity using high-throughput DNA metabarcoding. *Genome* 62: 147–159.

Roberts TJ. 1997. The mammals of Pakistan. Oxford University Press; Revised Edition. 525p.

Scheffers BR, Joppa LN, Pimm SL, and Laurance WF. 2012. What we know and don't know about earth's missing biodiversity. *Trends in Ecology & Evolution* 27: 501–510.

Stork NE. 2018. How many species of insects and other terrestrial arthropods are there on Earth? *Annual Review of Entomology* 63: 31–45.

Velasco-Cuervo SM, Aguirre-Ramirez E, Gallo-Franco JJ, Obando RG, Carrejo N, and Toro-Perea N. 2019. Saving DNA from museum specimens: the success of DNA mini-barcodes in haplotype reconstruction in the genus *Anastrepha* (Diptera: Tephritidae). *Journal of Advanced Research* 16: 123–134.

Virgilio M, Backeljau T, Nevado B, and De Meyer M. 2010. Comparative performances of DNA barcoding across insect orders. *BMC Bioinformatics* 11: 206.

Ward DF, and Lariviere M-C. 2004. Terrestrial invertebrate surveys and rapid biodiversity assessment in New Zealand: lessons from Australia. *New Zealand Journal of Ecology* 28: 151–159.

Weigand H, Beermann AJ, Čiampor F, Costa FO, Csabai Z, Duarte S, et al. 2019. DNA barcode reference libraries for the monitoring of aquatic biota in Europe: Gap-analysis and recommendations for future work. *Science of the Total Environment* 678: 499-524.

Werneck FP, Nogueira C, Colli GR, Sites JW, and Costa GC. 2012. Climatic stability in the Brazilian Cerrado: implications for biogeographical connections of South American savannas,

426 species richness and conservation in a biodiversity hotspot. *Journal of Biogeography* 39: 1695–
427 1706.

428 Wilson JJ, Sing KW, Floyd RM, and Hebert PDN. 2017. DNA barcodes and insect biodiversity.
429 https://doi.org/10.1002/9781118945568.ch17. In: Foottit RG & Adler PH, Editors. Insect
430 Biodiversity: Science and Society. Vol. 1, 2nd Edition. Blackwell Publishing Ltd., Oxford.

431 Wirta H, Varkonyi G, Rasmussen C, Kaartinen R, Schmidt NM, Hebert PDN, et al. 2016.
432 Establishing a community-wide DNA barcode library as a new tool for arctic research. *Molecular*
433 *Ecology Resources* 16: 809–822.

434 Zahiri R, Lafontaine JD, Schmidt BC, deWaard JR, Zakharov EV, Hebert PDN (2017) Probing
435 planetary biodiversity with DNA barcodes: The Noctuoidea of North America. *PLoS ONE* 12:
436 e0178548.

437 Zhou Z, Guo H, Han L, Chai J, Che X, Shi F (2019) Singleton molecular species delimitation based
438 on COI-5P barcode sequences revealed high cryptic/undescribed diversity for Chinese katydids
439 (Orthoptera: Tettigoniidae). *BMC Evolutionary Biology* 19: 79.

440

441 Figure legends:

442 Fig. 1: Map showing collection sites for insects examined in this study. The size and colour of
443 each site point indicates the number of specimens sampled.

444 Fig. 2: Pie chart showing the number of specimens barcoded from each of the 19 insect orders.

445 Fig. 3: Taxonomic (A) and BIN assignments (B) for the 12 insect orders represented by >50
446 specimens.

447 Fig. 4: BIN diversity and BIN/specimen ratio for the 15 insect families represented by >100 BINs.

448 Fig. 5: Sample-size-based rarefaction (solid line) and extrapolation (dashed line) sampling
449 curves for 49,363 specimens with barcodes from Pakistan. Solid dots represent the observed
450 richness of 6,590 species. Curved is estimated to reach an asymptote at 10,382 species.

451 Fig. 6: Percentage of insect BINs shared between Pakistan and the 70 other nations with >1,000
452 insect BINs on the Barcode of Life Data Systems (BOLD).

453

Table 1(on next page)

Table 1: Number of specimens belonging to 19 insect orders from Pakistan with DNA barcode records. The number of families, genera, species, and BINs is reported for each order.

* *For recognition as a new BIN, a sequence must include >500 bp of the barcode region (positions 70 bp to 700 bp in the BOLD alignment) and possess <1% ambiguous bases.*

Table 1: Number of specimens belonging to 19 insect orders from Pakistan with DNA barcode records. The number of families, genera, species, and BINs is reported for each order.

Order	Specimens with barcodes	Specimens assigned to BINs (%)	BINs recovered	OTUs without BIN	Singleton BINs (%)	BINs assigned to family (%)	Families recovered	BINs assigned to genus (%)	Genera recovered	BINs assigned to species (%)	Species recovered
Blattodea	64	84	19	5	36.8	100	5	78.9	9	52.6	10
Coleoptera	3889	93	819	123	45.2	100	56	21.9	118	13.3	119
Dermaptera	24	83.3	3	2	33.3	100	2	33.3	1	0.0	0
Diptera	20095	99	1684	94	40.1	99.0	68	29.8	212	13.8	222
Embioptera	28	96.4	7	1	14.3	100	2	14.3	1	14.3	1
Hemiptera	5859	96.5	642	73	41.9	98.3	59	31.6	132	22.6	135
Hymenoptera	10542	96	1711	177	47.7	99.4	50	34.7	226	10.2	170
Lepidoptera	6064	99.4	1233	24	42.5	99.6	62	71.9	516	41.5	514
Mantodea	113	97.3	36	2	50.0	100	2	13.9	4	5.6	2
Megaloptera	6	100	1	0	0.0	100	1	100	1	100	1
Neuroptera	559	92.3	99	6	39.4	99.0	7	54.5	30	36.4	32
Odonata	353	92.6	51	11	21.6	100	12	92.2	30	88.2	47
Orthoptera	1409	97.59	163	21	30.1	100	12	44.2	53	37.4	54
Phasmatodea	4	75	3	1	100.0	100	1	0.0	0	0.0	0
Psocodea	950	97.5	31	5	22.6	93.5	13	38.7	10	19.4	6
Strepsiptera	2	100	1	0	0.0	100	1	100	1	0.0	0
Thysanoptera	618	99.3	76	2	34.2	100	3	80.3	27	69.7	48
Trichoptera	11	100	10	0	90.0	100	6	60.0	4	40.0	4
Zygentoma	2	100	1	0	0.0	0.0	0	0.0	0	0.0	0
Total	*50,592*	*97.6%*	*6,590*	547	*42.9%*	*99%*	*362*	*40%*	*1,375*	*21%*	*1,364*

** For recognition as a new BIN, a sequence must include >500 bp of the barcode region (positions 70 bp to 700 bp in the BOLD alignment) and possess <1% ambiguous bases.*

Table 2(on next page)

Table 2: Species richness estimates based on the abundances of the 6,590 insect BINS encountered at 1,858 sites across Pakistan.

Seven estimates were calculated: Preston's log-normal (PRESTON), Chao1 (CHAO1), first-order jackknife (JACK1AB), second-order jackknife (JACK2AB), and their bias-corrected complements (CHAO1P, JACK1ABP, JACK2ABP).

Table 2: Species richness estimates based on the abundances of the 6,590 insect BINS encountered at 1,858 sites across Pakistan. Seven estimates were calculated: Preston's log-normal (PRESTON), Chao1 (CHAO1), first-order jackknife (JACK1AB), second-order jackknife (JACK2AB), and their bias-corrected complements (CHAO1P, JACK1ABP, JACK2ABP).

SPECIMENS	BINS	PRESTON	CHAO1	CHAO1P	JACK1AB	JACK1ABP	JACK2AB	JACK2ABP
49,363	6,590	9,253	10,377	12,285	9,416	11,147	11,189	12,246

Figure 1

Map showing collection sites for insects examined in this study. The size and colour of each site point indicates the number of specimens sampled.

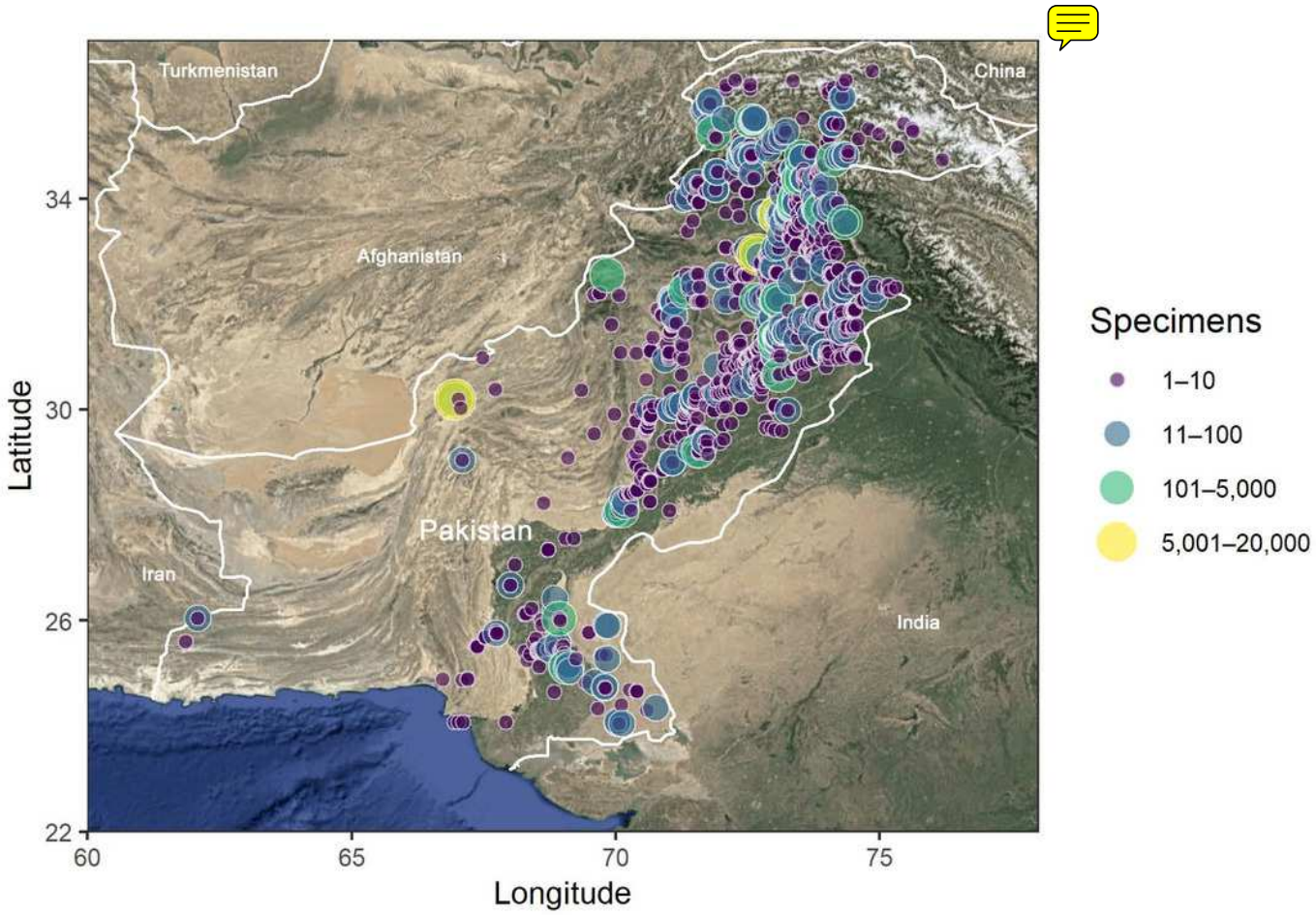


Figure 2

Pie chart showing the number of specimens barcoded from each of the 19 insect orders.

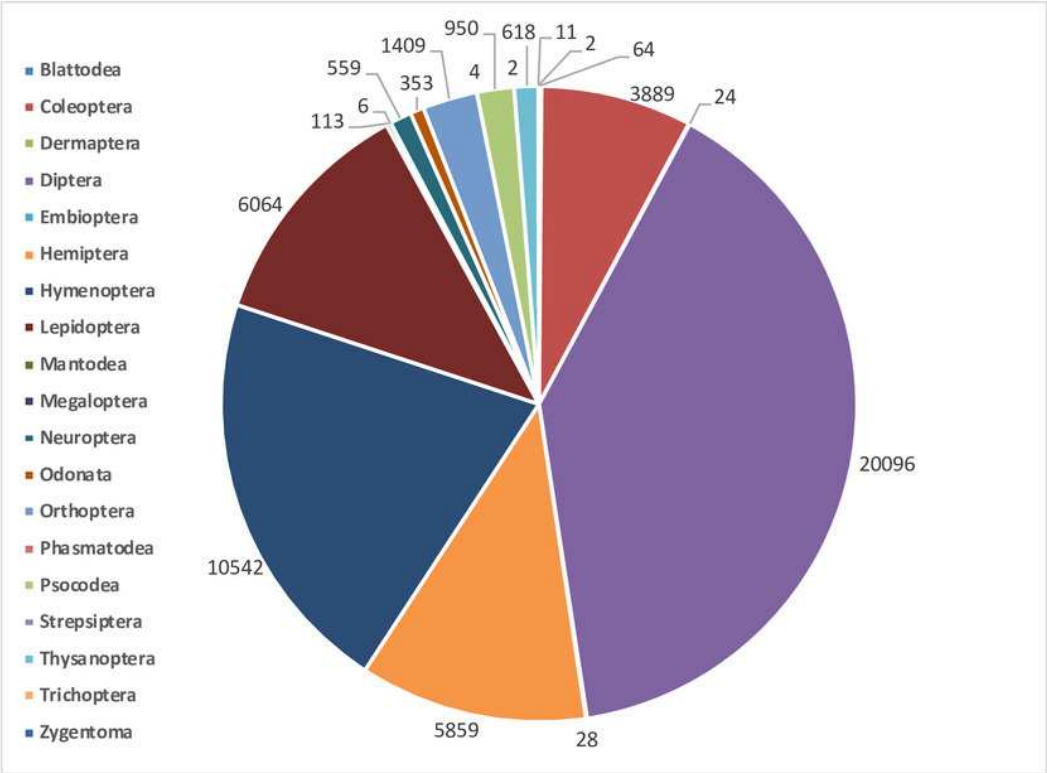
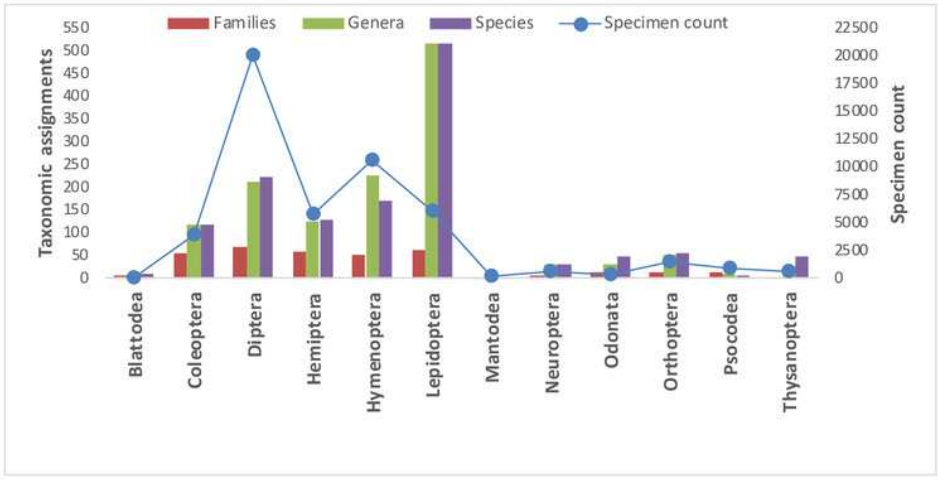
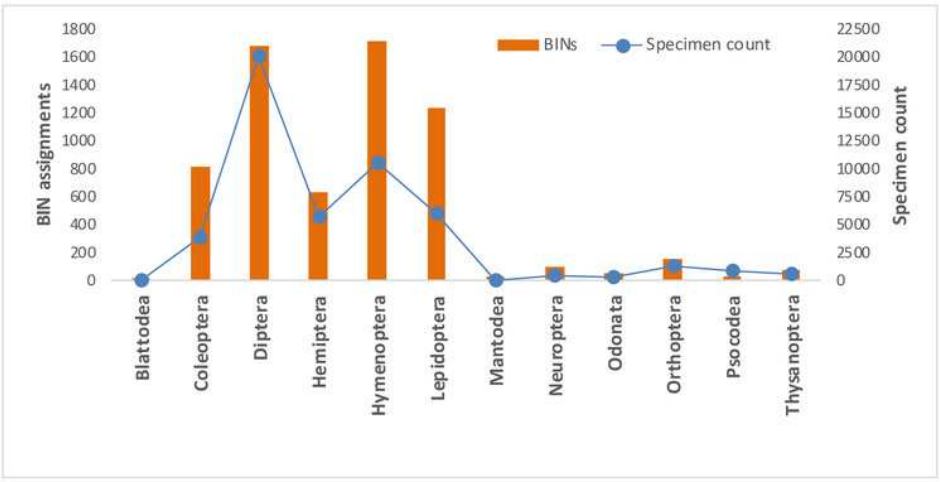

Blattodea
Coleoptera
Dermaptera
Diptera
Embioptera
Hemiptera
Hymenoptera
Lepidoptera
Mantodea
Megaloptera
Neuroptera
Odonata
Orthoptera
Phasmatodea
Psocodea
Strepsiptera
Thysanoptera
Trichoptera
Zygentoma
3889
24
20096
28
5859
10542
6064
113
6
559
1409
950
4
618
2
11
2
64

Figure 3

Taxonomic (A) and BIN assignments (B) for the 12 insect orders represented by >50 specimens.


Families
Genera
Species
Specimen count
Taxonomic assignments
Specimen count
550
500
450
400
350
300
250
200
150
100
50
0
22500
20000
17500
15000
12500
10000
7500
5000
2500
0
Blattodea
Coleoptera
Diptera
Hemiptera
Hymenoptera
Lepidoptera
Mantodea
Neuroptera
Odonata
Orthoptera
Psocodea
Thysanoptera

(A)


BINs
Specimen count
BIN assignments
Specimen count
1800
1600
1400
1200
1000
800
600
400
200
0
22500
20000
17500
15000
12500
10000
7500
5000
2500
0
Blattodea
Coleoptera
Diptera
Hemiptera
Hymenoptera
Lepidoptera
Mantodea
Neuroptera
Odonata
Orthoptera
Psocodea
Thysanoptera

(B)

Figure 4

BIN diversity and BIN/specimen ratio for the 15 insect families represented by >100 BINs.

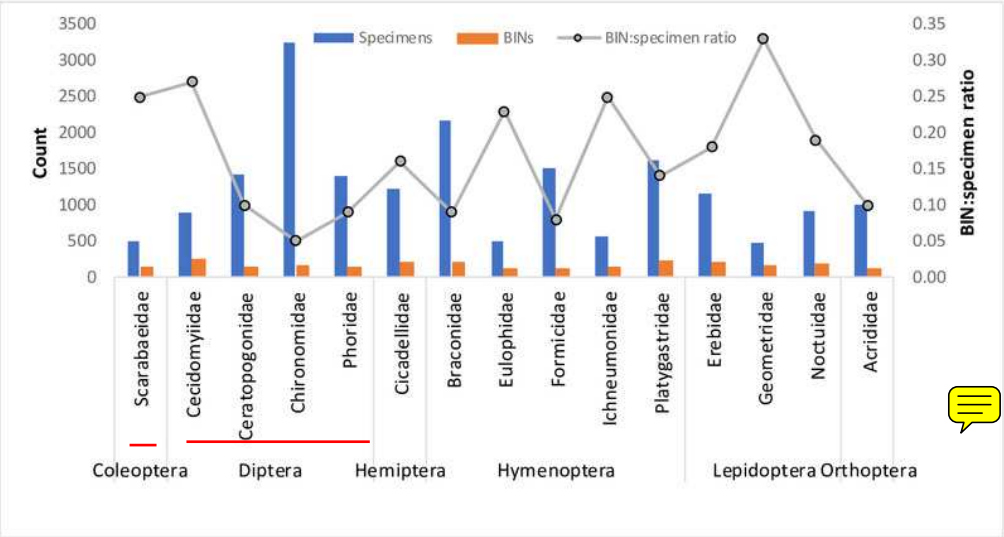

Specimens
BINs
BIN:specimen ratio
Count
BIN:specimen ratio
3500
3000
2500
2000
1500
1000
500
0
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
Scarabaeidae
Cecidomyiidae
Ceratopogonidae
Chironomidae
Phoridae
Cicadellidae
Braconidae
Eulophidae
Formicidae
Ichneumonidae
Platygastridae
Erebidae
Geometridae
Noctuidae
Acrididae
Coleoptera
Diptera
Hemiptera
Hymenoptera
Lepidoptera
Orthoptera

Figure 5

Sample-size-based rarefaction (solid line) and extrapolation (dashed line) sampling curves for 49,363 specimens with barcodes from Pakistan.

Solid dots represent the observed richness of 6,590 species. Curve is estimated to reach an asymptote at 10,382 species.

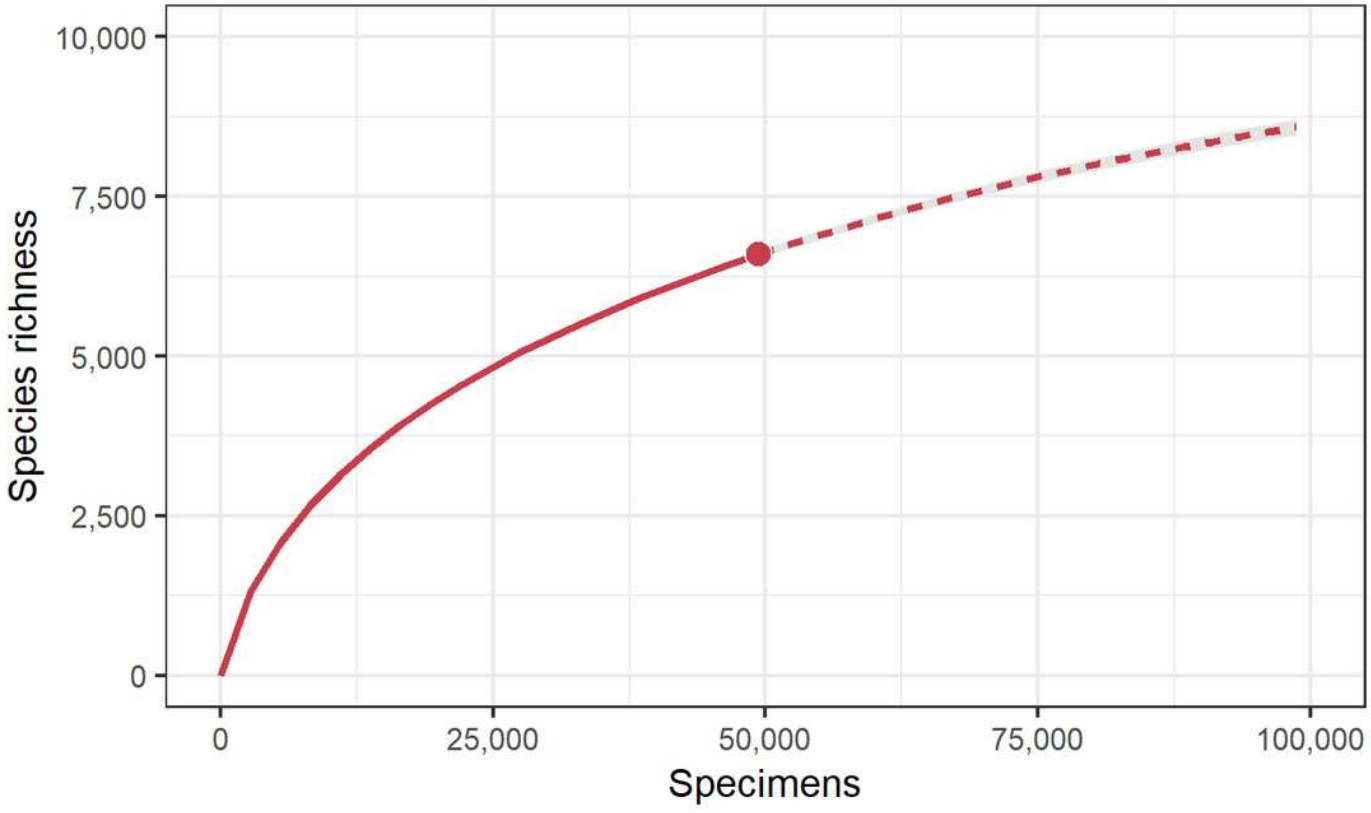
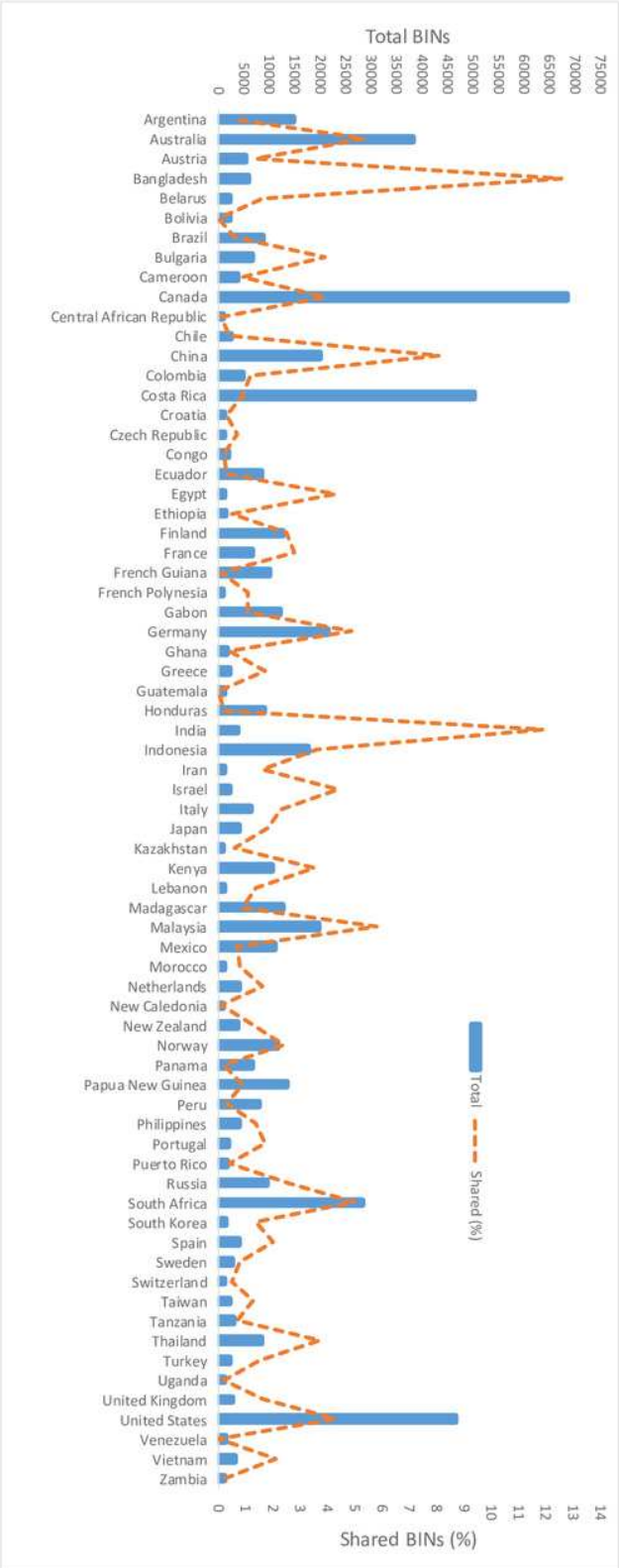


Figure 6

Percentage of insect BINs shared between Pakistan and the 70 other nations with >1,000 insect BINs on the Barcode of Life Data Systems (BOLD).


Total BINs
0
5000
10000
15000
20000
25000
30000
35000
40000
45000
50000
55000
60000
65000
70000
75000
Argentina
Australia
Austria
Bangladesh
Belarus
Bolivia
Brazil
Bulgaria
Cameroon
Canada
Central African Republic
Chile
China
Colombia
Costa Rica
Croatia
Czech Republic
Congo
Ecuador
Egypt
Ethiopia
Finland
France
French Guiana
French Polynesia
Gabon
Germany
Ghana
Greece
Guatemala
Honduras
India
Indonesia
Iran
Israel
Italy
Japan
Kazakhstan
Kenya
Lebanon
Madagascar
Malaysia
Mexico
Morocco
Netherlands
New Caledonia
New Zealand
Norway
Panama
Papua New Guinea
Peru
Philippines
Portugal
Puerto Rico
Russia
South Africa
South Korea
Spain
Sweden
Switzerland
Taiwan
Tanzania
Thailand
Turkey
Uganda
United Kingdom
United States
Venezuela
Vietnam
Zambia
Total
Shared (%)
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
Shared BINs (%)