Ghana BINs draft manuscript

Oct 17, 2023

# Abstract

# Introduction

## General statements on the issues of biodiversity-discovery in the tropics

### There’s lots of it

Like, loads. But the worst part is most of it is completely undescribed, basically unknown.

There have been many estimates of global arthropod richness, for example an article from the 2018 Annual Review of Entomology (Stork 2018) estimated 7 million species of terrestrial arthropods, with only 1 million of them formally named. Under this estimate, only 20% of insect species had been identified. The global dry biomass of this immense richness of insects is estimated at 300 million metric tons (Rosenberg et al. 2023), a similar mass to the combined biomass of humanity and its livestock.

### A lot of it’s under threat

The tropics are going through some things: climate change, widespread land-use change, introduced species, pollution, nitrification. The negative impacts of these threats can act synergistically, but with uneven impacts even for closely-related taxa (Williamson et al. 2022). The impacts of these changes are better-understood for more heavily-studied taxa such as vertebrates (Wagner et al. 2021). Low levels of baseline data and high per-taxa effort of identifying any observed taxa make estimates of insect declines very difficult. As a result, the few studies on insect declines tend to be clustered in relatively species-poor, wealthy western regions such as Europe and North America (Blüthgen et al. 2023)

### Formally describing any of it takes fucking ages

While collection of arthropod samples is laborious, the huge number of specimens that can be collected has long been an issue (Gotelli 2004), as a bottleneck is then hit with processing and identifying each individual arthropod .

It’s worthwhile, but doing so is a long and meticulous job, would be nice if we could lend a helping-hand somehow.

## DNA barcoding can be helpful for this

### Brief description of DNA barcoding

If you make a database of DNA sequences from things you’ve identified, you can compare unknown sequences to that database and infer what they probably came from.

### The BIN system: what it is **and what it fucking isn’t**

While DNA barcodes are invaluable when comparing unknown samples to a reference database of sequences from known-species, they can have an additional purpose: cataloguing unknown samples themselves. In addition to providing a publicly-available repository of DNA barcodes with associated metadata, the Barcode of Life Database (BOLD) (Ratnasingham and Hebert 2007) assigns a unique identifier to sequences uploaded to it, known as a Barcode Index Number or ‘BIN’ (Ratnasingham and Hebert 2013). BINs are clusters of similar DNA sequences, and so can act as a species-proxy when a species has yet to be formally described. While possessing a formal taxonomic description and binomial name remains the gold-standard for taxa delimitation, knowing that two samples have different (or identical) BINs can be of great value for ecological study, and sampling that generates novel BINs can highlight geographic and taxonomic areas of biodiversity that are deserving of increased attention. Uploading samples to BOLD’s BIN system provides a useful resource for further study, by allowing future researchers to match their sample to a previous record, with taxonomic, geographic and temporal metadata attached.

While some researchers have used BINs as proxies for species (CITE), or even as a major strand of evidence when describing new species outright (Sharkey et al. 2021), this remains controversial at best (Meier et al. 2022) due to the inherent limitations of using a single gene and unreproducible algorithm to delineate taxonomy. We here view BINs as Operational Taxonomic Units (OTUs): a valuable tool in biology and ecology where samples are classified into clusters based on shared or diverging traits.

In this form, BINs have been used to analyse ecological phenomena in the tropics (Ashfaq et al. 2018; Bukowski et al. 2022; D’Souza and Hebert 2018; deWaard et al. 2019), while simultaneously providing what is frequently the first scientific record of previously unknown taxa.

## Ghana, and Africa more generally

West African forests are designated as a biodiversity hotspot [IN FAIRNESS IT’S WEST AFRICAN FORESTS that are designated, dunno if our sites exactly qualify…]: a region with a high level of biodiversity, that is threatened by human actions (CITE). This issue is compounded by the relatively low proportion of the area’s species yet to be formally described; hampering ecological study, estimates of species distribution, extinction and more.

In this article we describe a high-intensity arthropod-sampling campaign in Ghana, capturing (insert number of samples here), and then generating a database of BINs from the material. We analyse the data to show rates of taxonomic discovery by trapping effort, and the geographic distribution of BINs detected in this dataset. This data is publicly available for use by the scientific community.

## Closing paragraph

*(Nice summary goes here)*

# Methods

## Field sites

Working at two villages in the Volta region of Ghana (Abutia Amegame and Mafi Agorve) we intensely collected arthropod samples during monthly visits February 2019-March 2020, and then April 2021-June 2021 (pandemic happened)(Check the full sampling dates with Talya, these are the unique sampling months in our dataset but they might not all have been created equal).

## Sampling

At each site, for every visit, *x* transects were set up, with *w* malaise traps, *x* CDC traps, *y* yellow pan traps and *z* heath traps per transect. Traps were left for approximately 24 hours, apart from malaise traps, whose bottles were exchanged approximately every 6 hours within the 24-hour period.

## DNA barcoding

* Sample sorting, shipment, timeframes etc
* Samples were DNA barcoded at the Canadian Centre for DNA Barcoding (CCDB), using their standard protocol of photographing each specimen, before performing a non-destructive DNA extraction, PCR using the CO1 ‘Folmer’ region (Folmer et al. 1994), and sequencing on a PacBio Sequel. Through the BOLD data-management platform (Ratnasingham and Hebert 2007), samples were assigned provisional taxonomy and a BIN (Ratnasingham and Hebert 2013).

## Data analyses

Our data was downloaded from BOLD on 14 September 2023, and BOLD was queried for information on the publicly available BINs that matched BINs in our dataset.

Using the iNEXT (Chao et al. 2014) R package, sampling completeness was calculated for each taxonomic order and trap type used; using the rate of BIN-accumulation to estimate the sampling effort required before all potential BINs in the landscape that would be captured by a given trap-type were obtained. Calculations only took place for combinations of taxonomic order and trap-type where at least 20 BINs were detected. I.e. calculations would not have taken place for spiders and malaise traps, if only 19 spider BINs were detected in malaise traps.

### NMDS

A series of non-metric multidimensional scaling (NMDS) analyses were ran for the dataset using the ‘vegan’ R package (Oksanen et al. 2022). Taxonomic groups were only included in the dataset if they contained a minimum of 10 samples (i.e. a taxonomic order would only be included in the order-level NMDS if the dataset contained 10 or more samples of that order). CDC traps were excluded from this analysis due to the low number of BINs obtained from them.

To analyse the extent of potential differences between the communities included in the NMDS analyses, we also ran a Permutational Multivariate Analysis of Variance Using Distance Matrices, using bray-curtis dissimilarity, using the vegan R package, and the betadispersion of the communities in each trap was calculated.

### Geographic analyses.

We queried BOLD for the available metadata all BINs in our dataset that already had public matches from other studies.The resulting dataset was used to identify countries and continents which shared many BINs with our dataset, and the taxonomic composition of those BINs. The 20 countries with the highest numbers of public BINs matching BINs in our dataset were queried between 18 September 2023 and 20 September 2023, to show areas where similar communities have been detected. This is more likely to represent geographic biases in sampling than similarities in the arthropod communities themselves.

### Closing sentence

All analyses took place using R (R Core Team 2023). All code for this manuscript is available at https://github.com/hemprichbennett/ghana\_bins.

# Results

## Basic summary statistics.

We detected 7,569 unique BINs across the 63,411 samples sequenced as of 14 September 2023 (see [Table 1](#tbl-per-traptype-summary)). For the 13 most abundant taxonomic orders, the average estimated sampling completeness was 54.9%.

Table 1: The number of samples and BINs sequenced for each trap type used

| type | Number of samples | Number of BINs |
| --- | --- | --- |
| Cdc | 3,035 | 16 |
| Heath | 41,278 | 5,066 |
| Malaise | 8,582 | 2,550 |
| Pitfall | 5,239 | 668 |
| Yellow Pan | 4,568 | 1,009 |

Of the 200 most-abundant BINs in our dataset, 78 were already publicly available, 72 had already been sequenced but no other representatives of that BIN were publicly available on BOLD, and 50 BINs were unique to our project.

## Trap types

### Trap types and novel BINs

Heath traps provided the majority of samples barcoded ([Table 1](#tbl-per-traptype-summary)), and also provided the majority of novel BINs in our analysis [Figure 1](#fig-bin_availability).

|  |
| --- |
| Figure 1: Public availability of BINs detected, by trap-type |

### BIN accumulation

No trap types or taxonomic orders approached completeness ([Figure 2](#fig-type1) and [Figure 3](#fig-type2)).

|  |
| --- |
| Figure 2: Type 1 iNEXT plot |

|  |
| --- |
| Figure 3: Type 2 iNEXT plot |

### NMDS

NMDS ordinations for both taxonomic Order and Family showed some overlap between the trap types ([Figure 4](#fig-order-nmds) and [Figure 5](#fig-family-nmds)). The Permutational Multivariate Analysis of Variance test showed a significant difference between trap types at both the order level (p =9.999^{-5}) and family level (p = 9.999^{-5}). However betadispersion was significant for both the order (p = 2.3136^{-34}) and family (p = 3.3501^{-21}) analyses. *(I’ll tidy those numbers up later, just not sure how to do it yet)*

|  |
| --- |
| Figure 4: Order-level NMDS plot |

|  |
| --- |
| Figure 5: Family-level NMDS plot |

## Geographic analyses

There were 1,846 distinct BINs in our project which are also found publicly on BOLD. Those BINs are found across 170 countries in 7 geographic regions, as shown in [Table 2](#tbl-geographic_matches):

Table 2: The number of BINs found in our dataset that have been found in each major geographic region.

| Order name | East Asia & Pacific | Europe & Central Asia | Latin America & Caribbean | Middle East & North Africa | North America | South Asia | Sub-Saharan Africa |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Araneae | 1 | 1 | 1 | 1 | 1 | 1 | 9 |
| Blattodea | 0 | 0 | 1 | 0 | 0 | 0 | 6 |
| Coleoptera | 20 | 14 | 13 | 10 | 14 | 28 | 136 |
| Diptera | 62 | 19 | 33 | 87 | 24 | 85 | 604 |
| Entomobryomorpha | 3 | 0 | 3 | 2 | 0 | 2 | 2 |
| Ephemeroptera | 0 | 0 | 0 | 1 | 0 | 0 | 1 |
| Hemiptera | 37 | 21 | 23 | 36 | 22 | 46 | 201 |
| Hymenoptera | 19 | 13 | 16 | 32 | 12 | 17 | 242 |
| Lepidoptera | 48 | 44 | 19 | 53 | 24 | 47 | 452 |
| Mantodea | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| Mesostigmata | 0 | 0 | 1 | 0 | 1 | 0 | 1 |
| Neuroptera | 0 | 0 | 0 | 0 | 0 | 0 | 3 |
| Odonata | 1 | 0 | 1 | 0 | 1 | 1 | 2 |
| Orthoptera | 2 | 2 | 0 | 6 | 1 | 3 | 27 |
| Poduromorpha | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| Psocodea | 4 | 0 | 9 | 2 | 3 | 5 | 15 |
| Thysanoptera | 2 | 0 | 0 | 1 | 0 | 3 | 9 |
| Trichoptera | 0 | 0 | 0 | 0 | 0 | 0 | 9 |

With that in mind, [Table 3](#tbl-top_countries) is a breakdown of the public BINs in common with our project, from the 20 countries which had the most BINs in common with our project:

Table 3: The top 20 countries in terms of sharing BINs with our project

| Ranking | Country | Number of shared BINs |
| --- | --- | --- |
| 1 | South Africa | 995 |
| 2 | Gabon | 551 |
| 3 | Kenya | 402 |
| 4 | Pakistan | 170 |
| 5 | Nigeria | 158 |
| 6 | Egypt | 139 |
| 7 | Ghana | 137 |
| 8 | Australia | 129 |
| 9 | Madagascar | 122 |
| 10 | Bangladesh | 111 |
| 11 | Cameroon | 105 |
| 12 | Saudi Arabia | 92 |
| 13 | United States | 90 |
| 14 | Costa Rica | 86 |
| 15 | India | 84 |
| 16 | Tanzania | 79 |
| 17 | Malaysia | 76 |
| 18 | Ethiopia | 58 |
| 19 | Indonesia | 58 |
| 20 | China | 56 |

It’s encouraging to see that several African countries are in the top 20, but noteworthy that Ghana is at number 7 and there are few representatives of nearby countries. Some of that will be due to varying habitat types, but it highlights just how undersampled West Africa currently is. A lot of these common BINs are also likely to be cosmopolitan/invasive taxa, rather than species only native to Ghana and the nearby area.

# Discussion

# Conclusion

# Supporting information

# Acknowledgements

All authors are members of the Target Malaria Not-for-Profit Research Consortium (www.targetmalaria.org), which is supported by a grant from the Bill & Melinda Gates Foundation and Open Philanthropy.

# References

Ashfaq, Muhammad, Jamal S. M. Sabir, Hosam O. El-Ansary, Kate Perez, Valerie Levesque-Beaudin, Arif M. Khan, Akhtar Rasool, Carlene Gallant, Joseph Addesi, and Paul D. N. Hebert. 2018. “Insect Diversity in the Saharo-Arabian Region: Revealing a Little-Studied Fauna by DNA Barcoding.” Edited by Bernd Schierwater. *PLOS ONE* 13 (7): e0199965. <https://doi.org/10.1371/journal.pone.0199965>.

Blüthgen, Nico, Lynn V. Dicks, Matthew L. Forister, Charlotte L. Outhwaite, and Eleanor M. Slade. 2023. “Insect Declines in the Anthropocene.” *Nature Reviews Earth & Environment*, September. <https://doi.org/10.1038/s43017-023-00478-x>.

Bukowski, Belén, Sujeevan Ratnasingham, Priscila E. Hanisch, Paul D. N. Hebert, Kate Perez, Jeremy deWaard, Pablo L. Tubaro, and Darío A. Lijtmaer. 2022. “DNA Barcodes Reveal Striking Arthropod Diversity and Unveil Seasonal Patterns of Variation in the Southern Atlantic Forest.” Edited by Janice L. Bossart. *PLOS ONE* 17 (4): e0267390. <https://doi.org/10.1371/journal.pone.0267390>.

Chao, Anne, Nicholas J. Gotelli, T. C. Hsieh, Elizabeth L. Sander, K. H. Ma, Robert K. Colwell, and Aaron M. Ellison. 2014. “Rarefaction and Extrapolation with Hill Numbers: A Framework for Sampling and Estimation in Species Diversity Studies.” *Ecological Monographs* 84 (1): 45–67. <https://doi.org/10.1890/13-0133.1>.

D’Souza, Michelle L., and Paul D. N. Hebert. 2018. “Stable Baselines of Temporal Turnover Underlie High Beta Diversity in Tropical Arthropod Communities.” *Molecular Ecology* 27 (10): 2447–60. <https://doi.org/10.1111/mec.14693>.

deWaard, Jeremy R., Valerie Levesque-Beaudin, Stephanie L. deWaard, Natalia V. Ivanova, Jaclyn T. A. McKeown, Renee Miskie, Suresh Naik, et al. 2019. “Expedited Assessment of Terrestrial Arthropod Diversity by Coupling Malaise Traps with DNA Barcoding.” Edited by John James Wilson. *Genome* 62 (3): 85–95. <https://doi.org/10.1139/gen-2018-0093>.

Folmer, O., M. Black, W. Hoeh, R. Lutz, and R. Vrijenhoek. 1994. “DNA Primers for Amplification of Mitochondrial Cytochrome c Oxidase Subunit 1 from Diverse Metazoan Invertebrates.” *Molecular Marine Biology and Biotechnology*, no. 5. <http://www.vliz.be/en/imis?refid=64543>.

Gotelli, Nicholas J. 2004. “A Taxonomic Wishlist for Community Ecology.” Edited by H. C. J. Godfray and S. Knapp. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* 359 (1444): 585–97. <https://doi.org/10.1098/rstb.2003.1443>.

Meier, Rudolf, Bonnie B. Blaimer, Eliana Buenaventura, Emily Hartop, Thomas Von Rintelen, Amrita Srivathsan, and Darren Yeo. 2022. “A Re-Analysis of the Data in Sharkey Et Al.’s (2021) Minimalist Revision Reveals That BINs Do Not Deserve Names, but BOLD Systems Needs a Stronger Commitment to Open Science.” *Cladistics* 38 (2): 264–75. <https://doi.org/10.1111/cla.12489>.

Oksanen, Jari, Gavin L. Simpson, F. Guillaume Blanchet, Roeland Kindt, Pierre Legendre, Peter R. Minchin, R. B. O’Hara, et al. 2022. *Vegan: Community Ecology Package*. Manual. <https://CRAN.R-project.org/package=vegan>.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Manual. Vienna, Austria: R Foundation for Statistical Computing. <https://www.R-project.org/>.

Ratnasingham, Sujeevan, and Paul D. N. Hebert. 2007. “Bold: The Barcode of Life Data System (http://www.barcodinglife.org).” *Molecular Ecology Notes* 7 (3): 355–64. <https://doi.org/10.1111/j.1471-8286.2007.01678.x>.

———. 2013. “A DNA-Based Registry for All Animal Species: The Barcode Index Number (BIN) System.” *PLOS ONE* 8 (7): e66213. <https://doi.org/10.1371/journal.pone.0066213>.

Rosenberg, Yuval, Yinon M. Bar-On, Amir Fromm, Meital Ostikar, Aviv Shoshany, Omer Giz, and Ron Milo. 2023. “The Global Biomass and Number of Terrestrial Arthropods.” *Science Advances* 9 (5): eabq4049. <https://doi.org/10.1126/sciadv.abq4049>.

Sharkey, Michael J., Daniel H. Janzen, Winnie Hallwachs, Eric G. Chapman, M. Alex Smith, Tanya Dapkey, Allison Brown, et al. 2021. “Minimalist Revision and Description of 403 New Species in 11 Subfamilies of Costa Rican Braconid Parasitoid Wasps, Including Host Records for 219 Species.” *ZooKeys* 1013 (February): 1–665. <https://doi.org/10.3897/zookeys.1013.55600>.

Stork, Nigel E. 2018. “How Many Species of Insects and Other Terrestrial Arthropods Are There on Earth?” *Annual Review of Entomology* 63 (1): 31–45. <https://doi.org/10.1146/annurev-ento-020117-043348>.

Wagner, David L., Eliza M. Grames, Matthew L. Forister, May R. Berenbaum, and David Stopak. 2021. “Insect Decline in the Anthropocene: Death by a Thousand Cuts.” *Proceedings of the National Academy of Sciences* 118 (2): e2023989118. <https://doi.org/10.1073/pnas.2023989118>.

Williamson, Joseph, Enoch Teh, Tommaso Jucker, Matilda Brindle, Emma Bush, Arthur Y. C. Chung, Jonathan Parrett, Owen T. Lewis, Stephen J. Rossiter, and Eleanor M. Slade. 2022. “Local-Scale Temperature Gradients Driven by Human Disturbance Shape the Physiological and Morphological Traits of Dung Beetle Communities in a Bornean Oil Palmforest Mosaic.” *Functional Ecology*, May, 1365–2435.14062. <https://doi.org/10.1111/1365-2435.14062>.