DNA barcodes provide insights into the diversity and biogeography of the non-biting midge Polypedilum (Diptera, Chironomidae) in South America.

Fabio Laurindo Silva[1], Luiz Pinho[2], Elisabeth Stur[3], Silvio Nihei[1], and Torbjorn Ekrem[3]

[1]Universidade de São Paulo Instituto de Biociências
[2]Universidade Federal de Santa Catarina - Campus Florianópolis
[3]Norwegian University of Science and Technology

February 9, 2023

Abstract

Aim The Neotropics, particularly South America, holds unparalleled high levels of species richness, when compared to other major biomes. Some neotropical areas are hotspots of a fragmentary known diversity of insects and are under manifest danger of biodiversity loss and climate change. Therefore, prompt estimates methods of its diversity are urgently required to complement slower traditional taxonomic approaches. Despite a variety of algorithms for delimiting species through single-locus DNA barcodes having been developed and applied for rapid estimates of species diversity in a wide array of taxa; however, tree-based and distance-based methods may lead to different group assignments, either overestimating or underestimating the number of putative species. Here, we investigate the performance of different DNA-based species delimitation approaches for a rapid biodiversity estimate of the diversity of Polypedilum (Chironomidae, Diptera) in South America. Location Worldwide Methods We analyze a mtDNA dataset comprising 1,492 specimens from 598 locations worldwide. Molecular operational taxonomic units (MOTUs) ranged from 267 to 520, based on the Barcode Index Number (BIN), Bayesian Poisson tree processes (bPTP), multi-rate Poisson tree processes (mPTP), single-rate Poisson tree processes (sPTP), and generalized mixed Yule coalescent (sGMYC) approaches. Results Our results highlight Polypedilum as a species-rich genus, yet incompletely documented, and found the sGMYC method to be the most adequate to estimate putative species in our dataset. Furthermore, based on these data, we describe the distribution of diversity and some biogeographical patterns of Polypedilum. Main Conclusions Findings imply the genus exhibited high levels of endemism and richness of species in the Neotropics, which confirmed our hypothesis that there are substantial differences in community structure between the Polypedilum fauna in South America and the neighboring regions.

DNA barcodes provide insights into the diversity and biogeography of the non-biting midge *Polypedilum* (Diptera, Chironomidae) in South America

Fabio Laurindo da Silva*

Department of Natural History, NTNU University Museum, Norwegian University of Science and Technology, Trondheim, Norway

Luiz Carlos Pinho

Laboratory of Systematics of Diptera, Department of Ecology and Zoology, Federal University of Santa Catarina, Florianópolis, Brazil.

Elisabeth Stur

Department of Natural History, NTNU University Museum, Norwegian University of Science and Technology, Trondheim, Norway

Silvio Shigueo Nihei

Laboratory of Systematic and Biogeography of Insecta, Department of Zoology, Institute of Biosciences, University of São Paulo, São Paulo, Brazil.

Torbjørn Ekrem

Department of Natural History, NTNU University Museum, Norwegian University of Science and Technology, Trondheim, Norway

*Corresponding author. E-mail:*fabiolaurindo@usp.br*

Acknowledgements

We are grateful to several colleagues and associates who contributed specimens, both South American and critical northern hemisphere taxa, towards this study:

- Brian Farrell, Harvard University, United Stated
- Anthony Eugene Kiszewski, Bentley University, United Stated
- Neusa Hamada, National Institute of Amazonian Research, Brazil
- Ana Maria Oliveira Pes, National Institute of Amazonian Research, Brazil
- Galileu Petronilo da Silva Dantas, National Institute of Amazonian Research, Brazil
- Susana Trivinho-Strixino, Federal University of São Carlos, Brazil
- Gilberto Gonçalves Rodrigues, Federal University of Pernambuco, Brazil

We thank all Argentinian, Brazilian, Chilean and Dominican Republic state and territory governments, and their relevant environment officers, park rangers and private owners who approved collecting permits and enabled access that allowed us to collect in protected areas. Our deepest gratitude goes out to Everton Santos Dias and Rogério Campos de Oliveira for their help and support during various fieldwork trips. We are also very grateful to Amanda Carvalho de Andrade and Paloma Helena Fernandes Shimabukuro for useful comments and valuable suggestions during the initial phase of this study. Special thanks go to Mark Miller and the CIPRES Science Gateway. The data reported in this paper were partly obtained thanks to the generosity of George Putnam, through a Putnam expedition Grant from the Museum of Comparative Zoology at Harvard University. F. L. da Silva was supported by fellowships of the São Paulo Research Foundation (FAPESP - 2016/07039–8, 2018/01507–5, 2021/08464–2, 2019/25567–0), which allowed the preparation of the present manuscript.

Conflict of Interest

The authors declare no conflict of interest.

Author contributions

F.L.S. conceived the study, performed the analyses, drafted and reviewed the manuscript; L.C.P contributed in data collation, taxonomical identification and reviewed the manuscript; E.S and T.E. conceived the study, drafted and reviewed the manuscript; S.S.N reviewed the manuscript

Aim

The Neotropics, particularly South America, holds unparalleled high levels of species richness, when compared to other major biomes. Some neotropical areas are hotspots of a fragmentary known diversity of insects and are under manifest danger of biodiversity loss and climate change. Therefore, prompt estimates methods of its diversity are urgently required to complement slower traditional taxonomic approaches. Despite a variety of algorithms for delimiting species through single-locus DNA barcodes having been developed and applied for rapid estimates of species diversity in a wide array of taxa; however, tree-based and distance-based methods may lead to different group assignments, either overestimating or underestimating the number of putative

species. Here, we investigate the performance of different DNA-based species delimitation approaches for a rapid biodiversity estimate of the diversity of *Polypedilum* (Chironomidae, Diptera) in South America.

Location

Worldwide

Methods

We analyze a mtDNA dataset comprising 1,492 specimens from 598 locations worldwide. Molecular operational taxonomic units (MOTUs) ranged from 267 to 520, based on the Barcode Index Number (BIN), Bayesian Poisson tree processes (bPTP), multi-rate Poisson tree processes (mPTP), single-rate Poisson tree processes (sPTP), and generalized mixed Yule coalescent (sGMYC) approaches.

Results

Our results highlight *Polypedilum* as a species-rich genus, yet incompletely documented, and found the sGMYC method to be the most adequate to estimate putative species in our dataset. Furthermore, based on these data, we describe the distribution of diversity and some biogeographical patterns of *Polypedilum* .

Main Conclusions

Findings imply the genus exhibited high levels of endemism and richness of species in the Neotropics, which confirmed our hypothesis that there are substantial differences in community structure between the*Polypedilum* fauna in South America and the neighboring regions.

Keywords : DNA barcode, GMYC, PTP, Chironomidae, Biogeography

1 Introduction

In the past decades, natural environments have been disturbed and destroyed worldwide at alarming rates, which results in a large loss of species (Barnosky et al., 2011; Stork, 2018). In hyperdiverse ecosystems, such as those in the Neotropical region, several species could go extinct before even being identified (Bradshaw et al., 2011; Laurance, 1999). This indicates that biodiversity evaluation needs to be accelerated by combining the strengths of molecular biology, sequencing technology, and bioinformatics to recognize previously known and described species (Gostel & Kress, 2022), and to allow new findings. In this context, DNA-based approaches have become increasingly useful and promising tools for estimating diversity and guaranteeing rapid and accurate identification of species. Since the proposal of the DNA barcoding technique, using a short standard genetic marker for species-level identification and cryptic species detection (Hebert et al., 2003; Hebert et al., 2004), the procedure has been becoming progressively popular among conservationists and taxonomists (Farooq et al., 2020; Pearlet al., 2022), and paved the way for biological monitoring using metabarcoding (e.g., Steinke et al., 2022).

Aquatic insects play a crucial role for the equilibrium of aquatic ecosystems because of their complex life cycle, which distinguishes them from exclusively aquatic or terrestrial life forms, and generates a differentiated potential for understanding biogeographical and ecological research. It is therefore paramount to invest in knowledge of the diversity of these organisms, as they are extremely rich both in functionality and species numbers. Non-biting midges (Diptera: Chironomidae) are true flies, and frequently dominate aquatic insect communities in both abundance and species richness. It is a cosmopolitan group, occurring in an enormous variety of aquatic ecosystems, in all biogeographical regions of the world, including Antarctica. Presumably, the great species and habitat diversity in this family is a product of its antiquity, relatively low vagility and evolutionary plasticity (Ferrington, 2008), which makes the family not only a valuable source of indicator species for lentic and lotic aquatic ecosystems, but also one of the most interesting groups for phylogenetic and biogeographical analyses (Silva & Ekrem, 2016).

The advantage of dealing with hyperdiverse taxa is that they surely exhibit several repeated patterns, which may provide evidence of underlying processes (Coscaron et al., 2009). Therefore, a genus such as*Polypedilum* , widespread and rich in species, may be suitable for biogeographical and ecological research, as the diversity

of these insects is strongly linked with the conservation of aquatic habitats.*Polypedilum* is one of the largest chironomid genera containing about 440 described species (Sæther et al., 2010). Larvae of*Polypedilum* occur in nearly all types of still and flowing waters. Knowledge of their community compositions is essential due to their potential as bioindicators, since natural or man-made shifts have impact on them, and consequently in ecosystem processes. However, whilst most studies rather focus on taxonomic or phylogenetic issues (e.g., Bidawid & Fittkau, 1995; Bidawid-Kafka, 1996; Sæther & Sundal, 1999; Vårdal et al., 2002; Sæther & Oyewo, 2008; Oyewo & Sæther, 2008; Shimabakuro, et al., 2019; Pinho & Silva, 2020), so far there have been only few detailed studies on the species richness and species turnover of the hyperdiverse Chironomidae (Lin et al., 2015; Song et al., 2018).

The biota of South America always has attracted the attention of naturalists because of the interesting distributional patterns exhibited by its flora and fauna (Hooker, 1844-47; Darwin, 1859; Wallace, 1876). For more than a century, biogeographers have proposed theories to explain the origin and relationships of the biodiversity found in South America and other southern temperate regions such as Australia, New Zealand and South Africa (Silva & Farrell, 2017). Moreover, the region is a preferred target for investigating the function of these components in the dynamic of diversification, both by harboring the majority of the Earth's species and extending across temperate and tropical belts. The high number of species in South America, on a regional as well as on a continental scale, makes the region an important reference mark for estimation of biodiversity loss. However, for the Neotropical non-biting midge fauna, the knowledge of the actual species diversity is fragmentary and formal identifications often are unachievable (Spies & Reiss, 1996).

Usually, automated species delimitation approaches are considered particularly useful in organisms with uncertain species boundaries, due to fragmentary taxonomic knowledge or signals in phylogenetic inferences being obscured by lineage sorting or introgression (O'Meara, 2010 and references therein). In this sense, several methods for species delimitation have been developed and applied, for instance, the Automatic Barcode Gap Discovery – ABGD (Puillandre et al., 2012), the Barcode Index Number – BIN (Ratnasingham & Hebert, 2013), the Generalized Mixed Yule Coalescent – GMYC (Pons et al., 2006), the Poisson Tree Processes – PTP (Zhang et al., 2013). Despite these approaches being suitable to delimit species, they can occasionally lead to uncertainty in genetic diversity estimates due to either oversplitting or overlumping of the taxa. Therefore, the integration of different algorithms is needed for accurate species delimitation. In this study, we first compare the performance of different methods of species estimation and evaluate how much these different approaches affect estimates of putative species richness in South America. We then test the hypothesis that there will be substantial differences in community structure between the *Polypedilum* fauna in South America, considered more diverse, and neighboring regions, particularly the Nearctic.

2 Material and methods

2.1 Taxon sampling and data collection

Specimens were collected between 2014 and 2017 from 34 localities, in a diverse array of habitats including small streams and ponds to lakes, rivers and bays in Argentina, Brazil, Chile and Dominican Republic. The main emphasis was on adult sampling, collected with a sweep net near aquatic systems. A 20 cm diameter D-frame kick net (mesh size 250 μm) was also used to collect immature stages at some localities. All sampled adults were preserved in 75%-85% ethanol and larvae in 96-100% ethanol and stored at 4°C in the dark prior to the extraction. Specimens were identified using the classification proposed by Townes (1945), Bidawid & Fittkau (1995), Bidawid-Kafka (1996), Shimabakuro et al. (2019), Pinho & Silva (2020), and eventual examination of type material. Voucher specimens are deposited in the Museum of Comparative Zoology (MCZ) at Harvard University and in the National Institute of Amazonian Research (INPA).

In addition to data generated for this publication, we also searched for public COI barcodes in the Barcode of Life Data Systems (BOLD,*www.boldsystems.org*) belonging to the genus *Polypedilum* that were longer than 300 base pairs and without stop codons. Searches were performed on 25 January 2022 in BOLD. In total, 9,540 COI barcodes were included in our dataset, of which 149 barcodes of 54 identified species

were not previously used in any molecular analysis. A reduced data set, containing 1,492 sequences, was generated based on the manual deletion of the highly similar sequences based on an UPGMA tree. Duplicate sequences occurring at different sampling localities were retained in our dataset. The detailed specimen records and sequence information, including trace files, are available in BOLD through the dataset 'DS-RPPPOL - Reduced personal and publicly available records of *Polypedilum* (Diptera: Chironomidae)" with DOI: https://doi.org/XXXX.

2.1 DNA extraction, PCR amplification, sequencing and alignment

The targeted taxa were sorted and dissected under a stereo microscope. Thorax and one pair of legs were used for genomic DNA extraction. All extraction procedures followed the Qiagen DNeasy Blood and Tissue kit protocol provided by the manufacturer. DNA was extracted from the thorax and head in a buffered solution with the enzyme proteinase-K at 56 °C overnight, and otherwise followed the manufacturer's protocol, except using a final elution volume of 100 µl. After digestion, the exoskeletons were removed carefully using a fine-tipped forceps and washed with 96% ethanol before mounting in Euparal on the same microscope slide as its corresponding head, antennae, wings, legs and abdomen following the procedure outlined by Sæther (1969).

A 658 bp fragment of the COI region was PCR-amplified in 25 µL reactions and containing 2 µL DNA template (concentration not measured), 2.5 µL 5X buffer, 2 µL $MgCl_2$ in 25 µM concentration, 0.2 µL of dNTPs in 10 mM concentration, 1 µL of each of the universal standard barcode primers (Folmer et al., 1994) LCO1490 (50-GGTCAACAAATCATAAAGATATTGG-30) and HCO2198 (50-TAAACTTCAGGGTGACCAAAAAATCA-30), in 10 µM concentration, 0.2 µL of HotStarTaq (Qiagen, Germany) and 16.1 µL of ddH2O. PCR amplification was performed in a thermocycler with an initial denaturation step of 95 ºC for 15 minutes, then followed by five cycles of 94 ºC for 30 seconds, 45 ºC for 30 seconds, 72 ºC for 1 minute, followed by 35 cycles of 94 ºC for 30 seconds, 51 ºC for 30 seconds, 72 ºC for 1 minute, and one cycle at 72 ºC for 5 minutes, then held at 4 ºC.

PCR products were checked visually by electrophoresis on a 1.5% agarose gel and purified using shrimp alkaline phosphatase and exonuclease I (USB Corp., USA). For bidirectional sequencing, we used the ABI PRISM BigDye Terminator version 3.1 Cycle Sequencing Kit (Life Technologies, USA), and cycle sequencing reactions were performed on ABI PRISM 3130xl or 3730xl automated sequencers (Life Technologies, USA) at Harvard University, or shipped to EurofinsGenomics (Ebersberg, Germany). Raw sequences were assembled and edited using Geneious 2021.2.2 (Kearse et al. , 2012), checked for stop codons and aligned as translated amino acids using the MUSCLE algorithm (Edgar, 2004) on amino acids as implemented in MEGA11 (Tamura et al. , 2021). The nucleotide compositions were calculated in MEGA11, while the pairwise genetic distances for each individual sequence were determined in BOLD, both using the K2P model (Kimura, 1980).

2.3 Phylogenetic analysis

Two phylogenetic trees were generated: a non-ultrametric phylogram using Maximum Likelihood (ML) (Felsenstein, 1981) and an ultrametric chronogram using Bayesian inference (BI) (Drummond et al., 2002). The ML tree was generated using Iq-Tree (Trifinopoulos et al., 2016). Node support was assessed with 1000 ultrafast bootstrap replicates (Hoang et al., 2018), using the GTR+G+I model, with the data partitioned according to codon position, as recommended by PartitionFinder version 2.1. (Lanfear et al., 2017).

The BI tree was generated using BEAST version 2.6.7 (Bouckaert et al., 2019), using default settings for all parameters. XML files were made with the BEAUti version 2.6.7 interface with the following settings: GTR+G+I substitution model, empirical base frequencies, 4 gamma categories and all codon positions partitioned with unlinked base frequencies and substitution rates. Since there is no agreement concerning the most appropriate clock and tree priors for reconstructing gene trees for species delimitation (Monaghan et al., 2009; Ratnasingham & Hebert 2013; Talavera et al., 2013; Tang et al., 2014), preliminary analyses to compare the use of two different clock (strict and relaxed lognormal) and two different tree priors (coalescent constant population and Yule) were undertaken (Rodrigues et al., 2020). The results of these exploratory analyses (data not shown) indicated the strict clock and Yule priors as the most suitable for our data set, thus these priors were used for the Bayesian inference analyses.

To account for mixing within chains and convergence among chains with reversible jump MCMC (Elworth et al., 2018), a total of 10 chains were run from different seeds for 100 million generations each. Log files from each run were combined in LogCombiner version 2.6.7 (Drummond et al., 2012) after removal of the first 10% of samples from each run as burn-in. Convergence of each run and the combined data were checked for proper mixing using effective sample size (ESS) > 200 in Tracer version 1.7.1 (Rambaut et al., 2018). Tree files from each run were resampled to retain only 10% of the total trees and combined using LogCombiner after removal of the first 10% of retained trees from each run as burn-in. A maximum clade credibility (MCC) tree was then produced using TreeAnnotator version 2.6.6 (Drummond et al., 2012) and FigTree version 1.4.4 (Rambaut, 2010) was used to visualize and edit the trees.

All phylogenetic analyses were conducted on the CIPRES Science Gateway High Performance Computing platform (Miller et al., 2011).

2.4 Putative species estimation

The two single-locus DNA barcoding methods consisted of two fundamentally different approaches. First, we implemented three distance-based approaches: (1) the Automatic Barcode Gap Discovery – ABGD (Puillandre et al., 2012) performed using the online version of the software (*https://bioinfo.mnhn.fr/abi/public/abgd/abgdweb.html*); (2) the Assemble Species by Automatic Partitioning – ASAP (Puillandre et al., 2021), an updated implementation of the ABGD hierarchical clustering algorithm, performed with the ASAP web version (*https://bioinfo.mnhn.fr/ abi/public/asap/*). Both methods used the MUSCLE aligned matrix as the input file and adopted the Kimura model, following the default settings for all parameters. (3) The Barcode Index Number (BIN), a method implemented in BOLD, in which newly submitted and already available sequences clustered in unique BINs using a refined single linkage analysis in which records with high sequence similarity and connectivity are clustered and separated from those with lower similarity and sparse connectivity (Ratnasingham & Hebert, 2013).

Second, we applied four tree-based approaches, which models speciation along the branches of an inferred phylogenetic: (1) the Single Poisson Tree Processes – sPTP (Zhang et al., 2013), implemented using the PTP online version (*http://species.h-its.org/ptp*); (2) the Bayesian Poisson tree process – bPTP (Zhang et al., 2013), also conducted on the PTP web-server. Both analyses were conducted with 500,000 MCMC generations and other parameters as default. (3) the Multi-rate Poisson Tree Processes – mPTP (Kapli et al., 2017), performed with the mPTP web-server (*https://mcmc-mptp.h-its.org/mcmc/*), using the multi-rate Poisson tree process model and following default settings. All PTP analyses (sPTP, mPTP and bPTP) used the ML trees calculated with Iq-Tree. (4) the Generalized Mixed Yule Coalescent – GMYC (Pons et al., 2006), performed by submitting the single ultrametric MCC tree resulting from BI obtained from BEAST to the online version of the GMYC software (*https://species.h-its.org/gmyc/*), following default single-threshold (sGMYC). We also tested the multiple-threshold model (mGMYC); however, it did not perform well, overestimating putative species (data not shown). Similar species estimates with the mGMYC algorithm were also observed in previous studies (Fujisawa & Barraclough, 2013; Schwarzfeld & Sperling, 2015).

2.5 Biogeographical analysis

The biogeographical relationships between samples were implemented with the software PRIMER 7 (Plymouth Routines In Multivariate Ecological Research). In the Neotropical region, we pre-assigned a set of smaller hierarchical geographical areas (zones), following Morrone et al. (2022), to enable hypothesis testing using non-metric multidimensional scaling (nMDS). The faunal similarity between zones and between larger regions was quantitatively measured using Sørensen similarity index of presence/absence data, and the significance of the geographical groupings was assessed using the ANOSIM test (see Appendices S1 & S2). Species accumulation curves were implemented by the vegan package (Oksanen et al., 2017) in the software R (R Development Core Team, 2021) using 100 randomizations. Accumulation curves are randomized plots of the presence or absence of each species (here delimited by the sGMYC algorithm) against the number of observations.

3 Results

3.1 Species delimitation

The complete data set consisted of 1,492 barcodes, ranging from 312 to 658 bp in length. In total, there were 319 variable sites (48.5%), of which 299 (93.7%) were parsimony informative. Most parsimony informative sites occurred in the third codon-position (Table 1). The sequences were heavily AT-biased, specifically in the third position, which exhibited a combined average AT-composition of 89.3% (Table 1). Average intraspecific and interspecific K2P-distances for all analyzed*Polypedilum* species were 1.3% and 15.2%, respectively. The barcode gap is an important concept in barcoding studies (Puillandre et al., 2012). It works well when the amount of intraspecific divergence is much smaller than the amount of interspecific variation between species. When this condition is met, a 'barcoding gap' exists (Meyer & Paulay, 2005). In general, our data showed clearly larger interspecific than intraspecific divergences, but we still could not observe the expected 'barcoding gap' in the pairwise K2P distances. On the contrary, a barcode overlap between the intraspecific and the interspecific distances was found, which may be attributable to the presence of cryptic species diversity and a few misidentifications. The lack of a gap is usually associated with recently diverged species with little genetic diversification, frequently coupled with incomplete lineage sorting and introgression (Wiemers & Fiedler, 2007; Dupuis et al., 2012).

Overall, most of the tested methods recovered similar groupings of molecular operational taxonomic units (MOTUs) (Figures 1-4), with the mPTP method being the most conservative, lumping the sequences into fewer MOTUs, and the bPTP algorithm the most relaxed, lumping the sequences into several MOTUs (Table 2). Two out of the three distance-based methods, ABGD and ASAP, yield unreliable delimitations with wide confidence intervals, with several clusters not reflecting relationships as understood based on the geographical sampling localities and others diverging into numerous lineages despite diminished divergence between them. ABGD and ASAP results were not included in the Figures 1-4. The BIN analysis returned a total of 415 MOTUs of which 174 were singleton BINs, 222 concordant BINs, and 19 discordant BINs. In total, 615 sequences of 143 morphospecies were assigned to 179 BINs, including 72 singleton BINs, 519 concordant BINs, and 24 discordant BINs. The unidentified 877 specimens, without binomial names, were assigned to 236 BIN-species, including 102 singleton BINs, 118 concordant BINs, and 16 discordant BINs.

DNA-based species delimitation applying bPTP, mPTP, sPTP, and sGMYC resulted in divergent number of clusters. The single-threshold general mixed Yule-coalescent calculations (sGMYC) recovered 370 MOTUs, while the sPTP model produced a more conservative number of MOTUs (411) compared to the bPTP method, which yielded 520 MOTUs (Table 2). The results from analyses using the multi-rate PTP (mPTP) model were also comparable to those of the other models, but revealed larger clusters, occasionally joining lineages belonging to different species in a single MOTU (Figure 1). Divergences in the number of clusters generated by the different species delimitation algorithms are caused by erroneously inferred splitting or lumping events (i.e., specimens of one morphospecies were divided or joined into two or more different MOTUs). However, regardless of the method applied, the total number of species delimited in *Polypedilum* in this study is at least twice as high (267–520) as the number of included morphospecies (143, see above).

3.2 Biogeography

The following biogeographical analyses were based on the MOTUs (370) delimited by the sGMYC approach (see discussion below). In our dataset, the number of species per sampling location varied from a single species to over 15 species across the different sampling sites at the studied biogeographical realms. Just a few of all sampled locations (3.8%) had 5 or more species present, while 203 (54.9%) of the 370 species included were only recorded at a single location. Since collecting methods, sampling sites, protocols and reporting varied in our dataset, comparisons of overall biodiversity between locations is challenging. Numbers of *Polypedilum* species and sampled locations varied between regions (Table 1). Our results indicate that 90.2 % of species were recorded only in a single major biogeographical region, while only 36 species spanned two or more of these regions.

The relative diversity and dominance of *Polypedilum* species as a proportion of the total number of species per

region shows a clear divergence between the geographical regions studied (Figure 5). Insufficiently sampled areas (Afrotropical, Australasian and Panamanian) with low numbers of recorded species present low levels of diversity and are dominated by few species. On the other hand, biogeographical regions exhaustively sampled exhibit high numbers of recorded species with the highest degree of species richness (Nearctic, Palearctic, Oriental and Sino-Japanese). Regarding the Neotropics, the region presented moderate levels of *Polypedilum* species diversity, particularly when compared to the neighboring Nearctic region (Figure 5); however, it is noteworthy that although only 6.2% of the sampling sites are located in the Neotropical region (mostly in South America), 19.1% of the total number of species occurred in this region. Moreover, based on our results, the Neotropical *Polypedilum* fauna can be considered endemic, since only one unidentified species was also recorded in the Nearctic region.

None of the species accumulation (rarefaction) curves for the biogeographical realms (Figure 6) exhibit asymptote for any area, although the Nearctic sequences may be approaching one. The Afrotropical, Australasian and Panamanian regions presented the lowest levels of diversity. The highest levels of diversity were seen in the Nearctic, Palearctic and Sino-Japanese regions, with the Neotropical and Oriental regions curves being noticeably lower, with levels of diversity which seems to be comparable. Biogeographical realm patterns across the entire assembly (Figure 7a) showed distinct groupings for Afrotropical and Australasian, while the ANOSIM (see Appendix S1) and nMDS results show some overlap between Nearctic and Palearctic regions. The Neotropical *Polypedilum* fauna despite the closeness to the Nearctic region presents distinct clustering. The different Neotropical zones compose distinct well supported groups (Figure 7b), with some degree of overlap between Southeastern Amazonia and Boreal Brazilian domination zones. In particular, the Palearctic region appears to display affinities for both the Nearctic and Oriental region (Figure 7b).

4 Discussion

4.1 Species delimitation

One of the objectives of this study was to explore the utility of a large-scale single-locus DNA barcode analysis of the genus*Polypedilum* to investigate its molecular diversity and compare the adequacy of molecular species delimitation approaches. Our results suggest that tree-based algorithms are more suitable than distanced-based because they are able to integrate evolutionary theory, not requiring arbitrary thresholds (Schwarzfeld & Sperling, 2015). In our study, ABGD and ASAP produced unreasonable delimitations, not consistently proposing species hypotheses. These approaches are known to over-lump, performing poorly on more speciose datasets such as ours, whereas the success rate increases remarkably for small populations (Dellicour & Flot, 2015; 2018). In contrast to ABGD and ASAP's over-lumping, the Barcode Index Number (BINs) method, assigned by BOLD, is known to oversplit species numbers due to the low intracluster distance (2.2%) at the initial clustering step of RESL algorithm (Ratnasingham & Hebert, 2013). Similar results were found by Song et al. (2018), when applying the BIN system also to delimit*Polypedilum* species, mostly from East Asia.

Among the drawbacks of distance-based methods is the lack of a universal threshold that fits all taxa (Yang & Rannala, 2017). Several DNA barcoding studies try to determine a fixed threshold value, Hebert et al. (2004) suggested the interspecific divergences at least 10 times as large as the intraspecific divergence the so-called "10 × rule,". However, it seems that different best-fit thresholds apply to different taxonomic groups (Havermans, et al., 2011). For example, a threshold of 2-3% was indicated for some for Ephemeroptera, Plecoptera and Trichoptera (Zhou et al., 2010), and 3-5% for some dipteran species groups (Lin et al., 2015; Nzelu et al., 2015), while a threshold 5-8% for species in *Polypedilum* was suggested by Song et al. (2018). Another downside of distance-based approaches is that they do not consider evolutionary relationships into their algorithms (Kapli et al., 2017). Tree-based methods are not influenced by such thresholds, since they use phylogenetic inference for a more precise barcode assignment (Song et al., 2018).

Applied to our dataset, sGMYC and PTP tended to over-perform when compared to delineations made with distance-based methods and the morphological species concept. The Poisson Tree Process (PTP) relies on the distribution of branch lengths in the gene tree in order to identify species status (Zhang et al., 2013).

The tree and branch lengths are inferred from a sequence alignment using maximum likelihood and then treated as lacking errors (Ranala & Yang, 2020). In our study, there was a large difference between recovered MOTUs among the PTP methods. There was a 109 MOTU difference between results based on the bPTP and sPTP methods. mPTP was the most conservative and commonly underestimated species by lumping singleton species, represented in our tree by isolated branches, into MOTUs. Along with our results, other studies have found that the mPTP algorithm leads to a lower number of recovered species when compared with other approaches (e.g., da Silva et al. 2018, Parslow et al. 2021).

The sGMYC analysis based on a single gene revealed the presence of 370 MOTUs (likelihood ratio: 600.4823, confidence interval: 349-383, threshold time: -0.01053644). This species-delimitation algorithm relies on the priors and parameters used to construct the ultrametric tree (Ceccarelli et al., 2012), and tends to overestimate species diversity compared to other methods (Paz & Crawford, 2012; Miralles & Vences, 2013; Talavera et al., 2013; Kekkonen & Hebert, 2014). In our study, the sGMYC method seems to be the most accurate since it recovered substantially fewer putative species than the bPTP and sPTP analyses despite its hypothesized oversplitting. Moreover, the sGMYC approach has been suggested to suit datasets with large numbers of singleton taxa (Talavera et al., 2013), which is what we observe for *Polypedilum* . Based on the aforementioned considerations, we chose the putative species delimited by the sGMYC method as the basis for the biogeographical analyses.

4.2 Biogeography

The level of taxonomic diversity present in an environment can be quantified by either enumerating numbers of species (e.g. Simpson's diversity) or estimating evolutionary divergences among species in which genetic divergences have been calculated (Webb, 2000). Moreover, besides the number of individuals sampled, the size of the local species pool, the evenness of species abundances in the community, size and environmental heterogeneity of the area, and the status of taxonomic understanding of the taxa investigated are parameters essential to the accuracy of estimates of taxonomic diversity (Antonelli et al., 2018). Although most measures of alpha and beta diversity rely on species numbers, DNA sequence data may provide an evolutionary framework to diversity estimates (Hebert et al., 2016). In this sense, genetic measures may also be used to evaluate species boundaries when compared with species richness in the same communities. Additionally, DNA barcodes can be used for species delimitation, assisting in documenting new species, and identifying targeted habitats for conservation (Faith, 1992; 2008). In geographic regions especially known for their unique lineages of organisms, biological diversity determined with DNA barcode sequence data can be essential for comparing diversity and establishing protected areas across the landscape (Shapcott et al., 2015, Hobner, 2021).

Numerical species delimitation methods require species to be sufficiently sampled (Dopheide et al., 2019) across geographical ranges to improve their ability to correctly delimit species (Parslow et al., 2021). In practice, this is a challenging task when it comes to *Polypedilum* due to its known worldwide diversity of ca. 440 described species and the expected number of undescribed species. Although recent taxonomic studies of regional fauna have been conducted (Song et al., 2016; 2018), particularly in East Asia, there are several regions that need modern taxonomic treatments, for example, Australia, Africa and South America. Therefore, it is difficult to determine the degree of sampling completeness of *Polypedilum* caused by the potentially large number of undescribed species. In the current study, many of the biogeographical differences in recorded species numbers can be ascribed to different sampling efforts and methods between regions. Usually, knowledge of species distributions and diversity patterns are strongly concentrated toward areas which are more easily accessible by roads, rivers, and research stations (Antonelli et al., 2018). This fact is evident in our investigation, as though we included all publicly available COI sequences for *Polypedilum* in BOLD, there was a bias towards Nearctic (33.2%) and Sino-Japanese (23.2%) taxa, with a reduced representation of Afrotropical (1.6%), Australasian (1.6%) and Panamanian (1.3%) species, regions known for receive less investment for research in Chironomidae.

Much of what we need to comprehend about biodiversity can be undertaken as a matrix of the presence or abundance of multiple species across time and space (Hobner, 2021). That said, plotting species accumulation

curves permit researchers to measure and compare diversity across populations or to assess the benefits of further sampling (Deng et al., 2015). In our study, the rarefaction curve analysis suggests that even when randomization sampling methods are considered there are regional differences in species richness. Noticeably, the most species-rich regions were the Nearctic and Sino-Japanese regions. This came as a little surprise, since we expected the Palearctic region also to be among the most specious biogeographical areas, due to the high number of*Polypedilum* sequences available in BOLD and the numerous studies performed on the family Chironomidae in this area. Although we used species accumulation curves to indicate the pattern of sequence accumulation within the current study, they are not expected to represent the accurate diversity of each region, as they are not based on actual random sampling (Schwarzfeld & Sperling 2015).

The Palearctic fauna overlaps partially with that of the Nearctic, especially in the north. This is similar to what is found in other studies (Ekrem et al., 2018; Marusik & Koponen, 2005) where distinct communities in the two regions share several species. This can be the result of numerous faunal interchanges that took place across the Bering land bridge (135 000 – 70 000 YBP). However, these migrations were mostly limited to large, cold-tolerant species (Rodriguez et al., 2006), and it is mainly these species which are found throughout the Holarctic realm today. Chironomids have also been observed as aerial plankton (Hardy & Milne, 1938; Gressitt et al., 1960; Cotoras & Zumbad, 2020) and one cannot rule out long distance dispersal as an explanation for trans-Atlantic distribution patterns in *Polypedilum* (Ekrem et al., 2018). Species overlap was also recorded between Palearctic and Oriental fauna, despite the Himalayas forming an altitudinal barrier between these realms, a pattern also previously recorded for butterflies (Larsen, 1984). Inasmuch as the majority of species (54.9%) were only recorded at a single location and only 3.8% of species were recorded at five or more locations, it is no surprise that a small number of wider distributed species are driving the regional and larger scale biogeographical patterns. The high number of species recorded only once is a typical result for understudied taxa (Velasco-Castrillon et al., 2014; Zhang et al., 2018).

The Neotropical region as one of the lesser studied regions with 71 species recorded from 37 localities, exhibited a higher species richness than that of the Palearctic and Oriental realms. Moreover, despite the Neotropical fauna being closely linked with that of the adjacent Nearctic fauna, from which it has received some, especial boreal components (e.g. *Polypedilum beckae* and *Paralauterborniella nigrohalteralis* , Silva et al., 2015), the results in the current study corroborate our hypothesis that there are significant differences in community structure between the *Polypedilum* fauna in South America, and the neighboring regions. Only a single unidentified species spanned from the Neotropics to the Nearctic region, recorded in Argentina and Mexico, which confirms our expectations of high levels of endemism and richness of *Polypedilum* species in the Neotropical region. The outstanding biodiversity there, when compared to other major biotic realms (Lundberg et al., 2000; Antonelli & Sanmartin, 2011) can be attributed to a complex process in which palaeogeographical and palaeoclimatic forces have been constantly interacting and new species have originated continuously in that area since the late Eocene/early Oligocene (Rull, 2008). As such, the Neotropics is paramount for research on the origin of biological diversity. Finally, some neotropical areas are under manifest danger of biodiversity loss (Antonelli, 2021). Our study shows that DNA-based species delimitation approaches can be used in rapid biodiversity estimates of poorly known taxonomic groups so these can be utilized as basis for biodiversity conservation strategies, and to unravel biogeographical patterns at both local and global scales.

4.3 Conclusion: implications of DNA barcoding to accelerate biogeography research

As different analytical methods have different theoretical foundations, it is advisable to test a wide variety of approaches of species delimitation, and to favor patterns that are congruent across the results. Moreover, the contrast of different methods helps to comprehend their propensity to either split or lump clusters. We evaluated some approaches for species delimitation in the genus *Polypedilum*through single-locus DNA barcodes and found the sGMYC as the method more adequate to estimate putative species on our dataset. Our results highlight *Polypedilum* as species-rich genus, yet incompletely documented, which implies in the need of increased taxon sampling, across geographical ranges, and the use of additional molecular data for greater resolution when using molecular species delimitation approaches for the group. Quantitative

species delimitation methods are sensitive to sampling effort. Since communities typically contain several species that are locally rare, observed species richness provides just an underestimate of the diversity actually present, except if the community is thoroughly sampled. Therefore, a reference COI sequence library derived from expert-identified reference material is fundamental to assign organisms into species by matching the sequence of an unknown sample to the reference library. Our hypothesis that there would be substantial differences in community structure between the*Polypedilum* fauna in South America and other neighboring regions, particularly the Nearctic region, was confirmed. The Neotropical region exhibited high levels of endemism and richness for *Polypedilum*species. Despite major advances in our understanding of Neotropical biodiversity in recent years, several questions remain to be answered: When did the Neotropics reach globally outstanding levels of species richness? Why do nearly all groups of organisms have more species in the Neotropics? What drives latitudinal patterns of diversity? When did the species observed today split from their most recent common ancestors? Further biological and geological data, associated with the integration of different DNA-based methods for estimating species richness, will advance the field of natural history and increase our ability to make knowledge-based decisions in conservation issues. The integration of biodiversity genomics in biogeography science therefore represents a major scientific priority.

References

Antonelli, A. (2021). The rise and fall of Neotropical biodiversity.*Botanical Journal of the Linnean Society* , 199 (1), 8–24. https://doi.org/10.1093/botlinnean/boab061

Antonelli, A., Ariza, M., Albert, J., Andermann, T., Azevedo, J., Bacon, C., Faurby, S., Guedes, T., Hoorn, C., Lohmann, L. G., Matos-Maravi, P., Ritter, C. D., Sanmartin, I., Silvestro, D., Tejedor, M., ter Steege, H., Tuomisto, H., Werneck, F. P., Zizka, A., & Edwards, S.V. (2018). Conceptual and empirical advances in Neotropical biodiversity research.*PeerJ.,* (6), e5644.

Antonelli, A., & Sanmartin, I. (2011). Why are there so many plant species in the Neotropics? *Taxon* , 60, 403–414.

Barnosky, A. D., Matzke, N., Tomiya, S., Wogan, G. O., Swartz, B., Quental, T. B., Marshall, C., McGuire, J. L, Lindsey, E.L., Maguire, K. C., Mersey, B., & Ferrer, E. A. (2011). Has the Earth's sixth mass extinction already arrived? *Nature* , 471, 51–57.

Bidawid, N., & Fittkau, E. J. (1995). Zur Kenntnis der neotropischen Arten der Gattung *Polypedilum* Kieffer, 1912. Teil I. (Diptera, Chironomidae). *Entomofauna* , 16(11), 465–536.

Bidawid-Kafka, N. (1996). Zur Kenntnis der neotropischen Arten der Gattung *Polypedilum* Kieffer, 1912. Teil II. (Diptera, Chironomidae). *Entomofauna* , 17(11), 165–240.

Bouckaert, R., Vaughan, T. G., Barido-Sottani, J., Duchene, S., Fourment, M., Gavryushkina, A., Heled, J., Jones, G., Kuhnert, D., de Maio, N., Matschiner, M., Mendes, F. K., Muller, N. F., Ogilvie, H. A., du Plessis, L., Popinga, A., Rambaut, A., Rasmussen, D., Siveroni, I., Suchard, M. A., Wu, C., Xie, D., Zhang, C., Stadler, T., & Drummond, A. J. (2019). Beast 2.5: An advanced software platform for Bayesian evolutionary analysis. *PLoS Computational Biology* , 15(4), e1006650.

Bradshaw, C. J., Sodhi, N. S., & Brook, B. W. (2009). Tropical turmoil: a biodiversity tragedy in progress. *Frontiers in Ecology and the Environment* , 7, 79–87.

Ceccarelli, F. S., Sharkey, M. J., & Zaldivar-Riveron, A. (2012). Species identification in the taxonomically neglected, highly diverse, neotropical parasitoid wasp genus *Notiospathius* (Braconidae: Doryctinae) based on an integrative molecular and morphological approach. *Molecular Phylogenetics and Evolution* , 62(1), 485–495.

Coscaron, M. C., Melo, M. C., Coddington, J., & Corronca, J. (2009). Estimating biodiversity: a case study on true bugs in Argentinian wetlands. *Biodiversity and Conservation* , 18, 1491–1507.

Cotoras, D. D., & Zumbad, M. A. (2020). Aerial plankton from the Eastern Tropical Pacific. *Revista de Biologia Tropical* , 68, 155–162.

da Silva, R., Peloso, P. L. V., Sturaro, M. J., Veneza, I., Sampaio, I., Schneider, H., & Gomes, G. (2018). Comparative analyses of species delimitation methods with molecular data in snappers (Perciformes: Lutjaninae). *Mitochondrial DNA Part A* , 29(7), 1108–1114.

Darwin, C. (1859). *On the Origin of Species by Means of Natural Selection* . London, UK: John Murray, London. 502 pp.

Dellicour, S., & Flot, J. F. (2015). Delimiting species-poor data sets using single molecular markers: A study of barcode gaps, haplowebs and GMYC. *Systematic Biology* , 64, 900–908.

Dellicour, S., & Flot, J. F. (2018). The hitchhiker's guide to single-locus species delimitation. *Molecular EcologyResources* , 18(6), 1234–46.

Deng, C., Daley, T., & Smith, A. D. (2015). Applications of species accumulation curves in large-scale biological data analysis.*Quantitative Biology* , 3(3),135–144.

Dopheide, A., Tooman, L. K., Grosser, S., Agabiti, B., Rhode, B., Xie, D., Stevens, M. I., Nelson, N., Buckley, T. R., Drummond, A. J., & Newcomb, R. D. (2019). Estimating the biodiversity of terrestrial invertebrates on a forested island using DNA barcodes and metabarcoding data. *Ecological Applications* , 29(4), e01877.

Drummond, A. J., Nicholls, G. K., Rodrigo, A. G., & Solomon, W. (2002). Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. *Genetics* , 161(3), 1307–1320.

Drummond, A. J., Suchard M. A., Xie, D. & Rambaut, A. (2012). Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Molecular Biology and Evolution* , 22(8), 1185–1192.

Dupuis, J. R., Roe, A. D., & Sperling, F. A. H. (2012). Multi-locus species delimitation in closely related animals and fungi: one marker is not enough. *Molecular Ecology* , (21), 4422–4436.

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* , (32), 1792–1797.

Ekrem, T., Stur, E., Orton, M. G., & Adamowicz, S. J. (2018). DNA barcode data reveal biogeographic trends in Arctic non-biting midges.*Trends in DNA Barcoding and Metabarcoding* , 1(1), 787–796.

Elworth, R. A. L., Ogilvie, H. A., Zhu, J., & Nakhleh, L. (2018). Advances in computational methods for phylogenetic networks in the presence of hybridization. *arXiv* 1808.08662v1.

Faith, D. P. (1992). Conservation evaluation and phylogenetic diversity.*Biology Conservation,* 61, 1–10.

Faith, D. P. (2008). *Phylogenetic diversity and conservation* . pp. 99–115 in Carroll, S. P., & Fox, C. (Eds) Conservation Biology: Evolution in Action. New York, Oxford University Press.

Farooq, Q., Shakir, M., Ejaz, F., Zafar, T., Durrani, K., & Ullah, A. (2020). Role of DNA Barcoding in Plant Biodiversity Conservation.*Scholars International Journal of Biochemistry* , 3(3), 48–52.

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: A maximum likelihood approach. *Journal of Molecular Evolution* , 17, 368–376.

Ferrington, L. C. Jr. (2008). Global diversity of non-biting midges (Chironomidae; Insecta-Diptera) in freshwater. *Hydrobiologia* , 595, 447–455.

Folmer, O., Black, M., Hoeh, W., Lutz, R., & Vrijenhoek, R. (1994). DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. *Molecular Marine Biology and Biotechnology* , 3, 294–299.

Fujisawa, T., & Barraclough, T. G. (2013). Delimiting species using single-locus data and the Generalized Mixed Yule Coalescent approach: A revised method and evolution on simulated data sets. *Systematic Biology* , 62, 707–724.

Gostel, M. R., & Kress, W. J. (2022). The Expanding Role of DNA Barcodes: Indispensable Tools for Ecology, Evolution, and Conservation.*Diversity* , 14(3), 213.

Gressitt, J. L., Leech, R. E. & O'Brien, C. W. (1960). Trapping of air-borne insects in the Antarctic area. *Pacific Insects* , 2, 245–250.

Havermans, C., Nagy, Z. T., Sonet, G., De Broyer, C. & Martin, P. (2011). DNA barcoding reveals new insights into the diversity of Antarctic species of *Orchomene* sensu lato (Crustacea: Amphipoda: Lysianassoidea). *Deep-Sea Research Part II Topical Studies in Oceanography* , 58(1–2), 230–241.

Hardy, A. C., & Milne, P. S. (1938). Studies in the Distribution of Insects by Aerial Currents. *Journal of Animal Ecology* , 7(2), 199–229.

Hebert, P. D. N., Cywinska, A., Ball, S. L., & deWaard, J. R. (2003). Biological identifications through DNA barcodes. *Proceedings of the Royal Society B: Biological Sciences* , 270, 313–321.

Hebert, P. D. N., Stoeckle, M. Y., Zemlak, T. S., & Francis, C. M. (2004). Identification of birds through DNA barcodes. *PLoS Biology* , 2, e312.

Hebert, P. D. N., Ratnasingham, S., Zakharov, E. V., Telfer, A. C., Levesque-Beaudin, V., Milton, M. A., Pedersen, S., Jannetta, P., & deWaard, J. R. (2016). Counting animal species with DNA barcodes: Canadian insects. *Philosophical Transactions of the Royal Society B* , 371, 20150333.

Hobern, D. (2021). BIOSCAN: DNA barcoding to accelerate taxonomy and biogeography for conservation and sustainability. *Genome* , 64, 161–164.

Hooker, J. D. (1844-47). The Botany of the Antarctic Voyage of H. M. Discovery Ships Erebus and Terror in the Years 1839–1843, under the Command of Captain Sir James Clark Ross. Flora Antarctica. London, Reeve Brothers

Hoang, D. T., Chernomor, O., von Haeseler, A., Minh, B. Q., & Vinh, L. S. (2018). UFBoot2: Improving the Ultrafast Bootstrap Approximation.*Molecular Biology and Evolution* , 35(2), 518–522.

Kapli, P., Lutteropp, S., Zhang, J. P., Kobert, K., Pavlidis, P., Stamatakis, A., & Flouri, T. (2017). Multi-rate Poisson tree processes for single-locus species delimitation under maximum likelihood and Markov chain Monte Carlo. *Bioinformatics* , 33, 1630–1638.

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., Duran, C., Thierer, T., Ashton, B., Meintjes, P., & Drummond, A. J. (2012). Geneious basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* , 28, 1647–1649.

Kekkonen, M., & Hebert, P. D. N. (2014). DNA barcode-based delineation of putative species: efficient start for taxonomic workflows.*Molecular Ecology Resources* , 14, 706–715.

Kimura, M. A. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences.*Journal of Molecular Evolution* , 16(2), 111–20.

Lanfear, R., Frandsen, P. B., Wright, A. M., Senfeld, T., & Calcott, B. (2017). PartitionFinder 2: New Methods for Selecting Partitioned Models of Evolution for Molecular and Morphological Phylogenetic Analyses.*Molecular Biology and Evolution* , 34(3), 772–773.

Larsen, T. B. (1984). The Zoogeographical Composition and Distribution of the Arabian Butterflies (Lepidoptera; Rhopalocera). *Journal of Biogeography* , 11(2), 119–158.

Laurance, W. F. (1999). Reflections on the tropical deforestation crisis. *Biological Conservation* , 91, 109–117.

Lin, X. L., Stur, E., & Ekrem, T. (2015). Exploring genetic divergence in a species-rich insect genus using 2790 DNA barcodes. *PLoS ONE* , 10(9), e0138993.

Lundberg, J. G., Kottelat, M., Smith, G. R., Stiassny, M. L. J., & Gill, A. C. (2000). So many fishes, so little time: an overview of recent ichthyological discovery in continental waters. Annals of the Missouri Botanical Garden, 87(1), 26–62.

Marusik, Y. M., & Koponen, S. (2005). A survey of spiders (Araneae) with Holarctic distribution. *Journal of Arachnology* , 33, 300–305.

Meyer, C. P., & Paulay G. (2005). DNA barcoding: Error rates based on comprehensive sampling. *Plos Biology* , 3, 2229–2238.

Miller, M. A., Pfeiffer, W., & Schwartz, T. (2011). Creating the CIPRES Science Gateway for inference of large phylogenetic trees.*Proceedings of the 2011 TeraGrid Conference: Extreme Digital Discovery* , 41, 1–8.

Miralles, A., & Vences, M. (2013). New metrics for comparison of taxonomies reveal striking discrepancies among species delimitation methods in *Madascincus* lizards. *PLoS ONE* , (8) e68242.

Monaghan, M. T., Wild, R., Elliot, M., Fujisawa, T., Balke, M., Inward, D. J., Lees, D. C., Ranaivosolo, R., Eggleton, P., Barraclough, T. G., & Vogler, A. P. (2009). Accelerated species inventory on Madagascar using coalescent-based models of species delineation. *Systematic Biology* , 58, 298–311.

Morrone, J. J., Escalante, T., Rodriguez-Tapia, G., Carmona, A., Arana, M., & Mercado-Gomez, J. D. (2022). Biogeographic regionalization of the Neotropical region: New map and shapefile. *Anais da Academia Brasileira de Ciencia* , 4(1), e20211167.

Nzelu, C. O., Caceres, A. G., Arrunategui-Jimenez, M. J., Lanas-Rosas, M. F., Yanez-Trujillano, H. H., Luna-Caipo, D. V., Holguin-Mauricci, C. E., Katakura, K., Hashiguchi, Y., & Kato, H. (2015). DNA barcoding for identification of sand fly species (Diptera: Psychodidae) from leishmaniasis-endemic areas of Peru. *Acta Tropica* , 145, 45–51.

Oksanen, J., Blanchet, F. G., Kindt, R., Legendre, P., Minchin, P. R., O'Hara, R. B., Simpson, G. L., Solymos, P., Henry, M., Stevens, H., & Wagner, H. (2017). Community Ecology Package. R package version 2.4-3.*http://cran.r-project.org/web/packages/vegan*

O'Meara, B. (2010). New heuristic methods for joint species delimitation and species tree inference. *Systematic Biology* , 59, 59–73.

Oyewo, E. A., & Saether, O. A. (2008). Revision of *Polypedilum*(*Pentapedilum*) Kieffer and *Ainuyusurika* Sasa et Shirasaki (Diptera: Chironomidae). *Zootaxa* , 1953, 1–145.

Parslow B.A., Schwarz M. P., & Stevens M.I. (2021). Molecular diversity and species delimitation in the family Gasteruptiidae (Hymenoptera: Evanioidea). *Genome* , 64, 253–264.

Paz, A., & Crawford, A. J. (2012). Molecular-based rapid inventories of sympatric diversity: a comparison of DNA barcode clustering methods applied to geography-based vs clade-based sampling of amphibians.*Journal of Biosciences* , 37, 887–896.

Pearl, H., Ryan, T., Howard, M., Shimizu, Y., & Shapcott, A. (2022). DNA Barcoding to Enhance Conservation of Sunshine Coast Heathlands.*Diversity* , 14, 436.

Pinho, L. C., & Silva, F. L. (2020). Description of two new species of*Polypedilum* (*Asheum*) and immature stages of*Polypedilum* (*A* .) *curticaudatum* (Diptera: Chironomidae). *Zootaxa* , 4759(2), 179–190.

Pons, J., Barraclough, T. G., Gomez-Zurita J., Cardoso, A., Duran, D. P., Hazell, S., Kamoun, S., Sumlin, W. D., & Vogler, A. P. (2006). Sequence-based species delimitation for the DNA taxonomy of undescribed insects. *Systematic Biology* , 55, 595–609.

Puillandre, N., Lambert, A., Brouillet, S., & Achaz, G. (2012). ABGD, Automatic Barcode Gap Discovery for primary species delimitation.*Molecular Ecology* , 21, 1864–1877.

Puillandre, N., Brouillet, S., & Achaz, G. (2021). ASAP: assemble species by automatic partitioning. *Molecular Ecology Resources* , 21, 609–620.

R Core Team. (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.*https://www.R-project.org/*

Rannala, B., & Yang, Z. (2020). Species Delimitation. Scornavacca, Celine; Delsuc, Frederic; Galtier, Nicolas. Phylogenetics in the Genomic Era, No commercial publisher | Authors open access book, pp. 5.5:1–5.5:18.

Rambaut, A. (2010). FigTree version 1.4.4. Institute of Evolutionary Biology, University of Edinburgh, Edinburgh.*http://tree.bio.ed.ac.uk/software/figtree/*

Rambaut, A., Suchard. M. A., Xie, D., & Drummond, A. J. (2018). Tracer v1.7. Available from*http://beast.bio.ed.ac.uk/Tracer*

Ratnasingham, S., & Hebert, P. D. N. (2013). A DNA-based registry for all animal species: the barcode index number (BIN) system. *PLoS One* , 8, e66213.

Rodrigues, B. L., Baton, L. A., & Shimabukuro, P. H. F. (2020). Single-locus DNA barcoding and species delimitation of the sandfly subgenus *Evandromyia* (*Aldamyia*). *Medical and Veterinary Entomology* , 34(4), 420–431.

Rodríguez, J., Hortal, J., & Nieto, M. (2006). An evaluation of the influence of environment and biogeography on community structure: the case of Holarctic mammals. *Journal of Biogeography* , 33(2), 291–303.

Rull, V. (2008). Speciation timing and neotropical biodiversity: the Tertiary-Quaternary debate in the light of molecular phylogenetic evidence. *Molecular Ecology* , 17, 2722–2729.

Schwarzfeld, M. D., & Sperling, F. A. H. (2015). Comparison of five methods for delimitating species in *Ophion* Fabricius, a diverse genus of parasitoid wasps (Hymenoptera, Ichneumonidae). *Molecular Phylogenetics and Evolution* , 93, 234–248.

Shapcott, A., Forster, P. I., Guymer, G. P., McDonald, W. J. F., Faith, D. P., Erickson, D., & Kress, W. J. (2015). Mapping biodiversity and setting conservation priorities for SE Queensland's rainforests using DNA barcoding. *PLoS ONE* , 10, e0122164.

Shimabukuro, E. M., Trivinho-Strixino, S. & Lamas, C. J. E. (2019). New*Polypedilum* Kieffer (Diptera: Chironomidae) from mountains of the Atlantic Forest, Brazil. *Zootaxa* , 4612(4), 518–532.

Silva, F. L., & Ekrem, T. (2016). Phylogenetic relationships of non-biting midges in the subfamily Tanypodinae (Diptera: Chironomidae) inferred from morphology. *Systematic Entomology* , 41, 73–92.

Silva, F. L., & Farrell, B. (2017). Non-biting midges (Diptera: Chironomidae) research in South America: subsidizing biogeographic hypotheses. Annales de Limnologie - International Journal of Limnology, 53, 111–128.

Silva, F. L., Wiedenbrug, S., & Farrell, B. (2015). A preliminary survey of the non-biting midges(Diptera: Chironomidae) of the Dominican Republic. *CHIRONOMUS Journal of Chironomidae Research* , 28, 12–19.

Song, C., Wang, Q., Zhang, R., Sun, B., & Wang, X. (2016). Exploring the utility of DNA barcoding in species delimitation of*Polypedilum* (*Tripodura*) non-biting midges (Diptera: Chironomidae). *Zootaxa* , 4079, 534–550.

Song, C., Lin, X., Wang, Q., & Wang, X. (2018). DNA barcodes successfully delimit morphospecies in a superdiverse insect genus.*Zoologica Scripta* , 47(3), 311–324.

Spies, M., & Reiss, F. (1996). Catalog and bibliography of Neotropical and Mexican Chironomidae (Insecta, Diptera). *Spixiana Supplement* , 22, 61–119.

Stork, N. E. (2018). How many species of insects and other terrestrial arthropods are there on earth? *Annual Review of Entomology* , 63, 31–45.

Sæther, O. A. (1969). Some Nearctic Podonominae, Diamesinae, and Orthocladiinae (Diptera: Chironomidae), Department of Fisheries and Oceans, Ottawa, 154 pp.

Sæther, O. A, Andersen, T., Pinho, L. C., & Mendes, H. F. (2010). The problems with *Polypedilum* Kieffer (Diptera: Chironomidae), with the description of *Probolum* subgen. n. *Zootaxa* , 2497, 1–36.

Sæther, O. A., & Oyewo, E. A. (2008). Keys, phylogenies and biogeography of *Polypedilum* subgen. *Uresipedilum* Oyewo et Sæther (Diptera: Chironomidae). *Zootaxa* , 1806, 1–34.

Sæther O.A., & Sundal, A. (1999). *Cerobregma* , a new subgenus of*Polypedilum* Kieffer, with a tentative phylogeny of subgenera and species groups within *Polypedilum* (Diptera: Chironomidae).*Journal of the Kansas Entomological Society* , 71, 315–382.

Talavera, G., Dinca, V., & Vila, R. (2013). Factors affecting species delimitations with the GMYC model: insights from a butterfly survey.*Methods in Ecology & Evolution* , 4, 1101–1110.

Tamura, K., Stecher, G., & Kumar S. (2021). MEGA11: Molecular Evolutionary Genetics Analysis Version 11. *Molecular Biology and Evolution* , 38(7), 3022–3027.

Tang, C. Q., Humphreys, A. M., Fontaneto, D., & Barraclough, T. G. (2014). Effects of phylogenetic reconstruction method on the robustness of species delimitation using single-locus data. *Methods in Ecology & Evolution* , 5, 1086–1094.

Townes, H. K. Jr. (1945). The Nearctic species of Tendipedini [Diptera, Tendipedidae (= Chironomidae)]. *American Midland Naturalist* , 34, 1–206.

Trifinopoulos, J., Nguyen, L.-T., von Haeseler, A., Minh, B. Q. (2016). W-IQ-TREE: a fast online phylogenetic tool for maximum likelihood analysis. *Nucleic Acids Research* , 44, W232–W235.

Vårdal, H., Bjørlo, A., & Sæther, O. A. (2002). Afrotropical*Polypedilum* subgenus *Tripodura* , with a review of the subgenus (Diptera: Chironomidae). *Zoologica Scripta* , 31, 331–402.

Velasco-Castrillón, A., Page, T. J., Gibson, J. A., & Stevens, M. I. (2014). Surprisingly high levels of biodiversity and endemism amongst Antarctic rotifers uncovered with mitochondrial DNA.*Biodiversity* , 15(2–3), 130–142.

Wallace, A. R. (1876). *The Geographical Distribution of Animals* . London, UK: Macmillans.

Webb, C. O. (2000). Exploring the phylogenetic structure of ecological communities: an example for rain forest trees. *The American Naturalist* , 156, 145–155.

Wiemers, M., & Fiedler, K. (2007). Does the DNA barcoding gap exist? - a case study in blue butterflies (Lepidoptera: Lycaenidae).*Frontiers in Zoology* , (4), 8.

Yang, Z. H., & Rannala, B. (2017). Bayesian species identification under the multispecies coalescent provides significant improvements to DNA barcoding analyses. *Molecular Ecology* , 26(11), 3028–3036.

Zhang, J., Kapli, P., Pavlidis, P., & Stamatakis, A. (2013). A general species delimitation method with applications to phylogenetic placements. *Bioinformatics* , 29, 2869–2876.

Zhang, F., Jantarit, S., Nilsai, A., Stevens, M. I., Ding, Y., & Satasook, C. (2018). Species delimitation in the morphologically conserved *Coecobrya* (Collembola: Entomobryidae): a case study integrating morphology and molecular traits to advance current taxonomy.*Zoologica Scripta* , 47(3), 342–356.

Zhou, X., Jacobus. L. M., DeWalt, R. E., Adamowicz, S. J., & Hebert, P. D. N. (2010). Ephemeroptera, Plecoptera, and Trichoptera fauna of Churchill (Manitoba, Canada): insights into biodiversity patterns from DNA barcoding. *Journal of the North American Benthological Society* , 29(3), 814–837.

Data Accessibility Statement

The detailed specimen records and sequence information, including trace files, are freely available in BOLD through the dataset 'DS-RPPPOL, Reduced personal and publicly available records of *Polypedilum*(Diptera: Chironomidae)' with DOI: https://doi.org/XXXX. Significant ANOSIM results per realm and region are given in Appendices S1 and S2.

Hosted file

`Table 1.docx` available at https://authorea.com/users/584361/articles/623607-dna-barcodes-provide-insights-into-the-diversity-and-biogeography-of-the-non-biting-midge-polypedilum-diptera-chironomidae-in-south-america

Hosted file

`Table 2.docx` available at https://authorea.com/users/584361/articles/623607-dna-barcodes-provide-insights-into-the-diversity-and-biogeography-of-the-non-biting-midge-polypedilum-diptera-chironomidae-in-south-america