






# Tiny insects, big troubles: a review of BOLD's COI database for Thysanoptera (Insecta)

cambridge.org/ber

Mariana F. Lindner<sup>1</sup> , Leonardo T. Gonçalves<sup>2</sup> , Filipe M. Bianchi<sup>1</sup> ,  
Augusto Ferrari<sup>3</sup>  and Adriano Cavalleri<sup>3</sup> 

## Research Paper

**Cite this article:** Lindner MF, Gonçalves LT, Bianchi FM, Ferrari A, Cavalleri A (2023). Tiny insects, big troubles: a review of BOLD's COI database for Thysanoptera (Insecta). *Bulletin of Entomological Research* 1–13. <https://doi.org/10.1017/S0007485323000391>

Received: 19 December 2022

Revised: 29 May 2023

Accepted: 22 July 2023

### Keywords:

Aeolothripidae; Barcode gap; DNA Barcoding; *Frankliniella*; Phlaeothripidae; Thripidae

### Corresponding author:

Mariana F. Lindner;

Email: [mflindner@hotmail.com](mailto:mflindner@hotmail.com)

<sup>1</sup>Department of Zoology, Laboratório de Entomologia Sistemática, Institute of Biosciences, Universidade Federal do Rio Grande do Sul (UFRGS), Porto Alegre, RS, Brazil; <sup>2</sup>Department of Genetics, Universidade Federal do Rio Grande do Sul (UFRGS), Porto Alegre, RS, Brazil and <sup>3</sup>Laboratório de Entomologia, Sistemática e Biogeografia (LESB), Matéria Zoologia, Institute of Biological Sciences, Universidade Federal do Rio Grande (FURG), Rio Grande, RS, Brazil

### Abstract

DNA Barcoding is an important tool for disciplines such as taxonomy, phylogenetics and phylogeography, with Barcode of Life Data System (BOLD) being the largest database of partial cytochrome *c* oxidase subunit I (COI) sequences. We provide the first extensive revision of the information available in this database for the insect order Thysanoptera, to assess: how many COI sequences are available; how representative these sequences are for the order; and the current potential of BOLD as a reference library for specimen identification and species delimitation. The COI database at BOLD currently represents only about 5% of the over 6400 valid thrips species, with a heavy bias towards a few species of economic importance. Clear Barcode gaps were observed for 24 out of 33 genera evaluated, but many outliers were also observed. We suggest that the COI sequences available in BOLD as a reference would not allow for accurate identifications in about 30% of Thysanoptera species in this database, which rises to 40% of taxa within Thripidae, the most sampled family within the order. Thus, we call for caution and a critical evaluation in using BOLD as a reference library for thrips Barcodes, and future efforts should focus on improving the data quality of this database.

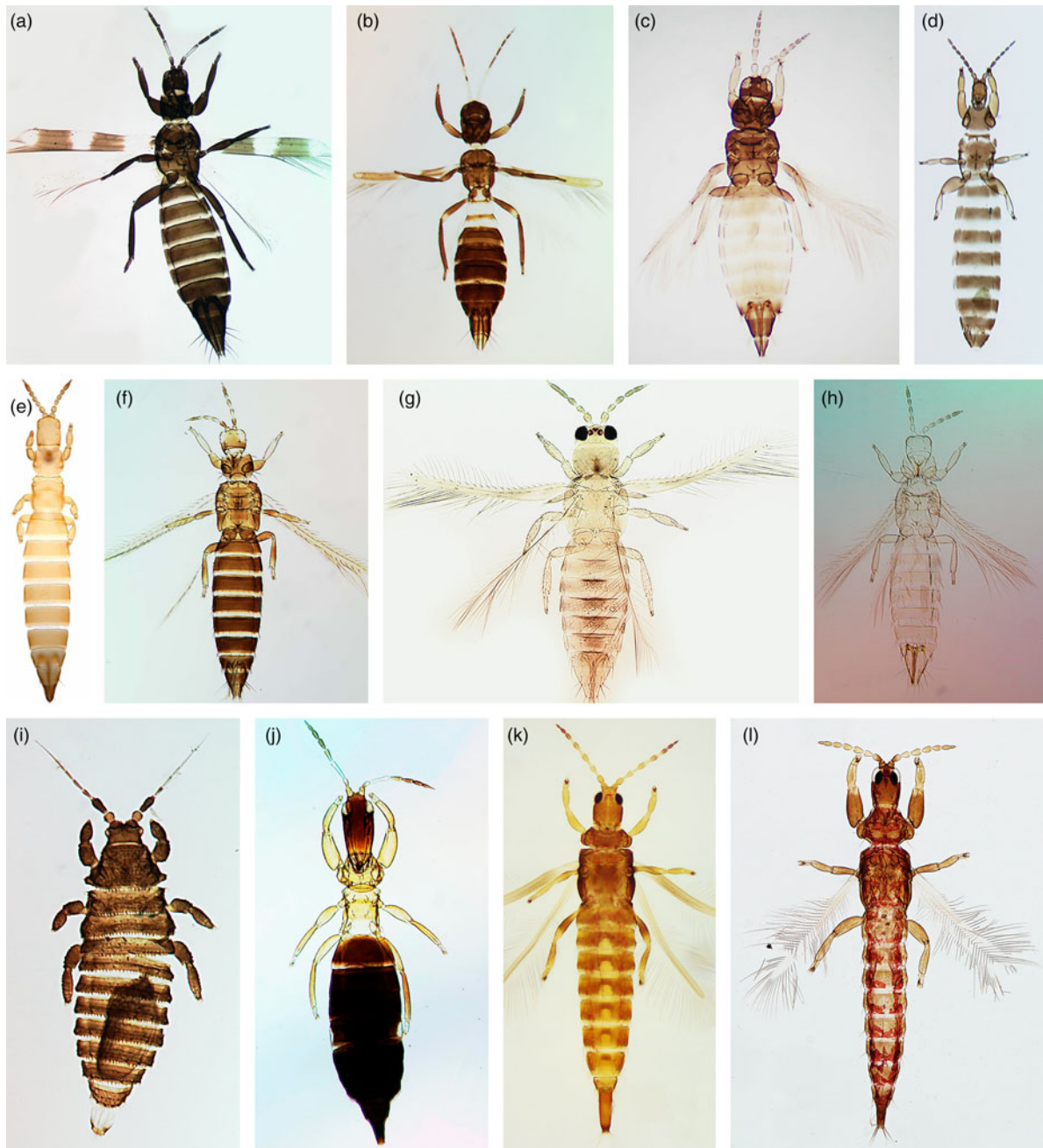
## Introduction

Since its first suggestion in the early 2000s (Hebert *et al.*, 2003), DNA Barcoding has received much attention due to its versatility as a global bioidentification system. The proposal of using a specific DNA sequence as a type of barcode for all life forms, allowing for quick comparisons and easier identification of specimens, is attractive in the context of fewer taxonomists and less time available for the careful study of specimens. In 2007, the Barcode of Life Data System (BOLD) was created as a freely available online workbench for collecting, analysing and sharing DNA Barcodes (Ratnasingham and Hebert, 2007).

In 15 years of existence, BOLD has gathered almost 14 million DNA sequences of over 345 thousand animals, plants and fungi species (as of May 16th, 2023; BOLD, 2023). It is a valuable repository allowing the association of voucher pictures with sequence data, which increases the repeatability and verification of information, two fundamental principles of scientific work (Vink *et al.*, 2012; Bianchi and Gonçalves, 2021b). BOLD also provides its own tool for species delimitation, the Barcode Index Number (BIN), which is based on cluster analysis of the sequences in the database, and compatible with the constant inclusion of new data (Ratnasingham and Hebert, 2013). Each BIN can represent a potential species, allowing the evaluation of such units and their use in the lack of a well-developed taxonomic frame.

However, some authors have pointed out some limitations currently found in BOLD for specific taxa (Sonet *et al.*, 2013; Lis *et al.*, 2016; Gonçalves *et al.*, 2021; Bianchi and Gonçalves, 2021a), and even questioning the quality of the data added to this database. For example, it has been shown that there are problems in the acquisition of reference data and its curation in BOLD and GenBank, as well as in the production of sequences to assess the reference data (Meiklejohn *et al.*, 2019; Pentinsaari *et al.*, 2020). Thus, the efforts to improve the quality of data of these online databases must be continuous, and should include revision and curation of available data.

While DNA Barcoding has been extensively utilised in many taxa, for the insect order Thysanoptera (popularly known as thrips; Fig. 1) it is still a rather incipient tool. With over 6400 species in the order and a cosmopolitan distribution, Barcode data are available only for a few species, most of them with some importance for agriculture (e.g., Karimi *et al.*, 2010; Chakraborty *et al.*, 2019). In fact, only a few works deal with a large variety of thrips taxa, and most studies focus on a limited geographical area (Iftikhar *et al.*, 2016; Tyagi *et al.*, 2017) or a specific family (Marullo *et al.*, 2020). Still, partial cytochrome *c* oxidase



**Figure 1.** Mounted specimens of a variety of Thysanoptera species and families. A-B: Aeolothripidae; A: *Aeolothrips fasciatus* (Linnaeus, 1758), B: *Franklinothrips vespiformis* (Crawford DL, 1909). C: Heterothripidae, *Heterothrips bicolor* Hood, 1954. D: Merothripidae, *Merothrips brunneus* Ward, 1969. E-H: Thripidae; E: *Aptinothrips stylifer* Trybom, 1894, F: *Frankliniella occidentalis* (Pergande, 1895), G: *Scirtothrips dorsalis* Hood, 1919, H: *Thrips palmi* Karny, 1925. I: Uzelothripidae, *Uzelothrips scabrosus* Hood, 1952. J-L: Phlaeothripidae; J: *Compsothrips graminis* (Hood, 1936), K: *Eschatothrips decoratus* Hood, 1957, L: *Haplothrips dissociatus* Cavalleri, Lindner & Mendonça, 2016. Photos A and E are from the site Thrips of California 2012 (Available at [https://keys.lucidcentral.org/keys/v3/thrips\\_of\\_california/Thrips\\_of\\_California.html](https://keys.lucidcentral.org/keys/v3/thrips_of_california/Thrips_of_California.html); accessed on September 16th, 2022), remaining photos from The Thrips of Brazil (Available at <http://thysanoptera.com.br/home>; accessed on September 16th, 2022).

subunit I (COI) sequences, especially at the 5' portion (COI-5P), have shown potential to be a useful identification tool for these insects, as shown in the recent revision of Ghosh *et al.* (2021) of molecular and electronic identification tools.

Thysanoptera specimens offer difficulties and limitations for their DNA extraction and sequencing. Most preserved specimens no longer contain any source of DNA, thus molecular studies of thrips require freshly collected specimens. Their small size requires the usage of whole specimens for DNA extraction, and

some procedures can easily damage the thrips, hampering specimen usage for molecular and morphological data concurrently. Finally, thrips often yield low quantities of DNA, further complicating molecular analyses (Dickey *et al.*, 2015).

This work aims to evaluate the available COI sequences for Thysanoptera in BOLD. Despite the existence of other databases for genetic sequences, such as GenBank, our focus on BOLD data is due to its emphasis on DNA Barcodes and implementation of several quality control steps. The objectives of this study

**Table 1.** Number of sequences and taxon representativity on BOLD, after each filtering step

Filtering step	N Sequences (% from total)	N Families	N Genus labels	N Species labels
0. Data labelled 'Thysanoptera' downloaded from BOLD (Nov. 2021)	30,581 (100)	7	139	323
1. Sequences other than COI-5P removed	29,920 (97.84)	7	125	300
2. Sequences lacking species identification removed	11,096 (36.29)	7	115	297
3. Genera with a single species removed	10,434 (34.12)	4	37	219
4. Sequences separated by family and aligned	10,434 (34.12)	4	37	219
5. Sequences with less than 400 bp removed	9816 (32.10)	4	37	198
6. Sequences separated into genera	9816 (32.10)	4	37	198
7. Genera with less than two species, or only with singleton species, removed	9810 (32.08)	3	33	193

were: (1) investigate the representativity of BOLD sequences compared to the valid taxa within Thysanoptera; (2) identify Barcode gaps at the generic level; and (3) assess the correct identifications of thrips specimens using DNA Barcodes. After these analyses, we highlight some taxa within Thysanoptera that need a careful taxonomic revision, sequences whose identity may need to be re-evaluated, and suggest ways to improve the overall quality of the database available in BOLD.

## Materials and methods

The workflow described below follows and adapts the methodology utilised in Gonçalves *et al.* (2021) and Bianchi and Gonçalves (2021a).

### Data acquisition and filtering

All sequences available on BOLD labelled as 'Thysanoptera' were manually downloaded in November 2021 (database 0). We curated this original database to remove sequences which did not fit the criteria needed for our analyses, and Table 1 lists how many sequences were removed at each step. The filtering steps are as follows: (1) removal of sequences of genes other than COI-5P; (2) all sequences without species-level identification removed, and names corrected whenever needed (synonymy, misspellings); (3) removal of all genera with a single species, as the probability of correct identification (PCI) analysis requires all genera to have at least two species; (4) remaining sequences divided into families and aligned using MAFFT v7.0 (Kato *et al.*, 2019); (5) alignments were trimmed to the canonical barcode region (Hebert *et al.*, 2003) using as reference the BOLD entry MAIMB460-09 (*Thrips palmi*), and all sequences with less than 400 bp were removed; (6) sequences were separated by genus; (7) genera with less than two species, or lacking any species with two or more sequences, were removed, to ensure intra- and interspecific comparisons for Barcode gap analysis. With these steps, Databases 1, 2 and 3 were generated (Table 2). All sequences were treated by their species name only, with subgenera or subspecies not being considered. Table 1 lists how many sequences, families, genera and species labels were available after each filtering step. All databases utilised in this work are given in Supplementary file 1. A dataset on BOLD has been generated with the majority of sequences downloaded in November 2021, under the name 'DS-THRIPS21'.

**Table 2.** Databases generated in this work, filtering steps completed on each, and analyses performed with them. All databases are available in Supplementary file 1.

	Filtering steps applied	Analyses performed
Database 0 (raw data from BOLD)	None	Distribution map A
Database 1 (COI sequences with ID)	1-2	Representativity
Database 2 (COI sequences by family)	1-5	Probability of correct identification (PCI)
Database 3 (COI sequences by genera)	1-7	Barcode gap analysis Boxplot outliers Distribution map B

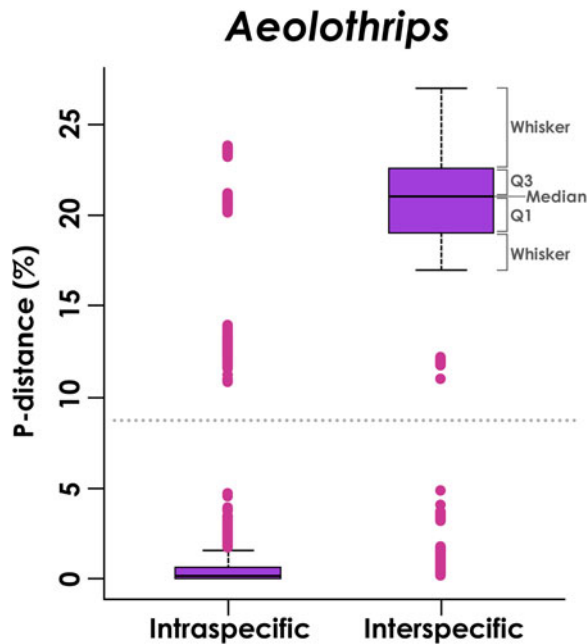
### Representativity of Thysanoptera data on BOLD

We assessed the representativeness of Database 1 for Thysanoptera taxa by determining the number of families, genera and species included. We also examined the distribution of sequences within these taxa to identify any potential biases. Geographical distribution data from databases 0 and 3 were obtained to generate global heat maps, to evaluate shifts in distribution patterns before and after filtering steps. The maps were created with MapChart, available at <https://www.mapchart.net>.

### Barcode gap analysis

To evaluate the presence of Barcode gaps in Thysanoptera, we used the function `dist.dna()` of the R package `ape` (Paradis and Schliep, 2019) on database 3 to estimate pairwise uncorrected p-distances for all sequences within each genus (Supplementary File 2). We used uncorrected p-distances because they yield better or similar results when compared to other nucleotide substitution models, such as Kimura 2-parameter (Collins *et al.*, 2012; Srivathsan and Meier, 2012). Intra- and interspecific distances were then represented in a boxplot for each evaluated genus, using the base R function `'boxplot()'`. This allows the automatic identification of outliers, which represent comparisons between two sequences whose distances fall outside the extent of the whiskers (fig. 2).

The boxplots allow visualisation of the Barcode gap, which were classified into one of the following three categories: Good,



**Figure 2.** Example of a boxplot graph, indicating its parts. The outliers are represented as pink dots. The horizontal dotted line represents the cut-off area we delimited for selecting outliers for visual inspection in this genus. The boxplots generated illustrate the lower (Q1) and upper (Q3) quartiles as the two parts of the box, divided in the middle by the median value of the observed data. The whiskers represent  $\pm 1.5$  Interquartile Range (IQR), the range from the lower limit of the boxplot (25%) to the upper limit of the boxplot (75%).

when there was no overlap between intraspecific and interspecific boxplots; Intermediate, when there was an overlap between boxplot whiskers only; and Poor, when the boxplot boxes overlapped (Badotti *et al.*, 2017; Bianchi and Gonçalves, 2021a).

Many of the boxplot graphs showed at least one outlier. To analyse these, we listed the intraspecific outliers above the upper whisker limit, and the interspecific outliers below the lower whisker limit (Supplementary Files 3–4). These were chosen due to their potential overlap with interspecific distances and intraspecific distances, respectively. Supplementary file 5 lists all the genera with outliers and how representative they are concerning the number of potential comparisons and sequences available.

Finally, to demonstrate the potential of outlier comparisons in detecting taxonomic inconsistencies, we conducted a detailed examination of select outliers for *Aeolothrips* Haliday, 1836 and *Frankliniella* Karny, 1910. These genera were chosen due to the abundance of available sequences, their economic importance and their history of challenging taxonomy.

### Probability of correct identification (PCI) analysis

To evaluate if the available sequences in BOLD allow for the correct identification of COI sequences within Thysanoptera, we calculated the PCI (Supplementary file 6) utilising database 2. The PCI is a ‘discrete species assignment’ and considers the maximum intraspecific distance and the minimum interspecific distance (or nearest-neighbour distance) for each recognised species (Erickson *et al.*, 2008). Then, these values are visualised in a scatterplot, where each dot represents a species name (Collins and Cruickshank, 2012). By drawing in the graph a line where  $x = y$ , it is possible to divide the species dots between two groups. Those above the  $x = y$  line have the nearest neighbour distance

**Table 3.** Number of Thysanoptera taxa with at least one sequence after filtering step 2 (database 1), compared to the number of taxa currently accepted in the order (following ThripsWiki 2023)

	Taxa with sequences (valid taxa in the order)	Barcode coverage (%)
Suborders	2 (2)	100
Families	7 (9)	77.78
Genera	115 (787)	14.61
Species	297 (6414)	4.63

higher than the maximum intraspecific distance, and thus are considered to provide a ‘correct’ identification (since there is a clear gap between the species and the closest one, thus a clear delimitation of that species). Those dots below the  $x = y$  line have the nearest neighbour distance lower than the maximum intraspecific distance, and thus are considered to provide an ‘incorrect’ identification (since there is an overlap between the species and the closest neighbour, therefore a query sequence could fall in this overlap and have an uncertain identity). By calculating the number of points above the line in relation to the total number of points in the graph, it is possible to calculate the PCI for a given taxon. Thus, PCI calculations were performed for Thysanoptera as a whole and for three families (*Aeolothripidae*, 12 species names; *Phlaeothripidae*, 52 species names; and *Thripidae*, 96 species names).

## Results

### Representativity of Thysanoptera data on BOLD

A total of 30,581 sequences were obtained from BOLD, of which about one third had any image record, and only 5% were barcode compliant. After removing non-COI sequences and those lacking species identification (steps 1 and 2 of the filtering procedure), 11,096 sequences remained, representing seven families, 115 genus labels and 297 species labels (Table 1). The overall representativity of these sequences was low, with less than 15% of genera and 5% of valid thrips species (*sensu* ThripsWiki, 2023) (Table 3). Representativity of species varied in each family and subfamily, but most families had only a third or less of their genera represented in BOLD (Table 4).

Out of the 11,096 sequences analysed, almost 90% belonged to Thripidae species (fig. 3A). Three genera comprise nearly 70% of the sequences: *Taeniothrips* Amyot & Serville, 1843 (32.81%), *Thrips* Linnaeus, 1758 (18.67%) and *Frankliniella* (17.59%) (fig. 3B). The species with most sequences in BOLD, *Taeniothrips inconsequens* (Uzel, 1895), represents almost 30% of all records in this database (fig. 3C). On the other hand, almost 70% of the species names have less than ten sequences each, and 27.6% of the species labels have a single COI sequence under their name (Supplementary file 7).

We also identified errors in at least 35 records, such as species labels with outdated names, typos, and even a sequence belonging to a beetle species mistakenly listed as a member of Thysanoptera. A complete list of the errors detected on the sequences obtained in November 2021 can be found in Supplementary File 8.

Geographical distribution data were available for 28,922 out of the 30,581 Thysanoptera sequences downloaded from BOLD (database 0), representing 69 countries which contributed with at least one sequence (fig. 4). Canada alone comprised about

**Table 4.** Number of Thysanoptera genera and species with at least one sequence after filtering step 2, compared to the number of genera and species currently accepted (following ThripsWiki 2023)

	Genera		Species	
	With sequences (valid)	% from valid	With sequences (valid)	% from valid
Suborder Terebrantia	75 (330)	22.72	191 (2615)	7.30
Family Aeolothripidae	8 (23)	34.78	19 (220)	8.64
Family Heterothripidae	1 (4)	25	2 (89)	2.25
Family Melanthripidae	2 (4)	50	2 (70)	2.86
Family Merothripidae	1 (3)	33.33	1 (18)	5.56
Family Stenurothripidae	1 (3)	33.33	1 (6)	16.67
Family Thripidae	62 (288)	21.53	166 (2206)	7.52
Subfamily Dendrothripinae	2 (13)	15.38	4 (111)	3.60
Subfamily Panchaethripinae	13 (42)	30.95	17 (146)	11.64
Subfamily Sericothripinae	3 (3)	100	13 (174)	7.47
Subfamily Thripinae	44 (230)	19.13	132 (1775)	7.44
Suborder Tubulifera	40 (458)	8.73	106 (3812)	2.78
Family Phlaeothripidae	40 (458)	8.73	106 (3812)	2.78
Subfamily Idolothripinae	5 (82)	6.10	10 (744)	1.34
Subfamily Phlaeothripinae	35 (376)	9.31	96 (3068)	3.13

40% of the total, with 11,756 sequences. The five countries with the most sequences (Canada, Costa Rica, South Africa, Australia and the United States) gather over 80% of the sequences (fig. 4A).

### Barcode gap

A total of 33 genera belonging to families Aeolothripidae, Phlaeothripidae and Thripidae could be evaluated for the presence and quality of Barcode gaps. Of these, 24 genera were classified as having a Good Barcode gap, four Intermediate, and five a Poor gap (fig. 5).

The median intraspecific distance varied widely among genera, with some presenting a median of 0% (i.e., *Franklinothrips* Back, 1912, *Hoplothrips* Amyot & Serville, 1843 and *Orothrips* Moulton, 1907), seven genera above 4%, and *Pseudodendrothrips* Schmutz, 1913 above 23%. The average median intraspecific distance was 2.49%.

The median interspecific distance also varied greatly among genera, with the lowest value being 3.44% for *Odontothrips* Amyot & Serville, 1843, and the highest value being 22.89% for *Pseudodendrothrips*. The average median interspecific distance was 13.27% (Table 5).

### Boxplot outliers

Among the 33 analysed genera, 22 exhibited outliers in the boxplots, indicating pairwise comparisons that fell at the extreme ends of the observed data range (fig. 5); and 19 genera had at least one outlier in the range listed by our R script (Table 6, Supplementary Files 4–5).

### Aeolothrips outliers

We found 2256 intraspecific outliers for *Aeolothrips*, of which 1727 outliers (those above the dotted line on fig. 2) were visually inspected. These outliers exclusively involved comparisons

between sequences of *Aeolothrips intermedius* Bagnall, 1934, which could be assigned to three distinct sequence clusters (Supplementary File 9).

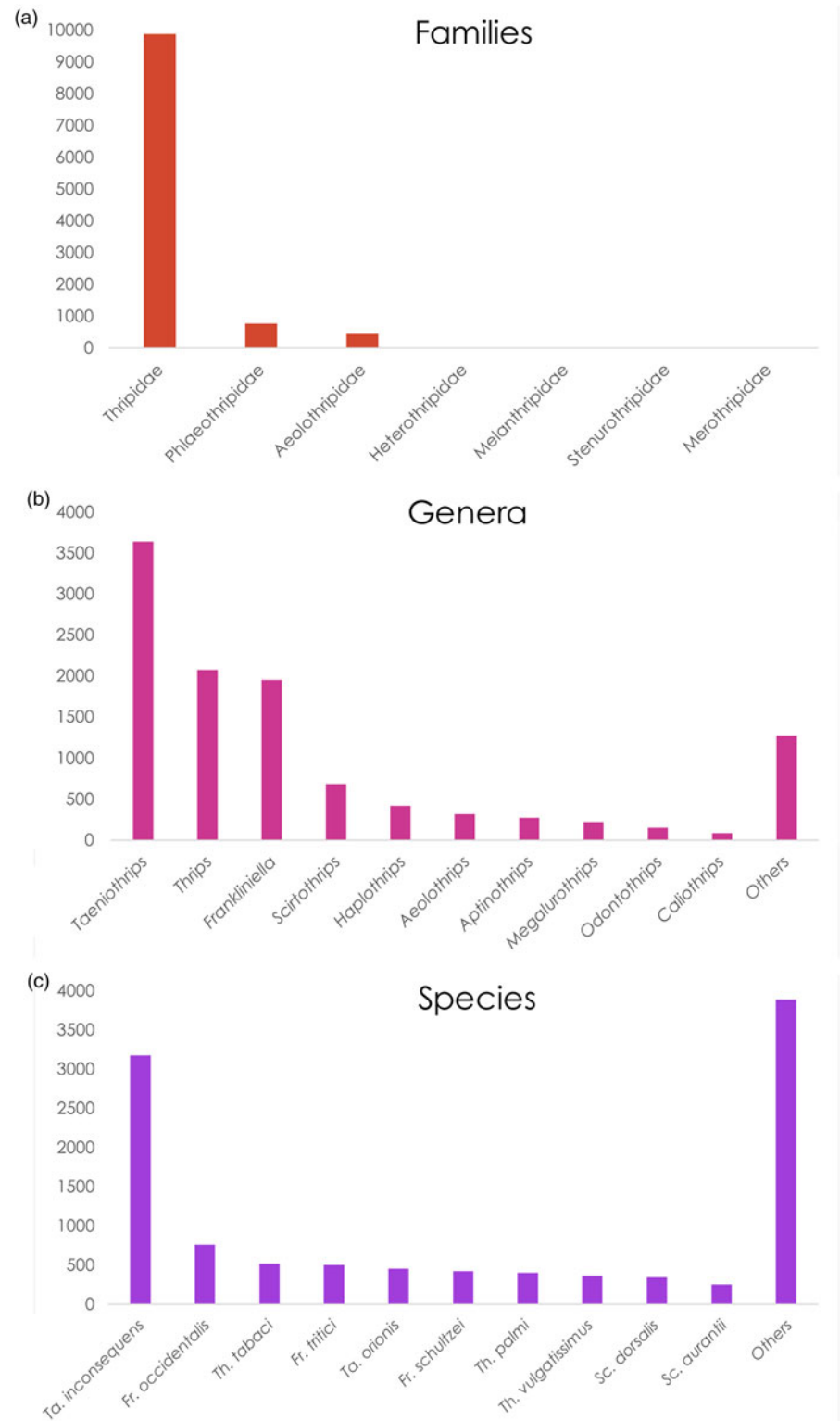
We also observed low interspecific distances involving some *Aeolothrips* sequences. The only sequence identified as *Aeolothrips melaleucus* Haliday, 1852 (BOLD ID: GBMIN39680-13) exhibited distances ranging from 0.17 to 1.73% when compared to sequences of *Aeolothrips fasciatus* (Linnaeus, 1758), whose highest observed intraspecific distance was 2.92%. Similarly, two entries of *Aeolothrips mongolicus* Pelikan, 1985 (BOLD ID: GBMIN91243-17 and GBMIN91244-17) displayed distances varying from 0.15 to 4.85% when compared to sequences of *A. intermedius*, and they even clustered with some *A. intermedius* sequences within the same BIN (BOLD:AAU0572).

### Frankliniella interspecific outliers

In the case of *Frankliniella*, a total of 16,717 interspecific outliers were identified, out of which 5403 (the ones which directly overlapped with the intraspecific boxplot) were considered for analysis. We observed outlier comparisons between five species pairs (Table 7). Moreover, the single sequence identified as *F. minuta* (Moulton, 1907) (BOLD ID: GBA8033-12) was identical to several sequences of *F. schultzei* (Trybom, 1910). Similarly, sequences labelled as *F. citripes* Hood, 1916 (BOLD ID: GBA8030-12) and *F. borinquen* Hood, 1942 (BOLD ID: GBMHT2007-19) exhibited very low distances when compared to *F. insularis* (Franklin, 1908) and *F. occidentalis* (Pergande, 1895), respectively. A complete list of the outlier comparisons between *Frankliniella* sequences is available in Supplementary File 9.

### PCI

The highest PCI value was observed for Aeolothripidae, with 83.33% of species labels allowing for 'correct' identifications



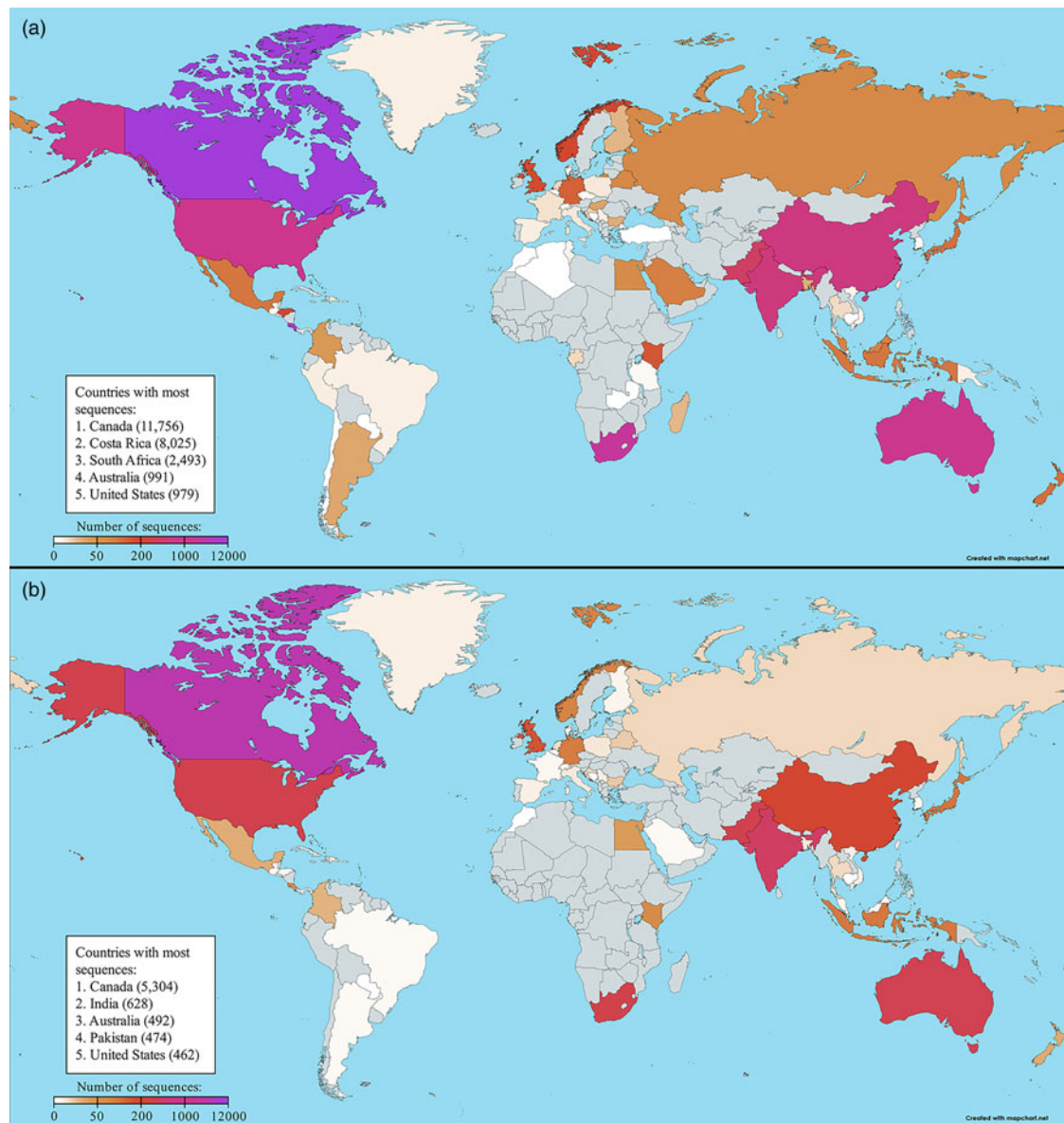
**Figure 3.** Distribution of Thysanoptera COI sequences (post filtering step 2) from BOLD, in different taxonomic levels. A: Family; B: Genus; C: Species.

(=maximum intraspecific distance < nearest neighbour distance). Meanwhile, the lowest value was observed for Thripidae, with 58.33% of species labels allowing for 'correct' identifications (fig. 6). The complete list of species names evaluated, and their maximum intraspecific and nearest neighbour distance values, can be found in Supplementary File 10.

## Discussion

### *Thysanoptera data on BOLD*

While there were over 30,000 sequences available on BOLD for Thysanoptera in November of 2021, only about a third of them matched the criteria to be included in the Barcode gap and PCI



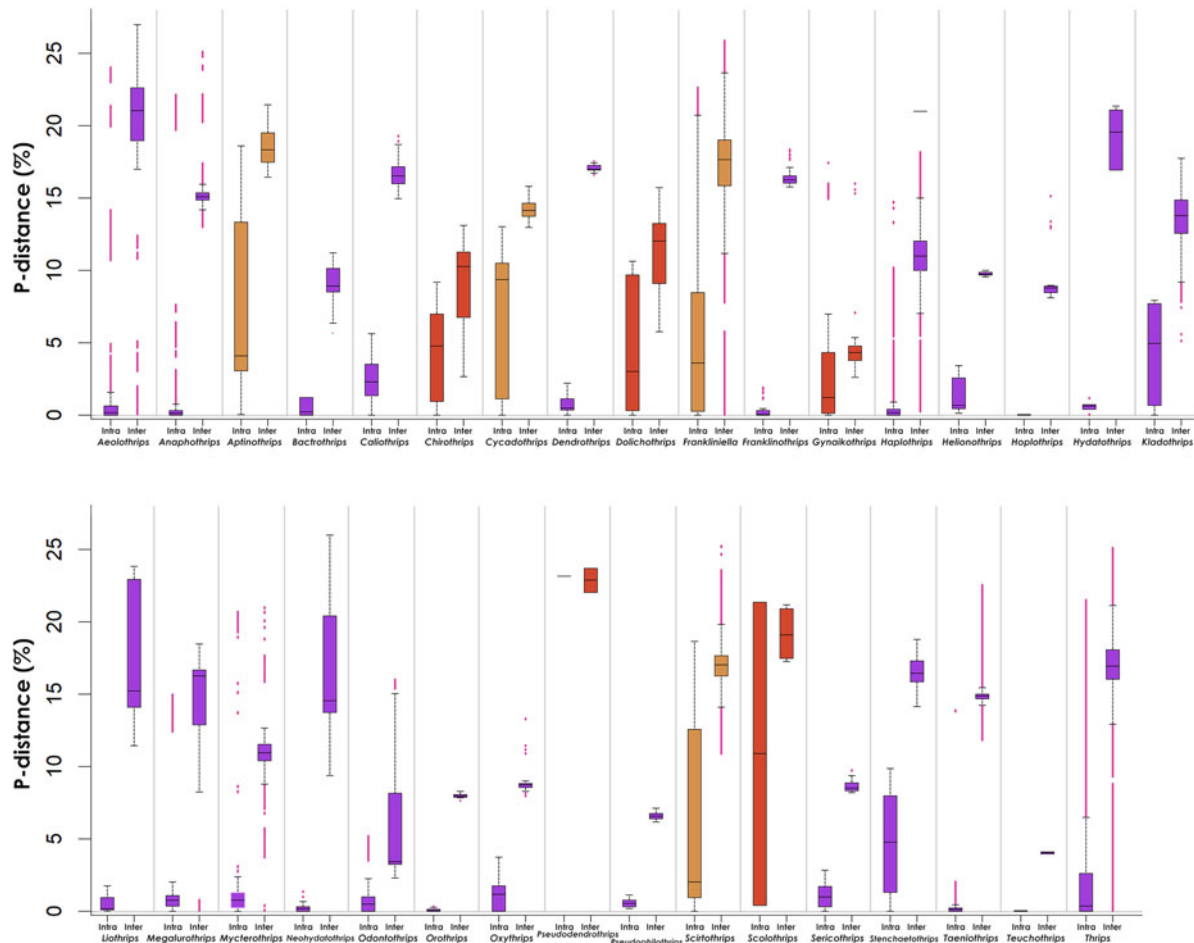
**Figure 4.** Heat maps showing the countries with most (purple) to least (white) Thysanoptera sequences added to BOLD. A: Geographical distribution of all Thysanoptera sequences downloaded from BOLD, before filtering (database 0). B: Geographical distribution of remaining Thysanoptera sequences, after all filtering steps (database 3). Countries in grey have zero sequences in the evaluated database. Image created with MapChart, available at <https://www.mapchart.net>.

analyses performed in this work; moreover, the sequences we utilised have a very limited representativity for Thysanoptera genera and species (about 15 and 5% of valid taxa, respectively). This is similar to what was observed for Pentatomomorpha (about 6% of valid species; Bianchi and Gonçalves, 2021a) and Orthoptera (about 3% of valid species; Timm *et al.*, 2022), but much less than what is available for insect groups with a higher focus on molecular studies, such as Apidae (around 17% of valid species; Gonçalves *et al.*, 2022) and Lepidoptera (almost two-thirds of valid species; Mutanen *et al.*, 2016). The usage of COI in Thysanoptera is usually focused on identification or population studies of a few pest species (e.g., Leão *et al.*, 2017; Chakraborty *et al.*, 2019; further references in Ghosh *et al.*, 2021).

Despite fungivorous thrips species representing about 50% of the current diversity in the order, most of them lack sequences in BOLD. For example, there is no molecular information for

the single extant species of Uzelothripidae, whose relationships within the order are still unknown. Subfamily Idolothripinae, which is the only group of thrips able to ingest and process whole fungal spores, is also underrepresented in BOLD.

Many sequenced specimens also lack any image records, and the available digital photographs were taken on a stereomicroscope or without enough magnification to examine thrips morphological traits. While it is possible to identify potential species units by utilising only the molecular data, for many taxa, including thrips, most species are still defined only by morphological traits. A good molecular library can work independently from morphology data, but we are still very far from using the BOLD database as a reliable identification tool for Thysanoptera. The lack of good quality pictures associated with the sequences or even voucher specimens hinders the possibility of reviewing and correcting potential misidentification. The lack of sequence metadata supporting



**Figure 5.** Boxplot graphs for 33 Thysanoptera genera, showing distribution of intraspecific and interspecific distances. Genera considered Good are in purple, Intermediate in orange and Poor in red. Pairwise comparisons which escape the area of significance indicated by the whiskers ( $\pm 1.5$  IQR), considered outliers, are represented in pink.

taxonomic identification plays against the basic scientific principle of reproducibility (Bianchi and Gonçalves, 2021b) and compromises the utility of reference sequences.

While many countries contributed sequences to BOLD's Thysanoptera database, most of these sequences are concentrated in a small number of countries (fig. 4). Most of Africa and several countries in Asia, Europe and Latin America do not have any data added to BOLD. After filtering our data, all sequences from 15 countries were removed, and the remaining data are even more concentrated on Canada, the largest source of sequenced thrips specimens.

### Barcode gap

A Good Barcode gap was observed for most of the genera, allowing species identification for these taxa based on this gap. However, many of them also had multiple outliers, which could potentially cloud identification efforts by increasing the observed intraspecific and interspecific ranges, creating overlaps. While the median intraspecific distance was below 1% in most genera classified as Good, both *Kladothrips* Froggatt, 1906 and *Stenchaetothrips* Bagnall, 1926 had a median intraspecific distance above 4%. If one were to use an arbitrary cut-off value to separate sequences into species for these groups (e.g., 2–3%; Hebert *et al.*, 2003), they would split a single species into different names. We

recommend caution in using arbitrary distance values for thrips species delimitations without a proper sampling and previous evaluation of the intraspecific diversity of the target group.

The median intraspecific distance in genera with Intermediate or Poor gaps was high in comparison to those genera with Good gaps, and in those cases the Barcode gap may not be a reliable tool for species delimitation. Within the genera with Intermediate Barcode gap, *Frankliniella* and *Scirtothrips* Shull, 1909 have a high number of species distributed worldwide (236 and 108, respectively; ThripsWiki 2023) as well as complex taxonomy (Mound and Palmer, 1981; Cavalleri and Mound, 2012), and some potential cryptic species (Rugman-Jones *et al.*, 2010; Dickey *et al.*, 2015).

### Boxplot outliers

The Barcode gap analysis resulted in frequent outlier comparisons, which can demonstrate the necessity of re-examining the sequences involved or even a taxonomic revision of some groups, especially when there are overlaps between intraspecific and interspecific distances.

For *A. intermedius*, one sequence (BOLD ID: GBMNC48112-20) had very high distances (above 20%) when compared to most the other sequences identified as *A. intermedius* (Supplementary File 9), suggesting this sequence is not conspecific with the



**Table 5.** Barcode gap category and median intra- and interspecific distances of the evaluated genera

Genus	Category	Median intraspecific distance (%)	Median interspecific distance (%)
<i>Aeolothrips</i>	Good	0.17	21.03
<i>Anaphothrips</i>	Good	0.17	15.09
<i>Aptinothrips</i>	Intermediate	4.12	18.36
<i>Bactrothrips</i>	Good	0.26	8.98
<i>Caliothrips</i>	Good	2.29	16.51
<i>Chirothrips</i>	Poor	4.80	10.28
<i>Cycadothrips</i>	Intermediate	9.29	14.10
<i>Dendrothrips</i>	Good	0.46	16.90
<i>Dolichothrips</i>	Poor	3.06	12.04
<i>Frankliniella</i>	Intermediate	3.58	17.65
<i>Franklinothrips</i>	Good	0	16.26
<i>Gynaikothrips</i>	Poor	1.24	4.34
<i>Haplothrips</i>	Good	0.15	10.97
<i>Helionothrips</i>	Good	0.70	9.70
<i>Hoplothrips</i>	Good	0	8.79
<i>Hydatothrips</i>	Good	0.70	19.55
<i>Kladothrips</i>	Good	4.89	13.77
<i>Liothrips</i>	Good	0.20	15.25
<i>Megalurothrips</i>	Good	0.77	16.25
<i>Mycterothrips</i>	Good	0.78	10.97
<i>Neohydatothrips</i>	Good	0.15	14.51
<i>Odontothrips</i>	Good	0.48	3.44
<i>Orothrips</i>	Good	0	7.90
<i>Oxythrips</i>	Good	1.20	8.86
<i>Pseudodendrothrips</i>	Poor	23.15	22.87
<i>Pseudophilothrips</i>	Good	0.54	6.59
<i>Scirtothrips</i>	Intermediate	1.99	16.98
<i>Scolothrips</i>	Poor	10.90	19.12
<i>Sericothrips</i>	Good	0.94	8.52
<i>Stenchaetothrips</i>	Good	4.79	16.51
<i>Taeniothrips</i>	Good	0.16	14.87
<i>Teuchothrips</i>	Good	0	4.09
<i>Thrips</i>	Good	0.33	16.95

other *A. intermedius* specimens. The other two observed sequence clusters also separate into different BINs (Group 1 = BOLD: ACD4587; Group 2 = BOLD:AAZ8618 and BOLD:AAU0572; see Supplementary File 9 for full composition of these groups), which indicates that what is currently identified morphologically as *A. intermedius* may represent three or four distinct species when utilising this COI fragment as reference. Tyagi *et al.* (2017) found support for two species within *A. intermedius* collected from India, when conducting single-locus delimitation. We also observed that the single sequence of *A. melaleucus* (BOLD ID: GBMIN39680-13) and the two sequences of *A. mon-golicus* (BOLD ID: GBMIN91243-17 and GBMIN91244-17) need

revision, as they may represent specimens of *A. fasciatus* and *A. intermedius*, respectively. None of these specimens have photos on BOLD, so we are unable to compare their morphologies to see if they match the identity suggested by molecular data.

Regarding *Frankliniella*, we suggest that at least the sequences GBMHT2007-19, GBA8030-12, and GBA8033-12 (labelled *F. borinquen*, *F. citripes* and *F. minuta*, respectively) are misidentified. Unfortunately, there are no available images of these records to verify their identity.

Taxonomic incongruencies may be the most probable explanation for many of the observed high intraspecific distances and outliers. Misidentification of thrips species is frequent, especially

**Table 6.** List of genera with observed outliers in the boxplot graphs

Genus	Type of outlier			
	Below intraspecific boxplot <sup>a</sup>	Above intraspecific boxplot <sup>b</sup>	Below interspecific boxplot <sup>b</sup>	Above interspecific boxplot <sup>a</sup>
<i>Aeolothrips</i>		x	x	
<i>Anaphothrips</i>		x	x	x
<i>Bactrothrips</i>			x	
<i>Caliothrips</i>				x
<i>Dendrothrips</i>			x	x
<i>Frankliniella</i>		x	x	x
<i>Franklinothrips</i>		x		x
<i>Gynaikothrips</i>		x	x	x
<i>Haplothrips</i>		x	x	x
<i>Hoplothrips</i>				x
<i>Hydatothrips</i>	x	x		
<i>Kladothrips</i>			x	
<i>Megalurothrips</i>		x	x	
<i>Mycterothrips</i>		x	x	x
<i>Neohydatothrips</i>		x		
<i>Odontothrips</i>		x		x
<i>Orothrips</i>		x	x	
<i>Oxythrips</i>			x	x
<i>Scirtothrips</i>			x	x
<i>Sericothrips</i>				x
<i>Taeniothrips</i>		x	x	x
<i>Thrips</i>		x	x	x

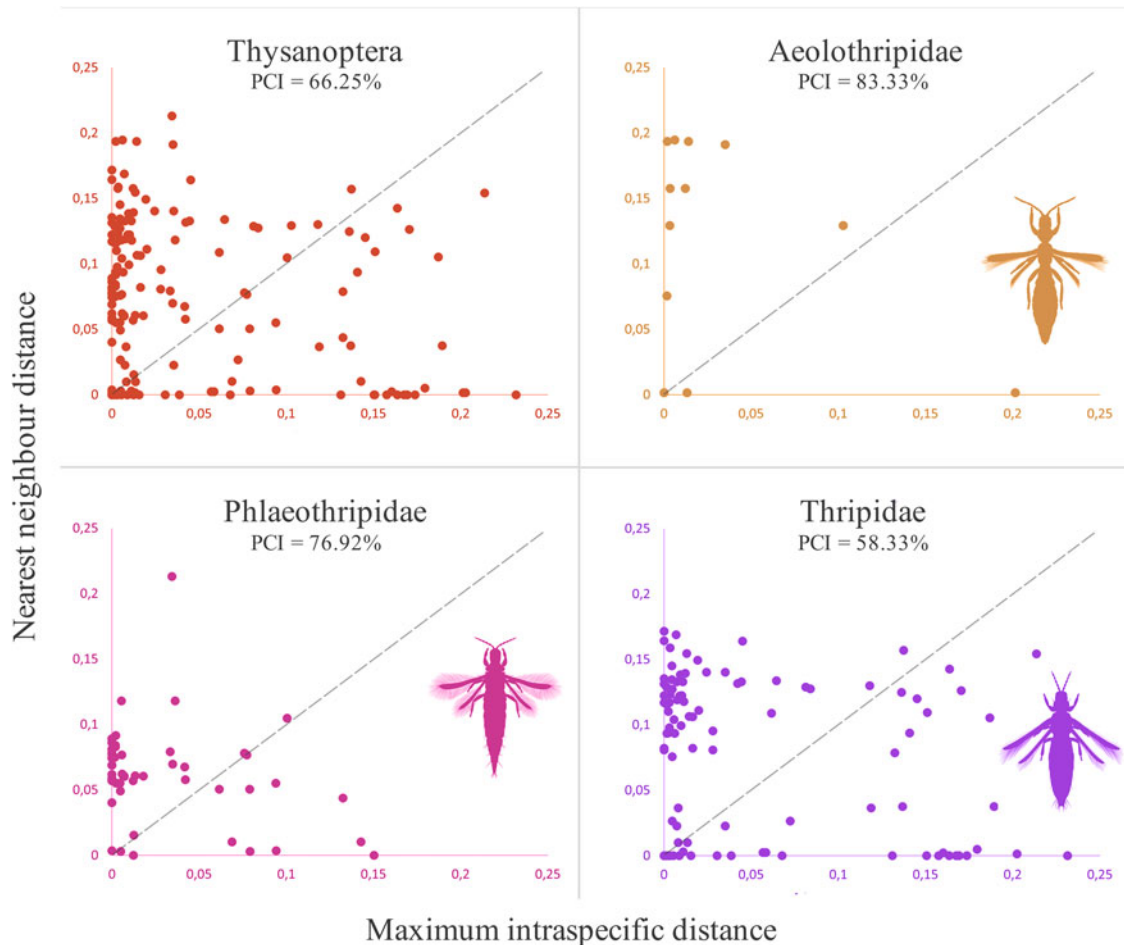
<sup>a</sup>Outliers not listed by the R script.<sup>b</sup>Outliers listed by the R script.

in groups with high reliance on minute and similar looking morphological characters, such as *Frankliniella* and *Scirtothrips*. Alternatively, cryptic species (i.e., when distinct species are lumped under the same name due to a lack of morphological, ecological or biological distinction) could also explain a high intraspecific variation, due to molecular divergence that has not

been translated into phenotypic differences yet (Struck *et al.*, 2018; Struck and Cerca De Oliveira, 2019). However, we cannot discard the possibility of the taxonomy being correct, and the high intraspecific variation in COI being explained by other underlying causes. Geographic distribution and events can have an influence, by allowing or limiting contact and genetic exchange

**Table 7.** *Frankliniella* species with analysed interspecific outlier comparisons

Species	N sequences involved in outliers	N sequences	Compared to	N comparisons	Average distance (%)
<i>F. bispinosa</i>	16	18	<i>F. tritici</i>	4530	4.544
<i>F. borinquen</i>	1	17	<i>F. occidentalis</i>	642	2.272
<i>F. citripes</i>	1	1	<i>F. insularis</i>	39	0.316
<i>F. insularis</i>	39	41	<i>F. citripes</i>	39	0.316
<i>F. minuta</i>	1	1	<i>F. schultzei</i>	164	0
<i>F. occidentalis</i>	642	762	<i>F. borinquen</i> , <i>F. panamensis</i>	670	2.519
<i>F. panamensis</i>	3	26	<i>F. occidentalis</i>	28	8.188
<i>F. schultzei</i>	164	424	<i>F. minuta</i>	164	0
<i>F. tritici</i>	284	505	<i>F. bispinosa</i>	4530	4.544



**Figure 6:** Probability of correct identification analysis for order Thysanoptera, and for Aeolothripidae, Phlaeothripidae and Thripidae. Each dot represents a species, and dots above the dashed line represent species where a correct identification, using the evaluated COI sequences as reference, would be possible (maximum intraspecific distance < nearest neighbour distance).

between different populations. The presence of parasites able to affect the host's reproduction, such as *Wolbachia* bacteria, could also influence the genetic composition of a species (Xiao *et al.*, 2012). Further studies can explore in more detail these or other potential explanations to the observed variation, but it is important to consider all available hypotheses and test them when reviewing the highly diverging sequences.

### PCI

The PCI analysis indicates that in over 30% of the cases, identifying Thysanoptera species using the sequences as a reference library could lead to incorrect names, if using the 'nearest neighbour' distance value as a cut-off. This is worrisome especially for Thripidae, the second largest family within the order and the one with the most sequences in BOLD: more than 40% of the species names analysed returned as 'incorrect' identifications. Furthermore, many of the Thripidae species labels with intraspecific and interspecific distances overlapping belong to large genera, with complex taxonomy (e.g., *Frankliniella*, *Scirtothrips*, *Thrips*). This could support the hypothesis of incorrect identifications of some reference specimens included in BOLD, but the possibility that multiple cryptic species may be under the same name cannot be discarded (Rebijith *et al.*, 2014; Dickey *et al.*, 2015; Tyagi *et al.*, 2017; see discussion above for other potential causes).

Curiously, *Haplothrips* Amyot & Serville, 1843 (PCI = 66.67%) and *Thrips* (PCI = 45.45%), despite their low PCI values, were both considered Good in the Barcode gap analysis, although with many outlier comparisons each. This suggests that most of the sequences within these genera have low enough distances for observing a clear Barcode gap between species; however, the PCI analysis can detect when there are a single or few sequences with a high intraspecific distance or low interspecific distance to another sequence.

The PCI analysis does not identify the causes for 'incorrect identifications' but can be used to detect taxa with a low percentage of correct identifications, which can then be further explored to identify such causes. A few potential causes for the 'incorrect identifications' include identification errors in the reference sequences, taxonomic incongruencies, human error during DNA extraction, sequencing or upload to databases, among others (Mutanen *et al.*, 2016).

### Conclusion

Undoubtedly BOLD serves as a valuable tool for various molecular studies, offering a freely accessible COI sequence library for many taxa and enabling specimen identification and species delimitation. However, caution is advised when using BOLD data, particularly for Thysanoptera, as the representativity of

thrips species in the database is low, with the majority lacking COI data. Additionally, the sampling effort has been limited to specific regions, restricting the usefulness of BOLD as a reference database for many geographical areas. Our analysis revealed a clear Barcode gap for most genera, yet numerous potential misidentifications and cryptic diversity were identified. We propose prioritising non-destructive DNA extraction methods and improving the photographic record to enhance taxonomic analysis. The hardest part – creating a global and freely accessible database of Barcode data – is done. It is up to us, researchers who use this database and populate it with new data, to work on identifying and correcting the inconsistencies and limitations currently present in BOLD, so that it can reach its full potential as a DNA-based species identification tool.

**Supplementary material.** The supplementary material for this article can be found at <https://doi.org/10.1017/S0007485323000391>.

**Acknowledgements.** MFL has received a doctorate fellowship from Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) during the development of this work. LTG was supported by a doctorate fellowship from Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

**Competing interests.** None.

## References

- Badotti, F., de Oliveira, F.S., Garcia, C F., Vaz, A.B.M., Fonseca, P.L.C., Nahum, L.A., Oliveira, G. and Góes-Neto, A. (2017) Effectiveness of ITS and sub-regions as DNA Barcode markers for the identification of Basidiomycota (Fungi). *BMC Microbiology* 17, 42.
- Bianchi, F.M. and Gonçalves, L.T. (2021a) Borrowing the Pentatomomorpha tome from the DNA Barcode library: scanning the overall performance of *cox1* as a tool. *Journal of Zoological Systematics and Evolutionary Research* 59, 992–1012.
- Bianchi, F.M. and Gonçalves, L.T. (2021b) Getting science priorities straight: how to increase the reliability of specimen identification? *Biology Letters* 17, 1–4.
- BOLD - Barcode of Life Data System V4** (2023). Available at <https://www.boldsystems.org/index.php>. Last accessed in May 2023.
- Cavalleri A and Mound LA (2012) Toward the identification of *Frankliniella* species in Brazil (Thysanoptera, Thripidae). *Zootaxa* 3270, 1–30.
- Chakraborty, R., Singha, D., Kumar, V., Pakrashi, A., Kundu, S., Chandra, K., Patnaik, S. and Tyagi, K. (2019) DNA Barcoding of selected *Scirtothrips* species (Thysanoptera) from India. *Mitochondrial DNA Part B* 4, 2710–2714.
- Collins, R.A. and Cruickshank, R.H. (2012) The seven deadly sins of DNA Barcoding. *Molecular Ecology Resources* 13, 969–975.
- Collins, R.A., Boykin, L.M., Cruickshank, R.H. and Armstrong, K.F. (2012) Barcoding's next top model: an evaluation of nucleotide substitution models for specimen identification. *Methods in Ecology and Evolution* 3, 457–465.
- Dickey, A.M., Kumar, V., Hoddle, M.S., Funderburk, J.E., Morgan, J.K., Jara-Cavieses, A., Shatters, R.G. Jr., Osborne, L.S. and McKenzie, C.L. (2015) The *Scirtothrips dorsalis* species complex: endemism and invasion in a global pest. *PLoS ONE* 10, e0123747.
- Erickson, D.L., Spouge, J., Resch, A., Weigt, L.A. and Kress, J.W. (2008) DNA Barcoding in land plants: developing standards to quantify and maximize success. *Taxon* 57, 1304–1316.
- Ghosh, A., Jangra, S., Dietzgen, R.G. and Yeh, W.-B. (2021) Frontiers approaches to the diagnosis of Thrips (Thysanoptera): how effective are the molecular and electronic detection platforms? *Insects* 12, 1–26.
- Gonçalves, L.T., Bianchi, F.M., Deprá, M. and Calegari-Marques, C. (2021) Barcoding a can of worms: testing *cox1* performance as a DNA Barcode of Nematoda. *Genome* 64, 705–717.
- Gonçalves, L.T., Francoso, E. and Deprá, M. (2022) Shorter, better, faster, stronger? Comparing the identification performance of full-length and mini-DNA Barcodes for apid bees (Hymenoptera: Apidae). *Apidologie* 53, 55. <https://doi.org/10.1007/s13592-022-00958-x>
- Hebert, P.D.N., Cywinska, A., Ball, S.L. and deWaard, J.R. (2003) Biological identifications through DNA Barcodes. *Proceedings of the Royal Society B* 270, 313–321.
- Ifitikhar, R., Ashfaq, M., Rasool, A. and Hebert, P.D.N. (2016) DNA Barcode analysis of thrips (Thysanoptera) diversity in Pakistan reveals cryptic species complexes. *PLoS ONE* 11, e0146014, 1–21.
- Karimi, J., Hassani-Kakhki, M. and Awal, M.M. (2010) Identifying thrips (Insecta: Thysanoptera) using DNA Barcodes. *Journal of Cell and Molecular Research* 2, 35–41.
- Katoh, K., Rozewicki, J. and Yamada, K.D. (2019) MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. *Briefings in Bioinformatics* 20, 1160–1166.
- Leão, E.U., Spadotti, D.M.A., Rocha, K.C.G., Lima, E.F.B., Tavella, L., Turina, M. and Krause-Sakate, R. (2017). Efficient detection of *Frankliniella schultzei* (Thysanoptera, Thripidae) by cytochrome oxidase I gene (mtCOI) direct sequencing and real-time PCR. *Brazilian Archives of Biology and Technology* 60, e17160425.
- Lis, J.A., Lis, B. and Ziaja, D.J. (2016) In BOLD we trust? A commentary on the reliability of specimen identification for DNA Barcoding: a case study on burrower bugs (Hemiptera: Heteroptera: Cydnidae). *Zootaxa* 4114, 83–86.
- Marullo, R., Mercati, F. and Vono, G. (2020) DNA Barcoding: a reliable method for the identification of thrips species (Thysanoptera, Thripidae) collected on sticky traps in onion fields. *Insects* 11, 1–10.
- Meiklejohn, K.A., Damaso, N. and Robertson, J.M. (2019) Assessment of BOLD and GenBank – their accuracy and reliability for the identification of biological materials. *PLoS ONE* 14, e0217084.
- Mound LA and Palmer JM (1981) Identification, distribution and host-plants of the pest species of *Scirtothrips* (Thysanoptera: Thripidae). *Bulletin of Entomological Research* 71, 467–479.
- Mutanen, M., Kivelä, S.M., Vos, R.A., Doorenweerd, C., Ratnasingham, S., Hausmann, A., Huemer, P., Dincă, V., van Nieuwerkerken, E.J., Lopez-Vaamonde, C., Vila, R., Aarvik, L., Decaëns, T., Efetov, K.A., Hebert, P.D.N., Johnsen, A., Karsholt, O., Pentinsaari, M., Rougerie, R., Segerer, A., Tarmann, G., Zahir, R. and Godfray, H.C.J. (2016) Species-level para- and polyphyly in DNA barcode gene trees: strong operational bias in European Lepidoptera. *Systematic Biology* 65, 1024–1040.
- Paradis, E. and Schliep, K. (2019) ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics (Oxford, England)* 35, 526–528.
- Pentinsaari, M., Ratnasingham, S., Miller, S.E. and Hebert, P.D.N. (2020) BOLD and GenBank revisited – Do identification errors arise in the lab or in the sequence libraries? *PLoS ONE* 15, e0231814.
- Ratnasingham, S. and Hebert, P.D.N. (2007) BOLD: the barcode of life data system ([www.barcodinglife.org](http://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.
- Rebijith KB, Asokan R, Krishna V, Ranjitha HH, Krishna Kumar NK and Ramamurthy VV (2014) DNA Barcoding and elucidation of cryptic diversity in thrips (Thysanoptera). *Florida Entomologist* 97, 1328–1347.
- Rugman-Jones, P.F., Hoddle, M.S. and Stouthamer, R. (2010) Nuclear-mitochondrial barcoding exposes the global pest Western flower thrips (Thysanoptera: Thripidae) as two sympatric cryptic species in its native California. *Journal of Economic Entomology* 103, 877–886.
- Sonet, G., Jordaens, K., Braet, Y., Bourguignon, L., Dupont, E., Backeljau, T., De Meyer, M. and Desmyter, S. (2013) Utility of GenBank and the Barcode of Life Data Systems (BOLD) for the identification of forensically important Diptera from Belgium and France. *Zookeys* 365, 307–328.
- Srivathsan, A. and Meier, R. (2012) On the inappropriate use of Kimura-2-parameter (K2P) divergences in the DNA-barcoding literature. *Cladistics* 28, 190–194.
- Struck, T.H. and Cerca De Oliveira, J. (2019) Cryptic species and their evolutionary significance. *Encyclopedia of Life Sciences*, 1–9. <https://doi.org/10.1002/9780470015902.a0028292>

- Struck, T.H., Feder, J.L., Bendiksby, M., Birkeland, S., Cerca, J., Gusarov, V.I. and ...Dimitrov, D. (2018). Finding evolutionary processes hidden in cryptic species. *Trends in Ecology & Evolution*, **33**, 153–163.
- ThripsWiki (2023) ThripsWiki - providing information on the World's thrips. Available from: [http://thrips.info/wiki/Main\\_Page](http://thrips.info/wiki/Main_Page) (Accessed May 2023).
- Timm, V.F., Gonçalves, L.T., Valente, V.L.D.S. and Deprá, M. (2022) The efficiency of the COI gene as a DNA Barcode and an overview of Orthoptera (Caelifera and Ensifera) sequences in the BOLD System. *Canadian Journal of Zoology* **100**, 710–718.
- Tyagi, K., Kumar, V., Singha, D., Chandra, K., Laskar, B.A., Kundu, S., Chakraborty, R. and Chatterjee, S. (2017) DNA Barcoding studies on thrips in India: Cryptic species and species complexes. *Scientific Reports* **7**, 1–14.
- Vink, C.J., Paquin, P. and Cruickshank, R.H. (2012) Taxonomy and irreproducible biological science. *BioScience* **62**, 451–452.
- Xiao, J.H., Wang, N.X., Murphy, R.W., Cook, J., Jia, L.Y. and Huang, D.W. (2012) *Wolbachia* infection and dramatic intraspecific mitochondrial DNA divergence in a fig wasp. *Evolution* **66**, 1907–1916.