The DNA barcoding project on German Diptera: An appreciative and critical analysis with four suggestions for improving the development and reliability of DNA-based identification

MARION KOTRBA

SNSB-Zoologische Staatssammlung München, Münchhausenstraße 21, D-81247 München, Germany; e-mail: kotrba@snsb.de

Key words. Diptera, DNA barcoding, GBOL, Germany, BOLD, biodiversity, identification

Abstract. The progress in constructing a DNA barcode library for German Diptera as published by Morinière et al. (2019, Mol. Ecol. Resour. 19: 900–928) is appraised from a dipterists’ perspective. The coverage of the diversity of German Diptera in terms of barcode index numbers (BINs) and identified barcodes is analysed and visualized in simple diagrams. The influence of the project setup, methodology and/or systematic effects on the emerging numbers and trends is elucidated and extensively discussed. In addition, the documentation on the species identification methods in the database is assessed. Based on this evaluation, four ways for improving the future development, utility and reliability of this DNA database and similar projects in general are identified: (1) Sample the collections of experts. This results in a greater and more reliable coverage within a limited time frame, as opposed to random collecting and relying on a posteriori identification. (2) Give priority to medically, agriculturally or ecologically important families. Addressing these gaps will meet the most pressing needs of the community and serve as a good advertisement for the usefulness and wide applicability of the DNA barcode library. (3) Allocate resources to recruiting established experts as opposed to trainees. The fact that half of the recovered BINs remained unidentified mostly results from the insufficient involvement of experts (and expert time). (4) Appropriately document the morphological identifications by experts in the database. This will allow to assess the reliability of DNA-based identifications and to prioritize conflicting identifications within a BIN accordingly.

INTRODUCTION

Morinière et al. (2019) published an extensive DNA barcode library for German Diptera with results from two major barcoding projects, the “Barcoding Fauna Bavarica” project (Hendrich et al., 2010) and the “German Barcode of Life” project (Geiger et al., 2016). The data set published includes 40,753 records and roughly 5,200 BINs (barcode index numbers) of 2,453 named species and 2,700 “dark taxa”, i.e. BINs that are unidentified. In their paper, the authors propose their DNA barcode library as an “intermediate taxonomic system” that will provide a foundation for subsequent taxonomic and biodiversity studies, but they also address problematic issues, such as dark taxa and the “taxonomic impediment”.

Undoubtedly DNA-based identification, i.e. the identification of specimens by comparing their DNA barcode to other, already identified DNA barcodes, is becoming an important and powerful tool for applied projects, such as assessing and comparing biodiversity patterns (Ratnsingham & Hebert, 2013; Morinière et al., 2019). Some methodological shortcomings, such as BIN sharing etc., are inherent in the system (Meier & Zhang, 2009; Ratnsingham & Hebert, 2013; Morinière et al., 2019) and are likely to prevent achieving an identification level of 100%, even if a complete and fully identified DNA barcode library was available. But these limitations may well be acceptable in a large range of studies and are not the focus of the present evaluation. Likewise, this review does not aim to discuss the claims and views brought forward in the introduction and discussion of the evaluated paper. Instead, it seeks to appraise the progress in constructing a DNA barcode library for German Diptera from a dipterists’ perspective and to identify ways to promote its future progress, utility and reliability. In these regards, the original paper does not fully evaluate the published data. Additional insights on this particular project, but also on the operation and evaluation of similar projects in general, can be expected from a more detailed scrutiny. For this purpose, the achieved coverage of the diversity of German Diptera in terms of recovered BINs and identified species, and the documentation of the respective species identifications, are statistically analysed and visualized in simple diagrams.

MATERIALS AND METHODS

This analysis is exclusively based on the information and data (including identification) published in Morinière et al. (2019, hereafter referred to as “source study”) without any additional information on the objectives or the subsequent development of this
Family | Species reported in Germany | BINs | Identified species | BIN Ratio | Identified. Ratio | Identified. Ratio
---|---|---|---|---|---|---
Acaridae | 2 | 1 | 2 | 0.50 | 2.00 | 1.00
Acrididae | 11 | 0 | 0 | 0.00 | N/A | 0.00
Agromyzidae | 552 | 218 | 65 | 0.39 | 0.30 | 0.12
Anisopodidae | 8 | 7 | 5 | 0.88 | 0.71 | 0.63
Anthomyiidae | 227 | 188 | 114 | 0.63 | 0.60 | 0.50
Antheridiidae | 14 | 5 | 5 | 0.36 | 1.00 | 0.36
Aletidae | 81 | 18 | 12 | 0.22 | 0.67 | 0.15
Aetidae | 7 | 3 | 3 | 0.43 | 1.00 | 0.43
Aletidae | 3 | 3 | 3 | 1.00 | 1.00 | 1.00
Athericeridae | 5 | 3 | 2 | 0.60 | 0.67 | 0.40
Aulacigasteridae | 1 | 0 | 0 | 0.00 | N/A | 0.00
Bibionidae | 21 | 12 | 8 | 0.57 | 0.67 | 0.38
Blaphegeriidae | 7 | 2 | 1 | 0.29 | 0.50 | 0.14
Bolitophilidae | 22 | 14 | 6 | 0.64 | 0.43 | 0.27
Bombyliidae | 40 | 6 | 5 | 0.15 | 0.83 | 0.13
Braulidae | 1 | 0 | 0 | 0.00 | N/A | 0.00
Calliphoridae | 62 | 35 | 33 | 0.56 | 0.94 | 0.53
Camelidae | 4 | 0 | 0 | 0.00 | N/A | 0.00
Canidae | 12 | 9 | 8 | 0.75 | 0.89 | 0.67
Cathalophilidae | 1 | 0 | 0 | 0.00 | N/A | 0.00
Camelidae | 11 | 7 | 6 | 0.64 | 0.00 | 0.00
Cecidomyiidae | 836 | 927 | 44 | 1.11 | 0.05 | 0.05
Ceratopogonidae | 332 | 131 | 31 | 0.39 | 0.24 | 0.09
Chaenopidae | 29 | 17 | 4 | 0.59 | 0.24 | 0.14
Chaoboridae | 7 | 2 | 2 | 0.29 | 1.00 | 0.29
Chironomidae | 696 | 455 | 152 | 0.65 | 0.33 | 0.22
Chloropidae | 198 | 101 | 42 | 0.51 | 0.42 | 0.21
Chytromyzidae | 5 | 2 | 2 | 0.40 | 1.00 | 0.40
Clypeidae | 9 | 6 | 4 | 0.67 | 0.67 | 0.67
Coeplidae | 2 | 0 | 0 | 0.00 | N/A | 0.00
Coneomyiidae | 1 | 0 | 0 | 0.00 | N/A | 0.00
Conopidae | 52 | 9 | 9 | 0.17 | 1.00 | 0.17
Cremastocheidae | 1 | 0 | 0 | 0.00 | N/A | 0.00
Cryptocheilidae | 1 | 0 | 0 | 0.00 | N/A | 0.00
Culicidae | 46 | 8 | 7 | 0.17 | 0.88 | 0.15
Cylindromyidae | 4 | 1 | 1 | 0.25 | 1.00 | 0.25
Dacnidae | 4 | 3 | 3 | 0.75 | 1.00 | 0.75
Diastatidae | 9 | 8 | 6 | 0.89 | 0.75 | 0.67
(Diptera) | 7 | 5 | 5 | 1.00 | 1.00 | 1.00
Diptera | 4 | 1 | 1 | 0.25 | 1.00 | 0.25
Dixidae | 16 | 4 | 3 | 0.25 | 0.75 | 0.19
Dolichopodidae | 362 | 112 | 55 | 0.31 | 0.49 | 0.15
(Hemiptera) | 59 | 26 | 23 | 0.47 | 0.82 | 0.39
Drosophilidae | 5 | 3 | 1 | 0.57 | 0.50 | 0.50
Eugeniidae | 1 | 0 | 0 | 0.00 | N/A | 0.00
Empididae | 383 | 161 | 54 | 0.42 | 0.34 | 0.14
Ephyridae | 177 | 130 | 116 | 0.73 | 0.89 | 0.66
Fannidae | 56 | 46 | 31 | 0.82 | 0.87 | 0.55
Gasterophilidae | 4 | 0 | 0 | 0.00 | N/A | 0.00
Helcormyidae | 3 | 0 | 0 | 0.00 | N/A | 0.00
Heliozidae | 74 | 58 | 29 | 0.78 | 0.50 | 0.39
(Hemiptera) | 1 | 0 | 0 | 0.00 | N/A | 0.00
Hesperidae | 1 | 0 | 0 | 0.00 | N/A | 0.00
Hilarimorphidae | 2 | 0 | 0 | 0.00 | N/A | 0.00
Hippoboscidae | 20 | 7 | 6 | 0.35 | 0.86 | 0.30
(Hydrorpidae) | 229 | 140 | 56 | 0.61 | 0.40 | 0.24
Hydrometridae | 5 | 0 | 0 | 0.00 | N/A | 0.00
Keroplatidae | 60 | 30 | 18 | 0.50 | 0.60 | 0.30
Laeoxiniidae | 67 | 25 | 14 | 0.37 | 0.56 | 0.21
Limoniidae | 280 | 96 | 41 | 0.34 | 0.43 | 0.15
Lonchocephidae | 47 | 17 | 6 | 0.34 | 0.44 | 0.15
Lonchopteridae | 9 | 5 | 6 | 0.56 | 1.20 | 0.50
Megalopidae | 1 | 1 | 1 | 1.00 | 1.00 | 1.00
Micropezidae | 13 | 5 | 3 | 0.38 | 0.60 | 0.23
Milibidae | 13 | 7 | 3 | 1.31 | 0.41 | 0.54
Musciidae | 317 | 174 | 101 | 0.55 | 0.58 | 0.30
Mycetophilliidae | 573 | 306 | 212 | 0.53 | 0.69 | 0.37
Neoptiliphilidae | 1 | 0 | 0 | 0.00 | N/A | 0.00
Oidiidae | 9 | 0 | 0 | 0.00 | N/A | 0.00

Table 1 (continued).

<table>
<thead>
<tr>
<th>Family</th>
<th>Species reported in Germany</th>
<th>BINs</th>
<th>Identified species</th>
<th>BIN Ratio</th>
<th>Identified. Ratio</th>
<th>Identified. Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

doi: 10.14411/eje.2020.037

and related projects. The data of Morinère et al. (2019: Table 1) were imported from the published PDF into a Microsoft Excel 2010 spreadsheet. Before subjecting them to statistical evaluation and discussion, some obvious revisions were applied. The first column “Family” lists Canacidae, Tethinidae, Ulidiidae, Otitidae, Diastatidae and Campioptera as separate families and the next column lists the respective number of species reported in Germany, according to the original Checklist of German Diptera (Schumann et al., 1999). The subsequent columns list the results of the source study itself. As evident from the spreadsheet with the individual records (Morinère et al., 2019: sup-0002-appendixS1) Tethinidae are now included in Canacidae, Otitidae in Ulidiidae, Campioptera in Diastatidae, Microptilinidae in Dolichopodidae and Nycteribiidae in Hippoboscidae (as not all family assignments were double checked, this list is possibly not exhaustive). This explains some suspicious numbers in the column “Ratio bar-
coded/species (%)

Table 2. Comparison of the Identification Ratio with the achieved coverage of the German Diptera diversity (only for families with Identification Ratio > 0.50).

<table>
<thead>
<tr>
<th>Family</th>
<th>E/C</th>
<th>E1/C</th>
<th>(E1/C)·100/C</th>
<th>E/C</th>
<th>(E/C)·100/E'</th>
<th>(E')·100/[C+(E1/C)]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Megamerineridae</td>
<td>1</td>
<td>1.00</td>
<td>100</td>
<td>0</td>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td>Opelitidae</td>
<td>1</td>
<td>1.00</td>
<td>100</td>
<td>0</td>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td>Pseudopomphyzida</td>
<td>1</td>
<td>1.00</td>
<td>100</td>
<td>0</td>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td>Tanyzeugidae</td>
<td>1</td>
<td>1.00</td>
<td>100</td>
<td>0</td>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td>Acartoprhalmidida</td>
<td>2</td>
<td>1.00</td>
<td>100</td>
<td>0</td>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td>Phaemomyzida</td>
<td>3</td>
<td>0.67</td>
<td>67</td>
<td>0</td>
<td>0</td>
<td>67</td>
</tr>
<tr>
<td>Platypheromidae</td>
<td>3</td>
<td>0.67</td>
<td>67</td>
<td>0</td>
<td>0</td>
<td>67</td>
</tr>
<tr>
<td>Atelitidae</td>
<td>3</td>
<td>1.00</td>
<td>100</td>
<td>0</td>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td>Diadocididae</td>
<td>4</td>
<td>0.75</td>
<td>75</td>
<td>0</td>
<td>0</td>
<td>75</td>
</tr>
<tr>
<td>Anisopodidae (Myctebiidae)</td>
<td>8</td>
<td>0.63</td>
<td>50</td>
<td>1</td>
<td>20</td>
<td>56</td>
</tr>
<tr>
<td>Diadocitidae (Campichoetidae)</td>
<td>9</td>
<td>0.67</td>
<td>56</td>
<td>1</td>
<td>17</td>
<td>60</td>
</tr>
<tr>
<td>Lonchophoridae</td>
<td>9</td>
<td>0.67</td>
<td>56</td>
<td>1</td>
<td>17</td>
<td>60</td>
</tr>
<tr>
<td>Rhinophoridae</td>
<td>10</td>
<td>0.60</td>
<td>60</td>
<td>0</td>
<td>0</td>
<td>60</td>
</tr>
<tr>
<td>Canacidae (Tethinidae)</td>
<td>12</td>
<td>0.67</td>
<td>58</td>
<td>1</td>
<td>13</td>
<td>62</td>
</tr>
<tr>
<td>Picophilidae</td>
<td>12</td>
<td>0.67</td>
<td>58</td>
<td>1</td>
<td>13</td>
<td>62</td>
</tr>
<tr>
<td>Milichiidae</td>
<td>13</td>
<td>0.54</td>
<td>46</td>
<td>1</td>
<td>14</td>
<td>50</td>
</tr>
<tr>
<td>Thaumaleidae</td>
<td>15</td>
<td>0.80</td>
<td>7</td>
<td>0</td>
<td>0</td>
<td>7</td>
</tr>
<tr>
<td>Fanniiidae</td>
<td>56</td>
<td>0.55</td>
<td>43</td>
<td>3</td>
<td>11</td>
<td>46</td>
</tr>
<tr>
<td>Tabanidae</td>
<td>58</td>
<td>0.72</td>
<td>39</td>
<td>7</td>
<td>3</td>
<td>69</td>
</tr>
<tr>
<td>Calliphoridae</td>
<td>62</td>
<td>0.53</td>
<td>47</td>
<td>4</td>
<td>12</td>
<td>50</td>
</tr>
<tr>
<td>Ephylidae</td>
<td>177</td>
<td>0.66</td>
<td>104</td>
<td>12</td>
<td>10</td>
<td>61</td>
</tr>
<tr>
<td>Anthomyiidae</td>
<td>227</td>
<td>0.50</td>
<td>77</td>
<td>34</td>
<td>37</td>
<td>43</td>
</tr>
<tr>
<td>Sciaridae</td>
<td>342</td>
<td>0.59</td>
<td>39</td>
<td>39</td>
<td>34</td>
<td>49</td>
</tr>
<tr>
<td>Syrphidae</td>
<td>440</td>
<td>0.62</td>
<td>227</td>
<td>52</td>
<td>18</td>
<td>73</td>
</tr>
</tbody>
</table>

Notes: "E" refers to the quantities established in Fig. 1. "E' = number of named species in the published spreadsheet with individual records (Morinière et al., 2019: sup-0002-appendixS1). For some families this slightly diverges from the number of identified species, E, presented in Table 1.

RESULTS

Fig. 1 illustrates the relevant quantities (= sets) and their general relations for all German species of Diptera. Across
Fig. 1. Relevant quantities (= sets) of dipteran taxa in Germany. The relative magnitude of the areas C (described dipteran species known to occur in Germany), D (BINs recorded in the source study) and E (species identified in study) corresponds to the respective species numbers (9,213, 5,207 and 2,462). The true magnitude of A (all dipteran species occurring in Germany) and B (described dipteran species occurring in Germany) and the intersections of D and E with B and C are unknown.

the individual families, the absolute magnitudes and intersections of these sets vary greatly. A represents the set of all species occurring in Germany, B is a subset of A (B ⊆ A) that represents the species that already have a taxonomic name. The exact magnitude of A and B is unknown. C is a subset of B (C ⊆ B) that represents the named species that have been reported to occur in Germany (2nd column in Table 1). In the source study the total for all families is 9,213.

D represents the set of BINs recorded in the source study (3rd column in Table 1). The total for all families is 5,207. As the study is predominantly based on material collected in Germany, this can be considered a subset of A. D is not a subset of B or C, because it can, and likely will, include species not yet described or not previously recorded in Germany (D ∩ C). Moreover, BINs do not translate directly into species, because some BINs comprise more than one species and, conversely, some species occur in more than one BIN (Morinière et al., 2019). This is reflected in the divergence between the number of “BINs” and the “Total number of taxa/with barcode” in Table 1 of Morinière et al. (2019). In fact, the incongruence of BINs and species is even greater than that, because mathematically the effects of BIN sharing and BIN splitting cancel each other out. The incongruence is hard to appraise, has a minor effect on the subsequent calculations and is therefore disregarded here. The intersections between D and B (D ∩ B) and between D and C (D ∩ C) are unknown.

E is a subset of D (E ⊆ D) that represents the taxonomically identified species (4th column in Table 1). It is calculated by subtracting the number “unnamed/with barcode” from “total number of taxa/with barcode” in Table 1 of Morinière et al. (2019). The total for all families is 2,462. E is also a subset of B (E ⊆ B), because only named species can be taxonomically identified. E is not a subset of C, however, because it can contain named species not previously reported for Germany (E ∩ C). Determining the intersection between E and C (E ∩ C) requires the cross checking of the individual species recorded in the source study with those reported for Germany. This was done only for some individual families (see below).

D/C relates the number of recovered BINs to the number of species reported to occur in Germany (BIN Ratio). The gross BIN Ratio is 0.57 (5th column in Table 1). Because D is not a subset of B or C, the ratio D/C neither indicates the fraction of species reported in Germany for which BINs were established, nor the respective fraction of all species occurring in Germany. It merely indicates that roughly half as many BINs were recovered as there were species reported in Germany before the source study.

The respective values are therefore given not in percent, but as a ratio. The average BIN Ratio across the families (0.42 ± 0.33 sd; median = 0.38) is smaller than the gross BIN Ratio, partially because for a large number of very small families the source study did not include any specimens (Fig. 2a). In Fig. 3a this is illustrated by the large number of data points in the bottom left. Conversely, the remainder of the very small families has a very high BIN Ratio of 1.00. Naturally, for families with only one species occurring in Germany, the BIN Ratio can only be either 0.00 or 1.00. The absence of many small families is also reflected by the trend line across all families declining towards the left (R² = 0.08, p < 0.01). If only those families with more than 10 species are analysed, the correlation is insignificant (R² = 0.00, p = 0.77). In this range the BIN Ratio varies widely and apparently independent of the size of the families. A BIN Ratio larger than 1.00 is a strong indicator of species occurring, but not yet reported, in Germany. After the above revisions of Table 1 the only remaining families with BIN Ratios exceeding 1.00 are Cecidomyiidae, Milichiidae and Trichoceridae (discussed below).

E/D relates the number of BINs that were identified to species to the number of all recovered BINs (BIN Identification Ratio). Roughly half of the recovered BINs were identified to species (0.47, 6th column in Table 1). The single BIN found for Acartophthalmidae combines two identified species (“BIN sharing”), resulting in an extraordinarily high BIN Identification Ratio of 2.00 for that family. To avoid distortion, this outlier was omitted from the following calculations and the diagrams. Still the average BIN Identification Ratio across the families (0.70 ± 0.26 sd;
median = 0.70) is considerably larger than the gross BIN Identification Ratio. This is partially due to the large number of small families with comparatively high values, whereas there are only a few, albeit speciose, families that have low BIN Identification Ratios (Figs 2b and 3b). The decline in the trend line in Fig. 3b towards the right (R² = 0.28, p < 0.01) remains significant, when the trend line is calculated only for those families with more than 10 species (R² = 0.06, p = 0.05), suggesting that species in large families are less likely to be identified.

D(E) is the set of BINs that remained unidentified (“unnamed with barcode”, “dark taxa” in Moriníere et al., 2019). It comprises unknown taxa that have not yet been formally described and given a Linnaean name (outside of B), as well as known taxa (inside B) that were not identified for other reasons.

E/C relates the number of BINs that were identified to species to the number of species known to occur in Germany (Identification Ratio). The gross Identification Ratio is 0.27 (7th column in Table 1). The average across the families is 0.29 (± 0.27 sd; median = 0.23). Because most of the BINs for the small families were identified (see above), their Identification Ratio is mostly identical to their BIN Ratio (left third of Figs 3a and 3c). For many more speciose families, however, only a fraction of the BINs were identified, resulting in a downward shift of the respective data points on the right side of Fig. 3c in comparison to Fig. 3a. The regression analysis shows no correlation between the Identification Ratio and the size of the family (R² = 0.00, p = 0.88) and the respective trend lines in Fig. 3c are almost level. But the data do not follow a Gaussian distribution. Instead, some outliers with particularly high Identification Ratios (discussed below) are balanced by the bulk of values being in the 0.00–0.30 range (Fig. 2c).

E\C/C is the achieved coverage of the species of Diptera known to occur in Germany by identified barcodes. This was assessed for the 24 families with a high Identification Ratio (>0.50) by individually cross checking the named species recorded in the source study with those reported for Germany (Schumann et al., 1999, Table 2, Fig. 4). Complete coverage was achieved for six small families, for which one to three species are known to occur in Germany, and these species were found and successfully barcoded. In most of the families with more than four species the actual coverage is considerably smaller than the Identification Ratio. The gap is somewhat reduced, however, when the species newly recorded for Germany in the source study itself are added to both sides of the fraction. This is particularly evident for Anthomyiidae and Sciaridae, for which comparatively high percentages of new records (E\C) were found (32% and 37% of the identified species).

Exemplary families

Because the values vary greatly across the families, the discussed averages are only partially informative for evaluating the progress of the project and the resulting implications for future strategies. For a more detailed analysis, the data for some exemplary families are identified by coloured lines in Figs 3a–c, and their characteristics summarized in Table 3. All other families can be located in the diagrams by their respective coordinates in Table 1. In

<table>
<thead>
<tr>
<th>BIN Ratio</th>
<th>BIN Identification Ratio</th>
<th>Identification Ratio</th>
<th>Colour in Figs 3a–c</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flagship families</td>
<td>high</td>
<td>high</td>
<td>high</td>
</tr>
<tr>
<td>Families of medical importance</td>
<td>low</td>
<td>high or low</td>
<td>low</td>
</tr>
<tr>
<td>Collector favourites</td>
<td>low</td>
<td>high</td>
<td>low</td>
</tr>
<tr>
<td>Dark taxa rich families</td>
<td>high</td>
<td>low</td>
<td>low</td>
</tr>
</tbody>
</table>
Fig. 3. Achieved coverage of individual families with different numbers of species reported in Germany in terms of (a) BIN Ratio (number of recovered BINs / number of species reported in Germany), (b) BIN Identification Ratio (number of BINs that were identified to species / number of recovered BINs) and (c) Identification Ratio (number of BINs that were identified to species / number of species reported in Germany). Some exemplary families are highlighted in green (flagship families), red (families of medical importance), blue (collector favourites), or black (families rich in dark taxa). The vertical portions of the lines in Fig. 3c indicate the deviation from the respective data points in Fig. 3a. The abscissa is scaled logarithmically for better resolution. The trend lines are likewise logarithmic. Dotted trend line for all families and solid trend line for families with more than 10 species.
Fig. 3c the lines have a vertical portion which indicates the deviation from the respective data points in Fig. 3a, i.e. the deviation of the Identification Ratio from the BIN Ratio. A short vertical portion signifies that most of the BINs recovered for that family were identified to species, whereas a long vertical portion generally indicates a large proportion of dark taxa.

The Tabanidae, Ephydridae and Syrphidae may rightfully be called the “flagship families” of the source study. They are indicated by the colour green in Figs 3a–c. All three are comparatively speciose and have a particularly high BIN Ratio, BIN Identification Ratio, and thus Identification Ratio. These families were analysed more closely, especially with respect to the origin and condition of the material studied and its taxonomic identification, to determine what factors contributed to the particularly high coverage achieved.

For Tabanidae the spreadsheet with the individual records contains 96 records, of which 91 (95%) are identified to species (Morinière et al., 2019: sup-0002-appendixS1). The number of taxa with a barcode is listed as 45, of which only three are unidentifi ed (Morinière et al., 2019: Table 1). This translates into 42 identifi ed species. As 58 species had been previously reported in Germany, the resulting Identification Ratio is 0.72, the highest among the speciose families. Cross checking with the Checklist of German Diptera (Schumann et al., 1999) reveals that only three of the identifi ed species (7%) are not listed there. After adding the new fi nds, the German fauna of Tabanidae now includes 61 known species of which 69% are represented by barcodes (Fig. 4). Notably, all but one of the 42 identifi ed species (98%) are represented by one or more specimens identifi ed and almost always also collected by W. Schacht (1939–2011), a dedicated dipterist, excellent collector and renowned expert on Tabanidae (Kotrba, 2011). The vouchers drawn from his collection are documented by photographs in the BOLD database. Different from the general methodology of the source study (specimens up to fi ve years old, stored in 96% EtOH before DNA extraction), all of these specimens were collected more than 20 years ago and, as evident from the photographs on the BOLD internet site, were pinned, i.e. dried. From almost all of them a full 658 bp sequence was retrieved.

For Ephydridae the spreadsheet with the individual records contains 548 records, of which 488 (89%) are identified to species (Morinière et al., 2019: sup-0002-appendixS1). The number of taxa with a barcode is listed as 132, of which 16 are not identifi ed (Morinière et al., 2019: Table 1). This translates into 116 identifi ed species. As 177 species had been previously reported in Germany, the resulting Identification Ratio is 0.66, the second best among the more speciose families. Cross checking with the Checklist of German Diptera (Schumann et al., 1999) shows, that only 12 of the identifi ed species (10%) are not listed there. After adding the new fi nds, the German fauna of Ephydridae now includes 189 known species of which 61% are represented by barcodes (Fig. 4). All but four of the 116 identifi ed species (97%) are represented by one or more specimens collected and identifi ed by J.-H. Stuke, a very successful contemporary German dipterist, collector and expert on Ephydridae (as well as Conopidae, Carnidae and several other acalyptrate families). Unfortunately, the preservation of the specimens sampled is generally not documented in the data set and could not be assessed from the photographs on the BOLD internet site. According to personal communications from J.-H. Stuke and D. Doczkal, the respective sequences were obtained from pinned material.

For Syrphidae the spreadsheet with the individual records contains 1,911 records, of which 1,381 (72%) are
identified to species (Morinière et al., 2019: sup-0002-appendixS1). The number of taxa with a barcode is listed as 297, of which 24 are unidentiﬁed (Morinière et al., 2019: Table 1). This translates into 273 identiﬁed species. As 440 species had been previously reported in Germany, the Identiﬁcation Ratio is 0.62, the third best among the speciose families. For this family, the vertical part of the respective line in Fig. 3c leads upwards, indicating that the number of found taxa with barcodes is considerably (23%) greater than the number of established BINs and therefore several BINs include more than one taxon (Morinière et al., 2019). Out of the 273 species identiﬁed, 106 (39%) are represented by one or more specimens identiﬁed by W. Schacht (see above). The vast majority of these were collected more than 20 years ago by various collectors, and for almost all of these a sequence longer than 550 bp was recovered. A large part of the remaining specimens were collected and identiﬁed by D. Doczkal, co-author of the source study and expert on Syrphidae.

A common factor for the three flagship families is that experts for the respective families were intensely involved in the collecting and identifying of the relevant material.

Three families of medical importance, i.e. Culicidae, Simuliidae and Ceratopogonidae, are indicated by the colour red in Figs 3a–c. The respective BIN Ratios are comparatively low, especially for Culicidae, indicating that only part of the diversity was sampled and/or included in the source study (Fig. 3a). Only for Culicidae was a high BIN Identiﬁcation Ratio achieved (Fig. 3b). As a result, the Identiﬁcation Ratios are very low for all three families (0.15, 0.18 and 0.09). The situation is somewhat similar for three families, which may be considered to be collector favourites, i.e., Asilidae, Conopidae and Stratiomyidae, indicated by the colour blue. Similar to Syrphidae (above), these families are very well studied and described for Germany, and several experts on these families are currently present in this country. Accordingly, high BIN Identiﬁcation Ratios were achieved. But the Identiﬁcation Ratios are still very low (0.15, 0.17 and 0.24), mainly due to the low input of material as is evident from the low BIN Ratios.

A common factor for the medically important and collector favourite families is that comparatively little material was included in the source study, and this appears to be the main reason for the low Identiﬁcation Ratios.
Black lines indicate some families that are characterized by large vertical portions of the respective lines in Fig. 3c. An extreme example is Cecidomyiidae, the most speciose of all families of Diptera in Germany. 836 species of Cecidomyiidae have been reported in this country (Table 1) and many more apparently occur: 927 BINs were recovered, resulting in a very high BIN Ratio of 1.11. At the same time, Cecidomyiidae have the lowest BIN Identification Ratio (0.05) and, accordingly, the lowest Identification Ratio (0.05) of the more speciose families. Phoridae and Chironomidae are other examples of very speciose families with comparatively high BIN Ratios, low BIN Identification Ratios and thus also low Identification Ratios, while among the smaller families the Trichoceridae, Milichiidae and Carnidae stand out with similar characteristics. A common factor for these families is that high BIN Ratios indicate a comparatively good sampling of the diversity (but see below), but only few of the records could be identified, leaving a large residual of dark taxa.

### Identification methods

The entries in the “identification methods” column of the spreadsheet with the individual records (Morinie r et al., 2019: sup-0002-appendixS1) are summarized in Fig. 5. For 15,372 records (59%) the entry indicates some kind of DNA-based identification (“BIN Taxonomy Match”, “BIN-taxonomy”, “BIN match”, “BIN comparison”, “BOLD ID Engine Manual”, “BOLD BIN match”, “BOLD Tree”, “DNA Barcoding”, “DNA plausibility”, “full DB manual search”, “Full DB, manual”, “Sequence match”, “Tree based identification”). For 472 records (2%) a morphological identification is indicated (“morphological”, “morphology”, “microscopy”, “Morphology + DNA barcodes”, “Digital Morphology”, “morpho-taxonomy”). The records based on morphological identification were checked for additional information on the literature and/or identification keys used, which was found for 18 records of Chironomidae in the column “taxonomic notes”. The remaining 10,066 identified records (39%) lack entries in the “identification methods” column.

### Identifiers

The entries in the “identifier” column of the spreadsheet with the individual records (Morinie r et al., 2019: sup-0002-appendixS1) are summarized in Table 4. Out of the 34 listed identifiers, 15 are members of the German society of dipterists “AK Diptera” and/or personally known to the author as experts on Diptera (just called “dipterists” below). The remaining identifiers cannot be classified as dipterists, but all or most of them are barcoding experts or technicians. Fig. 6 visualizes the individual contributions of the identifiers in terms of identified specimens (Fig. 6a) and identified species (Fig. 6b), showing that the work load was very unevenly distributed among a small number of main contributors. More than half of the specimens were identified by three non-dipterists, and among these, a single person provided the vast majority of the identifications.

In the dipterist fraction, eight experts provided substantial numbers of identifications, and among these, two were responsible for the bulk of the work. Because many species were independently identified by more than one of the identifiers, the sets of identified species intersect. It would be very time consuming to correct for this. Even without this correction, the species identified by dipterists add up to only 2,013 (Table 4). This implies that at least 449 out of the 2,462 species identified were not identified by a dipterist. Among the dipterists, one identified more species (1,92) than all others combined.

### DISCUSSION

**Constructing a DNA barcode library for German Diptera: Progress**

The evaluated study met its goal to provide a DNA barcode library for 5,200 BINs of Diptera. Without question, this is impressive, especially given the limited time frame and largely uniform collecting methods. Likely one of the most interesting figures for the community would be the achieved coverage of the diversity of all Diptera in Ger-

---

**Table 4.** Identifiers listed in Morinie r et al. (2019: sup-0002-appendixS1) with their respective contribution in terms of identified species and specimens.

<table>
<thead>
<tr>
<th>Name</th>
<th>Expert</th>
<th>Specimens identified</th>
<th>Species identified</th>
</tr>
</thead>
<tbody>
<tr>
<td>K. Perez</td>
<td></td>
<td>11,171</td>
<td>893</td>
</tr>
<tr>
<td>D. Doccказ</td>
<td></td>
<td>11,171</td>
<td>893</td>
</tr>
<tr>
<td>S. Prescher</td>
<td></td>
<td>11,171</td>
<td>893</td>
</tr>
<tr>
<td>J. Morinie</td>
<td></td>
<td>11,171</td>
<td>893</td>
</tr>
<tr>
<td>S. Wiedenbrug</td>
<td></td>
<td>11,171</td>
<td>893</td>
</tr>
<tr>
<td>V. Levesque-Beaudin</td>
<td></td>
<td>11,171</td>
<td>893</td>
</tr>
<tr>
<td>W. Adaschkwitz</td>
<td></td>
<td>11,171</td>
<td>893</td>
</tr>
<tr>
<td>K. Heller</td>
<td></td>
<td>11,171</td>
<td>893</td>
</tr>
<tr>
<td>J.-H. Stuke</td>
<td></td>
<td>11,171</td>
<td>893</td>
</tr>
<tr>
<td>W. Schacht</td>
<td></td>
<td>11,171</td>
<td>893</td>
</tr>
<tr>
<td>J. Kjaeransdien</td>
<td></td>
<td>11,171</td>
<td>893</td>
</tr>
<tr>
<td>O. Hawlitschek</td>
<td></td>
<td>11,171</td>
<td>893</td>
</tr>
<tr>
<td>T. Scheler</td>
<td></td>
<td>11,171</td>
<td>893</td>
</tr>
<tr>
<td>R. Hessing</td>
<td></td>
<td>11,171</td>
<td>893</td>
</tr>
<tr>
<td>C. Chimenon</td>
<td></td>
<td>11,171</td>
<td>893</td>
</tr>
<tr>
<td>M. Spies</td>
<td></td>
<td>11,171</td>
<td>893</td>
</tr>
<tr>
<td>C.D. Schubart</td>
<td></td>
<td>11,171</td>
<td>893</td>
</tr>
<tr>
<td>J. Mortelmans</td>
<td></td>
<td>11,171</td>
<td>893</td>
</tr>
<tr>
<td>R. Mueller</td>
<td></td>
<td>11,171</td>
<td>893</td>
</tr>
<tr>
<td>N.E. Woodley</td>
<td></td>
<td>11,171</td>
<td>893</td>
</tr>
<tr>
<td>J. Spelda</td>
<td></td>
<td>11,171</td>
<td>893</td>
</tr>
<tr>
<td>L. Hendrich</td>
<td></td>
<td>11,171</td>
<td>893</td>
</tr>
<tr>
<td>R. Wagner</td>
<td></td>
<td>11,171</td>
<td>893</td>
</tr>
<tr>
<td>J. Podhorn</td>
<td></td>
<td>11,171</td>
<td>893</td>
</tr>
<tr>
<td>M. Scheingraber</td>
<td></td>
<td>11,171</td>
<td>893</td>
</tr>
<tr>
<td>A. Reimann</td>
<td></td>
<td>11,171</td>
<td>893</td>
</tr>
<tr>
<td>F. Koehler</td>
<td></td>
<td>11,171</td>
<td>893</td>
</tr>
<tr>
<td>M. F. Geiger</td>
<td></td>
<td>11,171</td>
<td>893</td>
</tr>
<tr>
<td>E. Plassmann</td>
<td></td>
<td>11,171</td>
<td>893</td>
</tr>
<tr>
<td>S. Schmidt</td>
<td></td>
<td>11,171</td>
<td>893</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>11,171</td>
<td>893</td>
</tr>
</tbody>
</table>

Note: 1 *Identifiers characterized as dipterists are members of the German society of dipterists “AK Diptera” and/or personally known to the author as dipterists.*

---

doi: 10.14411/eje.2020.037
many in terms of identified barcodes (E/A). This cannot be assessed, however, because the true extent of the diversity of German Diptera is unknown. The coverage of the known part of the diversity (E/C/C) was assessed for all families with Identification Ratios larger than 0.50 by cross checking the list of found species with the species known to occur in Germany (Fig. 4). Complete coverage was achieved for six very small families. Among the larger families a very respectable coverage was achieved for the “flagship” families Tabanidae (69%), Ephyrididae (61%) and Syrphidae (53%). Among the medium sized families with up to 20 species, the Diastatidae (& Campichoetidae), Lonchopteridae, Rhinophoridae, Canacidae (& Tethinidae) and Piophilidae have a relatively high coverage of nearly 60%.

For the majority of families, which have Identification Ratios below 0.50, the coverage of the known diversity was not assessed. The Identification Ratio, i.e. the ratio between the number of identified BINs and the number of species known to occur in Germany (E/C), generally overestimates the achieved coverage (Fig. 4), but it is much easier to assess, and may still be used as a rough indicator of the achieved coverage. Across all families this ratio is 0.29, with the majority of values less than average (Fig. 2c). Regrettably, the Identification Ratio is also low for the medically important families, such as Culicidae, Simuliidae and Ceratopogonidae, and for some collector favourites, such as Asilidae, Conopidae and Stratiomyiidae, mainly due to low BIN Ratios showing that only a small fraction of the diversity was sampled (Figs 3a and 3c). The coverage provided by the entire BOLD database is probably higher, but this was not investigated here.

Unlike the quantity of identified BINs, i.e. species, the quality of the identifications is not as easily appraised. If indeed only 472 of all records were identified by traditional morphological methods (Fig. 5), then this would also apply to maximally 472 of the 2,462 identified BINs. Very likely, the proportion of morphological identifications is higher, but this is not evident from the published file and remains a matter of speculation. From the number of species identified by dipterists (Table 4) it follows that at least 449 species were not identified by a dipterist, but by a barcoding expert or technician. Some unknown quantity between 449 and 1,981 species must have been exclusively identified using DNA-based identification. Identifying specimens for the establishment of a DNA barcode library based on a DNA barcode library seems circuitous (Kotrba, 2019). Basically, the responsibility for the correct identification of these specimens is delegated and it is not immediately clear to whom (see below).

Extending the assessment of coverage to include the BINs that remained unidentified results in a ratio of roughly 1:2. This neither indicates that the source study “covers ~55% of the known Diptera fauna from Germany” (Morinière et al., 2019: 3), nor that it covers “half of the German Diptera fauna” (Morinière et al., 2019: 16), but merely that roughly half as many BINs were recovered as there were Diptera species reported in Germany before the source study (see above).

The BIN Ratio varies greatly across the families (Fig. 3a). This could be due to an uneven representation of the families in the original samples. Borkent et al. (2018) report that Malaise trap catches at Zurquí, Costa Rica, include only about half of the diversity of Diptera collected using a wider range of methods, and Karlsson et al. (2020) report that Malaise traps are comparatively inefficient at catching large, active insect fliers with good vision. The variation could also be a result of uneven processing of the individual families, favouring some and/or disregarding others. This may apply to the Scathophagidae, whose total absence in the source study comes as a surprise. There are 57 species of this family listed in the German checklist and at least Scathophaga stercoraria is very common. Moreover, the BIN Ratio is a function of the previous knowledge of the diversity of Diptera in Germany. In well studied families, where C in Fig. 1 includes also the rare species and thus approximates A, it is hard to achieve a high BIN Ratio. At the same time, in such cases the BIN Ratio is a good indicator of the degree to which the actual diversity of Diptera in Germany was covered by the source study. The collector favourite families, Asilidae, Conopidae and Stratiomyiidae, exemplify this correlation.

Syrphidae, which also undoubtedly qualify as collector favourites, have a higher BIN Ratio, possibly due to the substantial incorporation of specimens from the collections of experts. Conversely, it should be easy to achieve a high BIN Ratio for poorly studied families, where C is much smaller than A. In particular, a BIN Ratio larger than 1.00, as recorded for Cecidomyiidae, Milichiidae and Trichoceridae, is a strong indicator of species occurring, but not yet reported in Germany (Morinière et al., 2019) and even possibly undescribed. The same high BIN Ratio may thus indicate a good coverage of a well-studied family, as well as a moderate coverage of a very poorly studied family. Apart from the obvious underrepresentation of the smallest families with only up to 10 species, the size of the family has no significant effect on the BIN Ratio.

Roughly half of the recovered BINs were identified to species. The significant negative correlation of the BIN Identification Ratio with the size of the families (Fig. 3b) suggests that species belonging to large families are less likely to be identified. Naturally, species of very small families are easier to identify, because the respective identification keys will be short. Conversely, some very speciose families such as Cecidomyiidae, Phoridae and Chironomidae are notoriously challenging, even for experts. The respective literature is vast and the relevant characters tiny, making identifications more difficult. Nevertheless, quite a number of specialists have dedicated their working lives to these taxa (e.g. R. Gagne, N. Dorchin and M. Jaschhof for Cecidomyiidae). As exemplified by the “flagship” families Ephyrididae, Syrphidae and Tabanidae, a high BIN Identification Ratio can be achieved also in speciose families, if qualified experts are involved.
The other half of the recovered BINs remained unidenti-
ified. These are classified as dark taxa, comprising records
of unknown taxa that have not yet been formally described
and given a Linnaean name, as well as known taxa that
could not be identified, because they are “extremely dif-
ficult” to identify (Staatliche Naturwissenschaftlichen
Sammlungen Bayerns, 2019). The fact that about 7,000
further named species are known to occur in Germany in
addition to the 2,462 identified in the source study, sug-
gests that at least part of the remaining roughly 2,700 dark
taxa will ultimately turn out to belong to known species
(intersection of hatched area D:E with C in Fig. 1). Like-
wise, at the family level, the fact that 792 further named
Cecidomyiidae species are known to occur in Germany in
addition to the 44 identified in the source study, suggests
that a good part of the 882 unnamed BINs in this family
will ultimately turn out to be known and identifiable spe-
cies (as soon as a Cecidomyiidae expert is recruited and
dedicates the needed time to their identification). Among
the small families, Carnidae particularly stand out with
100% of dark taxa. Eleven Carnidae species are reported
in Germany, but out of the seven BINs recovered for this
family not a single one was identified. At least some of
these BINs will likely turn out to belong to known German
species. Of course, many new species and/or new records
for the German fauna are also likely to be discovered in
the process. For the 24 families, in which the found spe-
cies were cross checked with the list of species known to
occur in Germany (Table 2), the overall percentage of new
records for Germany is 18%. The highest individual values
are 34% (Sciaridae) and 32% (Anthomyiidae).

There are many possible reasons for specimens of
known species to remain unidentified. “Extremely difficult
to identify” may arguably apply, but difficulty, like beauty,
lies in the eye of the beholder. The vast majority, if not all,
of the more than 9,000 species listed in the checklist of
German Diptera (Schumann et al., 1999; Schumann, 2003,
2005, 2010) were identified based on classical morphologi-
cal characters, proving that this is feasible. This situation
is different, e.g., for fungi, where a growing proportion
of species are known only from sequence data and cannot
be linked to any physical specimen or resolved taxonomic
name; they are referred to as “dark taxa” or “dark matter
fungi” (Ryberg & Nilsson, 2018). Sometimes “difficult” is
extended to include options such as “requiring expertise”,
“time consuming”, or simply “tedious”. Page (2016) high-
lights taxonomic capacity as the main limiting factor for
BIN identification. This concerns not only a shortage of
taxonomic experts as such, but, maybe even more impor-
tantly, a shortage of expert time available and recruited for
this kind of studies. Clearly, therefore, it is important to
reveal what efforts were made to identify the records. Es-
specially, if the ratio of dark taxa is to be utilized to assess
deficiencies in taxonomic and faunistic knowledge.

Facing the well over 9,000 species in 111 dipteran fam-
ilies known to occur in Germany, and more than 40,000
records to be identified within a very limited time frame,
the present analysis indicates that way too few dipterists
were involved and that the work load was very unevenly
distributed among them. Specifically, only eight dipterists
contributed any substantial number of identifications, and
a single one accomplished the vast bulk of these (Table
4, Fig. 6a and b). For comparison, Karlsson et al. (2020)
report that more than 130 experts were actively recruited
for the Swedish Malaise Trap Project, which had a roughly
comparable scope and time frame, and that many of them
provided some number of identifications, which totalled
to over 4,000 species. Borkent et al. (2018) report the in-
volvement of 59 dipterists in a study of all Diptera col-
lected in one year in a patch of cloud forest in Costa Rica.

**Constructing a DNA barcode library: Ways of improving the developoment, utility and reliability**

Based on the above analysis it is possible to identify
some ways of improving the future development, utility
and reliability of this DNA database.

1. **Sample the collections of experts**

A high Identification Ratio can only be achieved by
combining a high BIN Ratio with a high BIN Identification
Ratio. The high Identification Ratios achieved for Tabani-
dae, Ephyridae and Syrphidae suggest that sampling the
collections of experts results in a higher coverage within
a reasonable time frame, as opposed to random collecting
and relying on a posteriori identification. Good results with
complete or nearly complete barcodes were obtained even
from comparatively old dry (pinned) material, and these
barcodes are linked to very reliably identified vouchers.
Similarly, good results were achieved by Hausmann et al.
(2016) and Dey et al. (2019) for Geometridae (Lepido-
ptera). Where possible, the barcoding of type material or
specimens verified by comparison with type material con-
stitutes the ideal approach.

2. **Prioritize important taxa**

It is unfortunate that some of the most important families
of Diptera are still very poorly covered. Naturally, achiev-
ing the same high coverage for all dipteran families oc-
curring in Germany within a limited time frame cannot be
expected. Giving priority to medically, agriculturally or
ecologically important families and to collector favourite
families will meet the most pressing needs of the com-
community and serve as a good advertisement for the useful-
ness and wide applicability of the DNA barcode library. At
the same time, prioritizing important families is likely to
increase the rate of progress, as experts and material are
more likely to be readily available. Moreover, these taxa
are well studied, well documented and sufficiently covered
in the literature. Achieving near complete coverage for
some important families this way might be preferable to a
random 30% coverage of the entire diversity.

3. **Allocate resources to recruiting experts**

In their response to Kotrba (2019), Chimeno et al. (2019)
state that “even authorities may fail (nobody is perfect),
although this is more improbable by the expert rather than
by the beginner”. This is true and constitutes a good reason
to preferentially rely on experts.
Of course, dipterology has not been spared the general restrictions of the “taxonomic impediment”, i.e. the deplorable shortage of both professional and amateur experts, which is the logical outcome of today’s failure to educate, employ, fund and generally support such scientists. Still, a considerable number of dipterists active in Germany and worldwide have recently been involved in comparable studies (see above), and can be contacted through a number of platforms such as AK Diptera (https://www.ak-diptera.de/), NADS (http://www.nadsdiptera.org/), or “the new Diptera site” (http://diptera.myspecies.info/). Morinière et al. (2019) explicitly “invite the global community of dipteran taxonomists to improve identifications for the many “dark taxa” encountered in our study by identifying these vouchers using reverse taxonomic approaches.” Those taxa with the most pressing deficits can be identified in Figs 3a-c and any contribution in that respect is strongly encouraged here for the common good. Surely, progress could be boosted by allotting money for remunerating such contributions.

It is also of great importance to educate new generations of experts. But it is not a promising strategy for the progress of the present project, i.e. establishing a DNA barcode library, to mix these tasks. Heavily relying on the contributions of trainees will negatively affect both the rate of progress and the reliability of the results.

4. Appropriately document identifiers and identification methods

The value of any DNA barcode library for identifying species is first and foremost dependent on the reliability of the stored identifications (Kotrba, 2019). If there is no conflicting taxonomy within a BIN, the taxonomic assignment is unanimous, but its reliability still depends on the expertise of the identifiers and the accuracy of the utilized literature and keys. Appropriate data about the identifying expert designates the scientific responsibility. Additional annotations regarding his or her field of expertise could further help in appraising the reliability of the identification. Documentation on the methodology is important, firstly, because only records identified by methods other than comparison with sequences already in the system add to the availability of taxonomic identification from that system. Secondly, for morphological identifications, documentation of the literature and/or keys used further helps in the appraisal of the reliability. Collins (2011) states that “... thoroughly demonstrating the characters used to identify your specimens will make the whole system more transparent and reliable”. More specifically, Meier (2017) establishes that species identifications in biological publications should be treated as a ‘Result’ and the literature that was used for that identification should be properly cited. This should also apply for species identifications presented in the form of a published DNA barcode library.

Matters are more complicated if there is conflicting taxonomy within a BIN. Ratnasingham & Hebert (2013) suggest that “Users will encounter discordant taxonomic assignments, especially among unpublished records, but majority rule is a useful way to gauge the validity of a particular identification. For example, if most specimens are assigned to one species and these identifications derive from several taxonomists, this assignment is more likely to be correct than any ‘outlier’ identifications.” This suggestion leads down a dark and dangerous path! On principle, “more likely correct” is not an acceptable degree of reliability in science. Moreover, the suggested approach depends on the user being able to recognize which records were identified by taxonomists, or even experts on the relevant taxon or clade, and which were not. This precondition is hardly ever met. The majority rule approach will be deceptive, if, e.g., a single ‘outlier’ correctly identified by an expert is outnumbered by numerous records identified by laymen, e.g. in the context of trainee programs, or, even worse, by DNA-based identification based on the very same BIN. The magnitude of this problem is constantly growing due to the uploading of thousands of sequences with incomplete documentation and/or with questionable or preliminary identifications.

According to Chimeno et al. (2019) “the truly crucial point is that the result of identification can be checked and thus can be falsified. By establishing voucher specimens...” But if it is unknown which identifications need to be checked and which are reasonably reliable, then all identifications may have to be considered doubtful. When in doubt, it will be more straightforward to directly identify the specimens at hand, than to get hold of the questionable vouchers and reidentify those. It might be naïve to assume the global scientific community will be able, let alone willing, to correct already uploaded misidentifications on a large scale a posteriori. Therefore “ultimately, the best prevention lies with collaboration, and working through identification uncertainties between labs before data are uploaded as reference specimens” (Collins, 2011).

CONCLUSION

Although based on the scrutiny of a specific, regionally and taxonomically restricted project, the resulting insights very likely apply to any comparable project. The progress in the construction and the subsequent success of a DNA barcode library for the purpose of DNA-based species identification is critically dependent on, and limited by, the expertise needed and available for the reliable identification of the vouchers. If the project is aiming for a rather complete coverage rather than a partial outcome with a huge residual of “dark taxa”, then the appropriate experts need to be actively recruited and involved from the beginning. Including already identified material from the collections of experts may help speed up the progress considerably. Conversely, any number of trainees and any amount of collecting cannot make up for the lack of experts.

The efforts made to identify the vouchers need to be appropriately documented together with the results, especially the identity of the identifier and the methodology utilized. It must be clearly evident, which identifications are based on expert morphological identification as opposed to, e.g., mere comparison with other sequences already in the system. Such documentation will greatly enhance the
value of the database as a dependable identification tool. For example, in cases of conflicting identifications within a BIN, this will help to decide which of the records are more reliably identified. The majority rule approach for BINs containing conflicting identifications is strongly discouraged. Instead, every resulting DNA-based species identification should be traceable to at least one reliable DNA barcode from a specimen that was identified by an expert taxonomist using traditional morphological methods.

ACKNOWLEDGEMENTS. I thank M. Spies (SNSB-Zoologische Staatssammlung München) for help with editing the manuscript and two anonymous reviewers for their valuable suggestions. H. Küchenhoff (Statistisches Beratungs-labor, LMU München) is acknowledged for his help with the statistics.

REFERENCES


Received May 7, 2020; revised and accepted June 29, 2020
Published online July 28, 2020