Using taxonomic consistency with semi-automated data pre-processing for high quality DNA barcodes

Björn Rulik | Jonas Eberle | Laura von der Mark | Jana Thormann | Manfred Jung | Frank Köhler | Wolfgang Apfel | Andreas Weigel | Andreas Kopetz | Jonas Köhler | Frank Fritzlar | Matthias Hartmann | Karl Hadulla | Joachim Schmidt | Thomas Hörren | Detlef Krebs | Florian Theves | Ute Eulitz | André Skale | Dirk Rohwedder | Andreas Kleeborg | Jonas J. Astrin | Matthias F. Geiger | J. Wolfgang Wägele | Peter Grobe | Dirk Ahrens

1 Zoologisches Forschungsmuseum Alexander Koenig, Zentrum für molekulare Biodiversitätsforschung (zmb), Bonn, Germany
2 Zoologisches Forschungsmuseum Alexander Koenig, Zentrum für Taxonomie und Evolution (ZTE), Bonn, Germany
3 Athenstedt, Germany
4 Bornheim, Germany
5 Naturkundemuseum Erfurt, Erfurt, Germany
6 Bonn, Germany
7 Jena, Germany
8 Trier, Germany
9 Admannshagen, Germany
10 Mülheim an der Ruhr, Germany
11 Aachen, Germany
12 Be, Germany
12 C/o Senckenberg Naturhistorische Sammlungen Dresden, Museum für Tierkunde, Dresden, Germany
13 Berlin, Germany

Correspondence
Björn Rulik
Email: b.rulik@leibniz-zfmk.de
and
Jonas Eberle
Email: eberle.jonas@gmail.com
and
Dirk Ahrens
Email: d.ahrens@leibniz-zfmk.de

Abstract
1. In recent years, large-scale DNA barcoding campaigns have generated an enormous amount of COI barcodes, which are usually stored in NCBI’s GenBank and the official Barcode of Life database (BOLD). BOLD data are generally associated with more detailed and better curated meta-data, because a great proportion is based on expert-verified and vouched material, accessible in public collections. In the course of the initiative German Barcode of Life data were generated for the reference library of 2,846 species of Coleoptera from 13,516 individuals.

2. Confronted with the high effort associated with the identification, verification and data validation, a bioinformatic pipeline, “TaxCI” was developed that (1) identifies taxonomic inconsistencies in a given tree topology (optionally including a reference dataset), (2) discriminates between different cases of incongruence in order to identify contamination or misidentified specimens, (3) graphically marks those cases in the tree, which finally can be checked again and, if needed, corrected or removed from the dataset. For this, “TaxCI” may use DNA-based species delimitations from other approaches (e.g. mPTP) or may perform implemented threshold-based clustering.

3. The data-processing pipeline was tested on a newly generated set of barcodes, using the available BOLD records as a reference. A data revision based on the first run of the TaxCI tool resulted in the second TaxCI analysis in a taxonomic match ratio very similar to the one recorded from the reference set (92% vs. 94%). The revised dataset improved by nearly 20% through this procedure compared to the original, uncorrected one.

4. Overall, the new processing pipeline for DNA barcode data allows for the rapid and easy identification of inconsistencies in large datasets, which can be dealt with
1 | INTRODUCTION

DNA barcoding can provide an efficient tool for rapid biodiversity assessments, because it meets needs for rapid and reproducible specimen identification in the era of massive habitat destruction, biodiversity loss, and climate change (Hebert & Gregory, 2005; Valentini, Pompanon, & Taberlet, 2009). Despite some well-known shortcomings (Funk & Omland, 2003; Ballard & Whitlock, 2004; Hebert & Gregory, 2005; Dowton, Meiklejohn, Cameron, & Wallman, 2014; Ross, 2014), barcoding is currently possibly the fastest and most straightforward way to assess the vast diversity of organisms such as invertebrates, which play a crucial role in ecosystem function, but are often poorly known taxonomically (Hebert & Gregory, 2005). Using standardized markers in DNA barcoding allows connecting the identities of different life stages such as eggs, larvae or adults—often a major difficulty in morphology-based taxonomy (e.g. Ahrens, Monaghan, & Vogler, 2007; Etzler, Wanner, Morales-Rodriguez, & Ivie, 2014; Freitag, 2013; García-Robledo, Kuprewicz, Staines, Kress, & Erwin, 2013; Šipek & Ahrens, 2011). Barcoding has been successfully applied to a vast number of taxa in many different geographic regions (www.boldsystems.org). It has become obvious that validated, comprehensive species libraries are the most fundamental basis for optimal barcode-based taxon identification (Kvist, 2013). Since the early days of DNA barcoding, barcodes have also been used for direct estimation of species boundaries (Carstens, Pelletier, Reid, & Satler, 2013; Meier, Shiayang, Vaidya, & Ng, 2006; Pons et al., 2006; Puillandre, Lambert, Brouillet, & Achaz, 2012; Ratnasingham & Hebert, 2013; Templeton, 2001; Zhang, Kapli, Pavlidis, & Stamatakis, 2013). Most of these methods attempt to infer species boundaries from the discontinuum between intraspecific and interspecific sequence variation, either visible as a “barcode gap,” or as a shift in branching rates (GMYC) or number of substitutions per branch (PTP). Therefore, barcode reference libraries should cover the genetic variation within species as comprehensive as possible. Bergsten et al. (2012) revealed increasing difficulties in specimen determination using DNA barcodes when scaling up the geographic scope from local or regional to continental focus. They found that in diving beetles (Dytiscidae) a minimum of 70 specimens needs to be analysed per species to sample 95% of its intraspecific variation. Consequently, building a proper barcode reference library that sufficiently reflects intraspecific variation and is thus able to correctly detect species boundaries is expected to be an elaborate procedure.

Completing existing barcode libraries and uploading larger datasets to the Barcode of Life Database (BOLD) can be challenging, particularly when a vast amount of barcode data already exists in public libraries. It is important to make sure that the newly submitted data match the existing data in order to avoid taxonomic confusion. Among different sources for inconsistencies, most likely are specimen misidentifications, altering nomenclature or misspellings, contaminations (with admixed alien DNA), sample confusion, or low-quality sequences. Especially in very diverse groups, checking big data (several thousand barcodes) requires major efforts. Recently, the Barcode Index Number (BIN) System, a DNA-based registry for all animal species (Ratnasingham & Hebert, 2013), has been implemented within BOLD which helps to detect taxonomic inconsistencies. However, once the data are imported into BOLD, the effort to revise and correct them online is higher than having done this beforehand and on user’s personal system. Depositing only high-quality sequences correctly annotated with correct species names in BOLD would be the “golden standard” and is crucial for keeping the global barcode library functional and reliable (Collins & Cruickshank, 2013).

A variety of tools for quality assessment of specimen’s taxonomy already exist. Some approaches do their taxonomy assessment based on threshold clustering (Köljalg et al., 2005, 2013; Meier et al., 2006; Ratnasingham & Hebert, 2007, 2013). Others detect inconsistencies (of node or terminal labels) based on tree topology alone (Kozlov, Zhang, Yilmaz, Glöckner, & Stamatakis, 2016; McDonald et al., 2012). In both latter approaches the species boundaries are not inferred from the sequencing data, but are defined by the user in form of terminal node annotations, and used as a topological constraint. However, generally a consistent and straightforward automated flagging and filtering for problematic or inconsistent cases, which is useful for fast processing large amount of data, is missing yet.

Therefore, and based on the previously successful use of taxonomic consistency in different phylogenetic studies to uncover conflicts between classification and phylogenetic tree topology (Ahrens & Vogler, 2008; Hunt & Vogler, 2008; Quicke et al., 2012), an r-package (TaxCI; R Core Team, 2014) was developed that helps to check data for inconsistencies before submission to BOLD.

Optionally, the data check can be done with reference data from BOLD, in order to increase data quality and to detect problems with
the new barcodes in a semi-automatic manner. This analysis pipeline was used to assist the publication of a second (see Hendrich et al., 2015) very large set of barcodes of beetles (Coleoptera) in Germany and central Europe that was created during the "German Barcode of Life (GBOL)" project, comprising nearly 2,900 species and more than 13,000 specimens (Figure 1).

2 | MATERIALS AND METHODS

2.1 | Sampling, sequencing and deposition

Sampling organized with the help of the GBOL web portal (https://www.bolgermany.de; Geiger et al., 2016), voucher treatment, and the taxonomy used (see Data S5) were similar to Hendrich et al. (2015). DNA extraction, polymerase chain reaction and sequencing followed standard protocols (Astrin et al., 2016) and are described together with the newly developed specific primers in detail in Data S5. All data are deposited in BOLD (https://doi.org/10.5883/gcolzfmk) and GenBank (accession numbers KU906118–KU919633) respectively.

2.2 | Data analysis

Our newly generated barcode data were counterchecked for taxonomic consistency against a reference dataset. As a reference a first larger dataset of German Coleoptera barcodes (3,514 species, 15,948 individuals) was used (Hendrich et al., 2015). Since the reference contained many short sequences (<500 bp), tree searches were performed using five different datasets (Table 1) with fastTree 2 (Price, Dehal, & Arkin, 2010): (1) newly generated data only, (2) all reference data, (3) reference data including only sequences longer than 500 bp, (4) combined new data and reference data including all sequences, and (5) combined new data and reference data including only sequences longer than 500 bp. Analyses were conducted on each full set of data (sets 1 to 5) and on separate families except for the monotypic Myxophagan family Sphaeriusidae (species poor families were combined with others with regard to their systematic position; Table S2). Tree searches used the GTR+CAT model. Midpoint rooting was applied before further processing using the R-packages phangorn (Schliep, 2011) and ape (Paradis, Claude, & Strimmer, 2004).

In order to pinpoint misidentifications, contaminations, mislabelling, but also potential cases of introgression, incomplete lineage sorting and instances of polyphyly, the phylogenetic trees obtained were used as phylogenetic backbone to investigate the consistency of the barcoded species, i.e. the monophyly of a priori assigned morphospecies. In a second step, a distance-based clustering was used as implemented in the R-package SPIDER (v.1.3.0; Brown et al., 2012) to gain further evidence for conspecificity or even stronger genetic differentiation (i.e. cryptic diversity).

These procedures were implemented in a new R-package "TaxCI" which is open-source and freely available at https://github.com/eberlejonas/TaxCI. This newly developed analysis tool had as input (1) a rooted dendrogram (fastTree tree), (2) the underlying DNA sequence alignment (PHYLIP format), and (3) a data sheet linking the dendrogram's tip-labels to the a priori given taxonomic information (i.e. assignment of family, genus and species). Based on the input tree topology and five consecutive evaluation steps, a penalty score for all specimens is returned that marks potentially misidentified or contaminated samples (Figure 2).

Step 1 The topology of the dendrogram is analysed with reference to monophyly of the a priori identified species. In analogy to the examination of homoplasy in morphological characters, the membership to a certain a priori identified species was coded as trait state "1" instead of trait state "0." The consistency index (Kluge & Farris, 1969) explores the degree of homoplasy of a trait (or here, of an a priori defined species), and is thus a good expression of "how polyphyletic" a species is in a given tree. All individuals of a given species not grouped as monophylum are marked and assigned with a tci penalty score. They also get highlighted during this step (Figures 3 and 4; 1st box), also those residing in a taxonomically and homogeneously composed clade. Members of the respective species within a heterogeneous cluster get an additional mark in the second step (2nd box). An approach similar to this step was used recently to detect non-monophyly in phylogenetic trees.

FIGURE 1 Map of Germany with the sample locations for the reference data from Hendrich et al. (2015; red circles) and newly added barcode data (blue triangles) (this study)
Table 1: Number of morphospecies, genetic barcode clusters ("OTUs"), tci positive species, homogeneous and heterogeneous cluster, species in multiple homogeneous cluster and the match ratio between genetic cluster and morphospecies from the initial and second TaxCI run on the reference data (Ref data), the newly generated barcodes (new data), and both combined (Comb)

|                  | Morpho-species | Cluster | tci-positive species | Cluster homogen. | Cluster heterogen. | Species in multiple homogen. clusters | Match ratio |
|------------------|----------------|---------|----------------------|------------------|-------------------|--------------------------------------|-------------|
| **Run 1**        |                |         |                      |                  |                   |                                      |             |
| Ref data (all)   | 3,513          | 3,883   | 50                   | 3,835            | 48                | 358                                  | 0.94        |
| Ref data (>500 bp)| 3,510          | 3,884   | 48                   | 3,837            | 47                | 358                                  | 0.94        |
| New data (all)   | 2,978          | 3,399   | 703                  | 2,959            | 440               | 644                                  | 0.73        |
| Comb             | 4,117          | 4,673   | 984                  | 4,079            | 594               | 831                                  | 0.74        |
| Comb (>500 bp)   | 4,115          | 4,676   | 982                  | 4,081            | 595               | 834                                  | 0.74        |
| **Run 2**        |                |         |                      |                  |                   |                                      |             |
| Ref data (all)   | 3,506          | 3,883   | 51                   | 3,836            | 47                | 362                                  | 0.94        |
| Ref data (>500 bp)| 3,503          | 3,884   | 50                   | 3,838            | 46                | 362                                  | 0.94        |
| New data (all)   | 2,846          | 3,272   | 42                   | 3,249            | 23                | 375                                  | 0.94        |
| Comb             | 4,002          | 4,621   | 168                  | 4,528            | 93                | 567                                  | 0.92        |
| Comb (>500 bp)   | 4,000          | 4,623   | 167                  | 4,531            | 92                | 567                                  | 0.92        |
| **Run 2 with mPTP** |              |         |                      |                  |                   |                                      |             |
| Ref data (all)   | 3,506          | 3,401   | 51                   | 3,197            | 204               | 161                                  | 0.88        |
| Ref data (>500 bp)| 3,503          | 3,392   | 50                   | 3,178            | 214               | 159                                  | 0.88        |
| New data (all)   | 2,846          | 2,768   | 42                   | 2,605            | 163               | 151                                  | 0.87        |
| Comb             | 4,002          | 4,211   | 168                  | 3,985            | 226               | 412                                  | 0.87        |
| Comb (>500 bp)   | 4,000          | 4,222   | 167                  | 4,003            | 219               | 411                                  | 0.87        |

and to ascertain the incidence of species non-monophyly in COI barcode sequences (Mutanen et al., 2016). Although species monophyly is not a stringent criterion of species validity (Zhang, Zhang, Zhu, & Yang, 2011), it has been widely shown that most species are monophyletic at least for the barcode gene COI (e.g. Funk & Omland, 2003; Mutanen et al., 2016; Ross, 2014).

Step 2 The TaxCI package uses the function localMinima from the R package spider (Brown et al., 2012) and derives a distance threshold (choice of evolutionary model for distance calculation selectable; here used the default option: K80) from the aligned sequence data for clustering. The actual value corresponds to the first dip in the density of pairwise distances and indicates the transition between intra- and inter-specific distances following the general concept of the barcoding gap. The newly derived clusters are then analysed to infer which of the specimens might be contaminants or wrongly identified. All individuals within a heterospecific cluster are tagged if—

and only if—there are other individuals of this species in at least one additional cluster (3rd box). All individuals marked during step two get also marked in the 3rd box, if this condition is fulfilled.

Step 4 Calculates the relative abundance of a nominal a priori identified species in a heterogeneous cluster as proportion of its individuals relative to the total number of specimens in this cluster. If a particular species has a lower relative abundance than any of the other species in the mixed cluster, all its individuals are tagged in the 4th box (Figure 3). A penalty score of 1–RA (RA—relative abundance of the species) is assigned. This step might be helpful to infer which of the specimens might be contaminants or wrongly identified.

Step 5 All individuals of a species in a homogeneous cluster with members in at least one other homogeneous cluster are marked in the 5th box. This step might indicate wrong identifications but also cryptic diversity.

All results for the specimens that were found to be conspicuous in any of the five steps are written into a text file table along with a penalty score based on the above five steps. For each analysis, we estimated additionally a match ratio (Ahrens et al., 2016): $MR = 2 \times N_{match} / (N_{cluster} + N_{morph})$, where $N_{match}$ is the number of species with exact matches, that is all specimens of one morphospecies (and only these) belong to one cluster entity, and $N_{cluster}$ and $N_{morph}$ are the number of clusters and morphospecies. The initially loaded dendrogram is plotted as text-searchable PDF file (without length limitations), now enriched with the details of the taxonomic congruence checks (Boxes 1–5). Thus, the modified tree file can easily be inspected visually on
FIGURE 2. Flow chart of principle procedures of the \texttt{taxci} script summarizing the four \texttt{R} functions \texttt{tci()}, \texttt{plotParamTCI()}, \texttt{BCCLUSTER()}, and \texttt{plotParamBCC()}. 
any computer and decoupled from the analysis platform. This is an important aspect for involving external collaborators or multiple project partners during the validation of taxon identifications. The output of the TAXCI analysis was used to re-inspect the affected specimens, to correct the species ID data, or to exclude specimens (temporally for further revision of ID, or definitively, in case of contaminations). Finally, the TAXCI analysis was run again with the cleaned dataset to assess the level of improvement in congruence.
3 | RESULTS

A full-length DNA barcode (658 bp) was obtained from 99.4% of the 14,330 specimens from a priori identified species representing 102 of the 103 beetle families known from Germany. The remaining 0.6% of sequences are between 657 and 557 bp. In the initial phylogenetic analysis and subsequent taxci run all 14,330 specimens were included representing 2,970 a priori defined species of 102 families. Based on poor taxci results we removed 120 species in 814 specimens from the BOLD submission pipeline, of which six specimens proved to be contaminations, three specimens had typing errors in their names, while 804 specimens were preliminary excluded to verify once again carefully the identification by a taxonomic specialist. In 32 cases taxonomic inconsistencies were found due to changes in nomenclature since 2012 (in these cases only the metadata file for taxci analysis was corrected, but not the terminal name in the tree; see Data S1 for reference of taxon nomenclature in 2012 vs. 2015). In 67 cases a taxonomic inconsistency was detected in the reference data (i.e. Hendrich et al., 2015). Five hundred and seventy-nine specimens were kept in the new dataset despite problems with taxonomic consistency as data do not allow for rejection of the sample as in the above cases (potential cases of introgression, incomplete lineage sorting, cryptic diversity). These cases require further research, ideally based on an integrative taxonomy toolset, in some cases possibly even resulting in comprehensive taxonomic revisions. In total, 13,516 specimens in 2,846 species were added to the BOLD library, which now includes >4,000 species of beetles occurring in Germany.

The taxci run 1 detected, among the newly generated data of 2,978 morphospecies, more than 23% of the species (n = 703) that were affected by taxonomic inconsistencies (Table 1). This percentage was reduced through dataset pruning and metadata (ID label) correction in the second taxci run to only 1.5% of the species. Accordingly, the number of homogeneous clusters increased by nearly 300 while the heterogeneous clusters dropped by more than 400. The overall match ratio increased from 0.73 (1st run) to 0.94. The same applies to the combined data, where the match ratio improved from 0.74 to 0.92. Most of the remaining taxonomic inconsistency can be ascribed to deep genetic differentiation. The numbers of species in multiple homogeneous genetic clusters remained relatively high (644 vs. 375), and increased particularly with the combination of the sampling of the reference and new data (n = 567). This must be either attributed to cryptic diversity and stronger divergence between populations due to historical isolation (e.g. in glacial refugia) or to possibly inaccurate clustering threshold and/or missing barcoding gap and/or sensitivity to the taxon sampling.

The number of resulting distance-based clusters is not comparable between the full and single family analyses, as the derived threshold for the different families differs strongly (Tables 1 and S2). Therefore, different total numbers of clusters between both approaches were found (4,621 vs. 4,494 for full and family analysis (2nd run) respectively). Nevertheless, match ratios were improved for nearly all families in the 2nd run as well (Table S2).

According to the BOLD BIN report, our dataset contained 337 globally new BINs (with 716 specimens), while 2,758 BINs were already present in the BOLD system (comprising 12,796 specimens from our new dataset and a total of 53,874 specimens) (see Table S1). A detailed overview of the resulting BINs, trees and taxonomic inconsistencies is given in Table S1 and Data S1 and S2.

4 | DISCUSSION

According to the widely accepted generalized lineage concept (GLC; de Queiroz, 1998, 2005, 2007), which is now commonly applied in DNA-based and integrative species delimitation (Camargo, Morando, Avila, & Sites, 2012; Carstens et al., 2013; Edwards & Knowles, 2014; Ence & Carstens, 2011; Jones, 2015; Jones, Aydin, & Oxelman, 2015; Knowles & Carstens, 2007; O’Meara, Ane, Sanderson, & Wainwright, 2006; Yang & Rannala, 2010, 2014), monophyly is not an essential prerequisite for a valid species. However, since a monophyletic species is with a certain probability also more likely a valid species (e.g. Kizirian & Donnelly, 2004), this criterion was used to explore our data for inconsistencies. Therefore, the topology of the obtained gene tree was analysed with reference to the degree of monophyly of the a priori identified species (step 1 of our pipeline). Species detected during this step are non-monophyletic. This can be caused by several natural phenomena (e.g. hybridization, incomplete lineage sorting) but can arise—as an artefact—also from human errors within the barcoding pipeline (e.g. contamination or misidentification of a sample). What is not detected in step 1 are (1) contaminated or misidentified singletons, (2) entire monophyla that appear as sister group of another clade within a cluster (Figure 3j, taxon “E”), or (3) specimens that mirror cryptic diversity (Figure 3m).

The use of a simple distance-based clustering in the steps following the monophyly assessment is because this pipeline was designed for extremely large datasets with thousands of specimens. Methodologically more appropriate methods like statistical parsimony analysis (Templeton, 2001), Poisson tree process modelling (Zhang et al., 2013), and General Mixed Yule Coalescent modelling (Pons et al., 2006) could potentially replace the distance-based clustering as they might be more accurate and also more appropriate since they follow a defined species concept (i.e. phylogenetic species concept). However, they are computationally much more demanding and thus time-consuming. As our approach is intended for data pre-processing, we opted for speed rather than accuracy, and implemented a distance clustering approach as default option. However, alternative species delimitations obtained from external programs can easily be implemented as they get more scalable like the recently developed mptp (Kapli et al., 2016).

The indication of the presence of additional specimens (steps 2 and 3) and their abundance (step 4) in another cluster may help to decide whether there is a case of contamination and which of the specimens are affected by contamination, mislabelling, or misidentification. This latter step is especially helpful to find misidentified singletons. The final analysis step (step 5) indicates those potential misidentifications.
that have gone unnoticed through steps 1–4 or that might be used as indicators for cryptic diversity or strong mtDNA differentiation among populations. Thus, the output from step 5 will be especially valuable for future studies to explore additional evidence on these observed patterns of differentiation.

Data quality, i.e. the match ratio of morphospecies vs. barcode clusters, greatly improved with the fast screening and the subsequent revision of the data and IDs from roughly 73% and 74% to 94 and 92% for our new data and all German beetle barcodes respectively. This is a significant improvement considering that the taxonomic consistency of the reference data was also 94%. Using multi-rate poison tree processing as external method with \textsc{tax2tree} resulting in a lower match ratio of 88% and 87% accompanied by lower amount of homogeneous clusters, but with more heterogeneous clusters than in our standard approach (Table 1). Mainly well separated singletons were lumped by the \textsc{mptp} clustering method (Data S1 and S5).

The inclusion of the reference dataset helped to find more taxonomic problems, which would remain undetected with the new dataset alone. This is especially true for wrongly identified or contaminated singletons. Starting with initial 829 singletons in our BNAII dataset, only 316 remained still as singletons after check with reference data, while 379 singleton cases were solved. However, 134 of the former singletons were now tci positive in the combined analysis. Only the analysis of entire dataset will reveal potential contaminations across family levels, because such specimens as far as in homogeneous clusters will not be noticed in the family analysis alone (1,059 less tci positives in the family analysis compared to the combined analysis). Family trees are suited for inspection by external collaborators. Given the opportunity to perform these quality assessments on-site within 24 hr allows for a proper and timely data curation by the data provider itself.

Tools of quality assessment of specimen’s taxonomy do already exist in \textsc{UNITe} (Köljlaj et al., 2005, 2013; https://unite.ut.ee/index.php), in \textsc{SpeciesIdentifier} (Meier et al., 2006), and in the BOLD systems (Ratnasingham & Hebert, 2007; http://www.boldsystems.org). Taxonomy assessments in these are exclusively based on threshold clusters, while monophyly and taxonomic consistency beyond clusters is not considered in detail. A flagging of problematic specimens is performed only in BOLD systems. Furthermore, other tools with similar tasks so far used for micro-organisms exist (Kozlov et al., 2016; McDonald et al., 2012). The \textsc{tax2tree} approach (McDonald et al., 2012) refers to topological constraints and does not implement the inference of species boundaries (e.g. by threshold clustering) as it was designed to transfer an existing taxonomy to newly generated sequences rather than critically checking species assignments of different sources of data critically against each other. \textsc{SATIVA} (Kozlov et al., 2016) detects inconsistencies (of node or terminal labels) based on the tree topology (i.e. taxon monophyly) considering various classification levels (except species) based on arbitrary threshold clustering, as species level often lacking in metadata annotations of Cyanobacteria. In both approaches the species boundary is taken as topological constraint given of terminals in relation to tree topology, rather than an inference of such a limit from the sequence data, and thus lacking the possibility to infer “cryptic diversity.” In all cases an automated straightforward high-throughput filtering for problematic or inconsistent cases, which is useful for large amount of data, is not yet performed.

\textsc{BIN} assignments from the BOLD system, similarly able to help with data quality assessment (in terms of correct species name assignments), generally have a turnaround time of several weeks. Furthermore, data curation—once uploaded to BOLD—requires additional work from submitters and data curators.

Another advantage of our \textsc{tax2ci} analysis pipeline lies in its capacity to produce tree files in a PDF format that is text-searchable and without page limitation, a feature that is lacking or only fragmentarily available in most above mentioned approaches. Even most current stand-alone tree editing programs do not allow this or have strong limitations for tree length. In the form of such portable files, readable on any desktop computer and also by in terms of molecular systematics untrained taxonomists or citizen scientists, they represent the most valuable linking element for using barcodes. This increases the acceptance and audience for these approaches and issues linked with the data (e.g. taxonomical problems, inconsistencies). Finally, \textsc{tax2ci} offers a more complex output for different ways of data control (numerical, graphical) which allows an either automated and computer-based but also manual/visual data inspection.

5 | CONCLUSIONS

The presented analysis pipeline for a semi-automatic data quality assessment and check, with subsequent revision and re-check, may provide a helpful tool when processing large to very large datasets that can be handled manually only with great difficulty and considerable time investment. While several so far existing tools or web portals offer partly similar functions, our pipeline combines in efficient way the presentation of data (tree output) with useful flagging of problematic samples, high data throughput, and the complex aspects of Molecular Operational Taxonomic Unit (MOTU) capture relying on tree based species delimitation as well as on fast clustering with distance-based algorithms. \textsc{tax2ci} will prove to be a useful tool that helps to generate at forefront high quality barcode data that keep away high correction workload from large reference databases, like BOLD systems, and to bring more people towards an integrative taxonomy where barcode data play a crucial role in biodiversity research, but also, where traditional taxonomy is deeply involved in biodiversity exploration.

ACKNOWLEDGEMENTS

This project was funded by the German Federal Ministry of Education and Research (work package GBOL1; FKZ: 01LI1101A/01LI1501A).

We are thankful to various students, who helped process the beetle specimens in our barcode pipeline. We are thankful also to the following colleagues who helped to complete the specimen library with additional samples: J. Beller, R. Bellstedt, R. Buchbach, H. Bußler, S. Chapuis, L. Brunk, C. Fehse, B. Feldmann, L. Friman, K.-H. Grübel, G. Haas, C. Herbig, O. Jäger, T. Klug, M. Koch, C. Morkel, M. Niehuis,
H.-P. Reike, E. Rößner, and A. Schäfer. Finally, we are thankful to M. Balke, L. Hendrich, and J. Morinière who inspired us for this paper. Our colleague J. Decher kindly checked and improved the English in this paper. We thank the two anonymous referees for their helpful comments on the manuscript. The 1KITE project (1KITE consortium: www.1kite.org) kindly provided access to mitogenome data for hundreds of samples that greatly helped in designing the new degenerate primers LCO1490-JJ3 and HCO2198-JJ3.

AUTHORS’ CONTRIBUTIONS
D.A., J.E., B.R. designed this study; B.R. led project administration and data curation together with D.A., M.F.G., P.G., L.M., J.T.; D.A., W.A., J.E., U.E., F.F., K.H., M.H., T.H., M.J., A.K., A.K., D.K., J.K., F.K., A.S., B.R., D.R., J.S., F.T., A.W. collected/identified samples and provided meta-data; L.M. and J.T. conducted laboratory work; J.E. programmed TaxCl; J.E. and B.R. improved part of code and tested all scripts with the different datasets; D.A., J.E., M.F.G., B.R. compiled the data and conducted the data analyses; D.A., J.J.A., J.E., M.F.G., B.R., J.W.W. drafted the manuscript; All authors contributed critically to the MS draft and gave final approval for publication.

DATA ACCESSIBILITY
Sequence data analysed in this paper are available at BOLD (https://doi.org/10.5883/ds-gcolzfmk) and at GenBank (accession numbers: KU906118–KU919633). The code used in this paper is available on Github (https://github.com/eberelejonas/TaxCl).

REFERENCES
Ahrens, D., Fujisawa, T., Krammer, H.-J., Eberle, J., Fabrizi, S., & Vogler, A. P. (2016). Rarity and incomplete sampling in DNA-based species delimitation. Systematic Biology, 65, 478–494.
Ahrens, D., Monaghan, M. T., & Vogler, A. P. (2007). DNA-based taxonomy for associating adults and larvae in multi-species assemblages of chafers (Coleoptera: Scarabaeidae). Molecular Phylogenetics and Evolution, 44, 436–449.
Ahrens, D., & Vogler, A. P. (2008). Towards the phylogeny of chafers (Sericini): Analysis of alignment-variable sequences and the evolution of segment numbers in the antennal club. Molecular Phylogenetics and Evolution, 47, 783–798.
Astrin, J. J., Höfer, H., Spelda, J., Holstein, J., Bayer, S., Hendrich, L., ... Muster, C. (2016). Towards a DNA barcode reference database for spiders and harvestmen of Germany. PLoS ONE, 11, e0162624. doi:10.1371/journal.pone.0162624.
Ballard, J. W., & Whitlock, M. C. (2004). The incomplete natural history of mitochondria. Molecular Ecology, 13, 729–744.
Bergsten, J., Bilton, D. T., Fujisawa, T., Elliott, M., Monaghan, M. T., Balke, M., ... Vogler, A. P. (2012). The effect of geographical scale of sampling on DNA barcoding. Systematic Biology, 61, 851–869. doi:10.1093/sysbio/sys037.
Blagoev, G. A., deWaar, J. R., Ratnasingham, S., deWaar, S. L., Lu, L., Robertson, J., ... Hebert, P. D. N. (2016). Untangling taxonomy: A DNA barcode reference library for Canadian spiders. Molecular Ecology Resources, 16, 325–341.
Brown, S. D. J., Collins, R. A., Boyer, S., Lefort, M.-C., Malumbres-Olarte, J., Vink, C. J., & Cruickshank, R. H. (2012). SPIDER: An R package for the analysis of species identity and evolution with particular reference to DNA barcoding. Molecular Ecology Resources, 12, 562–565.
Camargo, A., Morando, M., Avila, L. J., & Sites, J. W. Jr (2012). Species delimitation with ABC and other coalescent-based methods: A test of accuracy with simulations and an empirical example with lizards of the Liolemaus darwini complex (Squamata: Liolemaeidae). Evolution, 66, 2834–2849.
Carstens, B. C., Pelletier, T. A., Reid, N. M., & Satler, J. D. (2013). How to fail at species delimitation. Molecular Ecology, 22, 4369–4383.
Collins, R. A., & Cruickshank, R. H. (2013). The seven deadly sins of DNA barcoding. Molecular Ecology Resources, 13, 969–975. doi:10.1111/1755-0998.12046.
de Queiroz, K. (1998). The general lineage concept of species, species criteria, and the process of speciation. In D. J. Howard & S. H. Berlocher (Eds.), Endless forms: Species and speciation (pp. 57–75). New York, NY: Oxford University Press.
de Queiroz, K. (2005). Ernst Mayr and the modern concept of species. Proceedings of the National Academy of Sciences USA, 102, 6600–6607.
de Queiroz, K. (2007). Species concepts and species delimitation. Systematic Biology, 56, 879–886.
Dowton, M., Meiklejohn, K., Cameron, S. L., & Wallman, J. (2014). A preliminary framework for DNA barcoding, incorporating the multispecies coalescent. Systematic Biology, 63, 639–644.
Edwards, D. L., & Knowles, L. L. (2014). Species detection and individual assignment in species delimitation: Can integrative data increase efficacy? Proceedings of the Royal Society B, 281, 20132765.
Ence, D. D., & Carsten, B. C. (2011). SpedeSTEM: A rapid and accurate method for species delimitation. Molecular Ecology Resources, 11, 473–480.
Etzler, F. E., Wanner, K. W., Morales-Rodriguez, A., & Iwye, M. A. (2014). DNA barcoding to improve the species-level management of wireworms (Coleoptera: Elateridae). Journal of Economic Entomology, 107, 1476–1485.
Freitag, H. (2013). Ancyronyx Ericsson, 1847 (Coleoptera, Elmidae) from Mindoro, Philippines, with description of the larvae and two new species using DNA sequences for the assignment of the developmental stages. ZooKeys, 321, 35–64.
Funk, D. J., & Omland, K. E. (2003). Species-level paraphyly and polyphyly: Frequency, causes, and consequences, with insights from animal mitochondrial DNA. Annual Review of Ecology, Evolution, and Systematics, 34, 397–423.
García-Robledo, C., Kuprewicz, E. K., Staines, C. L., Kress, W. J., & Erwin, T. L. (2013). Using a comprehensive DNA barcode library to detect novel egg and larval host plant associations in a Cephaloleia rolled-leaf beetle (Coleoptera: Chrysomelidae). Biological Journal of the Linnean Society, 110, 189–198.
Geiger, M. F. G., Astrin, J. J., Borsch, T., Burkhardt, U., Grobe, P., Hand, R., ... Wägele, W. (2016). How to tackle the molecular species inventory for an industrialized nation—Lessons from the first phase of the German Barcode of Life initiative GBOL (2012-2015). Genome, 59, 661–670. doi:10.1139/gen-2015-0185.
Hebert, P. D. N., & Gregory, T. R. (2005). The promise of DNA barcoding for taxonomy. Systematic Biology, 54, 852–859.
Hendrich, L., Morinière, J., Haszprunar, G., Hebert, P. D. N., Hausmann, A., Köhler, F., & Balke, M. (2015). A comprehensive DNA barcode database for Central European beetles with a focus on Germany: Adding more than 3,500 identified species to BOLD. Molecular Ecology Resources, 15, 795–818.
Hunt, T., & Vogler, A. P. (2008). A protocol for large-scale rRNA sequence analysis: Towards a detailed phylogeny of Coleoptera. Molecular Phylogenetics and Evolution, 47, 289–301.
Jones, G. (2015). Species delimitation and phylogeny estimation under the multispecies coalescent. bioRxiv, https://doi.org/10.1101/010199 (accessed 1 March 2016).
Jones, G., Aydin, Z., & Oxelman, B. (2015). DISSECT: An assignment-free Bayesian discovery method for species delimitation under the multispecies coalescent. Bioinformatics, 31, 991–998.
