Received: 8 May 2019 | Revised: 9 September 2019 | Accepted: 11 September 2019

DOI: 10.1111/jzs.12347

ORIGINAL ARTICLE

JOURNAL OF ZOOLOGICAL SYSTEMATICS AND EVOLUTIONARY RESEARCH

WILEY

QUIDDICH: QUick IDentification of DIagnostic CHaracters

A. Luise Kühn[1] | Martin Haase[2]

[1]Institute of Mathematics and Computer Science, University of Greifswald, Greifswald, Germany

[2]Vogelwarte, Zoological Institute and Museum, University of Greifswald, Greifswald, Germany

Correspondence
A. Luise Kühn, Institute of Mathematics and Computer Science, University of Greifswald, Walther-Rathenau-Straße 47, D-17489 Greifswald, Germany.
Email: luise.kuehn@uni-greifswald.de

Abstract

With the advent of molecular genetic methods, an increasing number of morphologically cryptic taxa has been discovered. The majority of them, however, remains formally undescribed and without a proper name although their importance in ecology and evolution is increasingly being acknowledged. Despite suggestions to complement traditional descriptions with genetic characters, the taxonomic community appears to be reluctant to adopt this proposition. As an incentive, we introduce QUIDDICH, a tool for the QUick IDentification of DIgnostic CHaracters, which automatically scans a DNA or amino acid alignment for those columns that allow to distinguish taxa and classifies them into four different types of diagnostic characters. QUIDDICH is a system-independent, fast and user-friendly tool that requires few manual steps and provides a comprehensive output, which can be included in formal taxonomic descriptions. Thus, cryptic taxa do not have to remain in taxonomic crypsis and, bearing a proper name, can readily be included in biodiversity assessments and ecological and evolutionary analyses. QUIDDICH can be obtained from the comprehensive R archive network (CRAN, https://cran.r-project.org/package=quiddich).

KEYWORDS

Characteristic Attribute Organization System, diagnostic (genetic) characters, DNA taxonomy, integrative taxonomy, SPecies IDentity and Evolution in R

1 | INTRODUCTION

In traditional taxonomy, taxa are usually described based on morphological and anatomical features (Jörger & Schrödl, 2013). However, since the advent of molecular genetic methods, in particular PCR, an exponentially increasing number of morphologically hard or impossible to distinguish animal and plant species, so-called cryptic species, has been discovered and gained recognition in ecology, evolutionary and conservation biology (Bickford et al., 2007; Struck et al., 2018). Unfortunately, the vast majority of cryptic species remains formally undescribed and without a proper name (Schlick-Steiner et al., 2007). It has been suggested to complement the formal descriptions of such morphologically cryptic species by including, for example, behavioral, ecological, biogeographic or, in particular, genetic data, that is, the very same data that often have led to their discovery. But although this pluralistic approach, which is referred to as integrative taxonomy (Dayrat, 2005; Padial, Miralles, De la Riva, & Vences, 2010; Schlick-Steiner et al., 2010), is frequently applied to identify taxa, the formal descriptions remain undone in probably the majority of cases. Botanists and zoologists appear to be particularly reluctant to include genetic data, although the nomenclatural codes are indifferent regarding the nature of data used in descriptions (International Commission on Zoological Nomenclature, 1999; Turland et al., 2018). In fact, Renner (2016) has found only 98 descriptions of species of plants, animals and fungi containing DNA data as of November 2015. In a fundamental paper, Jörger & Schrödl (2013, 2014) have provided an important starting point on how to deploy genetic data given in the form of alignments in which the columns represent the positional homology assumptions. To extract diagnostic characters, that is, those columns that are suitable to distinguish a taxon of interest from the remaining ones, they used the Characteristic Attribute Organization System (CAOS; Sarkar, Planet, & Desalle, 2008). Alternatively, the R package SPIDER (Brown et al., 2012) could be used for this purpose (Jörger & Schrödl 2014). As neither of the programs has been specifically designed for taxonomic applications, it is not surprising that one encounters problems in this particular context.

Characteristic attribute organization system provides a detailed output and distinguishes between different types of diagnostic characters (see below). However, it requires a fully resolved phylogenetic tree, which, for the purpose of sequence-based taxonomy, has to be rearranged so that the taxon of interest becomes outgroup to the remaining sequences (Jörger & Schrödl, 2013, 2014). This essentially wrong or definitely suboptimal tree serves as a guiding structure for the sequence comparisons conducted along the tree hierarchy. Eventually, the only taxon whose sequences are compared to all other sequences is the outgroup taxon. Hence, only for the latter do we get the comprehensive collection of diagnostic positions. If we are interested in a second taxon, we have to rearrange the tree again and rerun the analysis. A second issue concerns the so-called symplesiomorphy filter. Since one of CAOS's original purposes is to classify a novel sequence into a tree, only apomorphic states are considered as useful information, whereas symplesiomorphic states are removed (Sarkar et al., 2002). The polarity, however, is tree-dependent. As the manipulated tree is arbitrary, the polarity of the character states is arbitrary as well. In addition, CAOS does not distinguish between gaps and masked alignment entries (Jörger & Schrödl, 2014). Both are considered as missing data, even though a masked entry indicates uncertainty about the true state, while a gap denotes an evolutionary event with a certain outcome, namely a deletion or insertion. A final disadvantage concerns the implementation in the software RubyCAOS (Sarkar et al., 2008), which is only available for Mac OS X 10.6 + and Linux. This system dependency limits its application. Additionally, as of November 2018, the Linux version produces an incomplete output as it is notably missing (among others) the CAOS groupFile.txt that is required to identify which part of the output belongs to the taxon of interest. Also, the online version of CAOS that was announced by Sarkar et al., (2008) is no longer available.

The alternative for CAOS that is currently available is the function *nucDiag* from the R package SPecies IDentity and Evolution in R (SPIDER) (Brown et al., 2012). On one side, *nucDiag* overcomes most of CAOS's flaws making the application easier, faster and system- as well as phylogeny-independent (i.e., SPIDER does not require a guide tree). On the other side, its output is not as comprehensive, because it considers only two types of diagnostic characters (see below) and only returns the alignment positions of the identified characters without any information on the states that are characteristic for the taxon of interest. Additionally, it extracts diagnostic characters for every single taxon that is contained in the dataset. This may cause unnecessary computational costs if the user is just interested in one or a few taxa. The third disadvantage is again the treatment of gaps and masked alignment entries. Both of them are considered as "valid" character states, although at least the latter should definitely not be treated as such.

In order to provide a tool for the QUick IDentification of DIagnostic CHaracters, which overcomes the drawbacks while at the same time preserving the useful conceptual aspects of existing software, we developed the R package QUIDDICH. QUIDDICH is system-independent, easy to implement, fast, and produces a detailed output. Extending the concepts of CAOS and SPIDER, it can also deliver pairwise diagnostic characters, that is, characters that are suitable to distinguish pairs of taxa (Zielske & Haase, 2015). As genetic data have identified also higher cryptic taxa, for example Ecdysozoa and Lophotrochozoa, the subclades of Protostomia (Aguinaldo et al., 1997; Philippe, Lartillot, & Brinkmann, 2005), which may be more robustly analyzed on the protein level, we implemented functions that can search through both DNA and amino acid alignments. We hope that with an appropriate tool at hand taxonomists will no longer hesitate to include genetic data in descriptions and diagnoses of morphologically or otherwise hard or impossible to define taxa. Thus, cryptic taxa do not have to remain in taxonomic crypsis (Schlick-Steiner et al., 2007) and, bearing a proper name, can readily be included in biodiversity assessments and ecological and evolutionary analyses.

2 | DATA INPUT AND OUTPUT

QUIDDICH requires an alignment (either nucleotides or amino acids) as well as a taxon vector whose *i*-th entry is the name of the taxon that the *i*-th row belongs to. Alignments may contain the IUPAC codes for the bases (resp. amino acids), – for gaps, and *N* (resp. *X*) for missing or ambiguous states or parts of the alignment that are to be masked, for example, in case of alignment ambiguities. The alignment must be stored using the classes DNAbin (nucleotides) or AAbin (amino acids) of the APE package (Paradis & Schliep, 2018). Fasta files can be imported and converted using adegenet's function fasta2DNAbin (Jombart, 2008). After specifying the taxa of interest, QUIDDICH's functions *diagCharNA* and *diagCharAA* can be used to extract four different types of diagnostic genetic characters. Assuming that $states\,(i,j)$ and $states\,(rest,j)$ denote the sets of all states that are present in the *j*-th column of the alignment in any row that belongs to taxon *i* or the remaining taxa, respectively, these types are defined as follows:

Definition 1 Type 1 characters distinguish each individual of the taxon *i* of interest from all individuals of the remaining taxa and are fixed for one state in taxon *i*. Mathematically, the *j*-the column of a given labeled alignment is a *type 1 character of taxon i* if (a) $N\left(resp.X\right) \notin states\,(rest,j)$, (b) $states\,(i,j) \cap states\,(rest,j) = \emptyset$, and (c) $states\,(i,j) = \{z\}$ with $z \neq N\left(resp.X\right)$.

Definition 2 Type 2 characters distinguish each individual of taxon *i* from all individuals of the remaining taxa and are not fixed for one state in taxon *i*. Mathematically, the *j*-th column of a given labeled alignment is a *type 2 character of taxon i* if (a)

$N\,(\text{resp.}X) \notin states\,(rest,j)$, (b) $states\,(i,j) \cap states\,(rest,j) = \emptyset$, (c) $N\,(\text{resp.}X) \notin states\,(i,j)$, and (d) $|states\,(i,j)| \geq 2$.

Definition 3 Type 3 characters distinguish some (but not all) individuals of taxon *i* from all individuals of the remaining taxa. Mathematically, the *j*-th column of a given labeled alignment is a *type 3 character of taxon i* if (a) $N\,(\text{resp.}X) \notin states\,(rest,j)$, (b) $\exists z \in states\,(i,j)$ with $z \neq N\,(\text{resp.}X)$ and $z \notin states\,(rest,j)$, and (c) it is not a type 1 or 2 character of taxon *i*.

Definition 4 Type 4 (or pairwise diagnostic) characters distinguish each individual of taxon *i* from all individuals of at least one (but not all) other taxon while being fixed in both the taxon of interest and the compared taxa. Mathematically, the *j*-th column of a given labeled alignment is a *type 4 (or pairwise diagnostic) character of taxon i* if (a) $states\,(i,j) = \{z\}$ with $z \neq N\,(\text{resp.}X)$, (b) there is a taxon *l* with $l \neq i$, such that $states\,(l,j) = \{y\}$ with $y \neq N\,(\text{resp.}X)$ and $y \neq z$, and (c) it is not a type 1, 2, or 3 character of taxon *i*.

It is to note that QUIDDICH's type 1 and type 2 characters are similar to CAOS's homogeneous and heterogeneous simple pure characteristic attributes (CAs), while type 3 characters are similar to simple private CAs. SPIDER combines the first two types as pure, simple diagnostic nucleotides, but does not consider type 3. Apart from this, neither CAOS nor SPIDER considers type 4 characters.

The reasoning behind the definitions is as follows: If an arbitrary taxon $l \neq i$ is masked at position *j*, it is impossible to know for sure that taxon *i* and taxon *l* do not share any character states at this position as $N\,(\text{resp.}X)$ denotes an unknown state that may be replaced by any other symbol. Hence, the first condition in Definitions 1 to 3 and the second condition $states\,(l,j) = \{y\} \neq \{N\,(\text{resp.}X)\}$ from Definition 4 are necessary to ensure the distinctness of taxon *i* and the taxa it is compared to. In addition to this, diagnostic characters must fulfill $z \neq N\,(\text{resp.}X)$ for at least one state $z \in states\,(i,j)$, see Condition (c) of Definitions 1 and 2, Condition (b) of Definition 3, and Condition (a) of Definition 4. This is necessary because a state that is unknown cannot be characteristic for taxon *i*. The last condition in each definition ensures that a character cannot be of more than one type.

The definition of type 4 characters extends the suggestion of Zielske and Haase (2015) by adding Condition c) and considering indels. It is also to note that type 4 characters are not "symmetric," that is, if the *j*-th column is found to be a type 4 character of taxon *i* when being compared to taxon *l*, it is not necessarily a type 4 character of taxon *l* when being compared to taxon *i*.

The output of the functions is for each taxon *i* of interest a set of tuples (j,t,Z,Y), each one representing one identified diagnostic character with *j* denoting its alignment position, *t* its type, Z the set of states that are characteristic for taxon *i*, that is, $Z = \{z \in states\,(i,j)\ |\, z \neq N\,(\text{resp.}X) \text{ and } z \notin states\,(rest,j)\}$ in case of type 1, 2, and 3 characters and $Z = states\,(i,j) \neq \{N\,(\text{resp.}X)\}$ in case of type 4 characters, and *Y* being the set of taxa that fulfill Condition (b) of Definition 4 (only relevant for type 4 characters). The algorithm on which the functions are based and its proof of correctness can be found in the Appendix S1. If the user chooses that gaps shall not be considered as "valid" character states, that is, they cannot be characteristic for a taxon of interest, the calculations and the output are adjusted accordingly. In addition to this, QUIDDICH's function *changesAA* can be used to identify those diagnostic characters in a nucleotide alignment of protein-coding loci that also cause diagnostic characters in the corresponding amino acid alignment.

3 | PERFORMANCE

Not only does QUIDDICH overcome CAOS's and SPIDER's drawbacks regarding the identification of diagnostic characters outlined above, it is also faster. Assume that a labeled alignment is given with *r* and *c* being the number of its rows and columns. Additionally, assume that the dataset contains *t* taxa, of which *s* are set as taxa of interest. To extract type 1, 2, or 3 characters, the functions *diagCharNA* and *diagCharAA* of the QUIDDICH package have an overall runtime in $O\,(rc+scr)$, because they first scan the alignment for polymorphic sites, which can be done in $O(rc)$, before extracting for each combination of taxon of interest and polymorphic site the two sets $states\,(i,j)$ and $states\,(rest,j)$, which can be done in $O(scr)$.

The runtime of SPIDER, which is in $O(rc+tcr)$, can be calculated similarly. The difference is that SPIDER considers all taxa in the dataset one after the other, while QUIDDICH restricts the calculation to the taxa of interest.

The runtime of CAOS is in $O\,(rs+scr^2)$ not including the manual adjustments to the tree that have to be made beforehand. The algorithm starts by numbering the nodes of the tree and conducting a Fitch optimization (Williams & Fitch, 1990) on it, both of which can be done in $O(r)$. Then, it proceeds from the root toward the leaves calculating for each inner node *n* with the children n_1 and n_2 and each alignment column *j* the sets $states\,(n_1,j)$ and $states\,(n_2,j)$. This has to be repeated for each of the $(r-1)$ inner nodes of the tree and each alignment column, leading to a calculation time in $O(rcr)$. In total, we have a runtime in $O\,(s \cdot (r+rcr))$, which can be rewritten to $O\,(rs+scr^2)$.

4 | APPLICATION OF QUIDDICH

To examine the practicality of QUIDDICH, we investigated three datasets. The first one was an alignment of *cytochrome c oxidase I (COI)* of nine *Pontohedyle* Golikov & Starobogatov, 1972, species, interstitial marine slugs, analyzed by Jörger and Schrödl (2013) using CAOS. It can be retrieved as electronic supplementary material of their paper. Applying the function *diagCharNA* searching for all type 1 and 2 characters delivered the same results as in the paper.

The second dataset (http://purl.org/phylo/treebase/phylows/study/TB2:S15532/) was an alignment of *COI*, *16S* rRNA gene sequences, and internal transcribed spacer 2 (*ITS2*) of small, inconspicuous New Caledonian freshwater gastropods of the family Tateidae

TABLE 1 Numbers of diagnostic characters of types 1–3 for four genera of Tateidae

Genus	Type 1	Type 2	Type 3
Meridiopyrgus	0	0	9
Rakiurapyrgus	6	0	0
Hadopyrgus	4	0	0
Opacuincola	0	0	17

that was analyzed by Zielske and Haase (2015) in order to complement diagnoses of morphologically practically indistinguishable genera, again using CAOS. Searching for all type 1 and 4 characters and setting the parameter *gapValid* to false delivered the same diagnostic positions as in the paper.

The third dataset (Appendix S1) was an alignment of *COI* comprising twelve tateid genera from New Zealand. A foregoing analysis of Haase (2008) found that *Meridiopyrgus* Haase, 2008, and *Rakiurapyrgus* Haase, 2008, as well as *Hadopyrgus* Haase, 2008, and *Opacuincola* Ponder, 1966, are almost identical regarding morphological features, while being phylogenetically very distinct. Thus, it was indicated to complement the morphological descriptions with a set of diagnostic genetic characters. The numbers of all type 1, 2, and 3 characters delivered is given in Table 1.

Additionally, QUIDDICH identified 26 type 4 characters that distinguish *Hadopyrgus* from *Opacuincola*, 28 type 4 characters that distinguish *Opacuincola* from *Hadopyrgus*, 37 type 4 characters that distinguish *Meridiopyrgus* from *Rakiurapyrgus*, and 33 type 4 characters that distinguish *Rakiurapyrgus* from *Meridiopyrgus*. The comprehensive output is given as (Tables S1–S3).

5 | OBTAINING QUIDDICH

QUIDDICH is a package of the statistical programming environment R (R Core Team, 2013), which can be downloaded from the comprehensive R archive network (CRAN, https://cran.r-project.org/package=quiddich) for all computing platforms. The package can also be obtained by entering the following commands into R's console:

```
>install.packages(“quiddich”)
>library(quiddich)
```

It depends on the package “APE,” which provides the necessary data structures and basic functions. If APE is not already installed on the system, it is automatically installed when the above commands are run. The download of QUIDDICH includes a manual.

ACKNOWLEDGEMENTS

We would like to thank Mareike Fischer for her input on this contribution and two reviewers for their helpful comments.

ORCID

Martin Haase https://orcid.org/0000-0002-9281-8752

REFERENCES

Aguinaldo, A. M. A., Turbeville, J. M., Linford, L. S., Rivera, M. C., Garey, J. R., Raff, R. A., & Lake, J. A. (1997). Evidence for a clade of nematodes, arthropods and other moulting animals. *Nature*, *387*, 489–493. https://doi.org/10.1038/387489a0

Bickford, D., Lohman, D. J., Sodhi, N. S., Ng, P. K. L., Meier, R., Winker, K., … Das, I. (2007). Cryptic species as a window on diversity conservation. *Trends in Ecology & Evolution*, *22*, 148–155. https://doi.org/10.1016/j.tree.2006.11.004

Brown, S. D. J., Collins, R. A., Boyer, S., Lefort, M., 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. https://doi.org/10.1111/j.1755-0998.2011.03108.x

Dayrat, B. (2005). Toward integrative taxonomy. *Biological Journal of the Linnean Society*, *85*, 407–415. https://doi.org/10.1111/j.1095-8312.2005.00503.x

Haase, M. (2008). The radiation of hydrobiid gastropods in New Zealand: A revision including the description of new species based on morphology and mtDNA sequence information. *Syst. Biodiv.*, *6*, 99–159. https://doi.org/10.1017/S1477200007002630

International Commission on Zoological Nomenclature (1999). *International Code of Zoological Nomenclature*, 4th ed. London, UK: The International Trust for Zoological Nomenclature.

Jombart, T. (2008). adegenet: A R package for the multivariate analysis of genetic markers. *Bioinformatics*, *24*, 1403–1405. https://doi.org/10.1093/bioinformatics/btn129

Jörger, K. M., & Schrödl, M. (2013). How to describe a cryptic species? Practical challenges of molecular taxonomy. *Frontiers in Zoology*, *10*, 59. https://doi.org/10.1186/1742-9994-10-59

Jörger, K. M., & Schrödl, M. (2014). How to use CAOS software for taxonomy? A quick guide to extract diagnostic nucleotides or amino acids for species descriptions. *Spixiana*, *37*, 21–26.

Padial, J. M., Miralles, A., De la Riva, I., & Vences, M. (2010). The integrative future of taxonomy. *Frontiers in Zoology*, *7*, 16. https://doi.org/10.1186/1742-9994-7-16

Paradis, E., & Schliep, K. (2018). ape 5.0: An environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*, *35*, 526–528. https://doi.org/10.1093/bioinformatics/bty633

Philippe, H., Lartillot, N., & Brinkmann, H. (2005). Multigene analyses of bilaterian animals corroborate the monophyly of Ecdysozoa, Lophotrochozoa, and Protostomia. *Molecular Biology and Evolution*, *22*, 1246–1253. https://doi.org/10.1093/molbev/msi111

R Core Team (2013). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Available at http://www.R-project.org/

Renner, S. (2016). A return to Linnaeus's focus on diagnosis, not description: The use of DNA characters in the formal naming of species. *Systematic Biology*, *65*, 1085–1095. https://doi.org/10.1093/sysbio/syw032

Sarkar, I. N., Planet, P. J., & Desalle, R. (2008). CAOS software for use in character-based DNA barcoding. *Molecular Ecology Resources*, *8*, 1256–1259. https://doi.org/10.1111/j.1755-0998.2008.02235.x

Sarkar, I. N., Thornton, J. W., Planet, P. J., Figurski, D. H., Schierwater, B., & DeSalle, R. (2002). An automated phylogenetic key for classifying homeoboxes. *Molecular Phylogenetics and Evolution*, *24*, 388–399. https://doi.org/10.1016/S1055-7903(02)00259-2

Schlick-Steiner, B. C., Seifert, B., Stauffer, C., Christian, E., Crozier, R. H., & Steiner, F. M. (2007). Without morphology, cryptic species stay in taxonomic crypsis following discovery. *Trends in Ecology & Evolution*, *22*, 391–392. https://doi.org/10.1016/j.tree.2007.05.004

Schlick-Steiner, B. C., Steiner, F. M., Seifert, B., Stauffer, C., Christian, E., & Crozier, R. H. (2010). Integrative taxonomy: A multisource approach to exploring biodiversity. *Annual Review of Entomology*, *55*, 421–438. https://doi.org/10.1146/annurev-ento-112408-085432

Struck, T. H., Feder, J. L., Bendiksby, M., Birkeland, S., Cerca, J., Gusarov, V. I., … Dimitrov, D. (2018). Finding evolutionary process hidden in cryptic species. *Trends in Ecology & Evolution*, *33*, 153–163. https://doi.org/10.1016/j.tree.2017.11.007

Turland, N. J., Wiersema, J. H., Barrie, F. R., Greuter, W., Hawksworth, D. L., Herendeen, P. S., … Smith, G. F. (2018): International Code of Nomenclature for algae, fungi, and plants (Shenzhen Code) adopted by the Nineteenth International Botanical Congress Shenzhen, China, July 2017. Regnum Vegetabile 159. Glashütten, Germany: Koeltz Botanical Books.

Williams and Fitch, 1990Williams, P. L., & Fitch, W. M. (1990). Finding the minimal change in a given tree. In: Dress, A., & von Haeseler, A. (Eds.), *Trees and Hierarchical Structures* (pp. 60–74). Berlin, Germany: Springer.

Zielske, S., & Haase, M. (2015). Molecular phylogeny and a modified approach of character-based barcoding refining the taxonomy of New Caledonian freshwater gastropods (Caenogastropoda, Truncatelloidea, Tateidae). *Molecular Phylogenetics and Evolution*, *89*, 171–181. https://doi.org/10.1016/j.ympev.2015.04.020

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Kühn AL, Haase M. QUIDDICH: QUick IDentification of DIagnostic CHaracters. *J Zool Syst Evol Res*. 2020;58:22–26. https://doi.org/10.1111/jzs.12347