MOLD, a novel software to compile accurate and reliable DNA diagnoses for taxonomic descriptions

A.E. Fedosov, Guillaume Achaz, Andrey Gontchar, Nicolas Puillandre

MOLECULAR ECOLOGY
RESOURCES

MOLD, a novel software to compile accurate and reliable DNA diagnoses for taxonomic descriptions

Journal:	*Molecular Ecology Resources*
Manuscript ID	MER-21-0531.R1
Manuscript Type:	Resource Article
Date Submitted by the Author:	n/a
Complete List of Authors:	Fedosov, Alexander; A N Severtsov Institute of Ecology and Evolution RAS, Morphology and Ecology of Marine Invertebrates Achaz, Guillaume; Institut Systématique Evolution Biodiversité (ISYEB), Muséum national d'Histoire naturelle, CNRS, Sorbonne Université, EPHE, 57 rue Cuvier, CP 26, 75005 Paris, France Gontchar, Andrey; Dmitry Rogachev National Medical Research Center of Pediatric Hematology Oncology and Immunology, Molecular Immunology Laboratory Puillandre, Nicolas; MNHN, Systematique & Evolution;
Keywords:	DNA character, DNA diagnosis, Taxonomy, Systematics

MOLD, a novel software to compile accurate and reliable DNA diagnoses for taxonomic descriptions

Alexander Fedosov[1,2], Guillaume Achaz[2,3,4], Andrey Gontchar[5], Nicolas Puillandre[2]

[1] A.N. Severtsov Institute of Ecology and Evolution, Russian Academy of Sciences, Leninsky prospect 33, 119071 Moscow, Russia.

[2] Institut Systématique Evolution Biodiversité (ISYEB), Muséum national d'Histoire naturelle, CNRS, Sorbonne Université, EPHE, Université des Antilles, 57 rue Cuvier, CP 26, 75005 Paris, France.

[3] UMR7206 Eco-Anthropologie, Université de Paris–CNRS–MNHN, Paris.

[4] UMR7241 Centre Interdisciplinaire de Recherche en Biologie, Collège de France–CNRS–INSERM, Paris.

[5] Molecular Immunology Laboratory, Dmitry Rogachev National Medical Research Center of Pediatric Hematology, Oncology and Immunology, Samory Mashela street 1, 117997 Moscow, Russia.

ABSTRACT

DNA data are increasingly used for phylogenetic inference, taxa delimitation and identification~~. Nevertheless~~, but scarcely~~, their use~~ for formal description of taxa, despite its incorporation in taxonomic routine promises undisputable merits ~~remains scarce and inconsistent~~. The uncertainty regarding the robustness of DNA diagnoses, however, remains a major impediment to their use. ~~Whether use of DNA sequence data will benefit taxonomy depends on our ability to transform it into accurate and robust diagnoses. However, the reliability of DNA diagnoses has never been addressed.~~ We develop~~ed~~ a new program, MOLD that identifies diagnostic nucleotide combinations (DNCs) in DNA sequence alignments for selected taxa, to be used as formal diagnoses of these taxa. To test the robustness of DNA diagnoses, we ~~carried~~ carry out iterated haplotype subsampling ~~on~~ for selected query species in published DNA data sets of varying complexity. We quantifi~~ed~~y the diagnosis' reliability by diagnosing each query subsample and then checking if this diagnosis remained valid against the entire data set. ~~Two subsampling regimes were tested: in h-sampling, haplotype per species composition varied, but the set of species remained constant; in hssp-sampling, samples varied in both species composition and the subset of haplotypes per species.~~

We demonstrate that widely used types of diagnostic DNA characters are often absent for a query taxa or are not sufficiently reliable. We thus propose a new type of DNA diagnosis, termed 'redundant DNC' (or rDNC), which takes into account unsampled genetic diversity and constitutes a much more reliable ~~diagnosis~~descriptor of a taxon. ~~We~~ MOLD successfully retrieve~~d~~s rDNCs for all but two species in the analyzed data sets, even in those comprising hundreds of species. MOLD shows unparalleled efficiency in large DNA data sets and is the only available software capable of compiling DNA diagnoses that suit pre-defined criteria of reliability.

Running title: MOLD: a novel tool for DNA diagnoses in taxonomy

Key words: DNA diagnosis, DNA character, DNA barcoding, taxonomy, description of taxa.

1. Introduction

The formal description of living organisms is an essential procedure to communicate their identities to the community of scientists and stakeholders, and is regulated by the relevant nomenclatural codes. Formally, a newly introduced name must be made *available*, the crucial requirement for which is the provision of a diagnosis and/or of a description. In practice, a diagnosis is a summary of the characters that differentiate the new nominal taxon from related or similar taxa (ICZN Article 13.1.3), and ideally is sufficient for the reliable recognition of this taxon.

Whereas traditionally descriptions of taxa are mainly based on morphological data (Dunn et al. 2003; Cook et al. 2010), non-morphological characters, in particular DNA characters, are equally accepted by all nomenclatural codes (Cook et al. 2010; Renner, 2016). The amount of DNA sequence data available to taxonomists has steadily grown over the last decades, and currently accumulates at an ever-increasing rate following the recent advent of high-throughput sequencing. Therefore, DNA sequence data are often more accessible than rare taxonomic expertise (Cook et al. 2010) and it is not surprising that DNA data is now widely used in phylogenetics, species delimitation (Fujita et al. 2012; Puillandre et al. 2012; Pante et al. 2015) and specimen identification (Herbert et al. 2003; Janzen et al. 2009; Goldstein & De Salle 2011). Conversely, the use of DNA data in formal descriptions remains scarce, the number of species described with DNA data being two orders of magnitude smaller than those described without (Renner 2016; Kühn & Haase 2020). However, the inclusion of DNA data in taxonomic descriptions has the potential to greatly increase their quality and usability, and is conceptually sound, as long as taxa discovery relies on a comprehensive integrative taxonomy framework. In this perspective, some recent publications strongly recommend that the relevant taxonomic codes should promote the use of DNA-based diagnoses (Renner 2016). There is a pressing need to establish clear universal requirements for DNA diagnoses, especially in view of the recent notorious article (Sharkey et al. 2021), where a misuse of molecular data led to the publication of multiple, arguably very problematic taxonomic acts (Meier et al. 2021).

Until recently, the lack of an efficient software for identification of diagnostic DNA characters was a major practical impediment to the use of DNA data in taxonomic descriptions. The recent releases of three novel software tools designed specifically for recovery of DNA diagnoses (R package QUIDDICH - Kühn & Haase 2020, DeSignate - Hütter et al. 2020 and FastaChar - Merkelbach & Borges 2020) have greatly improved the situation. Nevertheless, there persists an important methodological gap associated with the use of any currently

available software: the reliability of the obtained DNA-based diagnosis is not evaluated. Whichever of these tools one uses, by identifying signature nucleotide characters for members of a taxon in a given data set, one only retrieves a *draft* DNA diagnosis for this taxon. This draft diagnosis may or may not be a valid descriptor of the respective taxon in general, which depends on how accurately the initial data set conveys genetic diversity of both the query taxon and its parent taxon. If failures to accurately delimit taxa or to assemble adequate data sets are common, wide use of DNA characters in taxonomy may lead to an accumulation of inaccurate diagnoses, confusing taxa identification rather than enhancing it. Therefore, filling this methodological gap for the use of DNA data in taxonomy is currently a pressing need.

In the present paper, we analyze to what extent incomplete sampling of species genetic diversity affects the reliability of a resulting DNA diagnosis. Then, we assess whether unsampled genetic diversity can be 'predicted' and accounted for when compiling a DNA diagnosis. We developed a scoring algorithm that is able to overcome the effect of unsampled diversity on the reliability of the DNA diagnosis. We implement this algorithm in a new, powerful, scalable, and versatile software tool, MOLD (MOLecular Diagnosis), which recovers accurate DNA based diagnoses that also meet user-defined criteria of reliability. Below, we review the main types of signature DNA characters, their phylogenetic background, and the existing software tools that identify them. By doing so, we elaborate the conceptual basis on which MOLD capitalizes.

1.1. Single nucleotide Signature DNA Characters in Multispecies Alignments

Following the terminology of Hütter et al (2020), henceforth the taxon under diagnosis is referred to as the *query taxon*, whereas all other taxa in a data set are referred to as *reference taxa*. The most concise and comprehensive classification of single nucleotide characters was proposed by Kühn & Haase (2020). They considered three types of single-site characters that are compared across the data set (types 1-3)~~, as well as a Type 4, which is dedicated to pairwise taxa comparisons (and so not considered further)~~. *Type 1* characters (=pure diagnostic sites - Sarkar et al. 2008) are polymorphic sites in the nucleotide alignment for which all members of a query taxon have a given nucleotide that is not present in any member of the reference taxa (e.g., site 256 for the query *Conasprella* in Figure 1). Types 2 and 3 (sites 283 and 292 in Figure 1 respectively) correspond to sites that are polymorphic within the query taxon. The nucleotides in these sites are either unique to all members of the query taxon (type 2), or to a subset of members (type 3). We here introduce an additional category that we name *Type 5* character~~s~~.

This is a polymorphic site at which all members of a query taxon have the same nucleotide, which is also shared by some, but not all members of the reference taxa (e.g. sites 266 or 286 in Figure 1). Conceptually, Type 5 characters corresponds to “characters” in the population aggregation analysis (Davis & Nixon 1992). Although Type 5 characters considered individually cannot serve as a diagnosis, they can be combined to generate a composite diagnostic character - a combination of character states unique for a query taxon.

The R package SPIDER (Brown et al. 2012) is only capable of identifying Type 1 characters, FastaChar (Merkelbach & Borges 2020) – Types 1 and 2, the R package QUIDDICH (Kühn & Haase 2020) – Types 1 – 3. Only Type 1 and Type 2 characters allow unambiguous differentiation of a query taxon from all other taxa, and the output from QUIDDICH and FastaChar may or may not be sufficient to diagnose a query taxon explicitly, because such diagnostic characters are not necessarily present for each taxon in the analyzed alignment. Therefore, pitfalls are increasingly likely when diagnosing taxa with weak genetic differentiation (e.g. Marchan et al. 2020) or taxa in species-rich lineages that necessitate analysis of large multispecies alignments. DeSignate (Hütter et al. 2020) allows tackling such cases by pairing Type 5 characters into a single composite character, which, unlike Type 1 and 2 characters, can be recovered for taxa even in large DNA data sets.

1.2. Composite Signature DNA Characters in Multispecies Alignments

Despite composite DNA characters being previously implemented in cladistic haplotype analysis (CHA - Brower, 1999) and the characteristic attribute organization system (CAOS - Sarkar et al. 2008), both approaches rely on a tree-based algorithm with its inherent drawbacks (see Kühn & Haase 2020). Here we further develop the concept of composite diagnostic DNA characters by introducing a *minimal diagnostic nucleotide combination* or *mDNC* – a combination of nucleotides at selected sites that are shared by all members of the query taxon and by no member of the reference taxa (e.g. combinations 1-4 in Figure 1). An mDNC may comprise two (i.e., paired sites identified by DeSignate), or any larger number of sites (with a limit of ten sites by default). Furthermore, a Type 1 nucleotide character can be considered an mDNC with a single site (mDNC 1 in Figure 1). Consequently, mDNCs are a generalization of the concept of Type 1 characters. As any mDNC unambiguously defines a query taxon, mDNCs can be thought of as a minimal and sufficient condition to assign a specimen, through its DNA sequence, to a query taxon. As such, it is a proper diagnosis.

Unfortunately, any substitution in one of the mDNC sites, even in a single specimen, disqualifies the entire mDNC, as the remaining sites are not sufficient for proper query identification (Brower, 1999; Lim et al. 2012). An mDNC-based diagnosis can be invalidated either by a low frequency polymorphism in the query species or by a convergent emergence of the same nucleotide combination in any of the reference taxa. Therefore, a more robust diagnosis that could handle these situations is desirable. We thus explore combinations of DNA characters that contain more than the minimal number of nucleotide sites necessary to assign a sequence to a query taxon. We term such combination *redundant DNC*, or *rDNC*. Because rDNCs are longer than mDNCs, the probability of finding the same nucleotide combination among the reference taxa due to convergence is lower. Furthermore, if novel haplotypes that do not share some of the rDNC constituent nucleotides are discovered in a query taxon, a subset of sites from the rDNC that are shared by all query taxon members and are unique to them may still be retained. Therefore, incomplete match of an rDNC is acceptable. In this context, we developed the MOLD algorithm that compiles rDNCs. We applied it to empirical data sets in a series of tests to numerically assess the reliability of the resulting rDNC-based diagnosis compared with DNA diagnoses based on other types of DNA characters.

1.3. Phylogenetic background of signature DNA characters

Some concerns have been raised regarding the use of composite DNA characters as diagnoses because of the complex phylogenetic background of their constituent sites (Jörger & Schrödl, 2013; Merkelbach & Borges, 2020). We briefly address these issues and their relevance to alpha taxonomy, as it is important for setting the conceptual basis upon which we built MOLD. Jörger and Schrödl stated that 'compound characters can be unique for certain species, but they may have evolved from several independent mutation events' implying 'low probabilities of homology' (Jörger & Schrödl 2013: 20). We argue that as long as each constituent site of an mDNC or rDNC is fixed within a query taxon (and only such sites are used by MOLD), their homology in a broader phylogenetic context is not relevant when diagnosing a taxon. Likewise, we disagree with the assertion by Merkelbach & Borges (2020) that plesiomorphic characters should not be included in diagnoses, which would disqualify Type 5 characters from being used as signature characters. First, plesiomorphy or apomorphy cannot be simply deduced from the character type, and requires reconstruction of the character evolution across the data set (Jörger & Schrödl 2013). Second, reporting only apomorphic characters, or reporting only those characters resulting from a single mutation event, is not required by the nomenclatural codes.

The purpose of a diagnosis is to communicate the identity of a taxon, and therefore a diagnosis focuses on character states, but not on their evolutionary history. Most, if not all, traditional diagnoses comprise *informative* morphological characters irrespective of their apomorphic or plesiomorphic nature, or whether they may be homologous or analogous. A diagnosis in general, and *a fortiori* a tree-independent diagnosis, could thus be compared to an identification key – following the key enables allocation of a specimen to a certain taxon, but the consecutive dichotomies of the key are not expected to match events in the evolutionary history of that taxon.

2. Material and Methods

2.1. Overview of the MOLD software

The program MOLD (**MOL**ecular **D**iagnoses) constructs DNA-based diagnoses from an alignment of DNA sequences attributed beforehand to taxa. MOLD can be used to diagnose taxa from genera to species or even subspecies. MOLD is functionally subdivided into two modules (Fig. 2, boxes A and B). The first module identifies multiple mDNCs for the query taxon. Only two types of DNA characters are used to compile mDNCs: either Type 1 characters, each corresponding to a ready mDNC, or Type 5 characters that make up composite mDNCs. Both these character types do not vary across the query specimens, and this is essential for operationability of a diagnosis. The second module transforms the catalog of mDNCs into a set of rDNCs and calculates a score for each of them.

The current version of MOLD available at git-hub (https://github.com/SashaFedosov/MolD) is written in Python 3, but a Python 2.7 version is also available upon request. It does *not* require any dependencies besides standard python libraries. MOLD is also accessible through a graphical web-interface (beta version currently at https://mold.testapi.me/) and as an iTaxoTools module (Vences et al. 2021).

2.2. MOLD algorithm in detail

2.2.1. Building a list of mDNCs

First, all Type 1 and 5 sites are identified for a query taxon. The initial set of Type 5 characters can be filtered from sites where most of the reference taxa members have the same nucleotide as the query taxon. These sites are *a priori* poor candidates to construct short mDNCs, as they will need to be combined with many others to assemble an mDNC. Therefore, each site is assigned a *score* that corresponds to the number of reference taxa members that differ from the query taxa for the nucleotide at this site (Fig. 1, numbers below the alignment). The highest possible score ~~is reached for~~corresponds to Type 1 sites, in which all reference taxa members differ from the query taxon. The scores are then ranked in descending order and the user defines how many of the top-ranking sites are used for assembling a draft combination. The list of mDNCs is initiated as a list of Type 1 sites and then composite mDNCs are appended to it.

The algorithm that builds a composite mDNC from Type 5 sites (Fig. 2, box A) consists of two steps:

In the first step, Type 5 sites are sequentially randomly sampled and assembled into a draft combination. This draft generation process stops either when the combination of nucleotides in these sites is unique for the query taxon, or when the draft combination reaches a maximal length. In the former case the draft combination is directed to the second step; in the latter, it is discarded.

In the second step, the draft combination is refined by removing redundant sites. To do so, each site of the draft combination is discarded in successive~~ly~~ order (i.e. fist site of the draft DNC, then second one, third and so on), while making sure the combination remains diagnostic. When no more sites can be removed without losing the diagnostic property, the remaining combination of sites is an mDNC.

These two steps together constitute a *search iteration.* It is repeated multiple times to generate a collection of unique mDNCs. Users can tune the number of search iterations, the maximal lengths of draft and of final mDNCs, and the number of highest scoring Type 5 sites used for mDNC compilation. Greater maximal lengths for draft and final mDNCs, greater number of Type 5 sites used for draft DNC compilation and higher number of search iterations all lead to a more thorough search for diagnostic combinations, but increase computation time of MOLD. The resulting list of unique mDNCs is comparable to the outputs of existing software tools for identification of diagnostic DNA characters. In the second MOLD module, the list of mDNCs is converted into an rDNC of optimal length.

2.2.2. Compiling an rDNC from the list of mDNCs

The general principle of rDNC construction is illustrated in the Figure 1 (assembly of mDNCs 1-4) and in box B of Figure 2. First, mDNCs output from the first module are sorted by increasing lengths and mDNCs of the same length are 'binned'. In each bin, a given site can be shared by several mDNCs. We thus compute the frequency of occurrence of each site in each bin. Sites with frequency 1 are present in all mDNCs of the bin. Then the sites are ranked inside each bin, so that the top sites have the highest frequency among the shortest mDNCs. If Type 1 characters exist for the query taxa, they are ranked at the top, as they are considered as mDNCs of length 1.

A new rDNC is seeded using one random mDNC among the shortest ones. Then extra sites are picked from the top of the site ranking and are added to the rDNC one-by-one. After each addition of a site, the rDNC is scored for reliability (see below), and the score is recorded. The rDNC extension process stops either when two successive scores exceed the user-defined

reliability threshold (then the best-scoring rDNC is sent to output), or when the rDNC comprises 10 nucleotide sites. In the latter case, the rDNC is output with an alert message if at any step it scored above the reliability threshold. If the scores remain consistently below the reliability threshold, a message is output that no sufficiently reliable rDNC could be compiled.

2.2.3 rDNC scoring

To ~~evaluate~~ test an rDNC after each elongation step, MOLD repeatedly creates simulated *test data sets* that are generated by introducing artificial mutations into the original DNA sequences. This procedure aims to evaluate whether hypothetical larger data sets with sequences that were not sampled in the original data set would still validate a candidate rDNC. It evaluates which are the more relevant rDNCs, despite the limited number of sampled specimens/sequences.

Each artificial sequence is generated by introducing *p* nucleotide substitutions into an existing sequence, where *p* is a random natural number drawn from a uniform distribution [1, *k*L*/100]. In the latter expression k is the natural number corresponding to the desired maximum % of sequence divergence between the original and mutated DNA sequence, and *L* is the sequence length. Mutations are introduced only at polymorphic sites by substituting the original nucleotide by one of the three others, selected randomly with respect to their observed frequencies at this site in the original alignment. Ten artificially mutated sequences are created for each species in the original alignment from randomly sampled original sequences. For species with more than 10 DNA sequences in the original alignment, randomly sampled unchanged sequences are added to the test data set to match the original number of sequences for this species. Thus, a test data sets has at least 10 sequences per species.

For each rDNC evaluation step, MOLD generates 100 test data sets. For each of them, the rDNC under evaluation scores 1 if it unambiguously delimits the query taxon (unique combination defining the query taxon) or 0 otherwise. An rDNC score thus ranges from 0 (it failed in all 100 test data sets) to 100. Importantly, MOLD tolerates one discordant site when evaluating whether the query taxon is correctly diagnosed: if all but one site delineate the query taxon unambiguously, it scores 1. The threshold score after which the rDNC is output was set to 75 in all our analyses.

2.3. Testing MOLD

2.3.1. Testing MOLD on published data sets

In total, nine data sets were used to evaluate MOLD: the Pontohedyle (Mollusca: Gastropoda) *cox1* and Pontohedyle 28S data sets and seven additional published data sets. They correspond to genus-level taxa that proved to be challenging for species delimitation or for taxonomic description. Each of them includes complexes of closely related cryptic or pseudocryptic species with largely overlapping distributions and pronounced genetic structures. Three data sets: Xenuroturris (Mollusca: Gastropoda - Abdelkrim et al. 2018), Daphnia (Crustacea: Cladocera – Crease et al. 2012, plus a subset of sequences from GenBank), and Conus (Mollusca: Gastropoda – combined data of Duda et al. 2012 and Puillandre et al. 2014) comprise standard barcode fragments of the cytochrome oxidase subunit I (*cox1*) (Table 1). The remaining four data sets correspond to *cox1* and three nuclear protein-coding markers, AATS, CAD, and PDGI of the chironomid genus *Tanytarsus* (Insecta: Diptera - Lin et al. 2018).

For the *Pontohedyle* data set, the alignments supplied by the authors (Jörger & Schrödl, 2013) were used as MOLD input. For the other data sets, we generated alignments using MAFFT v.7 (https://mafft.cbrc.jp/alignment/server/) (Katoh et al. 2019) with FFT-NS-2 strategy. The alignments were then translated using MEGA v.6 (Tamura et al. 2014) to ensure consistency of the amino-acid (AA) sequences and lack of premature stop-codons. We ran RAxML v.8.2.12 (Stamatakis et al. 2006) on the Cypress Gateway (Miller et al. 2010), with three codon positions allocated to separate partitions, to check that sequences assigned to the same species formed a clade.

First, to evaluate the general performance of MOLD, we diagnosed all species from each analyzed data set, with 10,000 search iterations and 100 Type 5 sites considered for inclusion into mDNCs. When compiling rDNCs, mutated sequences in test data sets were a maximum of 1% different from the original sequences from which they were derived, which is within the typical ~~K2P~~ genetic distance for intra-specific comparisons of all analyzed taxa and genes (Abdelkrim et al. 2018; Puillandre et al. 2014; Lin et al. 2018; Hebert et al. 2003).

2.3.2. Comparison of MOLD with other tools

We compared the mDNCs identified by MOLD with the mDNCs recovered by the available programs for signature DNA character identification. The two Pontohedyle data sets (Jörger & Schrödl 2013), *cox1* and 28S rRNA, were selected for comparisons among QUIDDICH, DeSignate, FastaChar and MOLD since they were also used to test the three former tools. Here we only compared the Type 1 characters output by all currently available software, because this is the only character type identified by all these tools.

To compare 2-site mDNC outputs from MOLD and DeSignate, we also used the notably larger *Conus* data set (187 species, 984 unique sequences – Table 1). When running DeSignate, the size of the k-window was set to the alignment length, to obtain results comparable with those from MOLD.

When comparing MOLD with other relevant tools designed for the same purpose, we only focused on the consistency of output and, to a lesser extent, on the runtime performance. The additional features are reviewed in sufficient detail by Hütter et al. (2020) and not considered herein. We note that MOLD is the only existing tree-independent software tool capable of i) identifying mDNCs of three or more sites, and ii) assembling DNA diagnoses that fulfil user-defined criteria of reliability, rDNCs – these features could not therefore be assessed in comparison with other tools.

2.3.3. Assessment of the effect of sampling on the robustness of the DNA based diagnosis

We assessed the robustness of mDNC- and rDNC- based diagnoses by performing random haplotype subsampling (jackknifing) on six published data sets (Table 1). In each data set, subsampling was performed for two to four query taxa that contrast in their genetic diversity and phylogenetic distinctiveness (Table 2). The genetic diversity was estimated from the number of unique haplotypes per species, and phylogenetic distinctiveness was based on the length of the branch to the corresponding species in the phylogenetic tree.

In each subsampling run, we sampled an increasing fraction of genetic diversity for both the query and the reference taxa ~~(Fig. 3)~~. Two sampling regimes were performed: with constant species composition (h-sampling, performed for both mDNCs and rDNCs), and with varying species composition (hspp-sampling, only performed for rDNCs).

2.3.3.1. h-sampling

At each h-sampling iteration, a *partial data set* (as opposed to an *entire* data set) was initiated by randomly selecting *n* unique sequences of the query species, and complemented by unique sequences of each of the reference species. The number of unique sequences, *n,* sampled for the query species ranged from 2 to the total number of unique sequences available for this species. The number of sequences sampled per reference species was proportional to the representation of this species in the entire data set, but no less than one ~~(Fig. 2b, box C)~~. Ten independent subsampling replicates were performed for each *n*.

Each partial data set was analyzed by MOLD with 20,000 search iterations. The resulting mDNCs were tested for their ability to be used as proper diagnoses of the respective queries in the respective entire data set. The mDNCs recovered from a partial data set that retained their diagnostic property (i.e. remained a shared feature of all members of the query taxa and unique to them) in the respective entire data set were recorded. The proportion of such mDNCs in the output generated for each partial data set characterizes the reliability of the recovered mDNCs catalog, and is referred to as the *quantified reliability*. In addition to the reliability of the complete mDNC catalog, the *quantified reliability* was also recorded separately for each length of mDNCs to assess whether short or long mDNCs are generally more reliable. The proportion of polymorphic sites in the query sequences was also recorded as a measure of the genetic heterogeneity of each partial data set. Similarly, h-sampling was subsequently used to evaluate the robustness of the rDNCs.

2.3.3.2. hspp-sampling

Because h-sampling aimed at testing the effect of sampled genetic diversity per species on the reliability of the output diagnosis, the species composition is identical under this regime in all partial data sets. However, the taxonomic coverage of a partial data set (i.e. its completeness in terms of species) is expected to greatly affect the reliability of the diagnoses as well: the more species in the partial data set, the more reliable the diagnoses. To estimate the contribution of data set species composition on diagnosis reliability, we performed hspp-sampling (for rDNCs only). In hssp-sampling, the number of sequences per species is strictly proportional to its original abundance in the entire data set, meaning that a species may be not represented in a partial data set. In practice, all species with one or few sequences in the original data set are absent from the hssp-sampling partial data sets corresponding to small subsamples of query species. However, partial data sets always included 1-3 'indispensable' species that are the closest relatives of the query species in the reconstructed phylogenies; these are represented by a minimum of 1 sequence. In all other aspects, the h- sampling and hssp-sampling regimes are identical.

3. Results

3.1. DNA diagnoses recovered by MOLD

MOLD identified multiple mDNCs for all species in each of the nine analyzed data sets (Supplementary data 1). With the exception of *Tanytarsus brundini* (Tanytarsus *cox1* data set), at least 158 mDNCs were recovered for each of the diagnosed species in each data set. The smallest average number of mDNCs per species was in the Xenuroturris data set (1,012) and the largest was in the Conus data set (7,050 – Table S1).

Type 1 sites were detected ~~in all~~for each species in the smallest Pontohedyle *cox1* and Pontohedyle 28S data sets (entirely red top-left charts on the figure ~~(Fig. 4~~3). However, in larger data sets some species lacked Type 1 sites, and could only be diagnosed by 2-site or even 3-site mDNCs ~~(Figs 4, 5)~~. The proportion of species lacking Type 1 sites is highest in those data sets that include both a larger number of species and a larger number of unique haplotypes per species. Only 19 species (=22%) in the Daphnia *cox1* data set and 22 species (=12%) in the Conus *cox1* data set could be diagnosed by Type 1 sites (Figs ~~4, 5.~~ 3).

Each of the four analyzed *Tanytarsus* data sets contains a sufficient number of polymorphic sites to diagnose all the included species ~~(Fig. 5)~~. The proportion of species that possess at least one Type 1 site ranges from 25 % (AATS) to 61 % (CAD), and 84 species out of the total 105 analyzed have at least one Type 1 site in at least one marker (locus).

rDNCs that fulfill the pre-defined criteria of reliability with standard MOLD settings were successfully compiled for all but two species (~~Figs 4, 5,~~ Table S2, Supplementary data 1). The exceptions are *Tanytarsus brundini* in the Tanytarsus ~~CAD~~ *cox1* data set, and *Iotyrris conotaxis* in the Xenuroturris data set. *Tanytarsus brundini* is represented by two divergent mitochondrial lineages in the Tanytarsus *cox1* data set, which do not form a clade. *Iotyrris conotaxis* shows high *cox1* haplotype diversity and is weakly differentiated from its sister species, *I. musivum*, so these two species were reliably delimited only based on RAD-Seq data (Abdelkrim et al. 2018).

In the majority of species in the analyzed data sets, rDNCs are comprised of no more than four sites (Fig. ~~s 4, 5~~3), and shorter mDNCs generally translate into shorter rDNCs. When we diagnosed all species of a data set, MOLD runtime ranged from 92sec to 24h23min in Pontohedyle *cox1* and Conus data sets respectively.

MOLD uses random selection of alignment sites first at the step of mDNC recovery and later when building data sets of simulated sequences to score rDNCs. This could theoretically have a strong effect on the reproducibility of the resulting diagnostic combination. To evaluate

consistency of the output rDNC from run to run, we performed rDNC recovery in 10 replicates for 16 query taxa in seven data sets (totaling 20 series). The same rDNC was recovered in all 10 runs in 12 out of the 20 series (Table S3, Supplementary data 2). The rDNCs from different runs varied in length by one site, but otherwise were identical in six series, and only in one series, *Daphnia pulex* (Daphnia), did the rDNC length vary by three sites. Only in one series, *Iotyrris cingulifera* (Xenuroturris), did different runs employ alternative subsets of nucleotide sites in the production of rDNCs. In no instance was an rDNC identified successfully in some runs but not in others. Finally, to evaluate consistency of scoring from run to run, we performed rDNC recovery in 30 replicates for the query taxa *Pontohedyle brasilensis* (Pontohedyle 28S), *Xenuroturris legitima*, *Iotyrris cingulifera* (both Xenuroturris) and *Conus ebraeus* (Conus). Despite an often notable difference between the minimum and maximum scores (Fig. 4~~6~~), this difference is reduces with longer rDNCs to consistently fall above the selected reliability threshold of 0.75 (grey zone).

3.2. Performance of MOLD in comparison with previously available tools

The MOLD output of Type 1 characters in the Pontohedyle data sets (Table S4A, Supplementary data 3) was identical to that of other tools: *nucDiag* function of Spider, CAOS, QUIDDICH, DeSignate and FastaChar. A comprehensive search on the Pontohedyle *cox1* data set (50,000 search iterations across all informative alignment sites) produced 1,508 to 6,578 2-site mDNCs per species (i.e. 92-100% of the mDNCs returned by DeSignate - Table S4B, Supplementary data 3). When the Pontohedyle 28S data set with fewer informative sites was analyzed, the outputs from MOLD and DeSignate were identical for all species with only 10,000 MOLD search iterations.

An attempt to retrieve a species diagnosis in the larger Conus data set caused a gateway timeout error in the web server-based implementation of DeSignate. We therefore ran DeSignate via the Django server to identify 2-site mDNCs for four *Conus* species (Table S3). The same number of mDNCs comprising two nucleotide sites was obtained for each species using DeSignate and only 10,000 search iterations of MOLD. MOLD runtime increases almost linearly, from 2.86 seconds (2,000 search iterations) to 69.35 seconds (50,000 search iterations) for the query *Conus ebraeus* (Table S4). When we ran MOLD with 10,000 search iterations on four query species of *Conus*, the runtime varied from 11.61 seconds to 15.55 seconds, and was 5.5 to 9 times shorter than the DeSignate runtime for the same query.

439

440

441 *3.3. Reliability of the mDNC-based diagnoses*

442 We performed iterated haplotype subsampling with an increasing fraction of the data set's
443 genetic diversity sampled, to evaluate the reliability of the mDNC-based diagnoses associated
444 with each sample size. Our rationale was that if we access a sufficiently large DNA sequence
445 data set for a given taxon, this data set may be used to model finite genetic diversity of this
446 taxon. Then, by compiling diagnoses from sub-samples of this large data set, and checking if
447 they remain valid diagnoses of the query taxon in the context of the full data set (i.e. shared by
448 all query taxon members, and unique to them), we can quantify the reliability of the diagnosis
449 associated with each sub-sample. We expect small sub-samples of the genetic diversity to
450 produce low reliability diagnoses. With increasing sub-sample size, the diagnosis reliability will
451 increase until it finally reaches 100%, when all available records are included. The curve
452 describing the growing robustness of a diagnosis as a function of the fraction of diversity
453 sampled may reach a plateau earlier; in this case, the sampling fraction at which the plateau is
454 reached marks the minimum taxonomic sampling sufficient to provide a robust DNA diagnosis.
455 We also expect shorter mDNCs to be more robust, as the more sites are included in an mDNC,
456 the more probable it is that a yet undetected polymorphism exists in the query taxon at least at
457 one of these sites.

458 With an increasing number of sampled query species haplotypes, diagnosis reliability
459 grows almost linearly in nine of the ten analyzed query species (Figs ~~7~~ 5 a, b, c; Supplementary
460 data 4). In none of the ten query species does the curve come to a plateau until all or almost all
461 haplotypes are added to the partial data sets. Among the mDNCs identified for the smallest
462 partial data sets comprising only two query haplotypes, the quantified reliability ranged from 0
463 to 52%. This starting reliability value is higher in the phylogenetically more distinctive species:
464 *Xenuroturris legitima* (Fig. ~~7a~~5 a), *Daphnia longispina* and *D. laevis* (Fig. ~~7b~~5 b), *Conus*
465 *sanguinolentus* (Fig. ~~7c~~5 c) and/or in species represented in the entire data set by fewer unique
466 haplotypes: *Daphnia melanica* and *Daphnia longispina* (Fig. ~~7b~~5 b). An arbitrarily selected
467 reliability threshold of 0.75 (i.e. 3/4 recovered mDNCs are valid for the entire data set) is
468 reached when no less than 50% of query haplotypes are included in partial data sets.

469 When we performed subsampling of *Tanytarsus thomasi* and *T. tongmuensis* with nuclear
470 loci, a plateau was reached for each species, only after more than 75% of available haplotypes
471 were sampled (Figs ~~7~~ 5 d – f). It is noteworthy that, because each reference species was

represented by no more than 1-3 haplotypes, partial data sets were virtually identical with the final data set in representation of reference species genetic diversity after 2/3 of the query haplotypes were sampled. This introduced a bias compared to other data sets, which likely contributed to the observed faster increase of mDNC robustness.

In Figure ~~8~~6, each partial data set is represented on a respective scatterplot by three data points, which show mDNC reliability as a function of the query taxon sampled diversity, separately for 1-site mDNCs (blue), 2-site mDNCs (green) and 3- site mDNCs (orange). The areas occupied by green records are higher than those occupied by orange ones, indicating higher reliability of the 2-site mDNCs over 3- site mDNCs. In *Xenuroturris legitima* (Fig. ~~8a~~6a), which can be diagnosed by 25 1-site mDNCs in the entire Xenuroturris data set, all but four blue records are above the threshold of 0.75, suggesting that 1-site mDNCs allow for a very reliable diagnosis in this species. Therefore, shorter mDNCs are indeed generally more reliable than longer ones, but overall mDNC based diagnoses are weak, unless based on a thorough sampling of both the query and reference taxa.

3.4. Reliability of the rDNC-based diagnoses~~rDNC reliability analysis~~

When haplotype subsampling was performed to assess the rDNC reliability dynamics, the obtained graphs were notably different from those for mDNCs (Fig. ~~9a~~ 7 a – c, Supplementary data 5). In all analyzed query taxa, with the exception of *Daphnia pulex* (Fig. ~~9b~~7 b), the arbitrary threshold of 0.75 was reached earlier, compared to mDNC subsampling (marked with arrows for respective taxa), and a plateau was reached soon after. Failure to recover a reliable diagnosis for the assemblage of sequences here attributed to *D. pulex* based on smaller sampled diversity is most probably the result of complicated taxa delimitation. Even if *D. pulex* constitutes a monophyletic group, the maximum intraspecific K2P genetic distance for *D. pulex* as defined herein (0.039) exceeds more than two-fold the minimum genetic distances between *D. pulex* and *D. middendorfiana* (0.014) and between *D. pulex* and *D. melanica* (0.016).

The quick increase of rDNC reliability with growing sample of genetic diversity implies generally much higher robustness of rDNCs, compared to mDNCs. However, we suspected that the obtained results might be too optimistic. We designed our h-sampling regime to test the effect of sampled genetic diversity per species, where the number of haplotypes in each species was changing but the number of species was not. As all species of the final data set were represented in each partial data set, each partial data set might already reasonably well capture the genetic landscape of the entire data set. However, such an approach to subsampling might

not correctly reflect the selection of reference taxa in a real taxonomic study (for which MOLD is designed to be useful). Therefore, we performed hspp-sampling by assembling partial data sets that, in addition to varying in haplotype per species composition, also contained different subsets of reference species (Supplementary data 6). The curves describing changes in the proportion of rDNCs valid for the entire data set were similar to those obtained with h-sampling for most query species (Figs ~~9d~~ 7 d – f). Only in *Daphnia laevis* did the dynamics of the rDNC robustness differ notably depending on the sampling regime. Finally, we performed hspp-sampling for *Tanytarsus thomasi* and *T. tongmuensis* in three nuclear gene data sets, AATS, CAD, and PGDI. The results were consistent with those obtained with three *cox1* data sets: the curve of rDNC reliability reached the 0.75 value after 3 (out of 7) haplotypes were sampled in each data set for *T. thomasi*, and 3 to 5 haplotypes (out of 17) were sampled for *T. tongmuensis* (Figs ~~9g~~ 7 g – i). Therefore, rDNCs appear to constitute notably more reliable DNA diagnoses and can be efficiently compiled for both mitochondrial and nuclear loci.

4. Discussion

4.1. MOLD and other tools for diagnostic DNA character identification

We demonstrate that MOLD efficiently retrieves diagnostic combinations of nucleotides (mDNCs and rDNCs) for pre-defined groups of DNA sequences in data sets of varying complexity. All analyses performed for the present study (including computationally extensive iterative subsampling tasks) were run on a single CPU of a standard laptop. Therefore, even the unparallelized source code runs on virtually any reasonably performing computer with Python installed. A parallelized version of MOLD (currently being tested) should allow for improved performance.

We demonstrate that MOLD is capable of retrieving reliable DNA diagnoses for taxa of varying genetic diversity and distinctiveness. The user-defined parameters allow one to adjust depth and breadth of searches to match requirements posed by different data sets and taxa. Whereas most of our analyses were performed on nucleotide coding genes, non-coding DNA fragments can be analyzed equally well by coding alignment gaps as a fifth character. Finally, SNP data can be used more broadly for the identification of DNA based diagnoses in the future. In this case, a simple module would be required to extract filtered SNPs from variant call files and concatenate them into SNP haplotypes (Fourie et al. 2015; Marchán et al. 2020), which can then be used by MOLD.

We show that the shorter the mDNC, the more reliable it is. Also, MOLD core functions are designed in such a way that the shorter the mDNC, the higher the probability that it will be identified. All 1-site mDNCs and 97-100% of the 2-site mDNCs recovered by DeSignate are also identified by MOLD (Supp. data 6). However, finding all several thousand diagnostic combinations is not necessary for providing a reliable diagnosis from a taxonomic perspective. When such large numbers of equally powerful characters exists (each potentially sufficient to diagnose the query taxon), omitting a small percent of them does not affect the robustness of the resulting diagnosis. We therefore demonstrate that MOLD is a scalable and versatile program that returns reliable and reproducible results.

All currently existing tools that can be used for the identification of taxon signature characters in DNA alignments are capable of retrieving 1-site mDNCs (= Type 1 characters). This is a simple computational task (of complexity $n*L$, where n is the sample size and L the alignment length). It requires minimal CPU resources and 1-site mDNCs provide as robust a DNA diagnosis as is possible with mDNCs. DeSignate and MOLD are the two tree-independent

tools that are capable of identifying composite mDNCs, and despite being based on different approaches, they return very similar sets of 2-site mDNCs. In brief, DeSignate is faster and more efficient for simple data sets (such as the Pontohedyle *cox1*), whereas MOLD is faster and more powerful for medium or high complexity data sets (such as Daphnia and Conus). There are no alternatives to MOLD if identification of 3-site or longer DNCs is needed.

The main strength of MOLD is that it is the only available tool capable of compiling rDNC based diagnoses. The results of haplotype subsampling demonstrate that only 1-site mDNCs are sufficiently reliable to be useful for diagnosing taxa, and only when based on adequate sampling for both the query and the reference taxa. But as we demonstrate, the likelihood is low that even one such site per species exists in a monolocus data set comprising hundreds of species. Consequently, one should opt for either more taxonomically restricted data sets, longer or multiple DNA markers, or the use of composite characters. The latter approach is inevitable in highly diversified, poorly studied or taxonomically problematic groups, where available genetic resources are scarce, and for which defining a restricted scope of analysis may be difficult. In such data sets rDNCs offer a workable solution. In summary, all existing tools may potentially be used if only Type 1 characters are accepted as signature characters. When no Type 1 characters exist for a query taxon, rDNCs constitute a more reliable diagnosis than composite mDNCs, and in such cases MOLD is likely the best choice.

4.2. Taxonomic sampling and robustness of DNA-based diagnoses

The major impediment to the proposition of molecular diagnoses on a regular basis is the supposed lack of robustness, because of their inherent sampling-dependent nature. In this context, MOLD leverages to some extent unsampled genetic diversity, but properly designed sampling remains crucial for identification of a robust diagnosis. Tripp & Lendemer (2014) suggested that preferably 10 vouchers of any new taxon should be sequenced along with at least 15 of its closest relatives. These numbers, nevertheless, appear too generalized, because genetic diversity varies strongly from species to species, as does the distribution of this diversity across species distribution ranges (Pante et al. 2015b). Theoretically, in order to claim that a diagnostic character of a given taxon is truly fixed, every single individual of this taxon needs to be examined (Wiens & Servedio 2000), which will never be feasible. Furthermore, the number of new mutations per site per generation, N*mu, notably exceeds 1 for many species (Drake et al. 1998). However, only a tiny subset of possible polymorphisms reaches an appreciable

frequency among adults of the population. These are the sites that should be present in the sample of each species in a data set to ensure recovery of reliable diagnoses.

Rare species – those represented by a single specimen, or by few specimens acquired at one sampling event, pose a challenge to taxonomy. Rarity of a species may reflect its low population size or may result from inadequate sampling. The former scenario likely translates into reduced genetic diversity, which does not preclude usage of a single record for diagnosis recovery, but the latter implies greater unsampled diversity, thus impacting the robustness of the DNA diagnoses. Assessing the magnitude of unsampled diversity requires taxon-specific expertise, and so it will fall upon a taxonomist to decide whether to diagnose rare species or not. There is already a bulk of literature available that addresses sampling design for the purpose of species delimitation (e.g. Knowlton 2000; Eckert et al. 2008; Lim et al. 2011, and references therein), so we do not cover it in further detail. Nevertheless, we note that in MOLD we have made a first attempt to model unsampled diversity. Under default parameters, a single record of a species generates up to 1,000 unique simulated haplotypes, some of which by chance will match existing polymorphisms that are lacking from the empirical data.

We demonstrate that the reliability of DNA-based diagnoses can be estimated using a simple informatics toolkit – this is mainly due to the formal and universal language of DNA. Traditional morphological diagnoses, even theoretically, cannot be challenged in a similar manner because these are mainly based on taxon-specific features that are difficult to formalize (Lim et al. 2012) and are commonly subject to researcher bias (Fujita et al. 2012). From this perspective, and for a given sampling effort, DNA-based diagnoses compiled following a standardized protocol in a thoughtfully designed data set should be a more reliable descriptor of the identity of a taxon compared to traditional morphological diagnoses. Revision of a morphological diagnosis is common practice when novel data become available. Similarly, a DNA diagnosis will remain a reflection of the state-of-the-art in understanding the molecular identity of a taxon.

4.3. Which sources of genetic data can be used

Selection of DNA markers to be used for compilation of DNA diagnosis is an important task which has a strong impact on the credibility and usability of a resulting DNA diagnosis. Ideally it should enable their matching and verification in further analyses. We identify three main criteria that must be fulfilled by a candidate marker. First, it must be informative, i.e., comprise a sufficient number of variable sites to discriminate among taxa in a data set; second, it must

allow for high confidence reproducible alignment across data sets to ensure confident nucleotide homology hypotheses; third, it must generate a gene tree that is generally congruent with the species tree of the analyzed data set. Finding an ideal marker that satisfies these criteria may be difficult because different criteria imply contrasting patterns of molecular evolution. For example, informative non-coding markers often cannot be aligned confidently. For instance, internal transcribed spacers (ITS) widely used as barcode markers, especially in fungi, tend to produce alignments with multiple single-nucleotide columns flanked by gappy regions (e.g. Stielow et al. 2011; Garnica et al. 2016). Barcode matching in such cases relies on sequence similarity and does not require fixed homology hypotheses across the data set. The solution commonly used in phylogenetics - discarding poorly aligned columns from the alignment - is unacceptable for position-based diagnosis recovery because it disrupts base indexing. Furthermore, high rates of molecular evolution implied in informative markers may also result in artifacts in the gene tree topology due to LBA, high rates of homoplasy, paralogy, or marker specific biases (such as mitochondrial introgression).

There are some additional criteria that should be taken into consideration, in particular the ease and reproducibility of the laboratory protocols and the availability of comparative data. Most data sets analyzed in the present study comprise the widely used barcode marker *cox1*. The pros of using this fragment are well known: it is informative, it can be confidently aligned even among divergent taxa, degrees of its variation within and among taxa are well documented, and there is a wealth of data available for this fragment in the NCBI and BOLD databases. The cons, although mostly lineage-specific, are the low resolution in some basal metazoan lineages, such as sponges and corals (Huang et al. 2008; Vargas et al. 2012), taxa-specific mitochondrial introgression (Toews & Brelsford, 2012), or pseudogenization (Song et al. 2008). Therefore, examination of the gene tree is mandatory prior to any attempts to propose DNA based diagnosis even for such a broadly used marker as *cox1*.

Data ~~availability~~accessibility

The data that support the findings of this study (DNA alignments used in the present study, unedited output files, as well as the python scripts used to generate them, and scripts used to extract results from the output files and plot them) are ~~openly~~ available at https://github.com/SashaFedosov/Fedosov_et_al_MOLD_sripts_and_data. ~~The supplementary data include DNA alignments used in the present study, unedited output files, as well as the~~

~~python scripts used to generate them, and scripts used to extract results from the output files and plot them.~~

Acknowledgments

We are grateful to ~~two~~ three anonymous reviewers for their comments on the manuscript, and to Claudia Ratti (MNHN) for checking manuscript style. The present study was supported by the Russian Science Foundation~~,~~ (Grant #19-74-10020 to AF).

653 **References**

654 Abdelkrim J., Aznar-Cormano L., Buge B., Fedosov A., Kantor Y., Zaharias P., Puillandre N. 2018.
655 Delimiting species of marine gastropods (Turridae, Conoidea) using RAD-sequencing in an
656 integrative taxonomy framework. *Molecular Ecology*, 27:4591–4611.
657 Brower A.V.Z. 1999. Delimitation of phylogenetic species with DNA sequences: a critique of
658 Davis and Nixon's population aggregation analysis. *Systematic Biology*, 48: 199–213.
659 Brown S.D.J., Collins R.A., Boyer S. 2012. Spider: an R package for the analysis of species identity
660 and evolution, with particular reference to DNA barcoding. *Molecular Ecology Resources,*
661 12, 562–565.
662 Cook L.G., Edwards R.D., Crisp M.D., Hardy N.B. 2010. Need morphology always be required for
663 new species descriptions? *Invertebrate Systematics*, 24: 322–326.
664 Crease T.J., Omilian A.R., Costanzo K.S., Taylor D.J. 2012. Transcontinental Phylogeography of
665 the *Daphnia pulex* Species Complex. *PLOS ONE* 7(10): e46620.
666 Davis J.I., Nixon K.C. 1992. Populations, genetic variation, and the delimitation of phylogenetic
667 species. *Systematic Biology,* 41: 421–35.
668 Drake J.W., Charlesworth B., Charlesworth D., Crow J.F. 1998. Rates of spontaneous mutations.
669 *Genetics*, 148(4), 1667-1686.
670 Duda T.F.Jr., Terbio M., Chen G., Phillips S., Olenzek A.M., Chang D., Morris D.W. 2012. Patterns
671 of population structure and historical demography of Conus species in the tropical Pacific.
672 *American. Malacological Bulletin*, 30:175-187.
673 Dunn C.P. 2003. Keeping taxonomy based in morphology. *Trends in Ecology and Evolution*,
674 18(6), 270-271.
675 Eckert C.G., Samis K.E., Lougheed S.C. 2008. Genetic variation across species' geographical
676 ranges: the central–marginal hypothesis and beyond. *Molecular Ecology*, 17, 1170–1188.
677 Fujita M.K., Leaché A.D., Burbrink F.T., McGuire J.A., Moritz C. 2012. Coalescent-based species
678 delimitation in an integrative taxonomy. *Trends in Ecology and Evolution,* 27(9), 480-488.
679 Funk D.J., Omland K.E. 2003. Species-level paraphyly and polyphyly: frequency, causes, and
680 consequences, with insights from animal mitochondrial DNA. *Annual Review of Ecology*
681 *Evolution and Systematics* 34: 397‑423.
682 Garnica S., Schön M.E., Abarenkov K., Riess K., Liimatainen K., Niskanen T., Dima B., Soop K.,
683 Frøslev T.G., Jeppesen T.S., Peintner U., Kuhnert-Finkernagel R., Brandrud T.E., Saar G.,
684 Oertel B., Ammirati J.F. 2016. Determining threshold values for barcoding fungi: lessons

from *Cortinarius* (Basidiomycota), a highly diverse and widespread ectomycorrhizal genus. *FEMS Microbiol Ecol*. 92(4): fiw045.

Goldstein P., DeSalle R. 2011. Integrating DNA barcode data and taxonomic practice: Determination, discovery, and description. *Bioessays*, 33: 135-147.

Hebert P.D.N., Ratnasingham S., DeWaard J.R. 2003. Barcoding animal life: Cytochrome c oxidase subunit 1 divergences among closely related species. *Proc R Soc Lond B Biol Sci*, 270, S596–S599.

Huang D., Meier R., Todd P.A., Chou L.M. 2008. Slow mitochondrial COI sequence evolution at the base of the metazoan tree and its implications for DNA barcoding. *Journal of Molecular Evolution* 66(2): 167-74.

Hütter T., Ganser M.H., Kocher M., Halkic M., Agatha S., Augsten N. 2020. DeSignate: detecting signature characters in gene sequence alignments for taxon diagnoses. *BMC Bioinformatics*, 21, 151.

ICZN. 1999. International Code of Zoological Nomenclature. 4th ed. London, UK: The International Trust for Zoological Nomenclature. 306 pp. Available from: https://www.iczn.org/the-code/the-international-code-of-zoological-nomenclature/the-code-online/ (accessed 21 September 2020).

Janzen D.H., Hallwachs W., Blandin P., Burns J.M., et al. 2009. Integration of DNA barcoding into an ongoing inventory of complex tropical biodiversity. *Molecular Ecology Resources*, 9, 1–26.

Jörger K.M., Schrödl M. 2013. How to describe a cryptic species? Practical challenges of molecular taxonomy. *Frontiers in Zoology*, 10, 1–27.

Jörger K.M., Norenburg J.L., Wilson N.G., Schrödl M. 2012. Barcoding against a paradox? Combined molecular species delineations reveal multiple cryptic lineages in elusive meiofaunal sea slugs. *BMC Evolutionary Biology*, 12, 245.

Katoh K., Rozewicki J., Yamada K.D. 2019. MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. *Briefings in bioinformatics*, 20 (4), 1160–1166.

Knowlton N. 2000. Molecular genetic analyses of species boundaries in the sea. *Hydrobiologia*, 420, 73–90.

Kühn A.L., Haase M. 2020. QUIDDICH: QUick IDentification of DIagnostic CHaracters. *Journal of Zoological Systematics and Evolutionary Research*, 58: 22–26.

717 Lim G.S., Balke M., Meier R. 2011. Determining species boundaries in a world full of rarity:
718 singletons, species delimitation methods. *Systematic Biology*, 61, 165-169.
719 Lin X.-L., Stur E., Ekrem T. 2018. Exploring species boundaries with multiple genetic loci using
720 empirical data from non-biting midges. *Zoologica Scripta*, 47, 325– 341.
721 Marchán D.F., Fernández R., Domínguez J., Cosín D.J.D., Novo M. 2020. Genome-informed
722 integrative taxonomic description of three cryptic species in the earthworm genus
723 *Carpetania* (Oligochaeta, Hormogastridae). *Systematics and Biodiversity*, 18(3), 203-215.
724 Meier R., Blaimer B., Buenaventura E., Hartop E., von Rintelen T., Srivathsan A., Yeo D. 2021. A
725 re-analysis of the data in Sharkey et al.'s (2021) minimalist revision reveals that BINs do
726 not deserve names, but BOLD Systems needs a stronger commitment to open science.
727 *bioRxiv* doi:10.1101/2021.04.28.441626.
728 Merckelbach L.M., Borges L.M.S. 2020. Make every species count: FastaChar software for rapid
729 determination of molecular diagnostic characters to describe species. *Molecular Ecology*
730 *Resources,* 20: 1761–1768.
731 Miller M. A., Pfeiffer W., Schwartz T. 2010. Creating the CIPRES Science Gateway for inference
732 of large phylogenetic trees". In: Gateway Computing Environments Workshop (GCE), New
733 Orleans, pp. 1-8.
734 Pante E., Abdelkrim J., Viricel A., Gey D., France S., Boisselier M.-C., Samadi S. 2015a. Use of
735 RAD sequencing for delimiting species. *Heredity*, 114(5), 450–459.
736 Pante E., Puillandre N., Viricel A., Arnaud-Haond S., Aurelle D., Castelin M., Chenuil A.,
737 Destombe C., Forcioli D., Valero M., Viard F., Samadi S. 2015b. Species are hypotheses:
738 avoid connectivity assessments based on pillars of sand. *Molecular Ecology*, 24: 525-544.
739 Puillandre P., Bouchet P., Duda T. F., Kauferstein S., Kohn A. J., Olivera B. M., et al. 2014.
740 Molecular phylogeny and evolution of the cone snails (Gastropoda, Conoidea). *Molecular*
741 *Phylogenetics and Evolution*, 78, 290-303.
742 Puillandre P., Lambert A., Brouillet S., Achaz G. 2012. ABGD, Automatic Barcode Gap Discovery
743 for primary species delimitation. *Molecular Ecology* 21, 1864–1877.
744 Renner S.S. 2016. A return to Linnaeus's focus on diagnosis, not description: the use of DNA
745 characters in the formal naming of species. *Systematic Biology*, 65(6), 1085-1095.
746 Sarkar I.N., Planet P.J., DeSalle R. 2008. CAOS software for use in character-based DNA
747 barcoding. *Molecular Ecology Resources*. 8, 1256-1259.
748 Sharkey M.J., Janzen D.H., Hallwachs W., Chapman E.G., Smith M.A., Dapkey T., Brown A.,
749 Ratnasingham S, Naik S, Manjunath R, et al. 2021. Minimalist revision and description of

403 new species in 11 subfamilies of Costa Rican braconid parasitoid wasps, including host records for 219 species. *ZooKeys* 4.

Song H., Buhay J. E., Whiting M.F., Crandall K. A. 2008. Many species in one: DNA barcoding overestimates the number of species when nuclear mitochondrial pseudogenes are coamplified. *Proceedings of the National Academy of Sciences*, 105 (36), 13486-13491.

Stamatakis A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, 22, 2688-2690.

Stielow B., Bratek Z., Orczan A.K.I, Rudnoy S., Hensel G. 2011. Species Delimitation in Taxonomically Difficult Fungi: The Case of *Hymenogaster*. *PLoS ONE* 6(1), e15614.

Toews D.P.L., Brelsford A. 2012. The biogeography of mitochondrial and nuclear discordance in animals. *Molecular Ecology*, 21, 3907–3930.

Tripp E.A., Lendemer J.C. 2014. Sleepless nights: When you can't find anything to use but molecules to describe new taxa. *Taxon*, 63, 969–971.

Vargas S., Schuste A., Sache K., Büttner G., Schätzl S., Läuchli B., Hall K., Hooper J.N., Erpenbeck D., Wörheide G. 2012. Barcoding sponges: an overview based on comprehensive sampling. *PloS ONE*, 7(7), e39345.

Vences M., Miralles A., Brouillet B., Ducasse J., Fedosov A.E., Kharchev V., Kostadinov I., Kumari S., Patmanidis S., Scherz M.D., Puillandre N., Renner S.S. 2021. iTaxoTools 0.1: Kickstarting a specimen-based software toolkit for taxonomists. *Megataxa*, 6(2): 77-92.

Wiens J.J., Servedio, M.R. 2000. Species delimitation in systematics: inferring diagnostic differences between species. *Proceedings of the Royal Society series B*, 267, 631–636.

Captions

Figure 1. Major types of DNA characters in the alignment of *cox1* of the family Conidae; query taxon genus *Conasprella*. Invariable ~~nucleotides~~ sites are represented by dots. The character type (1, 2, 3 and 5) is indicated for informative characters; ~~v characters are marked with the respective digit above the alignment. V~~ariable sites not representing any type are marked by 'x'. All sites not used by MOLD are shaded. For Type 5 characters, sites in the reference taxa with a different nucleotide from that in the query taxon are marked in grey. Their count corresponds to the cut-off value reported below the alignment. Example mDNCs (1 – 3), and rDNC (4) are shown above the alignment; the numbers of constituent sites correspond to their position in the alignment.

Figure 2. Workflow of standard MOLD distribution. Box A represents the module for mDNC recovery. Box B represents the module ~~to~~ that transforms the catalog of mDNCs into a set of rDNCs and ~~to select~~outputs the rDNC with the highest score. Grey rectangles show analysis tasks, grey hexagons – intermediate outputs used by the program, yellow ellipses – final output available to users.~~The Legend is provided below the solid black line.~~

Figure 3. Results of MOLD application to the empirical datasets, when all species of a dataset were diagnosed. Each empirical data set is represented by a pair of charts, their size is proportional to the number of sequences in a dataset. In each pair, the top chart shows proportions of diagnosed species based on the length of their shortest recovered mDNCs (for example, a red segment corresponds to the proportion of species, for which at least one type one character (length = 1) is recovered. The bottom chart of each pair shows proportions of diagnosed species based on the length of their rDNCs; black segment in Xenuroturris dataset corresponds to the species *Iotyrris conotaxis,* for which no sufficiently robust rDNC could be recovered~~Workflow of haplotype subsampling analyses in which we tested reliability of mDNCs (orange elements), and rDNCs (green elements). Box C corresponds to the random haplotype sampling.~~

~~**Figure 4.** Alluvial diagram summarizing mDNC and rDNC species diagnoses retrieved from the analyzed data sets: Pontohedyle (*cox1* and 28S), Xenuroturris, Daphnia and Conus. The height of each block corresponds to the number of species: (left) in each data set, (centre) with~~

~~shortest retrieved mDNC comprising 1, 2, and 3 sites, (right) with rDNCs of varying length. The proportion of species for which no sufficiently reliable rDNC could be identified is marked with red triangle in the right column~~.

~~**Figure 5**. Alluvial diagram summarizing mDNC and rDNC species diagnoses retrieved from the four analyzed Tanytarsus data sets (*cox1*, AATS, CAD, PGDI). Column designation is the same as in fig. 4.~~

Figure ~~6~~4. Reproducibility of rDNC scoring. Dots connected by a thick line denote mean scores of the rDNCs (annotated at each dot); vertical bars correspond to the SD; thin lines connect data points showing minimal and maximal scores of respective rDNCs. Grey shading marks area above the reliability threshold of 75. A) *Pontohedyle brasilensis* 28S; B) *Xenuroturris legitima cox1*; C) *Iotyrris olangoensis cox1*; D) *Conus ebraeus cox1*.

Figure ~~7~~5. Haplotype h-subsampling and associated dynamics of mDNC reliability in the analyzed data sets~~:~~. In this analysis, we were sampling an increasing number of unique of a query species haplotypes, and of all reference taxa; 10 iterations were made for each tested sample size. The sampled haplotypes were combined in partial data sets that were passed to MOLD. In the output from each partial dataset we calculated proportion of the mDNCs that remained valid in the context of the entire query and reference taxa diversity (i.e. entire dataset). This proportion is plotted depending on the number of sampled haplotypes for query species in six analyzed empirical data sets: a) Xenuroturris ; b) Daphnia; c) Conus; d) Tanytarsus AATS; e) Tanytarsus CAD; f) Tanytarsus PGDI. Error bars correspond to the SD. The plots demonstrate that the mDNCs reliability grows slowly, and remains low when small fraction of the species diversity is sampled.

Figure ~~8~~6. Scatterplots of mDNC reliability for mDNCs of different lengths depending on the sampled genetic diversity of query taxon~~,~~. Here the proportion of mDNCs valid in the context of the entire dataset is plotted separately for mDNCs comprising one site ~~indicated separately for 1-site mDNCs~~ (blue), 2 -site~~s~~ ~~mDNCs~~ (green), and 3- site~~s~~ ~~mDNCs~~ (orange) a) *Xenuroturris legitima cox1*; b) *Iotyrris cingulifera cox1*; c) *Conus ebraeus cox1*. The plots demonstrate that shorter mDNCs are more reliable than the longer ones.

Figure 97. Different regimes of haplotype subsampling and associated dynamics of rDNC reliability. In this analysis, we were sampling an increasing number of unique haplotypes of a query species, but treated reference species differently in the h- and hspp- resampling. The sampled haplotypes were combined in partial data sets that were passed to MOLD. In the output from each partial dataset we checked, whether the recovered rDNCs remained valid in the context of the entire query and reference taxa diversity (i.e. entire dataset). This test was repeated 10 times for each sample size, the output of each iteration recorded as 1 or 0, and then divided by 10, to provide a measure of rDNC reliability associated with each sampled number of haplotypes. It is plotted depending on the number of sampled haplotypes for query species in analyzed empirical data sets. a – c. h-subsampling (each species represented in each partial data set). Arrows mark sampling fraction at which confidence threshold of 0.8 has been reached for respective species in mDNC subsampling. a) Xenuroturris *cox1*; b) Daphnia *cox1*; c) Conus *cox1*. d - i. hspp-subsampling (partial data sets varying in both the species and the haplotype per species composition). d) Xenuroturris *cox1*; e) Daphnia *cox1*; f) Conus *cox1*; g) Tanytarsus AATS; h) Tanytarsus CAD; i) Tanytarsus PGDI. The reliability of rDNCs grows notably faster than that of mDNCs.

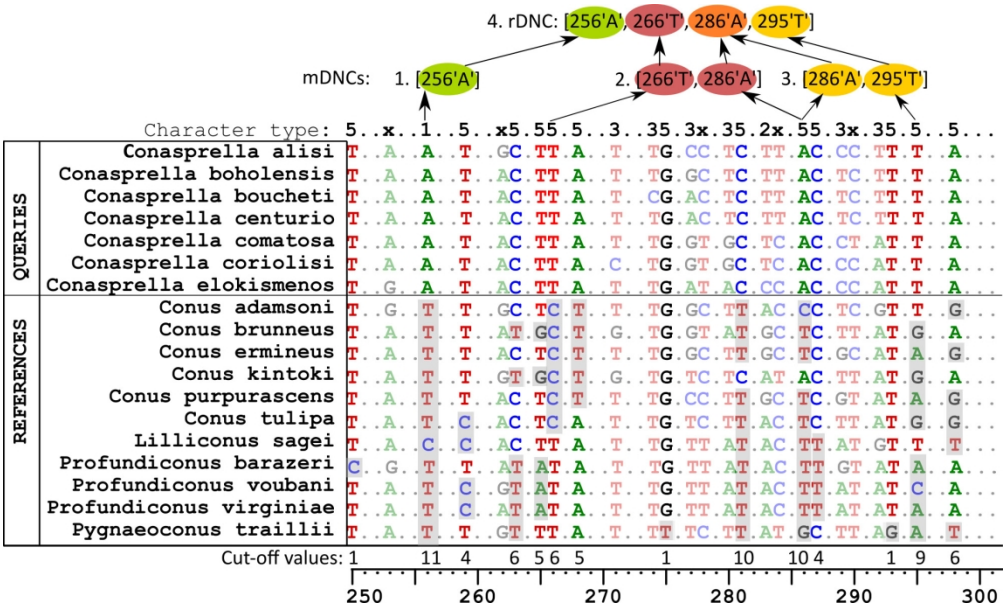

Figure 1. Major types of DNA characters in the alignment of cox1 of the family Conidae; query taxon genus Conasprella. Invariable sites are represented by dots. The character type (1, 2, 3 and 5) is indicated for informative characters; variable sites not representing any type are marked by 'x'. All sites not used by MOLD are shaded. For Type 5 characters, sites in the reference taxa with a different nucleotide from that in the query taxon are marked in grey. Their count corresponds to the cut-off value reported below the alignment. Example mDNCs (1 – 3), and rDNC (4) are shown above the alignment; the numbers of constituent sites correspond to their position in the alignment.

168x100mm (300 x 300 DPI)

Table 1. Datasets analysed in the present study

Dataset	Alignment length	Species	Unique Haplotypes	Variable positions	Query species in subsampling	Reference
Pontohedyle cox1	655	9	25	309	no	Jorger & Schrodl 2012
Pontohedyle 28S	984	10	15	105	no	Jorger & Schrodl 2012
X-I *cox1*	658	11	129	196	*X. legitima, I. olangoensis, I. cingulifera*	Abdelkrim et al. 2018
Daphnia cox1	657	87	573	373	*D. longispina, D. laevis, D. melanica, D. pulex*	Crease et al. 2012 extended
Conus cox1	658	187	984	361	*C. sanguinolentus, C. ebraeus, C. chaldaeus*	Puillandre et al. 2014, Duda et al. 2012
Tanytarsus AATS	405	99	180	219	*T. thomasi, T. tongmuensis*	Lin et al. 2018
Tanytarsus CAD	909	88	173	524	*T. thomasi, T. tongmuensis*	Lin et al. 2018
Tanytarsus PGDI	748	99	185	334	*T. thomasi, T. tongmuensis*	Lin et al. 2018

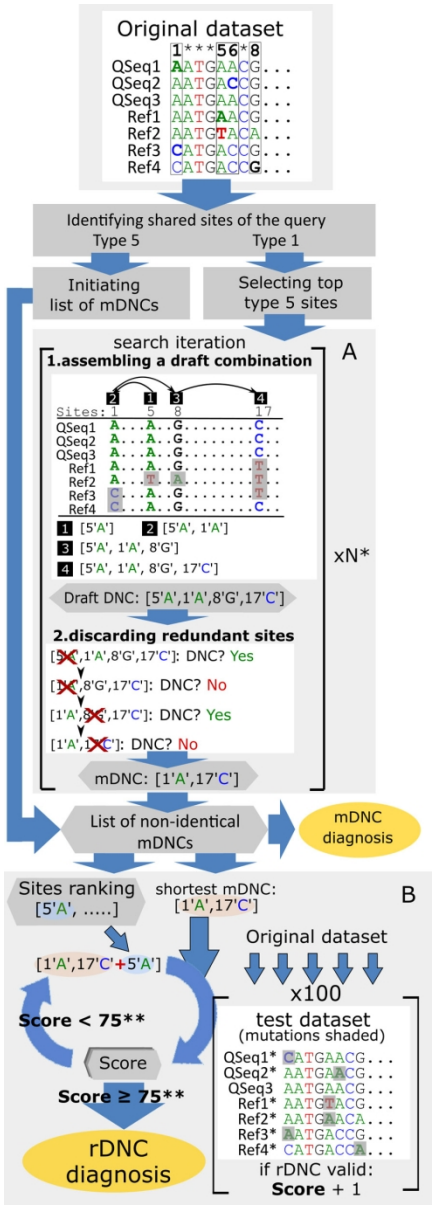

Figure 2. Workflow of standard MOLD distribution. Box A represents the module for mDNC recovery. Box B represents the module that transforms the catalog of mDNCs into a set of rDNCs and outputs the rDNC with the highest score. Grey rectangles show analysis tasks, grey hexagons – intermediate outputs used by the program, yellow ellipses – final output available to users.

75x209mm (300 x 300 DPI)

Table 2. Results of mDNCs recovery for tested query species

Data set	Query taxon	Number of unique haplotypes	Number of Type 1 characters	Remarks for the query
Xenuroturris	*X. legitima*	21	25	diversified, geographic structure & disinctive
Xenuroturris	*I. olangoensis*	17	2	diversified & part of complex
Xenuroturris	*I. cingulifera*	36	3	highly diversified, geographic structure & part of complex
Daphnia	*D. longispina*	8	no	moderately diversified & disinctive
Daphnia	*D. laevis*	36	no	highly diversified & distinctive
Daphnia	*D. melanica*	9	no	moderately diversified & part of complex
Daphnia	*D. pulex*	41	no	diversified & part of complex
Conus	*C. sanguinolentus*	23	no	moderately diversified & disinctive
Conus	*C. ebraeus*	48	no	diversified & part of complex
Conus	*C. chaldaeus*	41	no	diversified & part of complex

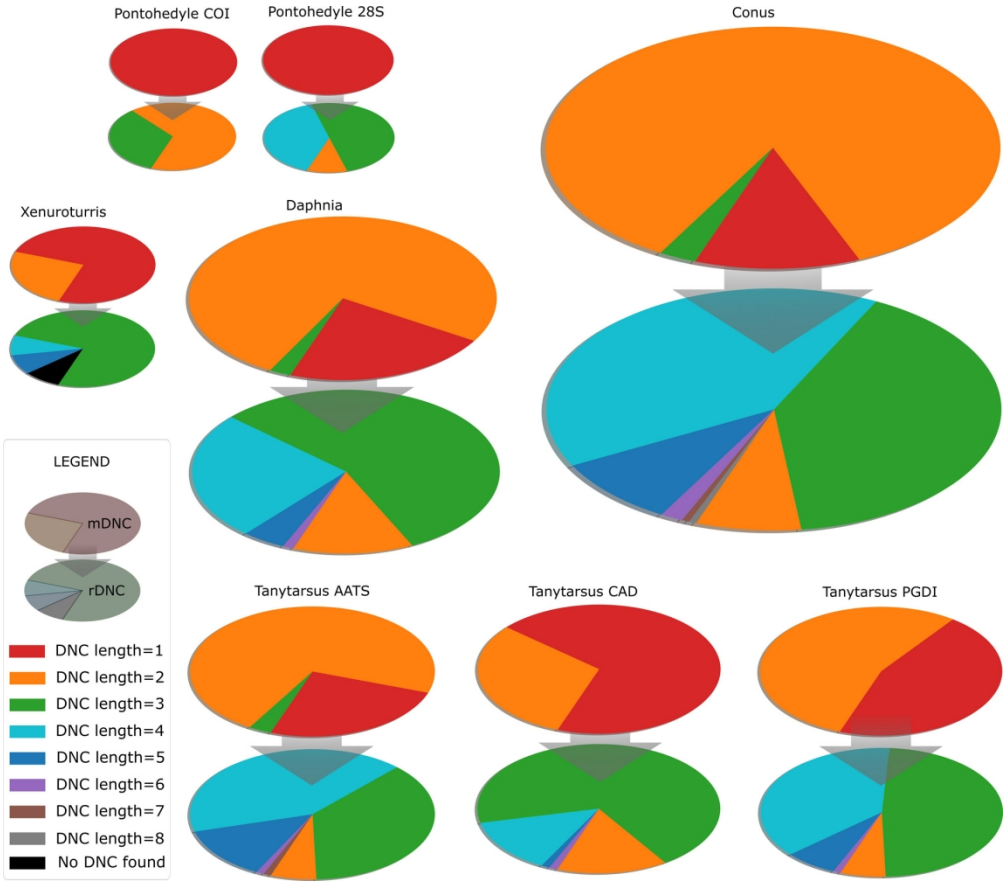

Figure 3. Results of MOLD application to the empirical datasets, when all species of a dataset were diagnosed. Each empirical data set is represented by a pair of charts, their size is proportional to the number of sequences in a dataset. In each pair, the top chart shows proportions of diagnosed species based on the length of their shortest recovered mDNCs (for example, a red segment corresponds to the proportion of species, for which at least one type one character (length = 1) is recovered. The bottom chart of each pair shows proportions of diagnosed species based on the length of their rDNCs; black segment in Xenuroturris dataset corresponds to the species Iotyrris conotaxis, for which no sufficiently robust rDNC could be recovered.

169x148mm (299 x 299 DPI)

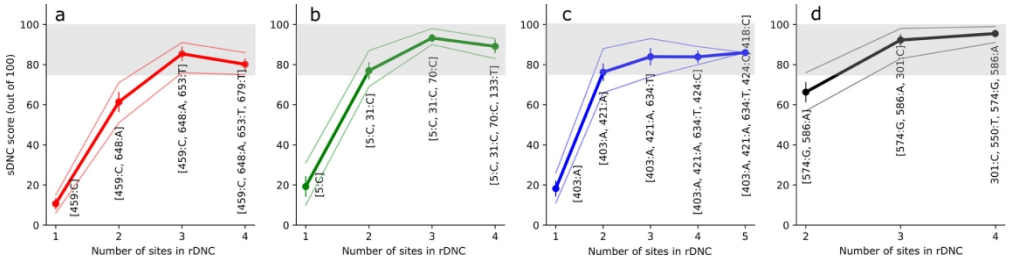

Figure 4. Reproducibility of rDNC scoring. Dots connected by a thick line denote mean scores of the rDNCs (annotated at each dot); vertical bars correspond to the SD; thin lines connect data points showing minimal and maximal scores of respective rDNCs. Grey shading marks area above the reliability threshold of 75. A) Pontohedyle brasilensis 28S; B) Xenuroturris legitima cox1; C) Iotyrris olangoensis cox1; D) Conus ebraeus cox1.

170x42mm (300 x 300 DPI)

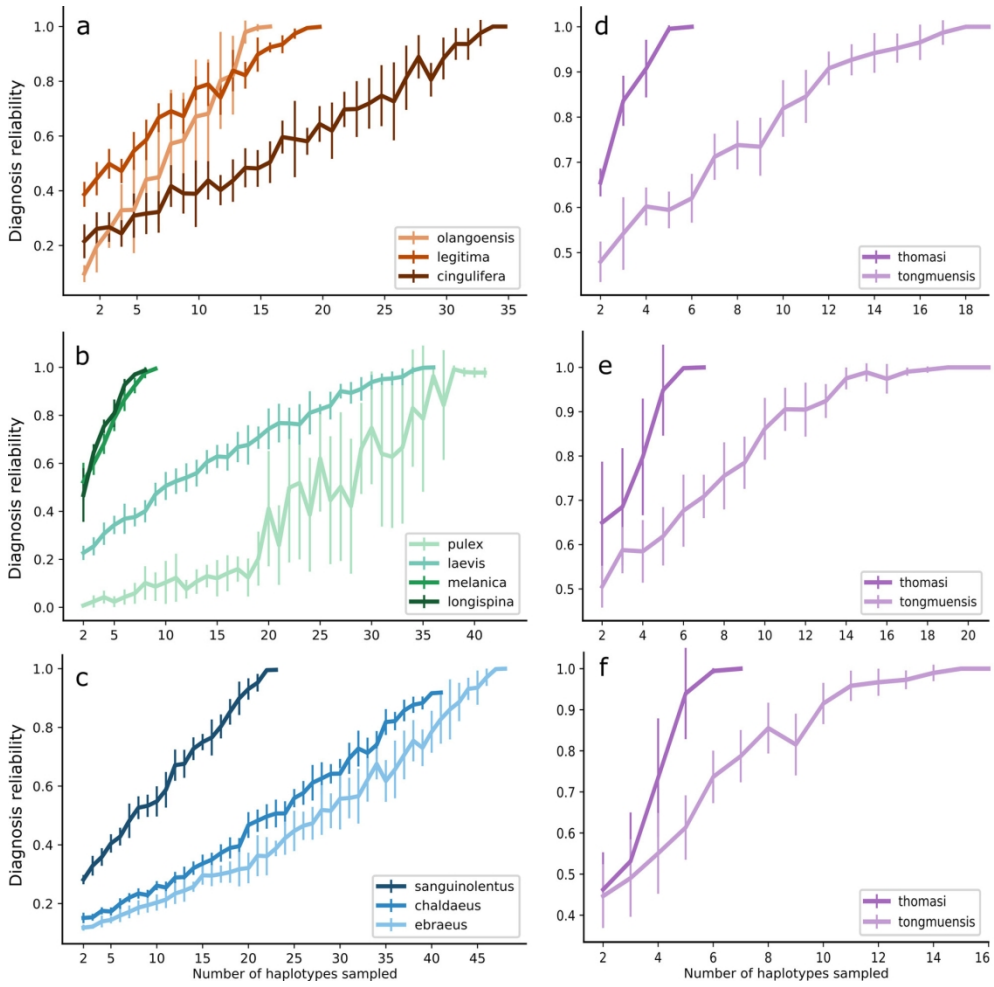

Figure 5. Haplotype h-subsampling and associated dynamics of mDNC reliability in the analyzed data sets. In this analysis, we were sampling an increasing number of unique of a query species haplotypes, and of all reference taxa; 10 iterations were made for each tested sample size. The sampled haplotypes were combined in partial data sets that were passed to MOLD. In the output from each partial dataset we calculated proportion of the mDNCs that remained valid in the context of the entire query and reference taxa diversity (i.e. entire dataset). This proportion is plotted depending on the number of sampled haplotypes for query species in six analyzed empirical data sets: a) Xenuroturris ; b) Daphnia; c) Conus; d) Tanytarsus AATS; e) Tanytarsus CAD; f) Tanytarsus PGDI. Error bars correspond to the SD. The plots demonstrate that the mDNCs reliability grows slowly, and remains low when small fraction of the species diversity is sampled.

169x165mm (300 x 300 DPI)

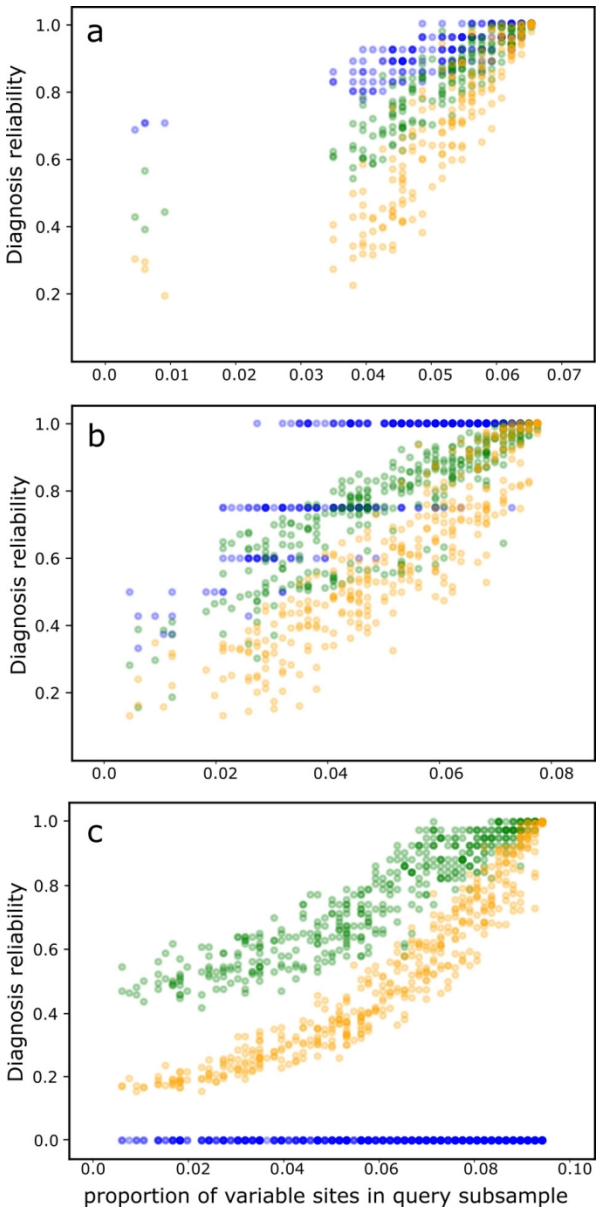

Figure 6. Scatterplots of mDNC reliability for mDNCs of different lengths depending on the sampled genetic diversity of query taxon. Here the proportion of mDNCs valid in the context of the entire dataset is plotted separately for mDNCs comprising one site (blue), 2 sites (green), and 3 sites (orange) a) Xenuroturris legitima cox1; b) Iotyrris cingulifera cox1; c) Conus ebraeus cox1. The plots demonstrate that shorter mDNCs are more reliable than the longer ones.

81x165mm (300 x 300 DPI)

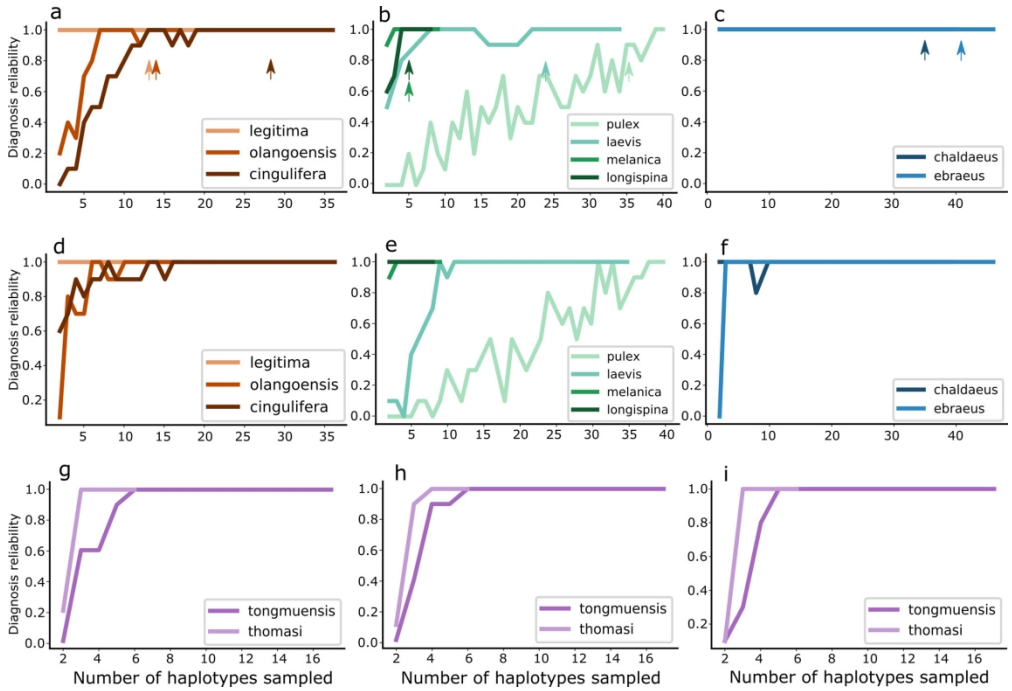

Figure 7. Different regimes of haplotype subsampling and associated dynamics of rDNC reliability. In this analysis, we were sampling an increasing number of unique haplotypes of a query species, but treated reference species differently in the h- and hspp- resampling. The sampled haplotypes were combined in partial data sets that were passed to MOLD. In the output from each partial dataset we checked, whether the recovered rDNCs remained valid in the context of the entire query and reference taxa diversity (i.e. entire dataset). This test was repeated 10 times for each sample size, the output of each iteration recorded as 1 or 0, and then divided by 10, to provide a measure of rDNC reliability associated with each sampled number of haplotypes. It is plotted depending on the number of sampled haplotypes for query species in analyzed empirical data sets. a – c. h-subsampling (each species represented in each partial data set). Arrows mark sampling fraction at which confidence threshold of 0.8 has been reached for respective species in mDNC subsampling. a) Xenuroturris cox1; b) Daphnia cox1; c) Conus cox1. d - i. hspp-subsampling (partial data sets varying in both the species and the haplotype per species composition). d) Xenuroturris cox1; e) Daphnia cox1; f) Conus cox1; g) Tanytarsus AATS; h) Tanytarsus CAD; i) Tanytarsus PGDI. The reliability of rDNCs grows notably faster than that of mDNCs.

170x115mm (300 x 300 DPI)