



HAL
open science

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

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

► **To cite this version:**

A.E. Fedosov, Guillaume Achaz, Andrey Gontchar, Nicolas Puillandre. MOLD, a novel software to compile accurate and reliable DNA diagnoses for taxonomic descriptions. *Molecular Ecology Resources*, 2022, 5, pp.2038-2053. 10.1111/1755-0998.13590 . hal-03663253

HAL Id: hal-03663253

<https://hal.science/hal-03663253>

Submitted on 10 May 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

MOLECULAR ECOLOGY RESOURCES

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

Journal:	<i>Molecular Ecology Resources</i>
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

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

3 Alexander Fedosov^{1,2}, Guillaume Achaz^{2,3,4}, Andrey Gontchar⁵, Nicolas Puillandre²

4

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

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

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

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

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

16 **ABSTRACT**

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

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

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

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

42

43 **1. Introduction**

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

51 Whereas traditionally descriptions of taxa are mainly based on morphological data (Dunn
52 et al. 2003; Cook et al. 2010), non-morphological characters, in particular DNA characters, are
53 equally accepted by all nomenclatural codes (Cook et al. 2010; Renner, 2016). The amount of
54 DNA sequence data available to taxonomists has steadily grown over the last decades, and
55 currently accumulates at an ever-increasing rate following the recent advent of high-
56 throughput sequencing. Therefore, DNA sequence data are often more accessible than rare
57 taxonomic expertise (Cook et al. 2010) and it is not surprising that DNA data is now widely used
58 in phylogenetics, species delimitation (Fujita et al. 2012; Puillandre et al. 2012; Pante et al.
59 2015) and specimen identification (Herbert et al. 2003; Janzen et al. 2009; Goldstein & De Salle
60 2011). Conversely, the use of DNA data in formal descriptions remains scarce, the number of
61 species described with DNA data being two orders of magnitude smaller than those described
62 without (Renner 2016; Kühn & Haase 2020). However, the inclusion of DNA data in taxonomic
63 descriptions has the potential to greatly increase their quality and usability, and is conceptually
64 sound, as long as taxa discovery relies on a comprehensive integrative taxonomy framework. In
65 this perspective, some recent publications strongly recommend that the relevant taxonomic
66 codes should promote the use of DNA-based diagnoses (Renner 2016). There is a pressing need
67 to establish clear universal requirements for DNA diagnoses, especially in view of the recent
68 notorious article (Sharkey et al. 2021), where a misuse of molecular data led to the publication
69 of multiple, arguably very problematic taxonomic acts (Meier et al. 2021).

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

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

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

95

96 1.1. Single nucleotide Signature DNA Characters in Multispecies Alignments

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

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

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

126

127 1.2. Composite Signature DNA Characters in Multispecies Alignments

128 Despite composite DNA characters being previously implemented in cladistic haplotype analysis
129 (CHA - Brower, 1999) and the characteristic attribute organization system (CAOS - Sarkar et al.
130 2008), both approaches rely on a tree-based algorithm with its inherent drawbacks (see Kühn &
131 Haase 2020). Here we further develop the concept of composite diagnostic DNA characters by
132 introducing a *minimal diagnostic nucleotide combination* or *mDNC* – a combination of
133 nucleotides at selected sites that are shared by all members of the query taxon and by no
134 member of the reference taxa (e.g. combinations 1-4 in Figure 1). An mDNC may comprise two
135 (i.e., paired sites identified by DeSignate), or any larger number of sites (with a limit of ten sites
136 by default). Furthermore, a Type 1 nucleotide character can be considered an mDNC with a
137 single site (mDNC 1 in Figure 1). Consequently, mDNCs are a generalization of the concept of
138 Type 1 characters. As any mDNC unambiguously defines a query taxon, mDNCs can be thought
139 of as a minimal and sufficient condition to assign a specimen, through its DNA sequence, to a
140 query taxon. As such, it is a proper diagnosis.

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

157

158 1.3. Phylogenetic background of signature DNA characters

159 Some concerns have been raised regarding the use of composite DNA characters as diagnoses
160 because of the complex phylogenetic background of their constituent sites (Jörger & Schrödl,
161 2013; Merkelbach & Borges, 2020). We briefly address these issues and their relevance to alpha
162 taxonomy, as it is important for setting the conceptual basis upon which we built MOLD. Jörger
163 and Schrödl stated that ‘compound characters can be unique for certain species, but they may
164 have evolved from several independent mutation events’ implying ‘low probabilities of
165 homology’ (Jörger & Schrödl 2013: 20). We argue that as long as each constituent site of an
166 mDNC or rDNC is fixed within a query taxon (and only such sites are used by MOLD), their
167 homology in a broader phylogenetic context is not relevant when diagnosing a taxon. Likewise,
168 we disagree with the assertion by Merkelbach & Borges (2020) that plesiomorphic characters
169 should not be included in diagnoses, which would disqualify Type 5 characters from being used
170 as signature characters. First, plesiomorphy or apomorphy cannot be simply deduced from the
171 character type, and requires reconstruction of the character evolution across the data set
172 (Jörger & Schrödl 2013). Second, reporting only apomorphic characters, or reporting only those
173 characters resulting from a single mutation event, is not required by the nomenclatural codes.

174 The purpose of a diagnosis is to communicate the identity of a taxon, and therefore a diagnosis
175 focuses on character states, but not on their evolutionary history. Most, if not all, traditional
176 diagnoses comprise *informative* morphological characters irrespective of their apomorphic or
177 plesiomorphic nature, or whether they may be homologous or analogous. A diagnosis in
178 general, and *a fortiori* a tree-independent diagnosis, could thus be compared to an
179 identification key – following the key enables allocation of a specimen to a certain taxon, but
180 the consecutive dichotomies of the key are not expected to match events in the evolutionary
181 history of that taxon.
182

For Review Only

183 2. Material and Methods

184 2.1. Overview of the MOLD software

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

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

199

200

201 2.2. MOLD algorithm in detail

202 2.2.1. Building a list of mDNCs

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

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

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

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

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

234

235 2.2.2. Compiling an rDNC from the list of mDNCs

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

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

248 reliability threshold (then the best-scoring rDNC is sent to output), or when the rDNC comprises
249 10 nucleotide sites. In the latter case, the rDNC is output with an alert message if at any step it
250 scored above the reliability threshold. If the scores remain consistently below the reliability
251 threshold, a message is output that no sufficiently reliable rDNC could be compiled.

252

253 2.2.3 rDNC scoring

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

260 Each artificial sequence is generated by introducing p nucleotide substitutions into an
261 existing sequence, where p is a random natural number drawn from a uniform distribution [1,
262 $k*L/100$]. In the latter expression k is the natural number corresponding to the desired
263 maximum % of sequence divergence between the original and mutated DNA sequence, and L is
264 the sequence length. Mutations are introduced only at polymorphic sites by substituting the
265 original nucleotide by one of the three others, selected randomly with respect to their
266 observed frequencies at this site in the original alignment. Ten artificially mutated sequences
267 are created for each species in the original alignment from randomly sampled original
268 sequences. For species with more than 10 DNA sequences in the original alignment, randomly
269 sampled unchanged sequences are added to the test data set to match the original number of
270 sequences for this species. Thus, a test data sets has at least 10 sequences per species.

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

278

279 2.3. Testing MOLD

280 2.3.1. Testing MOLD on published data sets

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

292 For the *Pontohedyle* data set, the alignments supplied by the authors (Jörger & Schrödl,
293 2013) were used as MOLD input. For the other data sets, we generated alignments using
294 MAFFT v.7 (<https://mafft.cbrc.jp/alignment/server/>) (Katoh et al. 2019) with FFT-NS-2 strategy.
295 The alignments were then translated using MEGA v.6 (Tamura et al. 2014) to ensure
296 consistency of the amino-acid (AA) sequences and lack of premature stop-codons. We ran
297 RAxML v.8.2.12 (Stamatakis et al. 2006) on the Cypress Gateway (Miller et al. 2010), with three
298 codon positions allocated to separate partitions, to check that sequences assigned to the same
299 species formed a clade.

300 First, to evaluate the general performance of MOLD, we diagnosed all species from each
301 analyzed data set, with 10,000 search iterations and 100 Type 5 sites considered for inclusion
302 into mDNCs. When compiling rDNCs, mutated sequences in test data sets were a maximum of
303 1% different from the original sequences from which they were derived, which is within the
304 typical ~~K2P~~ genetic distance for intra-specific comparisons of all analyzed taxa and genes
305 (Abdelkrim et al. 2018; Puillandre et al. 2014; Lin et al. 2018; Hebert et al. 2003).

306

307 2.3.2. Comparison of MOLD with other tools

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

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

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

325

326 2.3.3. Assessment of the effect of sampling on the robustness of the DNA based diagnosis

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

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

337

338 2.3.3.1. h-sampling

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

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

358

359 2.3.3.2. *hspp-sampling*

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

374 3. Results

375 3.1. DNA diagnoses recovered by MOLD

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

381 Type 1 sites were detected ~~in all~~ for each species in the smallest *Pontohedyle cox1* and
382 *Pontohedyle 28S* data sets (entirely red top-left charts on the figure (Fig. 43)). However, in larger
383 data sets some species lacked Type 1 sites, and could only be diagnosed by 2-site or even 3-site
384 mDNCs (Figs 4, 5). The proportion of species lacking Type 1 sites is highest in those data sets
385 that include both a larger number of species and a larger number of unique haplotypes per
386 species. Only 19 species (=22%) in the *Daphnia cox1* data set and 22 species (=12%) in the
387 *Conus cox1* data set could be diagnosed by Type 1 sites (Figs 4, 5, 3).

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

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

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

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

406 consistency of the output rDNC from run to run, we performed rDNC recovery in 10 replicates
407 for 16 query taxa in seven data sets (totaling 20 series). The same rDNC was recovered in all 10
408 runs in 12 out of the 20 series (Table S3, Supplementary data 2). The rDNCs from different runs
409 varied in length by one site, but otherwise were identical in six series, and only in one series,
410 *Daphnia pulex* (*Daphnia*), did the rDNC length vary by three sites. Only in one series, *Lotyrris*
411 *cingulifera* (*Xenuroturrus*), did different runs employ alternative subsets of nucleotide sites in
412 the production of rDNCs. In no instance was an rDNC identified successfully in some runs but
413 not in others. Finally, to evaluate consistency of scoring from run to run, we performed rDNC
414 recovery in 30 replicates for the query taxa *Pontohedyle brasiliensis* (*Pontohedyle* 28S),
415 *Xenuroturrus legitima*, *Lotyrris cingulifera* (both *Xenuroturrus*) and *Conus ebraeus* (*Conus*).
416 Despite an often notable difference between the minimum and maximum scores (Fig. 46), this
417 difference is reduced with longer rDNCs to consistently fall above the selected reliability
418 threshold of 0.75 (grey zone).

419

420 3.2. Performance of MOLD in comparison with previously available tools

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

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

438

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 7a5 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 *Xenuroturrus legitima* (Fig. 7a5 a), *Daphnia longispina* and *D. laevis* (Fig. 7b5 b), *Conus*
465 *sanguinolentus* (Fig. 7c5 c) and/or in species represented in the entire data set by fewer unique
466 haplotypes: *Daphnia melanica* and *Daphnia longispina* (Fig. 7b5 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 7d5 d – f). It is noteworthy that, because each reference species was

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

476 In Figure 86, each partial data set is represented on a respective scatterplot by three data
477 points, which show mDNC reliability as a function of the query taxon sampled diversity,
478 separately for 1-site mDNCs (blue), 2-site mDNCs (green) and 3-site mDNCs (orange). The areas
479 occupied by green records are higher than those occupied by orange ones, indicating higher
480 reliability of the 2-site mDNCs over 3-site mDNCs. In *Xenuroturris legitima* (Fig. 8a6a), which
481 can be diagnosed by 25 1-site mDNCs in the entire *Xenuroturris* data set, all but four blue
482 records are above the threshold of 0.75, suggesting that 1-site mDNCs allow for a very reliable
483 diagnosis in this species. Therefore, shorter mDNCs are indeed generally more reliable than
484 longer ones, but overall mDNC based diagnoses are weak, unless based on a thorough sampling
485 of both the query and reference taxa.

486

487 3.4. *Reliability of the rDNC-based diagnoses* ~~rDNC reliability analysis~~

488 When haplotype subsampling was performed to assess the rDNC reliability dynamics, the
489 obtained graphs were notably different from those for mDNCs (Fig. 9a-7 a – c, Supplementary
490 data 5). In all analyzed query taxa, with the exception of *Daphnia pulex* (Fig. 9b7 b), the
491 arbitrary threshold of 0.75 was reached earlier, compared to mDNC subsampling (marked with
492 arrows for respective taxa), and a plateau was reached soon after. Failure to recover a reliable
493 diagnosis for the assemblage of sequences here attributed to *D. pulex* based on smaller
494 sampled diversity is most probably the result of complicated taxa delimitation. Even if *D. pulex*
495 constitutes a monophyletic group, the maximum intraspecific K2P genetic distance for *D. pulex*
496 as defined herein (0.039) exceeds more than two-fold the minimum genetic distances between
497 *D. pulex* and *D. middendorffiana* (0.014) and between *D. pulex* and *D. melanica* (0.016).

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

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

518 **4. Discussion**

519 *4.1. MOLD and other tools for diagnostic DNA character identification*

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

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

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

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

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

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

568

569 *4.2. Taxonomic sampling and robustness of DNA-based diagnoses*

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

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

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

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

608

609 *4.3. Which sources of genetic data can be used*

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

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

629 There are some additional criteria that should be taken into consideration, in particular the
630 ease and reproducibility of the laboratory protocols and the availability of comparative data.
631 Most data sets analyzed in the present study comprise the widely used barcode marker *cox1*.
632 The pros of using this fragment are well known: it is informative, it can be confidently aligned
633 even among divergent taxa, degrees of its variation within and among taxa are well
634 documented, and there is a wealth of data available for this fragment in the NCBI and BOLD
635 databases. The cons, although mostly lineage-specific, are the low resolution in some basal
636 metazoan lineages, such as sponges and corals (Huang et al. 2008; Vargas et al. 2012), taxa-
637 specific mitochondrial introgression (Toews & Brelsford, 2012), or pseudogenization (Song et al.
638 2008). Therefore, examination of the gene tree is mandatory prior to any attempts to propose
639 DNA based diagnosis even for such a broadly used marker as *cox1*.

640 **Data ~~availability~~accessibility**

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

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

648

649 **Acknowledgments**

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

For Review Only

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., Loughheed 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

- 685 from *Cortinarius* (Basidiomycota), a highly diverse and widespread ectomycorrhizal genus.
686 *FEMS Microbiol Ecol.* 92(4): fiw045.
- 687 Goldstein P., DeSalle R. 2011. Integrating DNA barcode data and taxonomic practice:
688 Determination, discovery, and description. *Bioessays*, 33: 135-147.
- 689 Hebert P.D.N., Ratnasingham S., DeWaard J.R. 2003. Barcoding animal life: Cytochrome c
690 oxidase subunit 1 divergences among closely related species. *Proc R Soc Lond B Biol Sci*,
691 270, S596–S599.
- 692 Huang D., Meier R., Todd P.A., Chou L.M. 2008. Slow mitochondrial COI sequence evolution at
693 the base of the metazoan tree and its implications for DNA barcoding. *Journal of*
694 *Molecular Evolution* 66(2): 167-74.
- 695 Hütter T., Ganser M.H., Kocher M., Halkic M., Agatha S., Augsten N. 2020. DeSignate: detecting
696 signature characters in gene sequence alignments for taxon diagnoses. *BMC*
697 *Bioinformatics*, 21, 151.
- 698 ICZN. 1999. International Code of Zoological Nomenclature. 4th ed. London, UK: The
699 International Trust for Zoological Nomenclature. 306 pp. Available from:
700 [https://www.iczn.org/the-code/the-international-code-of-zoological-nomenclature/the-](https://www.iczn.org/the-code/the-international-code-of-zoological-nomenclature/the-code-online/)
701 [code-online/](https://www.iczn.org/the-code/the-international-code-of-zoological-nomenclature/the-code-online/) (accessed 21 September 2020).
- 702 Janzen D.H., Hallwachs W., Blandin P., Burns J.M., et al. 2009. Integration of DNA barcoding into
703 an ongoing inventory of complex tropical biodiversity. *Molecular Ecology Resources*, 9, 1–
704 26.
- 705 Jörger K.M., Schrödl M. 2013. How to describe a cryptic species? Practical challenges of
706 molecular taxonomy. *Frontiers in Zoology*, 10, 1–27.
- 707 Jörger K.M., Norenburg J.L., Wilson N.G., Schrödl M. 2012. Barcoding against a paradox?
708 Combined molecular species delineations reveal multiple cryptic lineages in elusive
709 meiofaunal sea slugs. *BMC Evolutionary Biology*, 12, 245.
- 710 Katoh K., Rozewicki J., Yamada K.D. 2019. MAFFT online service: multiple sequence alignment,
711 interactive sequence choice and visualization. *Briefings in bioinformatics*, 20 (4), 1160–
712 1166.
- 713 Knowlton N. 2000. Molecular genetic analyses of species boundaries in the sea. *Hydrobiologia*,
714 420, 73–90.
- 715 Kühn A.L., Haase M. 2020. QUIDDICH: QUick IDentification of Dlagnostic CHaracters. *Journal of*
716 *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., Kaufenstein 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

- 750 403 new species in 11 subfamilies of Costa Rican braconid parasitoid wasps, including
751 host records for 219 species. *ZooKeys* 4.
- 752 Song H., Buhay J. E., Whiting M.F., Crandall K. A. 2008. Many species in one: DNA barcoding
753 overestimates the number of species when nuclear mitochondrial pseudogenes are
754 coamplified. *Proceedings of the National Academy of Sciences*, 105 (36), 13486-13491.
- 755 Stamatakis A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with
756 thousands of taxa and mixed models. *Bioinformatics*, 22, 2688-2690.
- 757 Stielow B., Bratek Z., Orczan A.K.I, Rudnoy S., Hensel G. 2011. Species Delimitation in
758 Taxonomically Difficult Fungi: The Case of *Hymenogaster*. *PLoS ONE* 6(1), e15614.
- 759 Toews D.P.L., Brelsford A. 2012. The biogeography of mitochondrial and nuclear discordance in
760 animals. *Molecular Ecology*, 21, 3907–3930.
- 761 Tripp E.A., Lendemer J.C. 2014. Sleepless nights: When you can't find anything to use but
762 molecules to describe new taxa. *Taxon*, 63, 969–971.
- 763 Vargas S., Schuste A., Sache K., Büttner G., Schätzl S., Läubli B., Hall K., Hooper J.N., Erpenbeck
764 D., Wörheide G. 2012. Barcoding sponges: an overview based on comprehensive
765 sampling. *PLoS ONE*, 7(7), e39345.
- 766 Vences M., Miralles A., Brouillet B., Ducasse J., Fedosov A.E., Kharchev V., Kostadinov I., Kumari
767 S., Patmanidis S., Scherz M.D., Puillandre N., Renner S.S. 2021. iTaxoTools 0.1: Kickstarting
768 a specimen-based software toolkit for taxonomists. *Megataxa*, 6(2): 77-92.
- 769 Wiens J.J., Servedio, M.R. 2000. Species delimitation in systematics: inferring diagnostic
770 differences between species. *Proceedings of the Royal Society series B*, 267, 631–636.

771 **Captions**

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

781

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

787

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

799

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

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

806

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

810

811 **Figure 64.** Reproducibility of rDNC scoring. Dots connected by a thick line denote mean scores
 812 of the rDNCs (annotated at each dot); vertical bars correspond to the SD; thin lines connect
 813 data points showing minimal and maximal scores of respective rDNCs. Grey shading marks area
 814 above the reliability threshold of 75. A) *Pontohedyle brasiliensis* 28S; B) *Xenuroturrus legitima*
 815 *cox1*; C) *Lotyrris olangoensis cox1*; D) *Conus ebraeus cox1*.

816

817 **Figure 75.** Haplotype h-subsampling and associated dynamics of mDNC reliability in the
 818 analyzed data sets: In this analysis, we were sampling an increasing number of unique of a
 819 query species haplotypes, and of all reference taxa; 10 iterations were made for each tested
 820 sample size. The sampled haplotypes were combined in partial data sets that were passed to
 821 MOLD. In the output from each partial dataset we calculated proportion of the mDNCs that
 822 remained valid in the context of the entire query and reference taxa diversity (i.e. entire
 823 dataset). This proportion is plotted depending on the number of sampled haplotypes for query
 824 species in six analyzed empirical data sets: a) *Xenuroturrus* ; b) *Daphnia*; c) *Conus*; d) *Tanytarsus*
 825 *AATS*; e) *Tanytarsus CAD*; f) *Tanytarsus PGDI*. Error bars correspond to the SD. The plots
 826 demonstrate that the mDNCs reliability grows slowly, and remains low when small fraction of
 827 the species diversity is sampled.

828

829 **Figure 86.** Scatterplots of mDNC reliability for mDNCs of different lengths depending on the
 830 sampled genetic diversity of query taxon: Here the proportion of mDNCs valid in the context
 831 of the entire dataset is plotted separately for mDNCs comprising one site indicated separately
 832 for 1-site mDNCs (blue), 2-sites mDNCs (green), and 3-sites mDNCs (orange) a) *Xenuroturrus*
 833 *legitima cox1*; b) *Lotyrris cingulifera cox1*; c) *Conus ebraeus cox1*. The plots demonstrate that
 834 shorter mDNCs are more reliable than the longer ones.

835

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

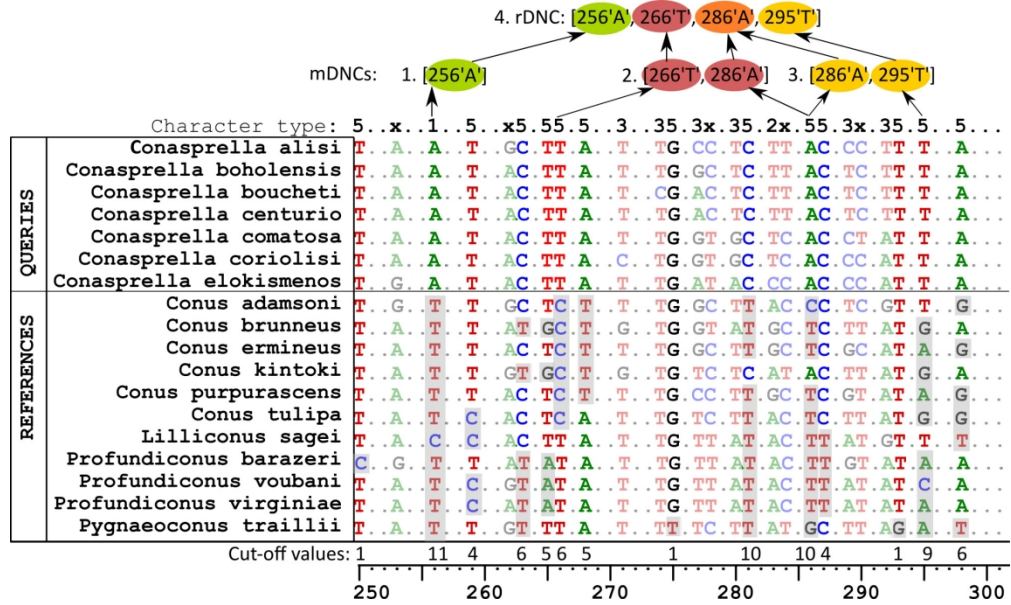


Figure 1. Major types of DNA characters in the alignment of *cox1* of the family Conidae; query taxon genus *Conasprella*. Invariable sites are represented by dots. The character type (1, 2, 3 and 5) is indicated for informative characters; variable sites not representing any type are marked by 'x'. 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
<i>Pontohedyle cox1</i>	655	9	25	309	no	Jorger & Schrodli 2012
<i>Pontohedyle 28S</i>	984	10	15	105	no	Jorger & Schrodli 2012
X-I <i>cox1</i>	658	11	129	196	<i>X. legitima</i> , <i>I. olangoensis</i> , <i>I. cingulifera</i>	Abdelkrim et al. 2018
<i>Daphnia cox1</i>	657	87	573	373	<i>D. longispina</i> , <i>D. laevis</i> , <i>D. melanica</i> , <i>D. pulex</i>	Crease et al. 2012 extended
<i>Conus cox1</i>	658	187	984	361	<i>C. sanguinolentus</i> , <i>C. ebraeus</i> , <i>C. chaldaeus</i>	Puillandre et al. 2014, Duda et al. 2012
<i>Tanytarsus</i> AATS	405	99	180	219	<i>T. thomasi</i> , <i>T. tongmuensis</i>	Lin et al. 2018
<i>Tanytarsus</i> CAD	909	88	173	524	<i>T. thomasi</i> , <i>T. tongmuensis</i>	Lin et al. 2018
<i>Tanytarsus</i> PGDI	748	99	185	334	<i>T. thomasi</i> , <i>T. tongmuensis</i>	Lin et al. 2018

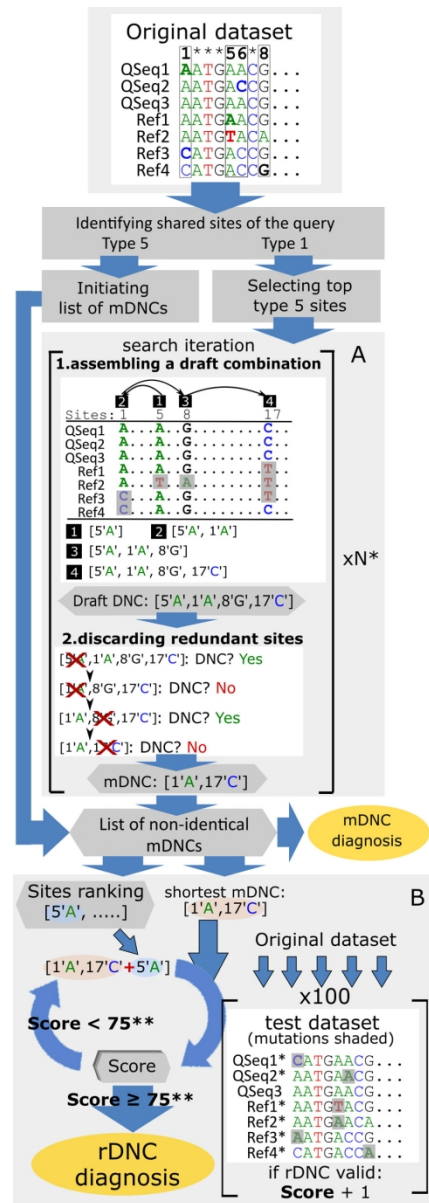


Figure 2. Workflow of standard MOLD distribution. Box A represents the module for mDNC recovery. Box B represents the module that transforms the catalog of mDNCs into a set of rDNCs and outputs the rDNC with the highest score. Grey rectangles show analysis tasks, grey hexagons – intermediate outputs 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
Xenuroturrus	<i>X. legitima</i>	21	25	diversified, geographic structure & disinctive
Xenuroturrus	<i>I. olangoensis</i>	17	2	diversified & part of complex
Xenuroturrus	<i>I. cingulifera</i>	36	3	highly diversified, geographic structure & part of complex
Daphnia	<i>D. longispina</i>	8	no	moderately diversified & disinctive
Daphnia	<i>D. laevis</i>	36	no	highly diversified & distinctive
Daphnia	<i>D. melanica</i>	9	no	moderately diversified & part of complex
Daphnia	<i>D. pulex</i>	41	no	diversified & part of complex
Conus	<i>C. sanguinolentus</i>	23	no	moderately diversified & disinctive
Conus	<i>C. ebraeus</i>	48	no	diversified & part of complex
Conus	<i>C. chaldaeus</i>	41	no	diversified & part of complex

For Review Only

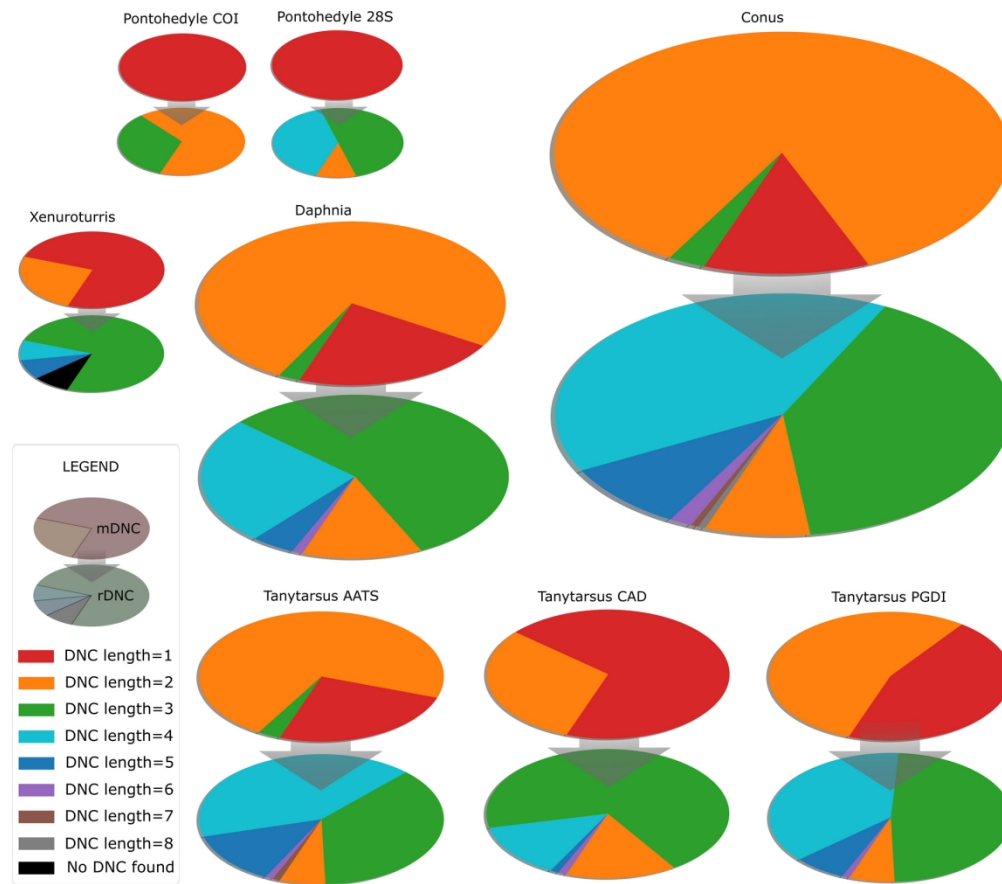


Figure 3. Results of MOLD application to the empirical datasets, when all species of a dataset were diagnosed. Each empirical data set is represented by a pair of charts, their size is proportional to the number of sequences in a dataset. In each pair, the top chart shows proportions of diagnosed species 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)

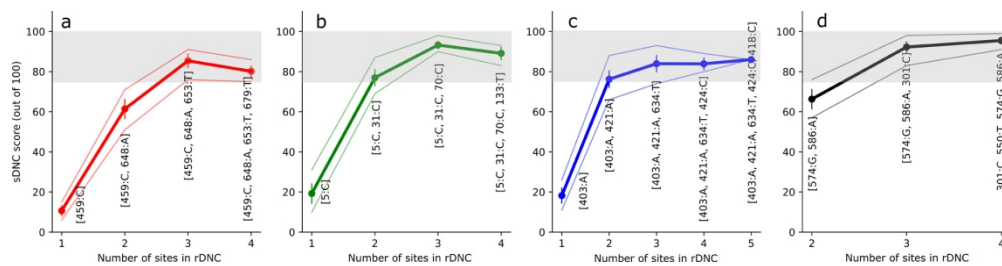


Figure 4. Reproducibility of rDNC scoring. Dots connected by a thick line denote mean scores of the rDNCs (annotated at each dot); vertical bars correspond to the SD; thin lines connect data points showing minimal and maximal scores of respective rDNCs. Grey shading marks area above the reliability threshold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

170x42mm (300 x 300 DPI)

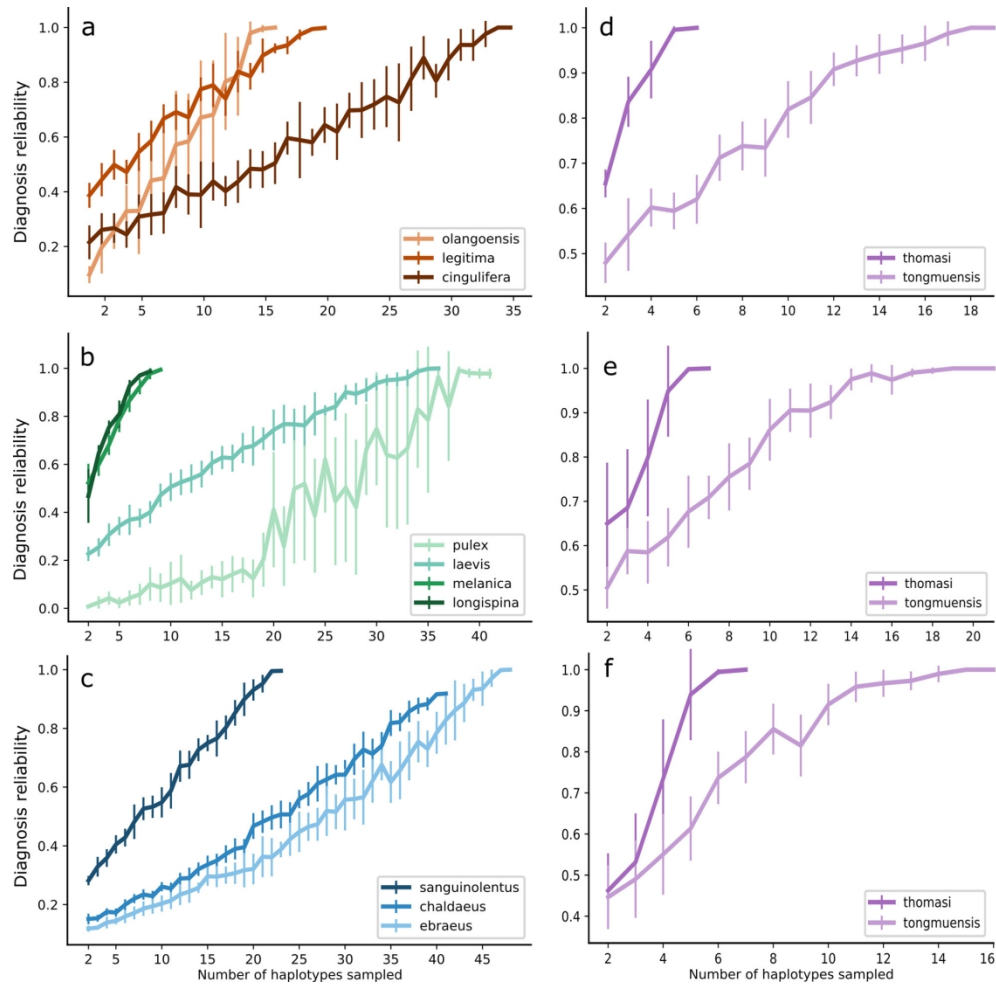


Figure 5. Haplotype h-subsampling and associated dynamics of mDNC reliability in the analyzed data sets. In this analysis, we were sampling an increasing number of unique of a query species haplotypes, and of all reference taxa; 10 iterations were made for each tested sample size. The sampled haplotypes 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) *Xenuroturrus*; 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)

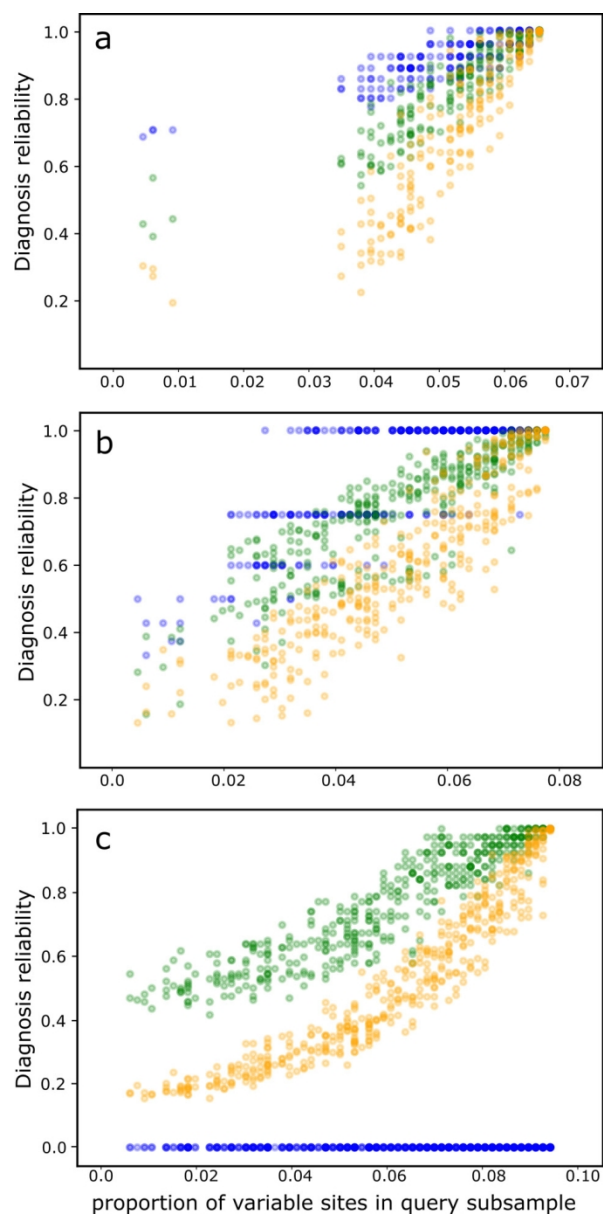


Figure 6. Scatterplots of mDNC reliability for mDNCs of different lengths depending on the sampled genetic diversity of query taxon. Here the proportion of mDNCs valid in the context of the entire dataset is plotted separately for mDNCs comprising one site (blue), 2 sites (green), and 3 sites (orange) 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)

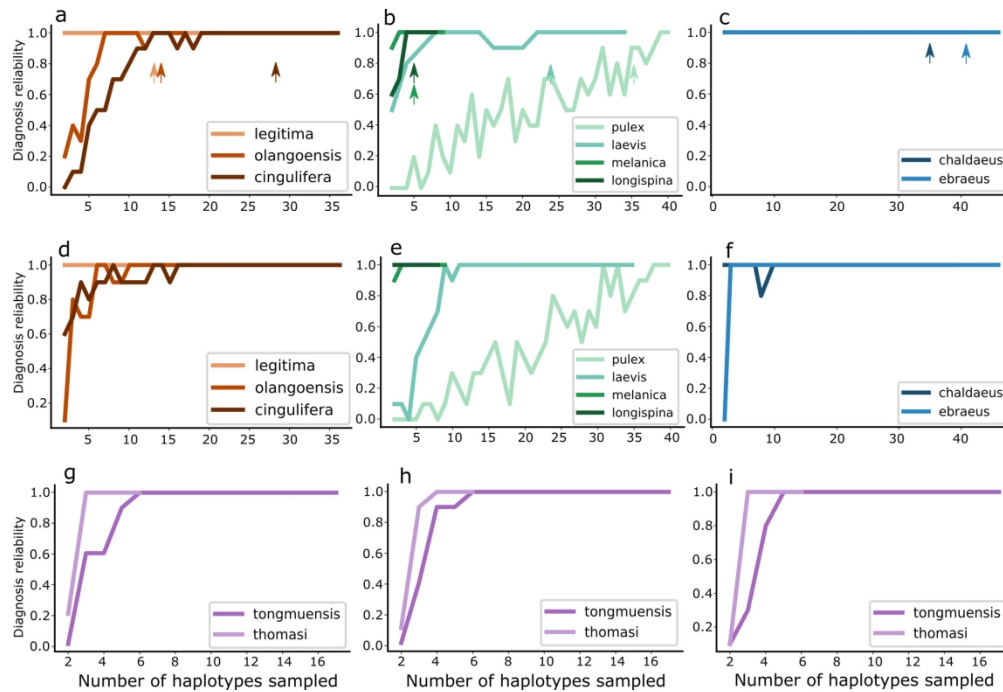


Figure 7. Different regimes of haplotype subsampling and associated dynamics of rDNC reliability. In this analysis, we were sampling an increasing number of unique haplotypes of a query species, but treated reference species differently in the h- and hspp- resampling. The sampled haplotypes were combined 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) *Xenuroturrus 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) *Xenuroturrus 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)