

1

2 **Quantifying uncertainty of taxonomic placement in DNA barcoding and** 3 **metabarcoding**

4

5 *Panu Somervuo*^{1*}, *Douglas Yu*^{2,3,**}, *Charles Xu*^{2,4}, *Yinqiu Ji*², *Jenni Hultman*⁵, *Helena*
6 *Wirta*⁶ and *Otso Ovaskainen*^{1,7}

7

8 ¹Department of Biosciences, P.O. Box 65, FI-00014, University of Helsinki, Finland.

9 ²State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy
10 of Sciences, 32 Jiaochang East Road, Kunming, Yunnan 650223, China

11 ³School of Biological Sciences, University of East Anglia, Norwich Research Park, Norwich, Norfolk, NR47TJ
12 UK

13 ⁴Groningen Institute for Evolutionary Life Sciences, University of Groningen, P.O. Box 11103, 9700 CC
14 Groningen, The Netherlands

15 ⁵Department of Food and Environmental Sciences, P.O. Box 56, FI-00014, University of Helsinki, Finland

16 ⁶Department of Agricultural Sciences, P.O. Box 27, FI-00014, University of Helsinki, Finland

17 ⁷Centre for Biodiversity Dynamics, Department of Biology, Norwegian University of Science and
18 Technology, N-7491 Trondheim, Norway

19 * Corresponding author

20 ** Co-corresponding author

21

22 **Abstract**

- 23 1. A crucial step in the use of DNA markers for biodiversity surveys is the assignment of Linnaean
24 taxonomies (species, genus, etc.) to sequence reads. This allows the use of all the information
25 known based on the taxonomic names. Taxonomic placement of DNA barcoding sequences is
26 inherently probabilistic because DNA sequences contain errors, because there is natural
27 variation among sequences within a species, and because reference databases are incomplete
28 and can have false annotations. However, most existing bioinformatics methods for taxonomic
29 placement either exclude uncertainty, or quantify it using metrics other than probability.
- 30 2. In this paper we evaluate the performance of a recently proposed probabilistic taxonomic
31 placement method PROTAX by applying it to both annotated reference sequence data as well
32 as unknown environmental data. Our four case studies include contrasting taxonomic groups
33 (fungi, bacteria, mammals, and insects), variation in the length and quality of the barcoding
34 sequences (from individually Sanger-sequenced sequences to short Illumina reads), variation
35 in the structures and sizes of the taxonomies (from 800 to 130 000 species), and variation in
36 the completeness of the reference databases (representing 15% to 100% of the species).
- 37 3. Our results demonstrate that PROTAX yields essentially unbiased assessment of probabilities
38 of taxonomic placement, and thus that its quantification of species identification uncertainty is
39 reliable. As expected, the accuracy of taxonomic placement increases with increasing coverage
40 of taxonomic and reference sequence databases, and with increasing ratio of genetic variation
41 among taxonomic levels over within taxonomic levels.
- 42 4. Our results show that reliable species-level identification from environmental samples is still
43 challenging, and thus neglecting identification uncertainty can lead to spurious inference. A
44 key aim for future research is the completion and pruning of taxonomic and reference
45 sequence databases, and making these two types of data compatible.

46 **Keywords**

47 DNA barcoding, DNA metabarcoding, molecular species identification, multinomial regression,
48 statistical model, taxonomic placement, taxonomic assignment

49

50

51

52 **Introduction**

53 In this paper, we use the term ‘DNA barcoding’ to refer to molecular species identification with
54 the help of ‘barcoding’ genes, which are short sequences of DNA that vary greatly between
55 species but little within species (Hebert et al. 2003). DNA barcoding has revolutionized biological
56 studies by increasing the speed and reliability of assigning Linnaean taxonomies to biological
57 specimens (Ratnasingham and Hebert 2007). When combined with high-throughput sequencing,
58 barcoding can be applied to bulk samples or environmental DNA, which approach we call here
59 ‘DNA metabarcoding’ (Taberlet et al. 2012; Yu et al. 2012).

60 In the metabarcoding pipeline, DNA is extracted from a bulk sample containing potentially
61 multiple species, a taxonomically informative gene is PCR-amplified, and the resulting PCR-
62 products are sequenced. The raw sequence output is processed through a bioinformatics pipeline
63 that includes denoising and removal of low quality and chimeric sequences, assignment of
64 sequences to their samples, and grouping similar sequences into ‘operational taxonomic units’
65 (OTUs). OTUs are meant to represent distinct biological taxa, usually distinct species. The term
66 OTU indicates that the clusters are not necessarily biological species but that they can be
67 considered as species hypotheses. This is because OTUs are typically defined phenetically using a
68 sequence-similarity threshold. Finally, in a crucial step, the researcher wishes to know the species
69 identities behind the OTUs, i.e. to place them into a Linnaean taxonomy.

70 Taxonomic placement of OTUs to high-level ranks (phylum, class, order) is relatively
71 straightforward (e.g. Yu et al. 2012), whereas placement to lower ranks (family, genus, species)
72 has remained more difficult. This is partly because of the limited information contained in the
73 short sequences generated by high-throughput sequencing platforms, and partly because of the
74 incomplete nature of reference databases, with missing taxa and limited within-taxon sampling
75 (Lou and Golding 2012). Furthermore, widely applied methods for low-level taxonomic placement
76 lack a proper assessment of identification reliability. For example, a user of the Barcode of Life
77 Database System (www.boldsystems.org, accessed 5 Aug 2016) encounters the warning “this
78 search only returns a list of the nearest matches and does not provide a probability of placement
79 to a taxon”. As we discuss in more detail below, the ability to conduct reliable low-level taxonomic
80 placement would make major contributions to species-level analyses, community-level analyses,
81 as well as metabarcoding methodology itself.

82 The value of assigning species names to barcoding sequences is that it allows one to link the
83 samples to the rest of our vast biological knowledge (Janzen et al. 2005). For instance, if
84 mammalian DNA isolated from a mosquito blood meal can be reliably assigned to red fox (*Vulpes*
85 *vulpes*), it enables one to combine the sample with many other kinds of information. These may
86 include information on the red fox's behaviour, population growth rate, age structure, geographic
87 distribution, habitat requirements, and trophic position, such as its top-down control of rodent
88 vectors of Lyme disease (Levi et al. 2012). More generally, accurate low-level taxonomic
89 placement of metabarcoding sequences improves many kinds of assessments of the structure and
90 function of communities, and how these change over space, time, or environmental gradients. For
91 example, species-level identifications of gut contents or faeces allows the construction of high-
92 resolution food webs (e.g. Wood et al. 2015). As another example, environmental change can be
93 inferred through species-level identification of ancient DNA, derived e.g. from lake sediments
94 (Pansu et al. 2015). As a further example, in food and medicine, DNA barcoding can be used to
95 improve food safety and wildlife forensics (Staats et al. 2016), e.g. through the detection of falsely
96 labelled products (Wong and Hanner 2008) and forbidden ingredients (Coghlan et al. 2012). As a
97 final example, metabarcoding can be used to monitor nature reserves and to detect endangered
98 species, e.g. rare rainforest mammals from the residual blood meals of leeches (Schnell et al.
99 2012).

100 The ability to conduct accurate low-level taxonomic placement would also contribute to the
101 metabarcoding methodology itself. Although OTUs are meant to represent single species,
102 biological species can unintentionally be split or merged during OTU clustering. Accurate species-
103 level taxonomic placement enables one to merge multiple OTUs that receive identical taxonomic
104 placements. Conversely, cases in which a single OTU is assignable with equal confidence to
105 multiple species can be used to identify taxonomic groups that would benefit the most from
106 better reference databases or where taxonomic revision may be needed. In addition, accurate
107 low-level taxonomic placement makes it easier to detect and remove contaminant OTUs.

108 As demonstrated by the above examples, for many kinds of purposes it is of critical importance to
109 know when we can and when we cannot reliably identify an OTU down to family, genus, or
110 species. Many kinds of bioinformatics programs are currently available for taxonomic placement.
111 These can be classified into three general categories: similarity-based, similarity/phylogeny-based,
112 and phylogenetic-placement-based. The most common are those that compare the similarities

113 between the environmental sequences and the sequences of the reference database: BLAST
114 (Altschul et al. 1997), MEGAN (Huson et al. 2007), BOLD (Ratnasingham and Hebert 2007), UTX
115 (Edgar 2013), NBC (Wang et al. 2007), and the Geneious Sequence Classifier (Kearse et al. 2012).
116 Similarity-based methods find the most phenetically similar reference sequence, and they do not
117 thus define taxonomic clades based on fundamental principles from systematic biology related to
118 synapomorphies (shared, derived characters). The second category of similarity/phylogeny-based
119 methods is represented by the Statistical Assignment Program (SAP), which first uses BLAST to
120 create a group of sequence homologues for an OTU (Munch et al. 2008). Multiple phylogenetic
121 trees are then generated for the OTU and its homologues, and taxonomic placement is guided by
122 the summarised position of the OTU within the trees. The third category of phylogenetic-
123 placement-based methods includes *pplacer* (Matsen et al. 2010) and the Evolutionary Placement
124 Algorithm (EPA) (Berger et al. 2011). These methods first construct a single maximum-likelihood
125 phylogeny from all available reference sequences, after which they place the OTUs within the
126 phylogenetic tree.

127 A major challenge affecting all taxonomic placement methods is that reference databases are
128 incomplete, and that they may contain mislabelled reference sequences. This is especially
129 problematic when trying to identify a sequence within a large taxonomic clade in regions of high
130 biodiversity where many organisms have yet to be sequenced. Ideally, uncertainty due to
131 incomplete or mislabelled reference sequences should result in taxonomic placement to higher
132 taxonomic ranks, not to the most similar reference sequence that happens to be available. Thus
133 far, only heuristic solutions to this problem have been proposed. For example, in MEGAN, a
134 lowest-common-ancestor (LCA) assignment algorithm uses several best BLAST hits to determine
135 the taxonomic level into which the assignment is given, but incomplete reference databases may
136 still lead to false annotations.

137 In our previous work, we developed the bioinformatics pipeline PROTAX (PRObabilistic TAXonomic
138 placement, Somervuo et al. 2016) which accounts explicitly for incompleteness of taxonomic and
139 reference databases. This is achieved by placing environmental sequences into a Linnaean
140 taxonomy that is typically only partly populated by reference sequences. The taxonomic
141 placements generated by PROTAX include known taxonomic units (species present in the Linnaean
142 taxonomy) for which reference sequences are available, known taxonomic units for which
143 reference sequences are not available, and unknown taxonomic units, such as species or genera

144 that are missing from the Linnaean taxonomy. A key feature of PROTAX is that it is probabilistic,
145 i.e. it decomposes the probability of one among all possible assignment outcomes. In the ideal
146 case, one of the outcomes obtains a high probability whereas the other taxonomic placements
147 obtain probabilities close to zero. In ambiguous cases, several outcomes obtain a non-negligible
148 classification probability, and thus reliable taxonomic placement can be achieved only at a higher
149 taxonomic rank. PROTAX is based on a statistically rigorous model, making its classification
150 probabilities unbiased, as shown in Somervuo et al. (2016) for simulated data and a small-scale
151 empirical case study. In other words, if PROTAX assigns an 80% probability for placement to a
152 given taxonomic unit for 100 sequences, the classification will be on average correct for 80 of
153 those sequences, whereas it will not be correct for 20 of the sequences.

154 This paper has two aims. The first aim is to evaluate the potential of DNA (meta)barcoding for
155 obtaining species-level identifications, given the current state of taxonomic databases, sequence
156 reference databases, and sequencing technologies. The second aim is to evaluate the
157 performance of PROTAX as a general tool for taxonomic placement. To address both aims, we
158 apply PROTAX to four contrasting case studies, which differ greatly in their taxonomic scope
159 (fungi, insects, mammals, and bacteria), the number of species involved, the coverage and quality
160 of the reference databases, and the sequencing technology applied to environmental data. For
161 each case study, we conduct two kinds of analyses. First, we examine how well PROTAX is able to
162 classify validation sequences sampled from the reference database. Second, we apply PROTAX to
163 environmental sequence data to examine the level of species identification resolution that can be
164 expected to be achieved by different kinds of empirical studies.

165 **Materials and methods**

166 We consider four case studies, for each of which we use three kinds of data: a taxonomy database,
167 a reference sequence database, and environmental sequences originating from an empirical study
168 (Table 1). The case studies vary greatly in many aspects: their taxonomic scopes (mammals, fungi,
169 insects and bacteria), the sizes and coverages of the taxonomies and the reference databases, the
170 barcoding gene used, and the sequencing technology applied. These influence e.g. the level of
171 overlap among genetic variation between consecutive taxonomic levels (Fig. 1), with obvious
172 implications to the possibility of species-level taxonomic placement. As the four case studies vary
173 simultaneously in many aspects, their comparison does not enable asking e.g. whether it is

174 generally easier to identify insects or fungi. Instead, they are selected to be diverse in order to
175 illustrate the many kinds of issues that influence the accuracy of taxonomic placement.

176 For each case study, we first utilized the taxonomy and reference sequence databases to
177 parameterize the PROTAX statistical model. To do so, we followed Somervuo et al. (2016), except
178 for small modifications that we describe below. We then used the parameterized model to classify
179 a set of well-identified reference sequences, with the aim of evaluating the classification accuracy
180 of PROTAX at different taxonomic levels, and to assess if the classification probabilities are
181 unbiased. Finally, we clustered the environmental data to OTUs, roughly at the species level,
182 picked the most common sequence to represent each cluster, and used the parameterized
183 PROTAX model for probabilistic taxonomic placement of these OTUs. The aim here was to assess
184 how large a fraction of environmental data can be reliably classified to each taxonomic level, and
185 to examine which fraction of environmental sequence data represents the two unknown
186 categories included in PROTAX: species that are present in the taxonomy but for which reference
187 sequences are available, and species that are missing from the taxonomy.

188 We first describe the three data types (taxonomy database, reference database, and
189 environmental data) that we acquired for each case study, as well as make some remarks about
190 the particularities of each case study. We then explain how PROTAX was fit to these data and how
191 we assessed PROTAX's performance in probabilistic taxonomic placement.

192 ***Identifying mammals from leech blood meals***

193 *Taxonomy database.* We used the NCBI taxonomy (NCBI Resource Coordinators 2016) of all clades
194 within Mammalia. This database has high coverage as it includes all 6674 species for which
195 molecular data are available. The taxonomy is classified to the four levels of order, family, genus
196 and species. For some species, classifications to intermediate levels or species-level were missing.

197 *Reference sequence database.* We used all available mammalian mitochondrial 16S rRNA gene
198 sequences (mt 16S rRNA) downloaded from GenBank (Clark et al. 2016). We removed ambiguous
199 bases and kept only sequences of length 300-1600 bp. We included at most 10 sequences of per
200 species, resulting in a database of 2627 sequences representing 1315 different species.

201 *Environmental sequences.* We used mammalian mt 16S rRNA gene sequences (see Schnell et al.
202 2012 for further details on primer) derived from residual blood meals of ~20,000 haematophagous

203 leeches collected in the central Annamite mountains of Vietnam and Laos (Yu et al., unpublished
204 data). DNA extraction was conducted using the Qiagen QIAquick PCR purification kit, and
205 sequencing by Illumina HiSeq 2000. Raw reads were denoised with *bfc* (Li 2015), chimeras
206 removed by UCHIME (Edgar et al. 2011), and the sequences assigned to samples using the QIIME
207 (Caporaso et al. 2010) script `split_libraries.py`. The reads were clustered into OTUs at 98%
208 similarity using CROP (Hao et al. 2011). OTUs that were not identified as vertebrate mt 16S rRNA
209 based on BLAST against GenBank were removed.

210 *Remarks.* As we use here individual, very short single-read (typically 100 bp) sequences provided
211 by Illumina HiSeq, we aim to demonstrate how PROTAX performs in the case of high identification
212 uncertainty, rather than attempting to identify the specimens as well as would be possible e.g. by
213 including an assembly step. To illustrate the effect of sequence length, we parameterized the
214 model both for full length and short length sequences.

215 ***Identifying insects from individually sequenced specimens***

216 *Taxonomy database.* We compiled a list of all species of the class Insecta (excluding Psocodea)
217 recorded in Greenland, based on Böcher et al. 2015, with additions from Wirta et al. (2016). The
218 environmental sequence data (see below) comes from the same study region as that of Wirta et
219 al. (2016), and thus the taxonomic database is expected to cover the species of the region
220 relatively well. The 1332 taxa were classified to the four levels of order, family, genus and species.
221 Most of the taxa were defined to species, but a fraction as a sole representative of a genus.

222 *Reference sequence database.* We used barcode sequences of specimens collected from
223 Zackenberg, Greenland. The reference database included the standard cytochrome *c* oxidase
224 1 (CO1) barcode sequence for 241 morphologically identified insect species (deposited in BOLD
225 under dataset dx.doi.org/10.5883/DS-ZACKANIM).

226 *Environmental sequences.* We used 7939 CO1 sequences from insect tissue caught on sticky traps
227 mimicking a flower in northeast Greenland. Each sequence (deposited in BOLD with the code
228 ZACKD) represents a separate specimen (Tiusanen et al., unpubl. data) that was Sanger sequenced
229 in one direction.

230 *Remarks.* As here both the taxonomy database as well as the reference database are specifically
231 tailored to the environmental data, and as here the environmental sequence data consist of high
232 quality sequences, this case study is aimed to illustrate a best case scenario.

233 ***Identifying wood-inhabiting fungi from saw dust samples***

234 *Taxonomy database.* We used the Index Fungorum database (www.indexfungorum.org), classified
235 into the six levels of phylum, class, order, family, genus, and species. We reduced the amount of
236 redundancy in the taxonomy by removing likely synonyms, such as old names of species that had
237 been renamed. The resulting taxonomy consists of 130 795 species.

238 *Reference sequence database.* To construct the reference database of 75 104 sequences, we used
239 the UNITE+INSD sequence database (<https://unite.ut.ee/>) consisting of fungal ITS region,
240 complemented with the database of Ovaskainen et al. (2013). In order to increase the coverage of
241 the reference sequences for poorly studied species groups, we also included those species
242 hypothesis (SH) from UNITE that were more than 97% divergent from the other reference
243 sequences. We extracted the ITS2 region of the reference sequences using ITSx software
244 (Bengtsson-Palme et al. 2013). The majority (73%) of the reference sequences were annotated to
245 the species level, but many only to the genus (11%) level or family or higher levels (16%). We
246 included at most five sequences per species.

247 *Environmental sequences.* We used fungal ITS2 sequences originating from the study of
248 Ovaskainen et al. (2013). The saw dust samples originate from 100 spruce logs sampled in autumn
249 2008 in a natural forest in southern Finland. DNA extraction was conducted using the Power Soil
250 DNA isolation kit (MoBio Laboratories, Inc., Carlsbad, CA, USA), and sequencing was done on a
251 Genome Sequencer FLX (454 Life Sciences, Roche, Branford, CT, USA). We removed all sequences
252 that were shorter than 150 bp, resulting in 259 327 sequences. We used cutadapt (Martin 2011)
253 to detect the presence of ITS4 primer in order to be sure that the sequence represented ITS2
254 region. To cope with homopolymer errors, all consecutive repetitions of the same nucleotide were
255 removed as in Ovaskainen et al. (2010, 2013), both for reference and environmental sequences.
256 Environmental sequences were clustered using UCLUST (Edgar 2010) with 99% identity threshold.

257 *Remarks.* This case study is aimed to illustrate how PROTAX copes with a very large taxonomy that
258 is only poorly covered by reference sequences. We further use the fungal case study to examine
259 how additional information can be incorporated into the PROTAX model: in addition to the
260 baseline model, we constructed an alternative model, where we gave more weight to species that
261 are expected to be found from the geographic area where the sampling was conducted (for more
262 details, see below).

263 ***Identifying bacteria from a food production pipeline***

264 *Taxonomy database.* The taxonomy used for bacteria is different from other taxonomies in the
265 sense that it is not an independent Linnaean taxonomy but it was generated from the Ribosomal
266 Database Project (RDP) reference sequences (Wang et al. 2007) and therefore fully coincides with
267 the reference sequence database (see below). The RDP reference taxonomy contains both
268 bacteria and archaea, and it is well curated down to the genus level. Here we included the six
269 levels of domain, phylum, class, order, family and genus. The taxonomy consists of 60 phyla
270 classified to 2175 genera.

271 *Reference sequence database.* For the reference sequence database, we used the RDPClassifier
272 training sequences, labeled to the genus level (trainset15_092015.fa from
273 RDPClassifier_16S_trainsetNo15_rawtrainingdata.zip, available at
274 rdp.cme.msu.edu/misc/resources.jsp).

275 *Environmental sequences.* We used bacterial 16S rRNA gene sequences from the study of Hultman
276 et al. (2015), who aimed to understand the effect of food-preparation-surface microbiomes on the
277 end product. The samples originate from surfaces of a food processing facility, from raw food
278 material, and from cooked food products. As detailed in Hultman et al. (2015), total DNA was
279 extracted from the samples using a bead beating method. The V1 to V3 region was PCR-amplified
280 and sequenced with 454 GS FLX. The reads were quality filtered, chimeras were removed, and
281 reads were assigned to OTUs with QIIME (Caporaso et al. 2010) using 97% similarity.
282 Homopolymers were treated as in the fungal data. The raw sequence reads can be downloaded
283 from Sequence Read Archive (SRA) of the NCBI under BioProject number PRJNA293141.

284 *Remarks.* As noted above, the bacterial case study differs fundamentally from the other case
285 studies as the taxonomy database is not independent of the reference sequence database.
286 Compared especially to mammals and Greenland insects, the taxonomy is likely to be incomplete.
287 Thus with this case study we were interested in examining whether the environmental sample
288 includes a high fraction of material that PROTAX would classify to belong to missing branches.

289 ***Fitting the PROTAX model***

290 PROTAX converts sequence similarities into probabilities of taxonomic classification in a
291 hierarchical manner, starting from the root node of the taxonomy and proceeding towards the
292 species nodes. Each node divides its probability into its child nodes by means of a multinomial

293 regression model. The predictors used in the multinomial regression can be chosen in many ways.
294 While the results of Somervuo et al. (2016) suggest that a combination of similarity-based and
295 phylogenetic-based predictors yields the best performance both for simulated and real data, in
296 this study we used solely similarity-based predictors.

297 The regression model for each taxonomic node containing seven predictors $!_1, \dots, !_7$. The baseline
298 case where all the seven predictors are zero corresponds to a child node that represents a missing
299 branch of the taxonomy. Predictor $!_1$ is an indicator variable for a known child node that contains
300 no reference sequences, whereas predictor $!_2$ is an indicator variable for a known child node that
301 contains at least one reference sequence. Predictors $!_3$ and $!_4$ are, respectively, the mean and the
302 maximum value of pairwise sequence similarities between the query sequence and the reference
303 sequences. To allow PROTAX to account in the predictions for the availability of the number of
304 reference sequences (with which e.g. maximal similarity is expected to increase just by chance),
305 we included as predictors also the log-transformed number of reference sequences representing
306 the child node ($!_5$), and the interactions between log-transformed number of reference sequences
307 and mean ($!_3$) and maximal ($!_4$) similarities.

308 We calculated pairwise sequence similarities using LAST (Kielbasa et al. 2011) with the following
309 deviations from the default parameters. We set the LAST argument `-T 1` to make the similarity
310 score represent the entire overlap alignment length between two sequences, excluding only the
311 possible overhangs. We set the gap open penalty to `(-a 1)`. In order to get meaningful values to the
312 mean sequence similarity predictor of the PROTAX model, we set the maximum number of initial
313 matches per query position (`-m`) values between 1000 and 3000 instead of the default value 10.
314 We replaced pairwise sequence similarities that were missing from LAST output by zeros, and
315 converted sequence similarities to the range $[0,1]$ by dividing the alignment score by the
316 alignment length.

317 We generated training data to parameterize the PROTAX model as described in Somervuo et al.
318 (2016), i.e. by modifying both the taxonomic tree itself as well as its coverage by the reference
319 sequences to mimic the different kinds of outcomes: (i) known species with reference sequences,
320 (ii) known species without reference sequences, and (iii) unknown species or unknown higher
321 taxonomic branches. For each case study, we generated in total 1000 training data points, out of
322 which 100 represented the category (iii), with an even distribution over the taxonomic levels. The
323 remaining 900 sequences representing categories (i) and (ii) were generated by randomly

324 selecting one of the species present in the database, and generating training data directly for that
325 species, or if not possible, for another species that was taxonomically as close to the selected
326 species as possible. For example, if the selected species had no reference sequences, we selected
327 the closest species that had at least one such sequence, selected one sequence to represent the
328 query sequence, and removed all the other sequences to mimic a species with no reference
329 sequences.

330 In our baseline analyses, we assumed a priori that all species that are part of the Linnaean
331 taxonomy are equally likely to be present in the empirical sample. If there is prior information
332 about which species are more likely to be found from an empirical sample than others, such
333 information can be incorporated into PROTAX by a weighting scheme, which can be considered as
334 an informative prior in the context of Bayesian analyses. To illustrate the influence of the prior, we
335 conducted an alternative version of the fungal analyses, where we gave more prior weight for
336 those species that are known to occur in Finland, as our environmental samples originate from
337 there. From the list of Finnish 6645 fungal species, we could map 4718 names to the 130 795
338 species taxonomy. In the weighted analysis, we assumed a priori that each sequence present in
339 our environmental sample represents one of the species known to occur in Finland with
340 probability 90%, and thus dividing the remaining probability of 10% among the remaining species.

341 We derived maximum *a posteriori* (MAP) parameter estimates for the PROTAX models using the
342 Bayesian approach presented in Somervuo et al. (2016), except that in the present study we
343 parameterized the models separately for each taxonomic level. The model parameters for each
344 level include the seven regression coefficients corresponding to each of the predictors, as well as
345 the probability by which the reference sequence is mislabeled (Somervuo et al. 2016).

346 ***Evaluating the performance of PROTAX***

347 We used the parameterized PROTAX models to perform taxonomic placements of both reference
348 sequences as well as environmental sequences. In the first set of analyses, we performed
349 taxonomic placements for 1000 validation sequences, which were chosen from the reference
350 sequence database in the same way as the training sequences described above. While PROTAX
351 yields for each of these the full probability distribution over possible outcomes, we selected here
352 only the outcome with the highest probability. We considered a taxonomic placement as
353 “plausible” if the classification probability was at least 50%, and as “reliable” if the classification

354 probability was at least 90%. To examine the overall confidence of classifications, we computed
355 the proportions of plausible and reliable classifications at each taxonomic level. To assess if the
356 probabilities of taxonomic placement were unbiased, we ordered the classification probabilities
357 from lowest to highest, and computed a cumulative sum of both these probabilities as well as the
358 indicator variables describing whether the outcome predicted with highest probability was a
359 correct one. We then plotted these two cumulative sums against each other. If the classification
360 probabilities are unbiased, such a plot should follow the identity line.

361 In the second set of analyses, we performed taxonomic placements for the environmental
362 sequence data. As the mammalian, fungal and bacterial case studied involved a large number of
363 sequences generated by high-throughput methods, we first clustered these sequences. The data
364 submitted to PROTAX involved 1514 (mammal), 4163 (fungi), and 6855 (bacteria) OTUs, and 7939
365 individual insect sequences.

366 To visualize community composition within each environmental data set, we used Krona (Ondov
367 et al. 2011) to plot for each case study a pie chart that shows the expected number of sequences
368 representing each taxonomic unit. To compute the expected abundances, we did not account only
369 for the highest probabilities, but we summed over the entire distribution of predicted probabilities
370 (ignoring values lower than 0.01 for computational reasons). To visualize the quality of the
371 classifications, we colored the charts to show six categories. The first three categories consisted of
372 well-identified taxonomic units for which the proportion of sequences for which the classification
373 was reliable was (1) in the range 50%-100%, (2) in the range 0%-50%, or (3) 0%. The remaining
374 three categories consisted of non-identified taxonomic units for which the proportion of
375 sequences for which the classification was reliable was (4) in the range 50%-100%, (5) in the range
376 0%-50%, or (6) 0%. Above, well-identified taxonomic unit refers to a single taxonomic unit for
377 which reference sequences were available, whereas non-identified taxonomic units refers to the
378 union of taxonomic units without reference sequences and unknown branches of the taxonomy.

379 **Results**

380 As expected based on our earlier results (Somervuo et al. 2016) and the fact that PROTAX is a
381 statistical model fitted to training data, PROTAX yielded essentially unbiased probabilities of
382 taxonomic placement for all the cases considered. This is evidenced by the fact that all lines in Fig.
383 2 generally follow the identity lines, the small deviations being attributable either to sampling

384 error due to finite sizes of the validation data sets, or to issues related to model misspecification,
385 the latter of which we return to in the Discussion. The probabilities shown in Fig. 2 are level-
386 specific, thus asking e.g. how well genera can be separated within a known family, or how well
387 species can be separated within a known genus. For high taxonomic levels, these probabilities are
388 lowest for fungi, which is consistent with the fact that for fungi there is the greatest amount of
389 overlap in sequence similarities among consecutive taxonomic levels (Fig. 1). For example, if
390 within-species similarities are sometimes lower than among-species similarities, accurate
391 taxonomic placement to the species-level is not always possible.

392 When performing a taxonomical placement of environmental samples, PROTAX works in a
393 hierarchical manner starting from the root of the tree, and proceeding level by level towards the
394 tips of the tree that represent typically species. The probabilities of taxonomic placement for a
395 given level (illustrated in Figs. 3 and 4) are thus obtained by multiplying the level-specific
396 conditional probabilities (illustrated in Fig. 2) for all levels lower than or equal to the focal level.
397 Figure 3 shows the proportions of the reference sequences (black lines) and environmental
398 sequences (gray lines) that were possible to identify reliably (dashed lines) or plausibly
399 (continuous lines). Let us first make two obvious remarks. First, as the threshold for plausible
400 identification (>50% probability of taxonomic placement) is lower than that of reliable
401 identification (>90% probability of taxonomic placement), the proportion of plausible
402 identifications is always higher than that of reliable identifications. Second, as the lower level
403 taxonomic placements are conditional on the higher level ones, the fraction of reliable (and
404 plausible) identifications decreases monotonously with taxonomic level.

405 Beyond the above made trivial remarks, Fig. 3 shows a number of interesting results. As the first
406 result, that we derive from the taxonomic placement of the validation sequences, reliable species-
407 level identification (dashed black lines in Fig. 3) was most successful for insects (74% of the
408 sequences), followed by mammals (46%) and fungi (15%). These numbers do not reflect only the
409 resolution of the barcoding sequences (Fig. 1), but also the fact that the insect taxonomy and
410 reference sequence databases were restricted to species occurring in Greenland, whereas the
411 mammalian and fungal databases were global and thus were larger and more heterogeneous
412 (Table 1). For mammals, full-length mt 16S sequences (black crosses in Fig. 3C) can be expectedly
413 classified with much higher confidence than fragmented sequences (black dots in Fig. 3C), the

414 latter corresponding to the nature of the environmental data. In case of bacteria, reliable genus-
415 level identification was possible for the majority (62%) of the cases.

416 As the second result, Fig. 3 shows that taxonomic placement of environmental sequences is often
417 less reliable than that of reference sequences (mammals and fungi), but sometimes environmental
418 sequences can be identified essentially equally reliably (insects) or even more reliably (bacteria)
419 than reference sequences. The main reason why taxonomic placement of environmental
420 sequences for mammals was much more difficult than that of reference sequences is simply that
421 in our case study the environmental sequences were very short fragments. If fragmenting the
422 reference sequences equally much (into 100 bp segments), their taxonomic placement became
423 essentially equally unreliable than that of reference sequences (lines with black dots in Fig. 3C). In
424 case of fungi (Fig. 3A), the reason for the difference between the taxonomic placement of the
425 reference and environmental sequences was not only a similar (though less pronounced)
426 difference in sequence length and quality as for mammals, but also the fact that the
427 environmental sequences are likely to represent many unknown units that are lacking from the
428 taxonomy. If bringing the prior information that, instead of any globally known fungi, the species
429 within the environmental sample are likely to represent species that are known to occur in
430 Finland, the proportion of reliable identifications increases dramatically from 3% to 14% (Fig. 3C).
431 The reason why for the insect data (Fig. 3D) the taxonomic placements are essentially equally
432 reliable for the reference and environmental sequences is that for this case study both kinds of
433 sequences were acquired by identical methods, i.e. Sanger sequencing of DNA sampled from
434 individual specimens. Thus, the only differences between the two were whether the specimens
435 were identified morphologically or not, and whether the specimens represent a random sample of
436 the community (environmental sequences) or whether they were targeted to represent the entire
437 community (reference sequence data). The most curious case is that of bacteria, where reliable
438 genus level taxonomic placements were more frequent for environmental sequences than for
439 reference sequences (Fig. 3B). The likely reason here is that in this case the environmental
440 sequences originated from the food production pipeline, the bacterial communities of which
441 represent one of the most well studied groups, and thus are better covered in the reference
442 sequence database than bacteria in general.

443 Let us then turn into the main question that motivates DNA (meta)barcoding studies: what are the
444 species behind the environmental samples? The answer to this question is given in Fig. 4, where

445 the pie charts show the proportions of sequences that belong to known and unknown taxonomical
446 unit at each hierarchical level. In this figure, the areas of the sectors show the expected number of
447 sequences that belong to each taxonomic unit, whereas the colors illustrate the proportions of
448 reliable identifications, and they thus echo the information shown by the grey dashed lines in Fig.
449 3. While our main interest here is not on the detailed results relating to the four case studies, let
450 us note that the overall patterns in Fig. 4 are consistent with expectations. Concerning fungi, the
451 majority of the species Agaricomycetes, and the reliably identified species (e.g. *Antrodia serialis*;
452 see the insert in Fig. 4) typically represent well known wood decomposers. Concerning mammals,
453 both Artiodactyla, Chiroptera, Rodentia and Carnivora were detected, as well as some primates.
454 While there are very few reliable or even plausible species-level taxonomical placements, among
455 possibly identified species are e.g. the endangered mammals *Muntiacus vuquangensis* (Giant
456 Muntjac; 43% identification probability) and *Rusa unicolor* (sambar; 27% identification
457 probability). Concerning bacteria, a large proportion of the sequences were assigned as
458 Lactobacillales, specifically to Streptococcaceae, Lactobacillaceae, and Leuconostocaceae (Figure
459 4). Further, the high proportion of *Brochotrhrrix* observed by Hultman (2015) was supported by the
460 PROTAX results. Concerning insects, the majority of the species belonged to Diptera and the
461 minority to Hymenoptera. Among the total of 104 distinct species that were reliably identified, the
462 most common one was *Drymeia segnis*, which has been observed to be common in the study area
463 also based on morphological identifications (Rasmussen et al. 2013).

464 In Supporting Information, we provide the same information as shown in Fig. 4 as interactive
465 HTML files, which allow the pie charts to be displayed using a standard web browser without any
466 additional plugins. This allows one to examine the taxonomic placements and their reliabilities in
467 much greater detail by e.g. using search tools and zooming to taxonomic clades of specific
468 interest.

469 **Discussion**

470 In this work, we have evaluated the potential of DNA barcoding for obtaining reliable taxonomic
471 placements at different taxonomic levels, and in particular illustrated how the PROTAX method
472 can be used as a general tool for quantifying uncertainty in such taxonomic placements. PROTAX
473 accounts for many kinds of uncertainties, including the possibilities of unknown taxonomic
474 branches, incomplete coverage of reference sequence databases, and mislabelling of reference
475 sequences. This makes its quantification of taxonomic placement uncertainty robust, as illustrated

476 by Fig. 2 and the simulations by Somervuo et al. (2016). However, it is important to understand
477 that the classification accuracy does not necessarily increase when taking all uncertainties into
478 account; it can rather be the opposite. To put it bluntly, it may be more tempting e.g. to claim that
479 the study detected the endangered mammal Giant Muntjac from a leech blood meal, rather than
480 to specify that this was the case with 43% probability, as the latter statement makes it explicit that
481 the species behind the sequence may actually have been some other one. However, making
482 uncertainty explicit is necessary for scientific reliability.

483 There are many choices to be done when applying DNA (meta)barcoding to an empirical case
484 study. As illustrated by our results, these choices can have a major influence on the reliability of
485 the resulting taxonomic placements. The first set of choices relates to the taxonomy and reference
486 databases used, which choices are in practice mostly guided by on what databases are available
487 rather than what might be optimal to use. Importantly, as PROTAX accounts for missing branches
488 in the taxonomy, the incompleteness of the taxonomy database should not lead to spurious false
489 positives, rather to decreased probabilities of taxonomic placement. This is because in the training
490 phase PROTAX generates situations in which some branches of the taxonomy are missing, making
491 it learn which kinds of values of the predictors (e.g. low values of sequence similarity) are
492 indicative of missing branches. Similarly, mislabeled reference sequences or inconsistencies
493 between the taxonomy and the reference databases are expected to decrease the probabilities of
494 taxonomic placement, but not to bias them. As one example, we used the RDP database for
495 bacteria. Since the reference taxonomy was constructed based on the reference sequences, 100%
496 of the taxa in the validation data were covered (Table 1). Somewhat surprisingly, the bacterial
497 reference database appeared to represent also the vast majority of the environmental sequences,
498 with only very few missing branch identified (Fig. 4). This however does not mean that the used
499 taxonomy would cover all the bacteria in the world, and novel phyla have indeed been discovered
500 in several recent metagenomic studies (e.g. Brown et al. 2015). The other commonly used
501 bacterial and archaeal databases are SILVA (Quast et al. 2013) and Greengenes (DeSantis et al.
502 2007). Compared to RDP, these two databases contain more representatives of the Candidate
503 divisions that have been recently found in various soil environments (Brown et al. 2015; Hug et al.
504 2016). Therefore, depending on the environment under analysis, the use of different reference
505 databases should be considered.

506 The second set of choices to be made relates to the DNA barcode applied, as well as the
507 sequencing technology. As has been long pointed out, an optimal barcoding gene should involve
508 much variation among species but only little within a species (Meyer and Paulay 2005). Further,
509 the environmental sequences should obviously have as long read length and as high quality as
510 possible. For example, if in the mammalian case study full length mt 16S rRNA sequences had
511 been available instead of the very short 100 bp fragments used here, the proportion of reliable
512 taxonomic placement would have been likely to increase from the present 0% to ca. 46%, where
513 the latter was the proportion of reference sequences that we could classify reliably. But even if
514 one would have full length sequences and complete taxonomic and reference sequence
515 information, some uncertainty will inevitably remain. For example, in the insect study the
516 mosquito species *Aedes impiger* and *Aedes nigripes* could not be disentangled since their COI
517 sequences are identical, and thus PROTAX assigned for some of the specimens a probability close
518 to 50% for both of these species. To resolve such cases, a deeper genomic approach (Bourke et al.
519 2013) than the single gene DNA barcoding approach should be used.

520 The third set of choices relates to the way in which the training data in PROTAX are generated,
521 technically the prior assumed for the empirical data. This is probably the most critical and at the
522 same time most difficult choice to be done by the user, as making a justified choice requires
523 biological knowledge and intuition. For example, one may assume either that each sequence in
524 the environmental sample represents any of the species present in the taxonomy with equal
525 probability (as we have done here), or utilize a hierarchical prior that assumes that each branch
526 under a given node is equally likely (as we did in Somervuo et al 2016). One may further give
527 additional weight for species that are known to occur in the geographic region where the samples
528 originated, as we did for the fungal case study. If such information is available, the prior can also
529 be adjusted e.g. based on the expected abundances of the species, or on the match between the
530 substrates sampled and the habitat requirements of the species. In addition to the known species,
531 the prior involves an assumption about the frequency of missing branches at different parts of the
532 taxonomic tree. As it may be difficult to make informative choices about all of the above
533 mentioned aspects, we recommend the user to test the sensitivity of the results against different
534 choices of the prior, as should be done with Bayesian analyses in general.

535 Finally, the fourth set of choices relates to the predictors used for the multinomial regression
536 underlying the PROTAX model. In this paper, we have used simply similarity-based predictors,

537 even if our previous work suggests that similarity-based predictors and phylogeny-based
538 predictors involve complementary information and thus their combination optimizes performance
539 (Somervuo et al. 2016). The reason behind the choice made for the present work was mainly
540 computational, as some of our databases were extensive, making LAST-based similarity the most
541 practical choice. For fungi, the use of phylogeny-based predictors is challenging also for the reason
542 that the construction of multiple sequence alignments is difficult with the ITS region only.
543 Phylogeny-based methods are easier/more suitable with conserved barcodes such as CO1 and
544 mt16S which allow sequences to be globally aligned even at high taxonomic levels. In more refined
545 studies focusing on any specific case study, the set of predictors should be optimized to maximize
546 the reliability of taxonomic placements. While there is no objective way to select the best prior,
547 the choice of the predictors can be optimized more or less objectively by examining which
548 predictors maximize unbiased probabilities of taxonomic placement for independent validation
549 sequences. The reason why for some choices of the predictors the classification probabilities can
550 be biased (as was to a limited extent a case for some of our case studies, Fig. 2) is that while the
551 PROTAX model is parameterized by training data, the model may be structurally misspecified. For
552 example, we have assumed that the model parameters are constant across the taxonomic tree.
553 Thus, when classifying an environmental sequence e.g. to the species level under a known genus,
554 the parameters (and thus the influences of the predictors, such as sequence similarity) are
555 assumed to be independent of the genus. This assumption is not likely to hold for large and
556 heterogeneous taxonomic groups, such as all mammals or all fungi. An indication of this in our
557 results was that, at the species level, the parameter estimates obtained for mislabeling probability
558 were much inflated, being ca. 80% for mammals and ca. 60% for fungi. This does not suggest that
559 there is such a vast amount of mislabeling, but that PROTAX used the mislabeling parameter to
560 correct for model misspecification. Thus, an important challenge for future work is to further
561 develop the statistical model underlining PROTAX, either by building a hierarchical structure that
562 allows for heterogeneity in the parameterization, or by finding predictors that are able to correct
563 for such heterogeneity.

564 To conclude, molecular species identification by DNA barcoding and metabarcoding is an exciting
565 and rapidly evolving research field, which has major potential to change our understanding of the
566 structure and functioning of ecological communities. To make the use of these methods practical
567 and reliable, a key challenge is the completion and pruning of taxonomic and reference sequence

568 databases, as well as making these two sources of information compatible. Similarly important is
569 the application and further development of statistical methods that allow one to make the most
570 out of such data by providing accurate taxonomic placements and reliable assessments of the
571 uncertainties inherent in such placements. Such methods are critical for providing a firm basis for
572 deriving species- and community-level inferences from DNA (meta)barcoding data, especially for
573 environmental DNA that by definition do not have physical specimens that could be verified
574 independently. Incorrect assignments can result in accumulated interpretation error, which can
575 result in wasted resources and social conflict in multiple social arenas, from conservation to food
576 safety. It is important to get the name right – or to be aware that it may be wrong.

577 **Acknowledgements**

578 We thank Mikko Tiusanen and Tomas Roslin for providing the empirical sequence data for the
579 insect case study, and Paul Kirk and Kessy Abarenkov for providing the taxonomy and reference
580 databases for the fungal case study, and Otto Miettinen for providing the list of Finnish fungal
581 species. OO and PS were supported by funding from the Academy of Finland (Grants no. 1273253
582 and 250444 to OO) and the Research Council of Norway (CoE grant no. 223257). DWY and YQJ
583 were supported by WWF-Vietnam (9S084701), the National Natural Science Foundation of China
584 (31400470, 41661144002), the Ministry of Science and Technology of China (2012FY110800), the
585 University of East Anglia and its GRACE computing cluster, and the State Key Laboratory of Genetic
586 Resources and Evolution at the Kunming Institute of Zoology (GREKF13-13, GREKF14-13, and
587 GREKF16-09). CX was supported by the MEME Erasmus Mundus Programme in Evolutionary
588 Biology, and the Marco Polo Exchange Fund of the University of Groningen.

589

590

591 **References**

592 Altschul, S., Madden, T., Schäffer, A., Zhang, J., Zhang, Z., & al. (1997) Gapped BLAST and PSI-
593 BLAST: a new generation of protein database search programs. *Nucleic Acids Research* 25, 3389-
594 3402.

595 Böcher, J., Kristensen, N., Pape, T. & Vilhelmsen, L. (eds) (2015) *The Greenland Entomofauna: an*
596 *identification manual of insects, spiders and their allies*. Brill, *Fauna Entomologica Scandinavica*,
597 vol 44.

598 Bengtsson-Palme, J. Ryberg, M., Hartmann, M., Branco, S., Wang, Z. & al. (2013) Improved
599 software detection and extraction of ITS1 and ITS2 from ribosomal ITS sequences of fungi and
600 other eukaryotes for analysis of environmental sequencing data. *Methods in Ecology and*
601 *Evolution* 4, 914–919.

602 Berger, S., Krompass, D. & Stamatakis, A. (2011) Performance, accuracy, and Web server for
603 evolutionary placement of short sequence reads under maximum likelihood. *Systematic Biology*
604 60, 291-302.

605 Bourke, B., Oliveira, T., Suesdek, L., Bergo, E. & Sallum, M. (2013) A multi-locus approach to
606 barcoding in the *Anopheles strodei* subgroup (Diptera: Culicidae). *Parasites & Vectors* 6, 111.

607 Brown, C., Hug, L., Thomas, B., Sharon, I., Castelle, C. & al. (2015) Unusual biology across a group
608 comprising more than 15% of domain Bacteria. *Nature* 523, 208–211.

609 Caporaso, J., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. & al. (2010) QIIME allows
610 analysis of high-throughput community sequencing data. *Nature Methods* 7, 335-336.

611 Clark, K., Karsch-Mizrachi, I., Lipman, D. Ostell, J. & Sayers, E. (2016) GenBank. *Nucleic Acids*
612 *Research* 44, D67–D72.

613 Coghlan, M., Haile, J., Houston, J., Murray, D., White, N. & al. (2012) Deep Sequencing of Plant and
614 Animal DNA Contained within Traditional Chinese Medicines Reveals Legality Issues and Health
615 Safety Concerns. *PLoS Genetics* 8, e1002657.

616 DeSantis, T., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E. & al. (2006) Greengenes, a chimera-
617 checked 16S rRNA gene database and workbench compatible with ARB. *Applied and*
618 *Environmental Microbiology* 72, 5069–5072.

619 Edgar, R. (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26,
620 2460-2461.

621 Edgar, R. (2013) UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nature*
622 *Methods* 10, 996-998.

623 Edgar, R., Haas, B., Clemente, J., Quince, C. & Knight, R. (2011) UCHIME improves sensitivity and
624 speed of chimera detection. *Bioinformatics* 27, 2194-2200.

625 Hao, X., Jiang, R. & Chen, T. (2011) Clustering 16S rRNA for OTU prediction: a method of
626 unsupervised Bayesian clustering. *Bioinformatics* 27, 611-618.

627 Hebert, P., Cywinska, A., Ball, S. & deWaard, J. (2003). Biological identifications through DNA
628 barcodes. *Proceedings of the Royal Society of London B* 270, 313–321.

629 Hug, L., Baker, B., Anantharaman, K., Brown, C., Probst, A. & al. (2016) A new view of the tree of
630 life. *Nature Microbiology* 1, 16048.

631 Hultman, J., Rahkila, R., Ali, J., Rousu, J. & Björkroth, K. (2015) Meat processing plant microbiome
632 and contamination patterns of cold-tolerant bacteria causing food safety and spoilage risks in the
633 manufacture of vacuum-packaged, cooked sausages. *Applied and Environmental Microbiology* 81,
634 7088-7097.

635 Huson, D., Auch, A., Qi, J. & Schuster, S. (2007) MEGAN analysis of metagenomic data. *Genome*
636 *Research* 17, 377-386.

637 Janzen, D., Hajibabaei, M., Burns, J., Hallwachs, W., Remigio, E. & Hebert, P. (2005) Wedding
638 biodiversity inventory of a large and complex Lepidoptera fauna with DNA barcoding.
639 *Philosophical Transactions of the Royal Society of London. Series B, Biological sciences* 360, 1835–
640 1845.

641 Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M. & al. (2012) Geneious Basic: an
642 integrated and extendable desktop software platform for the organization and analysis of
643 sequence data. *Bioinformatics* 28, 1647-1659.

644 Kielbasa, S., Wan, R., Sato, K., Horton, P. & Frith, M. (2011) Adaptive seeds tame genomic
645 sequence comparison. *Genome Research* 21, 487-493.

646 Levi, T., Kilpatrick, A., Mangel, M. & Wilmers, C. (2012) Deer, predators, and the emergence of
647 Lyme disease. *Proceedings of the National Academy of Sciences USA*, 109, 10942–10947.

648 Li, H. (2015) BFC: correcting Illumina sequencing errors. *Bioinformatics* 31, 2885–2887.

649 Lou, M. & Golding, G. (2012). The effect of sampling from subdivided populations on species
650 identification with DNA barcodes using a Bayesian statistical approach. *Molecular Phylogenetics*
651 *and Evolution* 65, 765-773.

652 Martin, M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads.
653 *EMBnet.journal* 17, 10-12.

654 Matsen, F., Kodner, R. & Armbrust, E. (2010) pplacer: linear time maximum-likelihood and
655 Bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC Bioinformatics*
656 11, 538.

657 Meyer, C. & Paulay, G. (2005) DNA barcoding: error rates based on comprehensive sampling. *PLoS*
658 *Biology* 3, e422.

659 Munch, K., Boomsma, W., Huelsenbeck, J., Willerslev, E. & Nielsen, R. (2008) Statistical assignment
660 of DNA sequences using Bayesian phylogenetics. *Systematic Biology* 57, 750-757.

661 NCBI Resource Coordinators (2016) Database resources of the National Center for Biotechnology
662 Information. *Nucleic Acids Research* 44, D7-D19.

663 Ondov, B., Bergman, N. & Phillippy, A. (2011) Interactive metagenomic visualization in a Web
664 browser. *BMC Bioinformatics* 12, 385.

665 Ovaskainen, O., Nokso-Koivisto, J., Hottola, J., Rajala, T., Pennanen, T. & al. (2010). Identifying
666 wood-inhabiting fungi with 454 sequencing – what is the probability that BLAST gives the correct
667 species. *Fungal Ecology*, 3, 274–283.

668 Ovaskainen, O., Schigel, D., Ali-Kovero, H., Auvinen, P., Paulin, L. & al. (2013) Combining high-
669 throughput sequencing with fruit body surveys reveals contrasting life-history strategies in fungi.
670 *ISME Journal* 7, 1696-1709.

671 Pansu, J., Giguet-Covex, C., Ficetola, G., Gielly, L., Boyer, F. & al. (2015) Reconstructing long-term
672 human impacts on plant communities: an ecological approach based on lake sediment DNA,
673 *Molecular Ecology*, 24, 1485-1498.

674 Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T. & al (2013) The SILVA ribosomal RNA gene
675 database project: improved data processing and web-based tools. *Nucleic Acids Research* 41,
676 D590-D596.

677 Rasmussen, C., Dupont, Y., Mosbacher, J., Trojelsgaard, K., Olesen, J. & al. (2013) Strong impact of
678 temporal resolution on the structure of an ecological network. *PLoS ONE* 8, e81694.

679 Ratnasingham, S. & Hebert, P. (2007) BOLD: The Barcode of Life Data System
680 (www.barcodinglife.org). *Molecular Ecology Notes* 7, 355–364.

681 Schnell, I., Thomsen, P., Wilkinson, N., Rasmussen, M., Jensen, L. & al. (2012) Screening mammal
682 biodiversity using DNA from leeches. *Current Biology* 22, R262–R263.

683 Somervuo, P., Koskela, S., Pennanen, J., Nilsson, H. & Ovaskainen, O. (2016) Unbiased probabilistic
684 taxonomic classification for DNA barcoding. *Bioinformatics*, btw346.

685 Staats, M., Arulandhu, A., Gravendeel, B., Holst-Jensen, A., Scholtens, I. & al. (2016) Advances in
686 DNA metabarcoding for food and wildlife forensic species identification. *Analytical and*
687 *Bioanalytical Chemistry* 408, 4615-4630.

688 Taberlet, P., Coissac, E., Pompanon, F., Brochmann, C. & Willerslev, E. (2012) Towards next-
689 generation biodiversity assessment using DNA metabarcoding. *Molecular Ecology* 21, 2045–2050.

690 Wang, Q., Garrity, G., Tiedje, J. & Cole, J. (2007) Naive Bayesian Classifier for Rapid Assignment of
691 rRNA Sequences into the New Bacterial Taxonomy. *Applied and Environmental Microbiology* 73,
692 5261–5267.

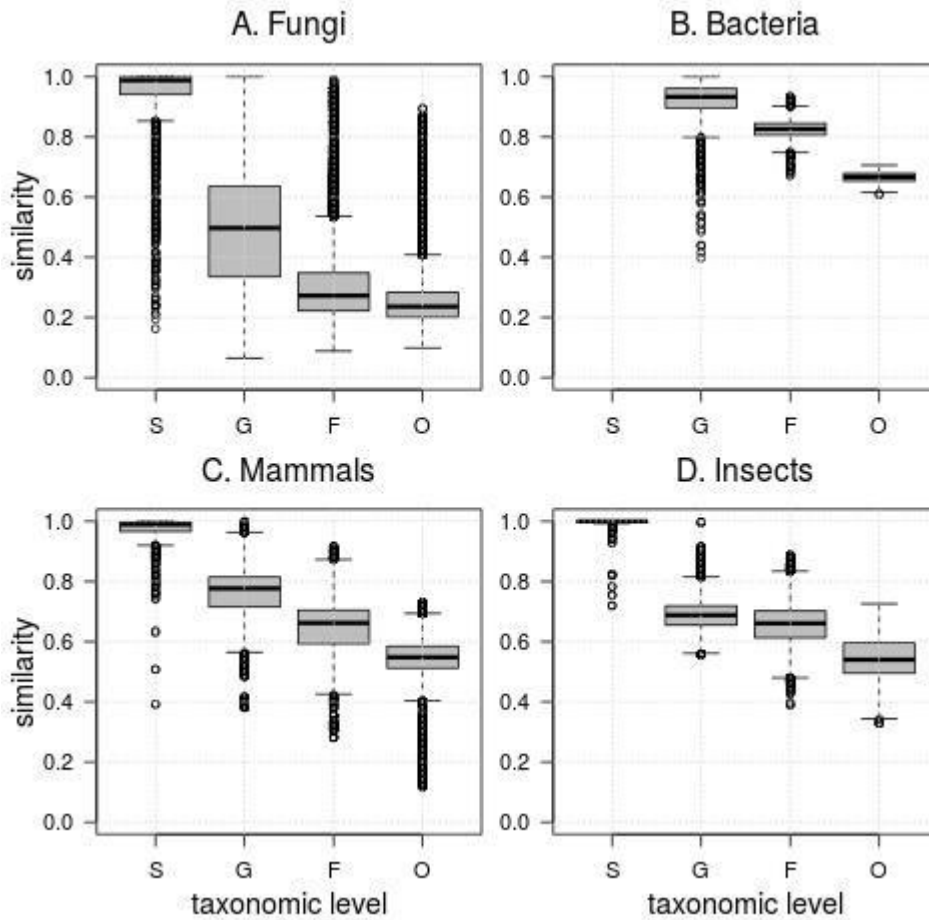
693 Wirta, H., Várkonyi, G., Rasmussen, C., Kaartinen, R., Schmidt, N. & al. (2016) Establishing a
694 community-wide DNA barcode library as a new tool for arctic research. *Molecular Ecology*
695 *Resources* 16, 809–822.

696 Wong E. & Hanner R. (2008) DNA barcoding detects market substitution in North American
697 seafood. *Food Research International* 41, 828–837.

698 Wood, T., Holland, J., & Goulson, D. (2015) Pollinator-friendly management does not increase the
699 diversity of farmland bees and wasps. *Biological Conservation* 187, 120-126.

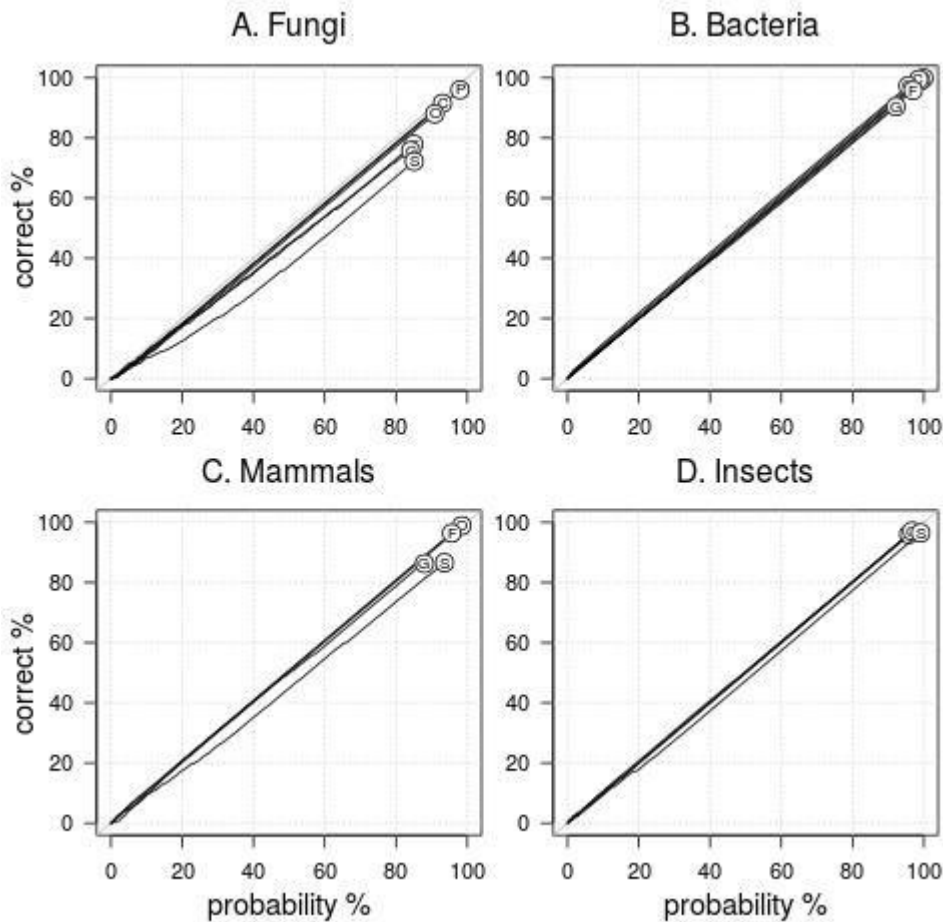
700 Yu, D., Ji, Y., Emerson, B., Wang, X., Ye, C. & al. (2012) Biodiversity soup: metabarcoding of
701 arthropods for rapid biodiversity assessment and biomonitoring. *Methods in Ecology and*
702 *Evolution* 3, 613–623.

703 **Figures**



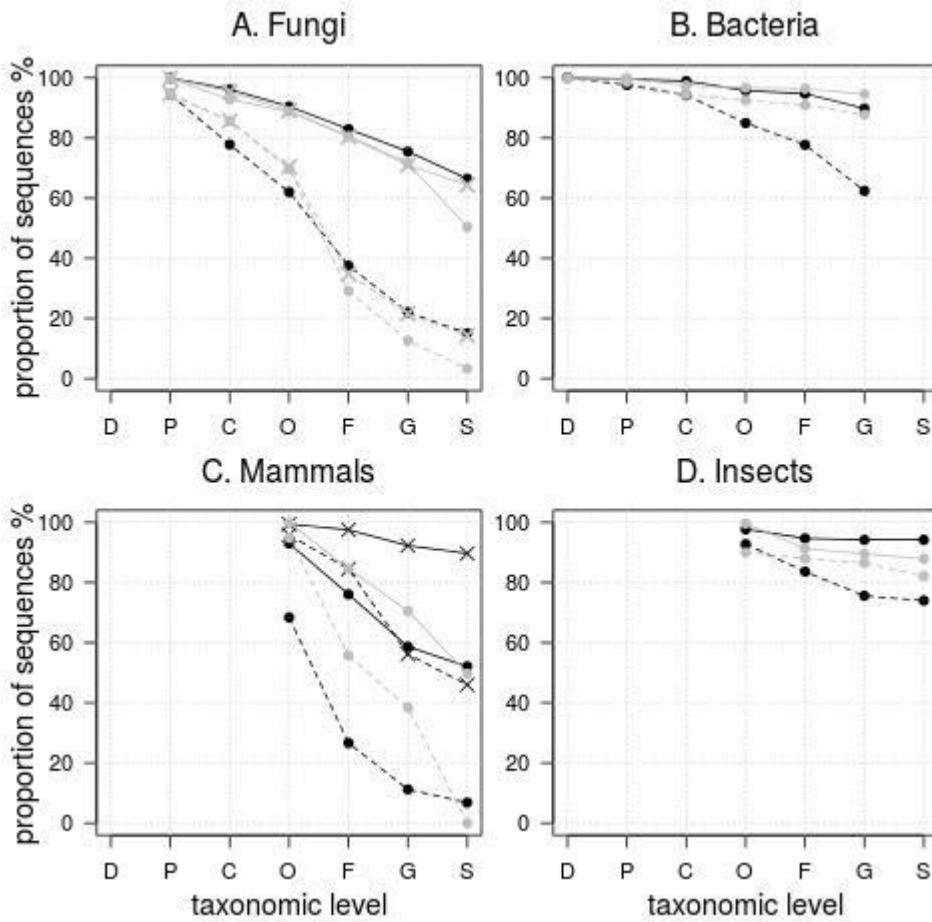
704

705 **Figure 1.** The distribution of pairwise LAST-similarities between reference sequences within each
706 taxonomical levels of species (S), genus (G), family (F) and order (O). The distribution of similarities
707 in a given taxonomical level originates from 1000 randomly selected sequence pairs. At the
708 species level, each sequence pair represents two different individuals of the same species. At the
709 genus, family, and order levels, each sequence pair represents, respectively, two different species,
710 genera, or families that belong to the same genus, family, or order.



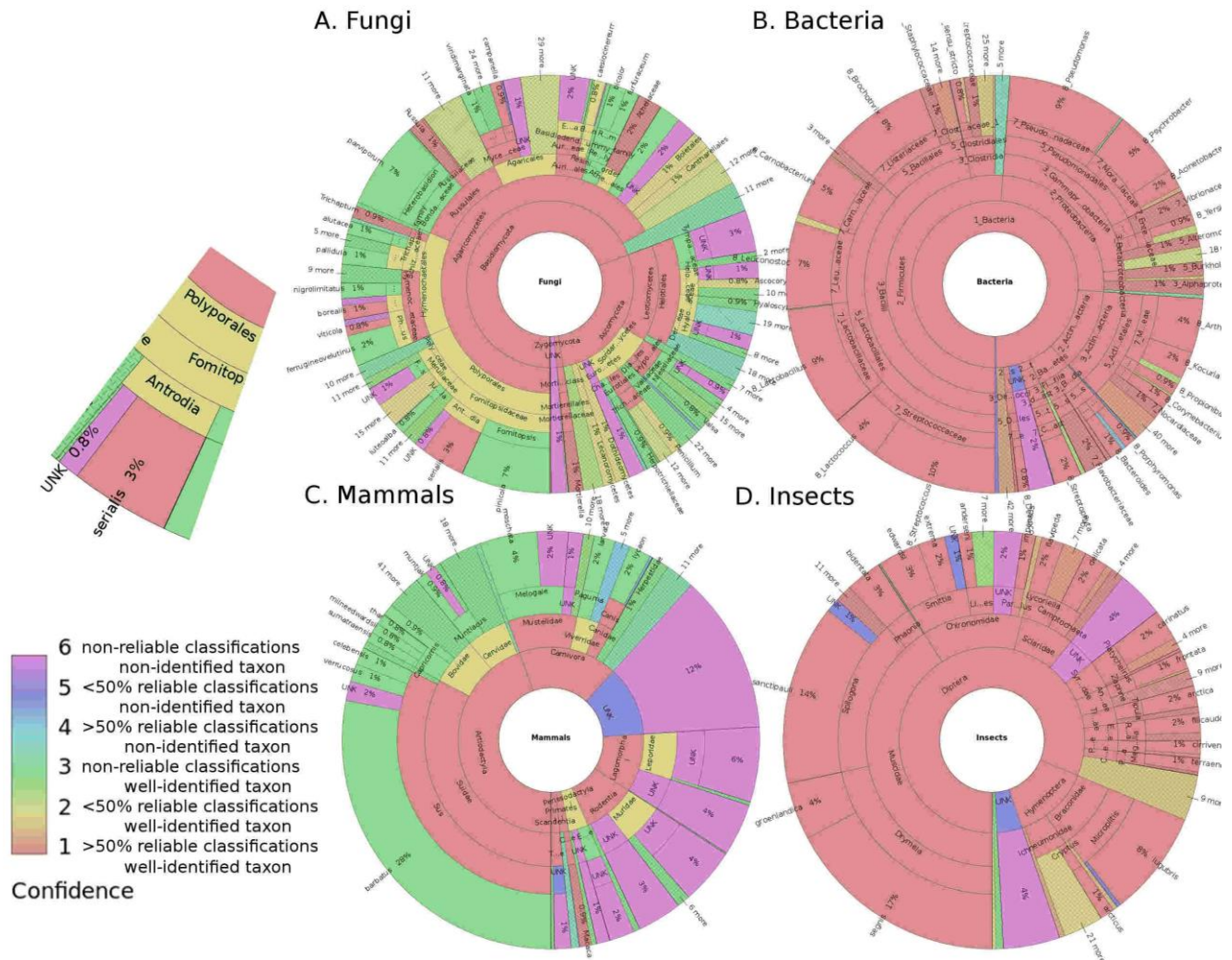
711

712 **Figure 2.** An assessment of bias and accuracy of the PROTAX algorithm for classifying well-
 713 identified sequence data to different taxonomic levels. We used PROTAX to classify well-identified
 714 reference sequences, with the focal sequence removed from the reference database to avoid
 715 circularity. The classification probabilities shown here are level-specific conditional probabilities,
 716 thus measuring e.g. the accuracy of species-level classifications conditional on knowing the true
 717 genus. While PROTAX yields a vector of identification probabilities for all possible outcomes, we
 718 considered here only the outcome with the largest identification probability, which we compared
 719 to the true identity of the species. For each taxonomic rank (indexed as S=species, G=genus,
 720 F=family, O=order, C=class, D=domain), panels show the cumulative number of correct
 721 identifications on the y-axis versus the cumulative sum of the identification probabilities on the x-
 722 axis (both normalized by the number of sequences). A curve matching with the identity line ($y=x$)
 723 indicates unbiased identification probabilities, both for small and large probabilities, as the
 724 identifications have been sorted in the order of increasing largest identification probability. The
 725 position of the dot gives the mean identification probability among the samples.



726

727 **Figure 3.** Confidence of taxonomic placement at different taxonomical levels. The value on the y-
 728 axis is the proportion of plausible (solid line) and reliable (dashed line) taxonomic placements.
 729 Results for validation data sampled from reference sequence database are shown in black and
 730 results for environmental query data are shown in gray. For fungi, gray crosses denote results
 731 from environmental data where species probabilities were weighted according to prior knowledge
 732 on which species exist in Finland. For mammals, black crosses denote results using full-length mt
 733 16S rRNA sequences as validation data. Taxonomic labels at x-axis from left to right: D=domain,
 734 P=phylum, C=class, O=order, F=family, G=genus, S=species.



735

736 **Figure 4.** Taxonomy pie charts of PROTAX output showing the composition of the environmental
 737 data sets. The width of each sector is proportional to the expected number of sequences that was
 738 placed to that taxonomic units. The colors code both the reliability of the identifications, and
 739 whether the identifications relate to taxonomic units that are part of the taxonomy or to unknown
 740 units (see color label). The enlarged insert illustrates species-level resolution for the fungal data.
 741 The charts are snapshots from interactive web pages (provided in the Supporting Information)
 742 generated by Krona software from the PROTAX output.

743

744 Tables

745 Table 1. Case studies used to evaluate the performance of PROTAX in probabilistic taxonomic placement of environmental sequence data.

Species group	Marker	Taxonomy	Reference database	Environmental data	Number of species (genera for bacteria) in taxonomy	Number of taxonomic levels used	Spatial extent of taxonomy	Reference sequences	Proportion of species with reference sequences	Median/Mean/Max number of sequences per species (for those which have >0 sequences)
Fungi	ITS2	Index Fungorum	UNITE and Ovaskainen et al. (2013)	Wood-inhabiting fungi sequenced from saw-dust samples from 100 spruce logs in Finland	130795	6	Global	75104	15%	2/2.8/5
Bacteria and Archaea	16S SSU rRNA gene	Ribosomal database project	RDP Release 11.4, trainset 15	Food processing plant factory microbiome, 101 samples	2175	6	Global	11127	100%	2/5.1/504
Insects	CO1	Greenland entomofauna, excluding Psocodea	Wirta et al. (2016)	Insect tissue from ~7000 specimens from North-East Greenland	844	4	Greenland	1853	26%	4/7.1/368
Mammals	16S mtDNA	NCBI	GenBank	iDNA of mammals from the residual blood meals of ~20,000 leeches from Vietnam and Laos	6675	4	Global	2627	20%	1/2.0/10

746

747