



**HAL**  
open science

## Delimiting species with single-locus DNA sequences

Nicolas Hubert, Jarrett D Phillips, Robert H Hanner

► **To cite this version:**

Nicolas Hubert, Jarrett D Phillips, Robert H Hanner. Delimiting species with single-locus DNA sequences. Robert DeSalle. DNA Barcoding: Methods and Protocols, 2744, Humana, pp.53-76, 2024, 978-1-0716-3583-4. 10.1007/978-1-0716-3581-0\_3 . hal-04579317

**HAL Id: hal-04579317**

**<https://hal.science/hal-04579317>**

Submitted on 17 May 2024

**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.

# 1 **Delimiting species with single-locus DNA sequences**

2

3 Nicolas Hubert<sup>1</sup>, Jarrett D. Phillips<sup>2,3</sup>, Robert H. Hanner<sup>3</sup>

4 <sup>1</sup>UMR ISEM (IRD, UM, CNRS), Université de Montpellier, Place Eugène Bataillon, 34095  
5 Montpellier cedex 05, France

6 <sup>2</sup>School of Computer Science, University of Guelph, 50 Stone Rd E, Guelph, ON N1G2W1, Canada

7 <sup>3</sup>Department of Integrative Biology, University of Guelph, 50 Stone Rd E, Guelph, ON N1G2W1,  
8 Canada

9 Contact: [nicolas.hubert@ird.fr](mailto:nicolas.hubert@ird.fr)

10

11 Running title: DNA-based species delimitation

12

## 13 **Abstract**

14 DNA sequences are increasingly used for large-scale biodiversity inventories. Because these  
15 genetic data avoid the time-consuming initial sorting of specimens based on their phenotypic  
16 attributes, they have been recently incorporated in taxonomic workflows for overlooked and  
17 diverse taxa. Major statistical developments have accompanied this new practice and several  
18 models have been proposed to delimit species with single-locus DNA sequences. However,  
19 proposed approaches to date make different assumptions regarding taxon lineage history,  
20 leading to strong discordance whenever comparisons are made among methods. Distance-  
21 based methods, such as Automatic Barcode Gap Discovery (ABGD) and Assemble Species by  
22 Automatic Partitioning (ASAP), rely on the detection of a barcode gap (i.e., the lack of overlap  
23 in the distributions of intraspecific and interspecific genetic distances) and the associated  
24 threshold in genetic distances. Network-based methods, as exemplified by the REfined Single

25 Linkage (RESL) algorithm for the generation of Barcode Index Numbers (BINs), use connectivity  
26 statistics to hierarchically cluster related haplotypes into molecular operational taxonomic units  
27 (MOTUs) which serve as species proxies. Tree-based methods, including Poisson Tree Processes  
28 (PTP) and the General Mixed Yule Coalescent (GMYC), fit statistical models to phylogenetic trees  
29 by maximum likelihood or Bayesian frameworks.

30 Multiple webservers and standalone versions of these methods are now available, complicating  
31 decision-making regarding the most appropriate approach to use for a given taxon of interest.

32 For instance, tree-based methods require an initial phylogenetic reconstruction, and multiple  
33 options are now available for this purpose such as RAxML and BEAST. Across all examined  
34 species delimitation methods, judicious parameter setting is paramount, as different model  
35 parameterizations can lead to differing conclusions. The objective of this chapter is to guide  
36 users step-by-step through all the procedures involved for each of these methods, while  
37 aggregating all necessary information required to conduct these analyses. The Materials section  
38 details how to prepare and format input files, including options to align sequences and conduct  
39 tree reconstruction with Maximum Likelihood and Bayesian inference. The Methods section  
40 presents the procedure and options available to conduct species delimitation analyses,  
41 including distance-, network- and tree-based models. Finally, limits and future developments  
42 are discussed in the Notes section. Most importantly, species delimitation methods discussed  
43 herein are categorized based on five indicators: reliability, availability, scalability,  
44 understandability, and usability, all of which are fundamental properties needed for any  
45 approach to gain unanimous adoption within the DNA barcoding community moving forward.

46

47 **Keywords**

48 Poisson Tree Processes, PTP, General Mixed Yule Coalescent model, GMYC, Barcode Index  
49 Number, BIN, Assemble Species by Automatic Partitioning, ASAP, Single threshold, Multiple  
50 threshold

51

## 52 **1. Introduction**

53 The use of mitochondrial DNA sequences to rapidly identify individuals to the species level has  
54 been increasingly used over the past two decades in the context of DNA barcoding [1]. Based  
55 on variation within ca. 650 base pairs of the cytochrome *c* oxidase subunit I gene, which serves  
56 as an internal molecular classification tag in animals, DNA barcoding proved to be operational  
57 for both specimen identification and species discovery in multiple groups as a means to address  
58 the longstanding taxonomic impediment [2, 3]. Robust identifications through DNA barcodes,  
59 used to assign unknown specimens to known species, rely heavily on the development of  
60 curated reference sequence libraries derived from adult vouchers which have previously been  
61 identified using morphological characters and current taxonomic literature [4]. While rapidly  
62 implemented for well-known faunas, these reference libraries proved to be challenging to  
63 develop for mega-diverse and poorly known taxa. Although not its initial goal, DNA barcoding  
64 has been widely employed in this context [5–7], and its integration with taxonomic workflows  
65 has led to new methodological and conceptual developments. Several factors set DNA-based  
66 identifications apart from those made on the basis of morphology alone: (1) the initial sorting  
67 of specimens is performed using DNA barcodes, as they constitute a quick alternative to the  
68 time- and labor-intensive categorization of specimens using morphological characters [8–10];  
69 (2) DNA sequences are combinations of four discrete states of known inheritance: A, T, G, and  
70 C. As such, the subjective procedure of standardization with phenotypes is avoided [4, 11]; (3)  
71 branching patterns reflective of the evolutionary history of nucleotide substitutions and DNA

72 sequences in a tree occur at different rates within and between species, as well as among  
73 different genomic loci; and, (4) morphologically indistinguishable diversity, namely cryptic  
74 species variation, has been extensively documented in largely unexplored regions, such as  
75 tropical, arctic, and marine ecosystems, thereby proving to bias our understanding of eco-  
76 evolutionary mechanisms and processes underlying diversity patterns [12–15]. This so-called  
77 “DNA-based workflow” has greatly enabled the development of statistical models to detect  
78 boundaries of genetically-isolated lineages [16–18].

79         Several statistical models have been specifically developed to perform standardized  
80 classification of DNA sequences into Molecular Operational Taxonomic Units (MOTUs)  
81 displaying genetic properties similar to that of species: (1) higher branching rates in a  
82 phylogenetic tree among sequences within species than between species; (2) higher genetic  
83 distances among sequences between, compared to within, species, i.e. leading to the formation  
84 of a DNA barcode gap; and, (3) detection of multiple diagnostic nucleotide substitutions of each  
85 MOTU (i.e., synapormorphies). These methods can be classified into three main categories,  
86 each implementing different strategies to capture species boundaries: (1) distance-based  
87 methods such as Automatic Barcode Gap Discovery (ABGD) [19] and Assemble Species by  
88 Automatic Partitioning (ASAP) [20], which examine pairwise genetic distances among  
89 sequences to detect the presence of a barcode gap, and the genetic distance at which it is  
90 expected to occur for delineating MOTUs, given available sequence data; (2) network-based  
91 methods, such as the REfined Single Linkage (RESL) algorithm, as implemented in the Barcode  
92 of Life Data System (BOLD) [21] to produce Barcode Index Numbers (BINs) [22], which employ  
93 a graph Markov clustering approach to explore connectivity among sequences through random  
94 walks of the network, a process that exposes regions of sparsity as potential taxon boundaries;  
95 (3) model-based approaches, such as the General Mixed Yule Coalescent (GMYC) model [23, 24]

96 and Poisson Tree Processes (PTP) model [25, 26], which apply mixture models with two distinct  
97 components within and between species (two Poisson distributions of branching events for PTP,  
98 or a coalescent, together with a Yule, diversification model for GMYC) to phylogenetic trees and  
99 partition clusters by selecting the best single threshold or multiple thresholds through  
100 Maximum Likelihood (ML) or Bayesian (B) parameter estimation methods.

101         These approaches are increasingly used to delineate MOTUs during biodiversity  
102 inventories as an initial step to group specimens according to their genetic similarities [27–32].  
103 However, each of these methods is prone to serious pitfalls: some oversplit singletons (i.e.  
104 MOTUs represented by a single sequence), while others are too conservative within lineages  
105 displaying higher diversification rates, thereby leading to frequent overlumping [33–35].  
106 Therefore, the abovementioned methods are often used concomitantly to circumvent these  
107 potential issues, where several strategies have been developed to estimate the robustness of  
108 their delimitation schemes: (1) concordance between methods is estimated by metrics  
109 quantifying the number of differences between one particular approach and all others, where  
110 the delimitation scheme resulting from the most discordant methods is then discarded [31, 32,  
111 35]; (2) model-based approaches provide an estimated probability supporting the classification  
112 scheme for each node in the tree, where poorly supported splits can be ignored [23, 25]; and,  
113 (3) a majority-rule consensus is derived from the delimitation schemes provided by all methods  
114 [12, 27, 28, 36, 37].

115         The present chapter focuses on the most commonly used methods in the literature  
116 (ASAP, GMYC, PTP, BIN, and ABGD). Procedures to perform distance-, network- and model-  
117 based delimitation analyses are detailed, from the preparation of the input data, to the collation  
118 and comparison of delimitation schemes. Data preparation and computational software  
119 packages are detailed for each category of analyses.

120

## 121 **2. Materials**

122 Species delimitation algorithms use multiple input files, ranging from sequence alignments and  
123 distance matrices, to phylograms and ultrametric trees (Fig. 1). Prepare all input files prior to  
124 running delimitation analyses, ensuring DNA barcode sequences are in a FASTA-formatted file.  
125 Packages listed below are available for all major operating systems (Windows, Mac OSX, Linux).

126

### 127 2.1 Aligning and formatting DNA barcodes

128 1. Unipro UGENE: Download the latest version of Unipro UGENE [38] at *ugene.net* for your  
129 operating system. Several alignment algorithms are available and accessible from the upper  
130 menu. Go to 'Tools', and 'Multiple sequence alignment'. Select 'Align with MUSCLE', select the  
131 input file (input.fasta) and name your output file (e.g. alignment). Alignment can be performed  
132 with default settings. For large alignments and limited computing resources, MUSCLE options  
133 can be changed (e.g., mode 'large alignment'). Faster alternatives are available. Select 'Align  
134 with ClustalW', define input (input.fasta) and output (alignment) files and perform the  
135 alignment with default settings. Aligned sequences in FASTA format are automatically saved in  
136 your source folder as alignment.aln. Aligned sequences can be exported in other formats such  
137 as NEXUS or PHYLIP using the window menu option 'Save alignment as'.

138

139 2. AliView: Download the most recent version of AliView [39] at <https://ormbunkar.se/aliview/>  
140 compatible with your operating system. Import your input file (alignment.fasta) with the option  
141 'File' and 'Open file' in the upper menu, or right click on your input file with 'open with' and  
142 select AliView. Go to 'Align', and 'Realign everything'. MUSCLE is used for alignment by default.  
143 Alternative algorithms are available from the menu 'AliView' and 'Preferences'. Save aligned

144 sequences with 'File' and 'Save as'. Multiple exporting format are available. Export the  
145 alignment in FASTA format (alignment.fasta).

146

147 3. SeaView: Download the last version of Seaview [40] at  
148 <https://drouot.prabi.fr/software/seaview> for your operating system. Import your input file  
149 (alignment.fasta) with the option 'File' and 'Open fasta' in the window menu or right click on  
150 your input file with 'open with' and select SeaView. Go to 'Align', 'Alignment options', and select  
151 the algorithm (Clustal or MUSCLE). Clustal is used by default. Run the alignment algorithm with  
152 'Align' and 'Align all'. Save aligned sequences with 'File' and 'Save as'. Multiple exporting format  
153 are available. Export as a FASTA file (alignment.fst).

154

155 2.2 Reconstructing a phylogram by Maximum Likelihood for PTP

156 1. Maximum Likelihood inferences with jModelTest2 and PhyML: Download the current version  
157 of jModelTest2 [41] at <https://github.com/ddarriba/jmodeltest2/releases> and PhyML 3.0 [42]  
158 at <http://www.atgc-montpellier.fr/phyml/versions.php>. Import your input file (alignment.fasta)  
159 with option 'File' and 'Load DNA alignment' in the window's menu. Compute likelihood scores  
160 for multiple DNA substitution models. Go to 'Analysis' in the window menu, select 'Compute  
161 likelihood score' and define parameters. Use default parameters for 'number of substitution  
162 schemes' (11 corresponding to 88 numerical models), 'base frequencies' (+F) and 'rate  
163 variation' (+I, +G, nCat=4). Depending on computing resources available, select either 'Fixed  
164 BIONJ-JC' or 'BIONJ' (used for low resources), or 'ML optimized' (employed for moderate to high  
165 resources) in the 'Base tree for likelihood calculations'. If 'ML optimized' is selected, several  
166 base tree search options are available with varying computing requirements from low (NNI) to  
167 high (best). Start with 'Fixed BIONJ-JC'. Once calculations are completed, perform model

168 selection using the 'Analysis' option in the menu and select 'Do BIC calculations', which uses  
169 the Bayesian Information Criterion to calculate "parsimony" scores for all considered models of  
170 DNA substitution. Alternative information criteria, such as the Akaike Information Criterion  
171 (AIC), are also available. The model producing the lowest information criterion value should be  
172 preferred to all other tested models. Before quitting, save the content of the 'PhyML-log' in a  
173 text file, and save the output using the 'Results' option of the menu, along with 'Build html log'.  
174 Check the html log in the 'log' folder of the jModelTest2 folder by opening the html file with a  
175 web browser. Once done, place a copy of the alignment in FASTA format (alignment.fasta) into  
176 the PhyML folder and open PhyML 3.0. Click on the exe file in windows or run the command  
177 line version in Mac OSX ('cd/home/.../PhyML-3.1' and launch with './phymIPhyML-3.1\_macOS-  
178 Mountailion') or Linux. Enter the sequence file name (alignment.fasta) and navigate in the  
179 menu with '+' and '-'. All the options to run the most likely model are available in the 'PhyML-  
180 log' window in jModelTest2. Browse the log to obtain all the options in the command line at  
181 the end of the file. Select most-likely model identified by jModelTest2 with 'm'. If the best  
182 nucleotide substitution model is not available, select 'custom' and define the model manually  
183 with 'K' (model code available in the PhyML-log of jModelTest2). Define all the parameters and  
184 run the analysis. The output tree is saved in the PhyML folder.

185  
186 2. Maximum Likelihood inferences with IQtree: Connect to the webserver version of IQtree [43,  
187 44] available at <http://iqtree.cibiv.univie.ac.at/>. At the 'tree inference' window, import your  
188 alignment in FASTA format (alignment.fasta) using 'Alignment file' in the 'Input Data' section.  
189 Set sequence type to 'DNA' or let IQtree detect the sequence type automatically. Assuming the  
190 alignment contains a single locus, no partition file is required. In the 'Substitution Model  
191 Options' section, set the 'Substitution model' to 'auto', and use default parameters for

192 'FreeRate heterogeneity' and "#rate categories'. The section 'Branch Support Analysis' is  
193 optional, single locus phylograms are expected to be poorly supported if ancient lineages are  
194 included. Use default parameters in the 'IQ-TREE Search Parameters', and then submit. Provide  
195 your email address to get an alert when the calculation is finished, or go to the window 'Analysis  
196 Results' and wait for completion. Once done, select your job and download results with the  
197 'Download selected jobs' option at the bottom of the window. Output files include IQtree log  
198 (alignment.fasta.log), treefile (alignment.fasta.treefile), and substitution model selection  
199 (alignment.fasta.iqtree).

200

201 3. Maximum Likelihood inferences with RAxML: RAxML is currently under active development  
202 and performance varies according to operating system. Download the pre-compiled binary of  
203 the latest version of RAxML-NG [45] at <https://github.com/amkozlov/raxml-ng>, and install it.  
204 The user manual is available at <https://github.com/amkozlov/raxml-ng/wiki>. Select the best  
205 nucleotide substitution model with jModelTest2 (see section above). In Mac OSX and Linux, go  
206 to source folder 'cd/home/.../raxml-ng' and run the tree inference with './raxml-ng --msa  
207 alignment.fasta --model GTR+I+G' (running with a general time reversible model with  
208 proportion of invariable sites and rate heterogeneity). In Mac OSX, both tree inference and  
209 model selection can be conducted with RAxML-NG and ModelTest-NG [46] using raxmlGUI 2.0  
210 [47] available at <https://antonellilab.github.io/raxmlGUI/>. Open raxmlGUI and import  
211 alignment in FASTA format. In the RAxML section, define the binary as 'modeltest-ng' to select  
212 the most likely model, which is provided in the output file 'alignment.raxml.bestModel'. Set  
213 RAxML binary to 'raxml-ng', define the best nucleotide substitution model and parameters in  
214 'Input' section, set the type of analysis in the 'Analysis' section to 'ML tree inference' (cf.

215 previous comments regarding single locus analyses and bootstrap support) and run the analysis.  
216 Output files include the best tree (alignment.raxml.bestTree) and log file (alignment.raxml.log).

217

## 218 2.3 Reconstructing an ultrametric tree with BEAST2 for GMYC

219 1. Collapse sequences to haplotypes: GMYC performs better if ultrametric trees are  
220 reconstructed using an alignment consisting of haplotypes only [48]. RAxML automatically  
221 produces a haplotype alignment (alignment.raxml.reduced.phy), which can be used here.  
222 Alternatively, the program Alignment Transformation EnviRonment (ALTER) at  
223 <http://www.sing-group.org/ALTER/> can be used to produce a haplotype alignment. First select  
224 the format of your alignment or autodetect it in the 'select format' section. Second, upload your  
225 alignment in FASTA format (alignment.fasta) in the 'upload or paste MSA' section. Third, select  
226 options and format in 'select output format and convert'. Set 'select program' to 'General',  
227 select 'Collapse sequences to haplotypes', set 'format' to 'FASTA' and proceed to the conversion  
228 with 'convert'. Fourth, export the haplotype alignment (alignment.fasta.alter.haps.fas) in the  
229 section 'save converted MSA'.

230

231 2. Preparing the input file with BEAUTY: BEAST2 [49] is available for Windows, Mac OS X and  
232 Linux; the most up-to-date version can be obtained at <https://www.beast2.org/>. The folder  
233 includes BEAUTY for preparing input files, BEAST2 to run Markov Chain Monte Carlo (MCMC),  
234 LogCombiner to combine multiple MCMC runs, and TreeAnnotator to reconstruct the  
235 consensus tree. Download Tracer [50] as well at <https://beast.community/tracer> for a graphical  
236 visualization and diagnostics of MCMC output. First, open BEAUTY and import the sequence  
237 alignment in FASTA format with haplotypes only (alignment.fasta.alter.haps.fas) by selecting  
238 the option 'File' and 'Import Alignment' in the upper menu. In 'Site Model' section, define the

239 nucleotide substitution model. Previous ML model selection analyses with jModelTest2 or  
240 ModelTest-NG or IQtree can be used to guide model selection and define tree priors. By default,  
241 only a subset of models is available. For more models, obtain the package 'substmodels'  
242 (<https://github.com/rbouckaert/substmodels>) by selecting 'File' and 'Manage Packages'. First  
243 select the model in the 'Subst Model' section (JC69 by default). If HKY is desired, set 'Kappa'  
244 using ML estimates and set 'Frequencies' to 'Empirical' (save computing time). Set 'Gamma  
245 Category Count', 'Shape' and 'Proportion Invariant' using ML estimates. If not included in the  
246 best substitution model by the ML algorithm, set 'Gamma Category Count' and 'Proportion  
247 Invariant' to zero. Alternatively, parameters of the substitution model can be estimated jointly  
248 with tree topology and age estimates by clicking on 'estimate' for each of the parameter. In the  
249 'Clock Model' section, set the clock model and rate. Set the model as 'Strict clock' for a clock-  
250 like model or 'Relaxed Clock Log Normal' if rate heterogeneity is expected among lineages. Set  
251 the clock with 'Clock.rate' and use 0.01 if a value of 1% genetic divergence per million years is  
252 applied. In the 'Priors' section, use default parameters. No prior regarding clade age is required  
253 here. In the 'MCMC' section, set the 'Chain Length' to at least 20 million, set 'Log Every' to  
254 10,000 to save 2,000 trees. Save your setting in 'File' with 'Save as' in the upper menu (xml file).

255  
256 3. Run the MCMC and check results: Open BEAST2 and select the input file (xml file) with  
257 'Choose File' in 'BEAST XML File'. Set 'default: only write new log files' to avoid overwriting  
258 results and run the MCMC. Once done, create a new folder (e.g. RUN1), place all the output  
259 files (input\_file.xml.state, input\_file.log, and input\_file.trees) and a run a second MCMC. Open  
260 Tracer to visualize the MCMC traceplot. In the 'Trace files' section, click on '+' to add a new  
261 traceplot, select the log file (input\_file.log) and repeat with the log file of the second run. The  
262 estimated parameter value (Mean) and effective sample size (ESS) of substitution and

263 diversification parameters are indicated in the left panel. Check the stability and convergence  
264 of the two chains by selecting 'Trace' in the right panel. If traceplots reach stability (only random  
265 fluctuations around a mean) and the two chains converge (parameter estimates are similar),  
266 check the combined traces in the 'Trace files' panel. If the two chains have mixed well,  
267 traceplots should look like "fuzzy caterpillars", which indicates that generated samples  
268 correspond well to draws from the posterior distribution. If all parameters have ESS > 200,  
269 proceed to the next step. If poor chain mixing and/or low ESS are present, run a third chain with  
270 a higher number of iterations (e.g., chain length = 50 million) and check these diagnostics again.  
271 For the next step, identify the number of states where the chains stabilize. All previous states  
272 should be discarded by defining a Burn-in amount in the upper left panel ('States', 'Burn-in'),  
273 corresponding to the number of states before stability is reached (e.g., if stability is reached at  
274 2 million steps, a burn-in of at least 2 million is appropriate). Setting the Burn-in is important to  
275 reduce dependence on initial conditions.

276

277 4. Combine MCMC, reconstruct the consensus, and prepare the tree: Open LogCombiner and  
278 set 'File type' to 'Tree Files'. Import trace files with '+' and set the burn-in for both runs. Burn-  
279 in should be expressed in percentage of the chain length. Ideally, the chain length, which  
280 expresses the number of MCMC iterations, should be as high as possible. For example, if the  
281 chain is 20 million steps long and the burn-in is 2 million states, set the burn-in to 10%. Each run  
282 may have a different burn-in. Name the output file (combined.trees) in the 'Output file' panel  
283 and run. Open TreeAnnotator to reconstruct the consensus tree. An additional burn-in  
284 percentage can be applied here, but if burn-in was previously defined correctly, this additional  
285 burn-in is not required. Set the 'Target tree type' to 'Maximum clade credibility tree' (MCT) and  
286 'node heights' to 'Common Ancestor heights', select 'Input Tree file' and choose the combined

287 tree file (combined.trees), name the output file (MCT.tree) and run. To visualize the MCT and  
288 prepare it for GMYC, download FigTree at <https://github.com/rambaut/figtree/releases>,  
289 available for Windows, Mac OS X and Linux. Open FigTree and import the MCT file (MCT.tre)  
290 with 'File' and 'open' in the upper menu. Export the tree using 'File' and 'Export Trees', set 'Tree  
291 file format' to 'Newick' and set as 'Save as currently displayed'. Name the output file  
292 (MCT\_GMYC.tre) and save. You can further explore the MCT with FigTree by visualizing 95%  
293 credibility intervals (select 'Node Bars' and set to 'CAheight\_95%\_HPD') or posterior  
294 probabilities (select 'Node Labels' and set to 'posterior').

295

### 296 **3. Methods**

#### 297 3.1 Running distance-based methods:

298 1. Delimiting MOTUs with ABGD: The ABGD webservice is available at  
299 <https://bioinfo.mnhn.fr/abi/public/abgd/abgdweb.html>. Import the complete alignment in  
300 FASTA format (alignment.fasta; without collapsing sequences to haplotypes) using 'parcourir'  
301 or copy-paste the alignment in the 'paste your data' panel. Set parameters 'Pmin' (minimum  
302 genetic distance) and 'Pmax' (maximum genetic distance) to default values (0.001 and 0.1,  
303 respectively). 'Steps' defines the number of iterations for optimizing local genetic thresholds  
304 (within primary MOTUs) and delimiting MOTUs. Start with a default value of 10. Set the genetic  
305 distance to JC69 (Juke-Cantor corrected p-distance; [51]) or K80 (Kimura-Two-Parameter (K2P);  
306 [52]) if kappa (ratio of transitions to transversions (ts/tv)) is known. If K80 is selected, a kappa  
307 of 2.0 is used by default, indicating that transitions are twice as likely to occur compared to  
308 transversions. 'Nb bins' defines the number of bins in which to pool genetic distances: the  
309 higher the value, the better the resolution. Start with a default value of 20 and increase if  
310 necessary. Run the analysis. Results include the histogram of distances, ranked distances, and

311 the number of partitions according to the prior intraspecific divergence (threshold between  
312 intra- and interspecific distributions). In the example in Fig. 2, a DNA barcode gap is observed  
313 between 0.02 and 0.05 JC69 genetic distance. Partition schemes are accessible by clicking on a  
314 symbol. Doing so will open a new window for the selected partition, with a link to download  
315 the annotated tree, which can be then opened with FigTree. A list of individuals for each  
316 inferred group is also included.

317

318 2. Delimiting MOTUs with ASAP: The ASAP webserver is available at  
319 <https://bioinfo.mnhn.fr/abi/public/asap/>. ASAP is now recommended over ABGD, which has  
320 largely been superseded. Import the complete alignment in FASTA format (alignment.fasta;  
321 without collapsing sequences to haplotypes) using 'Choose a file'. Set the genetic distance  
322 model to JC69, or K80 if kappa is known. Additional options are available which include the split  
323 groups probability threshold, number of best scores to keep, and Pmin and Pmax thresholds.  
324 Run the analysis. Results include a table with partition schemes ranked by their score (asap-  
325 score') with the number of MOTUs ('Nb of subsets'), the p-value, the relative gap width metric  
326 ('W') and the threshold distance ('Threshold dist.'). A list of individuals per MOTU (subset) can  
327 be download in the 'text' column by clicking on 'list' for an ABGD-like format or 'CSV' for a more  
328 flexible format. ASAP provides similar histograms of distance distribution and ranked  
329 distribution as ABGD (Fig. 2) with a barcode gap observed between 0.02 and 0.05 JC69 genetic  
330 distance. Outputs also include a plot of ASAP-score against genetic distance and an UPGMA  
331 dendrogram with split group probability for nodes. Mapping MOTUs onto the UPGMA tree for  
332 the 10 best partition schemes is also available at 'View/Save Boxed subsets graph here'.

333

334 3.2 Delimiting MOTUs with BINs in BOLD:

335 1. Create a project in BOLD: No stand-alone package is available to perform RESL analyses,  
336 which are only available in BOLD for deposited data sets with already existing BINs. The  
337 alignment should be first deposited in BOLD at <http://www.boldsystems.org/>. Projects are only  
338 accessible to registered users. To register, click on 'LOGIN' in the upper menu and select  
339 'CREATE AN ACCOUNT'. Depositing DNA barcode sequence requires creating a project.

340

341 2. Submit sequences in BOLD: Open 'Projects' in the left menu and create a project with '+ New  
342 Project'. Enter details about the project, set parameters and save. To access the newly created  
343 project, go to 'Projects' in the left menu, select 'View All Projects' and click on your project.  
344 Once done, go to 'Uploads' and select 'Specimen Data'. Once submitted and validated,  
345 sequences can be uploaded similarly with 'Sequences' option in 'Uploads'. Details about data  
346 formatting and submission can be found in the BOLD handbook (available at  
347 <http://www.boldsystems.org/index.php/Resources>).

348

349 3. Collect BIN numbers in BOLD: Newly submitted DNA barcode sequences are assigned to a  
350 BIN when the RESL algorithm is run every month. Once available, enter the project and select  
351 'Data Spreadsheets' in 'Downloads' from the left menu. In the Spreadsheet Download window,  
352 select 'Progress Report' and any additional information in 'Specimen Data' and click 'Download'.  
353 Open the Excel file and BINs are available in the 'BIN' column, together with specimen code  
354 ('Sample ID'). BIN details are available at 'View All Records' in the 'BIN' column by clicking on  
355 BIN numbers.

356

357 3.3 Delimiting MOTUs with GMYC, single and multiple thresholds

358 1. Maximum Likelihood version using the R package 'splits': RStudio (Allaire, 2012) is  
359 recommended when running GMYC analyses. It can be obtained at '<https://posit.co/>', where  
360 joint installation of RStudio Desktop and R is available. Open RStudio, define the folder with the  
361 haplotype alignment collected from RAxML or ALTER with 'Session' in the upper menu, 'Set  
362 Working Directory' and 'Choose directory'. Packages *ape* and *splits* are required. In the bottom  
363 right panel, select 'Packages' and 'Install', and type 'ape' and 'install'. The *splits* package might  
364 not appear among available package if it has not been previously installed. If this is the case,  
365 use the following command to download and install from the R-forge repository:

366

```
367 > install.packages("splits", repos="http://R-Forge.R-project.org")
```

368

369 Once *ape* and *splits* are installed, load them using the 'Packages' option in the bottom right  
370 panel by clicking. Once required packages are installed, they can be loaded using:

371

```
372 > library(ape)
```

```
373 > library(splits)
```

374

375 Once loaded, documentation for each package can be obtained with:

376

```
377 > library(help="ape")
```

```
378 > library(help="splits")
```

379

380 or more simply

381

```
382 > ?ape
```

```
383 > ?splits
```

384

385 Note that `library("package name")` needs to be called every time a new R session is initiated. To

386 import the BEAST2 ultrametric tree prepared for GMYC (`MCT_GMYC.tre`) and run GMYC with a

387 single threshold use:

388

```
389 > tree <- read.tree("tree.nwk")
390 > gmyc_single <- gmyc(tree, method = "single")
```

391

392 For the multiple threshold version, use the argument `method = "multiple"`. Details are

393 available using:

394

```
395 > help("gmyc")
```

396

397 or

398

```
399 > ?gmyc
```

400

401 Results can be collected using:

402

```
403 > summary.gmyc(gmyc_single) # summary statistics of the results
404 > plot.gmyc(gmyc_single) # lineage through time with inferred threshold
405 > MOTU_list<-spec.list(gmyc_single) # list of MOTUs and individuals
406 > write.csv(MOTU_list, file = "MOTU_gmyc_single.csv") # export the list in
407 csv format
408 > support <- gmyc.support(gmyc_single) # estimate support
409 > is.na(support[support == 0]) <- TRUE # select nodes
410 > plot(tree, cex = 0.4, no.margin = TRUE) # plot tree
```

```
411 > nodelabels(round(support, 2), cex = 0.9) # plot support on tree
```

412

413 In the example in Fig. 3, the lineage through time plot indicates a shift at 0.63 million years  
414 corresponding to the inferred threshold (Fig. 3B), which is the most likely (Fig. 3A), and used to  
415 delimit MOTUs (Fig. 3C). In most cases, GMYC's single threshold option should be sufficient for  
416 most species delimitation tasks, as studies have clearly demonstrated little difference  
417 compared to the multiple threshold option (e.g., [48]). Furthermore, the multiple thresholds  
418 option is considerably slower in terms of computation time than the single thresholds version;  
419 however, this depends strongly on data set size [25, 48, 54].

420

421 2. Maximum Likelihood version with the GMYC web server: The single threshold and multiple  
422 thresholds implementation of GMYC can be run online at <https://species.h-its.org/gmyc/>.  
423 Upload your Newick tree reconstructed with sequences collapsed to haplotypes with 'My  
424 ultrametric input tree (Newick format only)', select the method 'single' or 'multiple', provide  
425 your email address, and run. Results displayed include the lineage through time plot with  
426 inferred threshold, distribution of the likelihood score through time, the annotated tree with  
427 MOTUs, and summary statistics similar to that obtained with the `summary.gmyc()` function in  
428 'splits'.

429

### 430 3.5 Delimiting MOTUs with PTP

431 1. Bayesian standalone version of PTP: The Bayesian implementation of the single threshold PTP  
432 model (bPTP) is available at <https://github.com/zhangjjajie/PTP> for Linux. The package is  
433 expected to run in Mac OS X and Windows if the required Python package is installed. To install  
434 the latest version of Python, type:

435

```
436 Pip3 install -r requirements.txt
```

```
437 Python3 steup.py install
```

438

439 Within the command line (e.g., via the Bash terminal on Mac OSX).

440

441 All the scripts are in the folder 'PTP-master', located within the 'bin' folder. Place your input file

442 (ML tree from PhyML, IQtree or RAxML in Newick format named 'MLtree.tre') in the 'bin' folder,

443 in Bash mode, source your file and run bPTP. A list of available options is obtained with:

444

```
445 python3 bPTP.py
```

446

447 To run the analysis:

448

```
449 python3 bPTP.py -t MLtree.tre -o MLtree_output -s 12345678 -r -i 1000000 -n
```

```
450 1000 -b 0.3
```

451

452 Output files include the list of MOTUs and individuals with statistical support

453 'MLtree\_output.PTPSupportPartition.txt', annotated tree with MOTUs and support in

454 'MLtree\_output.PTPSupportPartition.txt.png', 'MLtree\_output.PTPSupportPartition.txt.svg',

455 and 'MLtree\_output.PTPSupportPartition.txt.sh.tre'. bPTP also performs the single threshold

456 ML version of PTP and results are provided as in files 'MLtree\_output.PTPMLpartition' with

457 extensions .txt, .txt.ml.tre, txt.png and txt.svg.

458

459 2. Maximum Likelihood standalone version of PTP: The Maximum Likelihood implementation

460 of the single and multiple threshold PTP model is available at <https://github.com/Pas->

461 [Kapli/mptp](#) for Linux. The package is expected to run in Mac OS X and Windows if required  
462 Python packages are installed. On Mac OS X, GNU Bison and Flex are required, see  
463 documentation for installing. Prepare the input file (ML tree from PhyML, IQtree or RAxML in  
464 Newick format). Open the ML tree with FigTree (see section 2.3.4), define the root by selecting  
465 the descending branch to the node and click on it, then click on 'Reroot' in the window menu.  
466 Export the tree using 'File' and 'Export Trees', set 'Tree file format' to 'Newick' and set as 'Save  
467 as currently displayed'. Name the output file (MLtree.tre) and save it in 'bin' of the 'mptp-0.204-  
468 ' folder. In bash mode source your file and run PTP. The list of available options is obtained with:

469

470 `mptp`

471

472 To run the analysis with multiple thresholds and a tree rooted with the taxa 'taxonA' and  
473 'taxonB':

474

```
475 mptp --ml --multi --tree_file MLtree.tre --output_file MLtree_output_mptp --  
476 outgroup taxonA, taxonB
```

477

478 Output files include the list of MOTUs and individuals in 'MLtree\_output\_mptp.txt', and  
479 annotated tree with MOTUs and support in 'MLtree\_output\_mptp.txt'. To run the single  
480 threshold version, replace the argument `--multi` with `-single` and specify the output file  
481 accordingly `--output_file MLtree_output_sptp`.

482

483 2. Bayesian and Maximum Likelihood version using web server: The Bayesian and ML versions  
484 of the single threshold PTP model can be run online at <https://species.h-its.org/>. Upload your  
485 ML tree 'MLtree.tre' with 'My phylogenetic input tree', select tree type 'unrooted' or 'rooted',

486 set 'No. MCMC generations' (default is 100,000), set the interval for sampling tree with  
487 'Thinning' (default is 100), set 'Burn-in' (default is 0.1), and set 'Seed' (default 123). Thinning is  
488 important to reduce autocorrelation within chains: a thinning value of 100 corresponds to  
489 sampling every hundredth observation. If the tree is unrooted, the tree will be rooted at the  
490 longest branch. If the tree is rooted, set outgroups with 'outgroup taxa names' and provide your  
491 email address. Results include the annotated tree 'output in SVG' and 'annotated tree', and  
492 delimitation results for bPTP and PTP.

493

### 494 3.6 MOTUs informed multi-species coalescent reconstructions with StarBEAST2:

495 1. Preparing the input file with BEAUTY: Mixed diversification (Yule, birth-death) and coalescent  
496 models are available for phylogenetic inferences in StarBEAST2 [55], which is available as a  
497 package for BEAST2. Open BEAUTY and choose 'File' and 'Manage packages'. Select the  
498 StarBEAST2 package and install it by clicking 'Install/Upgrade'. Several relaxed clock models are  
499 available, including random local clock (RLC), uncorrelated exponential clock (UCED) and  
500 uncorrelated lognormal clock (UCLN), which will be used here. Choose 'File' and 'Template', and  
501 select 'SpeciesTreeUCLN'. Import the initial alignment containing all sequences in FASTA format  
502 (alignment.fasta) with 'File' and 'Import alignment'. Now that MOTUs have been delimited, they  
503 can be declared to separate the diversification and coalescent component in 'Taxon sets' in the  
504 window menu. MOTUs can be defined by each of the delimitation methods described  
505 previously or by the majority rule consensus of several methods. MOTUs can be declared using  
506 sequence labels with the option 'Guess'. If sequence labels have been organized as  
507 'Genus\_species1\_MOTU01', click on the button 'guess', select 'split of character' and '\_', and  
508 set '3' in 'take group(s)'. 'Species/Population' will be defined based on the characters after the  
509 second '\_' . Set 'Gene Ploidy' to 0.5 if mitochondrial sequences are analyzed, and use default

510 parameters for 'Population Model'. Define the substitution model at 'Site Model' in the window  
511 menu. The number of available nucleotide substitution models is more limited than in BEAST2.  
512 If the most likely model selected by jModelTest2, ModelTest-NG, or IQtree (see section 2.2) is  
513 available, select it in the 'Subst Model' panel. If not, select 'GTR' model and use jModelTest2  
514 parameter estimates for GTR (see section ), which are accessible in the 'Model Optimization  
515 Results' table of the html log file saved in 'log' folder. It is recommended to use estimates  
516 provided by other methods and avoid estimating substitution model parameters jointly with  
517 tree topology age estimates, as overparameterization may prevent MCMC from initiating. Set  
518 'Gamma Category Count' to '4', and use ML estimates for 'Shape', and 'Proportion of Invariant'  
519 without clicking on estimate. In the 'Clock Model' section of the window menu, define  
520 'Clock.rate' (e.g. 0.01 for 1% of genetic divergence per million years). Use default parameter in  
521 the 'Priors' section, and set the MCMC to 'Chain Length' of 50,000,000, set 'Store Every' to 5,000  
522 and set 'trace log' to 5,000. Save with 'File' and 'Save as' (e.g. SpTree\_BEAUTY).

523

524 2. Check and combine MCMC, reconstruct the consensus: Open BEAST2, and select the input  
525 file (xml file) with 'Choose File' in 'BEAST XML File'. Set 'default: only write new log files' to avoid  
526 overwriting results and run the MCMC. Once done, create a new folder (e.g. RUN1), place all  
527 the output files (alignment.xml.state, starbeast.log, alignment.trees and species.trees) and a  
528 run a second MCMC. Open Tracer to visualize the MCMC traceplots. In the 'Trace files' section,  
529 click on '+' to add a new traceplot, select the log file (input\_file.log) and repeat with the log file  
530 of the second run. Estimated mean parameter value and ESS are indicated in the left panel.  
531 Browse results and check for stability and convergence of the two chains as indicated in section  
532 2.3.3. MCMC runs can be combined with LogCombiner following steps as described in section  
533 2.3.4, excepting that SpeciesTreeUCLN produces species and gene trees, and the whole

534 procedure to construct a consensus tree with LogCombiner and TreeAnnotator should be  
535 repeated for the species tree (species.trees) and the gene tree (alignment.trees). Each  
536 Maximum Credibility Tree can be further viewed with FigTree.

537

#### 538 **4. Notes**

539 1. Each of these methods has different properties regarding reliability, availability, scalability,  
540 understandability and usability (Table 1), which will determine which method will be favored  
541 and when. Ideally, any method should have high reliability, wide availability, broad scalability,  
542 extensive understandability, and global usability. A similar ranking scheme was conducted by  
543 Hleap et al. [56] in the context of specimen identification using metabarcoding for commonly  
544 used algorithms like the Basic Local Alignment Search Tool (BLAST) [57]. However, trade-offs  
545 are expected (e.g., scalability requires fast computations, reliability requires computationally-  
546 intensive calculations). As such, distance-based methods such as ABGD and ASAP have excellent  
547 availability (as webserver and raw C code), ample scalability (not computationally intensive)  
548 and large understandability (based on barcode gap recognition) but low reliability (no  
549 estimation of statistical support in the form of confidence estimates) and narrow usability (as a  
550 criterion for decision-making). Regarding BINs, the RESL algorithm is proprietary and currently  
551 remains unpublished. Furthermore, BINs are not static as their boundaries can shift once new  
552 sequences are submitted to BOLD (hence the method's low availability and low  
553 understandability) [36, 58]. In contrast, BINs require no resources as the framework is regularly  
554 run on the entire BOLD reference library (thus the framework has high scalability and high  
555 usability). The choice of which approach to favor and when is largely dependent on the  
556 computational resources available and the bioinformatic skills required of end users.

557

558 2. Computing a majority-rule consensus among several methods is currently an accessible way  
559 to circumvent individual pitfalls of each approach, as it is balanced by the properties of others  
560 regarding heterogeneous substitution/diversification rates, or uneven sampling [12, 13, 28, 36].  
561 However, each of these packages has adopted its own format for outputting files, which makes  
562 the establishment of the consensus a tedious task if data are formatted manually. The recent  
563 establishment of a universal format for species partitions, SPART [59], to ease data  
564 exchangeability and software to handle partition comparisons, such as LIMES [60], has opened  
565 new perspectives in terms of accessibility (see this volume). As of the date of publication of this  
566 chapter, ABGD, ASAP, PTP and GMYC already implement output files in SPART format.

567  
568 3. Single-locus delimitation of MOTUs in animals is largely conducted using mitochondrial  
569 markers, particularly COI, in the context of DNA barcoding, due to their ease of amplification,  
570 alignment, and sequencing thanks to their high copy number within cells, the wealth of  
571 available sequences and primers in public repositories, as well as low rates of recombination  
572 and high mutation rates. However, mitochondrial markers are maternally inherited, which  
573 accounts for their fast evolutionary rate due to the increased effect of genetic drift, meaning  
574 species-specific substitutions reach fixation within populations much more rapidly than other  
575 loci. Despite these desirable characteristics, the use of mitochondrial markers also has its limits.  
576 For a period following the disruption of gene flow between two lineages, taxa will still share  
577 polymorphisms by ancestry (Fig. 4). Due to the incomplete sorting of lineages, and even  
578 hybridization, species trees will not necessarily equate to gene trees, thereby complicating  
579 downstream recognition of monophyletic clusters that are diagnostic of species, particularly in  
580 a context of a uniparentally-inherited marker. Additionally, GMYC, ABGD, ASAP, PTP and all  
581 their variants do not use secondary lines of evidence, calling for caution when interpreting them

582 in a taxonomic context. However, unlike ABGD, ASAP produces a taxon partition score (i.e.,  
583 asap-score) based on observed barcode gap widths, along with the presumption of panmixia  
584 within species, which could aid decision making in the absence of detailed specimen  
585 examination. A similar index is employed by RESL in the form of the Silhouette Index when  
586 assigning sequences to BINs. In any case, external sources of evidence (e.g., nuclear markers,  
587 geography, morphology, ecology, and behaviour) are required to support putative clusters  
588 reminiscent of actual biological species. In such an integrative framework, MOTUs present a  
589 powerful lever to circumvent the taxonomic impediment through serving as primary  
590 hypotheses for species delineation. The framework detailed in this chapter has proven to  
591 speed-up taxonomy routines [8–10].

592

593 4. Species partitions suggested by the outlined delimitation methods are hypotheses and thus  
594 conditional on the extent of specimen sampling and haplotype coverage across the known  
595 geographic range of a species. In a context of spatially structured populations and isolation by  
596 distance, spatially restricted sampling will cause the maximum genetic distance to be  
597 underestimated within species, and the minimum genetic distance to be overestimated among  
598 species [33, 61]. Species delimitation methods based on the detection of a barcode gap using  
599 ABGD and ASAP will likely overestimate the number of MOTUs in this particular case.  
600 Alternatively, when large spatial scales are involved, as seen in oceanic organisms for instance,  
601 sampling from the most distant sites across species' range distributions may result in missing  
602 intermediate haplotypes and taxonomic oversplitting [36]. DNA barcoding uses practical sample  
603 sizes of 5-10 specimens per species, but many taxa in BOLD are only represented by singletons  
604 or doubletons, making barcode gap estimation and delimitation with distance-based  
605 approaches unreliable. Nonparametric methods like ABGD and ASAP require more data

606 because such methods make fewer assumptions regarding statistical distributions of genetic  
607 diversity and evolutionary history of species, since they use only the DNA sequences themselves  
608 for inference. In contrast, parametric approaches like GMYC and PTP are more robust because  
609 they make stronger assumptions about speciation and distribution of genetic diversity since  
610 they use a speciation model and require tree-building with an a priori model of nucleotide  
611 substitution. Although combining multiple delimitation algorithms into a majority rule  
612 consensus limits the impact of biased sampling on the most sensitive methods, it is no  
613 replacement for a comprehensive sampling of intraspecific diversity.

614

#### 615 **Acknowledgments**

616 We sincerely thank Rob DeSalle for the invitation to contribute to this important volume. This  
617 work was supported by IRD through incentive and recurrent fundings to NH.

618

#### 619 **5. References**

- 620 1. Hebert PDN, Cywinska A, Ball SL, de Waard JR (2003) Biological identifications through  
621 DNA barcodes. *Proc R Soc London Ser B* 270:313–321
- 622 2. Kerr KC, Stoeckle MY, Dove CJ, et al (2007) Comprehensive DNA barcode coverage of  
623 North American birds. *Mol Ecol Notes* 7:535–543
- 624 3. April J, Mayden L. R, Hanner RH, Bernatchez L (2011) Genetic calibration of species  
625 diversity among North America’s freshwater fishes. *Proc Natl Acad Sci USA* 108:10602–  
626 10607
- 627 4. Hubert N, Hanner R (2015) DNA barcoding, species delineation and taxonomy: a  
628 historical perspective. *DNA Barcodes* 3:44–58
- 629 5. Janzen DH, Hajibabaei M, Burns JM, et al (2005) Wedding biodiversity inventory of a

630 large and complex lepidoptera fauna with DNA barcoding. *Philos Trans R Soc Ser B*

631 360:1835–1845

632 6. Smith AM, Rodriguez JJ, Whitfield JB, et al (2008) Extreme diversity of tropical

633 parasitoid wasps exposed by iterative integration of natural history, DNA barcoding,

634 morphology, and collections. *Proc Natl Acad Sci USA* 105:12359–12364

635 7. Smith AM, Fisher BL, Hebert PDN (2005) DNA barcoding for effective biodiversity

636 assessment of a hyperdiverse arthropod group: the ants of Madagascar. *Philos Trans R*

637 *Soc Ser B* 360:1825–1834

638 8. Butcher BA, Smith MA, Sharkey MJ, Quicke DLJ (2012) A turbo-taxonomic study of Thai

639 *Aleiodes* (*Aleiodes*) and *Aleiodes* (*Arcaleiodes*) (Hymenoptera: Braconidae: Rogadinae)

640 based largely on COI barcoded specimens, with rapid descriptions of 179 new species.

641 *Zootaxa* 3457:1–232

642 9. Riedel A, Sagata K, Suhardjono YR, et al (2013) Integrative taxonomy on the fast track -

643 towards more sustainability in biodiversity research. *Front Zool* 10:15

644 10. Sharkey MJ, Janzen DH, Hallwachs W, et al (2021) Minimalist revision and description of

645 403 new species in 11 subfamilies of Costa Rican braconid parasitoid wasps, including

646 host records for 219 species. *Zookeys* 1–666

647 11. Blaxter M (2003) Molecular systematics: counting angels with DNA. *Nature* 421:122–

648 124

649 12. Sholihah A, Delrieu-Trottin E, Condamine FL, et al (2021) Impact of Pleistocene Eustatic

650 Fluctuations on Evolutionary Dynamics in Southeast Asian Biodiversity Hotspots. *Syst*

651 *Biol* 70:940–960. <https://doi.org/10.1093/sysbio/syab006>

652 13. Utami CY, Sholihah A, Condamine FL, et al (2022) Cryptic diversity impacts model

653 selection and macroevolutionary inferences in diversification analyses. *Proc R Soc B*

- 654 289:20221335
- 655 14. Bickford D, Lohman DJ, Sodhi NS, et al (2007) Cryptic species as a window on diversity  
656 and conservation. *Trends Ecol Evol* 22:148–155
- 657 15. Lohman DJ, Ingram KK, Prawiradilaga DM, et al (2010) Cryptic genetic diversity in  
658 “widespread” Southeast Asian bird species suggests that Philippine avian endemism is  
659 gravely underestimated. *Biol Conserv* 143:1885–1890
- 660 16. Barley AJ, Brown JM, Thomson RC (2018) Impact of model violations on the inference of  
661 species boundaries under the multispecies coalescent. *Syst Biol* 67:269–284
- 662 17. Chambers EA, Hillis DM (2020) The multispecies coalescent over-splits species in the  
663 case of geographically widespread taxa. *Syst Biol* 69:184–193
- 664 18. Sukumaran J, Knowles LL (2017) Multispecies coalescent delimits structure, not species.  
665 *Proc Natl Acad Sci* 114:1607–1612
- 666 19. Puillandre N, Lambert A, Brouillet S, Achaz G (2012) ABGD, Automatic Barcode Gap  
667 Discovery for primary species delimitation. *Mol Ecol* 21:1864–1877
- 668 20. Puillandre N, Brouillet S, Achaz G (2021) ASAP: assemble species by automatic  
669 partitioning. *Mol Ecol Resour* 21:609–620
- 670 21. Ratnasingham S, Hebert PDN (2007) BOLD: The Barcode of Life Data System  
671 ([www.barcodinglife.org](http://www.barcodinglife.org)). *Mol Ecol Notes* 7:355–364. [https://doi.org/10.1111/j.1471-](https://doi.org/10.1111/j.1471-8286.2006.01678.x)  
672 [8286.2006.01678.x](https://doi.org/10.1111/j.1471-8286.2006.01678.x)
- 673 22. Ratnasingham S, Hebert PDN (2013) A DNA-Based Registry for All Animal Species: The  
674 Barcode Index Number (BIN) System. *PLoS One* 8:.  
675 <https://doi.org/10.1371/journal.pone.0066213>
- 676 23. Fujiwasa T, Barraclough TG (2013) Delimiting species using single-locus data and the  
677 generalized mixed yule coalescent approach: a revised method and evaluation on

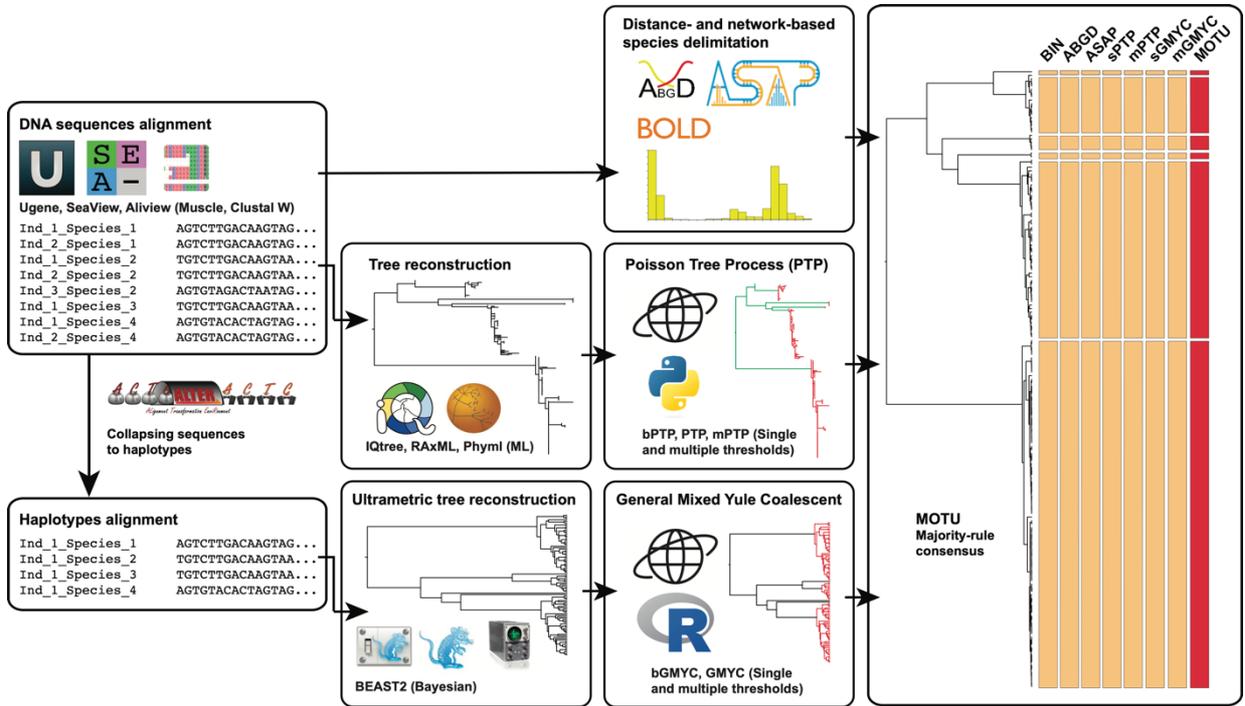
- 678 simulated data sets. *Syst Biol* 62:707–724
- 679 24. Pons J, Barraclough TG, Gomez-Zurita J, et al (2006) Sequence-based species  
680 delimitation for the DNA taxonomy of undescribed insects. *Syst Biol* 55:595–606
- 681 25. Zhang J, Kapli P, Pavlidis P, Stamatakis A (2013) A general species delimitation method  
682 with applications to phylogenetic placements. *Bioinformatics* 29:2869–2876.  
683 <https://doi.org/10.1093/bioinformatics/btt499>
- 684 26. Kapli P, S. L, Zhang J, et al (2017) Multi-rate Poisson Tree Processes for single-locus  
685 species delimitation under Maximum Likelihood and Markov Chain Monte Carlo.  
686 *Bioinformatics* 33:
- 687 27. Arida E, Ashari H, Dahrudin H, et al (2021) Exploring the vertebrate fauna of the Bird’s  
688 Head Peninsula (Indonesia, West Papua) through DNA barcodes. *Mol Ecol Resour*  
689 21:2369–2387
- 690 28. Shen Y, Hubert N, Huang Y, et al (2019) DNA barcoding the ichthyofauna of the Yangtze  
691 River: insights from the molecular inventory of a mega-diverse temperate fauna. *Mol*  
692 *Ecol Resour* 19:1278–1291
- 693 29. Kekkonen M, Mutanen M, Kaila L, et al (2015) Delineating Species with DNA Barcodes:  
694 A Case of Taxon Dependent Method Performance in Moths. *PLoS One* 10:e0122481.  
695 <https://doi.org/10.1371/journal.pone.0122481>
- 696 30. Kekkonen M, Hebert PDN (2014) DNA barcode-based delineation of putative species:  
697 efficient start for taxonomic workflows. *Mol Ecol Resour* 14:706–715
- 698 31. Blair C, Bryson JRW (2017) Cryptic diversity and discordance in single-locus species  
699 delimitation methods within horned lizards (Phrynosomatidae: Phrynosoma). *Mol Ecol*  
700 *Resour* 17:1168–1182
- 701 32. Miralles A, Vences M (2013) New metrics for comparison of taxonomies reveal striking

- 702 discrepancies among species delimitation methods in *Madascincus* lizards. *PLoS ONE*  
703 8:e68242
- 704 33. Chen W, Hubert N, Li Y, et al (2022) Large-scale DNA barcoding of the subfamily  
705 Culterinae (Cypriniformes: Xenocyprididae) in East Asia unveils a geographical scale  
706 effect, taxonomic warnings and cryptic diversity. *Mol Ecol* 31:3871–3887
- 707 34. Geiger MF, Herder . F, Monaghan MT, et al (2014) Spatial heterogeneity in the  
708 Mediterranean Biodiversity Hotspot affects barcoding accuracy of its freshwater fishes.  
709 *Mol Ecol Resour* 14:1210–1221
- 710 35. Arhens D, Fujisawa T, Krammer HJ, et al (2016) Rarity and incomplete sampling in DNA-  
711 based species delimitation. *Syst Biol* 65:478–494
- 712 36. Delrieu-Trottin E, Durand J, Limmon G, et al (2020) Biodiversity inventory of the grey  
713 mullets (Actinopterygii: Mugilidae) of the Indo-Australian Archipelago through the  
714 iterative use of DNA-based species delimitation and specimen assignment methods.  
715 *Evol Appl* 13:1451–1467
- 716 37. Limmon G, Delrieu-Trottin E, Patikawa J, et al (2020) Assessing species diversity of Coral  
717 Triangle artisanal fisheries: A DNA barcode reference library for the shore fishes  
718 retailed at Ambon harbor (Indonesia). *Ecol Evol* 10:3356–3366
- 719 38. Okonechnikov K, Golosova O, Fursov M, et al (2012) Unipro UGENE: A unified  
720 bioinformatics toolkit. *Bioinformatics* 28:1166–1167.  
721 <https://doi.org/10.1093/bioinformatics/bts091>
- 722 39. Larsson A (2014) AliView: a fast and lightweight alignment viewer and editor for large  
723 datasets. *Bioinformatics* 30:3276–3278
- 724 40. Gouy M, Guindon S, Gascuel O (2010) SeaView version 4: a multiplatform graphical user  
725 interface for sequence alignment and phylogenetic tree building. *Mol Biol Evol* 27:221–

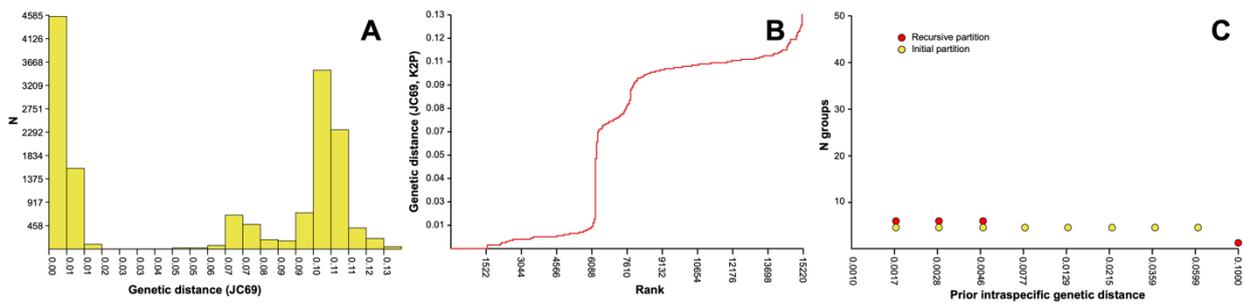
- 726 224
- 727 41. Darriba D, Taboada GL, Doallo R, Posada D (2012) jModelTest 2: more models, new  
728 heuristics and parallel computing. *Nat Methods* 9:772
- 729 42. Guindon S, Dufayard J-F, Lefort V, et al (2010) New algorithms and methods to estimate  
730 maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol*  
731 59:307–321
- 732 43. Trifinopoulos J, Nguyen L-T, von Haeseler A, Minh BQ (2016) W-IQ-TREE: a fast online  
733 phylogenetic tool for maximum likelihood analysis. *Nucleic Acids Res* 44:W232–W235
- 734 44. Nguyen L-T, Schmidt HA, Von Haeseler A, Minh BQ (2015) IQ-TREE: a fast and effective  
735 stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol*  
736 32:268–274
- 737 45. Stamatakis A (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis  
738 of large phylogenies. *Bioinformatics* 30:1312–1313
- 739 46. Darriba D, Posada D, Kozlov AM, et al (2020) ModelTest-NG: a new and scalable tool for  
740 the selection of DNA and protein evolutionary models. *Mol Biol Evol* 37:291–294
- 741 47. Edler D, Klein J, Antonelli A, Silvestro D (2021) raxmlGUI 2.0: A graphical interface and  
742 toolkit for phylogenetic analyses using RAxML. *Methods Ecol Evol* 12:373–377.  
743 <https://doi.org/https://doi.org/10.1111/2041-210X.13512>
- 744 48. Talavera G, Dinca V, Vila R (2013) Factors affecting species delimitations with the GMYC  
745 model: insights from a butterfly survey. *Methods Ecol Evol* 4:1101–1110
- 746 49. Bouckaert R, Heled J, Kühnert D, et al (2014) BEAST 2: A Software Platform for Bayesian  
747 Evolutionary Analysis. *PLoS Comput Biol* 10:1–6.  
748 <https://doi.org/10.1371/journal.pcbi.1003537>
- 749 50. Rambaut A, Drummond AJ, Xie D, et al (2018) Posterior summarization in Bayesian

- 750 phylogenetics using Tracer 1.7. *Syst Biol* 67:901–904.
- 751 <https://doi.org/10.1093/sysbio/syy032>
- 752 51. Jukes TH, Cantor CR (1969) Evolution of protein molecules. *Mamm protein Metab* 3:21–
- 753 132
- 754 52. Kimura M (1980) A simple method for estimating evolutionary rate of base substitutions
- 755 through comparative studies of nucleotide sequences. *J Mol Evol* 15:111–120
- 756 53. Allaire J (2012) RStudio: integrated development environment for R. Boston, MA
- 757 770:165–171
- 758 54. Fujita MK, Leaché AD, Burbrink FT, et al (2012) Coalescent-based species delimitation in
- 759 an integrative taxonomy. *Trends Ecol Evol* 27:480–488
- 760 55. Ogilvie HA, Bouckaert RR, Drummond AJ (2017) StarBEAST2 brings faster species tree
- 761 inference and accurate estimates of substitution rates. *Mol Biol Evol* 34:2101–2114.
- 762 <https://doi.org/10.1093/molbev/msx126>
- 763 56. Hleap JS, Littlefair JE, Steinke D, et al (2021) Assessment of current taxonomic
- 764 assignment strategies for metabarcoding eukaryotes. *Mol Ecol Resour* 21:2190–2203
- 765 57. Altschul SF, Gish W, Miller W, et al (1990) Basic local alignment search tool. *J Mol Biol*
- 766 215:403–410
- 767 58. Phillips JD, Gillis DJ, Hanner RH (2022) Lack of Statistical Rigor in DNA Barcoding Likely
- 768 Invalidates the Presence of a True Species' Barcode Gap. *Front Ecol Evol* 10 859099 doi
- 769 103389/fevo
- 770 59. Miralles A, Ducasse J, Brouillet S, et al (2022) SPART: A versatile and standardized data
- 771 exchange format for species partition information. *Mol Ecol Resour* 22:430–438
- 772 60. Ducasse J, Ung V, Lecointre G, Miralles A (2020) LIMES: a tool for comparing species
- 773 partition. *Bioinformatics* 36:2282–2283

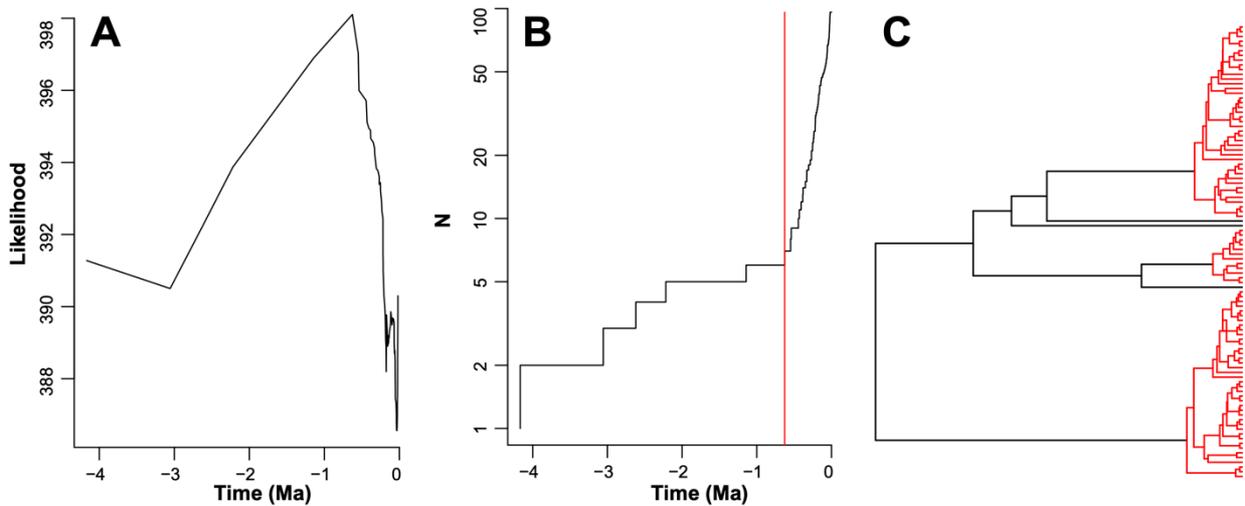
774 61. Bergsten J, Bilton DT, Fujisawa T, et al (2012) The effect of geographical scale of  
 775 sampling on DNA barcoding. Syst Biol 61:851–869



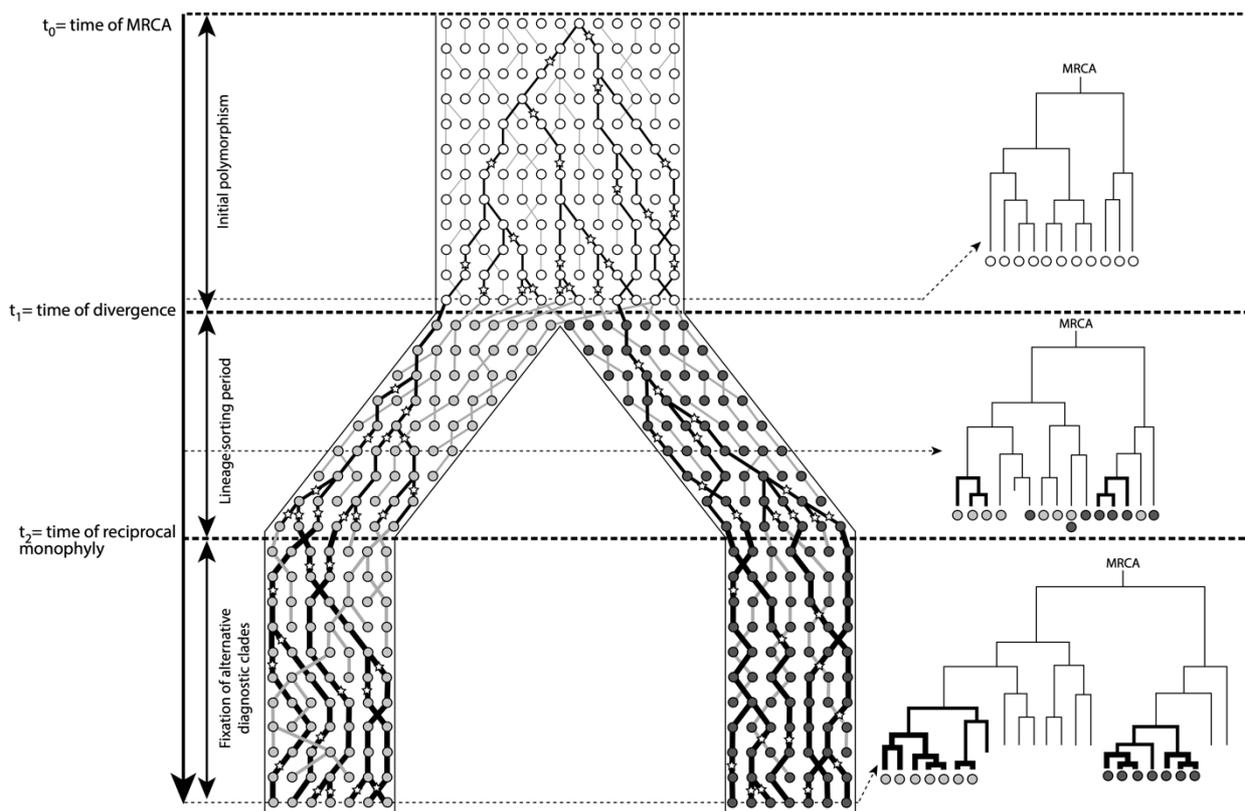
776



777



778



779

780 **Figures captions**

781 Fig. 1 Species delimitation workflow described in this chapter, including DNA sequence  
 782 alignment and data preparation, distance-based methods of species delimitation, phylogenetic  
 783 tree reconstruction and tree-based methods of species delimitation.

784 Fig. 2 Example of outputs from the distance-based methods ABGD and ASAP. (A) histogram of  
 785 genetic distances; (B) histogram of ranked distances; (C) distribution of the number of partitions  
 786 (initial and recursive) according to genetic distances (ABGD). Results were produced using the  
 787 dataset DS-BARBONYM in BOLD ([dx.doi.org/10.5883/DS-BARBONYM](https://dx.doi.org/10.5883/DS-BARBONYM)).

788 Fig. 3 Example of outputs from the tree-based method GMYC. (A) distribution of likelihood  
 789 according to time (million years ago); (B) cumulated number of lineages through time (million  
 790 years ago); (C) annotated ultrametric tree (BEAST2) with branches within MOTUs highlighted in  
 791 red. Results produced using the dataset DS-BARBONYM in BOLD ([dx.doi.org/10.5883/DS-](https://dx.doi.org/10.5883/DS-BARBONYM)  
 792 BARBONYM).

793 Fig. 4 Line of descent of mitochondrial genes between two lineages during their divergence,  
794 including the initial polymorphism prior to divergence, the lineage sorting period, and the  
795 fixation of alternative diagnostic clades. Stars represent mutation events leading to new  
796 haplotypes, circles represent individuals (white for ancestral population, light and dark grey for  
797 diverging lineages).

798

	<b>ABGD</b>	<b>ASAP</b>	<b>BIN</b>	<b>PTP</b>	<b>GMYC</b>
<b>Input data</b>	DNA sequence alignment or matrix of genetic distance	DNA sequence alignment or matrix of genetic distance	DNA sequence alignment	Phylogram	Ultrametric tree
<b>Support</b>	webservice	webservice	BOLD	webservice/Python code	webservice/R package
<b>Reliability</b>	+	++	++	+++	+++
	No estimation of statistical support for partition schemes and groups	Statistical support estimated for partition schemes and groups	No estimation of statistical support for partition schemes and groups but performed on the global COI library in BOLD	Statistical support estimated for each partition/node in the non-ultrametric tree	Statistical support estimated for each partition/node in the ultrametric (i.e., time-calibrated) tree
<b>Availability</b>	+++ Webservice available; C code available by request	+++ Webservice available; C code available by request	+ RESL algorithm proprietary; BOLD-registered users only; requires a personal project; no stand-alone available	++/+ Webservice version accessible/standalones require compiling C code and installing dependencies (mPTP); input trees need additional software (e.g., RAxML)	++/++ Webservice and standalone available; input trees need additional software (e.g., BEAST)
<b>Scalability</b>	+++ Not computationally intensive; can handle large datasets	+++ Not computationally intensive; can handle large datasets	+++ Run automatically by BOLD; no resources required; existing BINs updated monthly; can handle large datasets	+ /+++ Webservice has limited resources and running the Bayesian implementation is mandatory; ML versions in mPTP standalone are fast	+ /+++ Webservice has limited resources and running the Bayesian implementation is mandatory; ML version in R package 'splits' can handle larger datasets
<b>Understandability</b>	++++ Distance-based method with explicit analytical procedures; recommended for beginners	+++ Distance-based method with explicit analytical procedures; recommended for intermediate users	+ Underlying calculations not easily accessible; recommended for advanced users	+ Underlying calculations not easily accessible; knowledge of statistical probability distributions required; recommended for advanced users	+ Requires knowledge about coalescent, diversification theory, and statistical probability distributions; recommended for advanced users
<b>Usability</b>	+ No criterion available to compare alternative partitioning	+++ Decision-based with a weighted criterion (ASAP-score)	++ Decision-based with a criterion integrated in BOLD (Silhouette index)	+++ Decision-based with a maximum likelihood-based criterion	+++ Decision-based with a maximum likelihood-based criterion

799

800 **Tables captions**

801 Table 1. Properties of ABGD, ASAP, BIN, PTP and GMYC in terms of reliability, availability,  
802 scalability, understandability, and usability. Taxon delimitation methods are ranked from a  
803 value of + (low) to +++++ (high).