Delimiting species with single-locus DNA sequences

Nicolas Hubert, Jarrett D Phillips, Robert H Hanner

Delimiting species with single-locus DNA sequences

Nicolas Hubert[1], Jarrett D. Phillips[2,3], Robert H. Hanner[3]

[1]UMR ISEM (IRD, UM, CNRS), Université de Montpellier, Place Eugène Bataillon, 34095 Montpellier cedex 05, France

[2]School of Computer Science, University of Guelph, 50 Stone Rd E, Guelph, ON N1G2W1, Canada

[3]Department of Integrative Biology, University of Guelph, 50 Stone Rd E, Guelph, ON N1G2W1, Canada

Contact: nicolas.hubert@ird.fr

Running title: DNA-based species delimitation

Abstract

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

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

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

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

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

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

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

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

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

2. Materials

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

2.1 Aligning and formatting DNA barcodes

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

2. AliView: Download the most recent version of AliView [39] at *https://ormbunkar.se/aliview/* compatible with your operating system. Import your input file (alignment.fasta) with the option 'File' and 'Open file' in the upper menu, or right click on your input file with 'open with' and select AliView. Go to 'Align', and 'Realign everything'. MUSCLE is used for alignment by default. 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://doua.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 ‘./phymlPhyML-3.1_macOS-
178 MountailLion’) 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.

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

2.3 Reconstructing an ultrametric tree with BEAST2 for GMYC

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

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

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

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

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

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

tree file (combined.trees), name the output file (MCT.tree) and run. To visualize the MCT and prepare it for GMYC, download FigTree at *https://github.com/rambaut/figtree/releases*, available for Windows, Mac OS X and Linux. Open FigTree and import the MCT file (MCT.tre) with 'File' and 'open' in the upper menu. Export the tree using 'File' and 'Export Trees', set 'Tree file format' to 'Newick' and set as 'Save as currently displayed'. Name the output file (MCT_GMYC.tre) and save. You can further explore the MCT with FigTree by visualizing 95% credibility intervals (select 'Node Bars' and set to 'CAheight_95%_HPD') or posterior probabilities (select 'Node Labels' and set to 'posterior').

3. Methods

3.1 Running distance-based methods:

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

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

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

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

3.3 Delimiting MOTUs with GMYC, single and multiple thresholds

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

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

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

```
> library(ape)
> library(splits)
```

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

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

or more simply

```
> ?ape
> ?splits
```

Note that library("package name") needs to be called every time a new R session is initiated. To import the BEAST2 ultrametric tree prepared for GMYC (MCT_GMYC.tre) and run GMYC with a single threshold use:

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

For the multiple threshold version, use the argument `method = "multiple"`. Details are available using:

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

or

```
> ?gmyc
```

Results can be collected using:

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

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

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

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

3.5 Delimiting MOTUs with PTP

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

```
Pip3 install -r requirements.txt
Python3 steup.py install
```

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

All the scripts are in the folder 'PTP-master', located within the 'bin' folder. Place your input file (ML tree from PhyML, IQtree or RAxML in Newick format named 'MLtree.tre') in the 'bin' folder, in Bash mode, source your file and run bPTP. A list of available options is obtained with:

```
python3 bPTP.py
```

To run the analysis:

```
python3 bPTP.py -t MLtree.tre -o MLtree_output -s 12345678 -r -i 1000000 -n
1000 -b 0.3
```

Output files include the list of MOTUs and individuals with statistical support 'MLtree_output.PTPhSupportPartition.txt', annotated tree with MOTUs and support in 'MLtree_output.PTPhSupportPartition.txt.png', 'MLtree_output.PTPhSupportPartition.txt.svg', and 'MLtree_output.PTPhSupportPartition.txt.sh.tre'. bPTP also performs the single threshold ML version of PTP and results are provided as in files 'MLtree_output.PTPMLpartition' with extensions .txt, .txt.ml.tre, txt.png and txt.svg.

2. Maximum Likelihood standalone version of PTP: The Maximum Likelihood implementation 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',

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

3.6 MOTUs informed multi-species coalescent reconstructions with StarBEAST2:

1. Preparing the input file with BEAUTY: Mixed diversification (Yule, birth-death) and coalescent models are available for phylogenetic inferences in StarBEAST2 [55], which is available as a package for BEAST2. Open BEAUTY and choose ‘File’ and ‘Manage packages’. Select the StarBEAST2 package and install it by clicking ‘Install/Upgrade’. Several relaxed clock models are available, including random local clock (RLC), uncorrelated exponential clock (UCED) and uncorrelated lognormal clock (UCLN), which will be used here. Choose ‘File’ and ‘Template’, and select ‘SpeciesTreeUCLN’. Import the initial alignment containing all sequences in FASTA format (alignment.fasta) with ‘File’ and ‘Import alignment’. Now that MOTUs have been delimited, they can be declared to separate the diversification and coalescent component in ‘Taxon sets’ in the window menu. MOTUs can be defined by each of the delimitation methods described previously or by the majority rule consensus of several methods. MOTUs can be declared using sequence labels with the option ‘Guess’. If sequence labels have been organized as ‘Genus_species1_MOTU01’, click on the button ‘guess’, select ‘split of character’ and ‘_’, and set ‘3’ in ‘take group(s)’. ‘Species/Population’ will be defined based on the characters after the second ‘_’. Set ‘Gene Ploidy’ to 0.5 if mitochondrial sequences are analyzed, and use default

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

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

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

4. Notes

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

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

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

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

Acknowledgments

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

5. References

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

large and complex lepidoptera fauna with DNA barcoding. Philos Trans R Soc Ser B 360:1835–1845

6. Smith AM, Rodriguez JJ, Whitfield JB, et al (2008) Extreme diversity of tropical parasitoid wasps exposed by iterative integration of natural history, DNA barcoding, morphology, and collections. Proc Natl Acad Sci USA 105:12359–12364

7. Smith AM, Fisher BL, Hebert PDN (2005) DNA barcoding for effective biodiversity assessment of a hyperdiverse arthropod group: the ants of Madagascar. Philos Trans R Soc Ser B 360:1825–1834

8. Butcher BA, Smith MA, Sharkey MJ, Quicke DLJ (2012) A turbo-taxonomic study of Thai Aleiodes (Aleiodes) and Aleiodes (Arcaleiodes) (Hymenoptera: Braconidae: Rogadiniae) based largely on COI barcoded specimens, with rapid descriptions of 179 new species. Zootaxa 3457:1–232

9. Riedel A, Sagata K, Suhardjono YR, et al (2013) Integrative taxonomy on the fast track - towards more sustainability in biodiversity research. Front Zool 10:15

10. Sharkey MJ, Janzen DH, Hallwachs W, et al (2021) Minimalist revision and description of 403 new species in 11 subfamilies of Costa Rican braconid parasitoid wasps, including host records for 219 species. Zookeys 1–666

11. Blaxter M (2003) Molecular systematics: counting angels with DNA. Nature 421:122–124

12. Sholihah A, Delrieu-Trottin E, Condamine FL, et al (2021) Impact of Pleistocene Eustatic Fluctuations on Evolutionary Dynamics in Southeast Asian Biodiversity Hotspots. Syst Biol 70:940–960. https://doi.org/10.1093/sysbio/syab006

13. Utami CY, Sholihah A, Condamine FL, et al (2022) Cryptic diversity impacts model 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). Mol Ecol Notes 7:355–364. https://doi.org/10.1111/j.1471-
672 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 mathod and evaluation on

simulated data sets. Syst Biol 62:707–724

24. Pons J, Barraclough TG, Gomez-Zurita J, et al (2006) Sequence-based species delimitation for the DNA taxonomy of undescribed insects. Syst Biol 55:595–606

25. Zhang J, Kapli P, Pavlidis P, Stamatakis A (2013) A general species delimitation method with applications to phylogenetic placements. Bioinformatics 29:2869–2876. https://doi.org/10.1093/bioinformatics/btt499

26. Kapli P, S. L, Zhang J, et al (2017) Multi-rate Poisson Tree Processes for single-locus species delimitation under Maximum Likelihood and Markov Chain Monte Carlo. Bioinformatics 33:

27. Arida E, Ashari H, Dahruddin H, et al (2021) Exploring the vertebrate fauna of the Bird's Head Peninsula (Indonesia, West Papua) through DNA barcodes. Mol Ecol Resour 21:2369–2387

28. Shen Y, Hubert N, Huang Y, et al (2019) DNA barcoding the ichthyofauna of the Yangtze River: insights from the molecular inventory of a mega-diverse temperate fauna. Mol Ecol Resour 19:1278–1291

29. Kekkonen M, Mutanen M, Kaila L, et al (2015) Delineating Species with DNA Barcodes: A Case of Taxon Dependent Method Performance in Moths. PLoS One 10:e0122481. https://doi.org/10.1371/journal.pone.0122481

30. Kekkonen M, Hebert PDN (2014) DNA barcode-based delineation of putative species: efficient start for taxonomic workflows. Mol Ecol Resour 14:706–715

31. Blair C, Bryson JRW (2017) Cryptic diversity and discordance in single-locus species delimitation methods within horned lizards (Phrynosomatidae: Phrynosoma). Mol Ecol Resour 17:1168–1182

32. Miralles A, Vences M (2013) New metrics for comparison of taxonomies eveal striking

discrepancies among species delimitation methods in Madascincus lizards. PLoSONE 8:e68242

33. Chen W, Hubert N, Li Y, et al (2022) Large-scale DNA barcoding of the subfamily Culterinae (Cypriniformes: Xenocyprididae) in East Asia unveils a geographical scale effect, taxonomic warnings and cryptic diversity. Mol Ecol 31:3871–3887

34. Geiger MF, Herder . F, Monaghan MT, et al (2014) Spatial heterogeneity in the Mediterranean Biodiversity Hotspot affects barcoding accuracy of its freshwater fishes. Mol Ecol Resour 14:1210–1221

35. Arhens D, Fujisawa T, Krammer HJ, et al (2016) Rarity and incomplete sampling in DNA-based species delimitation. Syst Biol 65:478–494

36. Delrieu-Trottin E, Durand J, Limmon G, et al (2020) Biodiversity inventory of the grey mullets (Actinopterygii: Mugilidae) of the Indo-Australian Archipelago through the iterative use of DNA-based species delimitation and specimen assignment methods. Evol Appl 13:1451–1467

37. Limmon G, Delrieu-Trottin E, Patikawa J, et al (2020) Assessing species diversity of Coral Triangle artisanal fisheries: A DNA barcode reference library for the shore fishes retailed at Ambon harbor (Indonesia). Ecol Evol 10:3356–3366

38. Okonechnikov K, Golosova O, Fursov M, et al (2012) Unipro UGENE: A unified bioinformatics toolkit. Bioinformatics 28:1166–1167. https://doi.org/10.1093/bioinformatics/bts091

39. Larsson A (2014) AliView: a fast and lightweight alignment viewer and editor for large datasets. Bioinformatics 30:3276–3278

40. Gouy M, Guindon S, Gascuel O (2010) SeaView version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. Mol Biol Evol 27:221–

224

41. Darriba D, Taboada GL, Doallo R, Posada D (2012) jModelTest 2: more models, new heuristics and parallel computing. Nat Methods 9:772

42. Guindon S, Dufayard J-F, Lefort V, et al (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol 59:307–321

43. Trifinopoulos J, Nguyen L-T, von Haeseler A, Minh BQ (2016) W-IQ-TREE: a fast online phylogenetic tool for maximum likelihood analysis. Nucleic Acids Res 44:W232–W235

44. Nguyen L-T, Schmidt HA, Von Haeseler A, Minh BQ (2015) IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol 32:268–274

45. Stamatakis A (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312–1313

46. Darriba D, Posada D, Kozlov AM, et al (2020) ModelTest-NG: a new and scalable tool for the selection of DNA and protein evolutionary models. Mol Biol Evol 37:291–294

47. Edler D, Klein J, Antonelli A, Silvestro D (2021) raxmlGUI 2.0: A graphical interface and toolkit for phylogenetic analyses using RAxML. Methods Ecol Evol 12:373–377. https://doi.org/https://doi.org/10.1111/2041-210X.13512

48. Talavera G, Dinca V, Vila R (2013) Factors affecting species delimitations with the GMYC model: insights from a butterfly survey. Methods Ecol Evol 4:1101–1110

49. Bouckaert R, Heled J, Kühnert D, et al (2014) BEAST 2: A Software Platform for Bayesian Evolutionary Analysis. PLoS Comput Biol 10:1–6. https://doi.org/10.1371/journal.pcbi.1003537

50. Rambaut A, Drummond AJ, Xie D, et al (2018) Posterior summarization in Bayesian

phylogenetics using Tracer 1.7. Syst Biol 67:901–904. https://doi.org/10.1093/sysbio/syy032

51. Jukes TH, Cantor CR (1969) Evolution of protein molecules. Mamm protein Metab 3:21–132

52. Kimura M (1980) A simple method for estimating evolutionary rate of base subtitutions through comparative studies of nucleotide sequences. J Mol Evol 15:111–120

53. Allaire J (2012) RStudio: integrated development environment for R. Boston, MA 770:165–171

54. Fujita MK, Leaché AD, Burbrink FT, et al (2012) Coalescent-based species delimitation in an integrative taxonomy. Trends Ecol Evol 27:480–488

55. Ogilvie HA, Bouckaert RR, Drummond AJ (2017) StarBEAST2 brings faster species tree inference and accurate estimates of substitution rates. Mol Biol Evol 34:2101–2114. https://doi.org/10.1093/molbev/msx126

56. Hleap JS, Littlefair JE, Steinke D, et al (2021) Assessment of current taxonomic assignment strategies for metabarcoding eukaryotes. Mol Ecol Resour 21:2190–2203

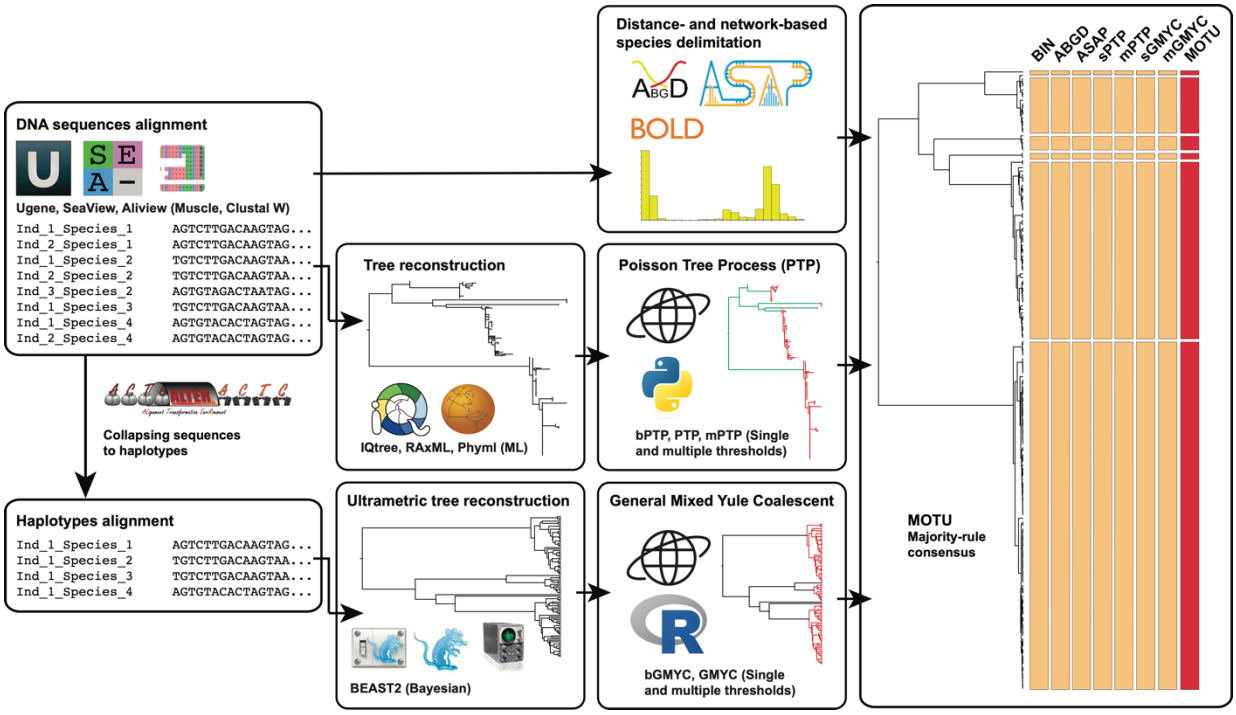
57. Altschul SF, Gish W, Miller W, et al (1990) Basic local alignment search tool. J Mol Biol 215:403–410

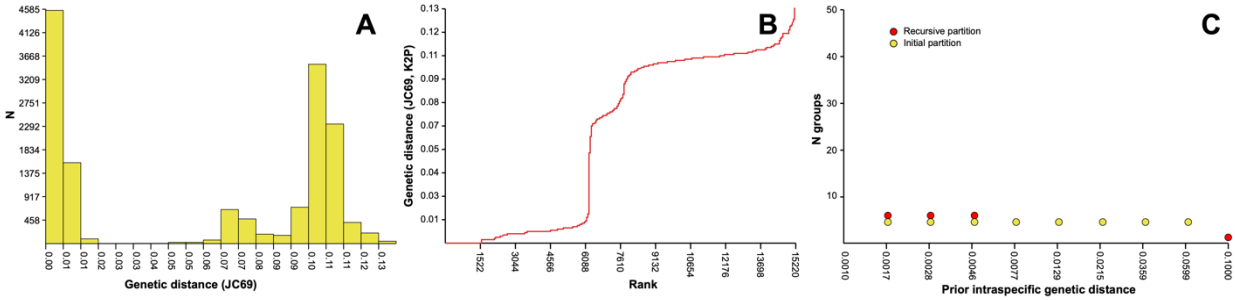
58. Phillips JD, Gillis DJ, Hanner RH (2022) Lack of Statistical Rigor in DNA Barcoding Likely Invalidates the Presence of a True Species' Barcode Gap. Front Ecol Evol 10 859099 doi 103389/fevo

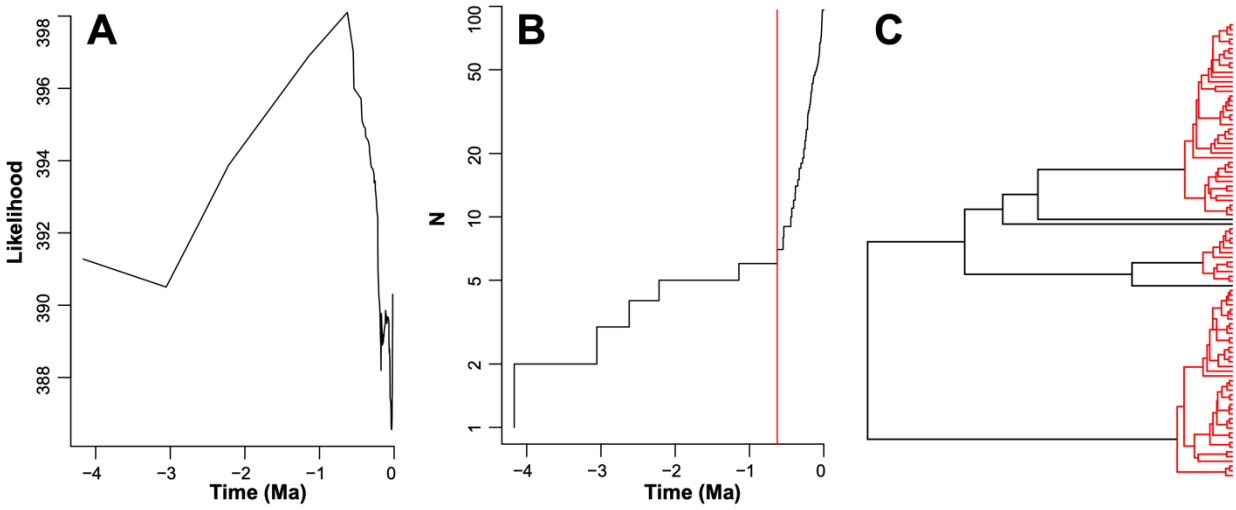
59. Miralles A, Ducasse J, Brouillet S, et al (2022) SPART: A versatile and standardized data exchange format for species partition information. Mol Ecol Resour 22:430–438

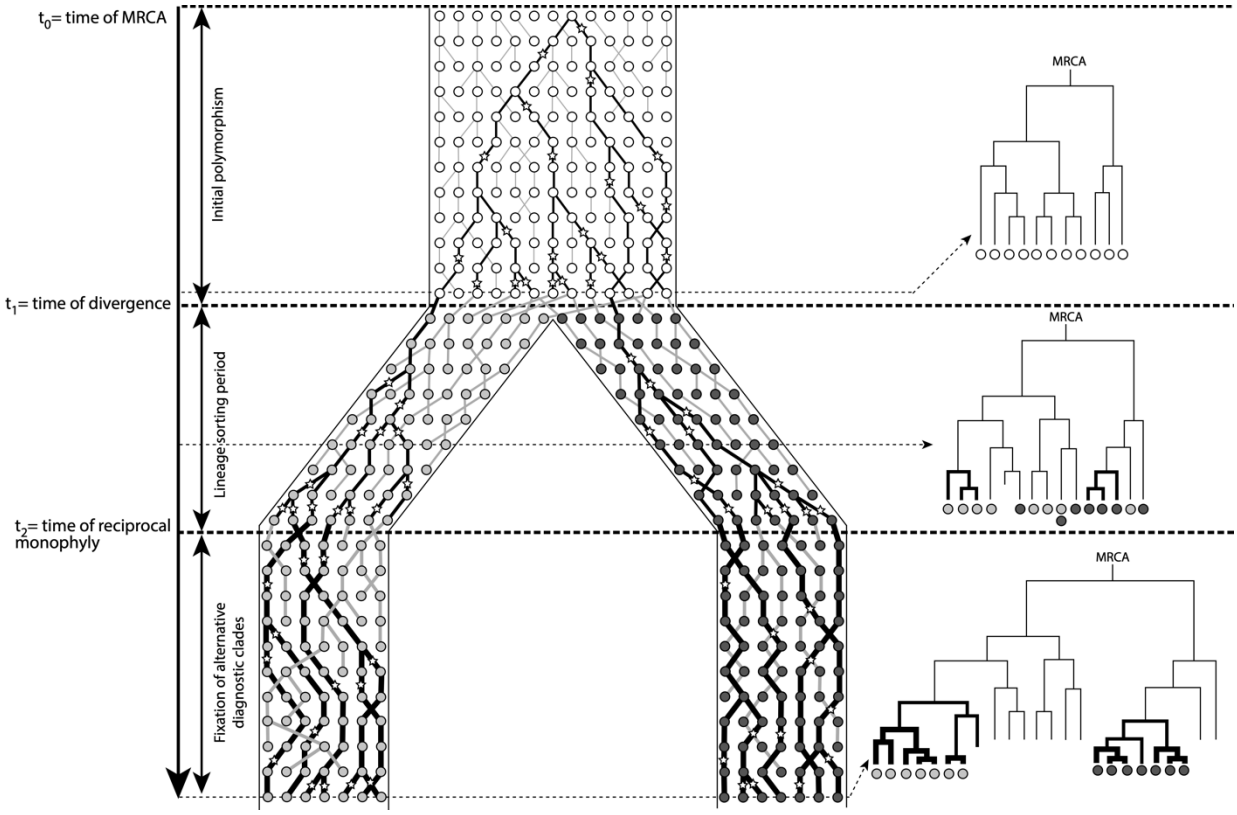
60. Ducasse J, Ung V, Lecointre G, Miralles A (2020) LIMES: a tool for comparing species partition. Bioinformatics 36:2282–2283

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).

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-
792 BARBONYM).

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

	ABGD	ASAP	BIN	PTP	GMYC
Input data	DNA sequence alignment or matrice of genetic distance	DNA sequence alignment or matrice of genetic distance	DNA sequence alignment	Phylogram	Ultrametric tree
Support	webserver	webserver	BOLD	webserver/Python code	webserver/R package
Reliability	+	++	++	+++	+++
	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
Availability	+++	+++	+	++/+	++/++
	Webserver available; C code available by request	Webserver available; C code available by request	RESL algorithm proprietary; BOLD-registered users only; requires a personal project; no stand-alone available	Webserver version accessible/standalones require compiling C code and installing dependencies (mPTP); input trees need additional software (e.g., RAxML)	Webserver and standalone available; input trees need additional software (e.g., BEAST)
Scalability	+++	+++	+++	+/+++	+/+++
	Not computationaly intensive; can handle large datasets	Not computationaly intensive; can handle large datasets	Run automatically by BOLD; no resources required; existing BINs updated monthly; can handle large datasets	Webserver has limited resources and running the Bayesian implementation is mandatory; ML versions in mPTP standalone are fast	Webserver has limited resources and running the Bayesian implementation is mandatory; ML version in R package 'splits' can handle larger datasets
Understandability	++++	+++	+	+	+
	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
Usability	+	+++	++	+++	+++
	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

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).