Journal (2050), Vol. 00, No. 000, pp. 1–10
DOI:xxx000

A pipeline for heuristic species delimitation under the multispecies coalescent model using multilocus sequence data

Daniel Kornai (orcid: 0000-0003-4919-2384) [1], Tomáş Flouri (orcid: 0000-0002-8474-9507) [1], and Ziheng Yang (orcid: 0000-0003-3351-7981) [1,*]

[1]*Department of Genetics, Evolution and Environment, University College London, UK*

Received on xxxx, revised on xxxx, accepted on xxxx

The multispecies coalescent (MSC) model accommodates genealogical variations across the genome and provides a natural framework for comparative analysis of genomic sequence data to infer the history of species divergence and gene flow. Given a set of populations, hypotheses of species delimitation (and species phylogeny) may be formulated as instances of MSC models (e.g., MSC for one species versus MSC for two species) and compared using Bayesian model selection. This approach, implemented in the Bayesian program bpp, has been found to be prone to over-splitting. Alternatively heuristic criteria based on population parameters under the MSC model (such as population/species divergence times, population sizes, and migration rates) estimated from genomic sequence data may be used to delimit species. Here we extend the approaches of Jackson *et al.* (2017) and Leaché *et al.* (2019) based on the genealogical divergence index (*gdi*) and develop hierarchical merge and split algorithms for species delimitation, and implemented them as a python pipeline. Applied to data simulated under a model of isolation by distance, the approach was able to recover the correct species delimitation, whereas model comparison by bpp failed. Analyses of empirical datasets suggested that the procedure may avoid the problem of over-splitting. We discuss possible strategies for accommodating gene flow in the procedure, as well as the challenges of species delimitation based on heuristic criteria.

bpp | genealogical divergence index | multispecies coalescent | species delimitation

Notes

- collect estimates of τ and θ into tables.
- Move all descriptions of MCMC setting etc. into SI. Provide data files in SI.

Introduction

Accurate delimitation of species boundaries is important to characterizing patterns of biological diversity, especially during the current global changes in climate and environment. Traditionally, species have been identified and distinguished using morphological characteristics. Molecular genetic data can provide additional information about many processes related to species delimitation and identification, including population identities, interspecific hybridization and gene flow, and phylogenetic relationships among the populations and their divergence times (Jiao *et al.*, 2021).

Given a set of populations, different species delimitations correspond to different ways of merging populations into the same species. Each species delimitation, combined with the phylogeny for the delimited species, can be formulated as an instance of the multispecies coalescent (MSC) model (Rannala and Yang, 2003) and fitted to genomic sequence data sampled from the modern species or populations. Competing models can then be compared via Bayesian model selection (i.e., using posterior model probabilities or Bayes factors) to find the best supported delimitation. In the Bayesian program bpp, this is accomplished by using a Markov chain Monte Carlo (MCMC) algorithm to estimate the posterior probabilities for different MSC models (Yang and Rannala, 2010; Yang, 2014, 2015; Flouri *et al.*, 2018). In simulations, bpp showed lower rates of species overestimation and underestimation than the generalized mixed Yule-coalescent or Poisson tree processes (Luo *et al.*, 2018). In empirical datasets, bpp was effective in identifying cryptic species in many ancient lineages that were not recognised by other molecular or morphological approaches. For example, Ramirez-Reyes *et al.* (2020) identified 13 new species of leaf-toed geckoes in a lineage that diverged 30 Ma.

However, bpp has been noted to often over-split, identifying more lineages as distinct species than many other methods (Sukumaran and Knowles, 2017). For example, Campillo *et al.* (2020) analyzed 99 population pairs in the genus *Drosophila* and found that bpp identified 80 pairs as distinct species, whereas reproductive isolation was identified in only 69 pairs. Similarly, Bamberger *et al.* (2022) examined 48

*to whom correspondence should be addressed

Albinaria cretensis land snail populations, and found that morphological delimitation ?? suggested 3–9 species, ADMIXTURE ?? suggested 15, while BPP suggested 45-48. Barley *et al.* (2018) simulated multiple populations from a single species that exhibits isolation by distance, and found that BPP delimits geographically separated populations as distinct species. Those results suggest that the lineages identified by BPP sometimes correspond to populations rather than species (Chambers and Hillis, 2020). Multiple studies using BPP have suggested significant taxonomic reassignments not supported using other methods (e.g., Wu *et al.*, 2018 in Yunnan Bananas). A number of authors have expressed concerns about the apparent over-splitting of BPP (MacGuigan *et al.*, 2021).

Rather than treating species delimitation as a model-selection problem, an alternative approach is to estimate population parameters, such as population split time (τ), population sizes (θ), and migration rates (M), and define species status using empirical criteria based on those parameters. For example, the '10× rule' specifies the interspecific divergence to be at least 10 times as large as the intraspecific diversity (Hebert *et al.*, 2004).

Jackson *et al.* (2017) suggested a criterion called the *genealogical divergence index* (*gdi*), defined using population parameters. Consider two sequences (a_1 and a_2) sampled population A and one sequence (b) from B (see fig. 1). Let the probability that the two sequences from population A coalesce first, so that the gene tree is $G_1 = ((a_1, a_2), b)$, be $P_1 = \mathbb{P}(G_1)$. In the case of no gene flow, this is given as

$$P_1 = 1 - \tfrac{2}{3}\,\mathrm{e}^{-2\tau_{AB}/\theta_A}, \qquad (1)$$

This is a simple function of $2\tau_{AB}/\theta_A = T_{AB}/(2N_A)$, the population divergence time in coalescent units (with one coalescent time unit to be $2N_A$ generations in population A). Jackson *et al.* (2017) rescaled P_1 so that the *gdi* ranges from 0 to 1.

$$gdi = 1 - \mathrm{e}^{-2\tau_{AB}/\theta_A} = 1 - \mathrm{e}^{-T_{AB}/(2N_A)}\,. \qquad (2)$$

Thus the *gdi* is the probability that the two A sequences coalesce before reaching species divergence (τ_{AB}) when we trace the genealogy of the sample backwards in time. A *gdi* close to 1 indicates a high level of population divergence. Based on a meta-analysis of data from Pinho and Hey (2010), Jackson *et al.* (2017) suggest that populations are likely to be a single species if $gdi < 0.2$, and separate species if $gdi > 0.7$. Intermediate values ($0.2 < gdi < 0.7$) indicate ambiguous species status.

When there is migration between the two populations, the probability for the gene tree G_1 depends on the parameters of the MSC-M model:

$$P_1 = \mathbb{P}(G_1 \big| \tau_{AB}, \theta_A, \theta_B, \theta_{AB}, M_{AB}, M_{BA}). \qquad (3)$$

Thus the minimum and maximum of P_1 used by Jackson *et al.* (2017) for rescaling P_1 depend on the model parameters. Instead, here we redefine *gdi* as the probability that the first coalescence is between the two A sequences and it occurs before reaching species divergence when we trace the genealogy backwards in time. This definition applies whether or not there is gene flow in the model (fig. 1), with $0 \le gdi \le 1$.

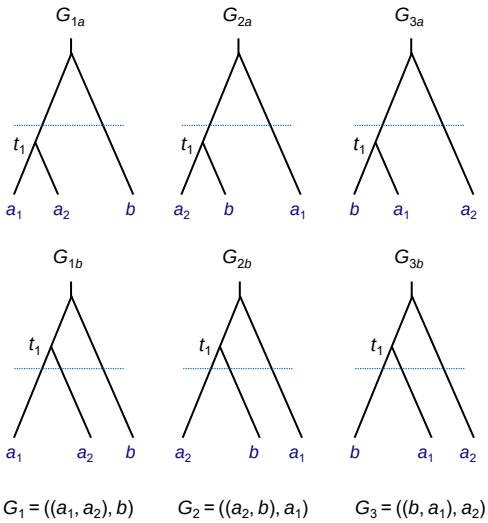
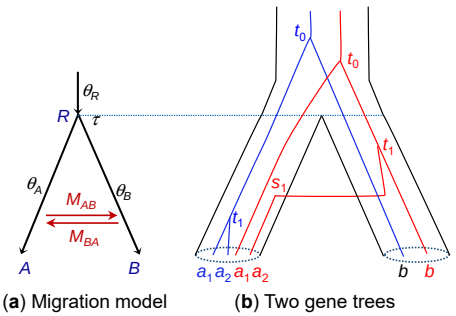


Figure 1: For a locus with two sequences a_1, a_2 from species A and one sequence b from B, there are three possible gene trees: $G_1 = ((a_1, a_2), b)$; $G_2 = ((a_2, b), a_1)$; and $G_3 = ((b, a_2), a_1)$. If the first coalescence time is more recent than the species divergence time (indicated by the dashed line), we label the gene tree as G_{1a}, G_{2a}, G_{3a}; otherwise they are labeled G_{1b}, G_{2b}, G_{3b}. We have $gdi = \mathbb{P}(G_{1a})$. The *gdi* is the probability that two A sequences coalesce first and before the population split. Note that if there is no gene flow between species A and B, gene trees G_{2b} and G_{3b} are impossible.

Leaché *et al.* (2019) described a hierarchical merge algorithm for species delimitation based on *gdi*. Given a set of populations and a guide tree for them, the procedure attempts to merge two populations into one species, judged by *gdi*. Here we develop a python pipeline to automate the procedure. We include a hierarchical split algorithm as well. We first describe the computation of *gdi* when there is gene flow in the model, following Leaché *et al.* (2019). Then we discuss our new pipeline. We apply the pipeline to four empirical datasets, for giraffes, snails, milksnakes, and sunfish.

Computation of gdi under the MSC-M model

Under the MSC-M model, the *gdi* can be computed analytically, using the Markov chain characterization of the backward-in-time process of coalescent and migration (Leaché *et al.*, 2019). For two populations (A and B) with gene flow and three sequences (a_1, a_2, and b), the genealogical process of coalescent and migration when one traces the history of the sample



(**a**) Migration model (**b**) Two gene trees

Figure 2: (**a**) An MSC-with-migration (MSC-M) model for two species or populations (A, B) showing the parameters. The two species diverged time $\tau \equiv \tau_{AB}$ ago and have since been exchanging migrants at the rate of $M_{AB} = m_{AB}N_B$ migrants per generation from A to B and at the rate M_{BA} from B to A. (**b**) Two gene trees at two loci, each with two sequences (a_1, a_2) from population A and one sequence (b) from B. In the blue tree, sequences a_1 and a_2 coalesce first, in population A, resulting in the gene tree $G_1 = ((a_1, a_2), b)$. In the red tree, sequence a_2 migrates into population B and coalesce with sequence b in population B, resulting in the gene tree $G_2 = ((a_2, b), a_1)$. The *gdi* index is defined as the probability that the first coalescence occurs between the two A sequences and before reaching species divergence when we trace the genealogy backwards in time.

backwards in time can be described by a Markov chain. The state of the chain is specified by the number of sequences remaining in the sample and the population IDs (A and B) and the sequence IDs (a_1, a_2, b, etc.). For example, The initial state is $A_{a_1}A_{a_2}B_b$, in which three sequences a_1, a_2, b are in populations A, A, and B, respectively. This is also written as 'AAB'. State $A_{a_1a_2}B_b$, abbreviated 'AB_b', means that two sequences remain in the sample, with the ancestor of sequences a_1 and a_2 in population A and sequence b in population B. There are 21 states in the Markov chain.

The transition rate matrix of the Markov chain $Q = \{q_{ij}\}$ is given in table S1. The transition probability matrix over time t is then $P(t) = \{p_{ij}(t)\} = \mathrm{e}^{Qt}$, where $p_{ij}(t)$ is the probability that the Markov chain is in state j at time t in the past given that it is in state i at time 0 (the present time). Suppose Q has the spectral decomposition

$$q_{ij} = \sum_{k=1}^{21} u_{ik}v_{kj}\lambda_k, \tag{4}$$

where $0 = \lambda_1 > \lambda_2 \geq \cdots \geq \lambda_{21}$ are the eigenvalues of Q, and columns in $U = \{u_{ij}\}$ are the corresponding right eigenvectors, with $V = \{v_{ij}\} = U^{-1}$. Then

$$p_{ij}(t) = \sum_{k=1}^{21} u_{ik}v_{kj}\,\mathrm{e}^{\lambda_k t}\,. \tag{5}$$

Consider the coalescent time t between sequences a_1 and a_2 given that they are to coalesce first and before τ (as in the blue gene tree of fig. 2b). This has density

$$f(t) = \left[p_{AAB,AAA}(t) + p_{AAB,AAB}(t)\right]\tfrac{2}{\theta_A} + \left[p_{AAB,BBA}(t) + p_{AAB,BBB}(t)\right]\tfrac{2}{\theta_B}, \quad t < \tau. \tag{6}$$

The two terms in the sum correspond to coalescence between a_1 and a_2 occurring in populations A and B, respectively. The first term is the probability, $p_{AAB,AAA}(t) + p_{AAB,AAB}(t)$, that sequences a_1 and a_2 are in A right before time t, times the rate for them to coalesce $\left(\frac{2}{\theta_A}\right)$. Similarly the second term is the probability density that a_1 and a_2 coalesce at time t in B.

By averaging over the distribution of t, we have

$$gdi = \int_0^\tau f(t)\,\mathrm{d}t, \tag{7}$$

where $f(t)$ is given in eq. 6. To calculate the integral in eq. 7, note that from eq. 5,

$$\int_0^\tau p_{ij}(t)\,\mathrm{d}t = u_{i1}v_{1j}\tau + \sum_{k=2}^{21} u_{ik}v_{kj}\frac{\mathrm{e}^{\lambda_k\tau} - 1}{\lambda_k}. \tag{8}$$

We have implemented this calculation of the *gdi* in the python pipeline for the case where the two populations are sister lineages exchanging migrants between themselves but not with other populations.

When populations A and B are involved in gene flow with other populations, analytical calculation of the *gdi* becomes complicated. It is simpler to simulate gene trees for sequences a_1, a_2, b under the extended migration model involving more than two populations to calculate the *gdi*. Specifically, given the fully specified MSC-M model for all species/populations (including the species tree topology and parameters such as τ, θ, M), simulate the gene trees with branch lengths (coalescent times) for a large number of loci ($R = 10^6$, say) , at which three sequences (a_1, a_2, b) are sampled. The *gdi* is simply the proportion of loci at which the gene tree is G_{1a}, that is, G_1 with $t_1 < \tau_{AB}$ (fig. 1).

Hierarchical merge and split (HMS?) algorithms

We implement both the hierarchical merge and hierarchical split algorithms in a python pipeline (fig. 3). Both algorithms require a guide tree for populations, possibly with migration events. In the merge algorithm, we progressively merge the populations into the same species, starting from the tips of the tree and moving towards the root. The merge is accepted if and only if the *gdi* < 0.2 for the population pair. The algorithm stops when no population/species pair can be merged (fig. 3a).

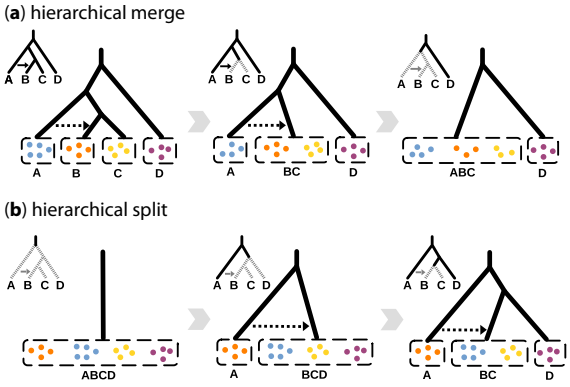


Figure 3: (**a**) Hierarchical merge and (**b**) hierarchical split algorithms applied to the same guide tree for four populations.
[Prepare a version of the figure without gene flow using the same guide tree for merge and plit.
Edit this version (with migration) to have the same guide tree for merge and split. I suppose with migration it's not obvious how merge should work, but we can discuss options in relation to the issue of non-monophyletic species.]

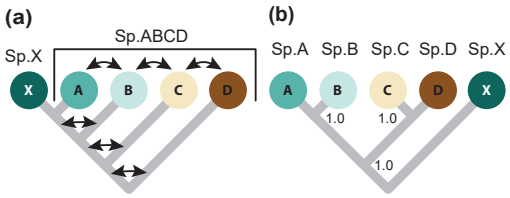


Figure 4: (**a**) An isolation-by-distance model used to simulate multilocus sequence data. A, B, C, D represent populations of a widely distributed species while X is a new species that split off from population A. (**b**) Incorrect species delimitation and phylogeny in Bayesian model selection using BPP under the MSC model assuming no gene flow. Use of the guide tree and the *gdi* criterion leads to delimitation of two species. Redrawn after Leaché *et al.* (2019, fig. 5).

In the hierarchical split algorithm, we start from the MSC model of one species and progressively split each species into distinct species, starting from the root and moving towards the tips of the tree (fig. 3b). The split is accepted if and only if the *gdi* > 0.7 for the species pair. The algorithm stops when no species can be split (fig. 3b).

If there are K populations on the guide tree, the merge algorithm arrives at a high number of species while the split algorithm arrives at a low number, with $1 \le K_l \le K_u \le K$.

Example with simulated data ($ABCDX$)

Leaché *et al.* (2019) simulated sequence data under the MSC-with-migration model for five populations of figure 4. Populations A, B, C, D represent a

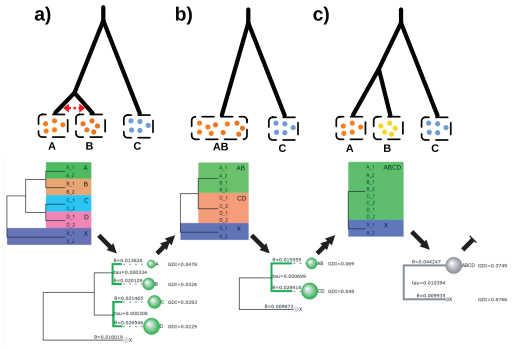


Figure 5: [Delete this. We are using the ABCDX example.] Simulated model with three populations and two species used for the tutorial. **a)** Simulated speciation scenario. **b)** Correct species delimitation. **c)** Incorrect species delimitation inferred by BPP A11. Output produced by the pipeline recording the iterations, using the model of figure 5.

single large paraphyletic species distributed across a wide geographic range. Migration between any two neighbouring populations occurs at the rate of $M = Nm = 2$ migrants per generation. X represents a new species that split off from population A, and there is no gene flow to or from X (fig. 4). The data consisted of $L = 100$ simulated loci, with two sequences sampled per species per locus, and 500 sites in the sequence. We use the dataset to illustrate our pipeline. The control file and the program output for the BPP analyses are shown in figures .

Tutorial with simulated data (ABC)

[delete this ABC simulation example and replace with the ABCDX example]

Simulation scenario. Sequence data was simulated using a model tree for three populations, $((A, B), C)$, with parameter $\tau_{ABC} = 0.02, \tau_{AB} = 0.001$, $\theta_A = \theta_B = \theta_C = 0.02$, with migration rates $M_{AB} = M_{BA} = 2$ (fig. **??**). A and B represent two geographic populations of a single species, which diverged recently and have been exchanging migrants. Five diploid individuals were sampled from each population per locus.

When BPP A11 mode is used to delimit species in this dataset, it incorrectly recovers three species with posterior probability > 0.999 (fig. **??**c).

The analyses is specified using a control file, as follows.

```
output_directory = Results
Imapfile = Sim_imap.txt
seqfile  = Sim_MSA.txt

guide_tree = ((A,B),C);
migration  = {A -> B, B -> A}
migprior   = 1 10

mode = merge
```

```
gdi_threshold = <0.2

burnin   = 200000
nsample = 500000
```

The variable `output_directory` specifies the folder where the results of the analysis will be written. `seqfile` points to phylip formatted multiple sequence alignment. `Imapfile` specifies the text file which describes the mapping of individuals to populations in the starting delimitation. `guide_tree` is a Newick representation of the guide tree topology. `migration` is used to specify the source and destination of migration events. `migprior` defines the gamma prior on migration rate. In this case, the value of `1 10` specifies a prior mean $M = 0.1$. `mode` specifies that the direction of iterative refinement. `gdi_threshold` specifies the gdi value below which two populations are merged into a candidate species. `burnin` specifies number of discarded iterations in the burn-in period of the MCMC analysis. `nsample` specifies the number of samples used in the estimation of τ, θ and M parameters. A detailed overview of all possible control file parameters is provided in the documentation for the program.

The program can be launched from a terminal, as follows

```
python HMDelimit.py --mcfile tutorial_sim_merge.txt
```

During the analysis, the pipeline provides written feedback to the user about the current state of the delimitation and the decisions and modifications made during each iteration:

```
Analysis started
Starting state of merge mode
Accepted species (3) in starting delimitation:
((A,B),C);

*** Iteration 1 ***
Inferred tau and theta parameters:
theta       tau
C    0.019639
A    0.018053
B    0.020506
AB   0.020471  0.000964
ABC  0.019852  0.020066

Migration rates:
source destination        M
A           B 2.138254
B           A 2.090163

Proposal results:
node 1 node 2  gdi 1  gdi 2  merge accepted
A      B   0.07   0.06            True

Accepted species (2) after iteration 1:
(AB,C);

*** Iteration 2 ***
Inferred tau and theta parameters:
theta       tau
```

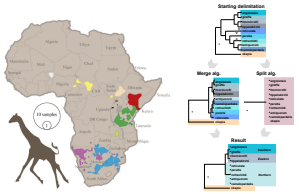


Figure 6: [For each of the empirical examples, perhaps include a picture of the animal, a map showing geographical distribution, the guide tree, and the merge and split results.] (**a**) The geographical distributions of nine subspecies of giraffa, modified from Petzold and Hassanin (2020), figure. (**b**) The hierarchical split algorithm supports 3 species, while (**c**) the hierarchical merge algorithm supports 6 species. [colours on the map and on the tree do not seem to match.]
Merge and Split analyses in the Giraffe and Okapia dataset. Top shows starting delimitation with migration patterns. Merge and split analyses converge to three giraffe species, which correspond to the geographically separated Southern, Eastern, and Northern metapopulations.

```
C    0.019649
AB   0.025840
ABC  0.019670  0.020204

Proposed results:
node 1 node 2  gdi 1  gdi 2  merge  accepted
    AB      C   0.79   0.87            False

Accepted species (2) after iteration 2:
(AB,C);

All modifications rejected.  Final delimitation reached.
```

During the first iteration of the merge analysis, the *gdi* score of the *A* and *B* populations was inferred to be below the threshold for distinct species, causing them to be merged into the new species *AB*. The merge proposal between the *AB* and *C* species in the second iteration is rejected, leaving two species, the known correct delimitation (fig. 5b).

The control files and data files are all provided in the supplementary materials.

Empirical examples

Species delimitation in the genus Giraffa

Data. Species delimitation in the genetically isolated, but phenotypically convergent Giraffes has generated sizeable controversy (Fišer et al., 2018). Based on morphological characters and molecular data, various hypotheses have classified the nine giraffe subspecies (*camelopardalis*, *angolensis*, *antiquorum*, *giraffa*, *peralta*, *reticulata*, *rothschildi*, *thornicrofti* and *tippelskirchi*.) into anywhere from one to six species.

Petzold and Hassanin (2020) compiled a multilocus dataset of 21 introns (average sequence length 808 bp),

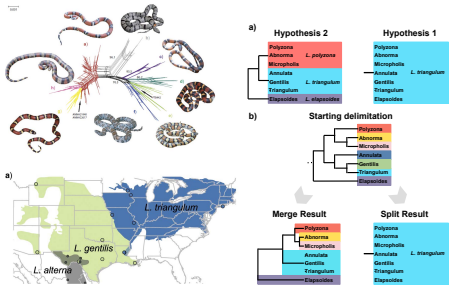


Figure 7: Species delimitation hypotheses in *Lampropeltis triangulum*. **a)** Three-species, and one-species delimitation hypotheses suggested by Chambers and Hillis (2020). **b)** Starting delimitation, and results from merge and split analysis.
Coloured dots represent the sampling location and original classification of individuals (blue: *triangulum*, green: *gentilis*).

sampled from 66 individuals from the nine subspecies. They found that population genetic approaches, such as the program STRUCTURE and the phylogenetic approaches implemented in MrBayes, PhyML, and SuperTRI all supported three species. However, the MSC based methods in *BEAST, STACEY, and BPP strongly supported five species, which was the maximum that was tested.

Based on the observations of mitochondrial haplotypes and hybridized individuals (Fennessy *et al.*, 2016; Petzold and Hassanin, 2020), bidirectional migration was specified between the *tippelskirchi* and *reticulata*, as well as the *reticulata* and *rothschildi* subspecies. The migration rate was assigned the prior $(\Gamma(1, 100))$ with mean 0.01 migrant individuals per generation. Merge and split analyses were conducted with the animal specific *gdi* thresholds of 0.3 and 0.7, as recommended by Jackson *et al.* (2017).

The split algorithm suggested three species while the merge algorithm suggested five (fig. 6). Both grouped the Eastern populations *thornicrofti* and *tippelskirchi* into one species, and the Southern populations *angolensis* and *giraffa* into another species. The split algorithm lumped the remaining five subspecies into a single Northern species, while the merge algorithm recognized four species.

Species delimitation in milksnakes (Lampropeltis triangulum)

The American milksnake *Lampropeltis triangulum* is a New World snake with one of the widest known geographic distributions within the squamates, with seven subspecies known: *abnorma*, *polyzona*, *micropholis*, *triangulum*, *gentilis*, *annulata*, *elapsoides* (fig. 7a). Ruane *et al.* (2014) analyzed 11 nuclear loci (average length 537 bp) for 164 individuals from the seven subspecies using BPP and found evidence for seven independent species.

Chambers and Hillis (2020) criticised these results, suggesting that several of the hypothesized species of milksnakes appear to represent arbitrary slices of continuous geographic clines. Based on a combination of phylogeographic and genetic evidence, they suggested two alternative hypotheses: a one-species hypothesis merging all subspecies into a single species, or a three-species hypothesis separating the *polyzona*, *triangulum*, and *elapsoides* lineages as species.

Chambers and Hillis (2020) also demonstrated that five different arbitrary east-west splits of the the *gentilis* and *triangulum* populations are all supported by BPP as being two separate species (Fig. 7). This result is highly concerning, as these five alternative species delimitations are not mutually compatible. These results also echo the simulations of Barley *et al.* (2018), who demonstrated that BPP will delimit geographically separated clusters of individuals from a single species as distinct species entities.

We analyzed the data using our pipeline, using the guide tree of ???, with no migration rates assumed (fig. 7). Merge and split algorithms were run using *gdi* thresholds of 0.3 and 0.7.

The merge algorithm established an upper bound of five species. When compared with the three-species hypothesis that was the suggested upper bound by Chambers and Hillis (2020), two of the species (*elapsoides* and *triangulum*) were identical, but HMDelimit identified additional diversity in the *polyzona* branch, marking each population as a distinct species. The split analysis only supported a single species.

We conducted a second analysis using only the 38 individuals from the *gentilis*, *triangulum*, and *alterna* populations (which acted as an outgroup in all analyses). The assignment of individuals to the *gentilis* and *triangulum* populations was varied in each analysis, according to the five arbitrary East-West splits of Chambers and Hillis (2020) (fig. 7). Merge and split analyses were ran using the settings as above.

For all five of the East-West geographic splits tested, our merge and split analyses converged on an identical result, merging the *gentilis* and *triangulum* populations into a single species, congruent with the suggestions of Chambers and Hillis (2020).

Introgression and species delimitation in the longear sunfish (Lepomis megalotis).

The longear sunfish (*Lepomis megalotis*) is a freshwater fish in the sunfish family, Centrarchidae, of order Perciformes. It is native to eastern North America from the Great Lakes down to northeastern Mexico. Due to widespread geographic distributions and frequent hybridizations, species delimitation in the longear sunfish poses considerable challenges.

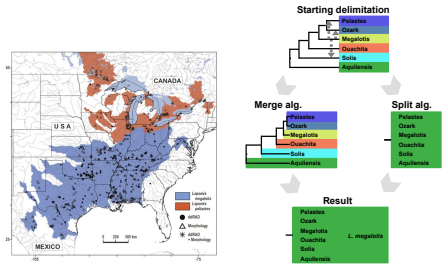


Figure 8: Species delimitation in the Longear Sunfish *Lepomis megalotis*. Both the merge and split algorithms support a single species.

Kim *et al.* (2022) analyzed a dataset of 163 ddRAD loci (average sequence length 89 bp) sampled from 50 individuals from the six subspecies: *aquilensis*, *solis*, *ouachita*, *megalotis*, *ozark*, *pelastes*. After determining a species tree using IQ-TREE, they used BPP A00 without migration to calculate τ and θ parameters for each of the subspecies, and used these values to calculate *gdi* scores, and delimit species in the group. They found that none of the populations have *gdi* values supporting distinct species status. Kim *et al.* (2022) also utilized FASTSIMCOAL2 to identify patterns of gene flow, and find evidence for multiple instances of significant historical or ongoing genetic exchange. This may be problematic, as their genetic delimitation procedure did not account for the patterns of gene flow observed.

We reanalyzed the data, taking into account migration between the subspecies. Based on presence of hybridized individuals, migration from the *megalotis* population to the *pelastes*, *solis*, and *ozark* populations was specified.

Both merge and split analyses supported a single species. This is congruent with the *gdi* based delimitation of Kim *et al.* (2022), who found that all populations have *gdi* values below the threshold for distinct species status.

Discussion

Challenges of heuristic species delimitation

Several issues with the *gdi* criterion have been noted before (Leaché *et al.*, 2019). First, given populations A and B, two *gdi* values may be calculated:

$$\begin{aligned} gdi_A &= 1 - \mathrm{e}^{-2\tau_{AB}/\theta_A} \\ gdi_B &= 1 - \mathrm{e}^{-2\tau_{AB}/\theta_B} \end{aligned} \tag{9}$$

These may not be consistent concerning the species status of populations A and B (Leaché *et al.*, 2019).

Second, the *gdi* may be large because the population is very small. Rannala and Yang (2020) recommended the use of absolute divergence time, such that two populations are considered distinct species only if their $gdi > 0.7$.

Assignment of individuals to populations and construction of the guide tree. Our pipeline requires the user to supply a guide tree. This may be inferred using a species tree estimation method under the MSC model with no gene flow (Yang, 2014; Rannala and Yang, 2017). Alternatives include maximum likelihood tree inference using concatenated data, or use of the mitochondrial genes.

The arbitrariness of the criterion. However, for mammals, a 10% CO1 (or cytb) divergence is a sure thing for distinct species.

Any empirical thresholds for particular criteria or properties will be imprecise, as it is recognized by multiple authors that such an attempt is futile Wells *et al.* (2022). While there can be no set of universal criteria or properties applicable to species delimitation, a heuristic approach does provide useful guide.

Molecular phylogenetic or population genetic analysis should always be integrated with an assessment of congruence with morphological and ecological data. Using genetic data, one should not exclude species generated by processes that do not automatically or immediately result in monophyly, such as hybrid speciation, polyploidy, or paraphyly in the case of recent ancestor-descendant speciation. Where molecular phylogenetic analysis is impractical due to inadequate samples or easily sequenced material, or where it fails to resolve well-supported relationships, species delimitation remains possible, but should be based on a strong hypothesis of phylogenetic relatedness resulting from multiple and unambiguous phenotypic and ecological traits.

The effects of sampling. Many empirical biologists emphasized the importance of sampling: Chambers and Hillis (2020); Wells *et al.* (2022). Yang and Rannala (2017) has pointed out that rarity and singletons should not be a major problem. Migration rates can also be estimated when some species are missing or unsampled. Zhang *et al.* (2011) through simulation illustrated that failure to sample the intermediate populations in a stepping-stones design does not cause false positives for species delimitation by BPP.

Gene flow and non-monophyletic species

Analyses of genomic data in the past two decades have demonstrated the prevalence of interspecific gene flow. Several studies suggested evidence for speciation despite ongoing gene flow, as in *Heliconius* butterflies (Martin *et al.*, 2013), Mangrove trees (He *et al.*, 2019), and Western Pacific abalones (Hirase *et al.*, 2021). Issues arise when we want to delimit species when there is gene flow between the species or populations.

Bayesian model selection. First consider Bayesian model selection. There are three models for two populations (A, B): M1: one species, M2-0 two species with no migration, and M2-m two species with migration. Leaché *et al.* (2019) compare M1 and M2-0 to decide whether there are one or two species, even

though the data were simulated with gene flow, and M2-m was not considered. Alternatively one may insist species status only if there is no significant amount of gene flow (i.e., only if M2-0 wins over M2-m), and consider M2-m as representing one species. This approach may suffer from over-lumping. [Generate an example. What about the $ABCDX$ dataset?]

It is not so clear how to incorporate gene flow in the hierarchical merge and split algorithms. In Leaché *et al.* (2019), we used MSC with no gene flow (M2-0) to construct the guide tree, and then the merge or split algorithms rely on the MSC model with no gene flow. The migration model is used to simulate data but not used in analysis of the data. This way the guide tree of figure 3b (Leaché *et al.*, 2019) was incorrect, but we arrived at the correct answer of two species: $ABCD$ and X. If we use the MSC-M model and use the correct guide tree with migration of figure 4a, there are two problems. First we will never recover the correct answer of two species by merging or splitting species and populations, keeping the migration events in the model. Second, when we merge populations according to the guide tree it may not be clear whether we want to keep the migration rate. For example if we merge X and A, it is unclear whether we want migration between XA and B since according to the guide tree there is gene flow between A and B but none between X and B. Another approach may be to populations that have high migration rates between them (with, $M > 1$, say), even if they are not sister lineages on the guide tree. For example, in the case of figure 4a, we will attempt to merge AB, BC, and CD, besides XA. Again there may be ambiguities in the specification of migration events in the new model with merged populations.

Program availability

The pipeline is written in python, which drives parameter estimation under the MSC or MSC-M models using BPP. The source code, documentation, and empirical datasets analyzed in the paper are available at `https://github.com/abacus-gene/xxx`.

Acknowledgements

We thank Bruce Rannala for code for calculating average pairwise sequence distances between species and Asif Tamuri for help reviewing the code. This study has been supported by Biotechnology and Biological Sciences Research Council grants (BB/T003502/1, BB/R01356X/1) to Z.Y.

References

Alexander, D. H. and Lange, K. 2011. Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. *BMC Bioinformat.*, 12: 246.

Bamberger, S., Xu, J., and Hausdorf, B. 2022. Evaluating species delimitation methods in radiations: The land snail *Albinaria cretensis* complex on crete. *Syst. Biol.*, 71(2): 439–460.

Barley, A. J., Brown, J. M., and Thomson, R. C. 2018. Impact of model violations on the inference of species boundaries under the multispecies coalescent. *Syst. Biol.*, 67(2): 269–284.

Bolotov, I. N., Aksenova, O. V., Bespalaya, Y. V., Gofarov, M. Y., Kondakov, A. V., Paltser, I. S., Stefansson, A., Travina, O. V., and Vinarski, M. V. 2017. Origin of a divergent mtDNA lineage of a freshwater snail species, *Radix balthica*, in Iceland: cryptic glacial refugia or a postglacial founder event? *Hydrobiologia*, 787(1): 73–98.

Campillo, L. C., Barley, A. J., and Thomson, R. C. 2020. Model-based species delimitation: are coalescent species reproductively isolated? *Syst. Biol.*, 69(4): 708–721.

Chambers, E. A. and Hillis, D. M. 2020. The multispecies coalescent over-splits species in the case of geographically widespread taxa. *Syst. Biol.*, 69(1): 184–193.

Chen, L., Qiu, Q., Jiang, Y., Wang, K., Lin, Z., Li, Z., Bibi, F., Yang, Y., Wang, J., Nie, W., Su, W., Liu, G., Li, Q., Fu, W., Pan, X., Liu, C., Yang, J., Zhang, C., Yin, Y., Wang, Y., Zhao, Y., Zhang, C., Wang, Z., Qin, Y., Liu, W., Wang, B., Ren, Y., Zhang, R., Zeng, Y., da Fonseca, R. R., Wei, B., Li, R., Wan, W., Zhao, R., Zhu, W., Wang, Y., Duan, S., Gao, Y., Zhang, Y. E., Chen, C., Hvilsom, C., Epps, C. W., Chemnick, L. G., Dong, Y., Mirarab, S., Siegismund, H. R., Ryder, O. A., Gilbert, M. T. P., Lewin, H. A., Zhang, G., Heller, R., and Wang, W. 2019. Large-scale ruminant genome sequencing provides insights into their evolution and distinct traits. *Science*, 364(6446): eaav6202.

Duda, T. F., J. 2021. Patterns of variation of mutation rates of mitochondrial and nuclear genes of gastropods. *BMC Ecol. Evol.*, 21(1): 13.

Fennessy, J., Bidon, T., Reuss, F., Kumar, V., Elkan, P., Nilsson, M. A., Vamberger, M., Fritz, U., and Janke, A. 2016. Multi-locus analyses reveal four giraffe species instead of one. *Curr. Biol.*, 26(18): 2543–2549.

Fiser, C., Robinson, C. T., and Malard, F. 2018. Cryptic species as a window into the paradigm shift of the species concept. *Mol. Ecol.*, 27(3): 613–635.

Flouri, T., Jiao, X., Rannala, B., and Yang, Z. 2018. Species tree inference with BPP using genomic sequences and the multispecies coalescent. *Mol. Biol. Evol.*, 35(10): 2585–2593.

He, Z., Li, X., Yang, M., Wang, X., Zhong, C., Duke, N. C., Wu, C. I., and Shi, S. 2019. Speciation with gene flow via cycles of isolation and migration: insights from multiple mangrove taxa. *Natl. Sci. Rev.*, 6(2): 275–288.

Hebert, P. D., Stoeckle, M. Y., Zemlak, T. S., and Francis, C. M. 2004. Identification of birds through DNA barcodes. *PLoS Biol.*, 2: 1657–1663.

Hirase, S., Yamasaki, Y. Y., Sekino, M., Nishisako, M., Ikeda, M., Hara, M., Merila, J., and Kikuchi, K. 2021. Genomic evidence for speciation with gene flow in broadcast spawning marine invertebrates. *Mol. Biol. Evol.*, 38(11): 4683–4699.

Jackson, N. D., Carstens, B. C., Morales, A. E., and O'Meara, B. C. 2017. Species delimitation with gene flow. *Syst. Biol.*, 66(5): 799–812.

Jiao, X. and Yang, Z. 2021. Defining species when there is gene flow. *Syst. Biol.*, 70(1): 108–119.

Jiao, X., Flouri, T., and Yang, Z. 2021. Multispecies coalescent and its applications to infer species phylogenies and cross-species gene flow. *Nat. Sci. Rev.*, 8(12): DOI: 10.1093/nsr/nwab127.

Kim, D., Bauer, B. H., and Near, T. J. 2022. Introgression and species delimitation in the longear sunfish *Lepomis megalotis* (Teleostei: Percomorpha: Centrarchidae). *Syst. Biol.*, 71(2): 273–285.

Leaché, A. D., Zhu, T., Rannala, B., and Yang, Z. 2019. The spectre of too many species. *Syst. Biol.*, 68(1): 168–181.

Luo, A., Ling, C., Ho, S. Y. W., and Zhu, C. D. 2018. Comparison of methods for molecular species delimitation across a range of speciation scenarios. *Syst. Biol.*, 67(5): 830–846.

MacGuigan, D. J., Hoagstrom, C. W., Domisch, S., Hulsey, C. D., and Near, T. J. 2021. Integrative ichthyological species delimitation in the Greenthroat Darter complex (*Percidae*: *Etheostomatinae*). *Zoologica Scripta*, 50(6): 707–733.

Martin, S. H., Dasmahapatra, K. K., Nadeau, N. J., Salazar, C., Walters, J. R., Simpson, F., Blaxter, M., Manica, A., Mallet, J., and Jiggins, C. D. 2013. Genome-wide evidence for speciation with gene flow in *Heliconius* butterflies. *Genome Res.*, 23(11): 1817–1828.

Petit, R. and Excoffier, L. 2009. Gene flow and species delimitation. *Trends Ecol. Evol.*, 24: 386–393.

Petzold, A. and Hassanin, A. 2020. A comparative approach for species delimitation based on multiple methods of multi-locus DNA sequence analysis: A case study of the genus *Giraffa* (Mammalia, Cetartiodactyla). *PLoS One*, 15(2): e0217956.

Pinho, C. and Hey, J. 2010. Divergence with gene flow: models and data. *Ann. Rev. Ecol. Evol. Syst.*, 41: 215–230.

Ramirez-Reyes, T., Blair, C., Flores-Villela, O., Pinero, D., Lathrop, A., and Murphy, R. 2020. Phylogenomics and molecular species delimitation reveals great cryptic diversity of leaf-toed geckos (Phyllodactylidae: *Phyllodactylus*), ancient origins, and diversification in Mexico. *Mol. Phylogenet. Evol.*, 150.

Rannala, B. and Yang, Z. 2003. Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. *Genetics*, 164(4): 1645–1656.

Rannala, B. and Yang, Z. 2017. Efficient Bayesian species tree inference under the multispecies coalescent. *Syst. Biol.*, 66: 823–842.

Rannala, B. and Yang, Z. 2020. Species delimitation. In N. Galtier, F. Delsuc, and C. Scornavacca, editors, *Phylogenetics in the Genomic Era*, book section 5.5, pages 5.5.1–18. No Commercial Publisher.

Ruane, S., Bryson, R. W., Pyron, R. A., and Burbrink, F. T. 2014. Coalescent species delimitation in milksnakes (genus lampropeltis) and impacts on phylogenetic comparative analyses. *Syst. Biol.*, 63(2): 231–250.

Schilthuizen, M. 2018. *Biogeography and Biodiversity of the Aegean*. Broken Hill Publishers, Nicosia, Cyprus.

Stankowski, S. and Ravinet, M. 2021. Defining the speciation continuum. *Evolution*,

75(6): 1256–1273.
Sukumaran, J. and Knowles, L. 2017. Multispecies coalescent delimits structure, not species. *Proc. Natl. Acad. Sci. USA.*, 114: 1607–1612.
Wells, T., Carruthers, T., Munoz-Rodriguez, P., Sumadijaya, A., Wood, J. R. I., and Scotland, R. W. 2022. Species as a heuristic: reconciling theory and practice. *Syst. Biol.*, 71(5): 1233–1243.
Wu, W., Ng, W. L., Yang, J. X., Li, W. M., and Ge, X. J. 2018. High cryptic species diversity is revealed by genome-wide polymorphisms in a wild relative of banana, *Musa itinerans*, and implications for its conservation in subtropical China. *BMC Plant Biol.*, 18(1): 194.
Yang, Z. 2014. *Molecular Evolution: A Statistical Approach*. Oxford University Press, Oxford, England.
Yang, Z. 2015. The BPP program for species tree estimation and species delimitation. *Curr. Zool.*, 61: 854–865.
Yang, Z. and Rannala, B. 2010. Bayesian species delimitation using multilocus sequence data. *Proc. Natl. Acad. Sci. USA*, 107: 9264–9269.
Yang, Z. and Rannala, B. 2017. Bayesian species identification under the multispecies coalescent provides significant improvements to DNA barcoding analyses. *Mol. Ecol.*, 26: 3028–3036.
Zhang, C., Zhang, D.-X., Zhu, T., and Yang, Z. 2011. Evaluation of a Bayesian coalescent method of species delimitation. *Syst. Biol.*, 60: 747–761.

Supplemental Materials

Extended Methods and Materials

Implementation details

We use the example of the simulated data of figure 4 to illustrate the details of implementation.

A control file is used to specify the analysis procedure.

Output. The program output is self-explanatory (fig. S1).

```
seqfile = Leache_2019_sequences.txt
Imapfile = Leache_2019_starting_populations.txt
guide tree = ((A, B), (C, D)), X);

mode = merge
GDI threshold = 0.2
decision criteria = all

burnin = 50000
nsample = 100000
```

`seqfile` is the sequence alignment file in PHYLIP formatt. `Imapfile` specifies the mapping of individuals to populations. `guide tree` is a Newick representation of the guide tree topology. `mode` specifies the algorithm, with `split` for hierarchical split and `merge` for hierarchical merge. `GDI threshold` specifies the *gdi* value below which two populations are merged into a candidate species. `decision criteria` specifies how the *gdi* values should be interpreted. As a mutation rate is not specified, the number of generations cannot be inferred. Accordingly, the decision criteria of `all` currently indicates that both *gdi* values need to be less than the threshold (0.2) for two populations to be merged. `burnin` and `nsample` specify the settings for the MCMC run.

```
ACCEPTED SPECIES IN STARTING DELIMITATION:
A, B, C, D, X

ITERATION 1:
PROPOSED MERGE   GDI 1   GDI 2   MERGE ACCEPTED
'A', 'B'         0.0478  0.0326  True
'C', 'D'         0.0283  0.0229  True

THE FOLLOWING PROPOSALS TO MERGE WERE ACCEPTED:
1) 'A', 'B' -> 'AB'
2) 'C', 'D' -> 'CD'

ACCEPTED SPECIES AFTER ITERATION 1:
AB, CD, X

ITERATION 2:
PROPOSED MERGE   GDI 1   GDI 2   MERGE ACCEPTED
'AB', 'CD'       0.0689  0.0475  True

THE FOLLOWING PROPOSALS TO MERGE WERE ACCEPTED:
1) 'AB', 'CD' -> 'ABCD'

ACCEPTED SPECIES AFTER ITERATION 2:
ABCD, X

ITERATION 3:
PROPOSED MERGE   GDI 1   GDI 2   MERGE ACCEPTED
'ABCD', 'X'      0.3749  0.8766  False

ACCEPTED SPECIES AFTER ITERATION 3:
ABCD, X

ALL PROPOSALS REJECTED, FINAL DELIMITATION REACHED
```

Figure S1: Screen output from running the pipeline to analyze the simulated dataset of figure 4.

Table S1: Rate matrix for Markov chain describing transitions between states in multispecies coalescent with migration model with two populations (A and B) and three sequences (a_1, a_2, and b).

	AAA	AAB	ABA	ABB	BAA	BAB	BBA	BBB	$A_{a_1}A$	$A_{a_2}A$	A_bA	$C_{a_1}B$	$B_{a_2}B$	B_bB	$A_{a_1}B$	$A_{a_2}B$	A_bB	AB_{a_1}	AB_{a_2}	AB_b	$A\|B$
AAA	·	w_{BA}	w_{BA}		w_{BA}				c_A	c_A	c_A										
AAB	w_{AB}	·		w_{BA}		w_{BA}														c_A	
ABA	w_{AB}		·	w_{BA}			w_{BA}												c_A		
ABB		w_{AB}	w_{AB}	·				w_{BA}							c_B						
BAA	w_{AB}				·	w_{BA}	w_{BA}											c_A			
BAB		w_{AB}			w_{AB}	·		w_{BA}								c_B					
BBA			w_{AB}		w_{AB}		·	w_{BA}									c_B				
BBB				w_{AB}		w_{AB}	w_{AB}	·				c_B	c_B	c_B							
$A_{a_1}A$									·						w_{BA}			w_{BA}			c_A
$A_{a_2}A$										·						w_{BA}			w_{BA}		c_A
A_bA											·						w_{BA}			w_{BA}	c_A
$B_{a_1}B$												·			w_{AB}			w_{AB}			c_B
$B_{a_2}B$													·			w_{AB}			w_{AB}		c_B
B_bB														·			w_{AB}			w_{AB}	c_B
$A_{a_1}B$									w_{AB}			w_{BA}			·						
$A_{a_2}B$										w_{AB}			w_{BA}			·					
A_bB											w_{AB}			w_{BA}			·				
AB_{a_1}									w_{AB}			w_{BA}						·			
AB_{a_2}										w_{AB}			w_{BA}						·		
AB_b											w_{AB}			w_{BA}						·	
$A\|B$																					·

Note.— $w_{AB} = 4M_{AB}/\theta_B = m_{AB}/\mu and w_{BA} = 4M_{BA}/\theta_A = m_{BA}/\mu$ are mutation-scaled migration rates, and $c_A = 2/\theta_A$ and $c_B = 2/\theta_B$ are the coalescent rates. The state of the chain is given by the population IDs (A or B) and sequence IDs (such as a_1, a_2, a_1a_2). For example the initial state $A_{a_1}A_{a_2}B_b$ means that the three sequences a_1, a_2, and b are from populations A, A, and B, respectively. States with three sequences are abbreviated, with the three sequences assumed to be in the order a_1, a_2, b so that the sequence IDs are suppressed. Thus $A_{a_1}A_{a_2}B_b$ is 'AAB'. State $A_{a_1a_2}B_b$ means that two sequences remain in the sample, with the ancestor of sequences a_1 and a_2 is in population A while sequence b is in population B. This is abbreviated 'AB_b', with the sequence ID 'a_1a_2' suppressed. '$A|B$' is an absorbing state in which only one sequence remains in the sample, in either A or B, after two coalescent events have occurred. From Leaché *et al.* (2019).