Journal (2050), Vol. 00, No. 000, pp. 1–14
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 fluctuations 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 heuristic species delimitation, and implement 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 suggest that the procedure may be less prone to 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

Introduction

Delineation 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 are informative about many processes related to species delimitation and identification, including population identities, interspecific hybridization and gene flow, and phylogenetic relationships and and divergence times among the populations (Fujita *et al.*, 2012; Degnan, 2018; Kubatko, 2019; Jiao *et al.*, 2021). Early methods that use genetic data to identify and delimit species rely on simple genetic-distance cutoffs (such as the 3x, 4x, or 10x rules), requiring interspecific divergence to be a few times larger than intraspecific diversit (Hebert *et al.*, 2004), or reciprocal monophyly in gene trees (e.g, Sites and Marshall, 2004). However, such criteria are too simplistic as they do not accommodate polymorphism in the ancestral populations and incomplete lineage sorting (Hudson and Turelli, 2003) or uncertainties in gene-tree reconstruction (Knowles and Carstens, 2007; Yang and Rannala, 2017).

The multispecies coalescent model (Rannala and Yang, 2003) provides the natural framework for analysis of sequence data from closely related species or populations. Likelihood-based implementations of the MSC accommodate incomplete lineage sorting and stochastic variation in gene trees (so that reciprocal monophyly is not needed) and gene-tree reconstruction errors, allowing species delimitation to proceed despite widespread incomplete lineage sorting or very little phylogenetic information at every locus.

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 and fitted to genomic sequence data sampled from the extant species or populations. Competing models of delimitation can then be compared via Bayesian model selection (i.e., using posterior model probabilities or Bayes factors) to find the best supported delimitation (Yang and Rannala, 2010). In the Bayesian program bpp, this is accomplished by using a Markov chain Monte Carlo (MCMC) algorithm to calculate the posterior probabilities for different MSC models (Yang and Rannala, 2010, 2014; Yang, 2015; Flouri *et al.*, 2018). In simulations, bpp showed lower rates of species overestimation and underestimation than the

*to whom correspondence should be addressed

generalized mixed Yule-coalescent or Poisson tree processes (Luo *et al.*, 2018). In empirical datasets, BPP was effective in identifying cryptic species among ancient lineages. For example, Ramirez-Reyes *et al.* (2020) identified 13 new species of leaf-toed geckoes in a lineage that diverged 30 Ma.

However, the approach of model selection implemented in BPP has often been noted to 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 *Albinaria cretensis* land snail populations, and found that morphological classifications suggested 3–9 species, ADMIXTURE suggested at least 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., in Yunnan Bananas; Wu *et al.*, 2018). 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 define species status using empirical criteria based on parameters that characterize the history of population divergence and gene flow, such as the population split time (τ), population sizes (θ_A, θ_B), and migration rates (M_{AB}, M_{BA}). Those parameters can be estimated under the MSC from the genomic data, with the stochasticity of the coalescent process and the phylogenetic uncertainty in genealogical trees accommodated. Jackson *et al.* (2017) introduced a criterion called the *genealogical divergence index* (gdi), by considering the probability that two sequences sampled from A (a_1 and a_2) coalesce before either coalesces with a sequence (b) sampled from B (see fig. 1). When a_1 and a_2 coalesce first, the resulting gene tree has the topology $G_1 = ((a_1, a_2), b)$. Let its probability be $P_1 = \mathbb{P}(G_1)$. In the case of no gene flow between A and B, this is given as

$$P_1 = \mathbb{P}(G_1|\Theta) = 1 - \tfrac{2}{3}\,\mathrm{e}^{-2\tau_{AB}/\theta_A}, \qquad (1)$$

where the parameter vector is $\Theta = (\tau, \theta_A, \theta_B, \theta_R)$, with $\tau_{AB} = T_{AB}\mu$ to be the the population divergence time and $\theta_A = 4N_A\mu$ to be the population size of A, with T_{AB} to be population split time in generations, and μ the mutation rate per site per generation. Both τ and θ are measured in the expected number of mutations per site. P_1 is a simple function of $2\tau_{AB}/\theta_A = T_{AB}/(2N_A)$, the internal branch length in the population phylogeny in coalescent units (with one coalescent time unit to be $2N_A$ generations in population A).

P_1 of eq. 1 ranges from $\frac{1}{3}$ (when populations A and B are at panmixia) to 1 (when A and B are completely isolated). This is rescaled so that the gdi ranges from 0 to 1 Jackson *et al.* (2017):

$$gdi = \frac{P_1 - \frac{1}{3}}{1 - \frac{1}{3}} = 1 - \mathrm{e}^{-2\tau_{AB}/\theta_A} = 1 - \mathrm{e}^{-T_{AB}/(2N_A)}. \qquad (2)$$

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.

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 definition and computation of gdi when there is gene flow in the model, following Leaché *et al.* (2019). Then we discuss our new pipeline and illustrate it using a simulated dataset. We apply the pipeline to three empirical datasets, for giraffes, milksnakes, and sunfish.

Definition and computation of gdi under the migration model

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

$$P_1 = \mathbb{P}(G_1\big|\Theta), \qquad (3)$$

where $\Theta = (\tau_{AB}, \theta_A, \theta_B, \theta_{AB}, M_{AB}, M_{BA})$ is the vector of parameters. Jackson *et al.* (2017) estimated the minimum and maximum values for P_1 to rescale P_1 so that gdi falls into (0, 1). Those limits depend on the model parameters.

Here we instead redefine gdi as the probability that the first coalescence is between the two A sequences and that it occurs before reaching population divergence when we trace the genealogy of the three sequences (a_1, a_2, b) backwards in time. In other words,

$$gdi = \mathbb{P}(G_{1a}\big|\Theta) \qquad (4)$$

(fig. 1). This definition applies whether or not there is gene flow in the model (fig. 1), with $0 \le gdi \le 1$; in the case of no gene flow, eq. 4 is given by eq. 2.

Under the MSC-M model, the gdi can be computed analytically, using the Markov chain characterization

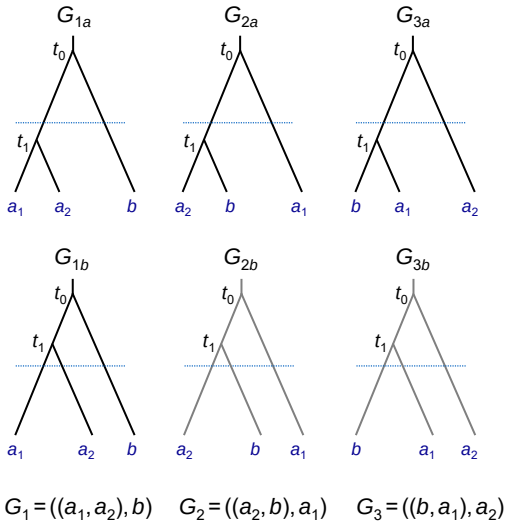

Figure 1: For a locus with two A sequences and one B sequence (a_1, a_2, b), there are three possible gene trees: $G_1 = ((a_1, a_2), b)$; $G_2 = ((a_2, b), a_1)$; and $G_3 = ((b, a_1), a_2)$. If the first coalescence is more recent than population divergence ($t_1 < \tau$), the gene trees are labelled G_{1a}, G_{2a}, G_{3a}; otherwise they are labelled G_{1b}, G_{2b}, G_{3b}. The gdi is the probability that the two A sequences coalesce first and before the population split: $gdi = \mathbb{P}(G_{1a})$. Note that if there is no gene flow between A and B, gene trees G_{2b} and G_{3b} are impossible.

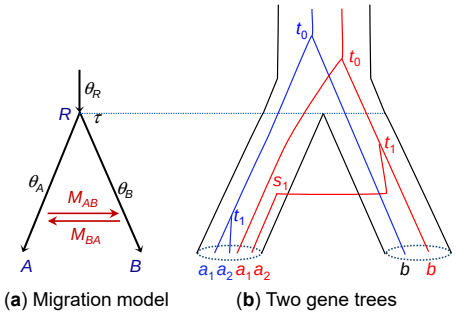

Figure 2: (**a**) An MSC-with-migration (MSC-M) model for two species or populations (A, B) showing the parameters. The two populations diverged time $\tau \equiv \tau_{AB}$ ago and have since been exchanging migrants at the rate of $M_{AB} = m_{AB}N_B$ migrants per generation from A to B and at the rate $M_{BA} = m_{BA}N_A$ from B to A. (**b**) Two gene trees, each for two A sequences and one B sequence (a_1, a_2, b). In the blue tree, a_1 and a_2 coalesce first, in population A, resulting in the gene tree $G_1 = ((a_1, a_2), b)$ (this is G_{1a} of fig. 1). In the red tree, a_2 migrates into B and coalesce with b in B, resulting in the gene tree $G_2 = ((a_2, b), a_1)$ (this is G_{2a} of fig. 1).

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 backwards in time can be described by a Markov chain (table S1). 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) (Hobolth *et al.*, 2011; Zhu and Yang, 2012; Jiao and Yang, 2021). For example, The initial state is $A_{a_1}A_{a_2}B_b$, with three sequences a_1, a_2, b in populations A, A, and B, respectively. This is also written 'AAB'. State $A_{a_1a_2}B_b$, abbreviated 'AB_b', means that two sequences remain in the sample, with the ancestor of a_1 and a_2 in A and b in B. Finally state $A|B$ is an artificial absorbing state, in which all three sequences have coalesced with the sole ancestral sequence in either A or B. There are 21 states in the Markov chain, with the transition rate (generator) matrix $Q = \{q_{ij}\}$ 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

$$q_{ij} = \sum_{k=1}^{21} u_{ik} v_{kj} \lambda_k, \tag{5}$$

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

$$p_{ij}(t) = \sum_{k=1}^{21} u_{ik} v_{kj} \, \mathrm{e}^{\lambda_k t} \,. \tag{6}$$

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 figure 2b). This has density

$$f(t) = \left[p_{AAB,AAA}(t) + p_{AAB,AAB}(t)\right] \tfrac{2}{\theta_A} + \left[p_{AAB,BBA}(t) + p_{AAB,BBB}(t)\right] \tfrac{2}{\theta_B}, \quad t < \tau. \tag{7}$$

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 that a_1 and a_2 are in A right before time t (states AAA or AAB depending on whether b is in A or B), times the rate for them to coalesce, $\frac{2}{\theta_A}$. 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

$$gdi = \int_0^\tau f(t) \, \mathrm{d}t. \tag{8}$$

To calculate the integral in eq. 8, note that from eq. 6,

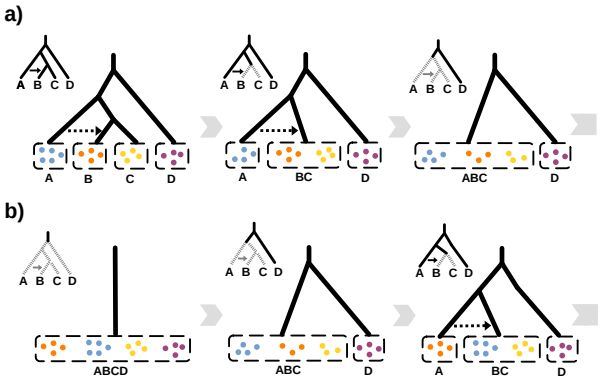

Figure 3: (**a**) Hierarchical merge and (**b**) hierarchical split algorithms applied to the same guide tree for four populations.

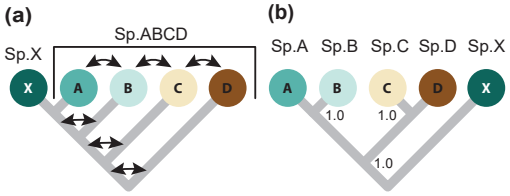

Figure 4: (**a**) An isolation-by-distance model used to simulate multilocus sequence data, in which A, B, C, D represent populations of a widely distributed species while X is a new species that split off from population A. (**b**) Incorrect species delimitation and phylogeny in Bayesian model selection 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).

$$\int_0^\tau p_{ij}(t)\,\mathrm{d}t = u_{i1}v_{1j}\tau + \sum_{k=2}^{21} u_{ik}v_{kj}\frac{\mathrm{e}^{\lambda_k\tau}-1}{\lambda_k}. \quad (9)$$

We have implemented this calculation of 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 $t_1 < \tau_{AB}$ (figs. 1&2).

The hierarchical merge and split 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 pair can be merged (fig. 3a).

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 $gdi > 0.7$. The algorithm stops when no species can be split (fig. 3b).

Both algorithms are implemented under either the MSC model with no gne flow or the MSC-M model with continuous migration. Under the MSC-M model, we retain the migration event when populations are merged as long as there is migration between the subpopulations. For example, there is migration from A to B in the guide tree (the initial delimitation) (fig. 3a). Later when B and C are merged into one species/population, we retain the migration event (now from population A to population BC).

Example with simulated data ($ABCDX$)

Leaché *et al.* (2019) simulated sequence data under the MSC-with-migration model for five populations of figure 4, in which A, B, C, D represent geographical populations for a paraphyletic species distributed across a wide geographic range while X is a new species that split off from population A. Migration between any two neighbouring populations of species $ABCD$ occurs at the rate of $M = Nm = 2$ migrants per generation, whereas there is no gene flow involving X. 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, and to discuss the challenges of species delimitation in presence of gene flow.

The control file for the program can be found in figure S1. During the analysis, the pipeline provides feedbacks about the current species delimitation and the decisions made during each iteration (figs. 5&S2). During the first iteration, attempt was made to merge the two current leaf node pairs (A, B, and $C - D$). As $gdi < 0.2$ for each pair, the merge was accepted. In the second iteration, a merge between the pair AB and CD was attempted, and again this was accepted. In the third iteration, a merge between the pair $ABCD$ and X was attempted. As $gdi > 0.2$, the merge was rejected. The final delimitation has two species, $ABCD$ and X.

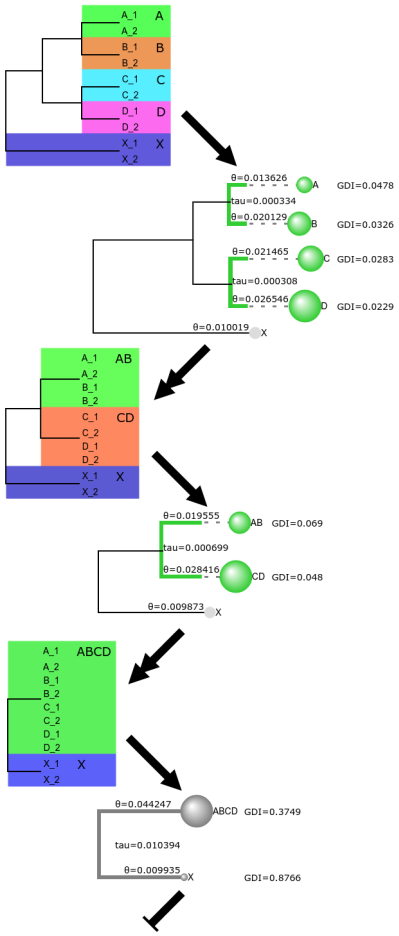

Figure 5: Screen output produced by the pipeline during the hierarchical merge iteration showing the currently accepted species delimitation and species phylogeny with parameter estimates (τ, θ).

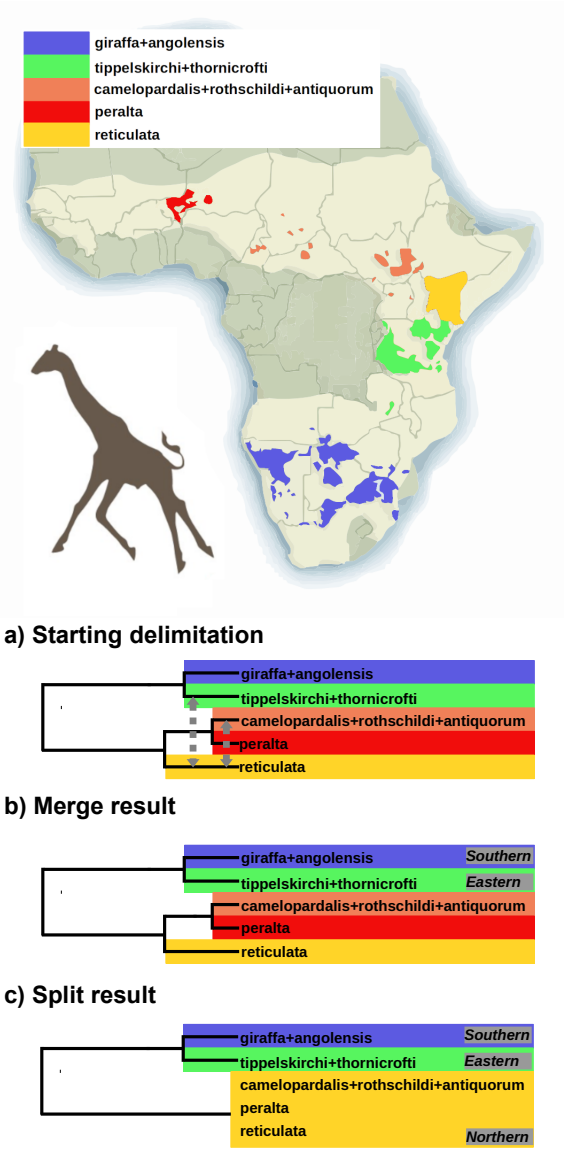

Figure 6: Geographical distributions of five species within Giraffa. Bright region on map shows historical (ca. 1700) giraffe ranges (modified from https://giraffeconservation.org/giraffe-species/). (**a**) The guide tree for five populations of giraffes, with dotted lines indicating bidirectional migration events (Petzold and Hassanin, 2020, fig. 1). (**b**) The merge algorithm supports five species, while (**c**) the split algorithm supports three. [Updated figure to use custom image. Updated colours to match between geographic distributions, and the phylogenetic trees.]

Empirical examples

Species delimitation in the genus Giraffa

Species delimitation in the genetically isolated, but phenotypically convergent Giraffes has generated considerable controversy (Fiser *et al.*, 2018). There are nine subspecies recognised: *camelopardalis*, *angolensis*, *antiquorum*, *giraffa*, *peralta*, *reticulata*, *rothschildi*, *thornicrofti* and *tippelskirchi*, which have been classified into one to six species in previous studies which used morphological characters and molecular data. Petzold and Hassanin (2020) compiled a multilocus dataset of 21 introns (average sequence length 808 bp), sampled from 66 individuals from the nine subspecies, and conducted a number of population genetic and phylogenetic analyses. The authors concluded that the best supported number of species is three, and found that BPP supports as many as five species.

This five species phylogeny (Fig. 6a) is used as the starting delimitation in our analysis. Based on phylogenetic analysis of mitochondrial haplotypes and hybridized individuals (Fennessy *et al.*, 2016; Petzold and Hassanin, 2020), bidirectional migration was specified between *tippelskirchi+thornicrofti* and *reticulata*, and between *reticulata* and *camelopardalis+rothschildi+antiquorum*. The migration rate was assigned the gamma prior $G(1, 100)$ 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) (control files in figs. S3 & S4).

The merge algorithm suggested five species while the split algorithm suggested three (fig. 6). Both methods recognized the Eastern (*thornicrofti* and *tippelskirchi* and Southern (*angolensis* and *giraffa*) species present in the starting delimitation as distinct species. For the remaining Northerns populations, the merge algorithm recognized three distinct species (fig. 6b), while the split algorithm recognized only a single Northern species combining the populations (fig. 6c).

[Also it maybe interesting to look at the estimates of parameters (τ, θ, M).] The migration rates interred during the delimitation process support the originally hypothesised patterns of hybridization and introgression between reticulated giraffes and their neighbouring populations (fig. 7). The highest levels of gene flow occurred within the Northern populations from *cam.+rot.+ant.* to *reticulata*, which notably affected four of the 21 introns in this dataset (Petzold and Hassanin, 2020)]

Source	Destination	M
tip.+tho.	reticulata	0.0024
reticulata	tip.+tho.	0.0016
reticulata	cam.+rot.+ant.	0.0272
cam.+rot.+ant.	reticulata	0.1233

Figure 7: Migration rates between the five giraffe species in the starting delimitation.

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. Seven subspecies are known: *abnorma*, *polyzona*, *micropholis*, *triangulum*, *gentilis*, *annulata*, and *elapsoides* (fig. 8a). 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 distinct species. Chambers and Hillis (2020) suggested that several species hypothesized by Ruane *et al.* (2014) may represent arbitrary slices of continuous geographic clines. Based on a combination of population genetic and phylogeographic evidence, they suggested two delimitation hypotheses, with one and three species, respectively (fig. 8**a**). Chambers and Hillis (2020) also constructed five arbitrary delimitation hypotheses, each with two species representing arbitrary east-west splits of the *gentilis* and *triangulum* populations. All the five delimitation hypotheses were supported by Bayesian model selection using BPP

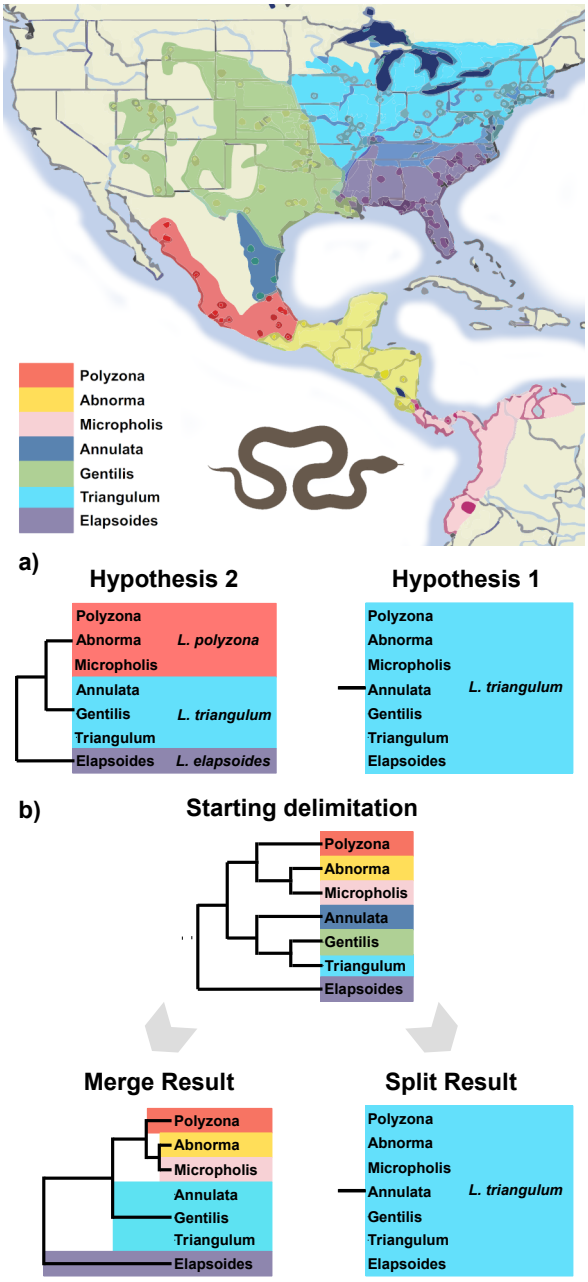

Figure 8: **a)** The one-species and three-species delimitation hypotheses for milksnakes suggested by Chambers and Hillis (2020). **b)** Starting delimitation, and inferred delimitations using the merge and split algorithms. [Updated figure to use custom image. Updated colours to match between geographic distributions, and the phylogenetic trees.]

(fig. 9), even though they are not mutually compatible.

We reanalyzed the data using our pipeline, using a guide tree for five populations of Chambers and Hillis (2020), with no migration rates assumed (fig. 8). Merge and split algorithms were run using *gdi* thresholds of 0.3 and 0.7 (control files in S5 and S6). The merge algorithm established an upper bound of five species. When compared with the three-species hypothesis

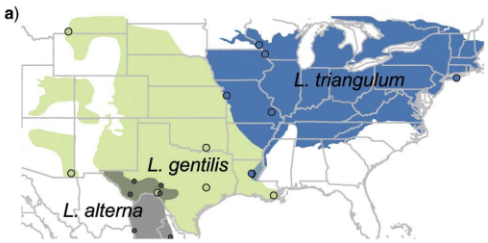

Figure 9: East-West splits for the milksnakes. Coloured dots represent the sampling location and original classification of individuals (blue: *triangulum*, green: *gentilis*).

of Chambers and Hillis (2020), two of the species (*elapsoides* and *triangulum*) were identical, but the *polyzona* lineage is split into distinct species. The split analysis supported only 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. 8). Merge and split analyses were ran using the settings as above (control files available in S7 and S8).

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

[Comment on the ~0 estimates of migration rates between geographic populations.][As patterns of hybridization or migration for nuclear genes were not specified in the Chambers and Hillis paper, and only nuclear genes were used in the analysis, the model was run without migration (just MSC), thus no migration rates were estimated (See control files in S5 an S6)]

*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. Six subspecies are recognised: *aquilensis*, *solis*, *ouachita*, *megalotis*, *ozark*, and *pelastes*. Due to widespread geographic distribution and frequent hybridization, species delimitation in the longear sunfish poses considerable challenges.

Kim *et al.* (2022) analyzed a dataset of 163 ddRAD loci (average sequence length 89 bp) sampled from 50 individuals from the six subspecies. 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.

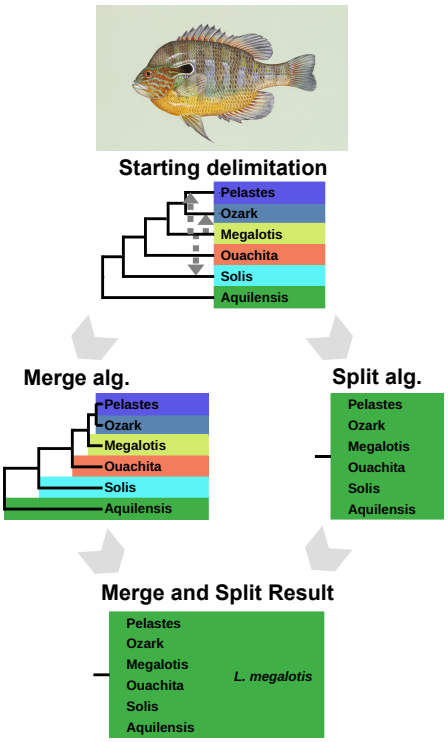

Figure 10: (**a**) The guide tree (starting delimitation) for the Longear Sunfish (*Lepomis megalotis*), with three migration events (from *megalotis* to *pelastes*, *solis*, and *ozark*) assumed in the BPP analysis, indicated by dotted lines. (**b**) Both merge and split algorithms support a single species. [show merge result, as in figures for the other datasets. We need care to detail, and consistency among datasets.][merge and split results are identical, updated figure subtitles to reflect this]

We reanalyzed the data, taking into account migration between the subspecies. Based on the hybridization patterns observed by Kim *et al.* (2022), migration from *megalotis* to *pelastes*, *solis*, and *ozark* was specified. The migration rate was assigned the gamma prior G(1, 100) with mean 0.01 migrant individuals per generation. Merge and split algorithms were run using gdi thresholds of 0.3 and 0.7 (control files in figs. S9 & S10).

Both merge and split analyses supported a single species. This is congruent with the delimitation of Kim *et al.* (2022), who calculated gdi under the MSC model without gene flow. [look at parameter estimates.]

Discussion

Challenges of heuristic species delimitation

In this paper we have developed a python pipeline to automate the procedure of hierarchical merge and split algorithms of heuristic species delimitation, initially discussed by Leaché *et al.* (2019). Our tests using both simulated and empirical datasets suggest that the implementation is correct, and the approach is less likely to suffer from over-splitting, which has been discussed extensively as a problem with the approach of Bayesian model selection (Yang and Rannala, 2010). We expect that the new implementation will be useful when evolutionary biologists want to integrate evidence based on multilocus genetic or genomic data in an integrated approach to species delimitation (Fujita *et al.*, 2012). The pipeline allows one to utilize the power of the MSC framework and the BPP program to estimate population parameters precisely and accurately using the ever-increasing genomic sequence data.

One should not expect the gdi or our pipeline to be a panacea that will apply to all cases of species delimitation problems using genomic data. First, we note that the gdi criterion and thus our pipeline may suffer from ambiguities. Suppose there are K populations on the guide tree, the merge algorithm may arrive at a high number of species (K_u) while the split algorithm at a low number (K_l), with $1 \le K_l \le K_u \le K$. When the two algorithms disagree, the gdi is ambiguous (Jackson *et al.*, 2017). Another ambiguity concerns the definition of gdi. In eqs. 2 or 4, we considered a sample of two A sequences and one B sequence. Similarly one may consider two B sequences and one A sequence. Under the MSC model of no gene flow, the two definitions are

$$\begin{aligned} gdi_A &= 1 - \mathrm{e}^{-2\tau_{AB}/\theta_A}, \\ gdi_B &= 1 - \mathrm{e}^{-2\tau_{AB}/\theta_B} \end{aligned} \tag{10}$$

(cf: eq. 2). When A and B have very different population sizes (θ_A, θ_B), the two definitions may be inconsistent concerning the species status of A and B. For example, A may appear to be a different species from B judged by gdi_A, but B may not appear to be a distinct species from A judged by gdi_B (Leaché *et al.*, 2019). One may insist on both gdi_A and gdi_B exceeding the threshold in our pipeline.

Second, a large gdi may result from a very small population size even if the divergence time is small. It may be advisable to examine the population size or the absolute divergence time together with the gdi (Rannala and Yang, 2020).

We note that alternative heuristic criteria can be similarly used in our pipeline. In particular, composite criteria can be used to determine the species status. For example, we may insist, besides the gdi cutoff, that the species split time reach a minimum of 10^4 generations (Rannala and Yang, 2020) or the migration rate between the two species cannot exceed $M = Nm = 0.1$ migrants per generation.

Gene flow and species status

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). Good species are recognised even if there is significant evidence for gene flow between them.

It is not so clear how to incorporate gene flow in the hierarchical merge and split algorithms. In the framework of Bayesian model selection, there are three models or scenarios for two populations A, B: (i) M_1 one species, (ii) M_{20} Two species with no gene flow, and (iii) M_{21} two species with gene flow. Leaché *et al.* (2019) compare M_1 and M_{20} to decide whether there are one or two species. Alternatively one may insist on species status only if there is no significant amount of gene flow (i.e., only if M_{20} wins over both M_1 and M_{21}). This approach may suffer from over-lumping.

In Leaché *et al.* (2019), we used MSC with no gene flow to construct the guide tree and to calculate the gdi. The resulting guide tree (Leaché *et al.*, 2019, figure 3b) was incorrect, even though it led to the correct inference of two species ($ABCD$ and X). If we use the correct MSC-M model and correct guide tree of figure 4a, the hierarchical merge and split algorithms will never recover the correct delimitation of two species. One approach may be to allow the merge of populations that exchange migrants at a certain rate even if they are not sister lineages on the guide tree. For example, in the case of figure 4a, we may attempt to merge AB, BC, and CD, if the migration rate between the species pairs exceed a certain threshold.

The guide tree and paraphyletic species

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 and Rannala, 2014; Rannala and Yang, 2017). Alternatives include maximum likelihood tree inference using concatenated data, or use of the mitochondrial genes.

We note that the hierarchical merge and split algorithms do not work when a species is paraphyletic. The guide tree of figure 4 is such an example. Can we think of any ideas for delimiting paraphyletic species? In the case of no gene flow, paraphyletic species does not seem to make sense. When there is gene flow between the geographic populations of the paraphyletic species, perhaps we can allow the merging of such populations if there is significant evidence for gene flow. This problem exist for the reversible-jump algorithms of Yang and

Rannala (2010) as well.

Note that the discussion here concerns the non-monophyly of the populations, rather than non-monophyly of gene trees, which is known to be problematic if used to delimit species Knowles and Carstens (2007).

Cite Sukumaran *et al.* (2021); Solis-Lemus *et al.* (2015).

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

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.

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.

Degnan, J. H. 2018. Modeling hybridization under the network multispecies coalescent. *Syst. Biol.*, 67(5): 786–799.

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.

Fujita, M. K., Leaché, A. D., Burbrink, F. T., McGuire, J. A., and Moritz, C. 2012. Coalescent-based species delimitation in an integrative taxonomy. *Trends Ecol. Evol.*, 27: 480–488.

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.

Hobolth, A., Andersen, L., and Mailund, T. 2011. On computing the coalescence time density in an isolation-with-migration model with few samples. *Genetics*, 187: 1241–1243.

Hudson, R. R. and Turelli, M. 2003. Stochasticity overrules the "three-times rule": genetic drift, genetic draft, and coalescence times for nuclear loci versus mitochondrial DNA. *Evolution*, 57: 182–190.

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.

Knowles, L. L. and Carstens, B. C. 2007. Delimiting species without monophyletic gene trees. *Syst. Biol.*, 56: 887–895.

Kubatko, L. 2019. The multispecies coalescent. In D. Balding, I. Moltke, and J. Marioni, editors, *Handbook of Statistical Genomics*, pages 219–245. Wiley, New York, 4th edition.

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.

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.

Sites, J. and Marshall, J. C. 2004. Delimiting species: A renaissance issue in systematic biology. *Trends Ecol. Evol.*, 18: 462–470.

Solis-Lemus, C., Knowles, L. L., and Ane, C. 2015. Bayesian species delimitation combining multiple genes and traits in a unified framework. *Evolution*, 69(2): 492–507.

Sukumaran, J. and Knowles, L. 2017. Multispecies coalescent delimits structure, not species. *Proc. Natl. Acad. Sci. USA.*, 114: 1607–1612.

Sukumaran, J., Holder, M. T., and Knowles, L. L. 2021. Incorporating the speciation process into species delimitation. *PLoS Comput. Biol.*, 17(5): e1008924.

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. 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. 2014. Unguided species delimitation using DNA sequence data from multiple loci. *Mol. Biol. Evol.*, 31: 3125–3135.

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.

Zhu, T. and Yang, Z. 2012. Maximum likelihood implementation of an isolation-with-migration model with three species for testing speciation with gene flow. *Mol. Biol. Evol.*, 29: 3131–3142.

*

```
# output
output_directory = res_sim_merge

# input files
Imapfile = Leache_2019_starting_populations.txt
seqfile = Leache_2019_sequences.txt

# guide tree
guide_tree = ((A, B), (C, D)), X);

# hierarchical algo. parameters
mode = merge
GDI threshold = <0.2

# MCMC settings
threads = 4
burnin = 50000
nsample = 100000
```

Figure S1: Control file for the ABCDX merge analysis (see figs. 5&S2). `output_directory` specifies the directory in which result files will be written. `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 (`merge` or `split`). `GDI_threshold` specifies the *gdi* value below which two populations are merged into a candidate species. `threads` specifies the number of cpu threads used to run BPP. `burnin`, `nsample` and `sampfreq` specify the MCMC settings for run BPP.

```
Number of species in starting delimitation:  5
((A, B), (C, D)), X);

*** Iteration 1 ***

Inferred tau and theta parameters:
        theta   tau
X       0.0098
A       0.0138
B       0.0202
C       0.0212
D       0.0263
ABCDX   0.0368  0.0101
ABCD    0.0432  0.0006
AB      0.0128  0.0003
CD      0.0204  0.0002

Proposal results:
Node pair        gdi 1   gdi 2   merge accepted?
'A', 'B'         0.05    0.03    True
'C', 'D'         0.03    0.02    True

Number of species after iteration 1:  3
((AB, CD), X);

*** Iteration 2 ***

Inferred tau and theta parameters:
        theta   tau
X       0.0098
AB      0.0194
CD      0.0286
ABCDX   0.0367  0.0101
ABCD    0.0428  0.0006

Proposal results:
Node pair        gdi 1   gdi 2   merge accepted?
'AB', 'CD'       0.07    0.05    True

Number of species after iteration 2:  2
(ABCD, X);

*** Iteration 3 ***

Inferred tau and theta parameters:
        theta   tau
X       0.0098
ABCD    0.0441
ABCDX   0.0365  0.0102


Proposal results:
Node pair        gdi 1   gdi 2   merge accepted?
'ABCD', 'X'      0.39    0.87    False

Number of species after iteration 3:  2
(ABCD, X);

Final delimitation reached.
```

Figure S2: Screen output from running the hierarchical merge algorithm to analyze the simulated dataset of figure 4.

```
# Notes: species are renamed as follows
#
# gir_ang = giraffa+angolensis
# tip_tho = tippelskirchi+thornicrofti
# cam_rot_ant = camelopardalis+rothschildi+antiquorum
# per = peralta
# ret = reticulata

# output
output_directory = res_giraffe_merge

# input files
Imapfile = Imap_Giraffe.txt
seqfile = MSA_Giraffe.txt

# guide tree
guide_tree = ((gir_ang,tip_tho),((cam_rot_ant,per),ret));

# migration events and priors
migration = {
    ret -> tip_tho,
    tip_tho -> ret,
    ret -> cam_rot_ant,
    cam_rot_ant -> ret,
}
migprior = 0.1 10

# hierarchical algorithm settings
mode = merge
gdi_threshold = <0.3

# MCMC settings
threads = 4
burnin = 50000
nsample = 200000
sampfreq = 2
```

Figure S3: Control file for the merge algorithm in analysis of the Giraffes data. The results are in figure 6.

```
# output
output_directory = res_giraffe_split

# input files
Imapfile = Imap_Giraffe.txt
seqfile = MSA_Giraffe.txt

# guide tree
guide_tree = ((gir_ang,tip_tho),((cam_rot_ant,per),ret));

# migration events and priors
migration = {
 ret -> tip_tho,
 tip_tho -> ret,
 ret -> cam_rot_ant,
 cam_rot_ant -> ret,
}
migprior = 0.1 10

# hierarchical algo. parameters
mode = split
gdi_threshold = >0.7

# MCMC settings
threads = 4
burnin = 50000
nsample = 200000
sampfreq = 2
```

Figure S4: Control file for the split analysis in Giraffes data. The results are in figure 6.

```
# Notes: species are renamed as follows
#
# Po = polyzona
# Ab = abnorma
# Mi = micropholis
# An = annulata
# Ge = gentilis
# Tr = triangulum
# El = elapsoides

# output
output_directory = res_milksnake_merge

# input files
Imapfile = Imap_Lampropeltis.txt
seqfile = MSA_Lampropeltis.txt

# guide tree
guide_tree = (((Po, (Ab, Mi)), (An, (Ge, Tr))), El);

# hierarchical algorithm settings
mode = merge
gdi_threshold = <0.3

# MCMC settings
threads = 4
burnin = 50000
nsample = 200000
sampfreq = 2
```

Figure S5: Control file for the merge analysis of the Milksnakes data. The results are presented in figure 8.

```
# output
output_directory = output_directory = res_milksnake_split

# input files
Imapfile = Imap_Lampropeltis.txt
seqfile = MSA_Lampropeltis.txt

# guide tree
guide_tree = (((Po, (Ab, Mi)), (An, (Ge, Tr))), El);

# hierarchical algorithm settings
mode = split
gdi_threshold = >0.7

# MCMC settings
threads = 4
burnin = 50000
nsample = 200000
sampfreq = 2
```

Figure S6: Control file for the split analysis of the Milksnakes data. The results are presented in figure 8.

```
# output
output_directory = # will be set from the command line

# input files
Imapfile = # will be set from the command line
seqfile = trigentalt.txt

guide_tree = ((Ge,Tr),Al);

mode = merge
gdi_threshold = <0.3

threads = 4
burnin = 50000
nsample = 100000
sampfreq = 2
```

Figure S7: Control file for the analysis of the milksnake data under delimitation hypotheses reflecting the East-West splits. The results are presented in figure 8. The `Imapfile` and `output_directory` parameters are left empty, as they will be provided at the command line, to ensure that the same basic control file can be used for each of the five alternative East-West delimitations.

```
HMDelimit --mcfile mcf_milksnake_EW.txt --mcfpor \
Imapfile = trigent1alt.Imap.txt, output_directory = res_EW_1

HMDelimit --mcfile mcf_milksnake_EW.txt --mcfpor \
Imapfile = trigent2alt.Imap.txt, output_directory = res_EW_2

HMDelimit --mcfile mcf_milksnake_EW.txt --mcfpor \
Imapfile = trigent3alt.Imap.txt, output_directory = res_EW_3

HMDelimit --mcfile mcf_milksnake_EW.txt --mcfpor \
Imapfile = trigent4alt.Imap.txt, output_directory = res_EW_4

HMDelimit --mcfile mcf_milksnake_EW.txt --mcfpor \
Imapfile = trigent5alt.Imap.txt, output_directory = res_EW_5
```

Figure S8: Shell script used to iterate through alternative East-West delimitation hypotheses in Milksnakes, presented in figure 8. The `-mcfpor` (control file parameter override) flag is used to override parameters of the mcf via the command line interface, setting the Imap file to one of the alternative East-West delimitations, and specifying the individual output directories for each analysis.

```
# Notes: species are renamed:
#
# PEL = pelastes
# OZK = ozark
# MEG = megalotis
# LIT = ouachita
# SOL = solis
# AQU = aquilensis

# output
output_directory = res_sunfish_merge

# input files
Imapfile =  Imap_Sunfish.txt
seqfile = MSA_Sunfish.txt

# guide tree
guide_tree = (((((PEL,OZK),MEG),LIT),SOL),AQU);

# migration events and priors
migration = {
    MEG -> PEL,
    MEG -> SOL,
    MEG -> OZK
}
migprior = 0.1 10

# hierarchical algo. parameters
mode = merge
gdi_threshold = <0.3

# MCMC settings
threads = 4
burnin = 50000
nsample = 200000
sampfreq = 2
```

Figure S9: Control file for the merge analysis of the sunfish data. The results are presented in figure 10.

```
# output
output_directory = res_sunfish_split

# input files
Imapfile =  Imap_Sunfish.txt
seqfile = MSA_Sunfish.txt

# guide tree
guide_tree = (((((PEL,OZK),MEG),LIT),SOL),AQU);

# migration events and priors
migration = {
    MEG -> PEL,
    MEG -> SOL,
    MEG -> OZK
}
migprior = 0.1 10

# hierarchical algo. parameters
mode = split
gdi_threshold = >0.7

# MCMC settings
threads = 4
burnin = 50000
nsample = 200000
sampfreq = 2
```

Figure S10: Control file for the split analysis of the sunfish data. The results are presented in figure 10.

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 (x, y, and k)