Species Delimitation

Bruce Rannala, UC Davis

Cave painting, Lascaux, France, 15,000 to 10,000 B.C.

Many Species Delimitations are Unambiguous

Cryptic Species

Eleutherodactylus ockendeni

Cryptic species complex

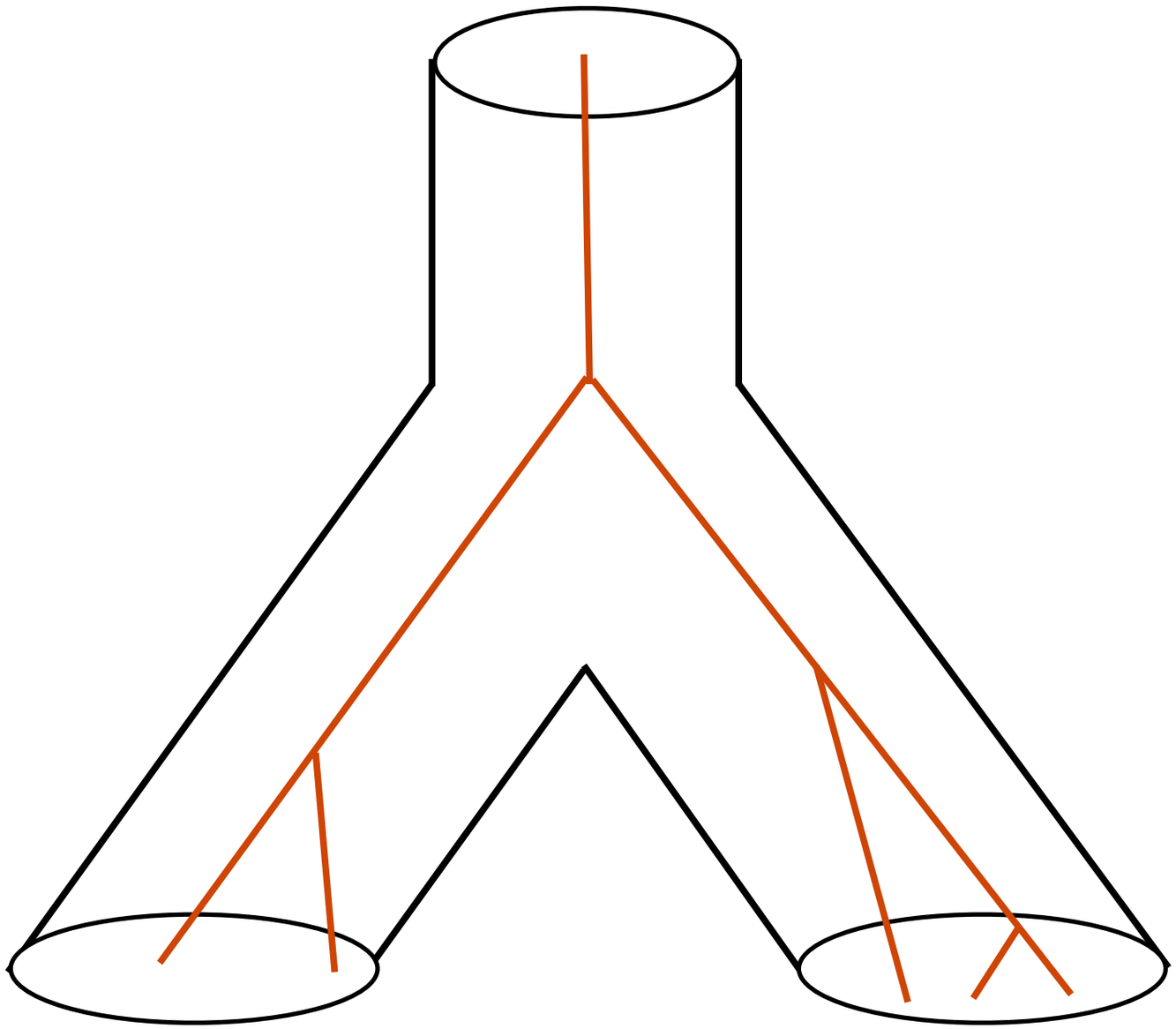
From Wikipedia, the free encyclopedia

a **cryptic species complex** is a group of species which satisfy the biological definition of species—that is, they are reproductively isolated from each other—but whose morphology is very similar (in some cases virtually identical).

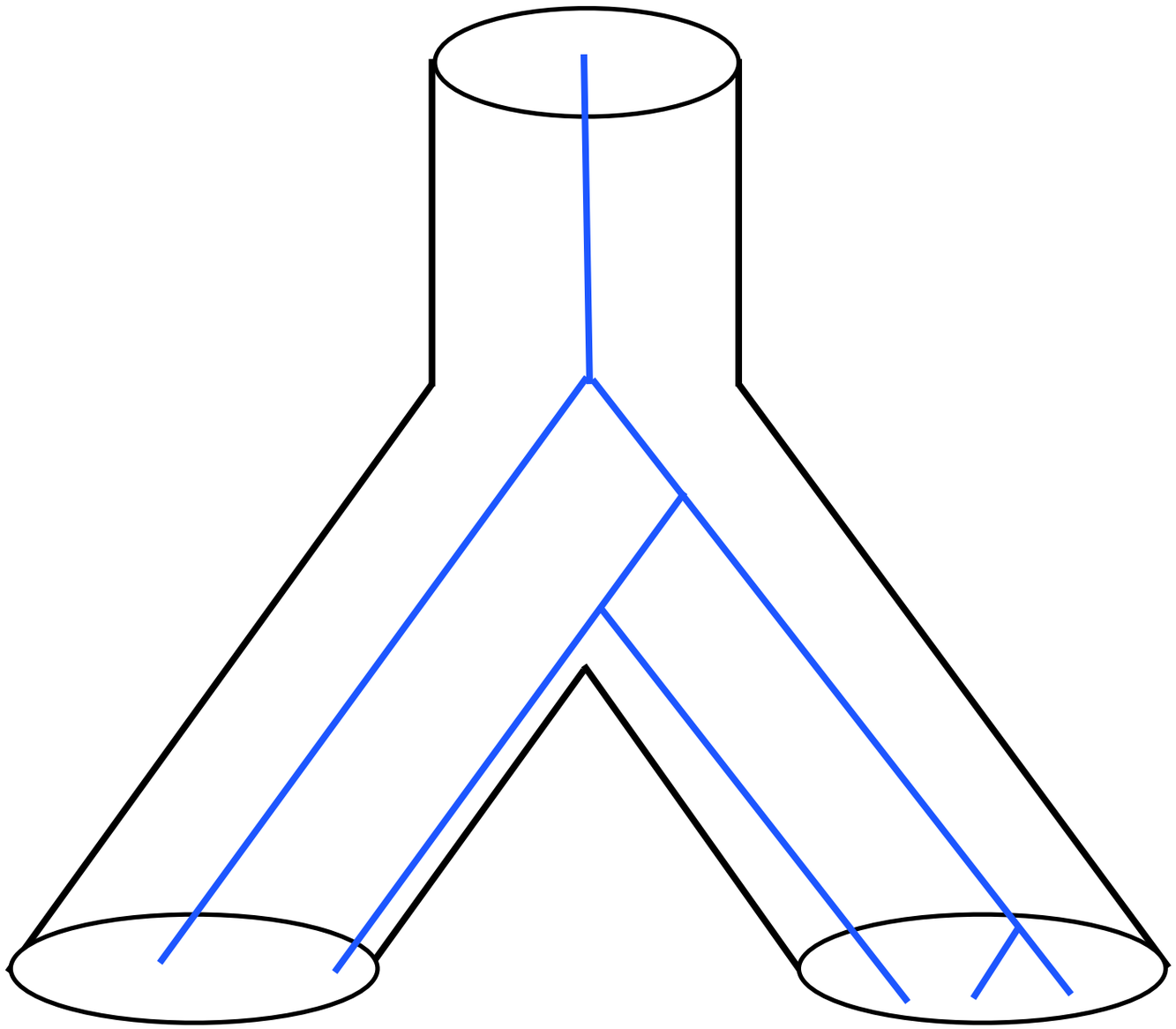
How to Delimit Species?

- morphological diagnostics
- phylogenetics (monophyly)
- population genetics (isolation)

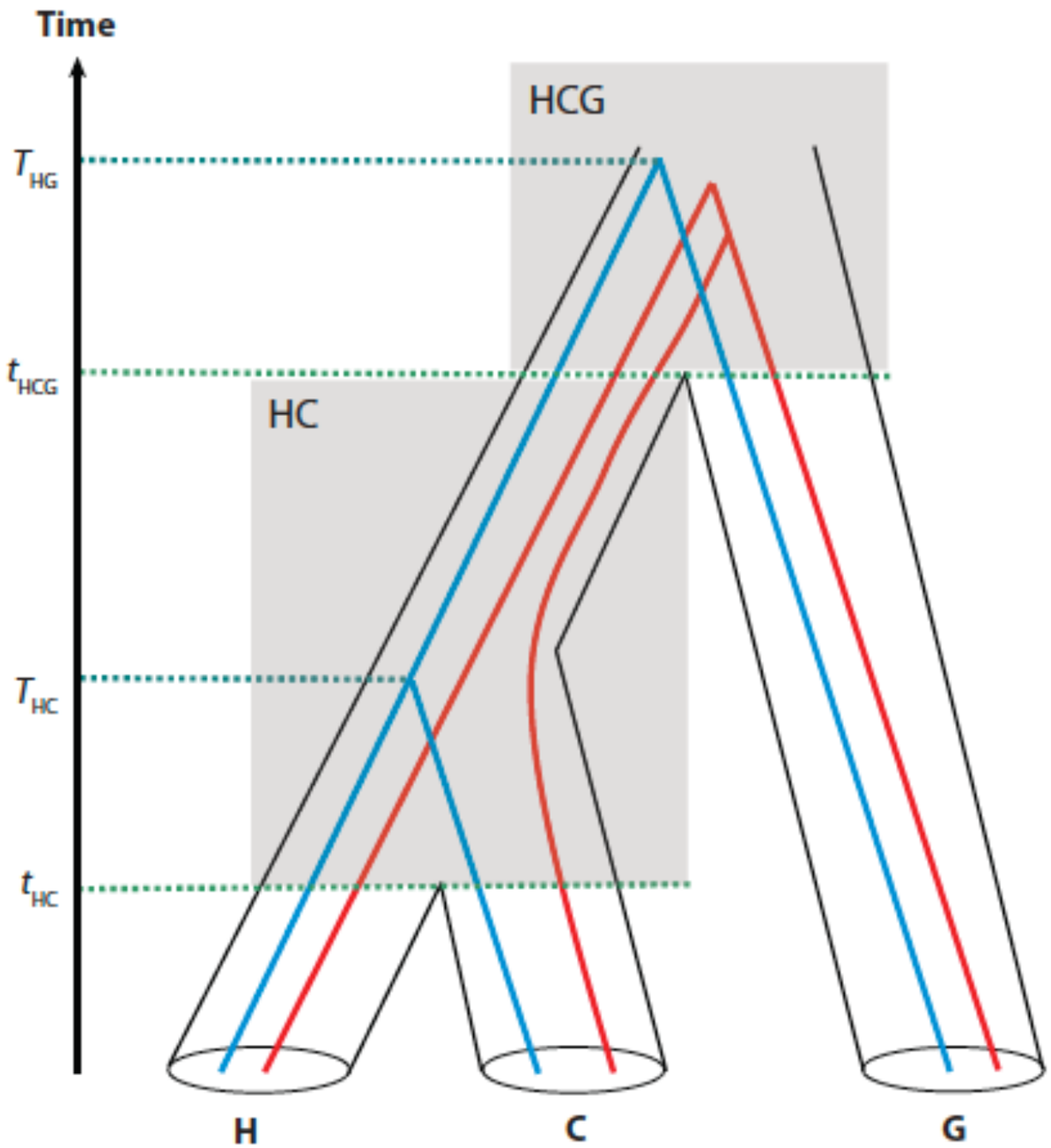
Reciprocal Monophyly



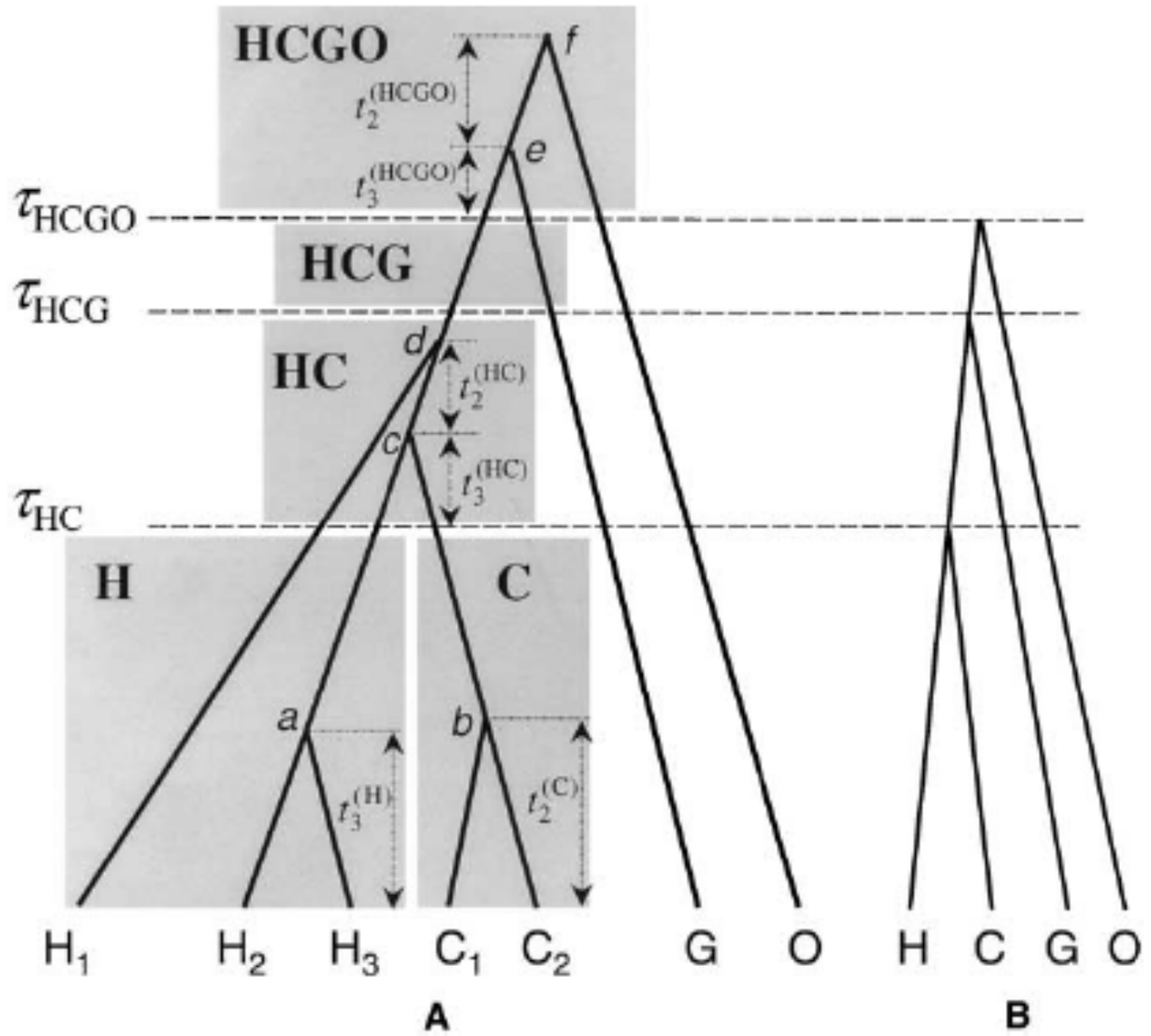
Lineage Sorting



Species Are Genetically Isolated



Multispecies Coalescent Processes



Bayesian species delimitation using multilocus sequence data

Ziheng Yang[a,b] and Bruce Ra

[a]Center for Computational and Evolu
University College London, London W
Davis, CA 95616

Edited by Scott V. Edwards, Harvard U

In the absence of recent admixtu
of individuals in gene trees t
potentially be used to infer the
This approach to species delimita
has been constrained by the fac
loci are often poorly resolved ar
hybridization, and other populat
discordant gene trees. Here we u
to generate the posterior prob
taking account of uncertainties
the ancestral coalescent process.
specified guide tree to avoid int
delimitations. The statistical perf
ined using simulations, and the
sequence data from rotifers, fence lizards, and

Bayesian phylogenetic inference | biological species c
Markov chain Monte Carlo | reversible jump

Accurate species delimitations are of crit
many areas of biology, such as conserva
ignating endangered species), epidemiology
pathogens), and evolutionary biology (describ
versification). Traditionally, species have be
described using morphological traits. Howev

Unguided Species Delimitation Using DNA Sequence Data from Multiple Loci

Ziheng Yang[1,2] and Bruce Rannala*[,1,3]

[1]Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing, China

[2]Department of Genetics, Evolution and Environment, University College London, London, United Kingdom

[3]Department of Evolution & Ecology, University of California, Davis

***Corresponding author:** E-mail: brannala@ucdavis.edu.

Associate editor: Yoko Satta

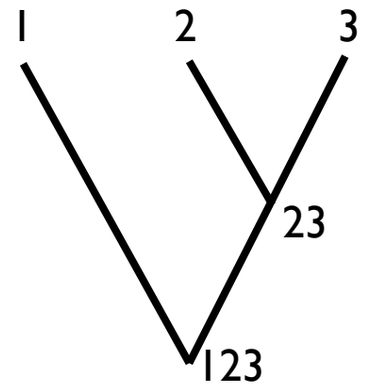
Improved Reversible Jump Algorithms for Bayesian Species Delimitation

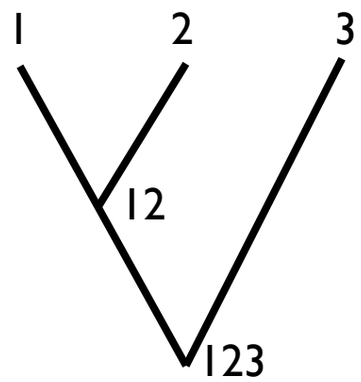
Bruce Rannala*[,†,‡] and Ziheng Yang*[,§,¶]

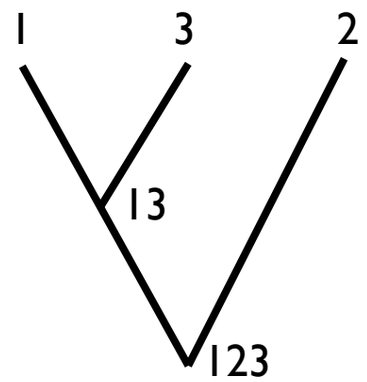
*Center for Computational and Evolutionary Biology, Institute of Zoology, Chinese Academy of Sciences, Beijing 100101, China, †Genome Center and Department of Evolution and Ecology, University of California, Davis, California 95616, ‡Laboratory of Alpine Ecology, Université Joseph Fourier, Grenoble 38041, France, and §Department of Biology, University College London, London WC1E 6BT, United Kingdom

ABSTRACT Several computational methods have recently been proposed for delimiting species using multilocus sequence data. Among them, the Bayesian method of Yang and Rannala uses the multispecies coalescent model in the likelihood framework to calculate the posterior probabilities for the different species-delimitation models. It has a sound statistical basis and is found to have nice statistical properties in simulation studies, such as low error rates of undersplitting and oversplitting. However, the method suffers from poor mixing of the reversible-jump Markov chain Monte Carlo (rjMCMC) algorithms. Here, we describe several modifications to the algorithms. We propose a flexible prior that allows the user to specify the probability that each node on the guide tree represents a true speciation event. We also introduce modifications to the rjMCMC algorithms that remove the constraint on the new species divergence time when splitting and alter the gene trees to remove incompatibilities. The new algorithms are found to improve mixing of the Markov chain for both simulated and empirical data sets.

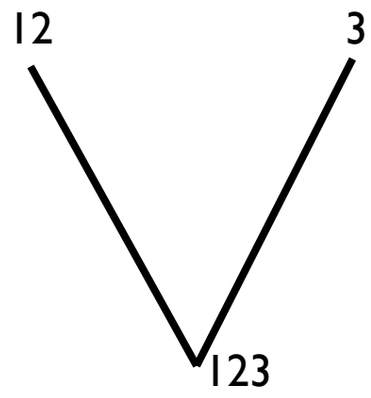
$$\Lambda_1 = \{1\}, \{2\}, \{3\}.$$



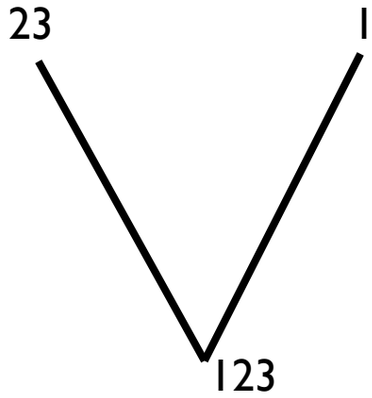




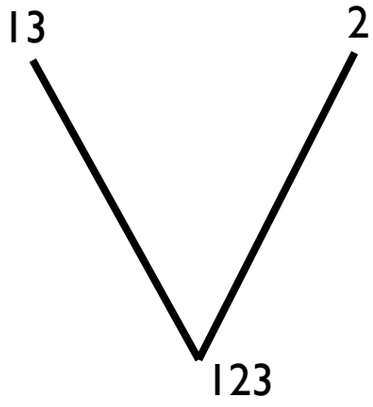
$$\Lambda_2 = \{1, 2\}, \{3\}.$$



$$\Lambda_3 = \{2, 3\}, \{1\}.$$



$$\Lambda_4 = \{1, 3\}, \{2\}.$$



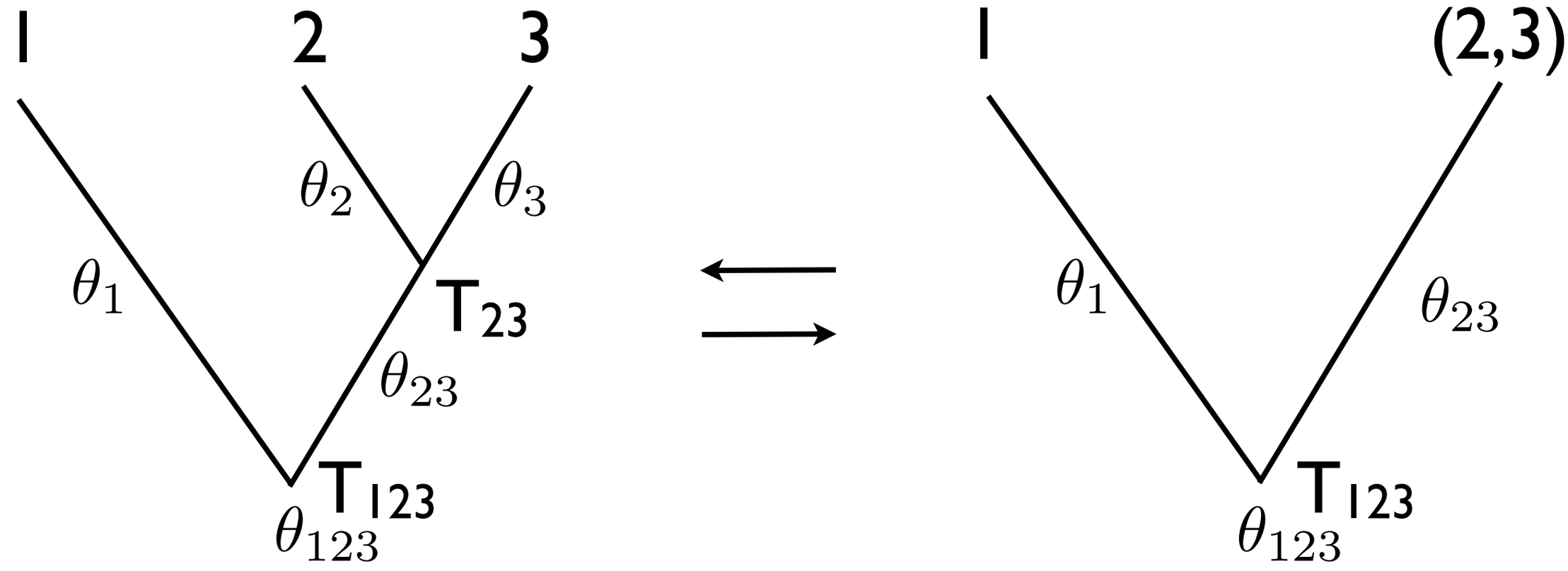
$$\Lambda_5 = \{1, 2, 3\}.$$

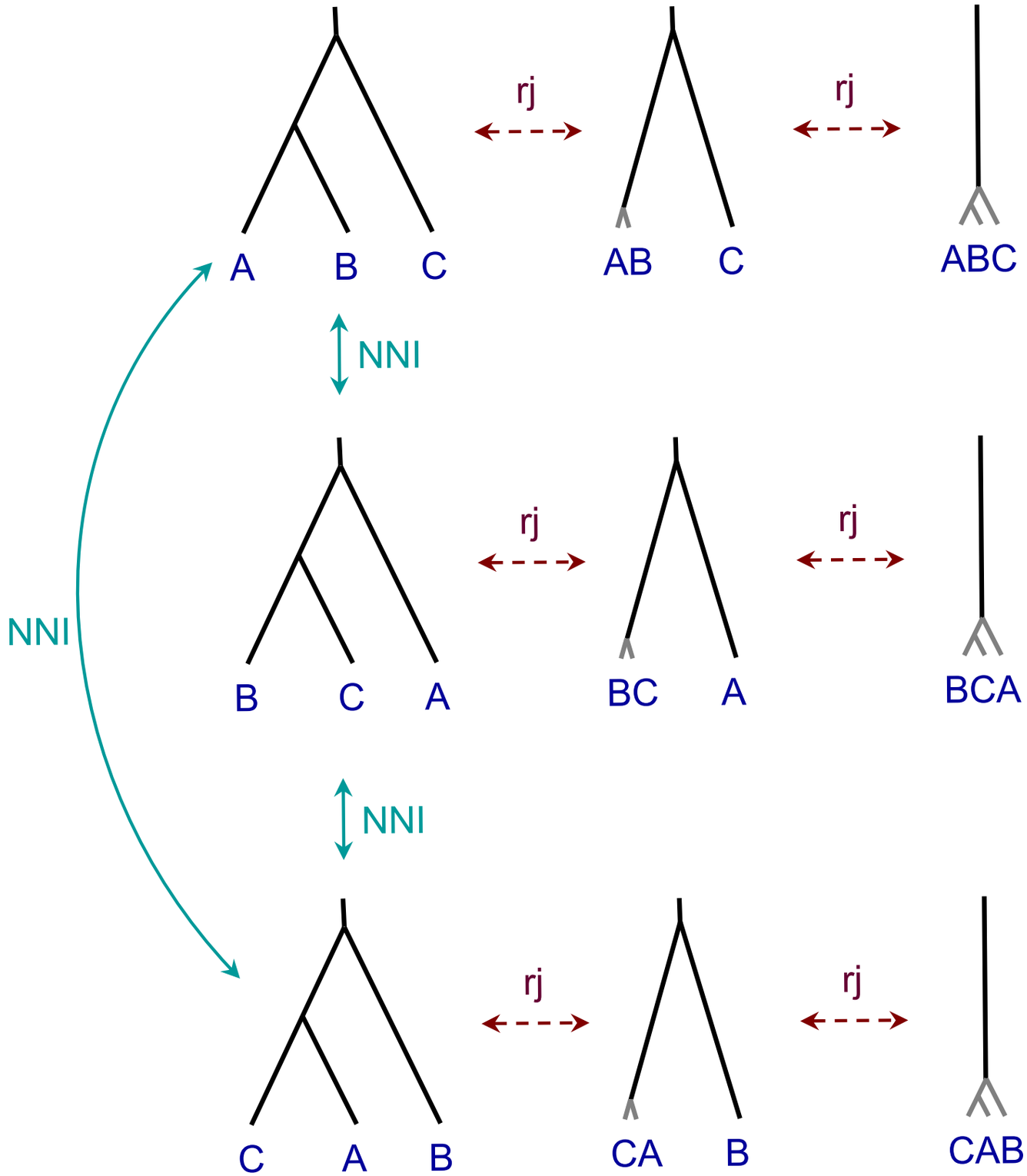
$$f(S, \Lambda|D) = \frac{1}{f(D)} f(D|S) f(S|\Lambda) f(\Lambda)$$

$$f(D|S) = \int_G f(D|G) f(G|S) dG$$

$$\begin{aligned} D &= \text{Multi-locus sequence data} \\ S &= \text{Species tree} \\ G &= \text{Multi-locus gene trees} \\ \Lambda &= \text{Species delimitation (partition)} \end{aligned}$$

Reversible-jump MCMC (rjMCMC)




rj
rj
A
B
C
AB
C
ABC
NNI
NNI
rj
rj
B
C
A
BC
A
BCA
NNI
rj
rj
C
A
B
CA
B
CAB

Priors on Model Parameters

tauprior = Gamma(a,b) distribution
(age of root of species tree)

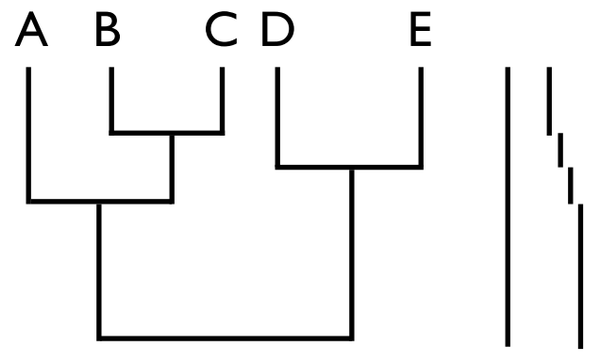
thetaprior = Gamma(a,b) distribution
$\theta = 4N_e\mu$

Topology prior:
(1) Uniform labeled history
(2) Uniform rooted trees
(3) Uniform delimitations

Prior on Species Trees:

Yule or Birth-death process (*Beast)

Dirichlet distribution conditioned on root age (BPP)



Prior on Topology:

Uniform on labelled histories (*Beast, BPP)

Uniform on trees (BPP)

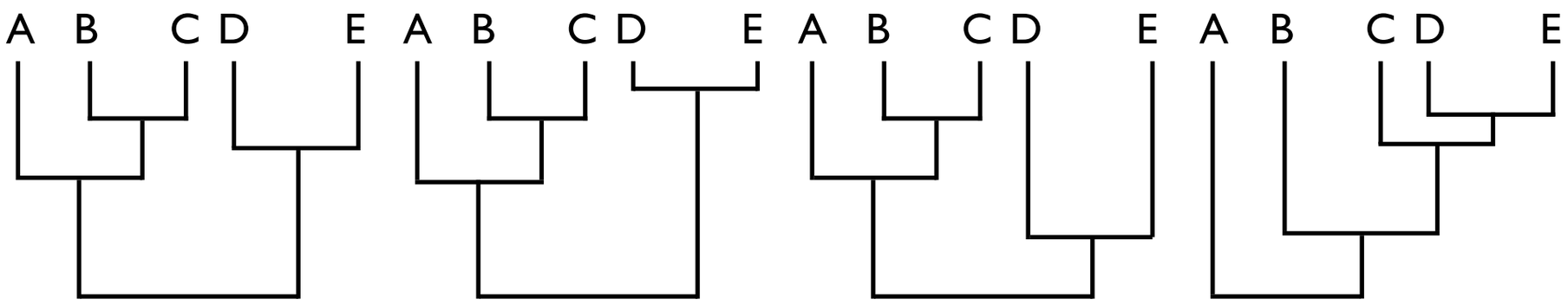


Table 1. Prior Probability for the Number of Delimited Species under Prior 1 (uniform distribution for rooted trees).

Number of Delimited Species	Number of Delimitations	Number of Rooted Trees	Number of Guide Trees	Product	Probability
$s = 3$ populations					
$d = 1$	1	1	3	3	$P_1 = 3/9 = 1/3 = 0.333$
$d = 2$	3 (1 2)	1	1	3	$P_2 = 3/9 = 1/3 = 0.333$
$d = 3$	1 (1 1 1)	3	1	3	$P_3 = 3/9 = 1/3 = 0.333$
$s = 4$ populations					
$d = 1$	1	1	15	15	$P_1 = 15/63 = 5/21 = 0.238$
$d = 2$	3 (2 2)	1	1	3	$P_2 = (3 + 12)/63 = 5/21 = 0.238$
	4 (1 3)	1	3	12	
$d = 3$	6 (1 1 2)	3	1	18	$P_3 = 18/63 = 6/21 = 0.286$
$d = 4$	1	15	1	15	$P_4 = 15/63 = 5/21 = 0.238$
$s = 5$ populations					
$d = 1$	1	1	105	105	$P_1 = 105/600 = 7/40 = 0.175$
$d = 2$	5 (1 4)	1	15	75	$P_2 = (75 + 30)/600 = 7/40 = 0.175$
	10 (2 3)	1	3	30	
$d = 3$	10 (1 1 3)	3	3	90	$P_3 = (90 + 45)/600 = 9/40 = 0.225$
	15 (1 2 2)	3	1	45	
$d = 4$	10 (1 1 1 2)	15	1	150	$P_4 = 150/600 = 10/40 = 0.250$
$d = 5$	1	105	1	105	$P_5 = 105/600 = 7/40 = 0.175$
$s = 6$ populations					
$d = 1$	1	1	945	945	$P_1 = 945/7245 = 3/23 = 0.130$
$d = 2$	6 (1 5)	1	105	630	$P_2 = (630 + 225 + 90)/7245$
	15 (2 4)	1	15	225	$= 3/23 = 0.130$
	10 (3 3)	1	9	90	
$d = 3$	15 (1 1 4)	3	15	675	$P_3 = (675 + 540 + 45)/7245$
	60 (1 2 3)	3	3	540	$= 4/23 = 0.174$
	15 (2 2 2)	3	1	45	
$d = 4$	20 (1 1 1 3)	15	3	900	$P_4 = (900 + 675)/7245$
	45 (1 1 2 2)	15	1	675	$= 5/23 = 0.217$
$d = 5$	15 (1 1 1 1 2)	105	1	1,575	$P_5 = 1575/7245 = 5/23 = 0.217$
$d = 6$	1	945	1	945	$P_6 = 945/7245 = 3/23 = 0.130$

Prior on Delimitation and Topology:

Uniform on labelled histories (*Beast, BPP)

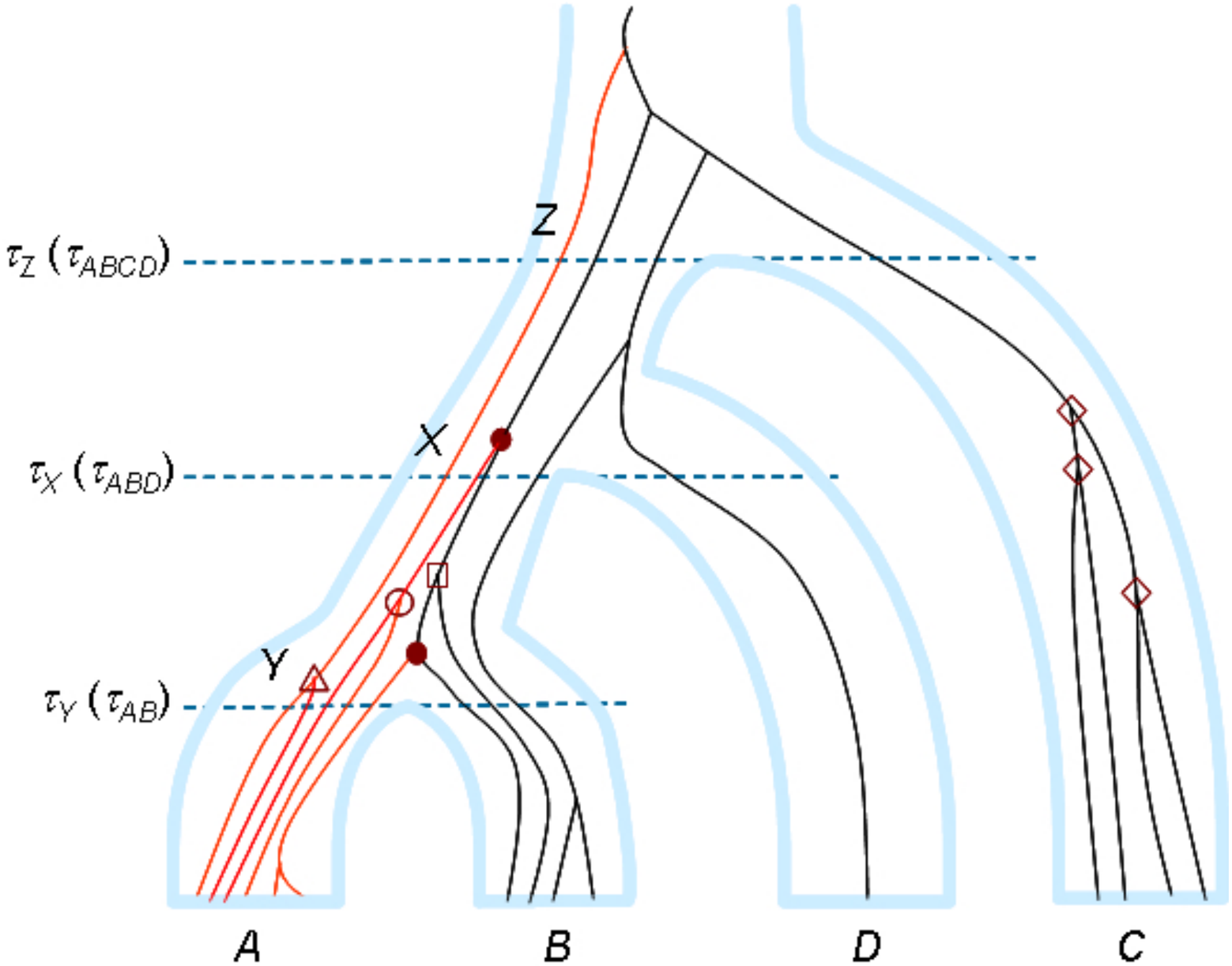
Uniform on rooted trees (BPP)

Uniform on number of delimited species

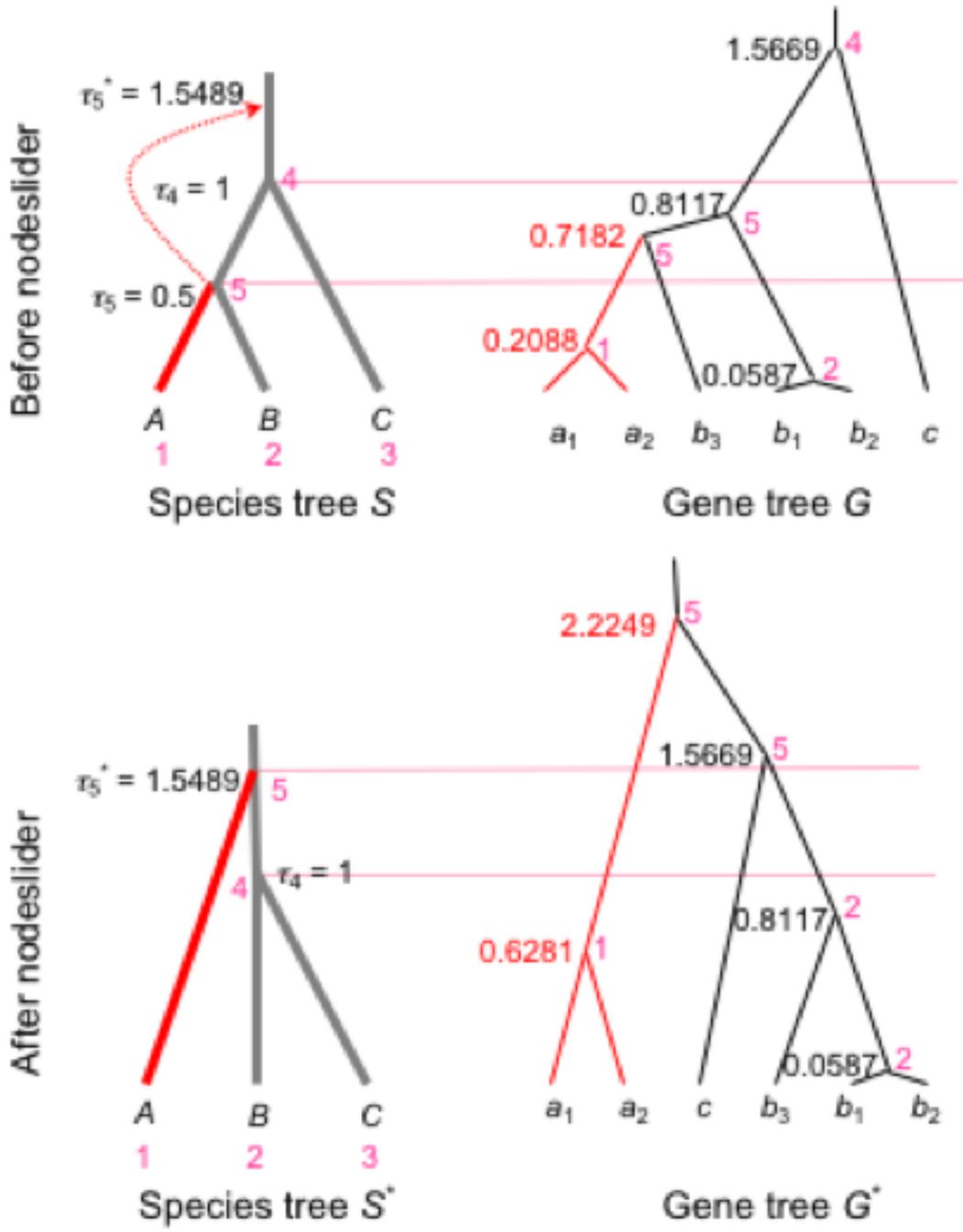
Making it work takes a little longer, Making it work takes a little time,…
Doug and the Slugs, 1982

<u>Making Species Tree Inference Work</u>:
Joint proposals that simultaneously alter all gene trees and the species tree.
(1) MSC|SPR move
(2) MSC|Node-slider move

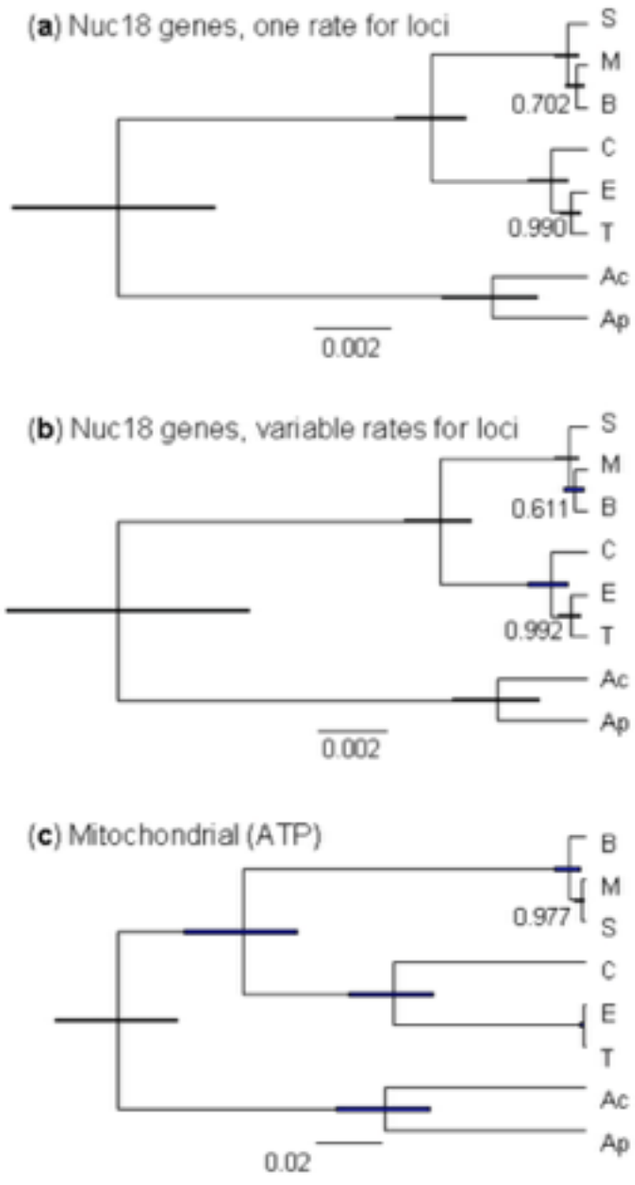
MSCISPR Move



MSCINode-slider Move



Species Tree Inference: Rattlesnakes



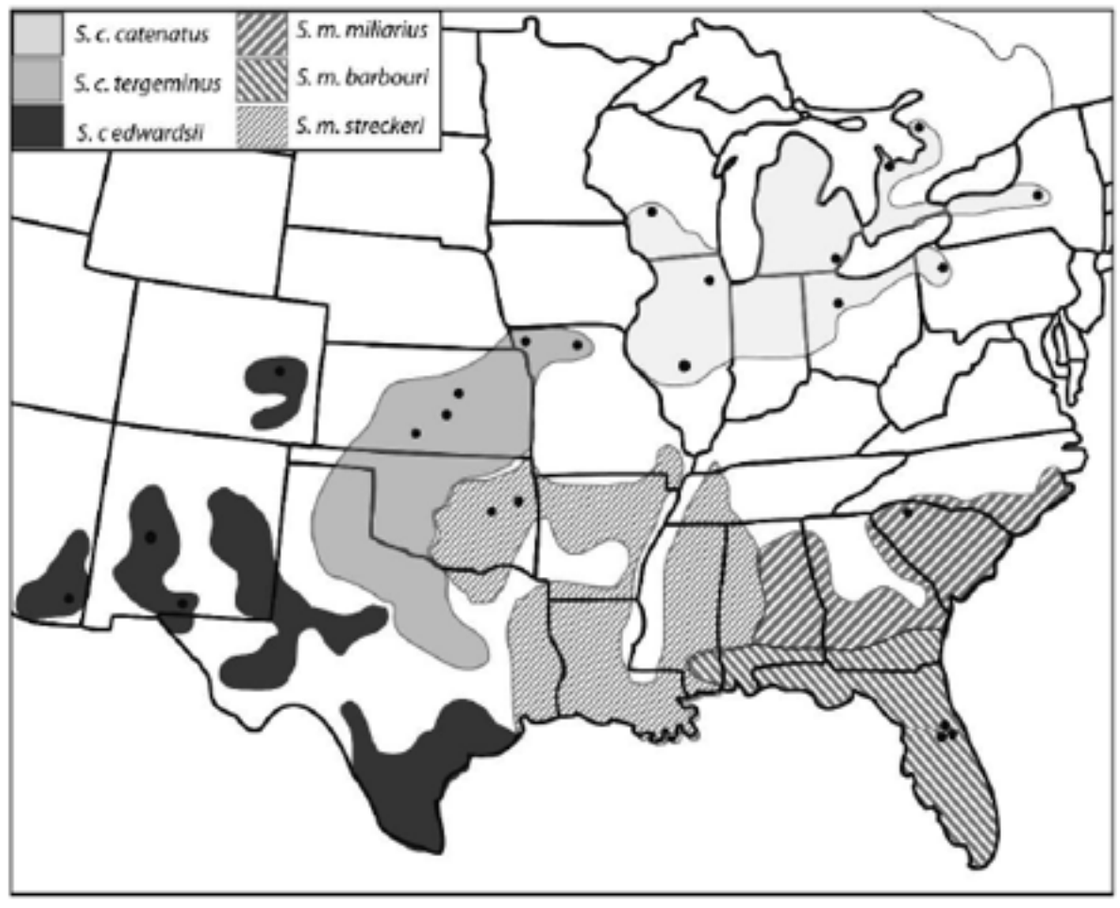
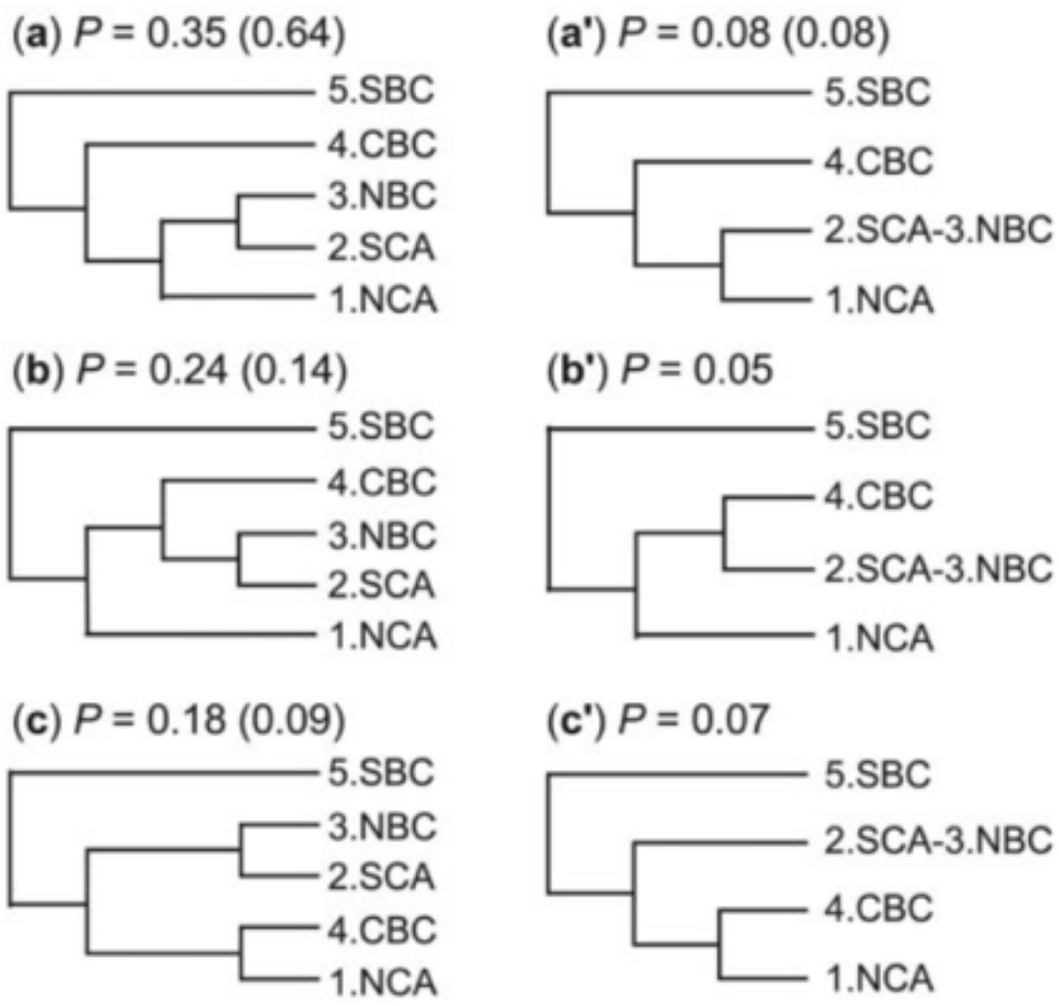


Figure 6: The MAP trees for the six subspecies of *Sistrurus* rattlesnakes and the outgroups in three different analyses of the nuclear (18 loci) and mitochondrial datasets. The three *S. catenatus* subspecies are *S. c. catenatus* (C), *S. c. tergeminus* (T), and *S. c. edwardsii* (E), while the three *S. miliarius* subspecies are *S. m. miliarius* (M), *S. m. barbouri* (B), and *S. m. streckeri* (S). The numbers next to the internal nodes are the posterior probabilities for the clades in the species tree (analysis A01: speciesdelimitation = 0, speciestree = 1). The branch lengths are drawn to represent the posterior means of the divergence times (τs) in the A00 analysis (speciesdelimitation = 0, speciestree = 0), with the phylogeny fixed, while the node bars represent the 95% HPD interval.

Species delimitation: Adam's Lizards



Analysis of Two Empirical Data Sets

The Coast Horned Lizard Data

The first data set we analyze includes two nuclear loci (*BDNF*: 132 sequences, 529 bp; and *RAG-1*: 136 sequences, 1,100 bp) sampled from coast horned lizards originally published by Leaché et al. (2009) and previously reanalyzed by Rannala and Yang (2013). Assignment is based on an mtDNA phylogeny, with five phylogeographic groups arranged latitudinally: North California (1.NCA), South California (2.SCA), Northern Baja California (3.NBC), Central Baja California (4.CBC), and South Baja California (5.SBC) (see fig. 8). There are thus five populations in the BPP analysis. We use the same priors as in Rannala and Yang (2013): $\tau_0 \sim G(2, 1000)$ for the root of the species tree and $\theta \sim G(2, 100)$. After a burn-in of 4,000 iterations, we took 2×10^5 samples, sampling every four iterations. Multiple runs using both rjMCMC algorithms 0 and 1 were used to ensure consistency between runs. Each run took about 9 h.

Bears in a Forest of Gene Trees: Phylogenetic Inference Is Complicated by Incomplete Lineage Sorting and Gene Flow

Verena E. Kutschera,*[,1] Tobias Bidon,[1] Frank Hailer,[1] Julia L. Rodi,[1] Steven R. Fain,[2] and Axel Janke*[,1,3]

[1]Biodiversity and Climate Research Centre (BiK-F), Senckenberg Gesellschaft für Naturforschung, Frankfurt am Main, Germany

[2]National Fish and Wildlife Forensic Laboratory, Ashland, OR

[3]Institute for Ecology, Evolution and Diversity, Goethe University Frankfurt, Frankfurt am Main, Germany

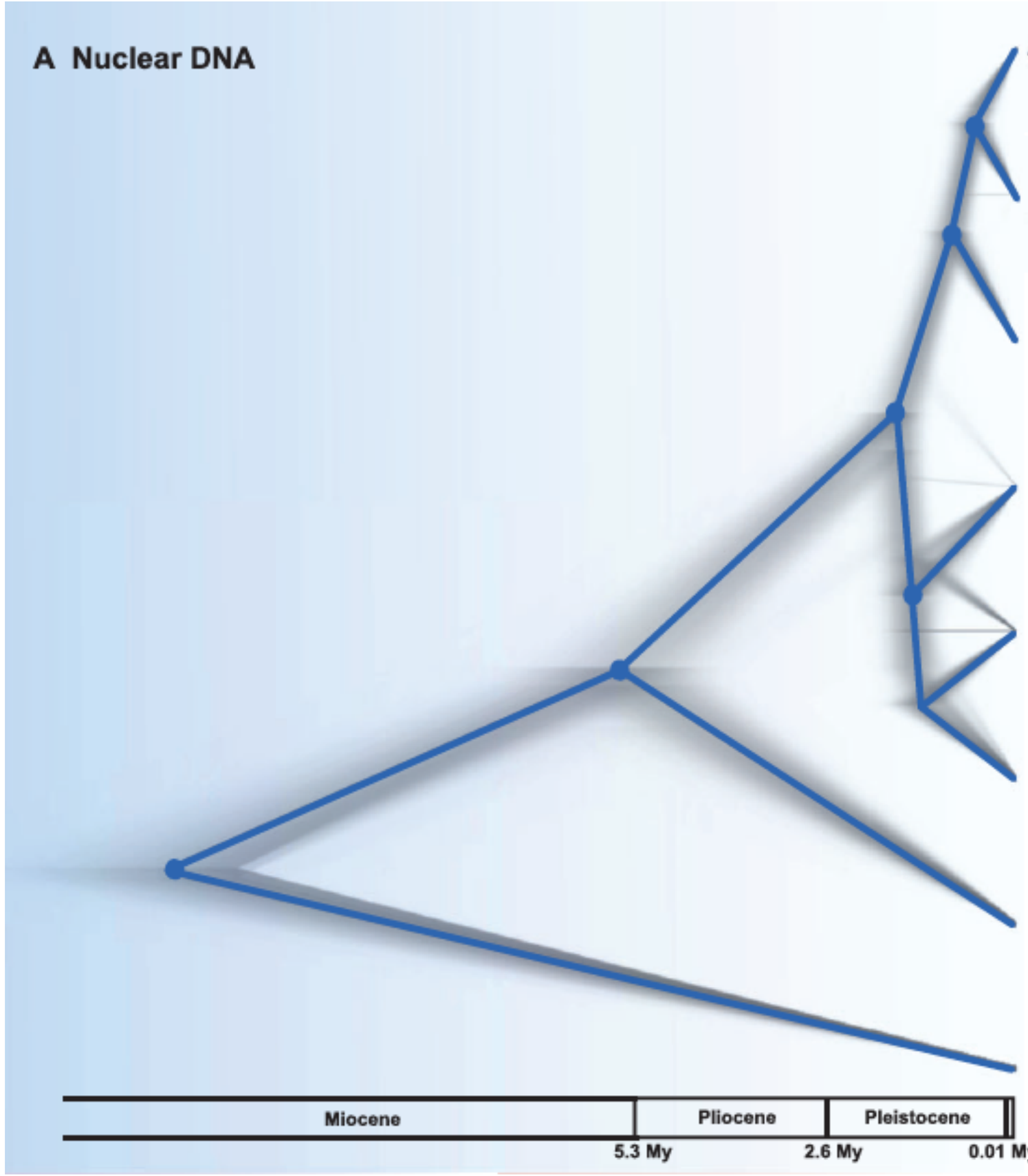
***Corresponding author:** E-mail: v.kutschera@gmx.net, axel.janke@senckenberg.de.

Associate editor: David Irwin

Abstract

Ursine bears are a mammalian subfamily that comprises six morphologically and ecologically distinct extant species. Previous phylogenetic analyses of concatenated nuclear genes could not resolve all relationships among bears, and appeared to conflict with the mitochondrial phylogeny. Evolutionary processes such as incomplete lineage sorting and introgression can cause gene tree discordance and complicate phylogenetic inferences, but are not accounted for in phylogenetic analyses of concatenated data. We generated a high-resolution data set of autosomal introns from several individuals per species and of Y-chromosomal markers. Incorporating intraspecific variability in coalescence-based phylogenetic and gene flow estimation approaches, we traced the genealogical history of individual alleles. Considerable heterogeneity among nuclear loci and discordance between nuclear and mitochondrial phylogenies were found. A species tree with divergence time estimates indicated that ursine bears diversified within less than 2 My. Consistent with a complex branching order within a clade of Asian bear species, we identified unidirectional gene flow from Asian black into sloth bears. Moreover, gene flow detected from brown into American black bears can explain the conflicting placement of the American black bear in mitochondrial and nuclear phylogenies. These results highlight that both

A Nuclear DNA




Polar bear — Uma


Brown bear — Uar


American black bear — Uam


Asian black bear — UTHL


Sloth bear — MURL


Sun bear — HMAL


Spectacled bear — Tor


Giant panda — Ame

Syst. Biol. 63(4):639–644, 2014

DOI:10.1093/sysbio/syu028
Advance Access publication March 28, 2014

A Preliminary Framework for DNA Barcoding, Incorporating the Multispecies Coalescent

MARK DOWTON[1,*], KELLY MEIKLEJOHN[2], STEPHEN L. CAMERON[3], AND JAMES WALLMAN[2]

[1]*Centre for Medical and Molecular Bioscience;* [2]*Institute for Conservation Biology and Environmental Management, School of Biological Sciences, University of Wollongong, NSW 2522; and* [3]*School of Earth, Environmental and Biological Sciences, Queensland University of Technology, QLD 4001, Australia*
**Correspondence to be sent to: Centre for Medical and Molecular Bioscience, School of Biological Sciences, University of Wollongong, NSW 2522, Australia; E-mail: mdowton@uow.edu.au.*

Received 27 February 2014; reviews returned 8 March 2014; accepted 18 March 2014
Associate Editor: Tanja Stadler

Known Knowns, Known Unknowns, Unknown Unknowns and Unknown Knowns in DNA Barcoding: A Comment on Dowton et al.

RUPERT A. COLLINS[1*] AND ROBERT H. CRUICKSHANK[2]

[1]*Laboratório de Evolução e Genética Animal, Departamento de Biologia, Universidade Federal do Amazonas, Av. Rodrigo Otávio, Manaus, Amazonas, Brazil and* [2]*Department of Ecology, Faculty of Agriculture and Life Sciences, Lincoln University, Lincoln 7647, Canterbury, New Zealand*
**Correspondence to be sent to: Departamento de Biologia, Universidade Federal do Amazonas, Av. Rodrigo Otávio, Manaus, Amazonas, Brazil; E-mail: rupertcollins@gmail.com*

Received 5 June 2014; reviews returned 1 August 2014; accepted 4 August 2014
Associate Editor: Tanya Stadler

Can delimitation methods identify rare species?

Collins and Cruickshank (2015):

“The species delimitation literature shows a surprising lack of awareness for the commonness of rarity.”

“it is questionable whether such statistics would be reliable where they would be most useful—that is, for singletons such as S . australis KM673—*due to the sampling and parameter estimation problems* associated with taxon rarity in species delimitation methods (Lim et al. 2012).”

Determining Species Boundaries in a World Full of Rarity: Singletons, Species Delimitation Methods

GWYNNE S. LIM[1], MICHAEL BALKE[2], AND RUDOLF MEIER[1,3,*]

[1]*Department of Biological Sciences, National University of Singapore, Science Drive 4, Singapore 117543, Singapore;*
[2]*Zoologische Staatsammlung, Muenchhausenstrasse 21, 81247 Munich, Germany; and* [3]*University Scholars Programme, National University of Singapore, Science Drive 4, Singapore 117543, Singapore;*
**Correspondence to be sent to: Department of Biological Sciences and University Scholars Programme, National University of Singapore, Science Drive 4, Singapore 117543, Singapore; E-mail:meier@nus.edu.sg.*

Received 14 May 2010; reviews returned 29 July 2010; accepted 7 January 2011
Associate Editor: Mark Fishbein

"Not all methods and algorithms are explicit about how densely species have to be sampled in order for these methods to be successful, but some of the more commonly used software for coalescence analysis including "BEST," "COAL," or "Brownie" assume sampling frequencies of 5 individuals per species. Otherwise, an inadequate representation of intraspecific variability ***will lead to incorrect inferences***. However, our survey of the biodiversity and taxonomic literature reveals that such sampling is unattainable for ca. 30% of all species, that is, the failure to account for rarity in coalescence analyses is ***likely to yield incorrect results for a large proportion of the species diversity***."

New Results

Species Identification by Bayesian Fingerprinting: A Powerful Alternative to DNA Barcoding

Ziheng Yang, Bruce Rannala

doi: http://dx.doi.org/10.1101/041608

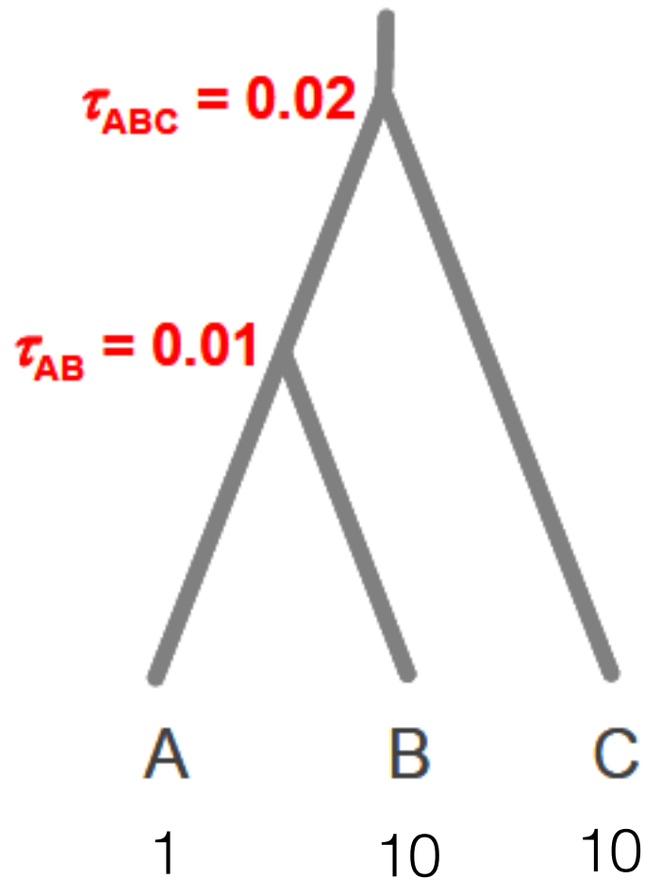
Species Delimitation versus DNA Barcoding

DNA barcoding typically uses 1 mtDNA locus and a distance threshold for assigning individuals to species

Questions:

(1) Can delimitation methods identify rare species?
(2) Does a universal barcoding “gap” exist?
(3) Can delimitation methods identify cryptic species?

(1) Identifying rare species

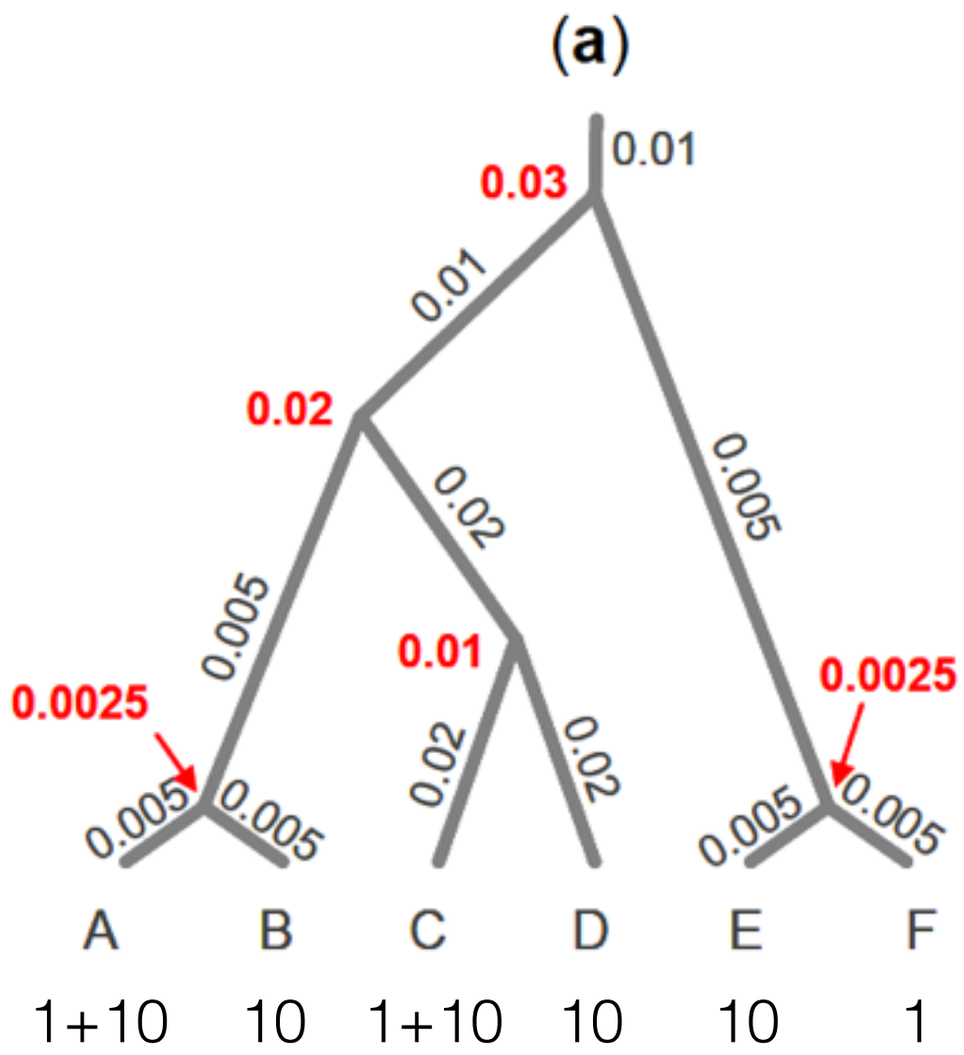


Avg PP of 3 species (2 loci): 0.998

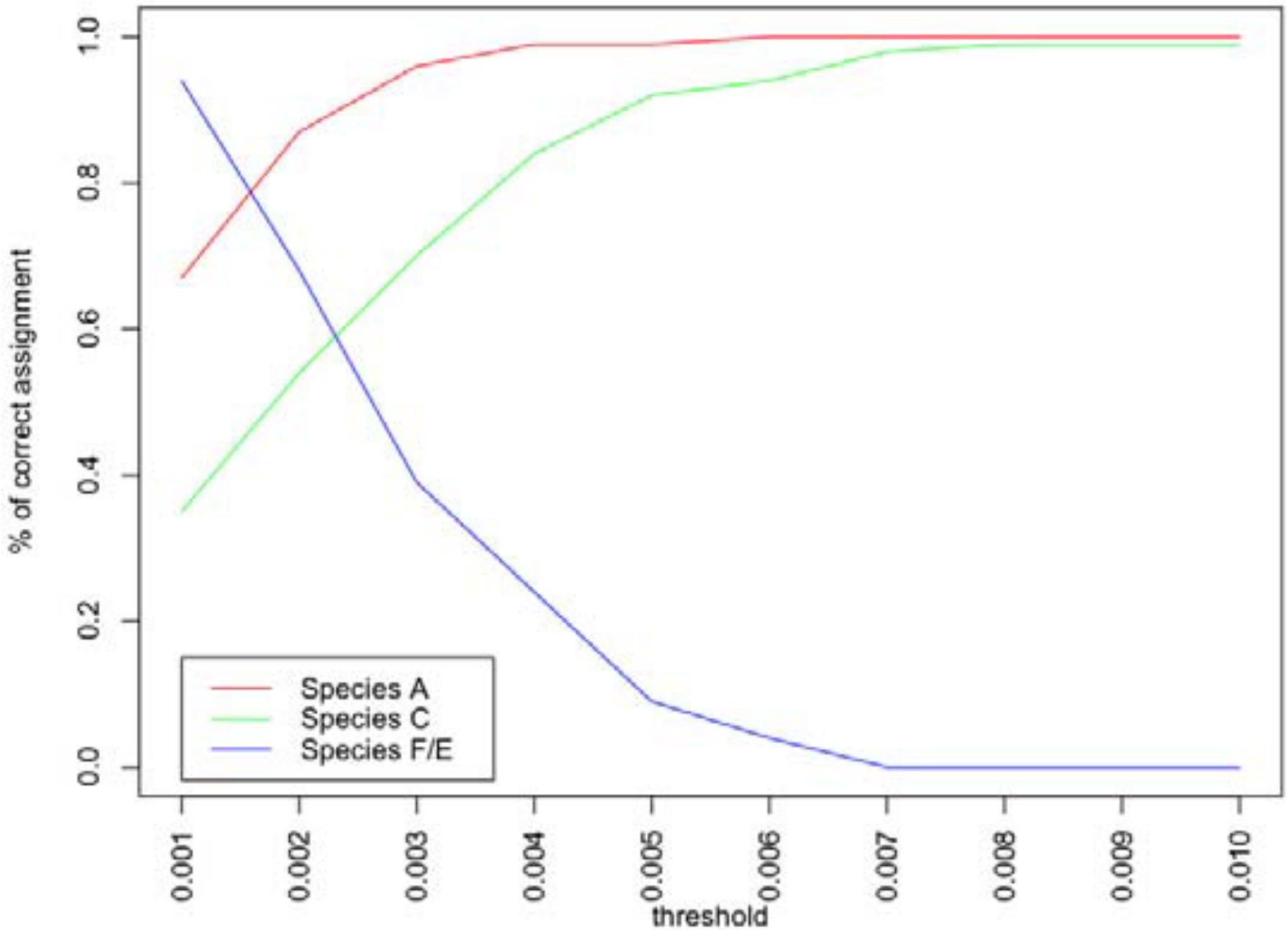
Avg PP of 3 species (10 loci): 1.000

(2) No barcode gap for identifying all species exists

Simulation 1: 1 Locus 1000 bps

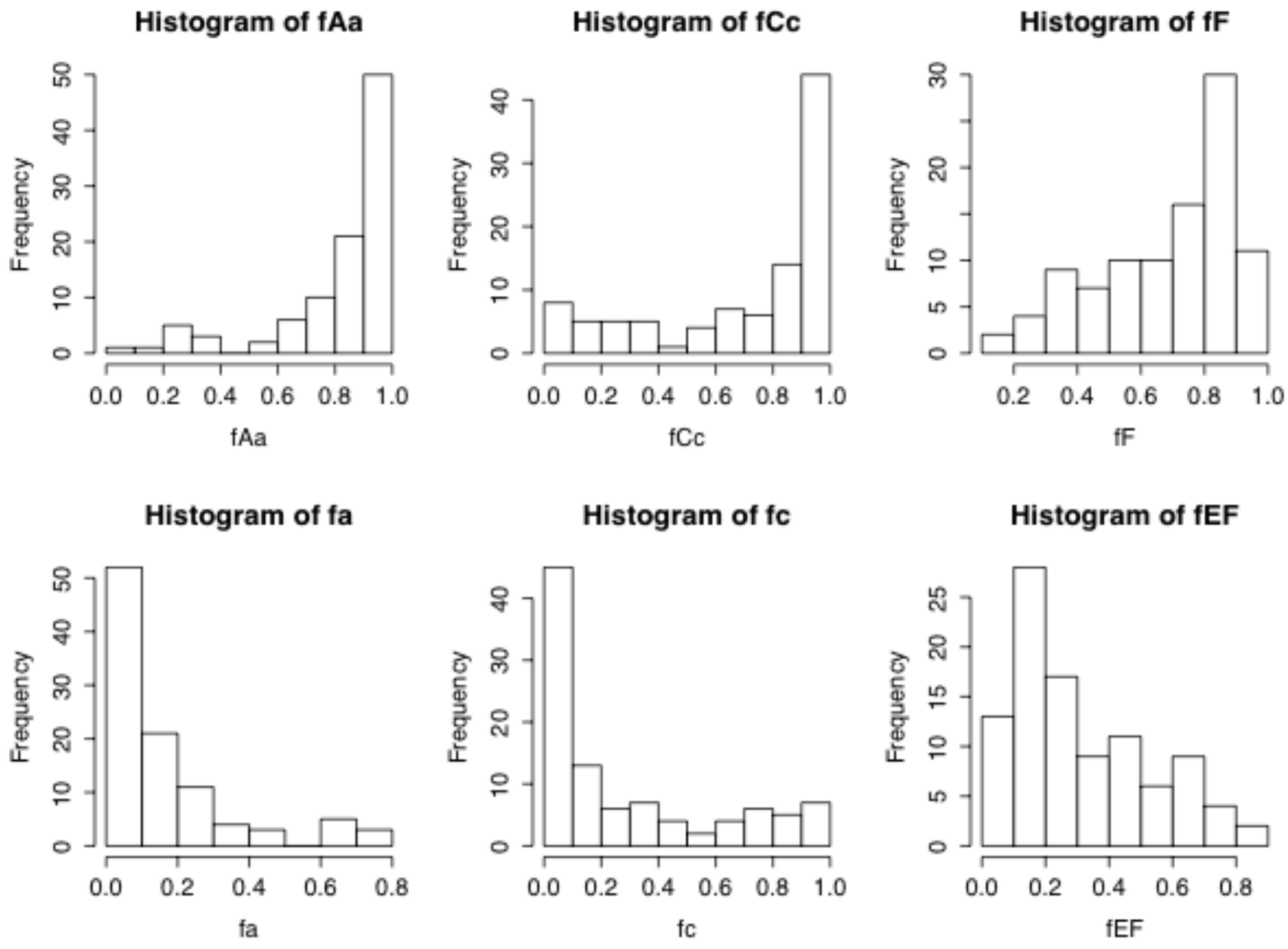


No barcode gap for identifying all species exists



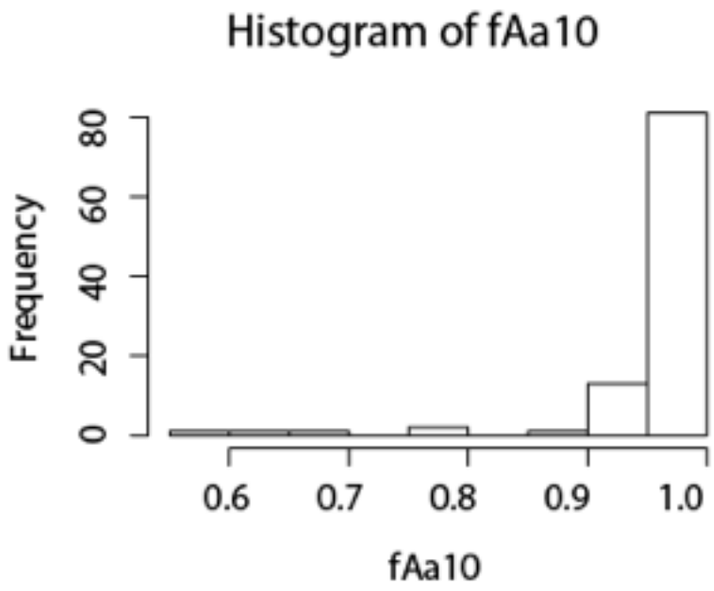
BPP Posterior Probabilities of Delimitations

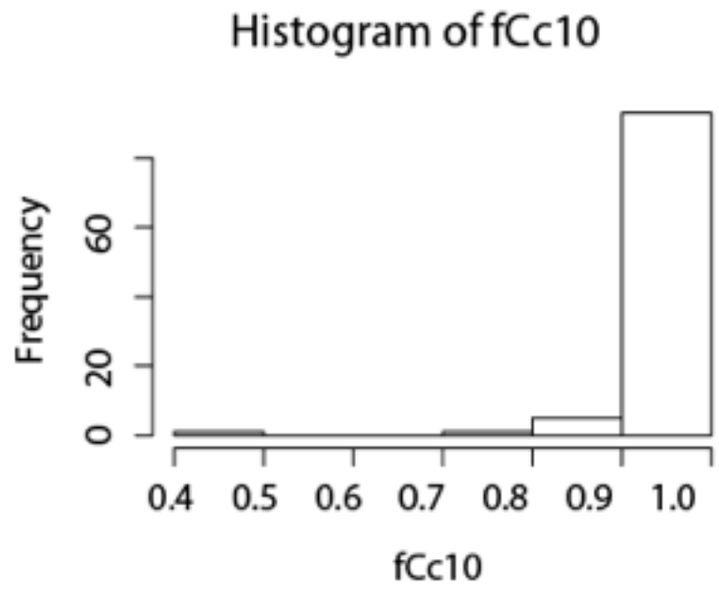
1 Locus

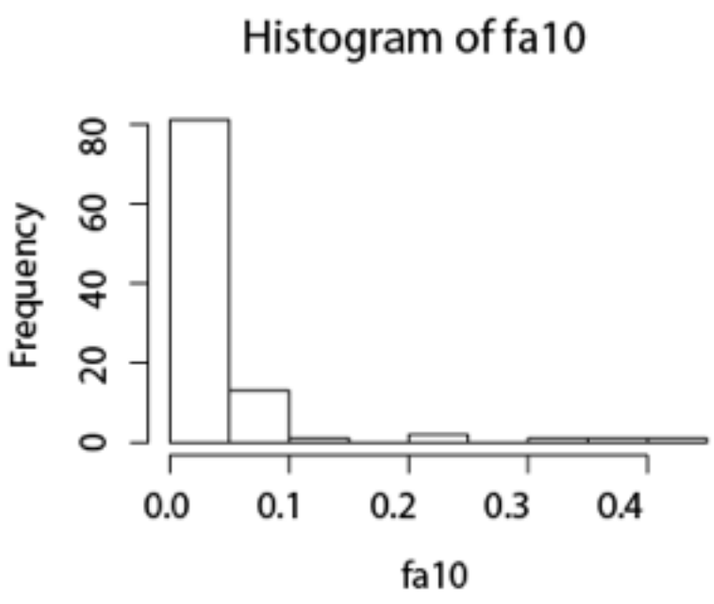


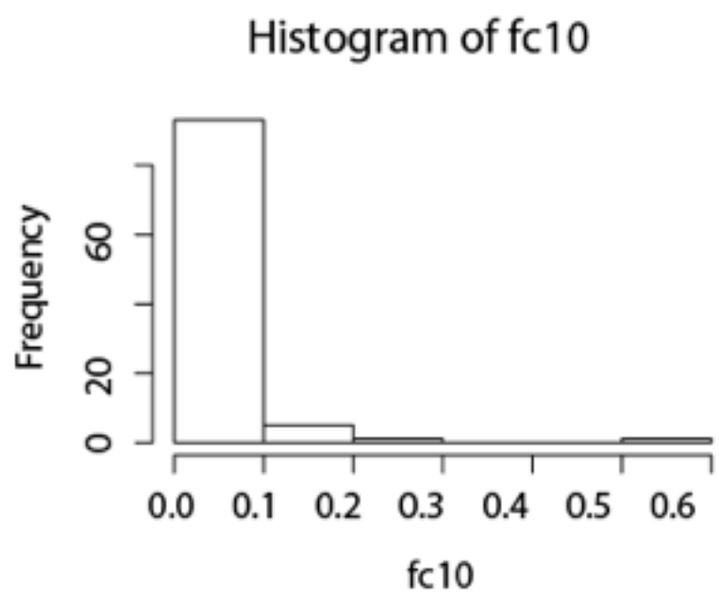
BPP Posterior Probabilities of Delimitations

10 Loci

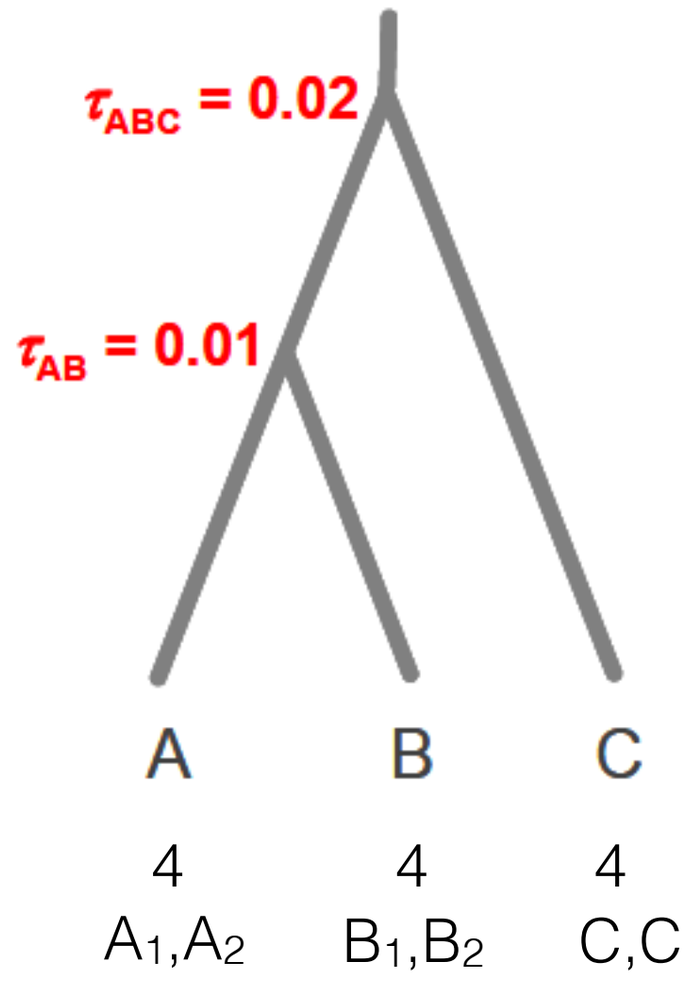




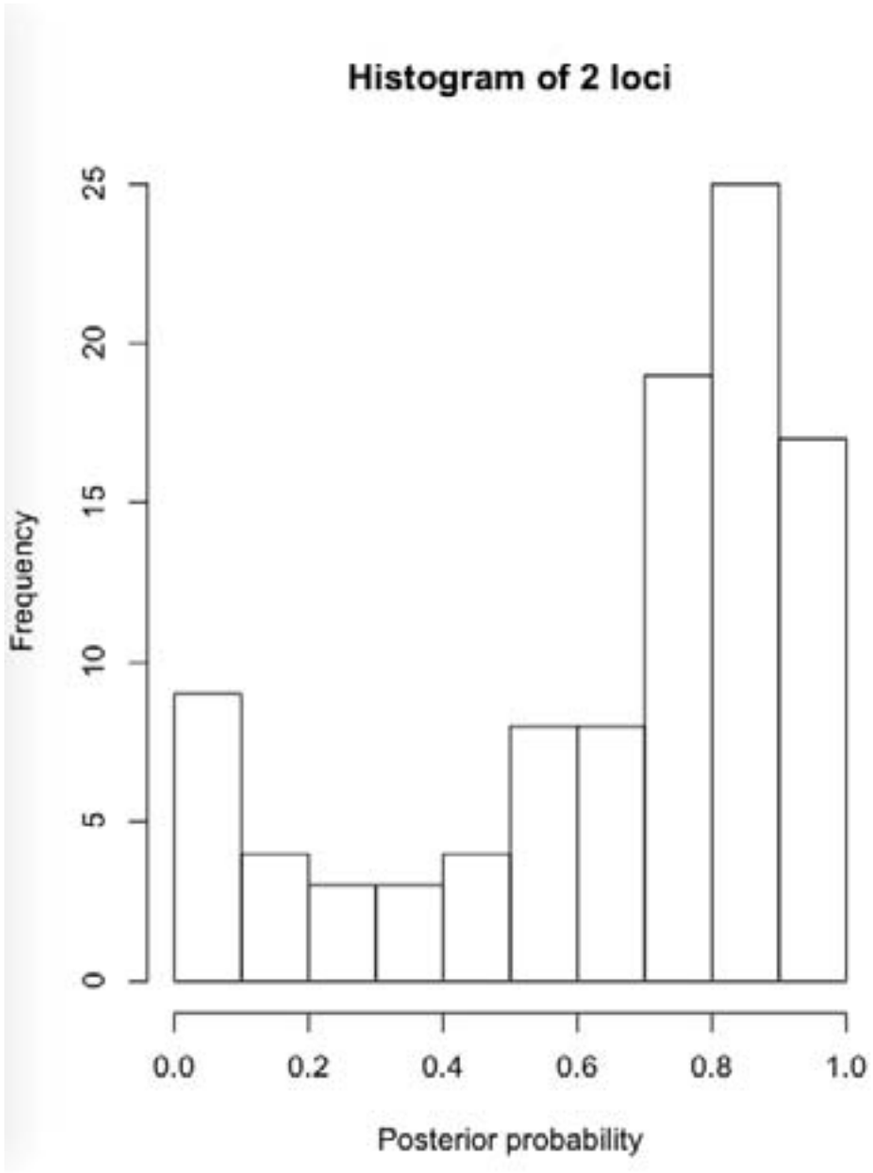


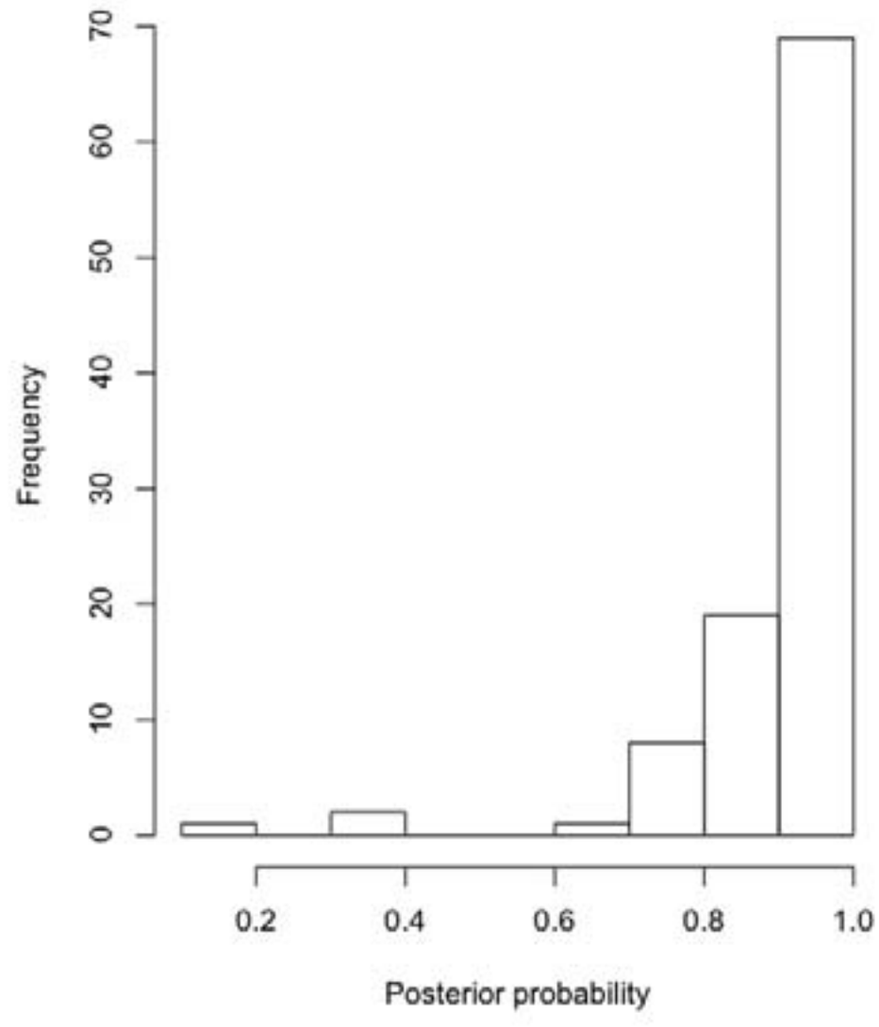


(3) Identifying cryptic species



Identifying cryptic species





The Future of BPP

- Introgression (possibly locus specific)
- Efficiency (improved proposals for gene trees, faster likelihood calculations)
- Parallel programming (calculate gene tree likelihood on different compute nodes?)
- Recombination?