Phylogenetics: Species trees and gene tree variation

George P. Tiley
University of Antananarivo
DBEV Phylogenomics Workshop
8 March 2022

Learning Goals

Gene tree variation

Coalescent Theory

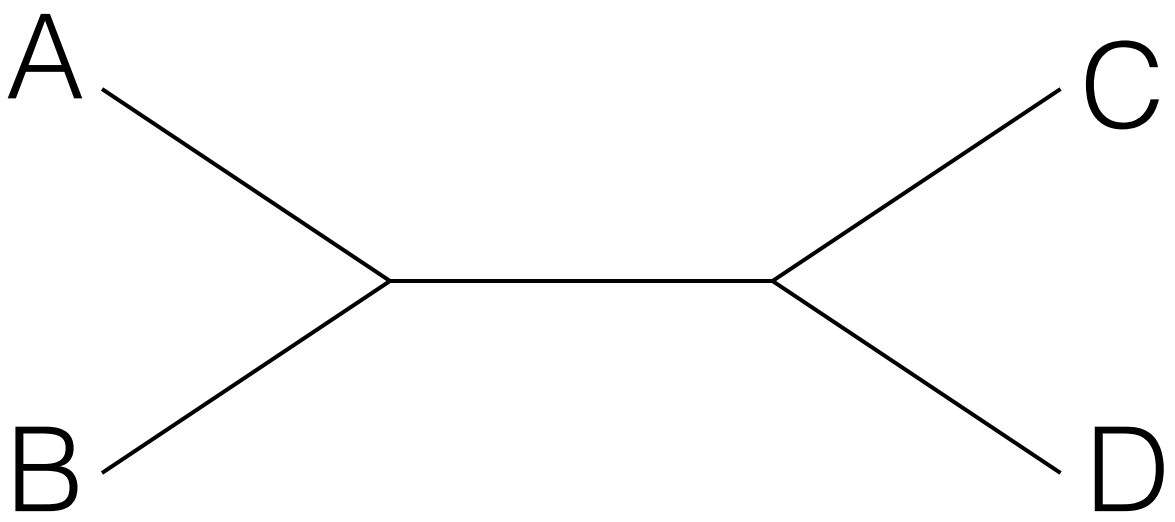
The multispecies coalescent

Species tree estimation

Application to taxonomy

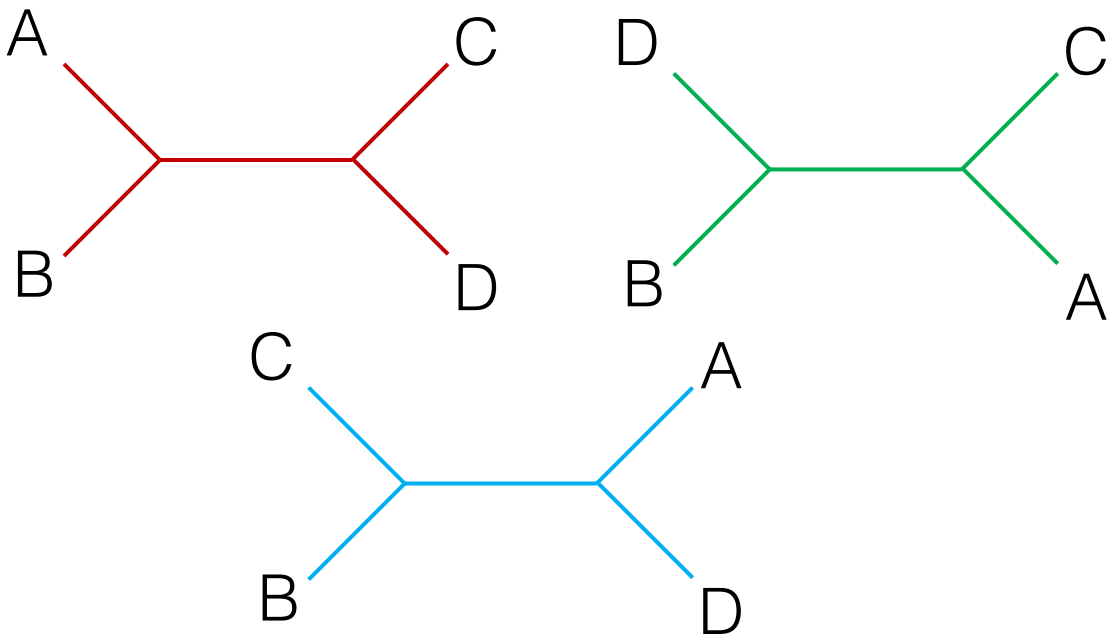
Gene tree variation

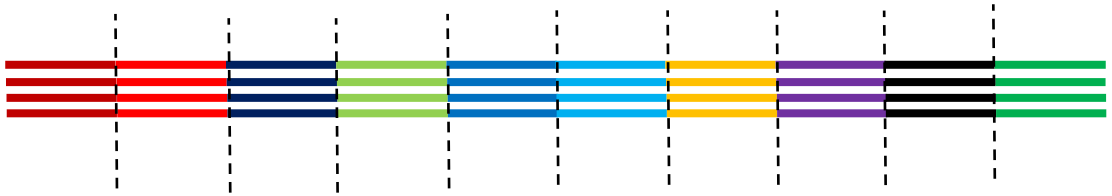
Maximum likelihood can be used to estimate a phylogeny.



Gene tree variation

Maximum likelihood can be used to estimate a phylogeny.
Ideally collecting a lot of data allows us to do this accurately.





But gene trees can vary among loci

Gene tree variation

Syst. Biol. 46(3):523–536, 1997

GENE TREES IN SPECIES TREES

WAYNE P. MADDISON

Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, Arizona 85721, USA

Multiple biological sources of gene tree variation

- Incomplete lineage sorting
- gene duplication and loss
- horizontal transfer (gene flow)

Gene tree variation

Review

Cell PRESS

Computational approaches to species phylogeny inference and gene tree reconciliation

Luay Nakhleh[1,2]

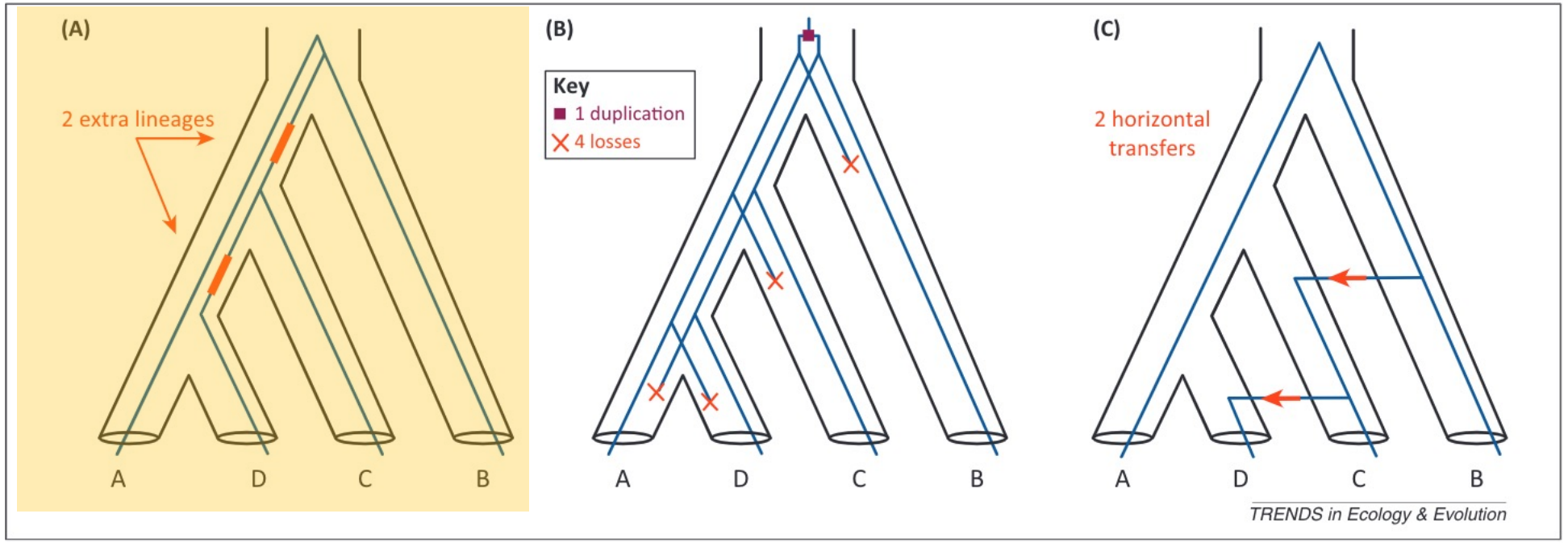


Figure 4. Reconciliation of a gene tree with a species tree. **(A)** Reconciliation assuming incomplete lineage sorting (ILS) results in two extra lineages, highlighted with thick red lines. **(B)** Reconciliation assuming gene duplication and loss (DL) results in a single duplication event and four losses. **(C)** Reconciliation assuming horizontal gene transfer (HGT) (or hybridization) results in two horizontal transfer events, highlighted with red arrows.

Gene tree variation

Concatenation can be misleading when incomplete lineage sorting (ILS) is high

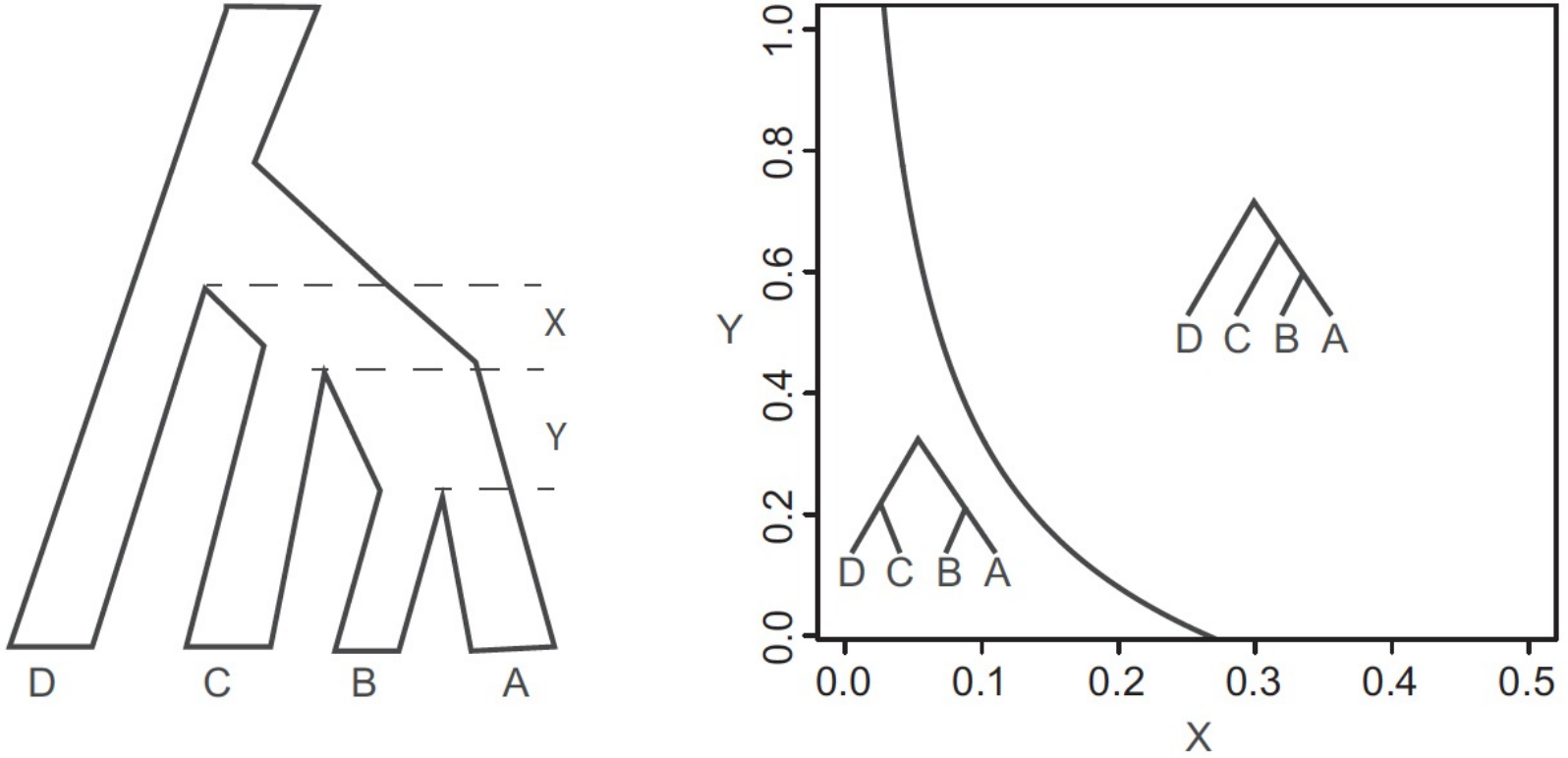
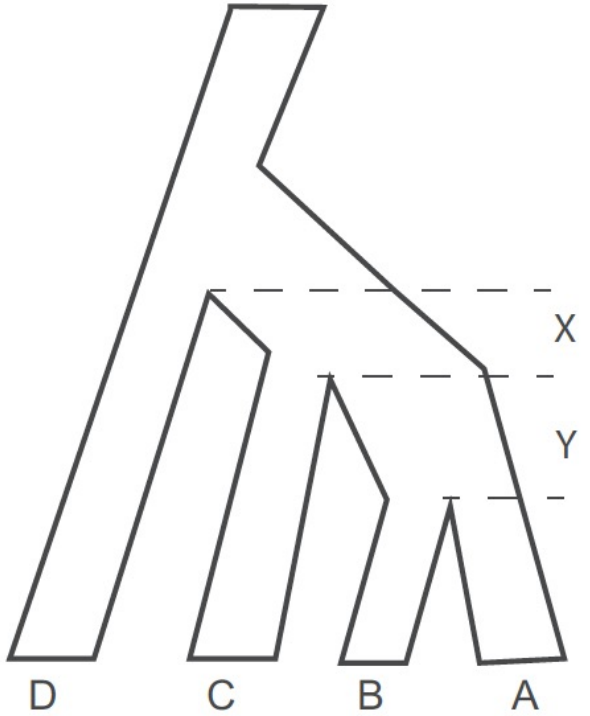


FIGURE 1. The lengths of branches X and Y in coalescent units in the species tree determine the probability of the gene tree topology. For branches under the anomaly zone curve, the symmetric AGT will have a higher probability than the asymmetric gene tree that matches the species tree.

Linkem et al. 2016

Gene tree variation

Concatenation can be misleading when incomplete lineage sorting (ILS) is high



As the internal branch lengths become very small, the most common gene tree will no longer be the species tree

Kubatko and Degnan 2007

Linkem et al. 2016

Learning Goals

Gene tree variation

Coalescent Theory

The multispecies coalescent

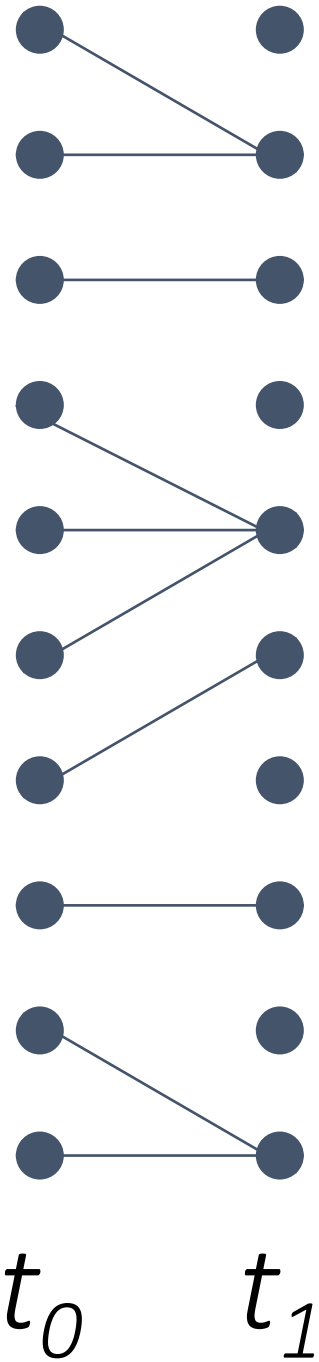
Species tree estimation

Application to taxonomy

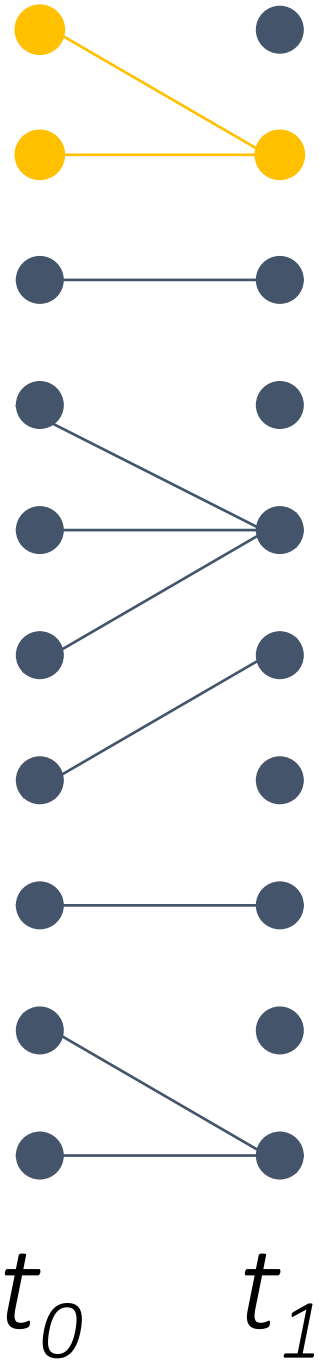
Coalescent Theory

The Coalescent for a Single Population

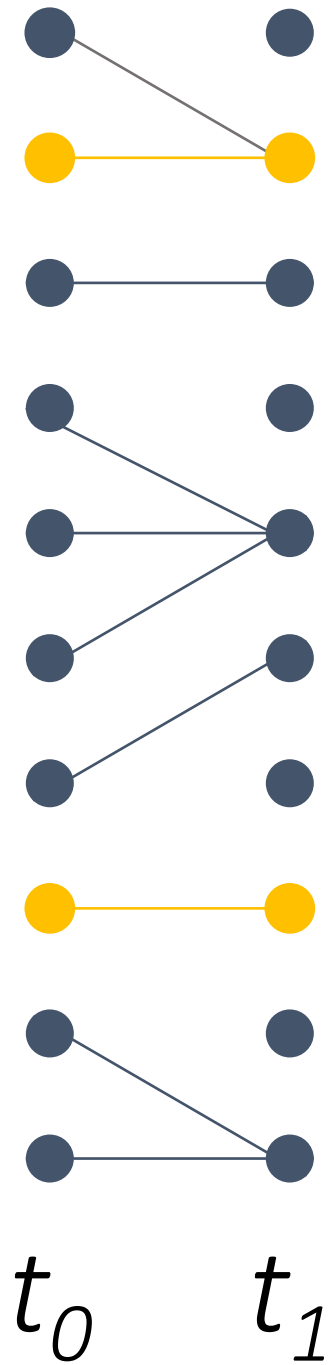
A sample of 10 individuals
in the present



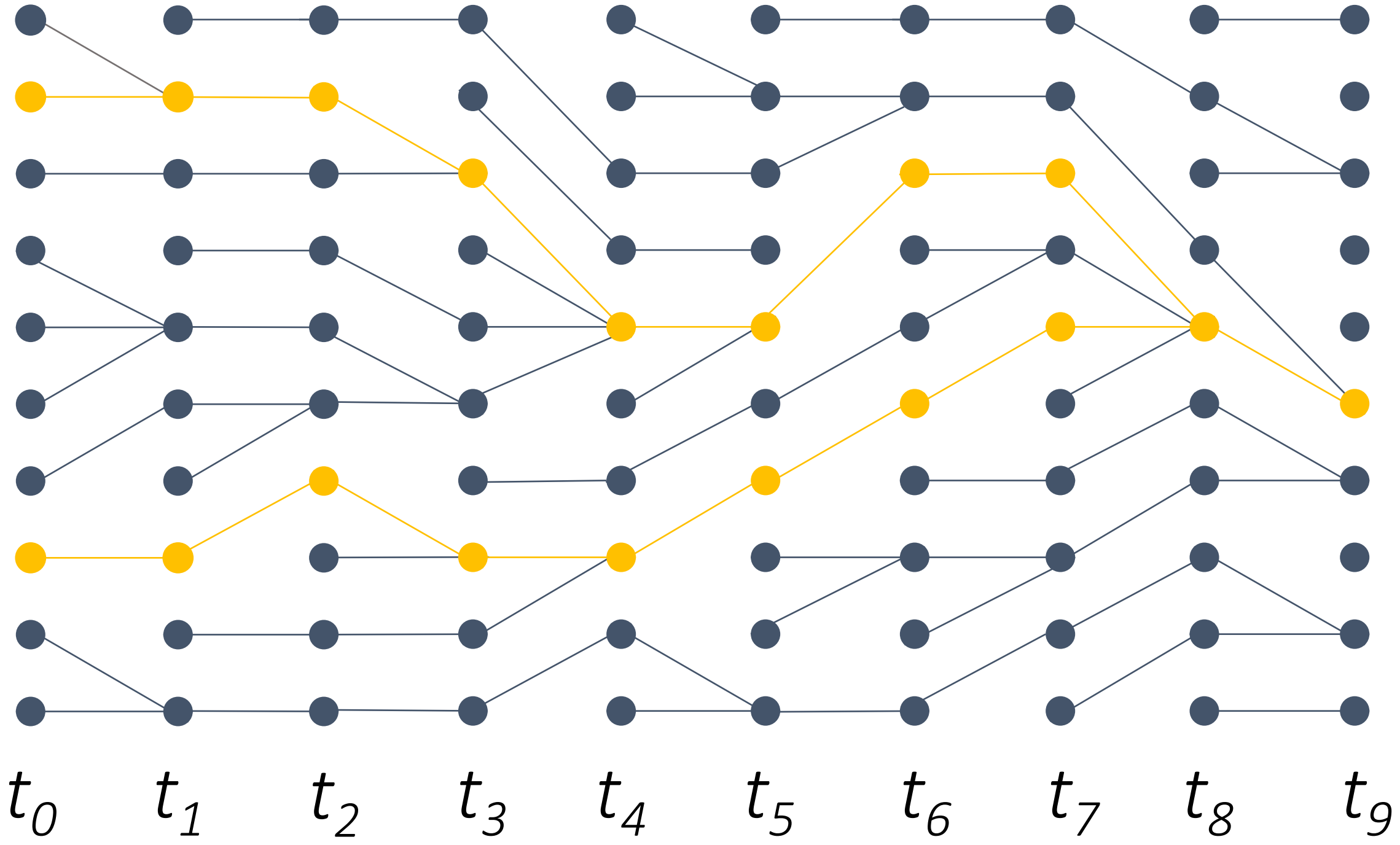
Sampled from the previous generation under Wright-Fisher

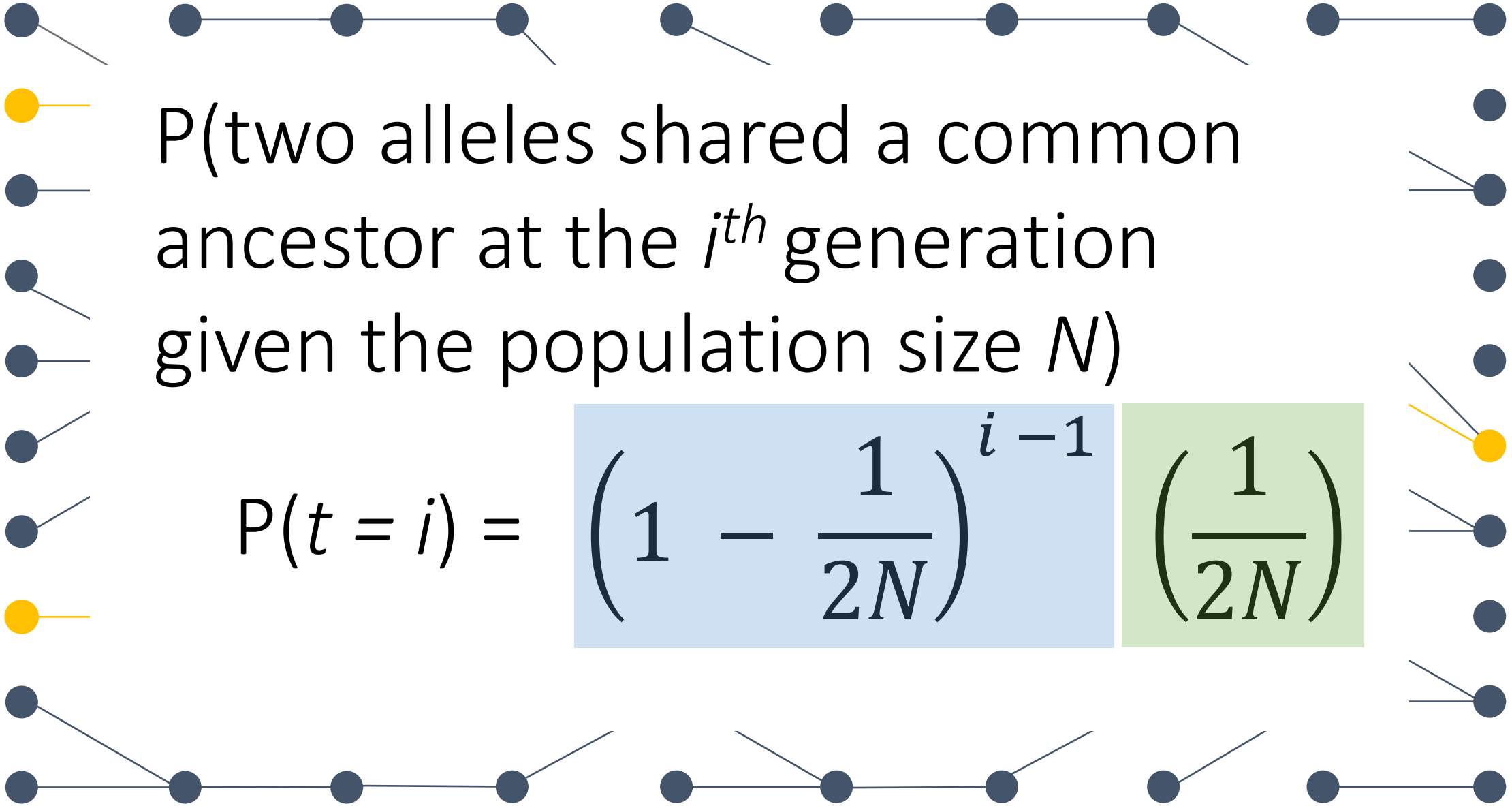


P(Two Alleles Share Ancestor 1 generation ago)
= 1/2N



P(Two Alleles **Do Not** Share Ancestor 1 generation ago)
= 1 – (1/2N)


t_0
t_1
t_2
t_3
t_4
t_5
t_6
t_7
t_8
t_9



P(two alleles shared a common ancestor at the i^{th} generation given the population size N)

$$P(t = i) = \left(1 - \frac{1}{2N}\right)^{i-1} \left(\frac{1}{2N}\right)$$

t_0 t_1 t_2 t_3 t_4 t_5 t_6 t_7 t_8 t_9

Geometric distribution with

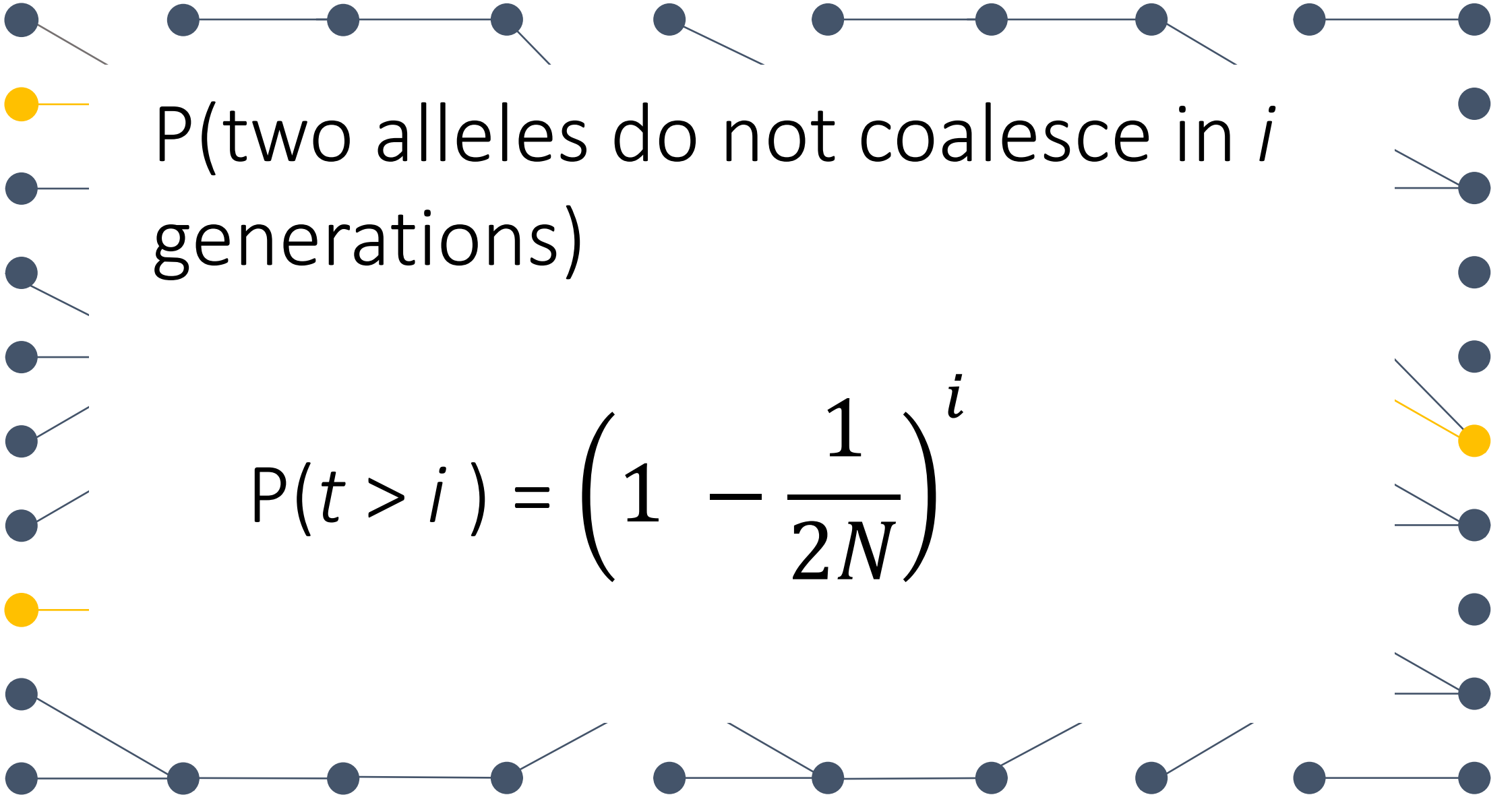
$p = 1/2N$

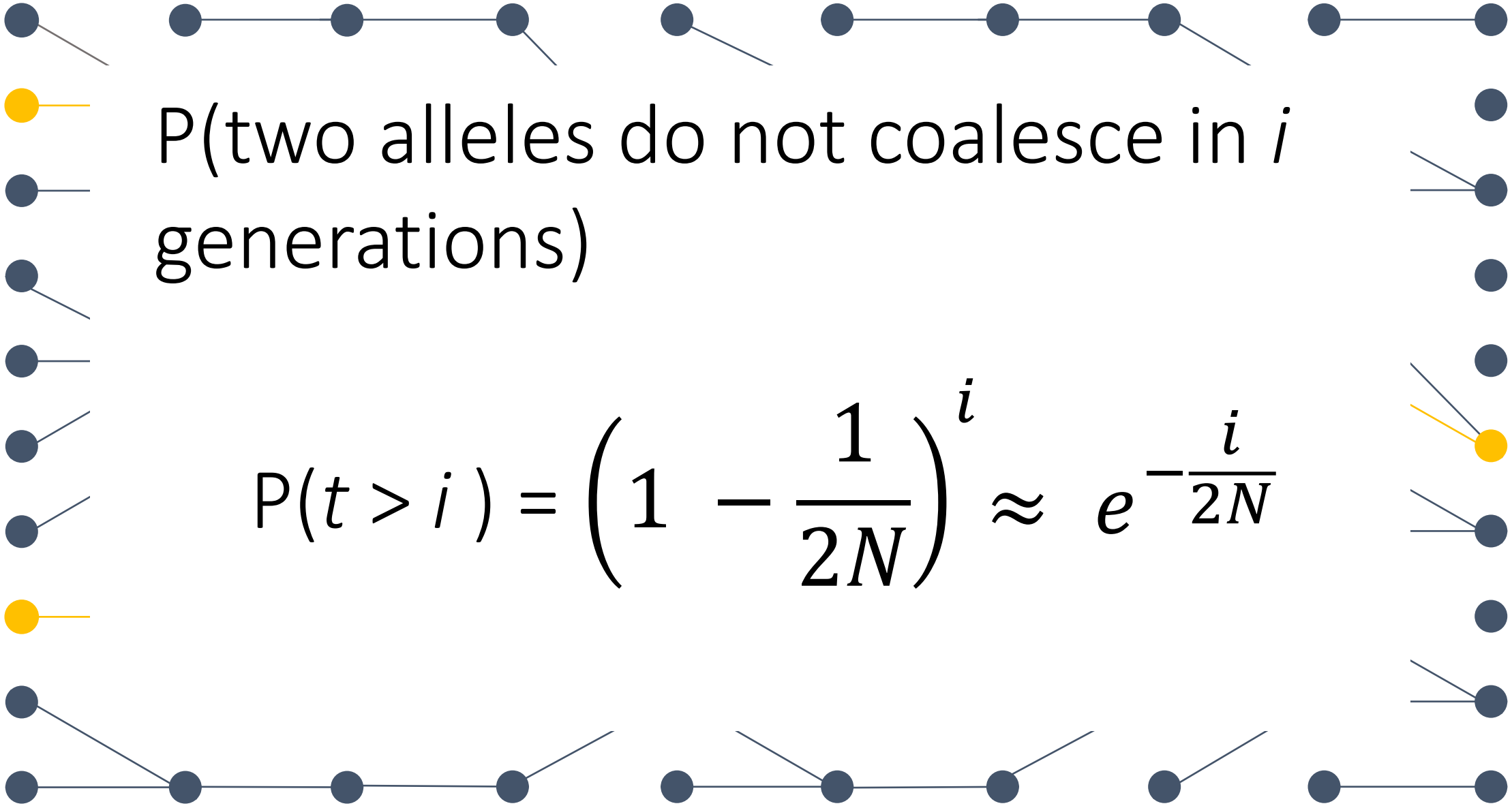
$$P(t|N) = \left(1 - \frac{1}{2N}\right)^{i-1} \left(\frac{1}{2N}\right)$$

$E[t] = 1/p = 2N$

P(two alleles do not coalesce in *i* generations)

$$P(t > i) = \left(1 - \frac{1}{2N}\right)^{i}$$

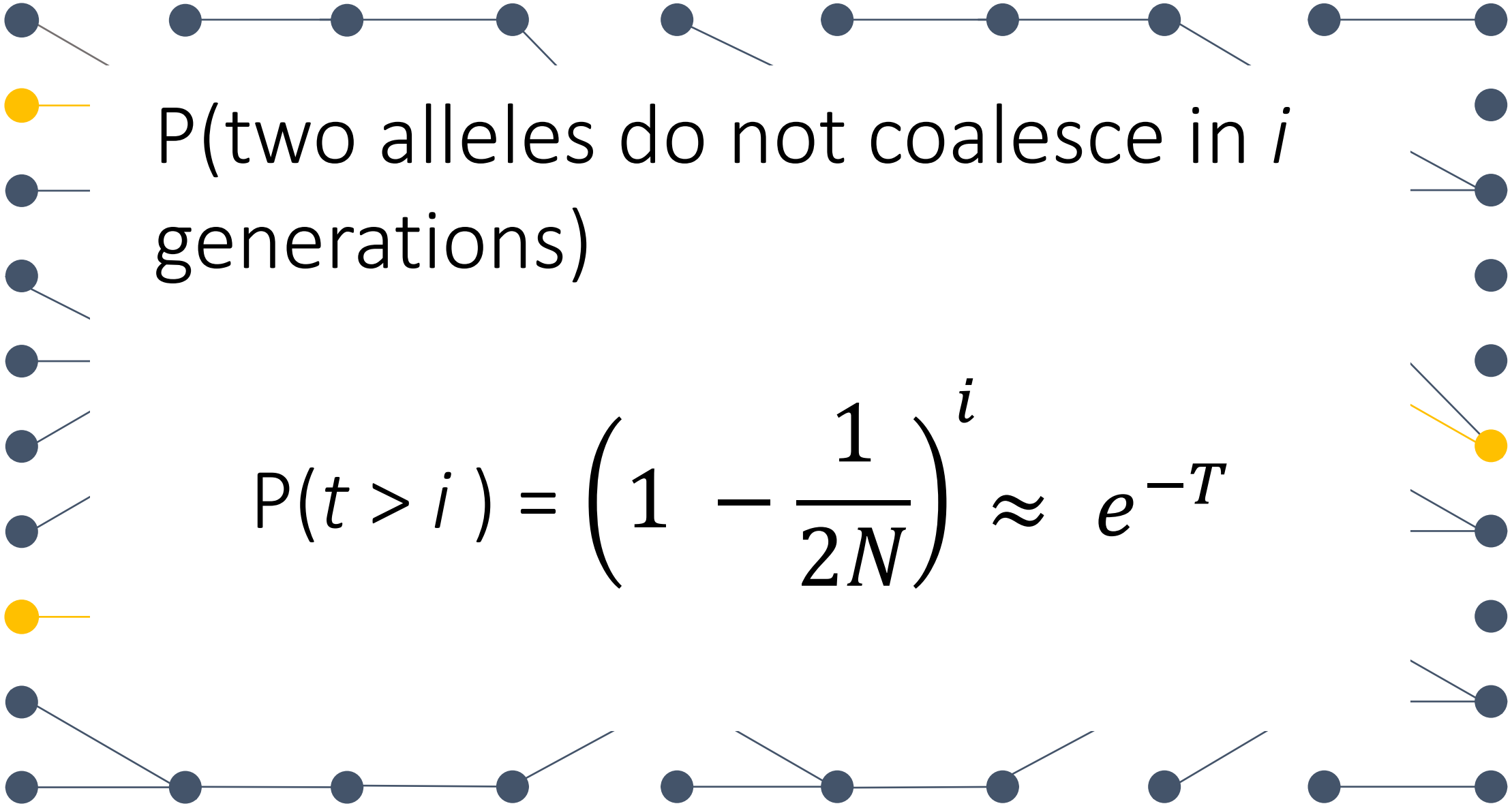




P(two alleles do not coalesce in *i* generations)

$$P(t > i) = \left(1 - \frac{1}{2N}\right)^{i} \approx e^{-\frac{i}{2N}}$$

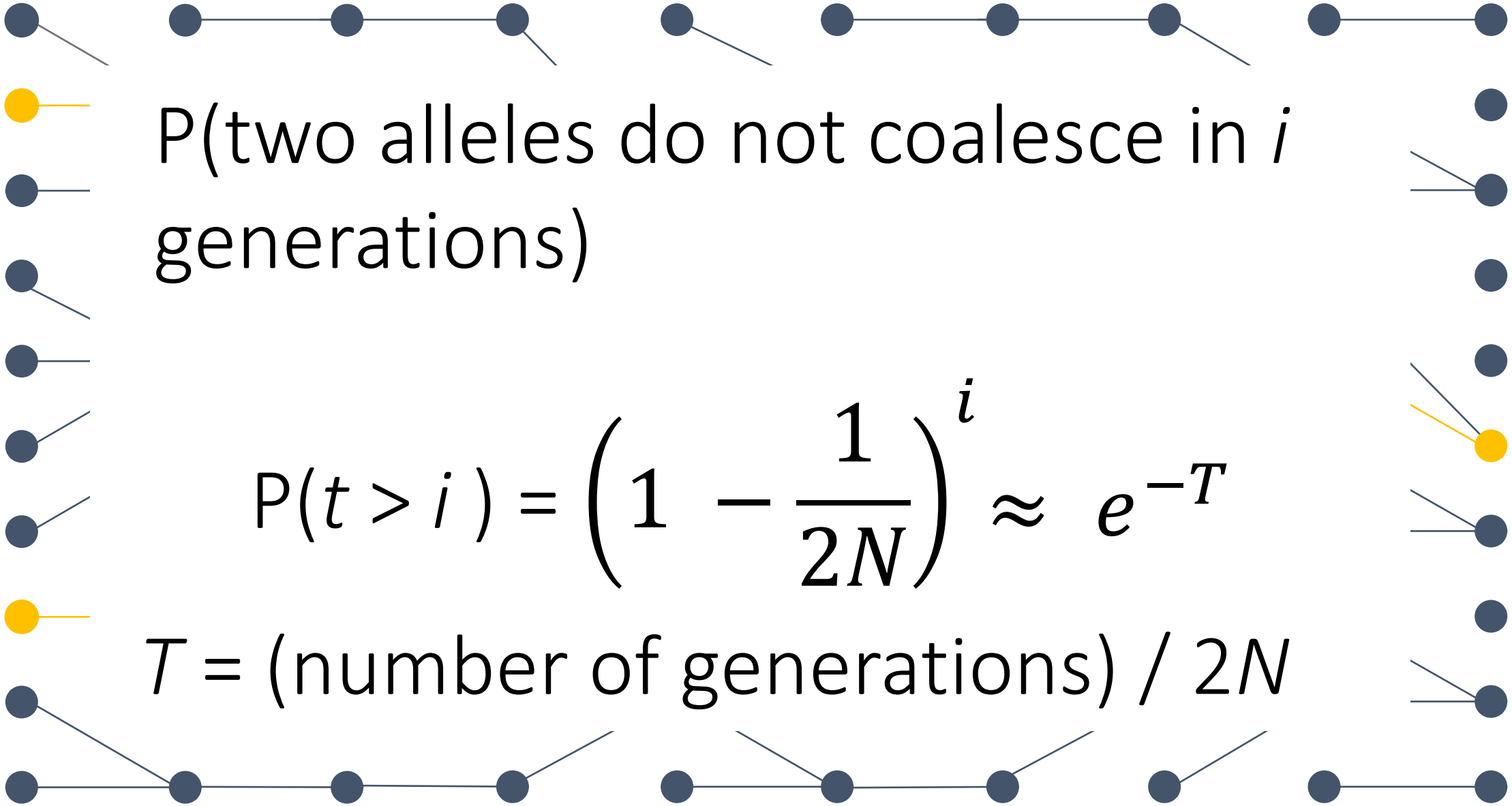
t_0 t_1 t_2 t_3 t_4 t_5 t_6 t_7 t_8 t_9



P(two alleles do not coalesce in *i* generations)

$$P(t > i) = \left(1 - \frac{1}{2N}\right)^{i} \approx e^{-T}$$

t_0 t_1 t_2 t_3 t_4 t_5 t_6 t_7 t_8 t_9



P(two alleles do not coalesce in *i* generations)

$$P(t > i) = \left(1 - \frac{1}{2N}\right)^{i} \approx e^{-T}$$

T = (number of generations) / 2*N*

t_0 t_1 t_2 t_3 t_4 t_5 t_6 t_7 t_8 t_9

Coalescent Theory

How does this relate to species trees?

Learning Goals

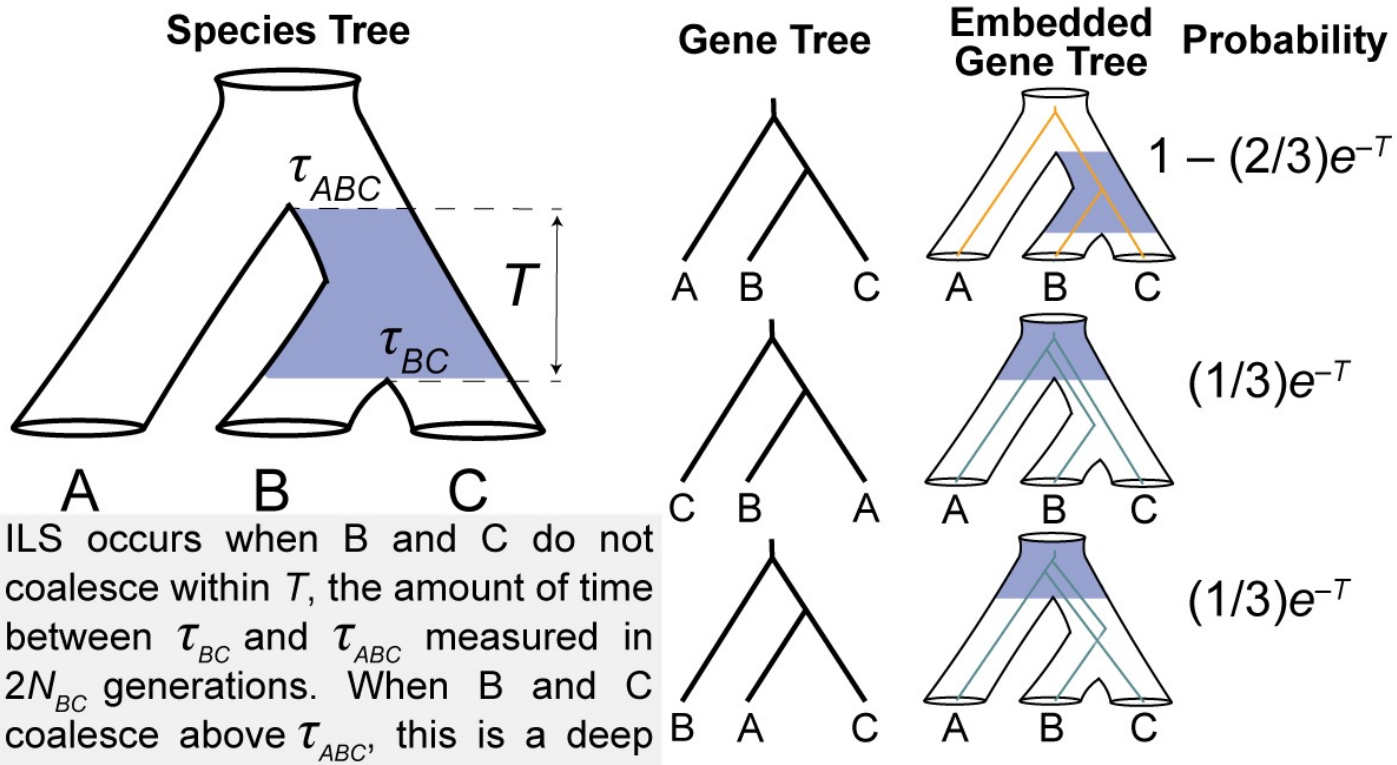
Gene tree variation

Coalescent Theory

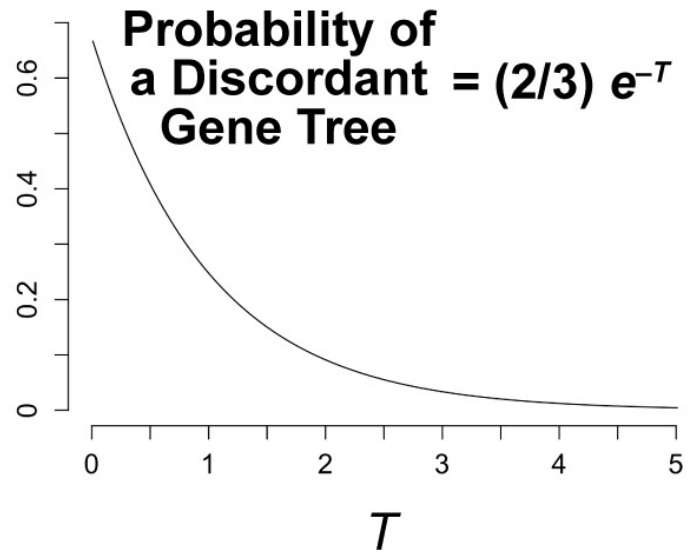
The multispecies coalescent

Species tree estimation

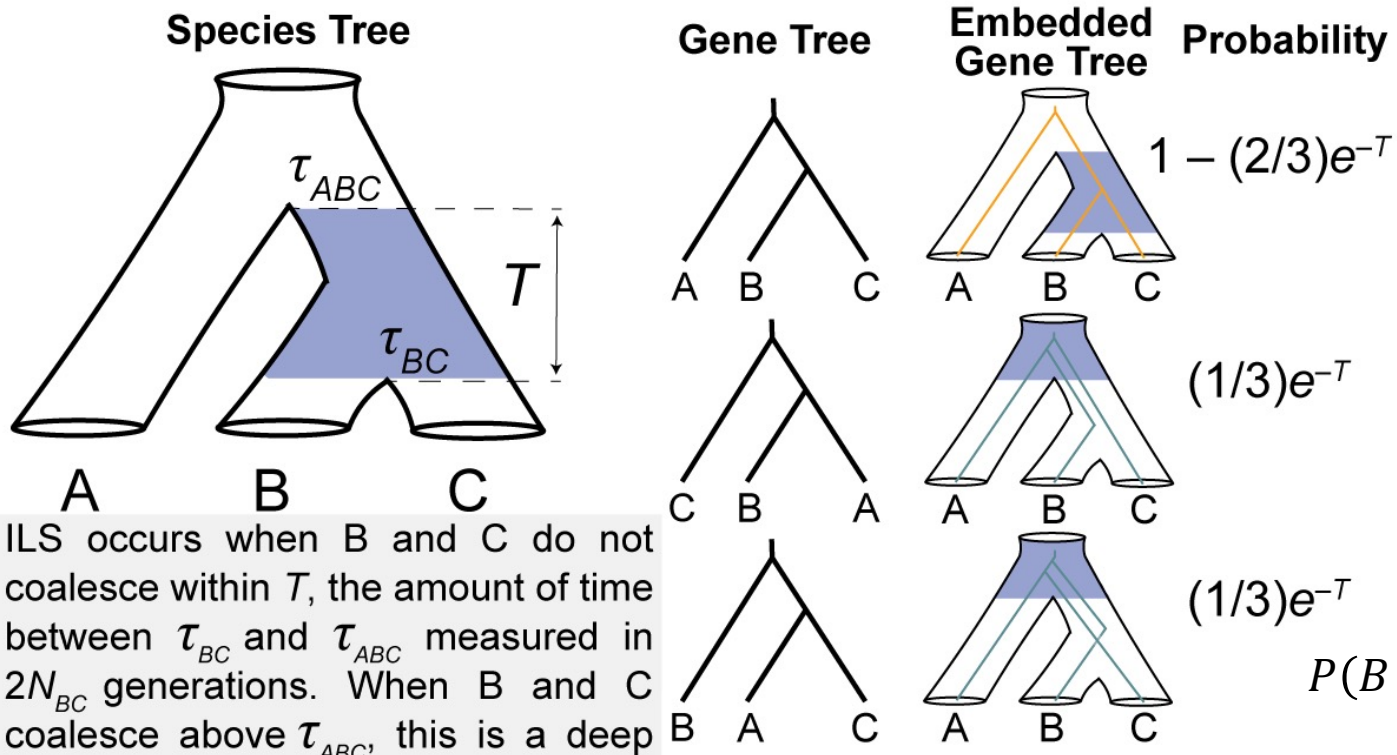
Application to taxonomy



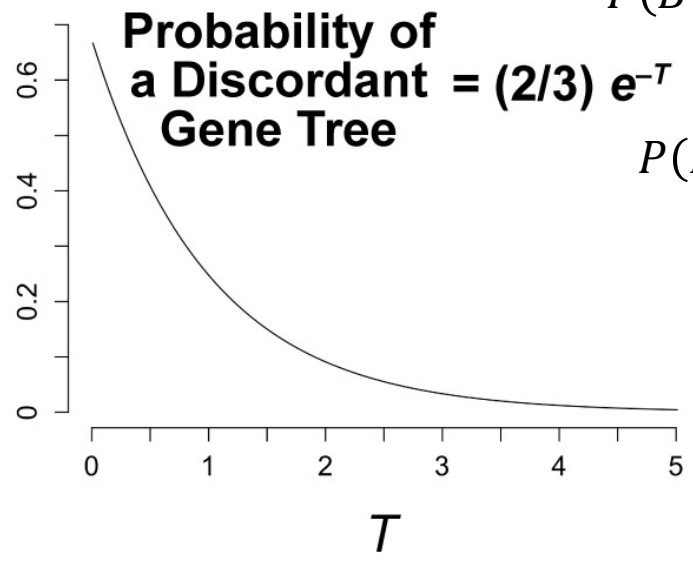
ILS occurs when B and C do not coalesce within T, the amount of time between τ_{BC} and τ_{ABC} measured in $2N_{BC}$ generations. When B and C coalesce above τ_{ABC}, this is a deep coalescence and lineage sorting is incomplete. ILS can be identified visually by embedding gene trees within the species tree. The neutral coalescent provides expectations for the frequency with which ILS occurs that are dependent on T alone [97]. When T is 0, 2/3 of gene trees are expected not to match the species tree due to ILS. Less than 1% of gene trees are expected to be discordant around T of 5. The absolute divergence time does not affect T.



Tiley et al. 2020



ILS occurs when B and C do not coalesce within *T*, the amount of time between τ_{BC} and τ_{ABC} measured in $2N_{BC}$ generations. When B and C coalesce above τ_{ABC}, this is a deep coalescence and lineage sorting is incomplete. ILS can be identified visually by embedding gene trees within the species tree. The neutral coalescent provides expectations for the frequency with which ILS occurs that are dependent on *T* alone [97]. When *T* is 0, 2/3 of gene trees are expected not to match the species tree due to ILS. Less than 1% of gene trees are expected to be discordant around *T* of 5. The absolute divergence time does not affect *T*.



$$P(2\ alleles\ do\ not\ coalesce) = e^{-T}$$

$$P(B\ and\ C\ do\ not\ coalesce) = e^{-T}$$

$$P(B\ and\ C\ do\ coalesce) = 1 - e^{-T}$$

$$P(B\ and\ C\ do\ not\ coalesce\ and\ A\ and\ B\ do\ first) = \frac{1}{3}e^{-T}$$

$$P(B\ and\ C\ do\ not\ coalesce\ and\ A\ and\ C\ do\ first) = \frac{1}{3}e^{-T}$$

$$P(B\ and\ C\ do\ coalesce) = 1 - e^{-T} + \frac{1}{3}e^{-T} = 1 - \frac{2}{3}e^{-T}$$

Tiley et al. 2020

Bayes Estimation of Species Divergence Times and Ancestral Population Sizes Using DNA Sequences From Multiple Loci

Bruce Rannala* and Ziheng Yang[†,1]

**Department of Medical Genetics, University of Alberta, Edmonton, Alberta T6G 2H7, Canada and [†]Galton Laboratory, Department of Biology, University College London, London WC1E 6BT, England*

Manuscript received December 4, 2002
Accepted for publication April 18, 2003

Bayes Estimation of Species Divergence Times and Ancestral Population Sizes Using DNA Sequences From Multiple Loci

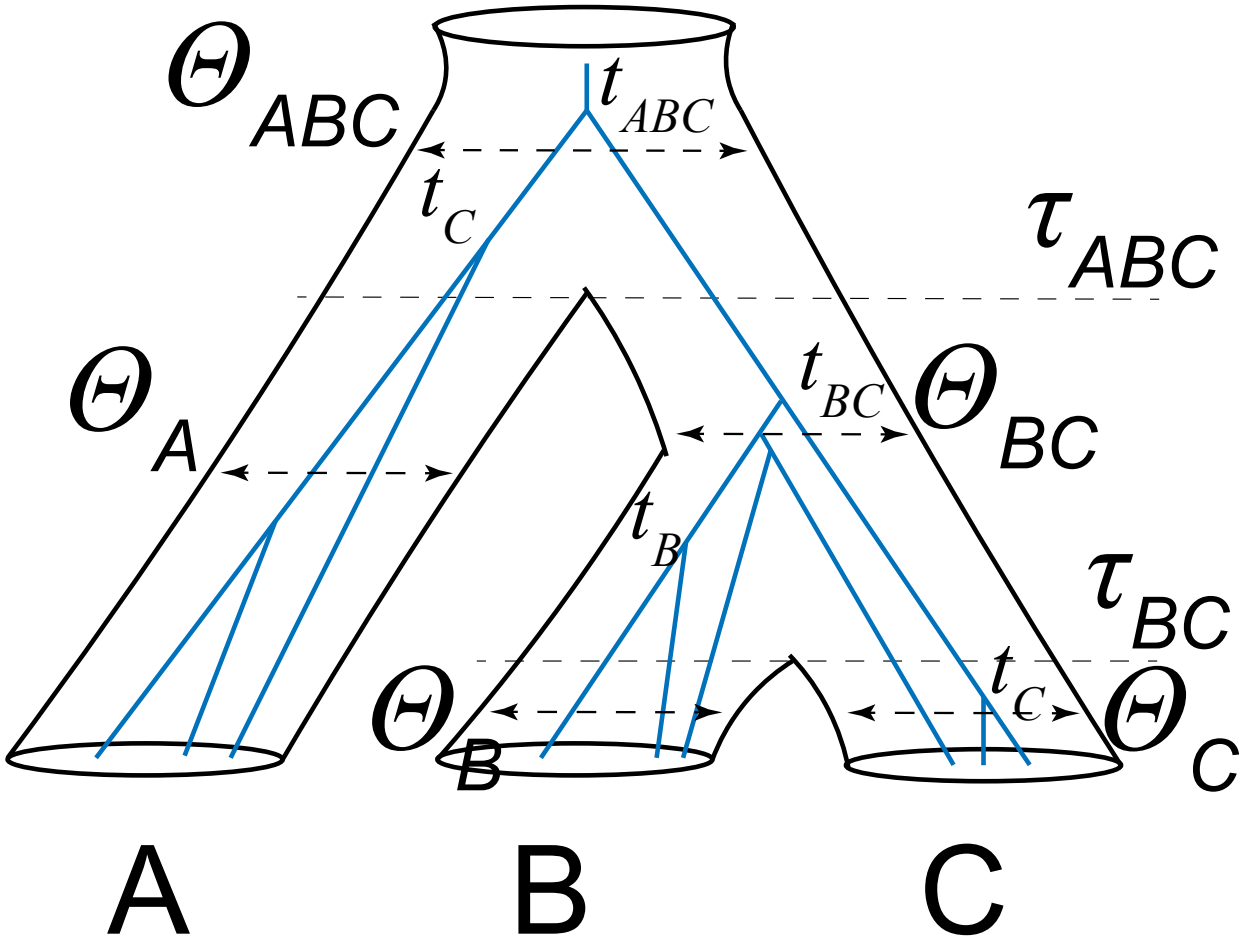
Bruce Rannala* and Ziheng Yang†,1

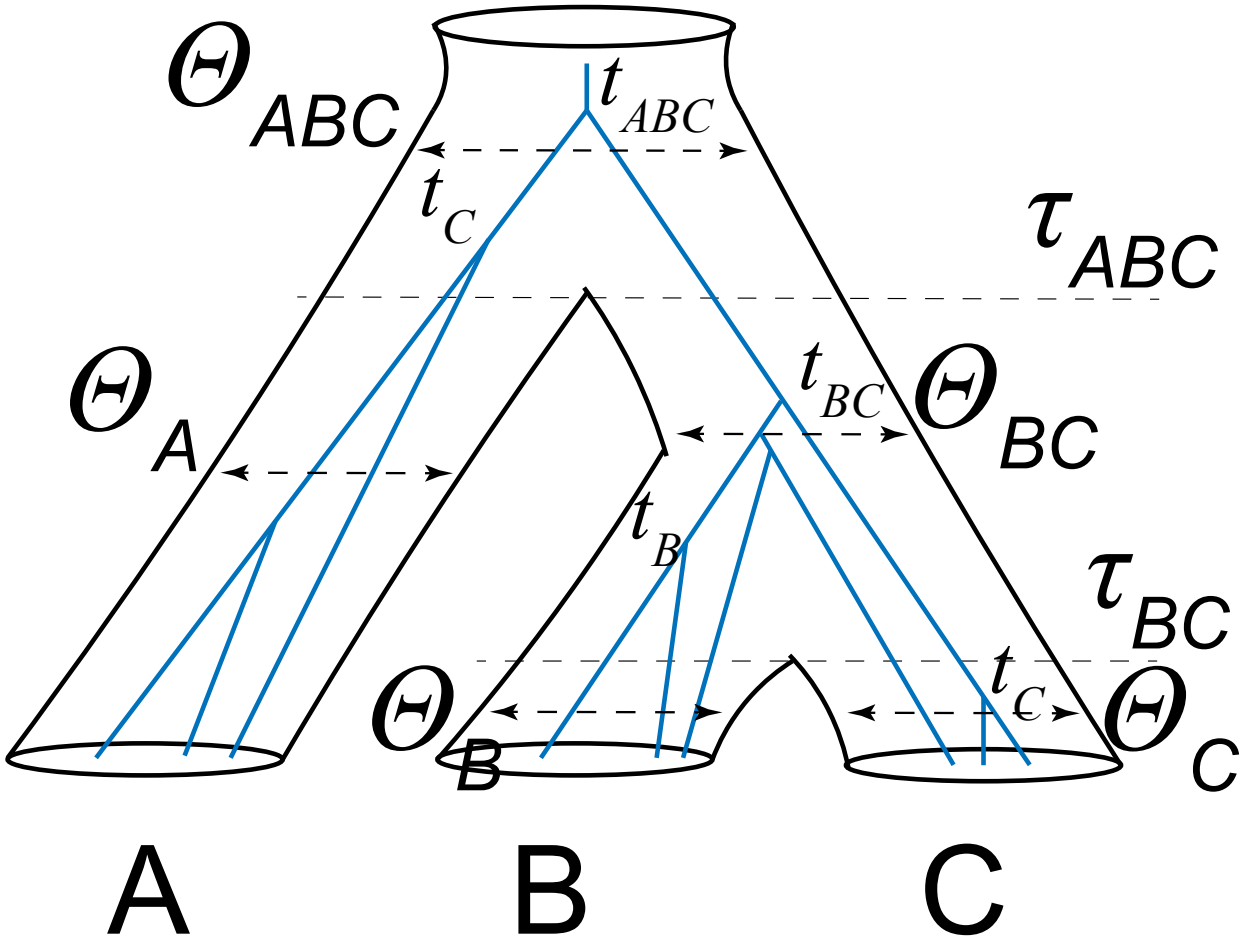
**Department of Medical Genetics, University of Alberta, Edmonton, Alberta T6G 2H7, Canada and †Galton Laboratory, Department of Biology, University College London, London WC1E 6BT, England*

Manuscript received December 4, 2002
Accepted for publication April 18, 2003

The Multispecies Coalescent (MSC)

The Censored Coalescent


Θ_{ABC}
t_{ABC}
t_C
τ_{ABC}
Θ_A
t_{BC}
Θ_{BC}
t_B
τ_{BC}
Θ_B
t_C
Θ_C
A
B
C



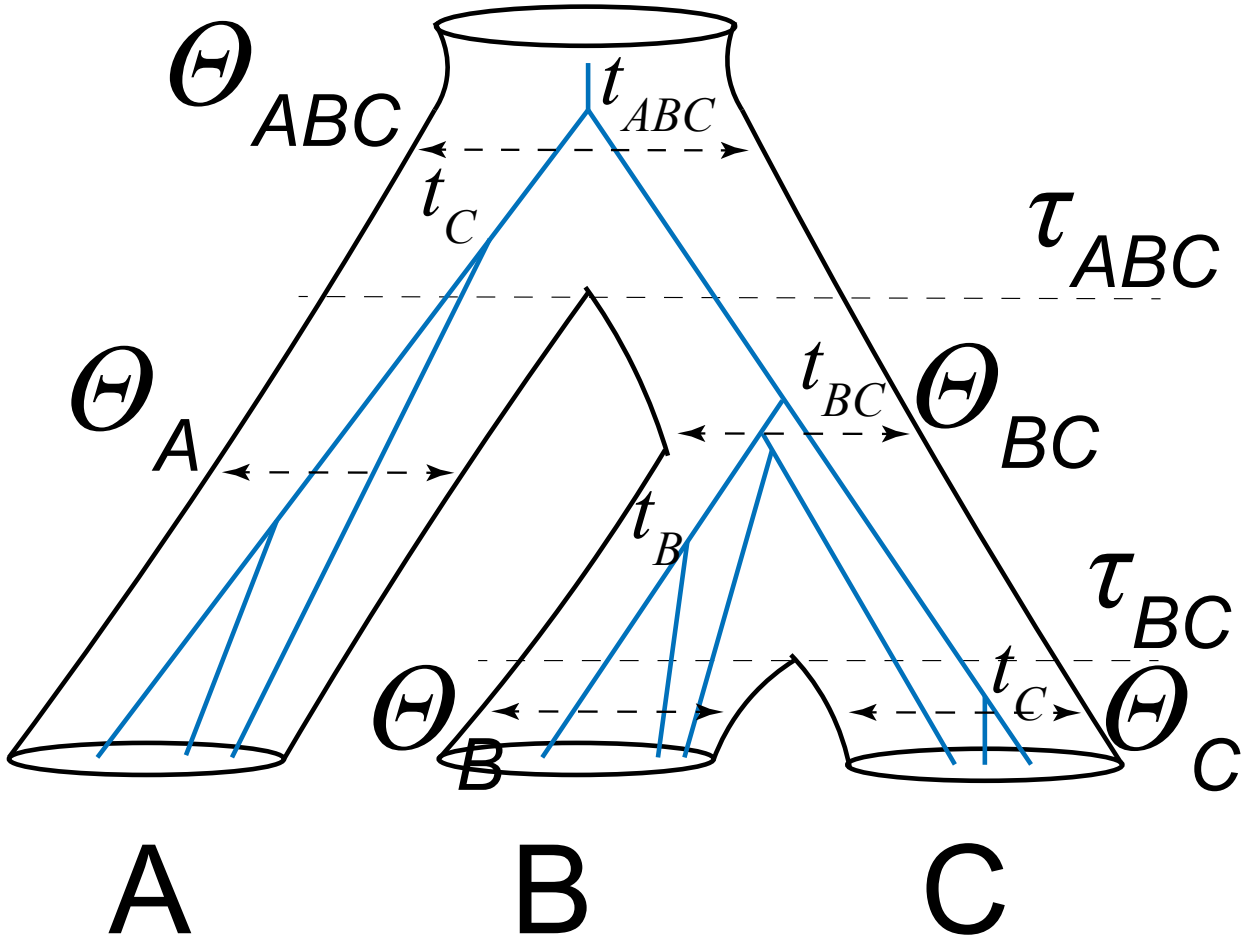
Simple Extension of coalescent for a single population

Lineages join independently in different populations

Coalescent rate is reset when lineages enter a new species

Coalescent times are measured by the expected number of mutations per site

Two Species Tree Parameters:
τ – Divergence Times
θ – Population Sizes

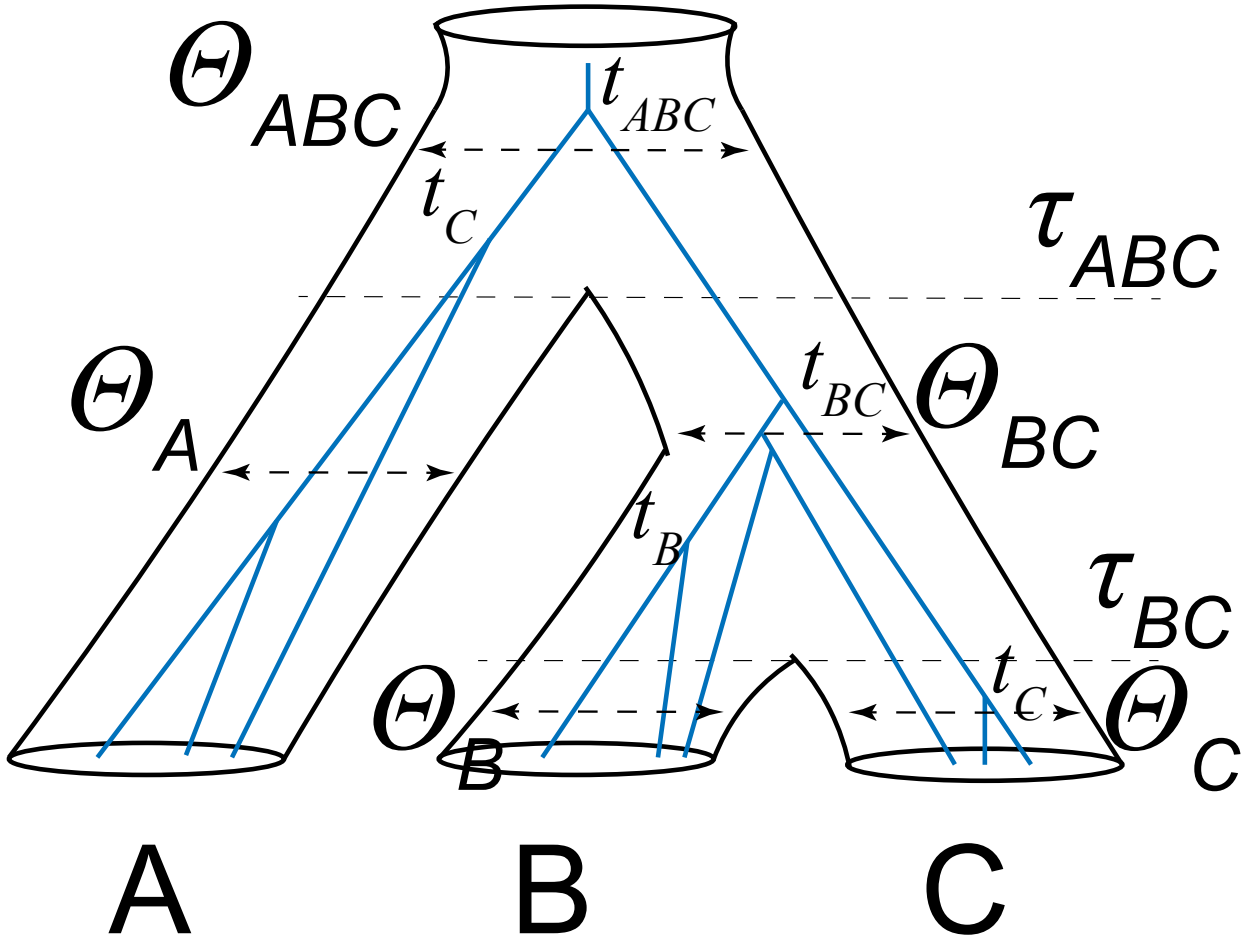


MSC Likelihood Function

Integrate over coalescent times

Felsenstein Likelihood

$$\ell(\Theta) = \sum_{i=1}^{L} log f(X_i|\Theta) = \sum_{i=1}^{L} log \left\{ \sum_{G_i} \int_{t_i} f(G_i, t_i|\Theta) f(X_i|G_i, t_i) dt_i \right\}$$



Note that the MSC accommodates multiple individuals per species

Learning Goals

Gene tree variation

Coalescent Theory

The multispecies coalescent

Species tree estimation

Application to taxonomy

Species tree estimation

Many methods and software available

Full-likelihood

- Only possible with Bayesian methods
- Can potentially be very time consuming or even impossible for many loci with a moderate number of tips

Species tree estimation

Many methods and software available

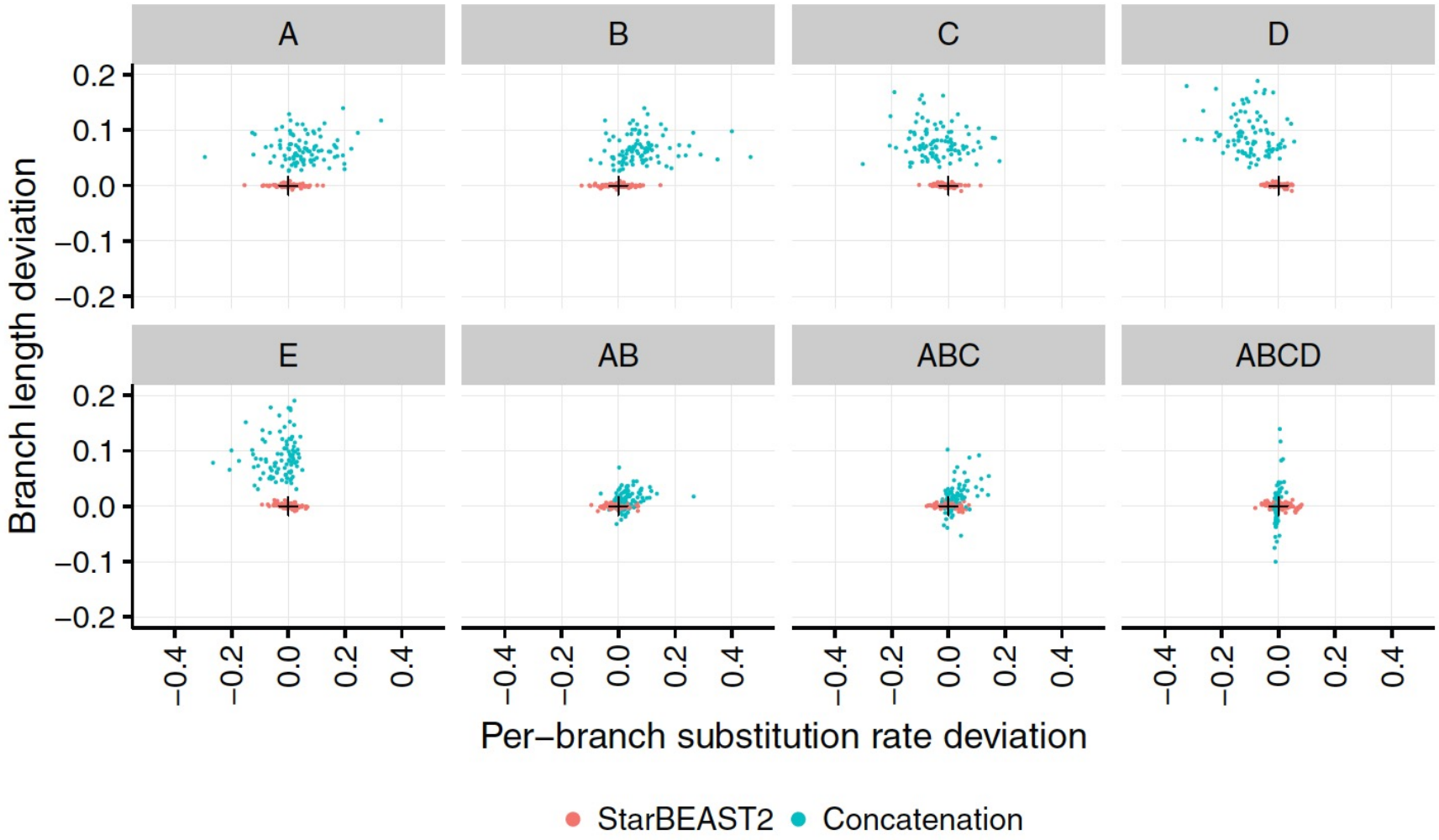
StarBeast2

$$\ell(\Theta) = \sum_{i=1}^{L} log f(X_i|\Theta) = \sum_{i=1}^{L} log \left\{ \sum_{G_i} \int_{t_i} f(G_i, t_i|\Theta) f(X_i|G_i, t_i) dt_i \right\}$$

You provide the sequence data only and the model jointly estimates the gene trees and species tree

Species tree estimation

The MSC can estimate branch lengths with less error than concatenation in the presence of ILS



Ogilvie et al. 2017

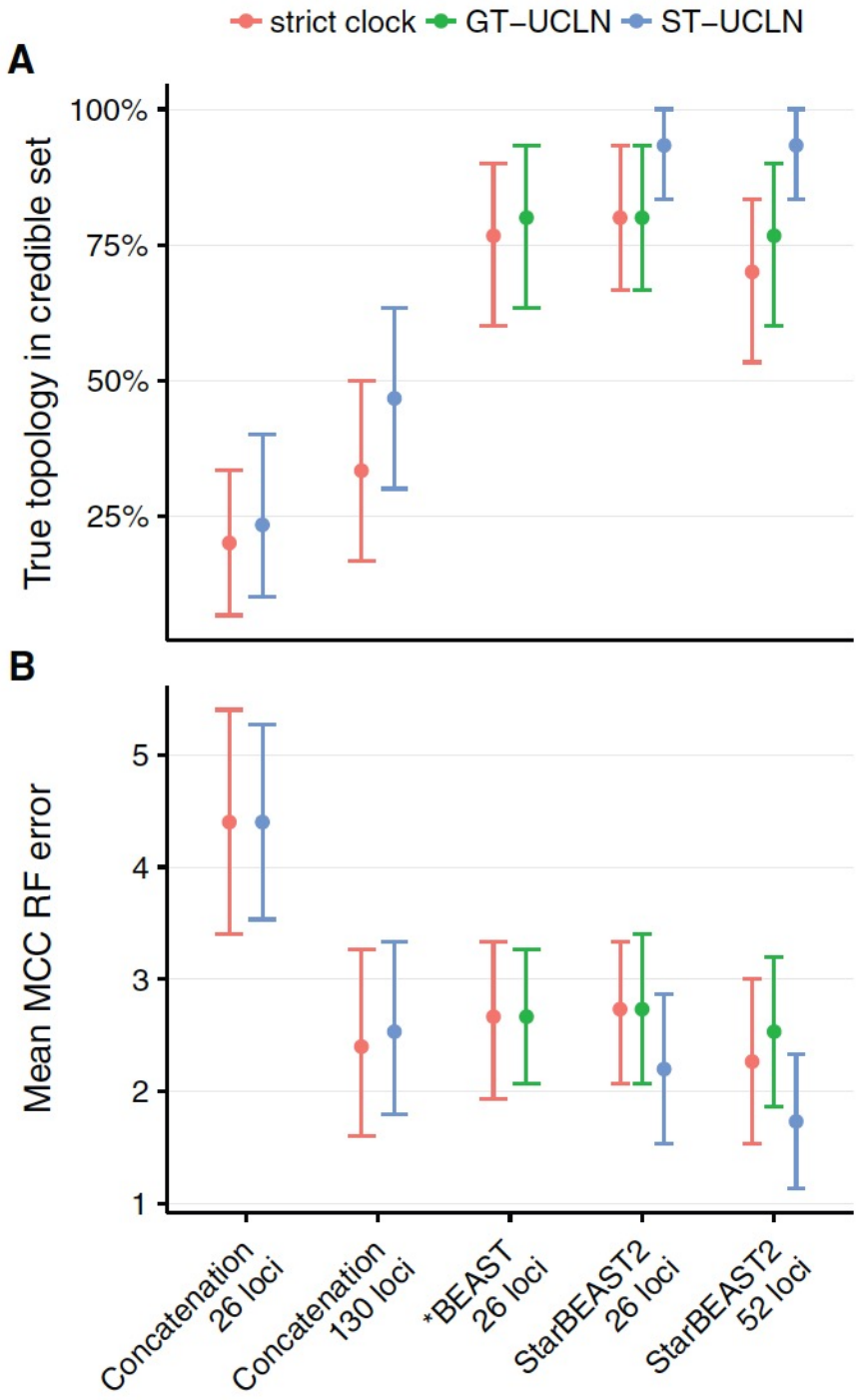
FIG. 2. Accuracy of branch substitution rates and lengths inferred by BEAST concatenation and StarBEAST2. Deviation is the difference of each estimated rate and length from the true value. Estimated rates and lengths are the posterior expectation of the overall substitution rate and length for each species tree branch. Black crosses in each panel indicate the point of perfect accuracy. Each panel shows the distributions for the labeled extant or ancestral branch. $N = 96$.

Species tree estimation

Concatenation would be misleading in their simulations as it often recovers an incorrect topology.

*BEAST is an older version of StarBEAST2

The software is flexible with many options, but takes time to learn



Ogilvie et al. 2017

Species tree estimation

Many methods and software available

BPP

$$\ell(\Theta) = \sum_{i=1}^{L} log f(X_i|\Theta) = \sum_{i=1}^{L} log \left\{ \sum_{G_i} \int_{t_i} f(G_i, t_i|\Theta) f(X_i|G_i, t_i) dt_i \right\}$$

You provide the sequence data only and the model jointly estimates the gene trees and species tree

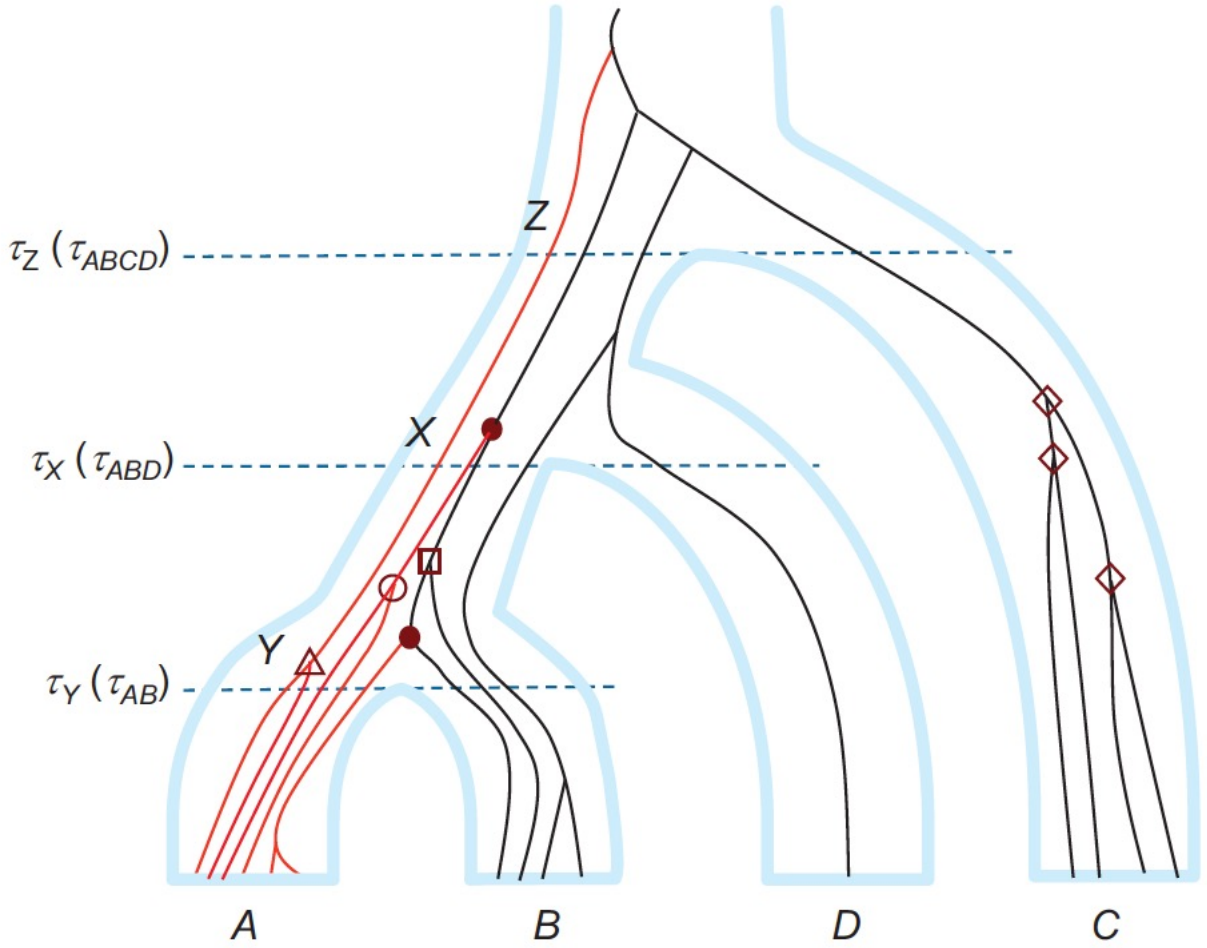
Species tree estimation

Many methods and software available

BPP

Works similarly to starBEAST2, just different software

I prefer using this one



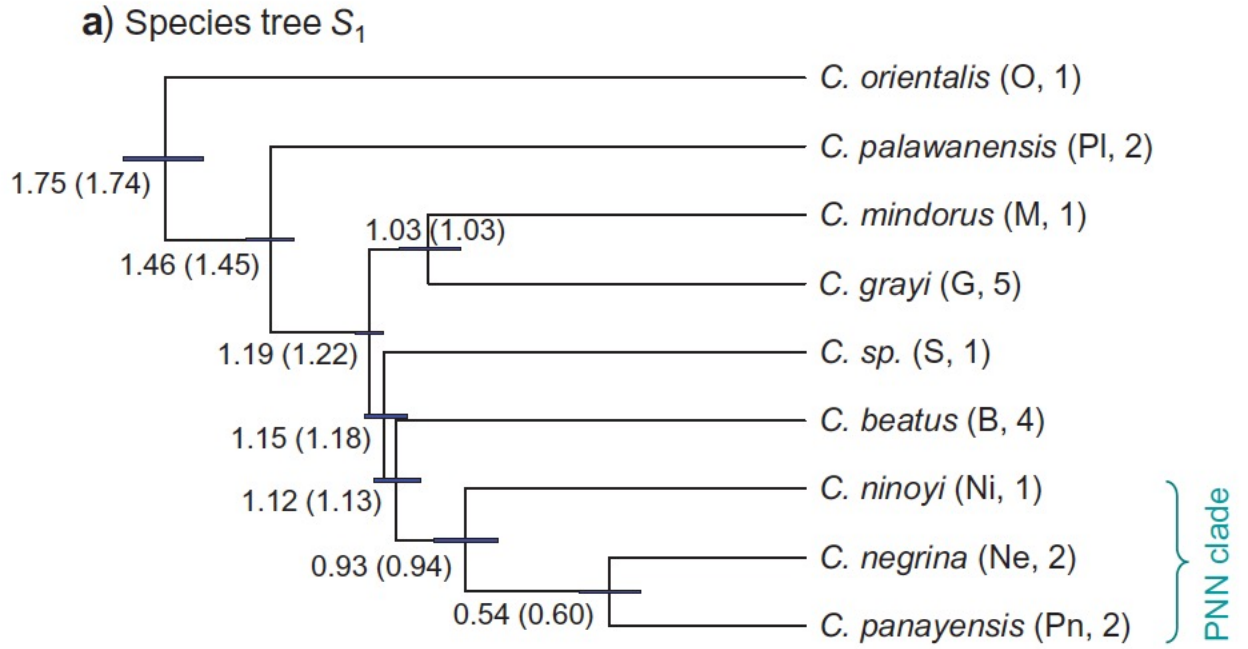
Rannala and Yang 2017

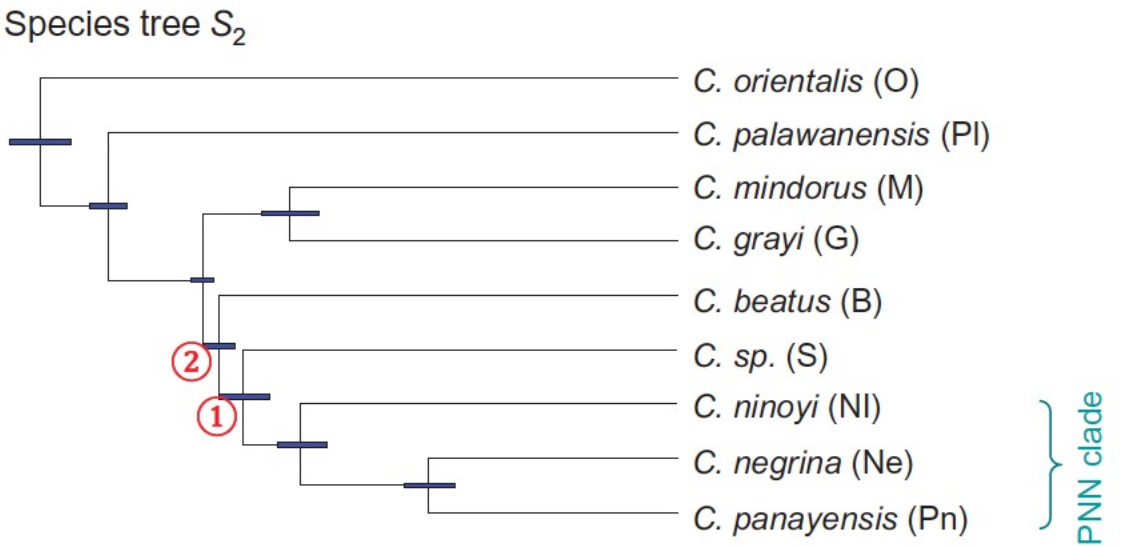
Species tree estimation

BPP

An added benefit of Bayesian software is that you will get multiple species trees with different posterior probabilities as a result

Useful for evaluating uncertainty





Rannala and Yang 2017

Species tree estimation

Many methods and software available

Summary methods

- Use pre-estimated gene trees to bypass that green part of the MSC likelihood function
- Scalable to very large phylogenomics data
- There are software that exist or have existed for this, but we will only focus on one

Species tree estimation

Many methods and software available

ASTRAL

$$\ell(\Theta) = \sum_{i=1}^{L} log f(X_i|\Theta) = \sum_{i=1}^{L} log \left\{ \sum_{G_i} \int_{t_i} f(G_i, t_i|\Theta) f(X_i|G_i, t_i) dt_i \right\}$$

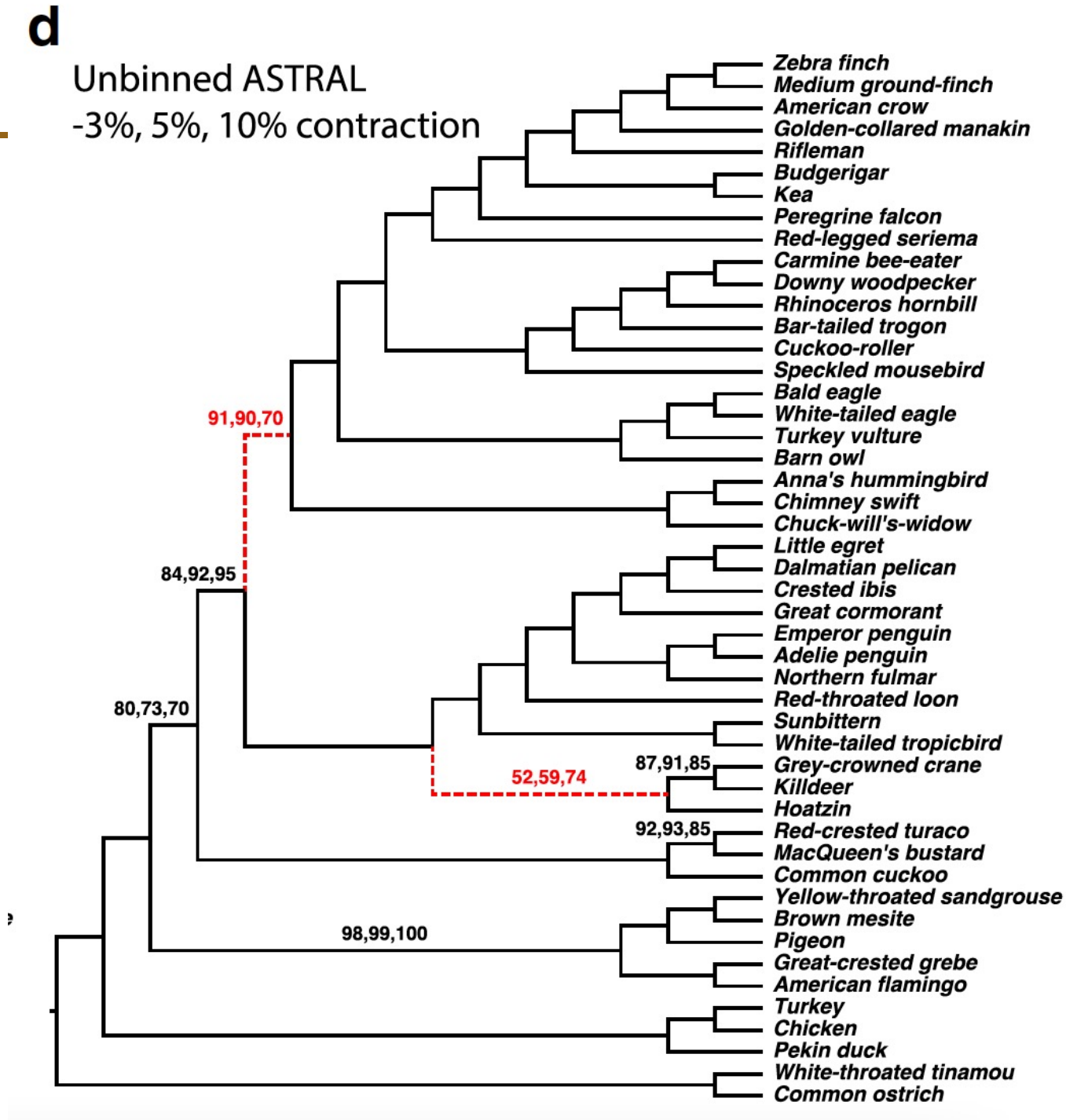
You provide the gene trees only and the model estimates species tree assuming the gene trees are correct.

Zhang et al. 2018

Species tree estimation

ASTRAL

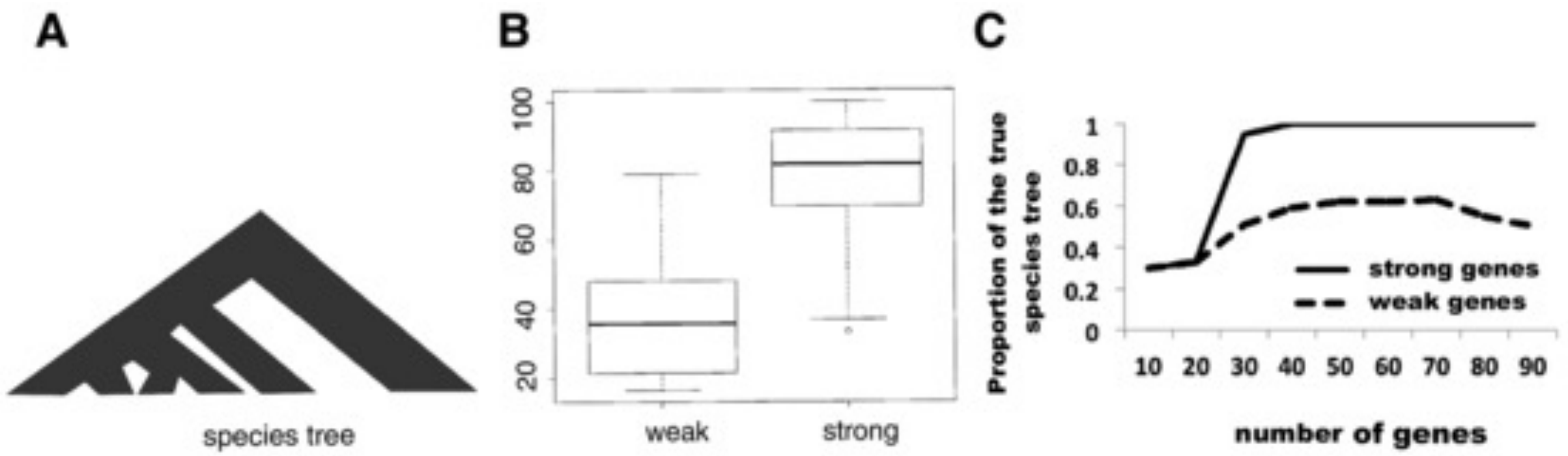
Summary methods make it possible to estimate large species trees with thousands of loci!



Zhang et al. 2018

Species tree estimation

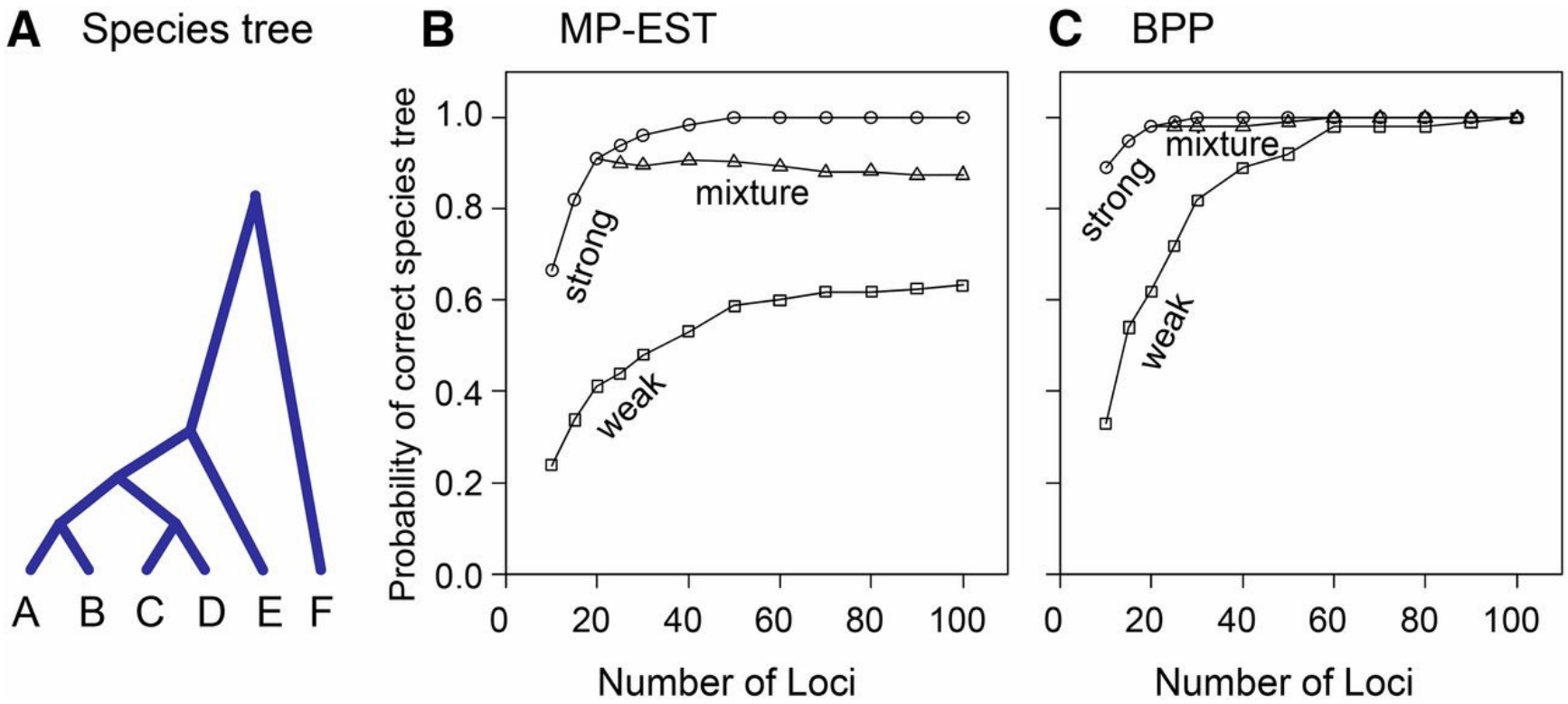
There are valid concerns that summary methods can be wrong when the gene trees are wrong or are uninformative



Liu et al. 2015. Annals of the New York Academy of Sciences. 1360:36-53

Species tree estimation

There are valid concerns that summary methods can be wrong when the gene trees are wrong or are uninformative



Xu and Yang. 2016. Genetics. 204:1353-1368

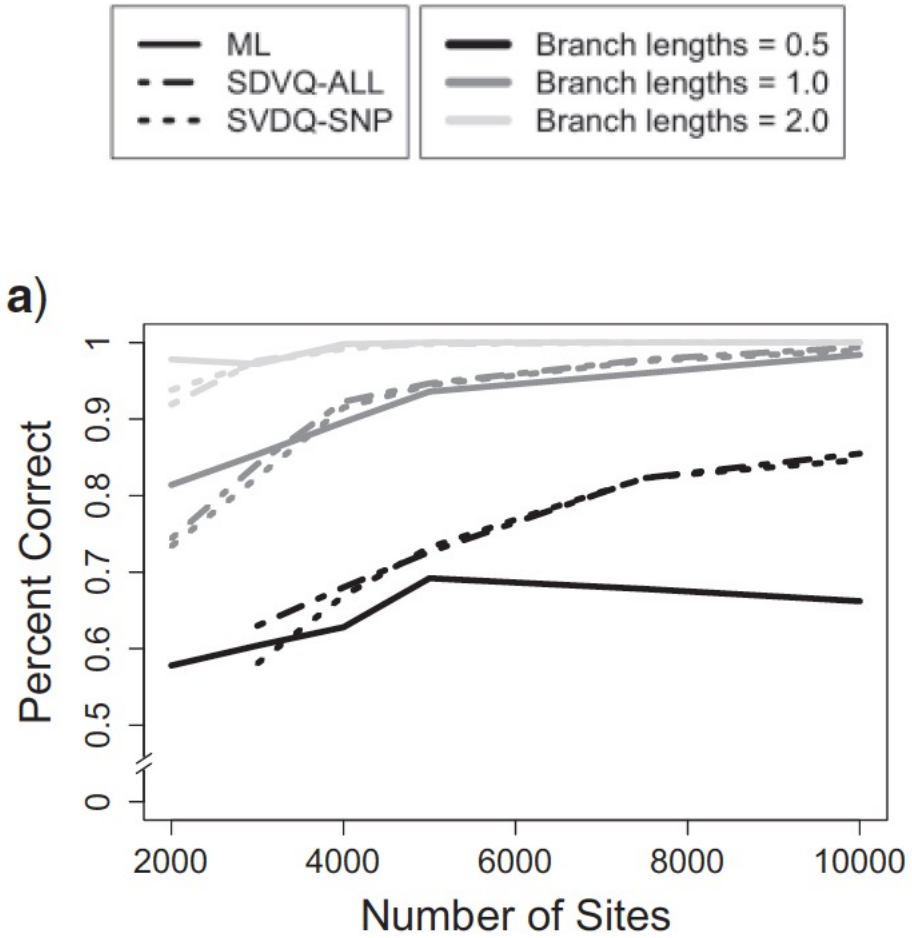
Species tree estimation

Many methods and software available

SVDquartets

An alternative to the other two approaches that does not rely on gene trees!

Has been shown to work well for whole sequences or only single nucleotide polymorphisms (SNPs), which is good for populations.



$\theta = 0.001$

Wascher and Kubatko 2021

Learning Goals

Gene tree variation

Coalescent Theory

The multispecies coalescent

Species tree estimation

Application to taxonomy

Application to taxonomy

Syst. Biol. 56(6):879–886, 2007

ISSN: 1063-5157 print / 1076-836X online
DOI: 10.1080/10635150701701083

Species Concepts and Species Delimitation

KEVIN DE QUEIROZ

Department of Vertebrate Zoology, National Museum of Natural History, Smithsonian Institution, Washington, DC 20560-0162, USA; E-mail: dequeirozk@si.edu

Abstract.— The issue of species delimitation has long been confused with that of species conceptualization, leading to a half century of controversy concerning both the definition of the species category and methods for inferring the boundaries and numbers of species. Alternative species concepts agree in treating existence as a separately evolving metapopulation lineage as the primary defining property of the species category, but they disagree in adopting different properties acquired by lineages during the course of divergence (e.g., intrinsic reproductive isolation, diagnosability, monophyly) as secondary defining properties (secondary species criteria). A unified species concept can be achieved by treating existence as a separately evolving metapopulation lineage as the only necessary property of species and the former secondary species criteria as different lines of evidence (operational criteria) relevant to assessing lineage separation. This unified concept of species has several consequences for species delimitation, including the following: First, the issues of species conceptualization and species delimitation are clearly separated; the former secondary species criteria are no longer considered relevant to species conceptualization but only to species delimitation. Second, all of the properties formerly treated as secondary species criteria are relevant to species delimitation to the extent that they provide evidence of lineage separation. Third, the presence of any one of the properties (if appropriately interpreted) is evidence for the existence of a species, though more properties and thus more lines of evidence are associated with a higher degree of corroboration. Fourth, and perhaps most significantly, a unified species concept shifts emphasis away from the traditional species criteria, encouraging biologists to develop new methods of species delimitation that are not tied to those properties. [Species concept; species criteria; species delimitation.]

TABLE 1. Alternative contemporary species concepts (i.e., major classes of contemporary species definitions) and the properties upon which they are based (modified from de Queiroz, 2005). Properties (or the converses of properties) that represent thresholds crossed by diverging lineages and that are commonly viewed as necessary properties of species are marked with an asterisk (*). Note that under the proposal for unification described in this paper, the various ideas summarized in this table would no longer be considered distinct species concepts (see de Queiroz, 1998, for an alternative terminology). All of these ideas conform to a single general concept under which species are equated with separately evolving metapopulation lineages, and many of the properties (*) are more appropriately interpreted as operational criteria (lines of evidence) relevant to assessing lineage separation.

Species concept	Property(ies)	Advocates/references
Biological	Interbreeding (natural reproduction resulting in viable and fertile offspring)	Wright (1940); Mayr (1942); Dobzhansky (1950)
Isolation	*Intrinsic reproductive isolation (absence of interbreeding between heterospecific organisms based on intrinsic properties, as opposed to extrinsic [geographic] barriers)	Mayr (1942); Dobzhansky (1970)
Recognition	*Shared specific mate recognition or fertilization system (mechanisms by which conspecific organisms, or their gametes, recognize one another for mating and fertilization)	Paterson (1985); Masters et al. (1987); Lambert and Spencer (1995)
Ecological	*Same niche or adaptive zone (all components of the environment with which conspecific organisms interact)	Van Valen (1976); Andersson (1990)
Evolutionary	Unique evolutionary role, tendencies, and historical fate	Simpson (1951); Wiley (1978); Mayden (1997)
(some interpretations)	*Diagnosability (qualitative, fixed difference)	Grismer (1999, 2001)
Cohesion	Phenotypic cohesion (genetic or demographic exchangeability)	Templeton (1989, 1998a)
Phylogenetic	Heterogeneous (see next four entries)	(see next four entries)
Hennigian	Ancestor becomes extinct when lineage splits	Hennig (1966); Ridley (1989); Meier and Willmann (2000)
Monophyletic	*Monophyly (consisting of an ancestor and all of its descendants; commonly inferred from possession of shared derived character states)	Rosen (1979); Donoghue (1985); Mishler (1985)
Genealogical	*Exclusive coalescence of alleles (all alleles of a given gene are descended from a common ancestral allele not shared with those of other species)	Baum and Shaw (1995); see also Avise and Ball (1990)
Diagnosable	*Diagnosability (qualitative, fixed difference)	Nelson and Platnick (1981); Cracraft (1983); Nixon and Wheeler (1990)
Phenetic	*Form a phenetic cluster (quantitative difference)	Michener (1970); Sokal and Crovello (1970); Sneath and Sokal (1973)
Genotypic cluster (definition)	*Form a genotypic cluster (deficits of genetic intermediates; e.g., heterozygotes)	Mallet (1995)

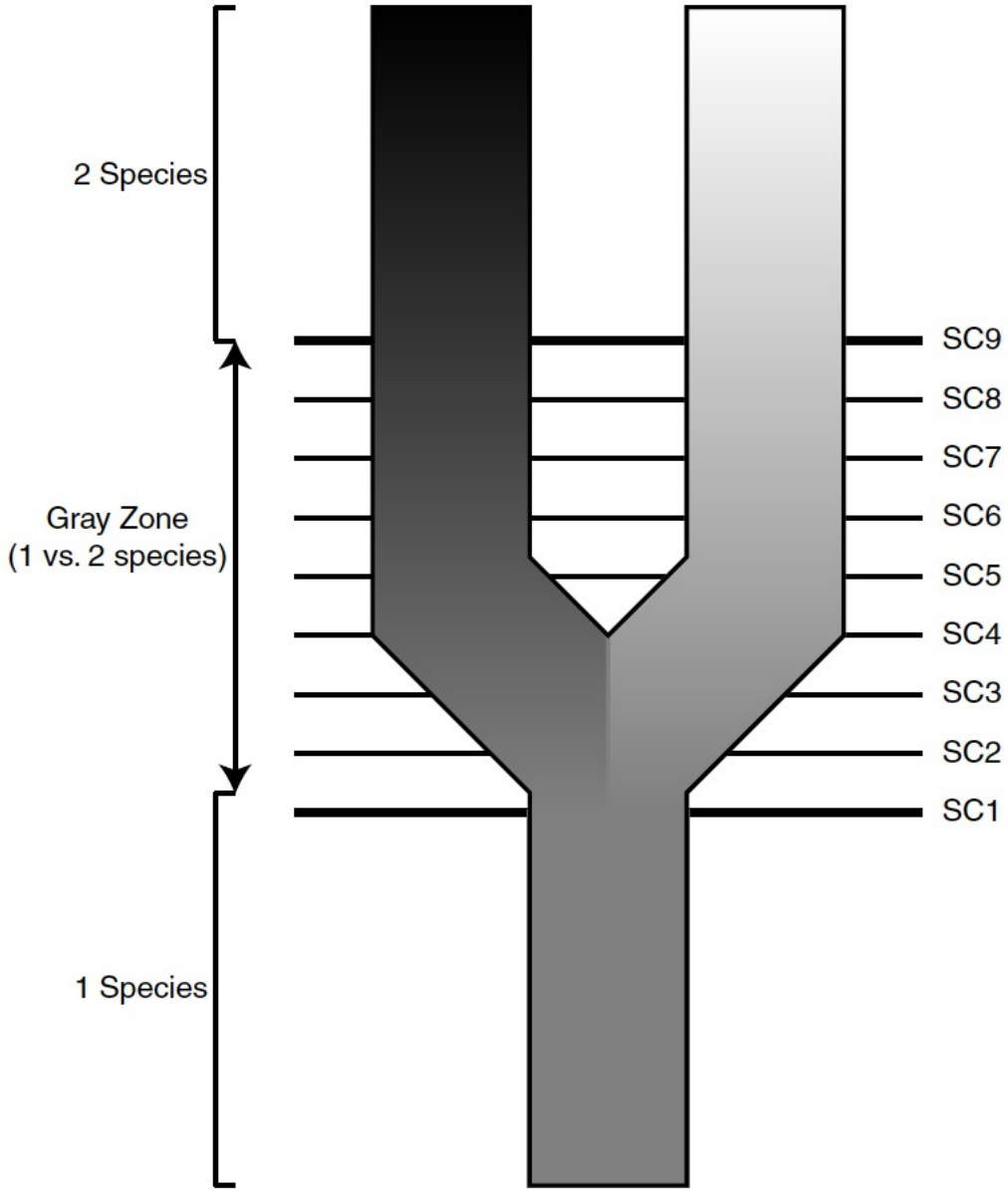


FIGURE 1. Lineage separation and divergence (speciation) and species concepts (after de Queiroz, 1998, 1999, 2005a). This highly simplified diagram represents a single lineage (species) splitting to form two lineages (species). The gradations in shades of gray represent the daughter lineages diverging through time, and the horizontal lines labeled SC (species criterion) 1 to 9 represent the times at which they acquire different properties (i.e., when they become phenetically distinguishable, diagnosable, reciprocally monophyletic, reproductively incompatible, ecologically distinct, etc.). The entire set of properties forms a gray zone within which alternative species concepts come into conflict. On either side of the gray zone, there will be unanimous agreement about the number of species. Before the acquisition of the first property, everyone will agree that there is a single species, and after the acquisition of the last property, everyone will agree that there are two. In between, however, there will be disagreement. The reason is that different contemporary species concepts adopt different properties (represented by the horizontal lines) as their species criteria—that is, as their cutoffs for considering a separately evolving lineage to have become a species.

Barcoding
And
Single Locus Delimitation

THE ROYAL SOCIETY

Received 29 July 2002
Accepted 30 September 2002
Published online 8 January 2003

Biological identifications through DNA barcodes

Paul D. N. Hebert*, Alina Cywinska, Shelley L. Ball and Jeremy R. deWaard

Department of Zoology, University of Guelph, Guelph, Ontario N1G 2W1, Canada

Although much biological research depends upon species diagnoses, taxonomic expertise is collapsing. We are convinced that the sole prospect for a sustainable identification capability lies in the construction of systems that employ DNA sequences as taxon 'barcodes'. We establish that the mitochondrial gene cytochrome *c* oxidase I (COI) can serve as the core of a global bioidentification system for animals. First, we demonstrate that COI profiles, derived from the low-density sampling of higher taxonomic categories, ordinarily assign newly analysed taxa to the appropriate phylum or order. Second, we demonstrate that species-level assignments can be obtained by creating comprehensive COI profiles. A model COI profile, based upon the analysis of a single individual from each of 200 closely allied species of lepidopterans, was 100% successful in correctly identifying subsequent specimens. When fully developed, a COI identification system will provide a reliable, cost-effective and accessible solution to the current problem of species identification. Its assembly will also generate important new insights into the diversification of life and the rules of molecular evolution.

A DNA barcode for land plants

CBOL Plant Working Group[1]

Communicated by Daniel H. Janzen, University of Pennsylvania, Philadelphia, PA, May 27, 2009 (received for review March 18, 2009)

DNA barcoding involves sequencing a standard region of DNA as a tool for species identification. However, there has been no agreement on which region(s) should be used for barcoding land plants. To provide a community recommendation on a standard plant barcode, we have compared the performance of 7 leading candidate plastid DNA regions (*atpF–atpH* spacer, *matK* gene, *rbcL* gene, *rpoB* gene, *rpoC1* gene, *psbK–psbI* spacer, and *trnH–psbA* spacer). Based on assessments of recoverability, sequence quality, and levels of species discrimination, we recommend the 2-locus combination of *rbcL+matK* as the plant barcode. This core 2-locus barcode will provide a universal framework for the routine use of DNA sequence data to identify specimens and contribute toward the discovery of overlooked species of land plants.

matK | *rbcL* | species identification

■ large-scale standardized sequencing of the mitochondrial

intergenic spacers *trnH–psbA* and *psbK–psbI*, in part attributable to a high frequency of mononucleotide repeats disrupting individual sequencing reads.

Species Discrimination. Among 397 samples successfully sequenced for all 7 loci, species discrimination for single-locus barcodes ranged from 43% (*rpoC1*) to 68%–69% (*psbK–psbI* and *trnH–psbA*), with *rbcL* and *matK* providing 61% and 66% discrimination respectively (rank order: *rpoC1*<*rpoB*<*atpF–*

Author contributions: P.M.H., L.L.F., J.L.S., M.H., S.R., M.v.d.B., M.W.C., R.S.C., D.L.E., A.J.F., S.W.G., K.E.J., K.-J.K., W.J.K., H.S., S.C.H.B., C.v.d.B., M.C., T.A.J.H., B.C.H., G.P., J.E.R., G.A.S., V.S., O.S., M.J.W., and D.P.L. designed research; D.L.E., A.J.F., K.E.J., J.v.A.S., D.B., K.S.B., K.M.C., J.C., A.C., J.J.C., F.C., D.S.D., C.S.F., M.L.H., L.J.K., P.R.K., J.S.K., Y.D.K., R.L., H.-L.L., D.G.L., S.M., O.M., I.M., S.G.N., C.-W.P., D.M.P., and D.-K.Y. performed research; L.L.F., J.L.S., M.H., S.R., and D.P.L. analyzed data; and P.M.H., S.W.G., S.C.H.B., and D.P.L. wrote the paper.

NCBI Resources How To

Nucleotide

Nucleotide (Magnoliopsida[Organism]) AND rbcL[Gene Name]

Create alert Advanced

COVID-19 is an emerging, rapidly evolving situation.
Get the latest public health information from CDC: https://www.coronavirus.gov .
Get the latest research from NIH: https://www.nih.gov/coronavirus.

Species clear
- ✓ **Plants** (135,947)
- Customize ...

Molecule types clear
- ✓ **genomic DNA/RNA** (135,947)
- mRNA (0)
- Customize ...

Source databases
- INSDC (GenBank) (135,947)
- Customize ...

Sequence Type
- Nucleotide (135,947)

Genetic compartments
- Chloroplast (127,213)
- Mitochondrion (64)
- Plastid (134,648)

Sequence length clear
- ✓ **From 500 to 5000** (135,947)

Release date
- Custom range...

Revision date
- Custom range...

Clear all

Show additional filters

Summary 20 per page Sort by Default order Send to:

Items: 1 to 20 of 135947

<< First < Prev Page 1 of 6798 Next > Last >>

Filters activated: Plants, genomic DNA/RNA, Sequence length from 500 to 5000. Clear all

1. Silene aprica isolate XNR ribulose-1,5-bisphosphate carboxylase/oxygenase large subunit (**rbcL**) gene, partial cds; chloroplast
   553 bp linear DNA
   Accession: MK534839.1 GI: 1774218687
   GenBank FASTA Graphics PopSet

2. Silene fortunei isolate DZ1R ribulose-1,5-bisphosphate carboxylase/oxygenase large subunit (**rbcL**) gene, partial cds; chloroplast
   553 bp linear DNA
   Accession: MK534838.1 GI: 1774218685
   GenBank FASTA Graphics PopSet

3. Silene fortunei isolate DZ3R ribulose-1,5-bisphosphate carboxylase/oxygenase large subunit (**rbcL**) gene, partial cds; chloroplast
   553 bp linear DNA
   Accession: MK534837.1 GI: 1774218683
   GenBank FASTA Graphics PopSet

4. Silene fortunei isolate WZ1R ribulose-1,5-bisphosphate carboxylase/oxygenase large subunit (**rbcL**) gene, partial cds; chloroplast
   553 bp linear DNA
   Accession: MK534836.1 GI: 1774218681

Open access, freely available online PLoS BIOLOGY

DNA Barcoding: Error Rates Based on Comprehensive Sampling

Christopher P. Meyer*, Gustav Paulay

Florida Museum of Natural History, University of Florida, Gainesville, Florida, United States of America

DNA barcoding has attracted attention with promises to aid in species identification and discovery; however, few well-sampled datasets are available to test its performance. We provide the first examination of barcoding performance in a comprehensively sampled, diverse group (cypraeid marine gastropods, or cowries). We utilize previous methods for testing performance and employ a novel phylogenetic approach to calculate intraspecific variation and interspecific divergence. Error rates are estimated for (1) identifying samples against a well-characterized phylogeny, and (2) assisting in species discovery for partially known groups. We find that the lowest overall error for species identification is 4%. In contrast, barcoding performs poorly in incompletely sampled groups. Here, species delineation relies on the use of thresholds, set to differentiate between intraspecific variation and interspecific divergence. Whereas proponents envision a "barcoding gap" between the two, we find substantial overlap, leading to minimal error rates of ~17% in cowries. Moreover, error rates double if only traditionally recognized species are analyzed. Thus, DNA barcoding holds promise for identification in taxonomically well-understood and thoroughly sampled clades. However, the use of thresholds does not bode well for delineating closely related species in taxonomically understudied groups. The promise of barcoding will be realized only if based on solid taxonomic foundations.

Citation: Meyer CP, Paulay G (2005) DNA barcoding: Error rates based on comprehensive sampling. PLoS Biol 3(12): e422.

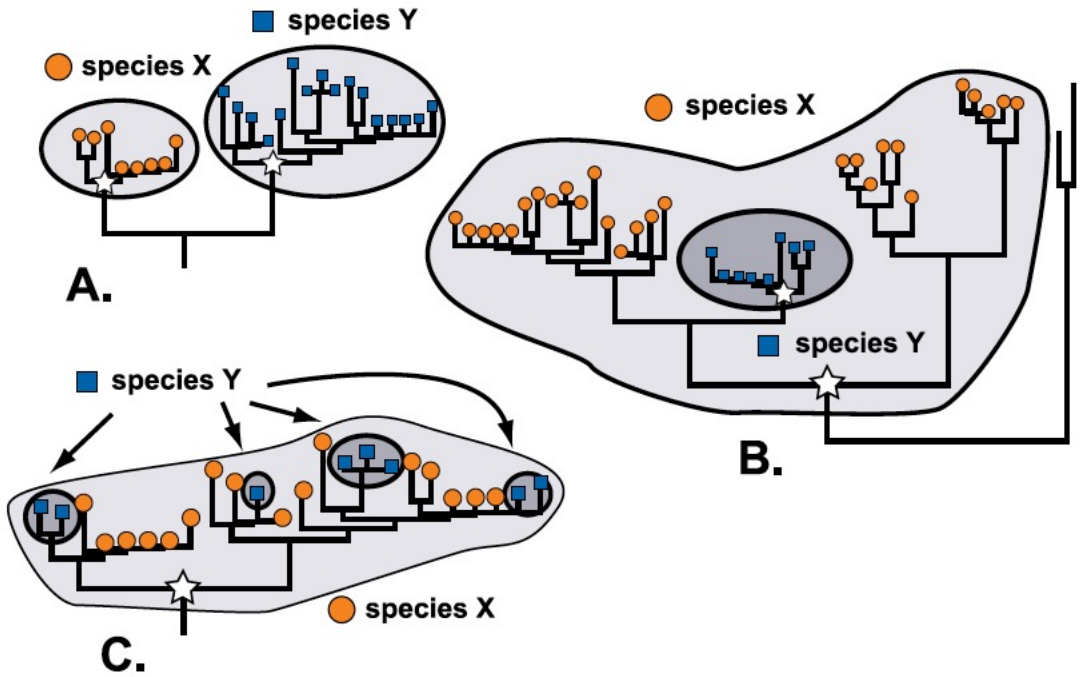


Figure 1. Phylogenetic Relationships and Terminology

(A) Reciprocal monophyly. Members of each species share a unique common ancestor. For each species, the white star represents the coalescent, the point at which all extant haplotypes share a common ancestry.

(B) Paraphyly. One species (Y), is monophyletic, but nests within another recognized species (X). Thus, the coalescent of species Y (small star) is contained within the coalescent of species X (large star).

(C) Polyphyly. Neither species X or Y are monophyletic, and both coalesce to the white star.

DOI: 10.1371/journal.pbio.0030422.g001

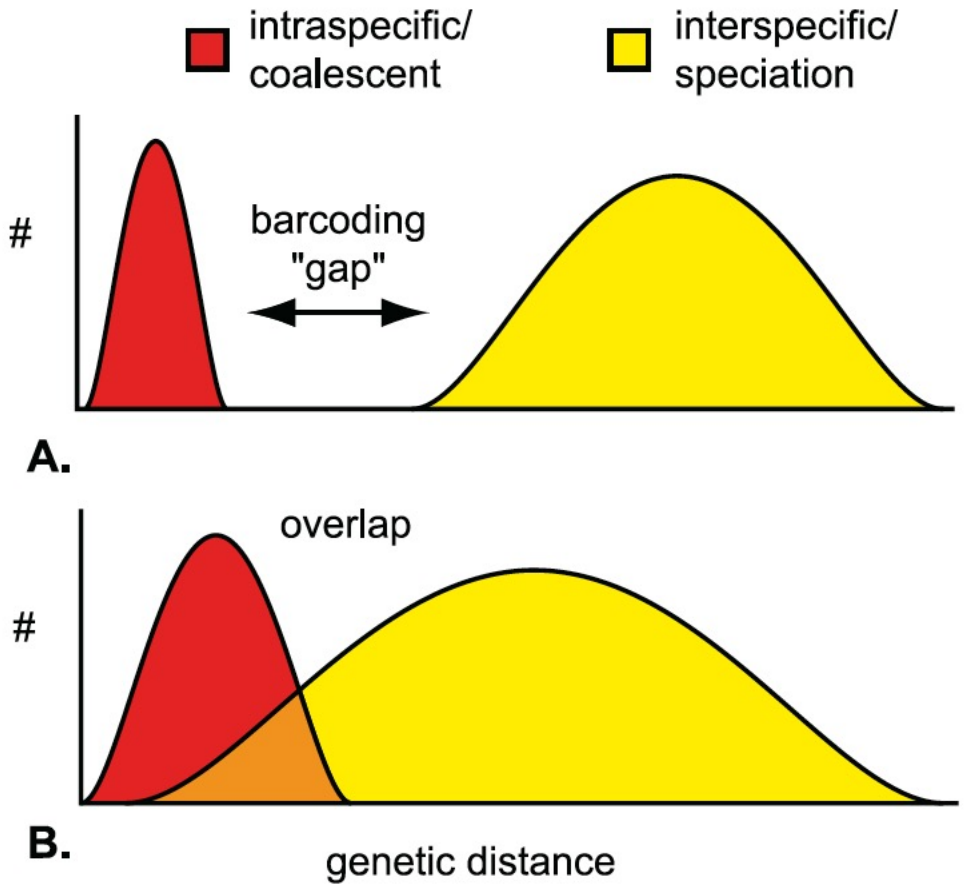


Figure 2. Schematic of the Inferred Barcoding Gap

The distribution of intraspecific variation is shown in red, and interspecific divergence in yellow. (A) Ideal world for barcoding, with discrete distributions and no overlap. (B) An alternative version of the world with significant overlap and no gap.

DOI: 10.1371/journal.pbio.0030422.g002

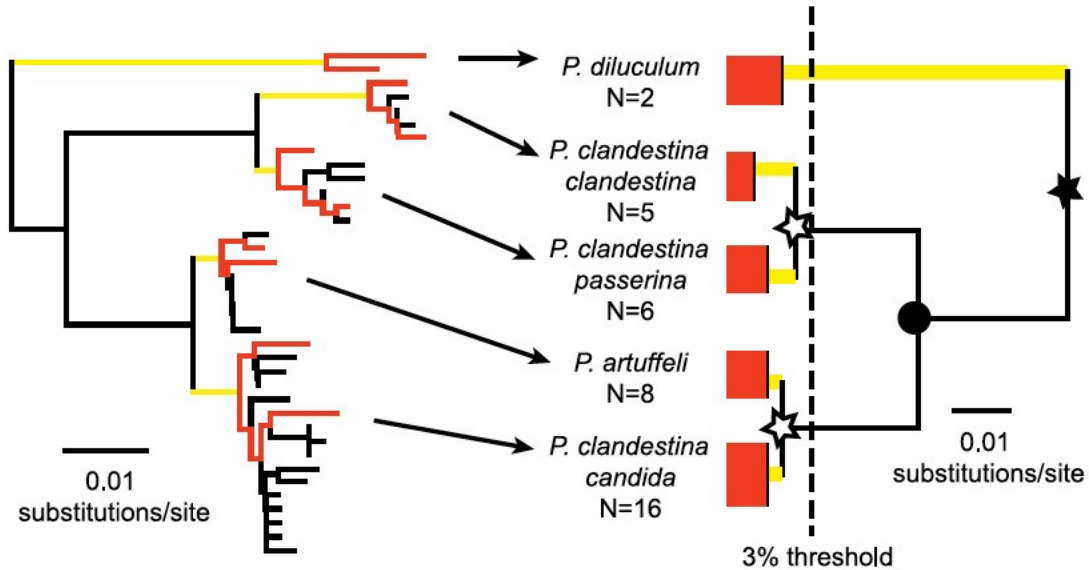


Figure 3. Intraspecific and Interspecific Estimations

A subclade of five cowrie ESUs shows how both coalescent and divergence depths are generated. The two most disparate individuals are culled from within each ESU (left—red) and used in a constrained phylogeny with a molecular clock enforced (right) to recover both the maximum coalescent depth (red) and the divergence depths between sisters (yellow). Two young ESUs (stars) would be missed (false negatives) if a 3% threshold cutoff (shown) was employed. Note that *Palmadusta artuffeli,* a Japanese endemic species, is nested among monophyletic subspecies of the paraphyletic species *P. clandestina*. The black circle indicates the coalescent for the species *P. clandestina,* and the black star indicates the interspecific divergence for species-level analyses.

DOI: 10.1371/journal.pbio.0030422.g003

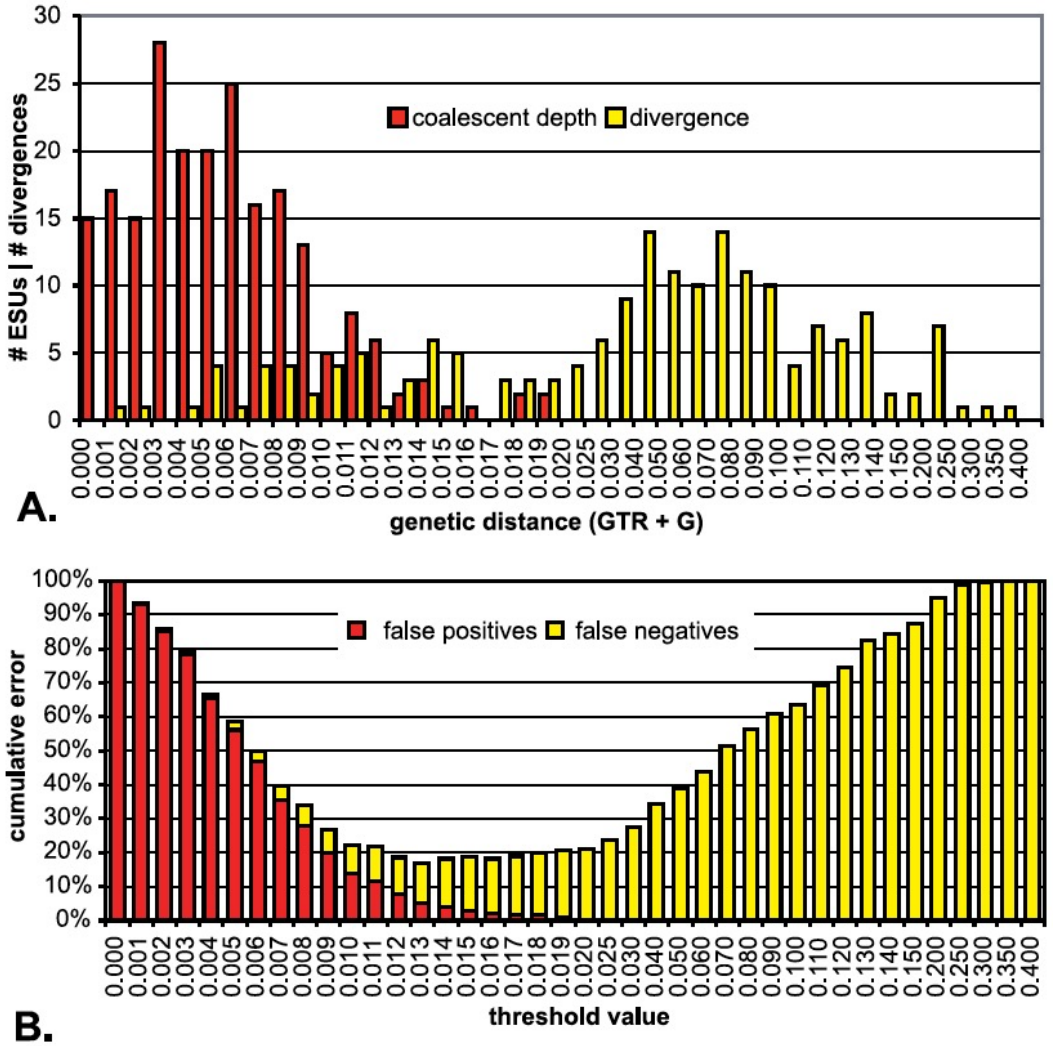


Figure 7. Barcoding Overlap: Cowrie ESUs

(A) Relative distributions of intraspecific variability (coalescent depth—red) and interspecific divergence between ESUs (yellow), demonstrating significant overlap and the lack of a barcoding gap. Note that the *x*-axis scale shifts to progressively greater increments above 0.02.

(B) Cumulative error based on false positives plus false negatives for each threshold value. The optimum threshold value is 0.013 (2.6%), where error is minimized at 17%.

DOI: 10.1371/journal.pbio.0030422.g007

A number of ad hoc distance thresholds were proposed for delimiting species

A single distance may not work for all taxonomic groups

Optimizing the threshold based on established taxonomy is possible

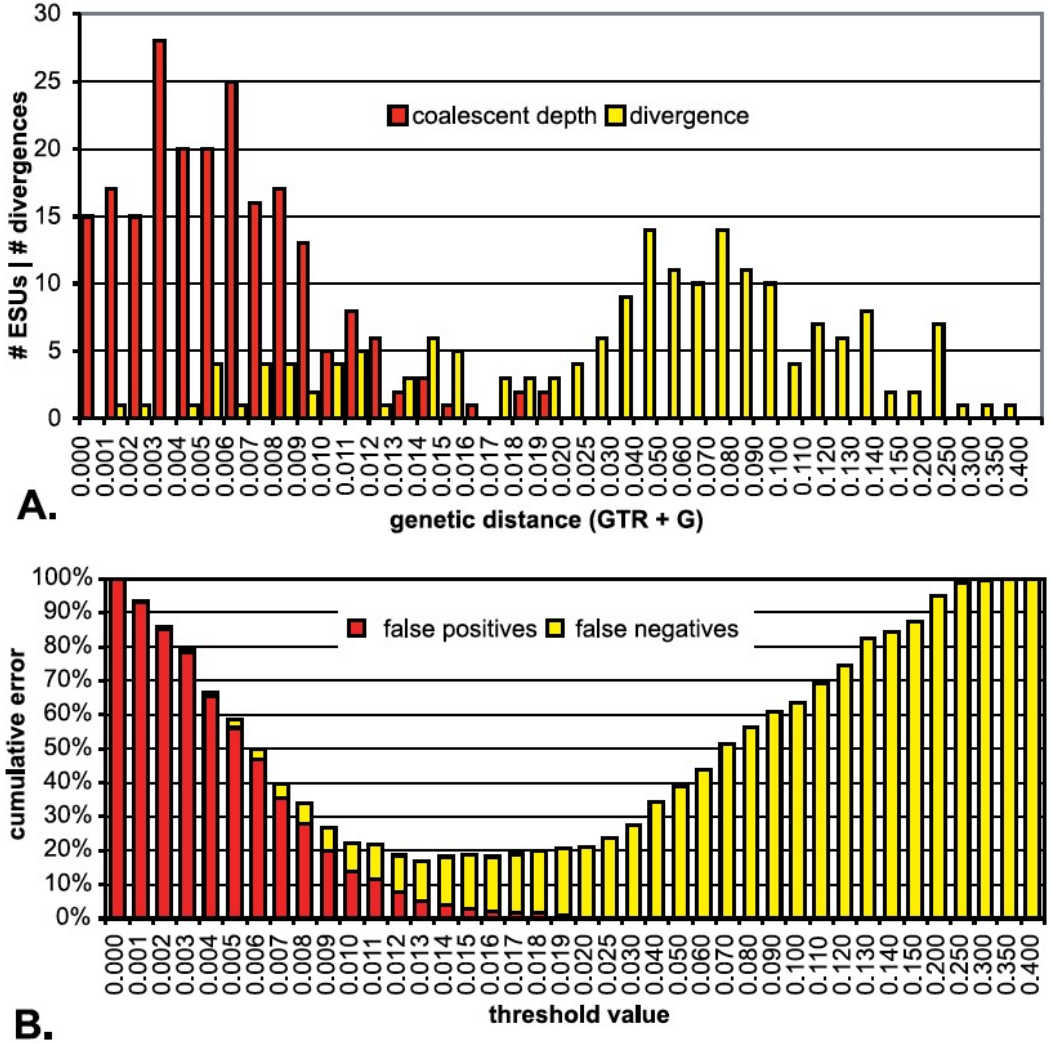


Figure 7. Barcoding Overlap: Cowrie ESUs

(A) Relative distributions of intraspecific variability (coalescent depth—red) and interspecific divergence between ESUs (yellow), demonstrating significant overlap and the lack of a barcoding gap. Note that the *x*-axis scale shifts to progressively greater increments above 0.02.

(B) Cumulative error based on false positives plus false negatives for each threshold value. The optimum threshold value is 0.013 (2.6%), where error is minimized at 17%.

DOI: 10.1371/journal.pbio.0030422.g007

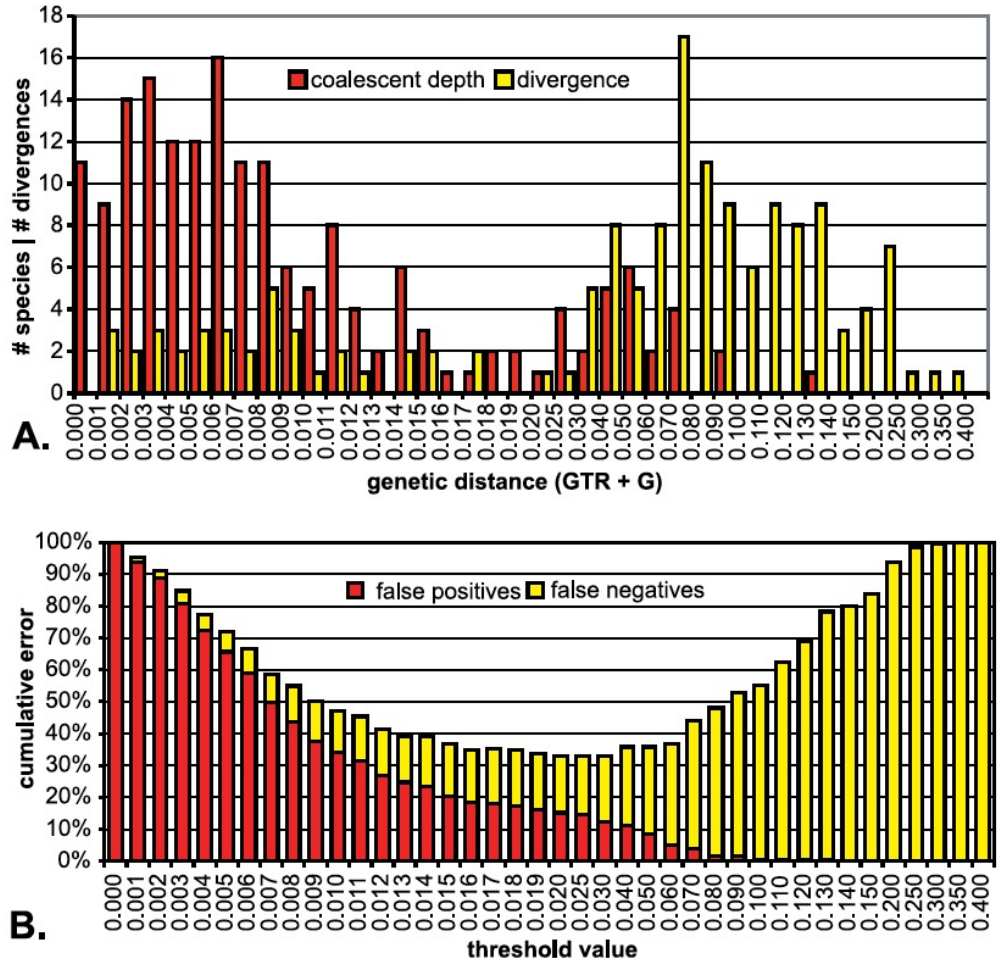


Figure 9. Barcoding Overlap: Cowrie Species

Data are presented as in Figure 7; however, estimates of intraspecific variation and interspecific divergence are based on traditionally recognized cowrie species.

(A) Relative distributions of intraspecific variability (coalescent depth—red) and interspecific divergence between species (yellow), demonstrating a more pronounced overlap than when utilizing ESUs. Note that the *x*-axis scale shifts to progressively greater increments above 0.02.

(B) Cumulative error based on false positives plus false negatives for each threshold value. The optimum threshold value is 0.025 (5%), where error is minimized at 33%.

DOI: 10.1371/journal.pbio.0030422.g009

Advances in Testing Species Hypotheses with the Multispecies Coalescent

Syst. Biol. 56(6):887–895, 2007

ISSN: 1063-5157 print / 1076-836X online
DOI: 10.1080/10635150701701091

Delimiting Species without Monophyletic Gene Trees

L. LACEY KNOWLES AND BRYAN C. CARSTENS

Department of Ecology and Evolutionary Biology, Museum of Zoology, 1109 Geddes Avenue, University of Michigan, Ann Arbor, MI 48109-1079, USA; E-mail: knowlesl@umich.edu (L.L.K.)

Abstract.—Genetic data are frequently used to delimit species, where species status is determined on the basis of an exclusivity criterium, such as reciprocal monophyly. Not only are there numerous empirical examples of incongruence between the boundaries inferred from such data compared to other sources like morphology—especially with recently derived species, but population genetic theory also clearly shows that an inevitable bias in species status results because genetic thresholds do not explicitly take into account how the timing of speciation influences patterns of genetic differentiation. This study represents a fundamental shift in how genetic data might be used to delimit species. Rather than equating gene trees with a species tree or basing species status on some genetic threshold, the relationship between the gene trees and the species history is modeled probabilistically. Here we show that the same theory that is used to calculate the probability of reciprocal monophyly can also be used to delimit species despite widespread incomplete lineage sorting. The results from a preliminary simulation study suggest that very recently derived species can be accurately identified long before the requisite time for reciprocal monophyly to be achieved following speciation. The study also indicates the importance of sampling, both with regards to loci and individuals. Withstanding a thorough investigation into the conditions under which the coalescent-based approach will be effective, namely how the timing of divergence relative to the effective population size of species affects accurate species delimitation, the results are nevertheless consistent with other recent studies (aimed at inferring species relationships), showing that despite the lack of monophyletic gene trees, a signal of species divergence persists and can be extracted. Using an explicit model-based approach also avoids two primary problems with species delimitation that result when genetic thresholds are applied with genetic data—the inherent biases in species detection arising from when and how speciation occurred, and failure to take into account the high stochastic variance of genetic processes. Both the utility and sensitivities of the coalescent-based approach outlined here are discussed; most notably, a model-based approach is essential for determining whether incompletely sorted gene lineages are (or are not) consistent with separate species lineages, and such inferences require accurate model parameterization (i.e., a range of realistic effective population sizes relative to potential times of divergence for the purported species). It is the goal (and motivation of this study) that genetic data might be used effectively as a source of complementation to other sources of data for diagnosing species, as opposed to the exclusion of other evidence for species delimitation, which will require an explicit consideration of the effects of the temporal dynamic of lineage splitting on genetic data. [Coalescence; genealogical discord; genealogical species concept; gene trees; incomplete lineage sorting.]

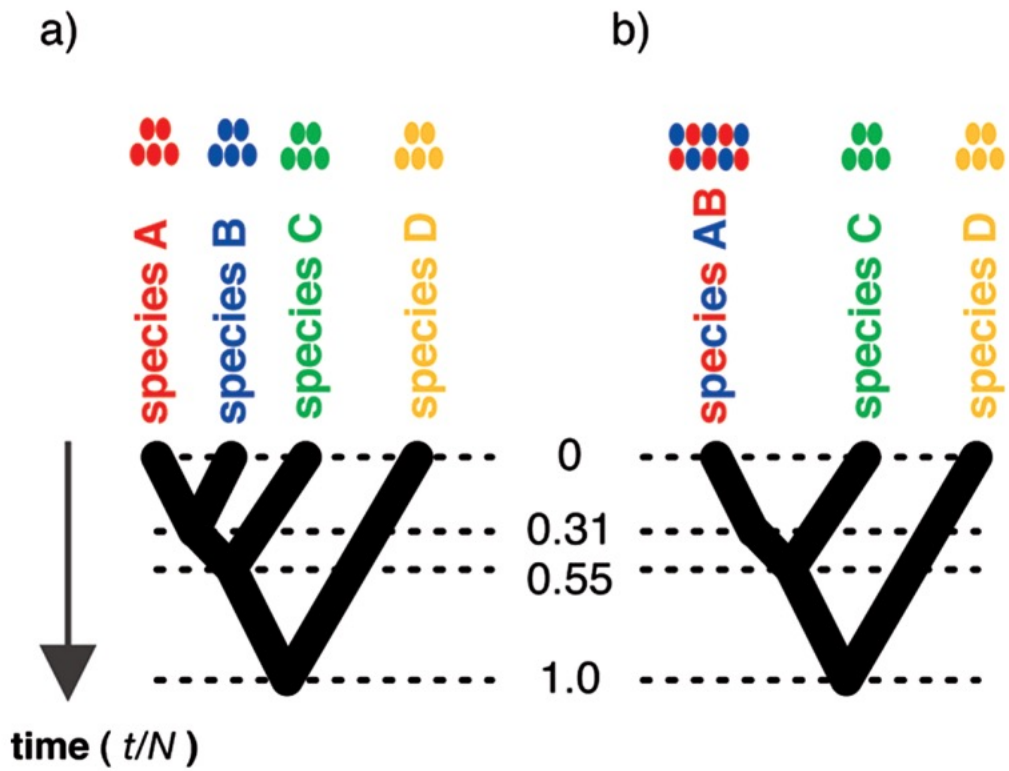


FIGURE 2. Models of the histories used in the simulations to evaluate the coalescent-based approach to species delimitation, where the focus of the study is on whether (a) the history of species divergence of the A and B species lineages can be distinguished from (b) the lack of divergence of the AB lineage.

H_0: Lump A and B into a single species
H_A: Split A and B into separate species (TRUE)

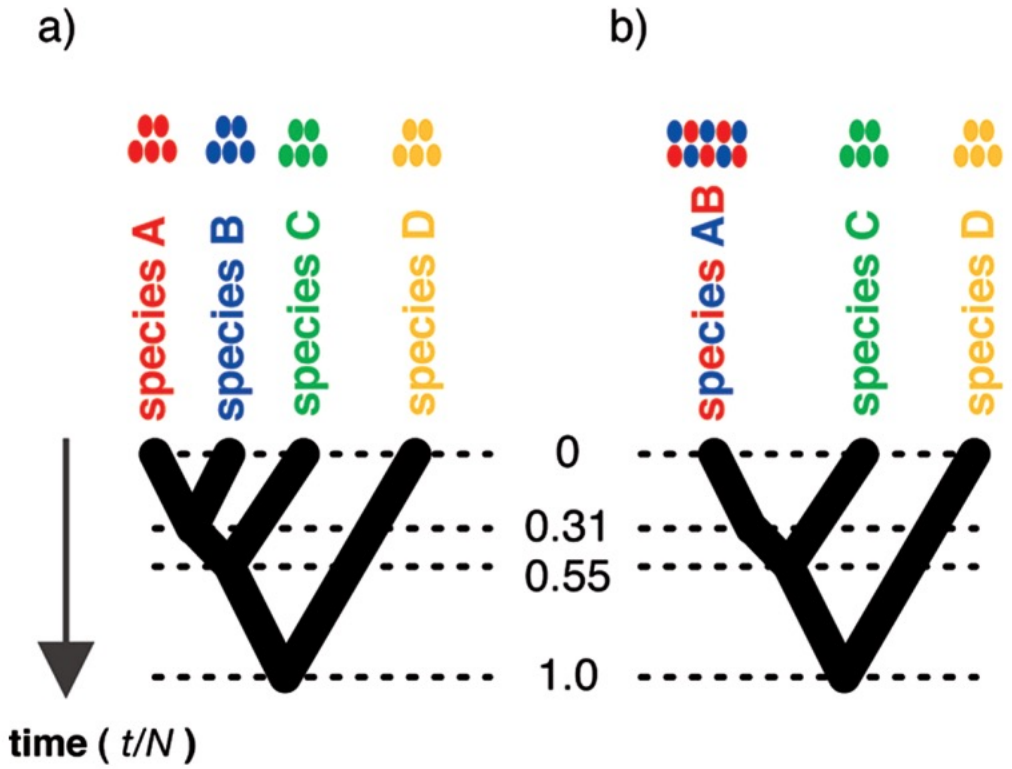


FIGURE 2. Models of the histories used in the simulations to evaluate the coalescent-based approach to species delimitation, where the focus of the study is on whether (a) the history of species divergence of the A and B species lineages can be distinguished from (b) the lack of divergence of the AB lineage.

H_0: Lump A and B into a single species

H_A: Split A and B into separate species (TRUE)

$$LRT = 2*(\ln L_A - \ln L_0)$$

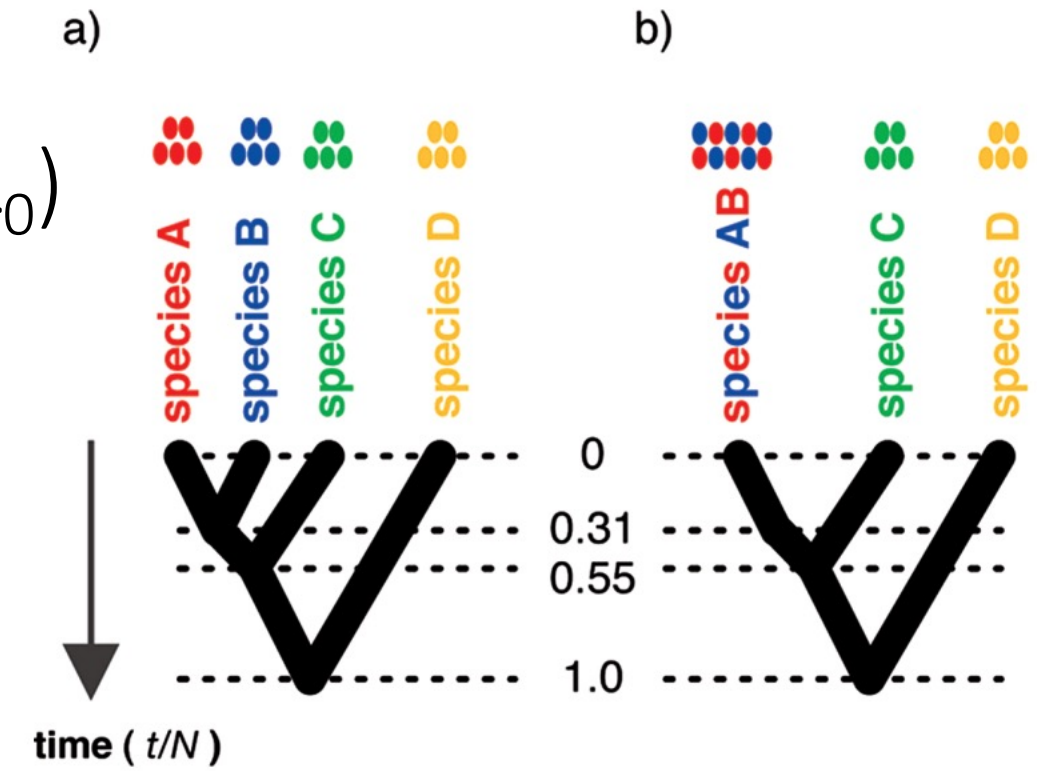


FIGURE 2. Models of the histories used in the simulations to evaluate the coalescent-based approach to species delimitation, where the focus of the study is on whether (a) the history of species divergence of the A and B species lineages can be distinguished from (b) the lack of divergence of the AB lineage.

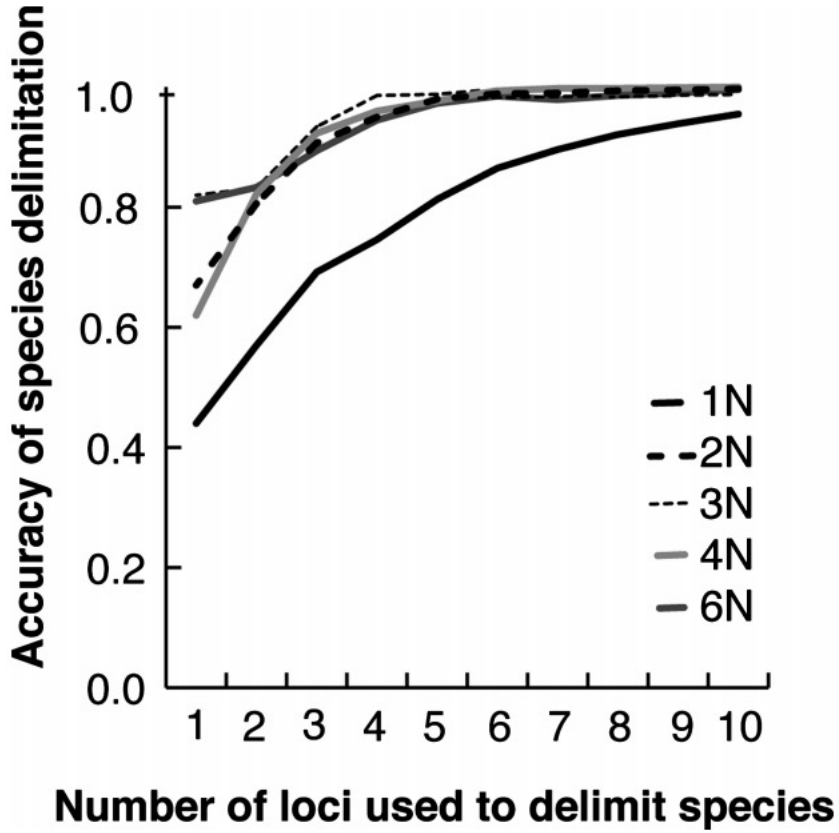


FIGURE 3. Accuracy of the coalescent-based approach for delimiting species A and B with different sampling efforts, showing the false-negative error rate decreases as the number of loci sampled increases from 1 to 10 loci; each line represents a set of simulations for a specific divergence time, ranging from a total species tree depth (see Fig. 2) of 1N to 6N.

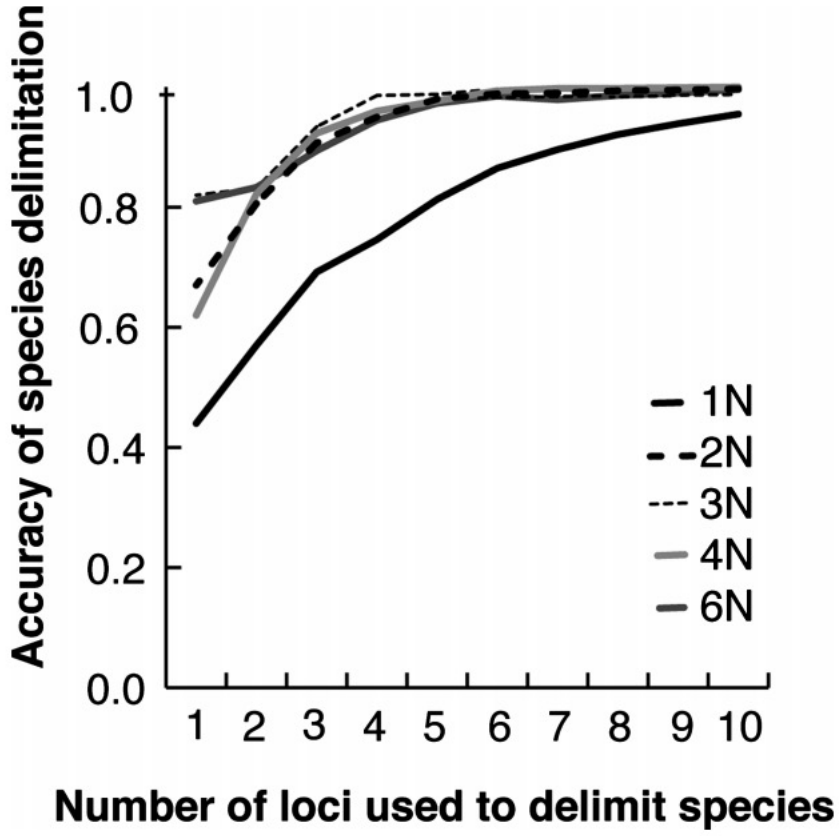
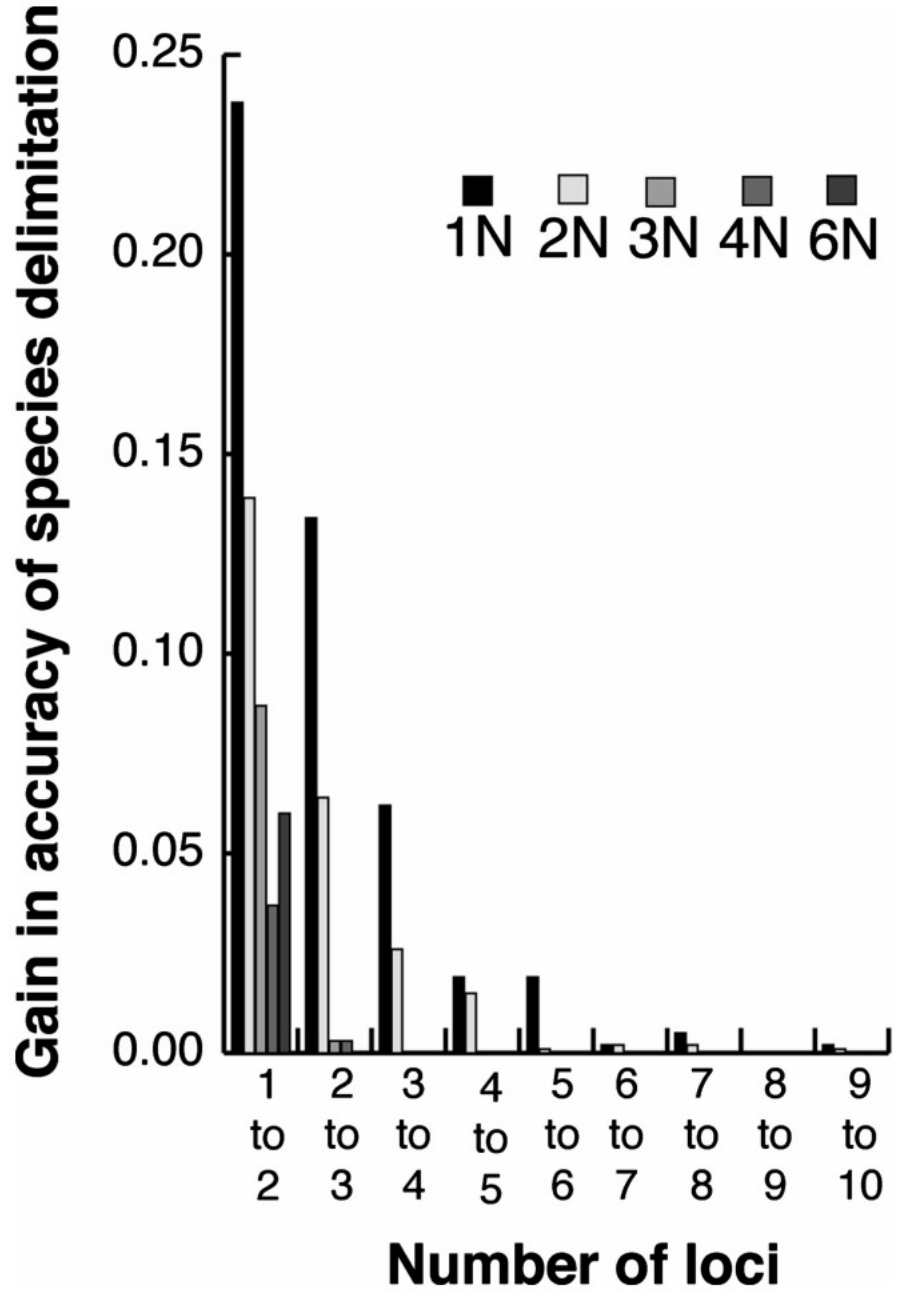


FIGURE 3. Accuracy of the coalescent-based approach for delimiting species A and B with different sampling efforts, showing the false-negative error rate decreases as the number of loci sampled increases from 1 to 10 loci; each line represents a set of simulations for a specific divergence time, ranging from a total species tree depth (see Fig. 2) of 1N to 6N.



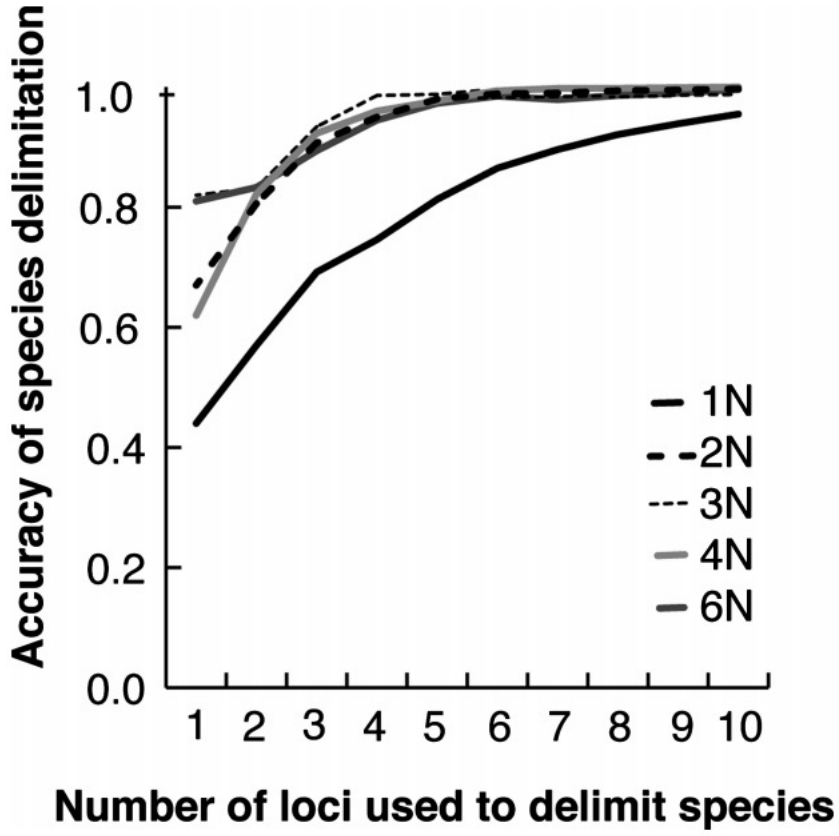
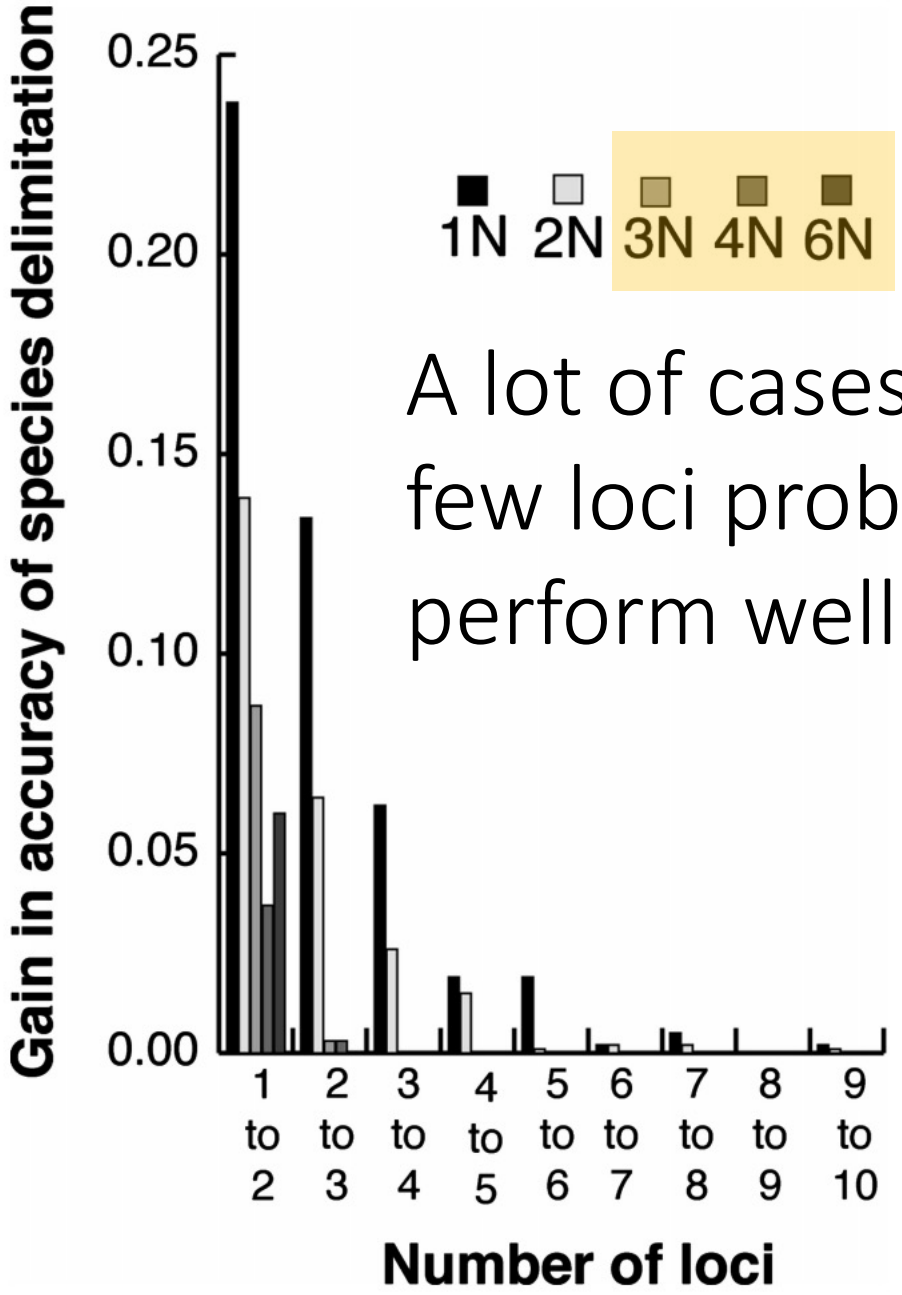


FIGURE 3. Accuracy of the coalescent-based approach for delimiting species A and B with different sampling efforts, showing the false-negative error rate decreases as the number of loci sampled increases from 1 to 10 loci; each line represents a set of simulations for a specific divergence time, ranging from a total species tree depth (see Fig. 2) of 1N to 6N.



A lot of cases where few loci probably perform well

Bayesian species delimitation using multilocus sequence data

Ziheng Yang[a,b] and Bruce Rannala[a,c,1]

[a]Center for Computational and Evolutionary Biology, Institute of Zoology, Chinese Academy of Sciences, Beijing 100101, China; [b]Department of Biology, University College London, London WC1E 6BT, United Kingdom; and [c]Genome Center and Department of Evolution and Ecology, University of California, Davis, CA 95616

Edited by Scott V. Edwards, Harvard University, Cambridge, MA, and accepted by the Editorial Board April 2, 2010 (received for review November 11, 2009)

Bayesian species delimitation using multilocus sequence data

Ziheng Yang[a,b] and Bruce Rannala[a,c,1]

[a]Center for Computational and Evolutionary Biology, Institute of Zoology, Chinese Academy of Sciences, Beijing 100101, China; [b]Department of Biology, University College London, London WC1E 6BT, United Kingdom; and [c]Genome Center and Department of Evolution and Ecology, University of California, Davis, CA 95616

Edited by Scott V. Edwards, Harvard University, Cambridge, MA, and accepted by the Editorial Board April 2, 2010 (received for review November 11, 2009)

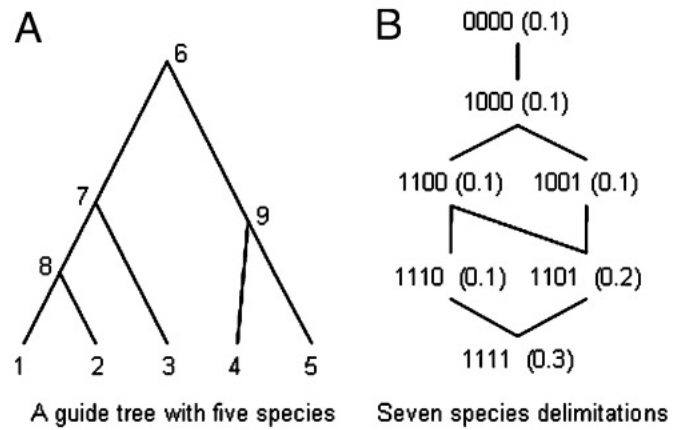


Fig. 1. Given the guide species tree (*A*), each species delimitation is represented by a set of flags indicating whether each of the four ancestral nodes (6, 7, 8, 9) is collapsed (0) or resolved (1). For this guide tree, there exist seven species delimitations, shown in *B*, where 0000 indicates all nodes are collapsed so that there is only one species, and 1111 indicates the fully resolved tree with five species. The reversible-jump algorithm allows moves between species delimitations connected in *B*. The probabilities of the species delimitations under the uniform Dirichlet prior with equal probabilities for each labeled history are shown in parentheses. For example, tree 1101 has prior probability 0.2 because this tree corresponds to two labeled histories, with node 9 being older or younger than node 7, respectively (node 8 is collapsed in tree 1101). The prior with equal probabilities for the rooted trees assigns probability 1/7 for each of these species delimitations.

Bayesian species delimitation using multilocus sequence data

Ziheng Yang[a,b] and Bruce Rannala[a,c,1]

[a]Center for Computational and Evolutionary Biology, Institute of Zoology, Chinese Academy of Sciences, Beijing 100101, China; [b]Department of Biology, University College London, London WC1E 6BT, United Kingdom; and [c]Genome Center and Department of Evolution and Ecology, University of California, Davis, CA 95616

Edited by Scott V. Edwards, Harvard University, Cambridge, MA, and accepted by the Editorial Board April 2, 2010 (received for review November 11, 2009)

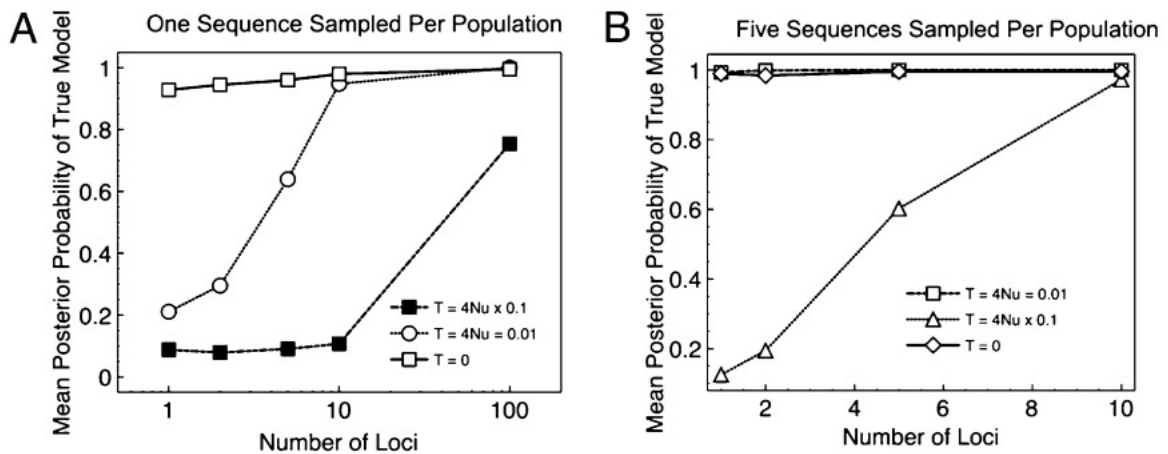


Fig. 2. Mean posterior probability of the correct model across 100 replicate datasets as a function of the number of unlinked loci. The sequence at each locus was 1 kb in length. In all cases $\theta = 0.01$. The divergence time $\tau = 0$ corresponds to a single species, whereas $\tau = \theta$ and $\tau = \theta/10$ correspond to two species with ancient and recent divergence times. In *A*, one sequence was sampled from each of two populations. In *B*, five sequences were sampled from each population.

Evaluation of a Bayesian Coalescent Method of Species Delimitation

CHI ZHANG[1,2,3], DE-XING ZHANG[1,2], TIANQI ZHU[4], AND ZIHENG YANG[1,5,*]

[1]*Center for Computational and Evolutionary Biology, Institute of Zoology, Chinese Academy of Sciences, Beijing 100101, China;* [2]*State Key Laboratory of Integrated Management of Pest Insects and Rodents, Institute of Zoology, Chinese Academy of Sciences, Beijing 100101, China;* [3]*Graduate University of Chinese Academy of Sciences, Beijing 100049, China;* [4]*School of Mathematical Sciences, Peking University, Beijing 100871, China; and* [5]*Department of Biology, University College London, Darwin Building, Gower Street, London WC1E 6BT, UK;*
**Correspondence to be sent to: Department of Biology, University College London, Darwin Building, Gower Street, London WC1E 6BT, UK; E-mail: z.yang@ucl.ac.uk.*

Received 16 November 2010; reviews returned 10 February 2011; accepted 20 April 2011
Associate Editor: Thomas Buckley

Abstract.—A Bayesian coalescent-based method has recently been proposed to delimit species using multilocus genetic sequence data. Posterior probabilities of different species delimitation models are calculated using reversible-jump Markov chain Monte Carlo algorithms. The method accounts for species phylogenies and coalescent events in both extant and extinct species and accommodates lineage sorting and uncertainties in the gene trees. Although the method is theoretically appealing, its utility in practical data analysis is yet to be rigorously examined. In particular, the analysis may be sensitive to priors on ancestral population sizes and on species divergence times and to gene flow between species. Here we conduct a computer simulation to evaluate the statistical performance of the method, such as the false negatives (the error of lumping multiple species into one) and false positives (the error of splitting one species into several). We found that the correct species model was inferred with high posterior probability with only one or two loci when 5 or 10 sequences were sampled from each population, or with 50 loci when only one sequence was sampled. We also simulated data allowing migration under a two-species model, a mainland-island model and a stepping-stone model to assess the impact of gene flow (hybridization or introgression). The behavior of the method was diametrically different depending on the migration rate. Low rates at < 0.1 migrants per generation had virtually no effect, so that the method, while assuming no hybridization between species, identified distinct species despite small amounts of gene flow. This behavior appears to be consistent with biologists' practice. In contrast, higher migration rates at $\geqslant 10$ migrants per generation caused the method to infer one species. At intermediate levels of migration, the method is indecisive. Our results suggest that Bayesian analysis under the multispecies coalescent model may provide important insights into population divergences, and may be useful for generating hypotheses of species delimitation, to be assessed with independent information from anatomical, behavioral, and ecological data. [Species delimitation; coalescent; Bayesian inference; simulation; stepping-stone model; Lindley's paradox.]

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

Abstract

A method was developed for simultaneous Bayesian inference of species delimitation and species phylogeny using the multispecies coalescent model. The method eliminates the need for a user-specified guide tree in species delimitation and incorporates phylogenetic uncertainty in a Bayesian framework. The nearest-neighbor interchange algorithm was adapted to propose changes to the species tree, with the gene trees for multiple loci altered in the proposal to avoid conflicts with the newly proposed species tree. We also modify our previous scheme for specifying priors for species delimitation models to construct joint priors for models of species delimitation and species phylogeny. As in our earlier method, the modified algorithm integrates over gene trees, taking account of the uncertainty of gene tree topology and branch lengths given the sequence data. We conducted a simulation study to examine the statistical properties of the method using six populations (two sequences each) and a true number of three species, with values of divergence times and ancestral population sizes that are realistic for recently diverged species. The results suggest that the method tends to be conservative with high posterior probabilities being a confident indicator of species status. Simulation results also indicate that the power of the method to delimit species increases with an increase of the divergence times in the species tree, and with an increased number of gene loci. Reanalyses of two data sets of cavefish and coast horned lizards suggest considerable phylogenetic uncertainty even though the data are informative about species delimitation. We discuss the impact of the prior on models of species delimitation and species phylogeny and of the prior on population size parameters (θ) on Bayesian species delimitation.

Key words: **Bayesian species delimitation, species tree, multispecies coalescent, reversible-jump MCMC, guide tree, nearest-neighbor interchange.**

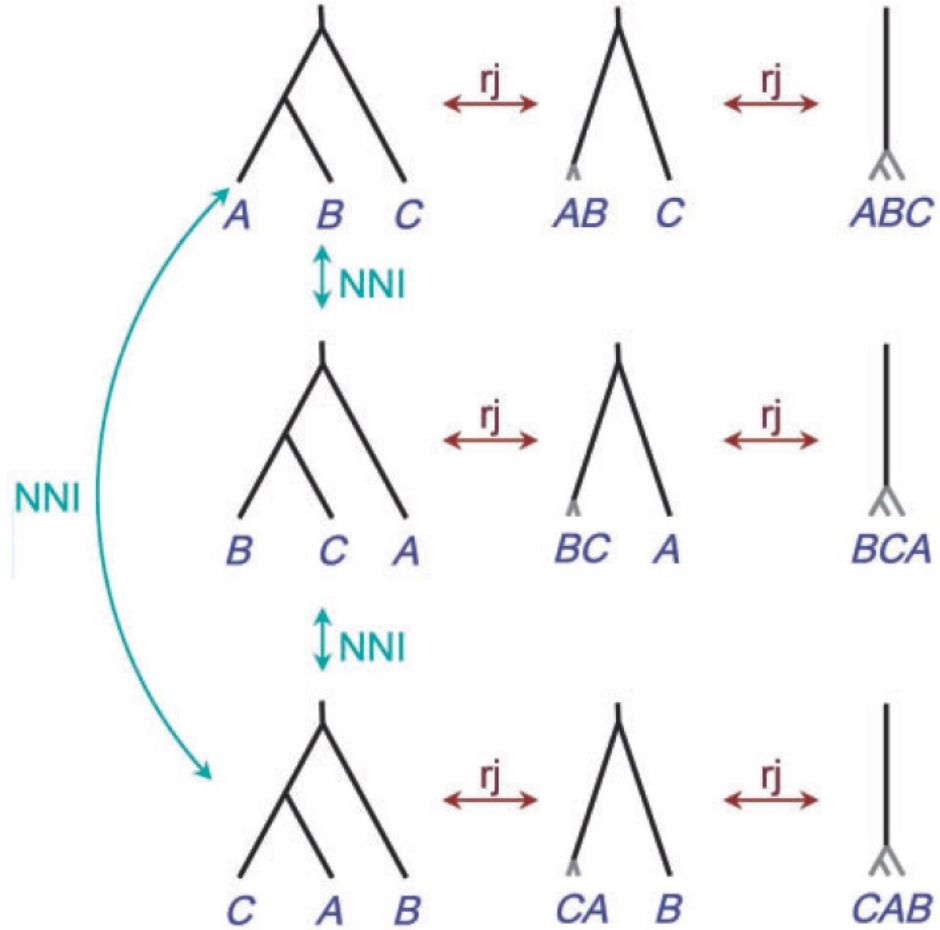


FIG. 3. Models of species delimitation and species phylogeny for three populations *A*, *B*, and *C*. Models on the same row correspond to different species delimitation models given the same guide tree, formed by collapsing internal nodes on the guide tree (represented by short gray branches). The one-species model is represented three times, and there are nine models in our MCMC algorithm even though there are only seven biologically distinct models. The two priors constructed in this article assign equal probabilities ($\frac{1}{9}$) to the nine models. An NNI algorithm is used to move between the guide trees, whereas rjMCMC is used to move between species-delimitation models.

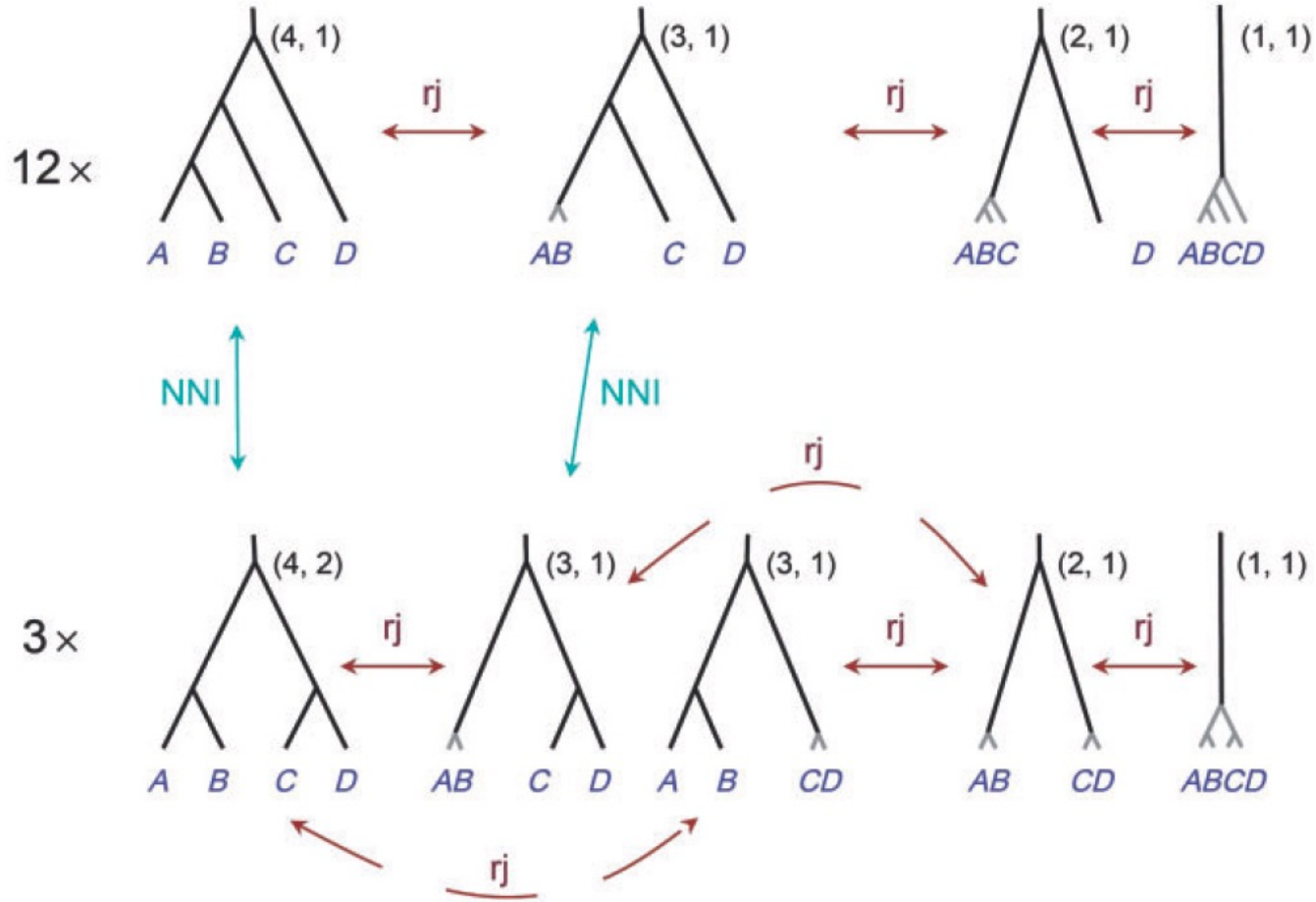


Fig. 4. The models of species delimitation and species phylogenies for four populations *A*–*D*. There should be 15 rows, but only two rows are shown here, to represent the two guide tree shapes. On the same row are the species delimitation models generated by collapsing internal nodes on the same guide tree. The pair of numbers next to each model is the number of species and the number of labeled histories for the species tree. rjMCMC moves between different species-delimitation models are shown, but most of the NNI moves changing species phylogenies are not shown here.

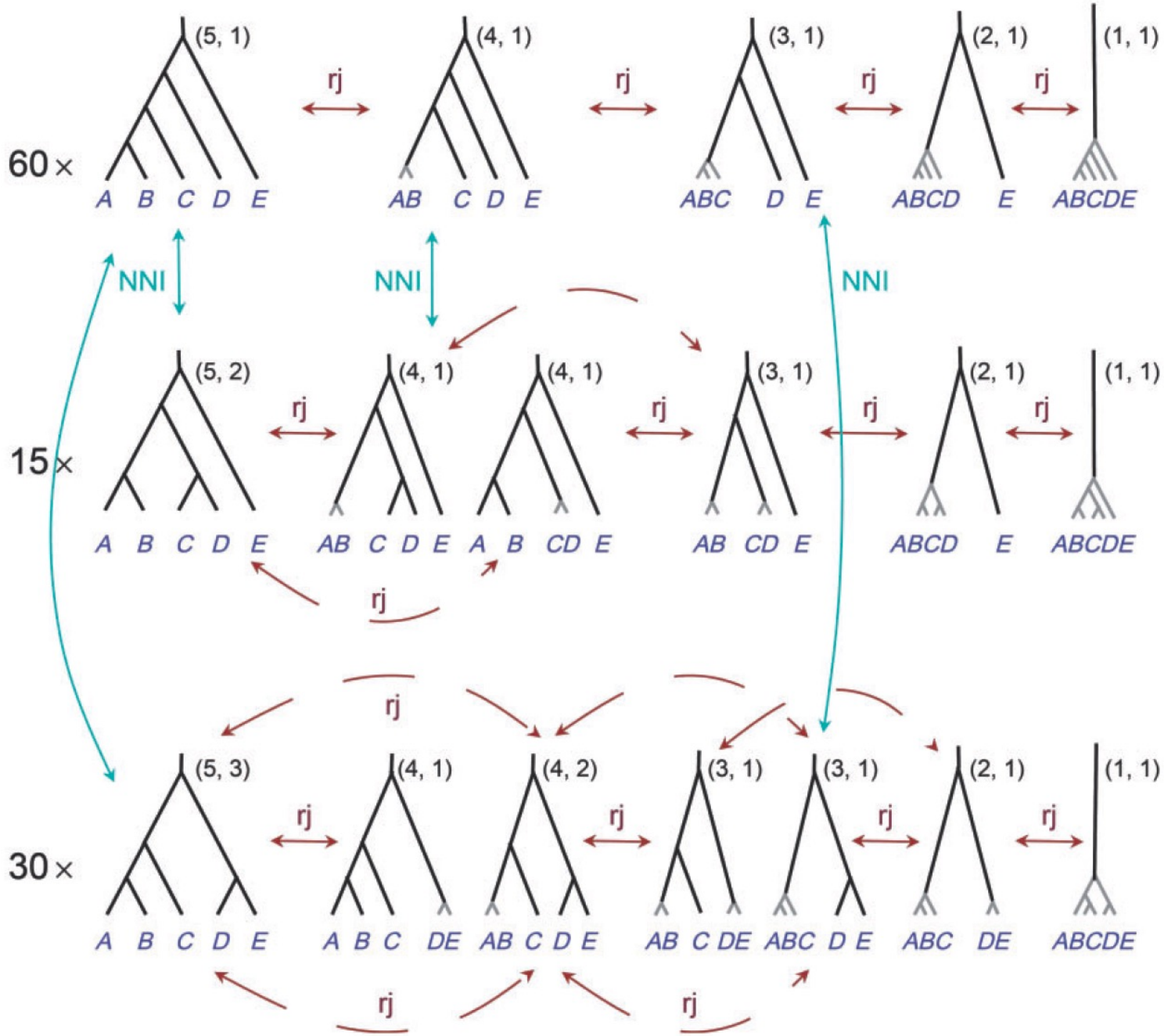


Fig. 5. The models of species delimitation and species phylogeny for five populations *A*–*E*. There should be 105 rows but only three are shown here, to represent the three different guide tree shapes. See legends to figures 3 and 4.

Table 2. Average MAP Probability of Model versus Percent Correct.

Number of Loci	$\theta = 0.005, \tau_0 = 5 \times \theta$		$\theta = 0.005, \tau_0 = 1.25 \times \theta$		$\theta = 0.001, \tau_0 = 1.25 \times \theta$	
	Prob	% Correct	Prob	% Correct	Prob	% Correct
1	0.53	0.64	0.28	0.30	0.17	0.06
2	0.77	0.92	0.40	0.52	0.24	0.32
5	0.83	0.88	0.53	0.76	0.34	0.42
10	0.89	1.0	0.61	0.70	0.39	0.52
20	0.93	1.0	0.78	0.94	0.57	0.78

NOTE.—Prob is the average probability of the MAP model over all the simulated data sets for each specific combination of simulation parameters and % correct is the proportion of these data sets for which the delimitation and phylogeny both matched the true model used in the simulation (i.e., the MAP model is the true model).

Table 2. Average MAP Probability of Model versus Percent Correct.

Number of Loci	$\theta = 0.005, \tau_0 = 5 \times \theta$		$\theta = 0.005, \tau_0 = 1.25 \times \theta$		$\theta = 0.001, \tau_0 = 1.25 \times \theta$	
	Prob	% Correct	Prob	% Correct	Prob	% Correct
1	0.53	0.64	0.28	0.30	0.17	0.06
2	0.77	0.92	0.40	0.52	0.24	0.32
5	0.83	0.88	0.53	0.76	0.34	0.42
10	0.89	1.0	0.61	0.70	0.39	0.52
20	0.93	1.0	0.78	0.94	0.57	0.78

NOTE.—Prob is the average probability of the MAP model over all the simulated data sets for each specific combination of simulation parameters and % correct is the proportion of these data sets for which the delimitation and phylogeny both matched the true model used in the simulation (i.e., the MAP model is the true model).

Syst. Biol. 67(2):269–284, 2018

DOI:10.1093/sysbio/syx073
Advance Access publication September 4, 2017

Impact of Model Violations on the Inference of Species Boundaries Under the Multispecies Coalescent

ANTHONY J. BARLEY[1,*], JEREMY M. BROWN[2], AND ROBERT C. THOMSON[1]

[1]*Department of Biology, University of Hawai'i, 2538 McCarthy Mall, Edmondson Hall 216, Honolulu, HI 96822, USA;* [2]*Department of Biological Sciences and Museum of Natural Science, Louisiana State University, 202 Life Sciences Building, Baton Rouge, LA 70803, USA*
**Correspondence to be sent to: Department of Biology, University of Hawaii, 2538 McCarthy Mall, Edmondson Hall 216, Honolulu, HI 96822, USA; E-mail: ajbarley@hawaii.edu.*

Received 15 March 2017; reviews returned 21 June 2017; accepted 31 August 2017
Associate Editor: Laura Kubatko

Abstract.—The use of genetic data for identifying species-level lineages across the tree of life has received increasing attention in the field of systematics over the past decade. The multispecies coalescent model provides a framework for understanding the process of lineage divergence and has become widely adopted for delimiting species. However, because these studies lack an explicit assessment of model fit, in many cases, the accuracy of the inferred species boundaries are unknown. This is concerning given the large amount of empirical data and theory that highlight the complexity of the speciation process. Here, we seek to fill this gap by using simulation to characterize the sensitivity of inference under the multispecies coalescent (MSC) to several violations of model assumptions thought to be common in empirical data. We also assess the fit of the MSC model to empirical data in the context of species delimitation. Our results show substantial variation in model fit across data sets. Posterior predictive tests find the poorest model performance in data sets that were hypothesized to be impacted by model violations. We also show that while the inferences assuming the MSC are robust to minor model violations, such inferences can be biased under some biologically plausible scenarios. Taken together, these results suggest that researchers can identify individual data sets in which species delimitation under the MSC is likely to be problematic, thereby highlighting the cases where additional lines of evidence to identify species boundaries are particularly important to collect. Our study supports a growing body of work highlighting the importance of model checking in phylogenetics, and the usefulness of tailoring tests of model fit to assess the reliability of particular inferences. [Populations structure, gene flow, demographic changes, posterior prediction, simulation, genetics.]

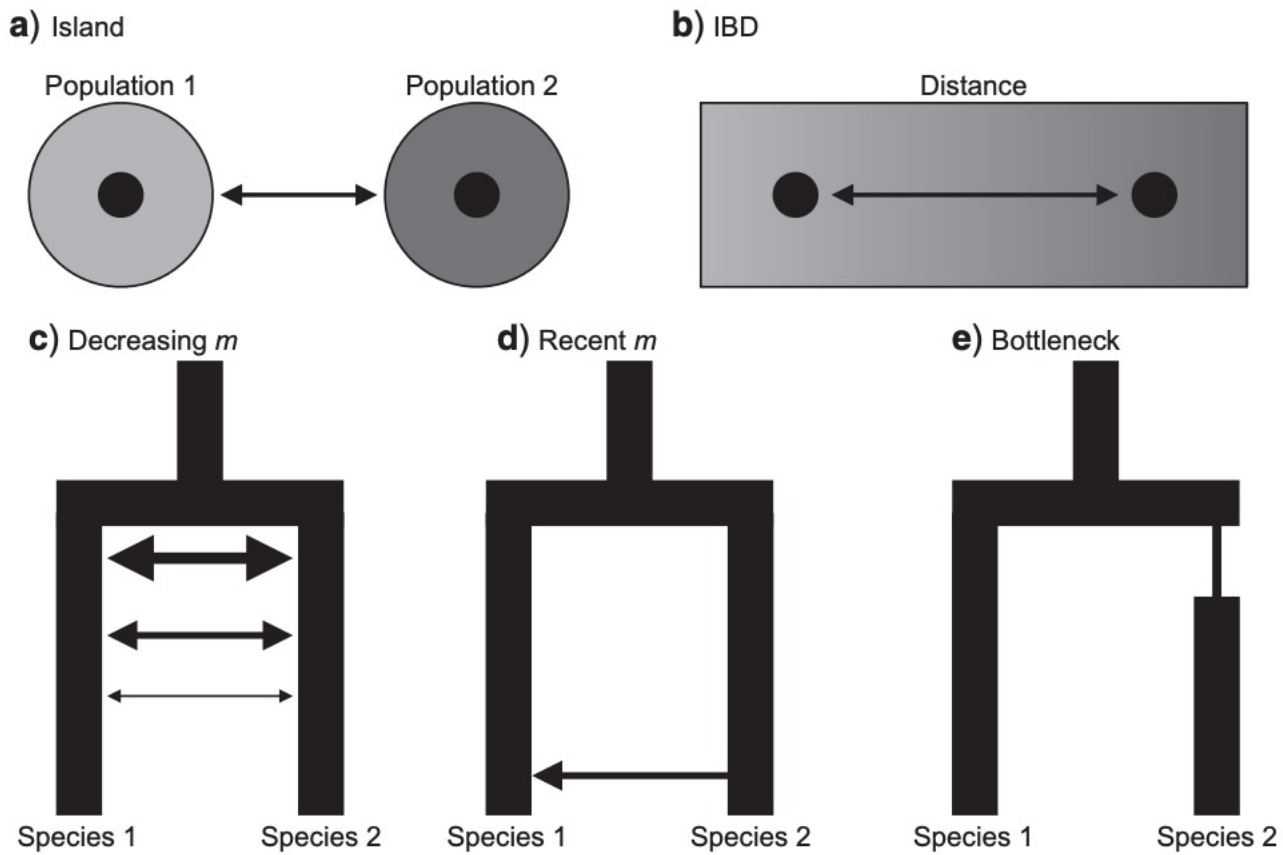


FIGURE 1. Illustration of the simulation scenarios used to test for sensitivity of inference under the MSC to model violations; arrows indicate migration. a) An island model of population structure in which two populations are connected by migration. b) IBD in a single, continuous population sampled at two disjunct geographic locations. c) Species divergence with gene flow that decreases through time. d) Recent migration (or secondary contact) between two species following a period of isolation. e) A population bottleneck experienced by one species following lineage separation.

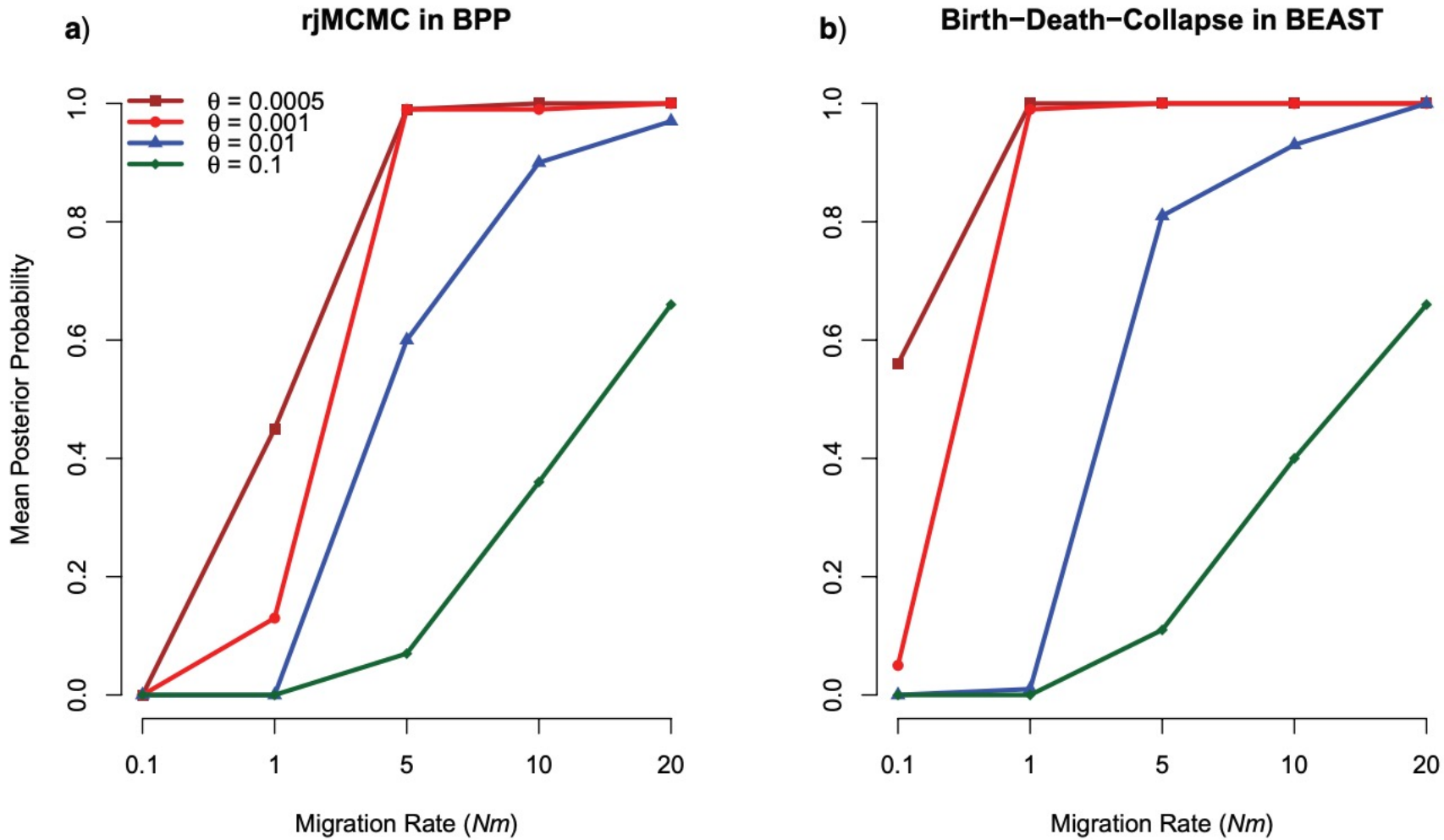


FIGURE 2. Results of the population structure simulation analyses. Data were simulated for two populations connected by varying levels of symmetrical migration and species delimitation was performed using a) BPP and b) STACEY. Each point represents the mean posterior probability for 1 species model across 100 replicates under a wide range of parameter values for θ per site and the migration rate (expressed as *Nm*).

Comparison of Methods for Molecular Species Delimitation Across a Range of Speciation Scenarios

ARONG LUO[1,2,*], CHENG LING[3], SIMON Y. W. HO[2], AND CHAO-DONG ZHU[1,4]

[1]*Key Laboratory of Zoological Systematics and Evolution, Institute of Zoology, Chinese Academy of Sciences, Beijing 100101, China;* [2]*School of Life and Environmental Sciences, University of Sydney, Sydney, New South Wales 2006, Australia;* [3]*Department of Computer Science and Technology, College of Information Science and Technology, Beijing University of Chemical Technology, Beijing 100029, China; and* [4]*College of Life Sciences, University of Chinese Academy of Sciences, Beijing 100049, China*

[*]*Correspondence to be sent to: Key Laboratory of Zoological Systematics and Evolution, Institute of Zoology, Chinese Academy of Sciences, Beijing 100101, China;*
E-mail: luoar@ioz.ac.cn.

Simon Y. W. Ho and Chao-Dong Zhu contributed equally to this article.

Received 25 June 2017; reviews returned 31 August 2017; accepted 10 February 2018
Associate Editor: Rachel Mueller

Abstract.—Species are fundamental units in biological research and can be defined on the basis of various operational criteria. There has been growing use of molecular approaches for species delimitation. Among the most widely used methods, the generalized mixed Yule-coalescent (GMYC) and Poisson tree processes (PTP) were designed for the analysis of single-locus data but are often applied to concatenations of multilocus data. In contrast, the Bayesian multispecies coalescent approach in the software Bayesian Phylogenetics and Phylogeography (BPP) explicitly models the evolution of multilocus data. In this study, we compare the performance of GMYC, PTP, and BPP using synthetic data generated by simulation under various speciation scenarios. We show that in the absence of gene flow, the main factor influencing the performance of these methods is the ratio of population size to divergence time, while number of loci and sample size per species have smaller effects. Given appropriate priors and correct guide trees, BPP shows lower rates of species overestimation and underestimation, and is generally robust to various potential confounding factors except high levels of gene flow. The single-threshold GMYC and the best strategy that we identified in PTP generally perform well for scenarios involving more than a single putative species when gene flow is absent, but PTP outperforms GMYC when fewer species are involved. Both methods are more sensitive than BPP to the effects of gene flow and potential confounding factors. Case studies of bears and bees further validate some of the findings from our simulation study, and reveal the importance of using an informed starting point for molecular species delimitation. Our results highlight the key factors affecting the performance of molecular species delimitation, with potential benefits for using these methods within an integrative taxonomic framework. [Molecular species delimitation; speciation; multispecies coalescent; simulation; generalized mixed Yule-coalescent; Poisson tree processes; Bayesian phylogenetics.]

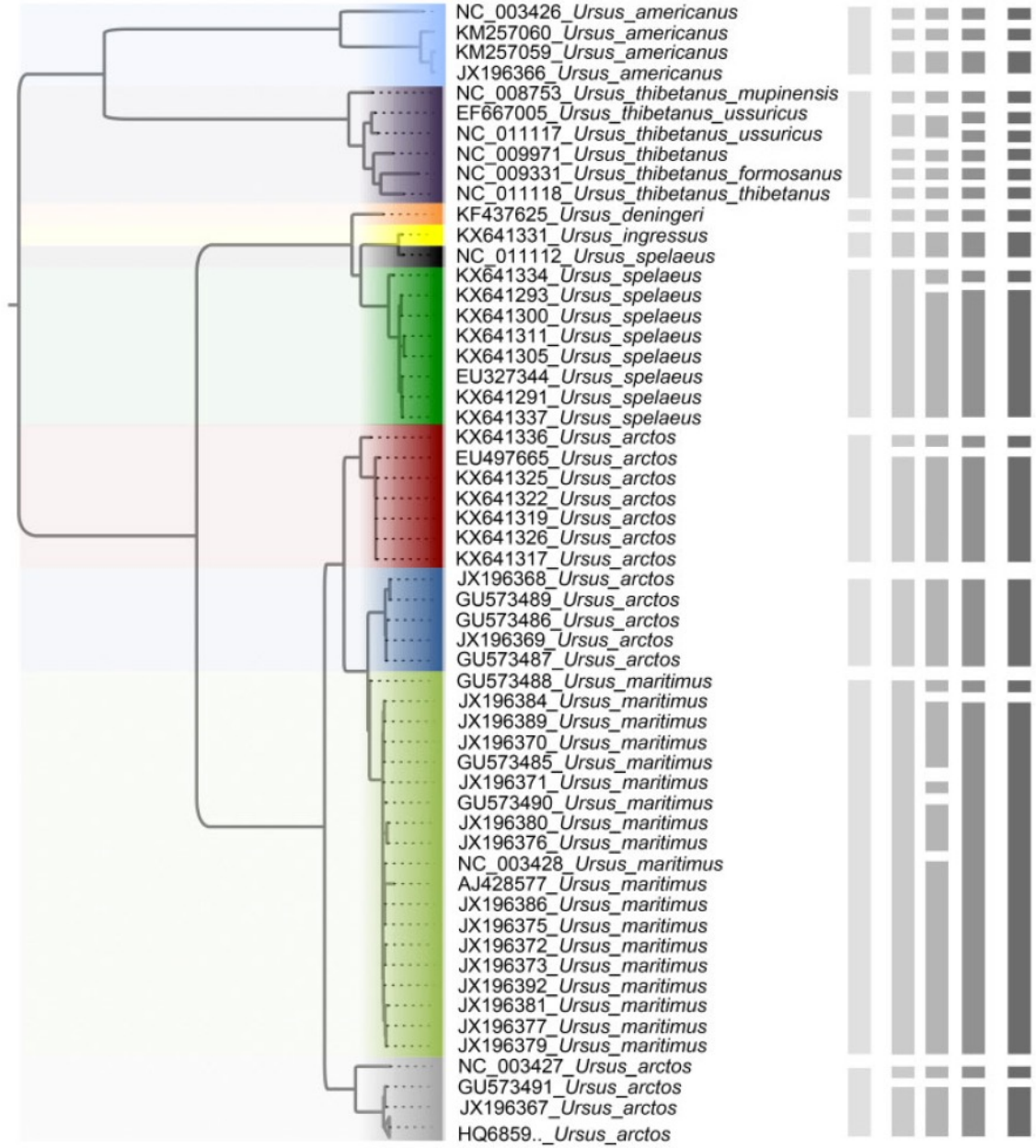


FIGURE 7. Species delimitations estimated for a dataset comprising 89 sequences from bears (genus *Ursus*). The maximum-likelihood tree is shown on the left. The vertical bars, from left to right, indicate the OTUs inferred by BPP, bPTP-ML, bPTP-h, PTP-h, and GMYC, respectively. Clades (of different colors in online version) in the tree indicate the 10 taxa in the guide tree for BPP delimitation, and a collapsed clade at the bottom with the label "HQ6859.._ *Ursus_ arctos*" represents 34 sequences of *Ursus arctos* with accession numbers beginning with "HQ6859" (Supplementary Appendix S4 available on Dryad).

Bayesian species identification under the multispecies coalescent provides significant improvements to DNA barcoding analyses

ZIHENG YANG*† and BRUCE RANNALA†‡

*Department of Genetics, Evolution and Environment, University College London, Gower Street, London WC1E 6BT, UK, †College of Life Sciences, Beijing Normal University, Beijing 100875, China, ‡Department of Evolution and Ecology, University of California at Davis, One Shields Avenue, Davis, CA 95616, USA

Abstract

DNA barcoding methods use a single locus (usually the mitochondrial COI gene) to assign unidentified specimens to known species in a library based on a genetic distance threshold that distinguishes between-species divergence from within-species diversity. Recently developed species delimitation methods based on the multispecies coalescent (MSC) model offer an alternative approach to individual assignment using either single-locus or multiloci sequence data. Here, we use simulations to demonstrate three features of an MSC method implemented in the program BPP. First, we show that with one locus, MSC can accurately assign individuals to species without the need for arbitrarily determined distance thresholds (as required for barcoding methods). We provide an example in which no single threshold or barcoding gap exists that can be used to assign all specimens without incurring high error rates. Second, we show that BPP can identify cryptic species that may be misidentified as a single species within the library, potentially improving the accuracy of barcoding libraries. Third, we show that taxon rarity does not present any particular problems for species assignments using BPP and that accurate assignments can be achieved even when only one or a few loci are available. Thus, concerns that have been raised that MSC methods may have problems analysing rare taxa (singletons) are unfounded. Currently, barcoding methods enjoy a huge computational advantage over MSC methods and may be the only approach feasible for massively large data sets, but MSC methods may offer a more stringent test for species that are tentatively assigned by barcoding.

Keywords: BPP, coalescent, DNA barcoding, species delimitation, species identification

Multispecies coalescent delimits structure, not species

Jeet Sukumaran[a,1,2] and L. Lacey Knowles[a,1]

[a]Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor MI 48109-1079

Edited by David M. Hillis, University of Texas at Austin, Austin, TX, and approved December 29, 2016 (received for review May 23, 2016)

A number of valid criticisms can be made about using the MSC to delimit species

Multispecies coalescent delimits structure, not species

Jeet Sukumaran[a,1,2] and L. Lacey Knowles[a,1]

[a]Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor MI 48109-1079

Edited by David M. Hillis, University of Texas at Austin, Austin, TX, and approved December 29, 2016 (received for review May 23, 2016)

A number of valid criticisms can be made about using the MSC to delimit species

Although, the MSC model performs as it should and in an expected way

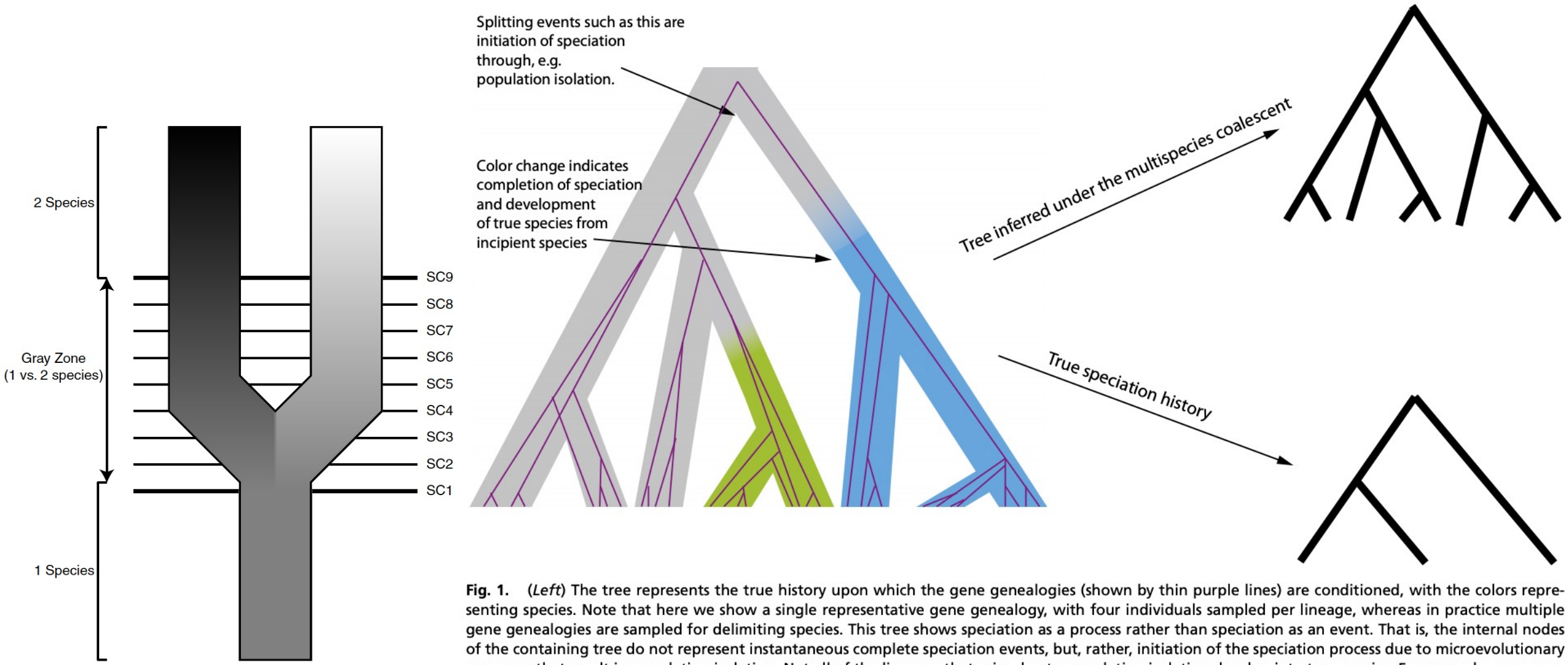


Fig. 1. (*Left*) The tree represents the true history upon which the gene genealogies (shown by thin purple lines) are conditioned, with the colors representing species. Note that here we show a single representative gene genealogy, with four individuals sampled per lineage, whereas in practice multiple gene genealogies are sampled for delimiting species. This tree shows speciation as a process rather than speciation as an event. That is, the internal nodes of the containing tree do not represent instantaneous complete speciation events, but, rather, initiation of the speciation process due to microevolutionary processes that result in population isolation. Not all of the lineages that arise due to population isolation develop into true species. For example, some may merge back into the other lineages of the same species in the future if whatever barriers led to their isolation were to be removed (i.e., the evolutionary independence will be ephemeral because two populations belonging to the same species do not have an impediment to reproduction). Some of these isolated incipient species lineages, however, do stochastically develop into true species (indicated by shift in color), so that they will remain distinct lineages with independent evolutionary trajectories that do not merge back into their parental species, even if the isolation barriers were to disappear. (*Upper Right*) The tree shows the results of inference under the multispecies coalescent using BPP, which includes the structuring both due to species boundaries as well as due to lineage splitting as a result of population isolation. As such, the inference corresponds to the full structural history, but not the true speciation history, which is shown by the tree in *Lower Right*. That is, the multispecies coalescent does not distinguish between structuring due to population isolation vs. structuring due to speciation: It only identifies genetic structure.

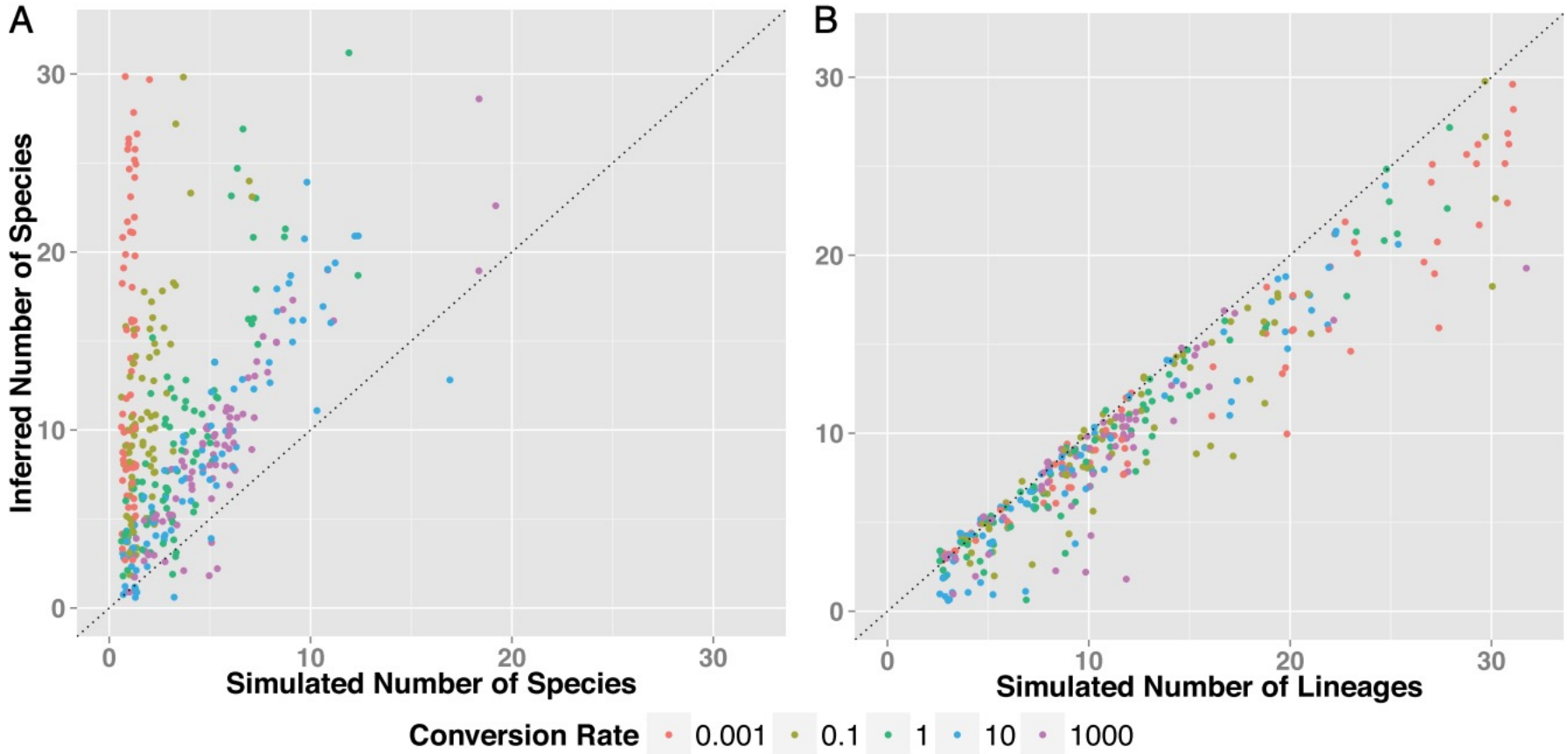


Fig. 2. The performance of species delimitation under the multispecies coalescent when the data are generated under the protracted speciation model, from simulations run for a fixed duration of time (5 units) under different species conversion rates. Speciation initiation rate was fixed at 0.5, while extinction rates were either 0.0 and 0.2 (the plot does not distinguish between these different extinction rates, because these had no meaningful effect on the main results or our argument). (*A*) Shown is the number of species per replicate inferred at a 0.95 probability vs. the number of true species on the input tree. (*B*) Shown is the number of species per replicate vs. the number of lineages (i.e., both true species as well as lineages representing incipient species or population structure). Generally, across all conversion rates, BPP tends to overestimate the number of species. However, what is striking is that BPP does not track species, as seen in *A*, but, rather, tracks structure of any sort, whether incipient species or true species, as seen in *B*.

Bayes Factor Delimitation

$$BF_{10} = \frac{p(D|M_1)}{p(D|M_0)}$$

$p(D|M_1)$ = Marginal Probability of the Data under Model 1

$p(D|M_0)$ = Marginal Probability of the Data under Model 0

$$BF_{10} = \frac{p(D|M_1)}{p(D|M_0)}$$

$$p(D|M_1) = \int p(D|\theta, M_1)p(\theta)\, d\theta$$

$$BF_{10} = \frac{p(D|M_1)}{p(D|M_0)}$$

$$p(D|M_1) = \int p(D|\theta, M_1)p(\theta)\, d\theta$$

Posterior

Prior

Syst. Biol. 60(2):150–160, 2011

DOI:10.1093/sysbio/syq085
Advance Access publication on December 27, 2010

Improving Marginal Likelihood Estimation for Bayesian Phylogenetic Model Selection

WANGANG XIE[1], PAUL O. LEWIS[2,*], YU FAN[2], LYNN KUO[3] AND MING-HUI CHEN[3]

[1]*Abbott, 100 Abbott Park, R436/AP9A-2, Abbott Park, IL 60064, USA;*
[2]*Department of Ecology and Evolutionary Biology, University of Connecticut, 75 North Eagleville Road, Unit 3043, Storrs, CT 06269, USA; and*
[3]*Department of Statistics, University of Connecticut, 215 Glenbrook Road, Unit 4120, Storrs, CT 06269, USA;*
**Correspondence to be sent to: Paul O. Lewis, Department of Ecology and Evolutionary Biology, University of Connecticut, 75 North Eagleville Road, Unit 3043, Storrs, CT 06269, USA; E-mail: paul.lewis@uconn.edu.*

Received 9 February 2009; reviews returned 24 June 2009; accepted 20 September 2010
Associate Editor: Marc Suchard

Abstract.—The marginal likelihood is commonly used for comparing different evolutionary models in Bayesian phylogenetics and is the central quantity used in computing Bayes Factors for comparing model fit. A popular method for estimating marginal likelihoods, the harmonic mean (HM) method, can be easily computed from the output of a Markov chain Monte Carlo analysis but often greatly overestimates the marginal likelihood. The thermodynamic integration (TI) method is much more accurate than the HM method but requires more computation. In this paper, we introduce a new method, steppingstone sampling (SS), which uses importance sampling to estimate each ratio in a series (the "stepping stones") bridging the posterior and prior distributions. We compare the performance of the SS approach to the TI and HM methods in simulation and using real data. We conclude that the greatly increased accuracy of the SS and TI methods argues for their use instead of the HM method, despite the extra computation needed. [Bayes factor; harmonic mean; phylogenetics, marginal likelihood; model selection; path sampling; thermodynamic integration; steppingstone sampling.]

Accurate Model Selection of Relaxed Molecular Clocks in Bayesian Phylogenetics

Guy Baele,*[,1] Wai Lok Sibon Li,[2] Alexei J. Drummond,[3,4,5] Marc A. Suchard,[2,6,7] and Philippe Lemey[1]

[1]Department of Microbiology and Immunology, Rega Institute, KU Leuven, Leuven, Belgium
[2]Department of Human Genetics, David Geffen School of Medicine, University of California, Los Angeles
[3]Bioinformatics Institute, University of Auckland, Auckland, New Zealand
[4]Department of Computer Science, University of Auckland, Auckland, New Zealand
[5]Allan Wilson Centre for Molecular Ecology and Evolution, University of Auckland, Auckland, New Zealand
[6]Department of Biostatistics, School of Public Health, University of California, Los Angeles
[7]Department of Biomathematics, David Geffen School of Medicine, University of California, Los Angeles

***Corresponding author:** E-mail: guy.baele@rega.kuleuven.be.

Associate editor: Barbara Holland

Abstract

Recent implementations of path sampling (PS) and stepping-stone sampling (SS) have been shown to outperform the harmonic mean estimator (HME) and a posterior simulation-based analog of Akaike's information criterion through Markov chain Monte Carlo (AICM), in Bayesian model selection of demographic and molecular clock models. Almost simultaneously, a Bayesian model averaging approach was developed that avoids conditioning on a single model but averages over a set of relaxed clock models. This approach returns estimates of the posterior probability of each clock model through which one can estimate the Bayes factor in favor of the maximum a posteriori (MAP) clock model; however, this Bayes factor estimate may suffer when the posterior probability of the MAP model approaches 1. Here, we compare these two recent developments with the HME, stabilized/smoothed HME (sHME), and AICM, using both synthetic and empirical data. Our comparison shows reassuringly that MAP identification and its Bayes factor provide similar performance to PS and SS and that these approaches considerably outperform HME, sHME, and AICM in selecting the correct underlying clock model. We also illustrate the importance of using proper priors on a large set of empirical data sets.

Key words: model comparison, marginal likelihood, Bayes factors, path sampling, stepping-stone sampling, model averaging, molecular clock, Bayesian inference, phylogeny, BEAST.

Table 1. Model Selection Performance for 100 Simulated Data Sets under either a Balanced or Yule Tree and Two Relaxed Molecular Clock Models Using HME, sHME, AICM, PS, SS, and the MAP Estimated under BMA.

Tree	Clock	HME	sHME	AICM	PS	SS	MAP
Balanced	UCED	92	100	100	94	94	90
Balanced	UCLD	28	5	1	99	99	99
Yule	UCED	92	100	100	99	99	97
Yule	UCLD	11	1	1	61	61	65

NOTE.—The columns report the number of correct classifications obtained out of 100 simulations.

Syst. Biol. 69(2):209–220, 2020

DOI:10.1093/sysbio/syz046
Advance Access publication August 28, 2019

19 Dubious Ways to Compute the Marginal Likelihood of a Phylogenetic Tree Topology

MATHIEU FOURMENT[1], ANDREW F. MAGEE[2], CHRIS WHIDDEN[3], ARMAN BILGE[3], FREDERICK A. MATSEN IV[3], AND VLADIMIR N. MININ[4,*]

[1]*University of Technology Sydney, ithree Institute, Ultimo NSW 2007, Australia;* [2]*Department of Biology, University of Washington, Seattle, WA 98195, USA;* [3]*Fred Hutchinson Cancer Research Center, Seattle, WA 98109, USA; and* [4]*Department of Statistics, University of California, Irvine, CA 92697, USA*
[*]*Correspondence to be sent to: Department of Statistics, University of California, Irvine, CA 92697, USA; E-mail: vminin@uci.edu.*

Frederick A. Matsen and Vladimir N. Minin supervised this research project.

Received 28 November 2018; reviews returned 27 June 2019; accepted 2 July 2019
Associate Editor: David Posada

Abstract.—The marginal likelihood of a model is a key quantity for assessing the evidence provided by the data in support of a model. The marginal likelihood is the normalizing constant for the posterior density, obtained by integrating the product of the likelihood and the prior with respect to model parameters. Thus, the computational burden of computing the marginal likelihood scales with the dimension of the parameter space. In phylogenetics, where we work with tree topologies that are high-dimensional models, standard approaches to computing marginal likelihoods are very slow. Here, we study methods to quickly compute the marginal likelihood of a single fixed tree topology. We benchmark the speed and accuracy of 19 different methods to compute the marginal likelihood of phylogenetic topologies on a suite of real data sets under the JC69 model. These methods include several new ones that we develop explicitly to solve this problem, as well as existing algorithms that we apply to phylogenetic models for the first time. Altogether, our results show that the accuracy of these methods varies widely, and that accuracy does not necessarily correlate with computational burden. Our newly developed methods are orders of magnitude faster than standard approaches, and in some cases, their accuracy rivals the best established estimators. [Bayesian inference; evidence; importance sampling; model selection; variational Bayes.]

Considerations for SNP data - SNAPP

Inferring Species Trees Directly from Biallelic Genetic Markers: Bypassing Gene Trees in a Full Coalescent Analysis

David Bryant,*[,1] Remco Bouckaert,[2] Joseph Felsenstein,[3] Noah A. Rosenberg,[4] and Arindam RoyChoudhury[5]

[1]Department of Mathematics and Statistics and the Allan Wilson Centre for Molecular Ecology and Evolution, University of Otago, Dunedin, New Zealand

[2]Computational Evolution Group, Department of Computer Science, University of Auckland, Auckland, New Zealand

[3]Department of Genome Sciences and Department of Biology, University of Washington

[4]Department of Biology, Stanford University

[5]Department of Biostatistics, Mailman School of Public Health, Columbia University

***Corresponding author:** E-mail: david.bryant@otago.ac.nz.

Associate editor: Rasmus Nielsen

Abstract

The multispecies coalescent provides an elegant theoretical framework for estimating species trees and species demographics from genetic markers. However, practical applications of the multispecies coalescent model are limited by the need to integrate or sample over all gene trees possible for each genetic marker. Here we describe a polynomial-time algorithm that computes the likelihood of a species tree directly from the markers under a finite-sites model of mutation effectively integrating over all possible gene trees. The method applies to independent (unlinked) biallelic markers such as well-spaced single nucleotide polymorphisms, and we have implemented it in SNAPP, a Markov chain Monte Carlo sampler for inferring species trees, divergence dates, and population sizes. We report results from simulation experiments and from an analysis of 1997 amplified fragment length polymorphism loci in 69 individuals sampled from six species of *Ourisia* (New Zealand native foxglove).

Key words: multispecies coalescent, species trees, SNP, AFLP, effective population size, SNAPP.

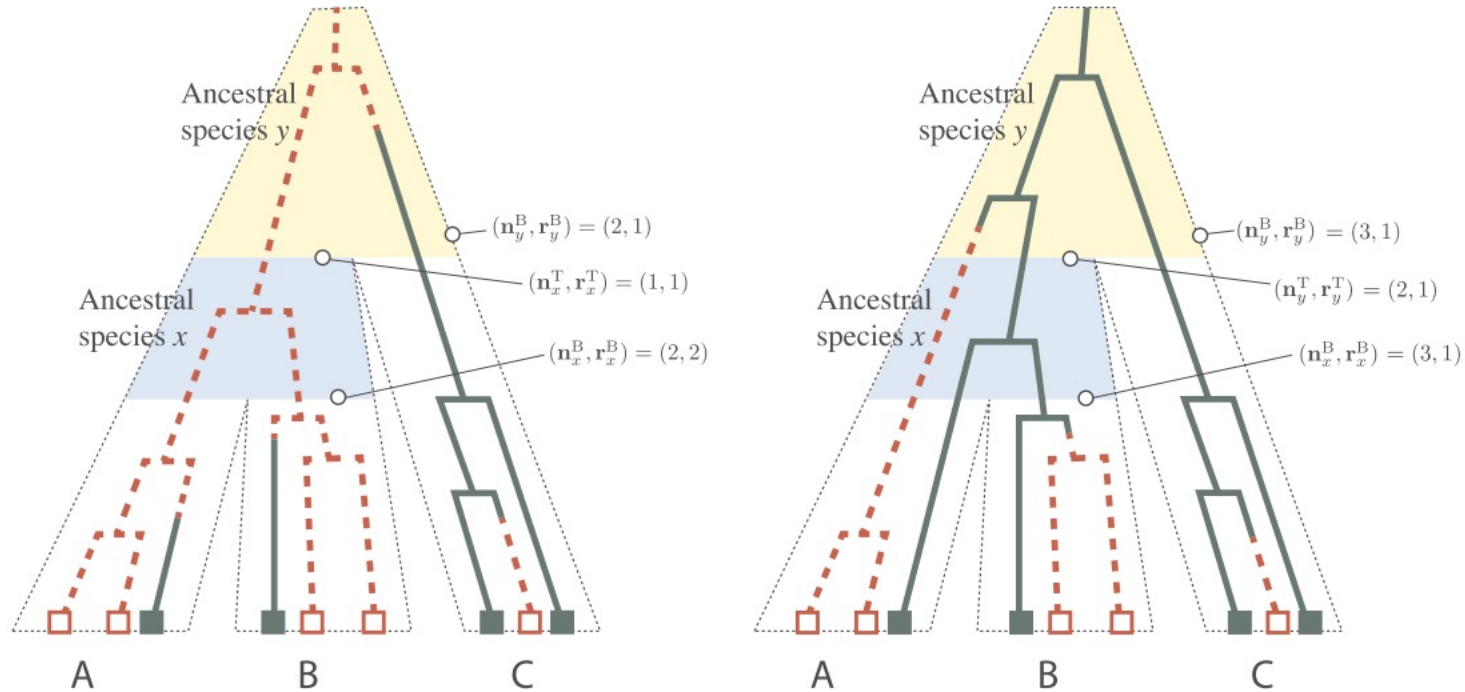


FIG. 1. Gene trees in species trees. Each branch in the species trees corresponds to a species that is either contemporary (A,B,C) or ancestral (x, y). The present-day samples are represented by green (solid) and red (hollow) squares along the lower edge of the tree. The red (dashed) and green (solid) lines trace out two possible gene trees for these individuals, the red–green coloring indicating which allele is carried by a lineage at any particular time. The random variables $\mathbf{n}_x^{\mathrm{B}}$ and $\mathbf{r}_x^{\mathrm{B}}$ equal the number of lineages and the number of red lineages, respectively, at the bottom of the branch for ancestral species x. The corresponding values at the top of this branch are denoted $\mathbf{n}_x^{\mathrm{T}}$ and $\mathbf{r}_x^{\mathrm{T}}$, respectively.

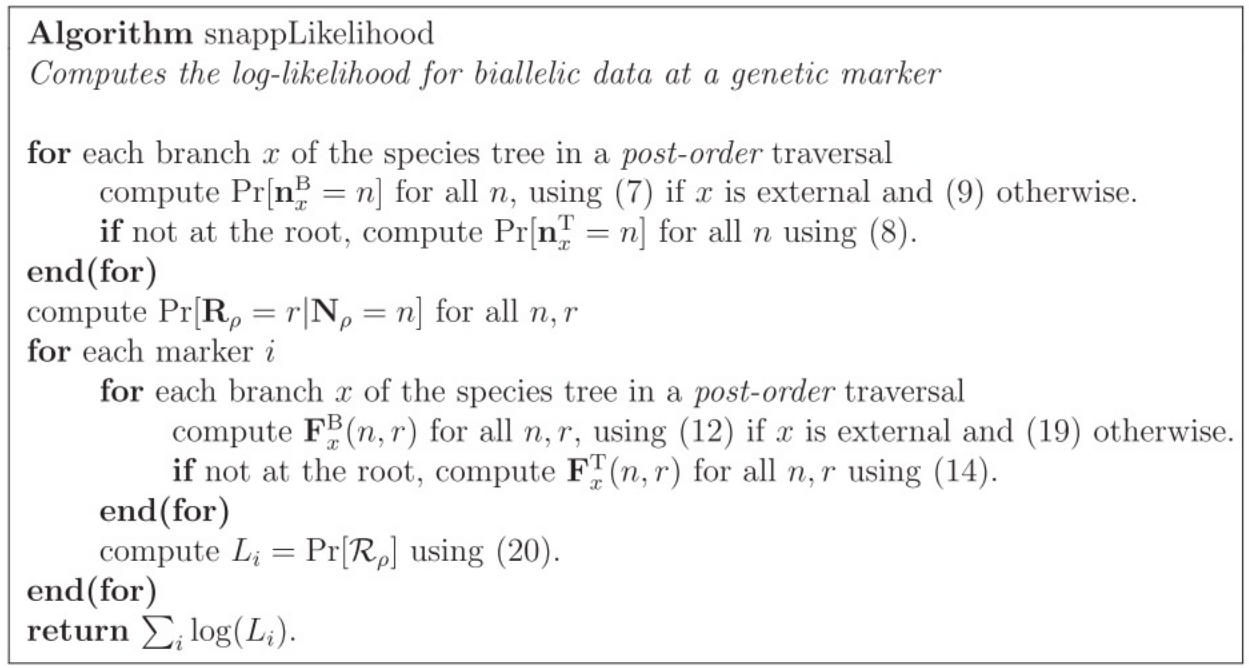


FIG. 2. High-level outline of the algorithm to compute the log-likelihood of a set of unlinked biallelic markers, given the species tree. A branch *x* in the species tree is external if it is adjacent to a leaf; otherwise, it is internal. In equations (9) and (19), we use *y* and *z* to denote the branches attached to the base of branch *x*.

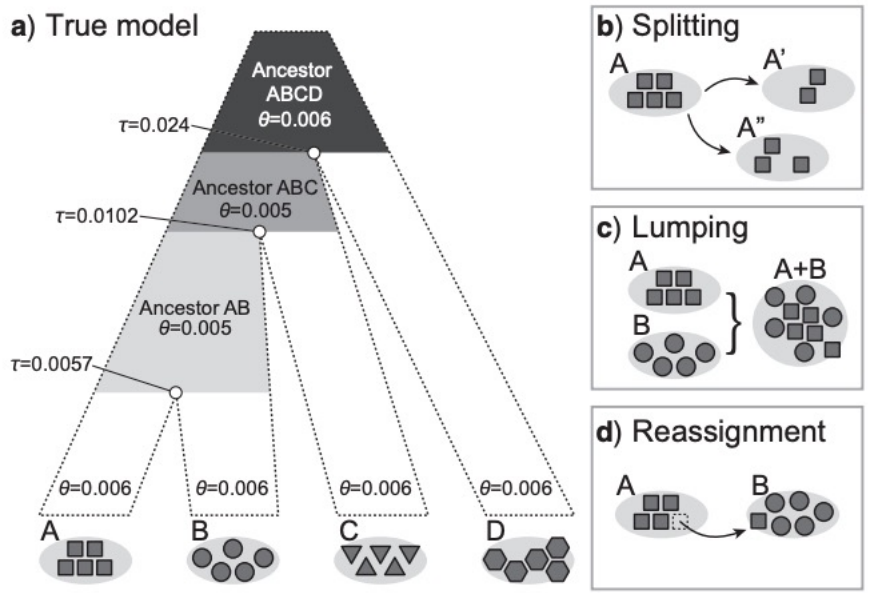


FIGURE 1. The fully specified species tree used to simulate SNP data for Bayes factor species delimitation (a). Perturbations to the true model include (b) splitting a species into two false species, (c) lumping two distinct species into one, and (d) reassigning a sample into the wrong species. Simulations are conducted with SNP matrices of different sizes (100, 500, 1000), variable sampling within species (2, 5, 10), and with different theta priors (correct, high, low). Species tree divergence times are in units of expected mutations per site.

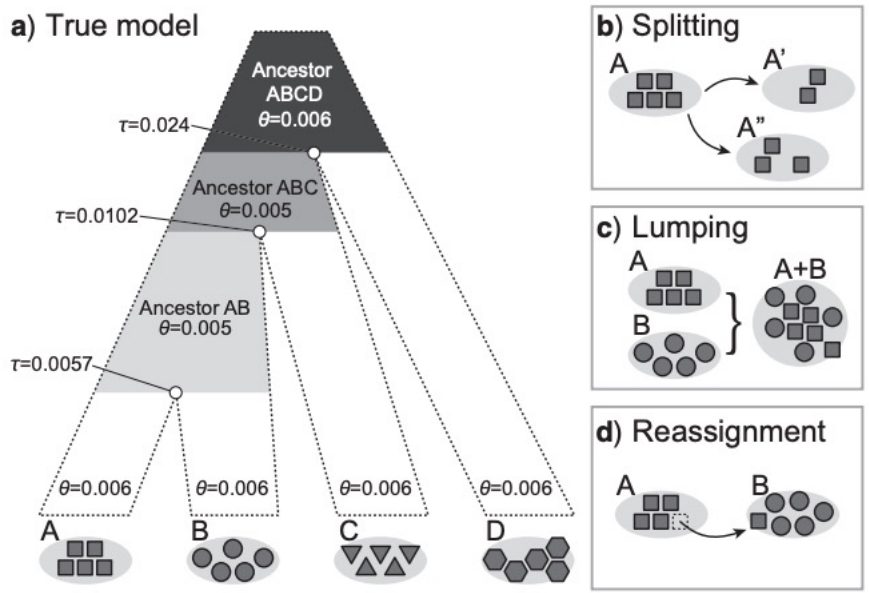


FIGURE 1. The fully specified species tree used to simulate SNP data for Bayes factor species delimitation (a). Perturbations to the true model include (b) splitting a species into two false species, (c) lumping two distinct species into one, and (d) reassigning a sample into the wrong species. Simulations are conducted with SNP matrices of different sizes (100, 500, 1000), variable sampling within species (2, 5, 10), and with different theta priors (correct, high, low). Species tree divergence times are in units of expected mutations per site.

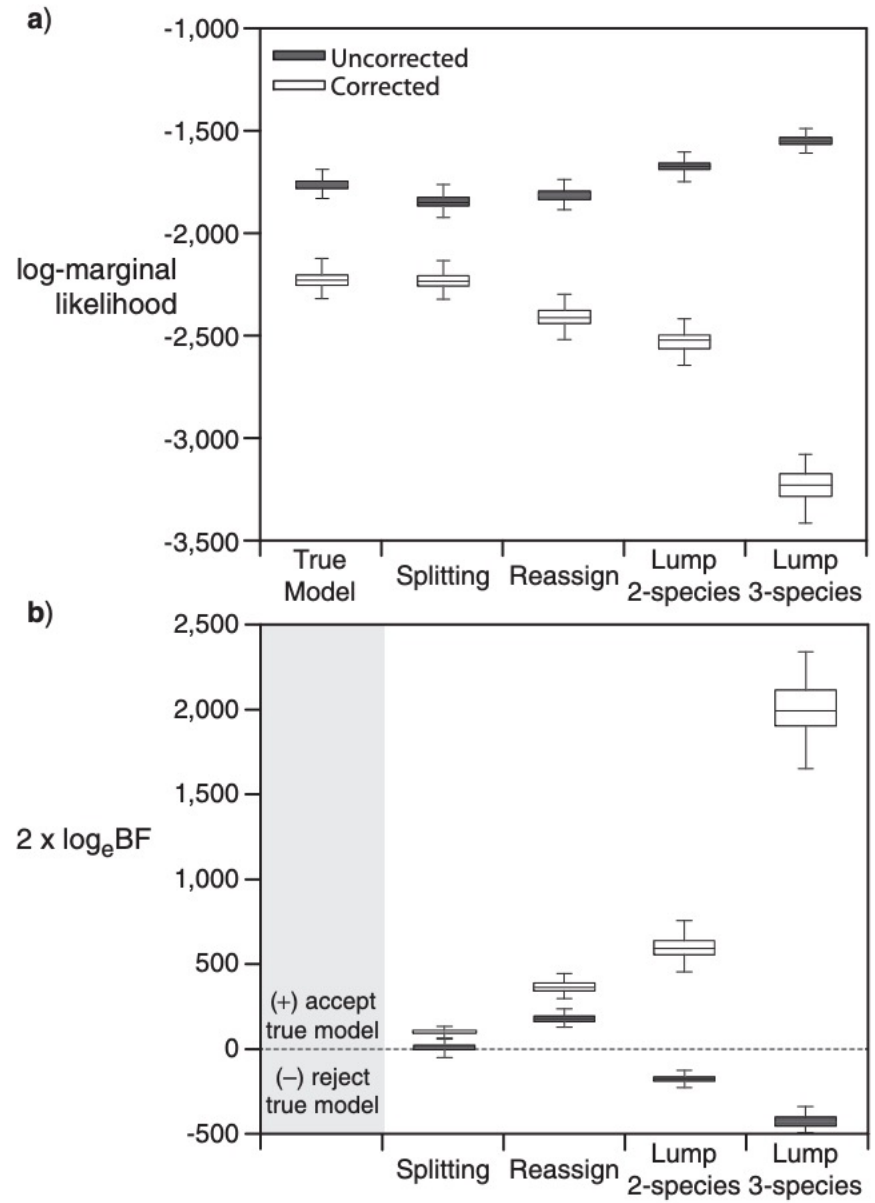


FIGURE 3. Comparisons of the behavior of corrected and uncorrected marginal likelihoods (a), and their influence on Bayes factor comparisons of candidate species trees (b). The simulated data used in this comparison include 500 SNPs and 5 samples per species.

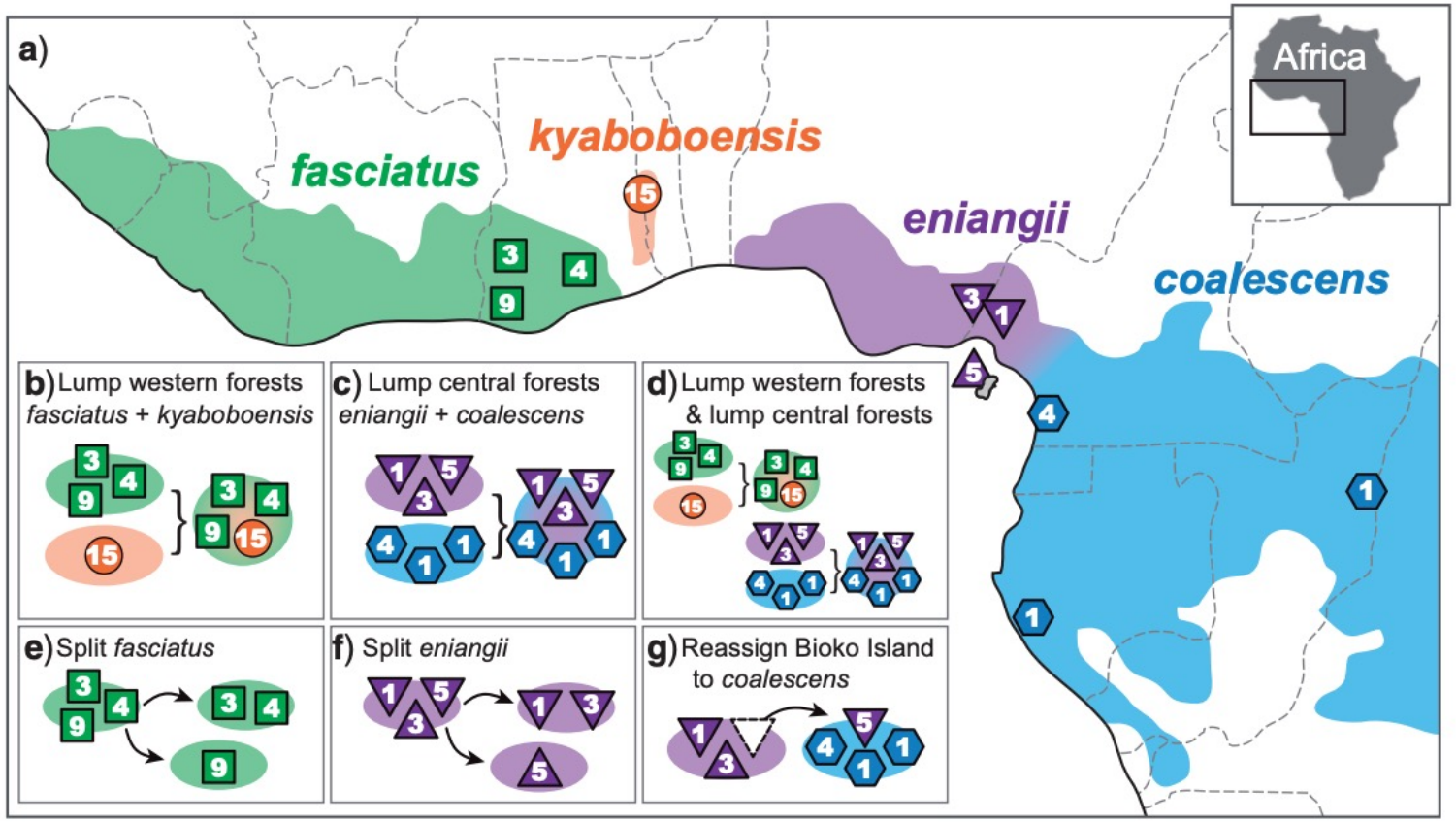


FIGURE 2. Geographic sampling of *Hemidactylus fasciatus* complex geckos (numbers in symbols indicate sample sizes), and our preferred current taxonomy (a). BFD* is used to test the alternative species delimitation models outlined in b–g.

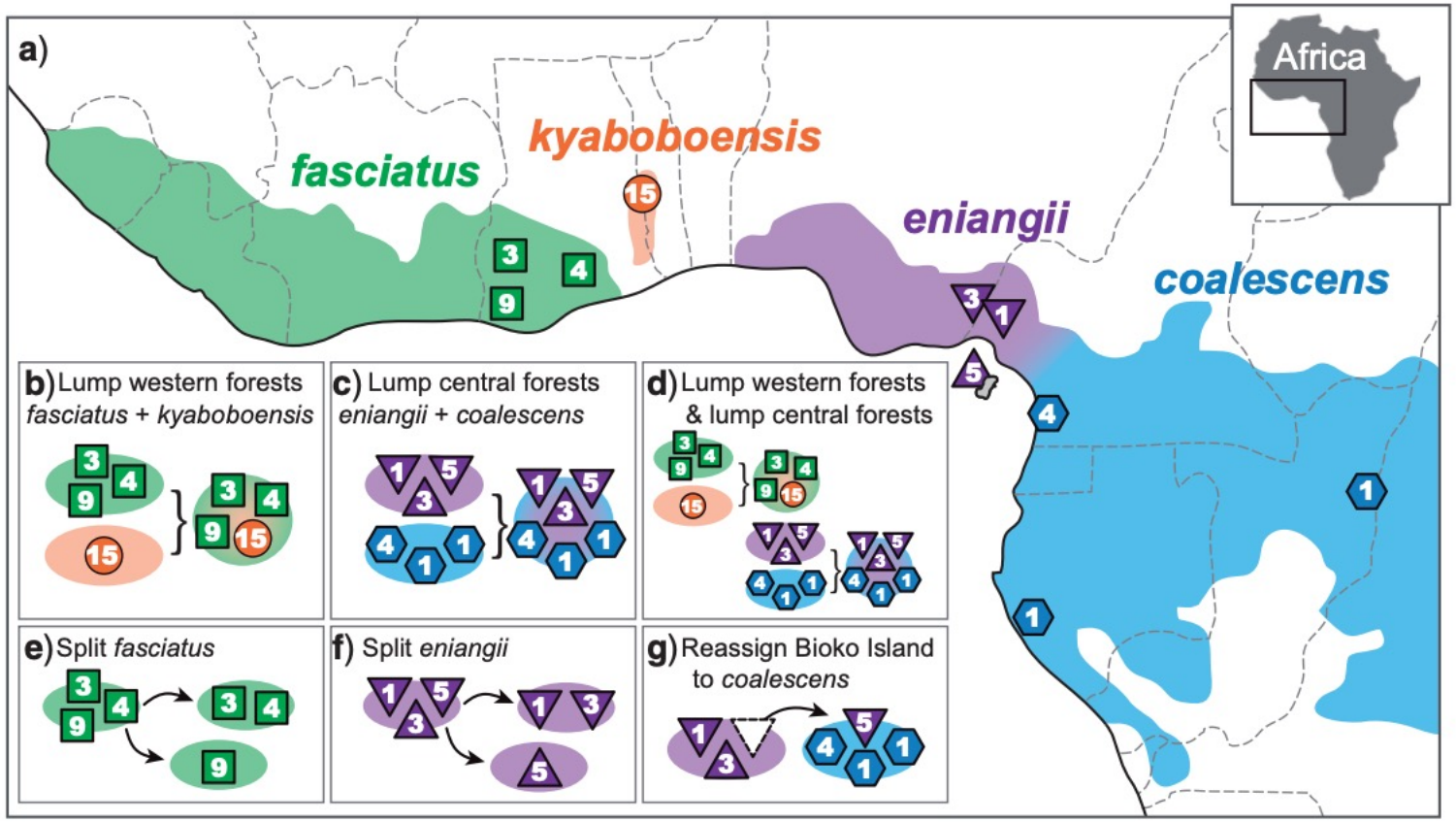


FIGURE 2. Geographic sampling of *Hemidactylus fasciatus* complex geckos (numbers in symbols indicate sample sizes), and our preferred current taxonomy (a). BFD* is used to test the alternative species delimitation models outlined in b–g.

TABLE 2. Empirical results for BFD* species delimitation in the *Hemidactylus fasciatus* complex

		129 SNPs			1087 SNPs		
Model	Species	ML	Rank	BF	ML	Rank	BF
a. Current taxonomy	4	−1673.4	2	–	−12890.3	2	–
b. Lump western forests	3	−1724.2	5	+101.5	−15024.5	6	+4268.3
c. Lump central forests	3	−1788.0	6	+229.2	−14094.0	5	+2407.4
d. Lump western & central forests	2	−1842.9	7	+339.0	−16190.4	7	+6600.3
e. Split *fasciatus*	5	−1713.2	4	+79.7	−13088.0	3	+395.5
f. Split *eniangii*	5	−1625.9	1	−95.1	−12615.3	1	−550.0
g. Reassign Bioko Island	4	−1712.6	3	+78.4	−13434.4	4	+1088.2

Biases towards over-splitting with BFD

Syst. Biol. 63(2):119–133, 2014

DOI:10.1093/sysbio/syt069
Advance Access publication November 20, 2013

Species Delimitation Using Bayes Factors: Simulations and Application to the *Sceloporus scalaris* Species Group (Squamata: Phrynosomatidae)

JARED A. GRUMMER[1,2,*], ROBERT W. BRYSON JR.[2], AND TOD W. REEDER[1]

[1]*Department of Biology, San Diego State University, San Diego, CA 92182-4614, USA and* [2]*Department of Biology and Burke Museum of Natural History and Culture, University of Washington, Box 351800, Seattle, WA 98195-1800, USA*
**Correspondence to be sent to: Department of Biology, University of Washington, Box 351800, Seattle, WA 98195-1800, USA; E-mail: grummer@uw.edu*

Received 25 March 2013; reviews returned 7 June 2013; accepted 11 November 2013
Associate Editor: Laura Kubatko

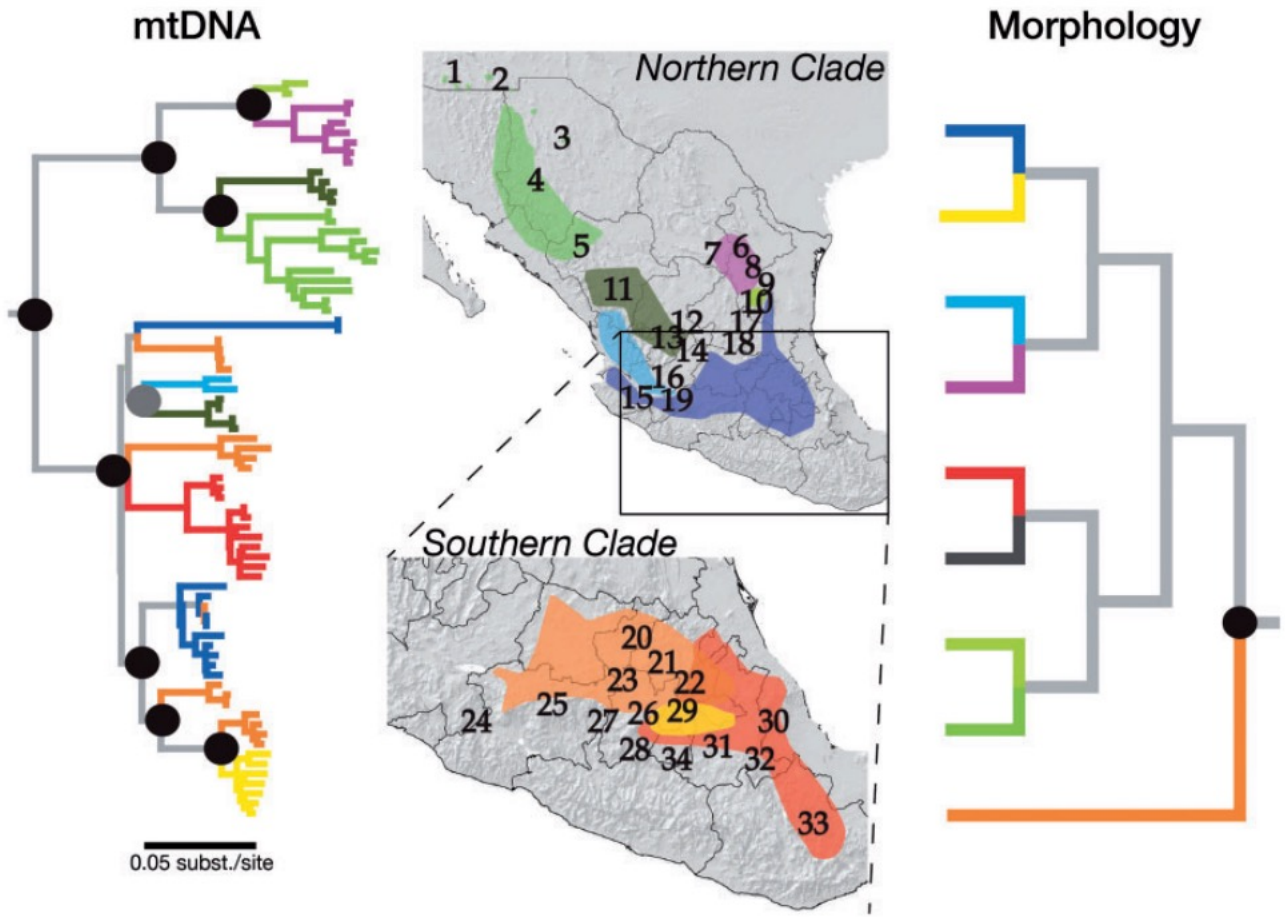


FIGURE 1. Phylogenetic relationships inferred for the *S. scalaris* species group based on maximum-likelihood analysis of the concatenated mtDNA and previously published morphological phylogeny (Wiens and Reeder 1997). The range map indicates sampling localities (Supplementary Table S1) and geographic ranges. The branch leading to *S. goldmani* in the morphological tree is represented in dark gray to reflect that the species is likely extinct and is not sampled for DNA. Black dots on nodes represent bootstrap proportions ≥85, whereas gray dots represent bootstrap proportions between 70 and 85.

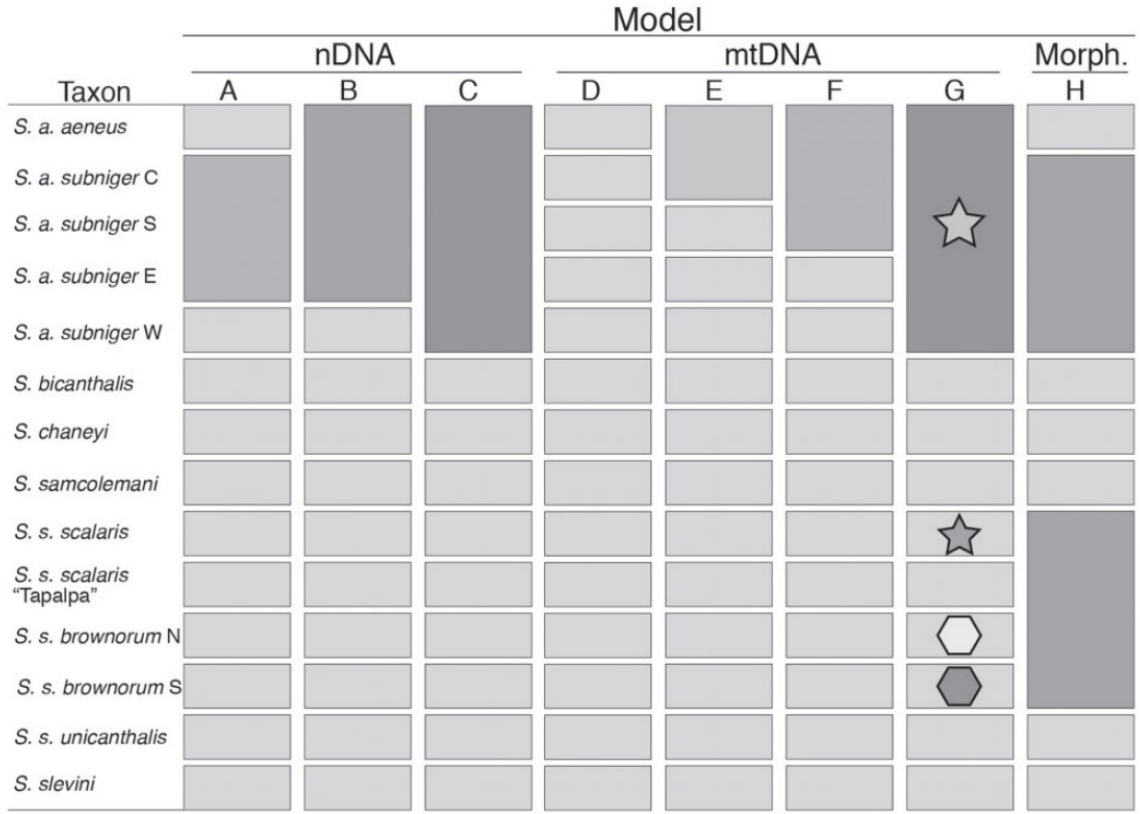


TABLE 3. Marginal-likelihood estimates and Bayes factor testing results (2*ln*Bf) from the combined analyses of four independent runs with sHME, HME, PS, and SS methods

Model	sHME		HME		PS		SS	
	MLE	2*ln*Bf	MLE	2*ln*Bf	MLE	2*ln*Bf	MLE	2*ln*Bf
A	–11 485.08	59.37	–11 457.45	3.26	**–12 057.53**	N/A	**–12 060.91**	N/A
B	**–11 455.40**	N/A	**–11 455.82**	N/A	–12 072.20	29.33	–12 076.38	30.93
C	–11 460.99	11.17	–11 461.19	10.75	–12 108.66	102.26	–12 108.86	95.89
D	–12 128.89	1346.99	–12 128.80	1345.97	–12 764.55	1414.05	–12 763.71	1405.59
E	–11 462.35	13.91	–11 462.83	14.02	–12 178.85	242.63	–12 173.89	225.95
F	–11 461.78	12.75	–11 462.60	13.56	–12 180.29	245.53	–12 175.92	230.02
G	–11 464.56	18.33	–11 465.04	18.44	–12 216.95	318.83	–12 210.81	299.79
H	–11 459.50	8.21	–11 460.44	9.24	–12 236.92	358.78	–12 232.97	344.11

The model receiving the best marginal-likelihood score for each estimation method is indicated by a 2*ln*Bf score = N/A, and its associated marginal likelihood is in bold.

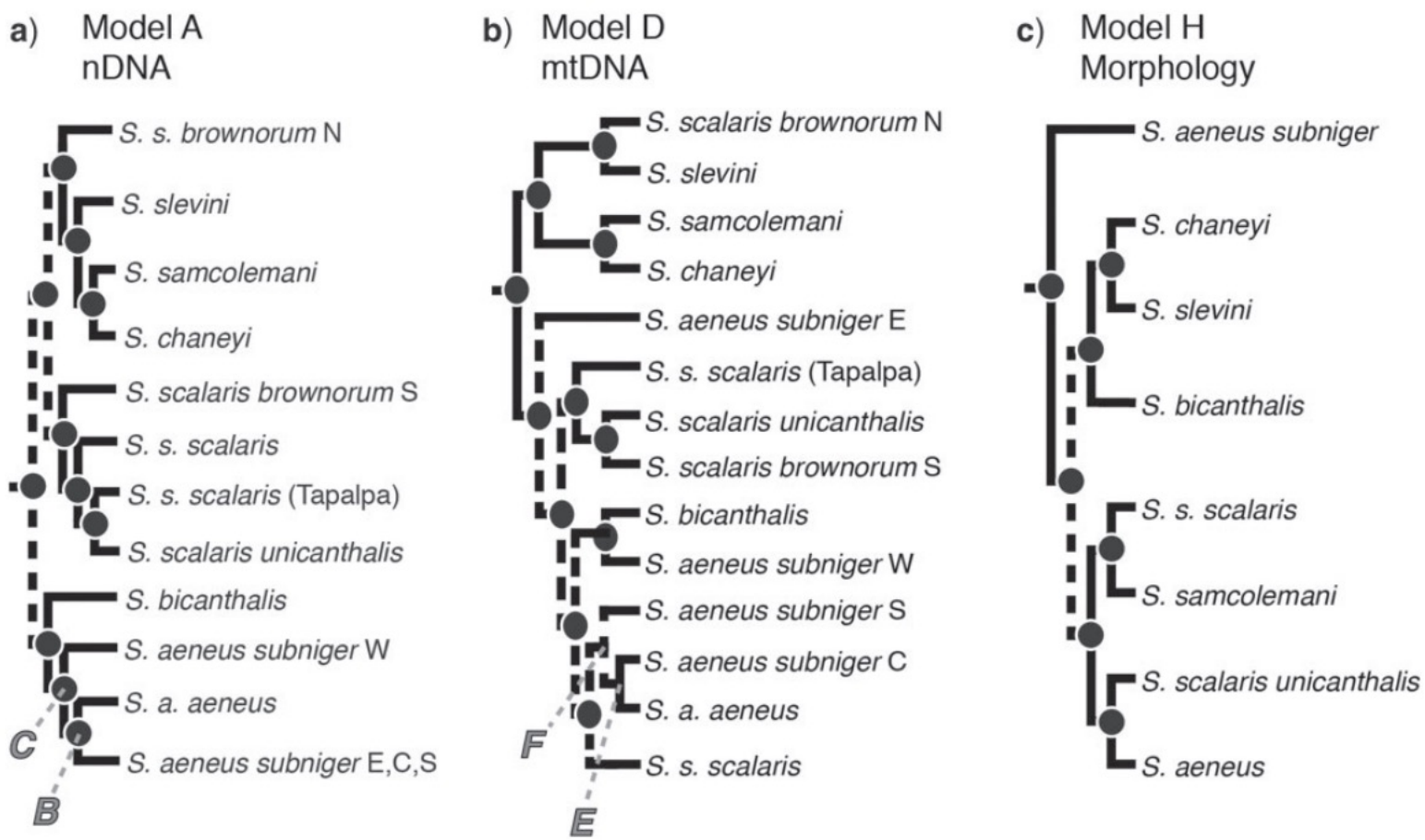


FIGURE 7. Species delimitation results from BP&P. The concatenated a) nDNA, b) mtDNA, and c) morphology guide trees are equivalent to Models A, D, and H in Figure 3, respectively. Letters at nodes represent the species delimitation model resulting from collapsing daughter taxa from that node into a single composite lineage. Model G is not shown because lineage compositions differ from Model D due to individual reassignments. Black dots represent nodes that received a Pp-value $\geq$0.95 across the three prior combinations examined (see Methods section for details). Nodes that did not receive Pp $\geq$0.95 across all prior combinations examined are not labeled. The dashed branches represent where the subclades that served as guide trees were joined together (Supplementary Fig. S1).

MOLECULAR ECOLOGY

Molecular Ecology (2016) **25**, 2029–2045 doi: 10.1111/mec.13604

Species discovery and validation in a cryptic radiation of endangered primates: coalescent-based species delimitation in Madagascar's mouse lemurs

SCOTT HOTALING,* MARY E. FOLEY,* NICOLETTE M. LAWRENCE,* JOSE BOCANEGRA,* MARINA B. BLANCO,† RODIN RASOLOARISON,‡ § PETER M. KAPPELER,§ MEREDITH A. BARRETT,¶ ANNE D. YODER** and DAVID W. WEISROCK*

**Department of Biology, University of Kentucky, Lexington, KY 40506, USA, †Duke Lemur Center, Durham, NC 27705, USA, ‡Département de Biologie Animale, Université d'Antananarivo, BP 906 Antananarivo (101), Madagascar, §Behavioral Ecology and Sociobiology Unit, German Primate Center (DPZ), 37077 Göttingen, Germany, ¶Center for Health and Community, University of California, San Francisco, CA 94118, USA, **Department of Biology, Duke University, Durham, NC 27708, USA*

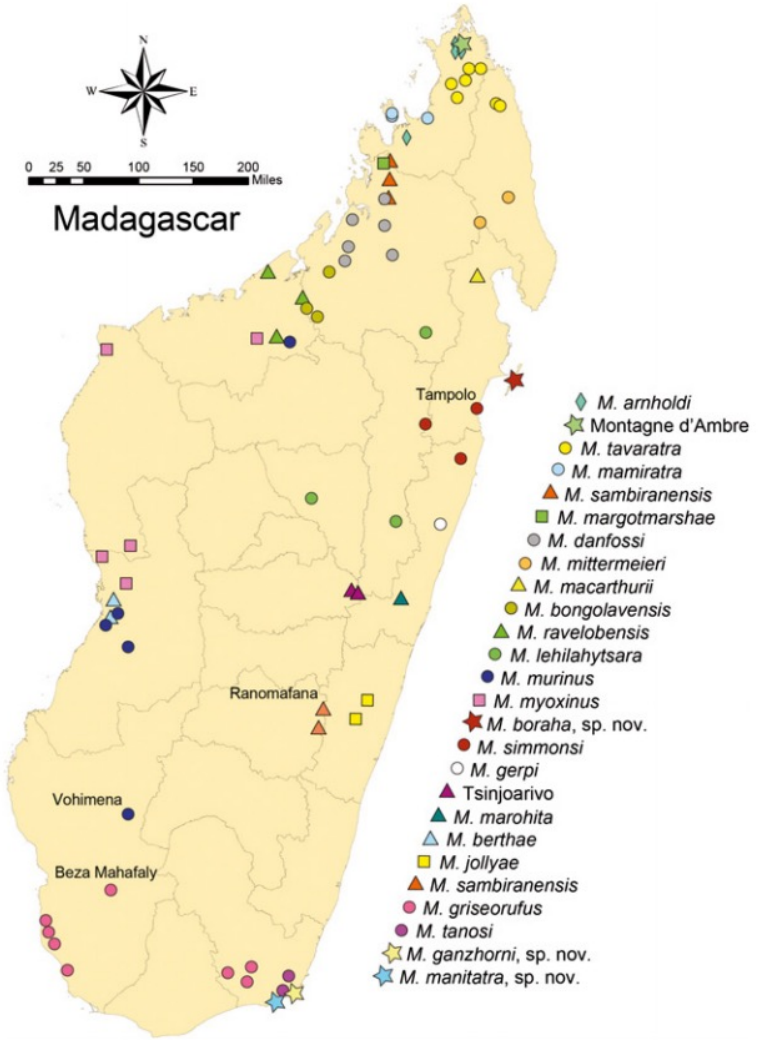


Fig. 1 A map of Madagascar showing the distribution of described *Microcebus* species according to localities used in Yoder *et al.* (2000), Louis *et al.* (2006, 2008), Olivieri *et al.* (2007), Radespiel *et al.* (2012) and Weisrock *et al.* (2010). Also included is the newly sampled Tsinjoarivo population. Not all localities presented here are those used in this study. All coordinate information for localities used here are available in the study's Dryad accession (doi:10.5061/dryad.h6s5j). Locality names specifically mentioned in the text are designated on the map.

(B)

	M. murinus	*M. murinus* - Mandena	*M. murinus* - Bemanasy	*M. murinus* - Vohimena	*M. griseorufus*	*M. griseorufus* - Beza Mahafaly	**TEST**: path sampling MLE	**TEST**: path sampling 2*ln*Bf	**RANDOM**: path sampling MLE
Model A							–6581.4	88.8	n/a
Model B		■					–6564.8	55.6	–6584.1
Model C			■				–6574.1	74.2	–6584.4
Model D				■			–6571.6	69.2	–6584.7
Model E					■	■	–6577.5	81.0	–6584.2
Model F		■	■				–6545.1	16.2	–6586.0
Model G		■		■			–6557.6	41.2	–6586.0
Model H		■			■	■	–6563.2	52.4	–6589.4
Model I			■	■			–6567.7	61.4	–6585.6
Model J			■		■	■	–6573.0	72.0	–6588.1
Model K				■	■	■	–6569.3	64.6	–6587.4
Model L	■	■	■	■			–6537.8	1.6	–6592.9
Model M		■	■		■	■	–6543.0	12.0	–6591.4
Model N		■		■	■	■	–6557.8	41.6	–6593.3
Model O			■	■	■	■	–6564.8	55.6	–6590.2
Model P	■	■	■	■	■	■	–6537.0	★	–6596.1

Syst. Biol. 68(1):168–181, 2019

DOI:10.1093/sysbio/syy051
Advance Access publication July 5, 2018

The Spectre of Too Many Species

ADAM D. LEACHÉ[1], TIANQI ZHU[2,3], BRUCE RANNALA[4], AND ZIHENG YANG[2,5,6,*]

[1]*Department of Biology & Burke Museum of Natural History and Culture, University of Washington, Seattle, WA 98195, USA;*
[2]*National Center for Mathematics and Interdisciplinary Sciences, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China;*
[3]*Key Laboratory of Random Complex Structures and Data Science, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China;*
[4]*Department of Evolution and Ecology, University of California Davis, One Shields Avenue, Davis, CA 95645, USA;*
[5]*Department of Genetics, University College London, London WC1E 6BT, UK;*
and [6]*Radcliffe Institute for Advanced Studies, Harvard University, Cambridge, MA 02138, USA*
**Correspondence to be sent to: Department of Genetics, University College London, London WC1E 6BT, UK; E-mail: z.yang@ucl.ac.uk.*

Adam D. Leaché and Tianqi Zhu contributed equally to this article.

Received 7 November 2017; reviews returned 29 June 2018; accepted 29 June 2018
Associate Editor: Matthew Hahn

Abstract.—Recent simulation studies examining the performance of Bayesian species delimitation as implemented in the BPP program have suggested that BPP may detect population splits but not species divergences and that it tends to over-split when data of many loci are analyzed. Here, we confirm these results and provide the mathematical justifications. We point out that the distinction between population and species splits made in the protracted speciation model (PSM) has no influence on the generation of gene trees and sequence data, which explains why no method can use such data to distinguish between population splits and speciation. We suggest that the PSM is unrealistic as its mechanism for assigning species status assumes instantaneous speciation, contradicting prevailing taxonomic practice. We confirm the suggestion, based on simulation, that in the case of speciation with gene flow, Bayesian model selection as implemented in BPP tends to detect population splits when the amount of data (the number of loci) increases. We discuss the use of a recently proposed empirical genealogical divergence index (*gdi*) for species delimitation and illustrate that parameter estimates produced by a full likelihood analysis as implemented in BPP provide much more reliable inference under the *gdi* than the approximate method PHRAPL. We distinguish between Bayesian model selection and parameter estimation and suggest that the model selection approach is useful for identifying sympatric cryptic species, while the parameter estimation approach may be used to implement empirical criteria for determining species status among allopatric populations. [BPP; multispecies coalescent; Species delimitation; taxonomy.]

Splitting should be expected

Splitting should be expected

$$KL = q(x) log \frac{q(x)}{p(x|\phi)} dx$$

Splitting should be expected

$$KL = q(x) log \frac{q(x)}{p(x|\phi)} dx$$

Even if speciation is not complete, the model with divergence will always be preferred over one without that disregards that information

Heuristic Criteria informed by MSC parameter estimates

Heuristic Criteria informed by MSC parameter estimates

$$gdi = 1 - exp\left\{-\frac{2\tau}{\theta_A}\right\}$$

Heuristic Criteria informed by MSC parameter estimates

$$gdi = 1 - exp\left\{-\frac{2\tau}{\theta_A}\right\}$$

Has to be computed twice for the two θ estimates – once from the left child and once from the right

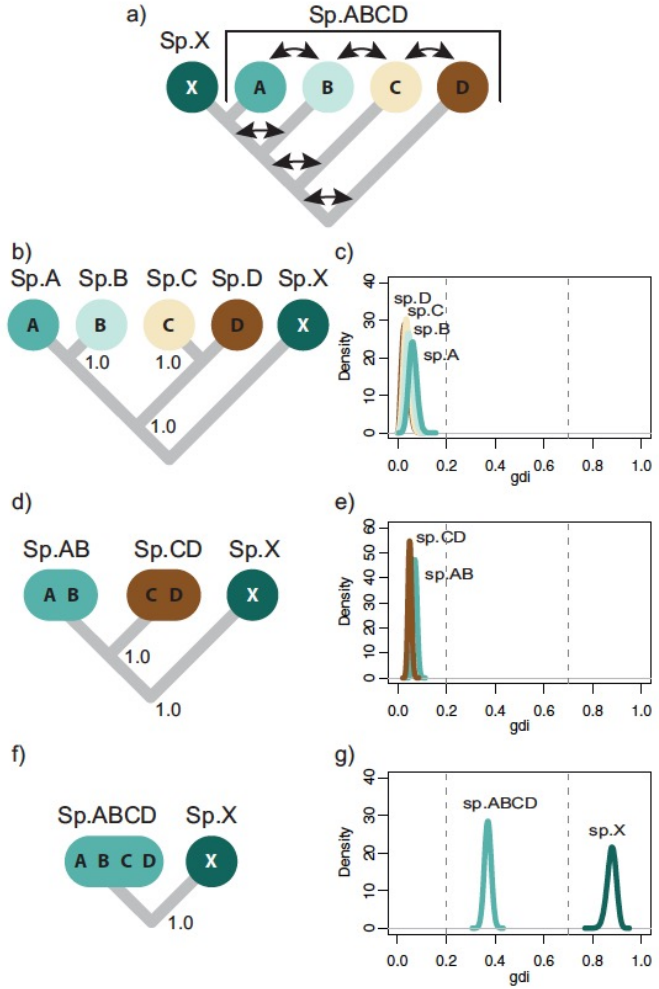


FIGURE 5. Species delimitation applying heuristic index *gdi* to parameter estimates from BPP. a) Species tree used for simulation allows migration between populations A, B, C, and D and their ancestors (indicated by arrows), but no gene flow involving species X. b) Species (guide) tree inferred from A11 analysis of BPP. In (b–g), *gdi* is used to collapse populations on guide tree into same species in a hierarchical procedure, with BPP used to estimate MSC parameters (θ and τ) and generate posterior distribution of *gdi*. For example, *gdi* calculated using population A of panel b, based on $2\tau_{AB}/\theta_A$ (equation 7), is shown in panel c (labeled 'sp. A'). Sister populations inferred to belong to same species by *gdi* are collapsed, and resulting species tree is used to conduct a new BPP analysis. Procedure is repeated until distinct species are inferred or until root of tree is reached. According to Jackson et al. (2017), $gdi < 0.2$ indicates a single species, $gdi > 0.7$ indicates distinct species, and *gdi* values between 0.2 and 0.7 represent ambiguous species status.

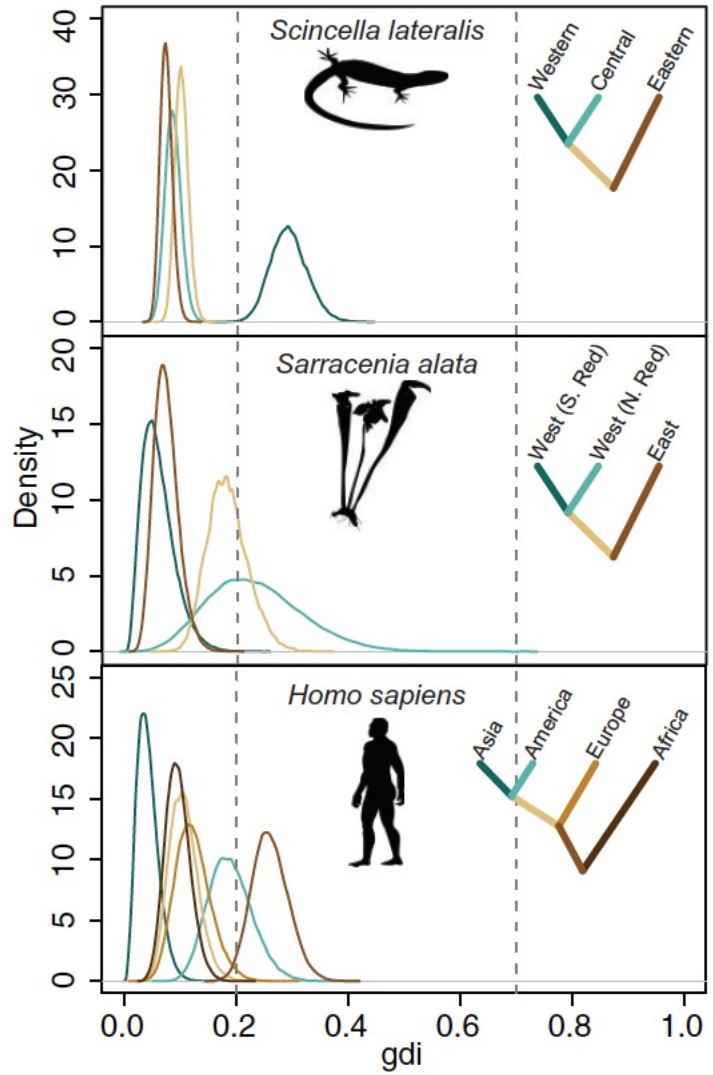
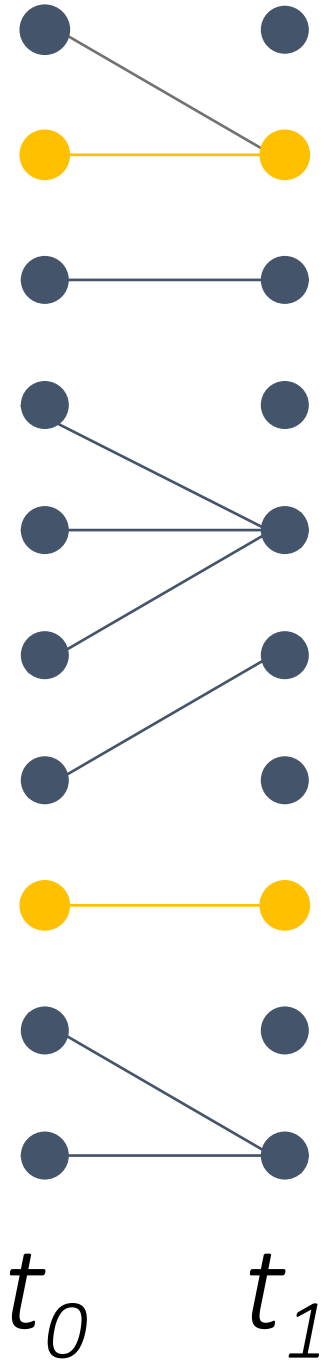


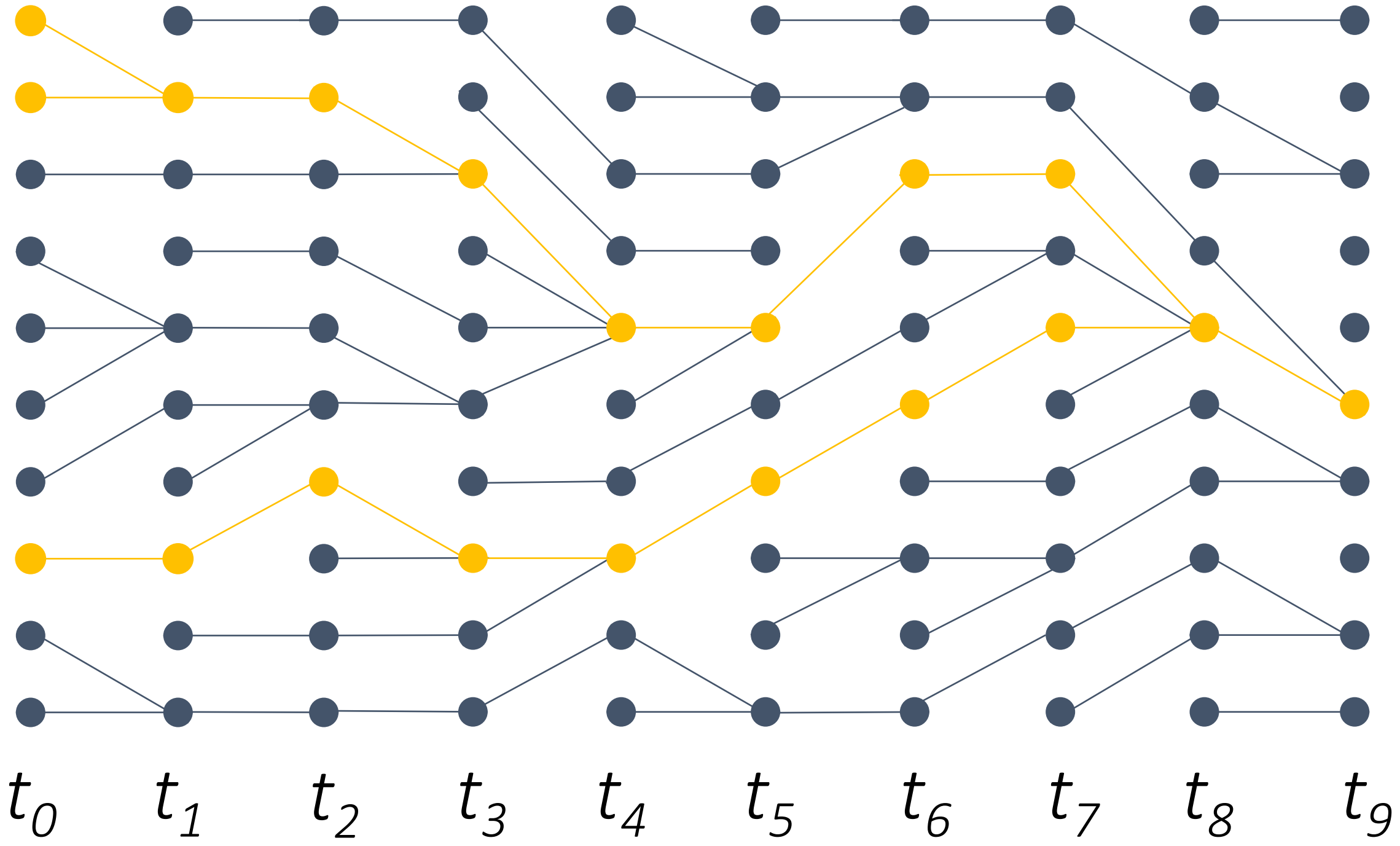
FIGURE 6. Posterior distribution of genealogical divergence index (*gdi*), generated in BPP analysis of three real data sets of Jackson et al. (2017). Silhouettes of species are from phylopic.org http://phylopic.org. Colored ancestral branches were analyzed by collapsing descendent species and conducting new MCMC analyses.

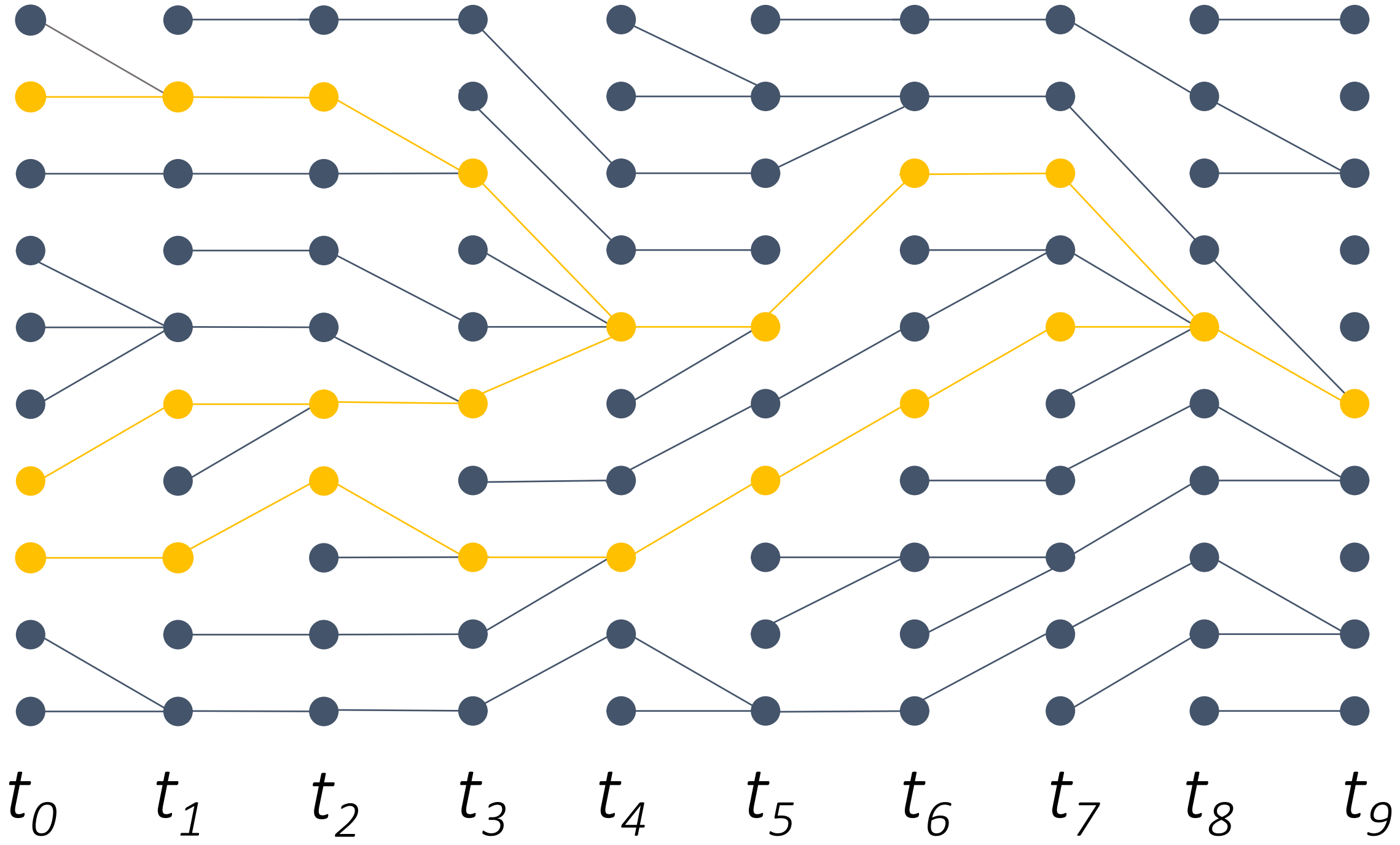
End

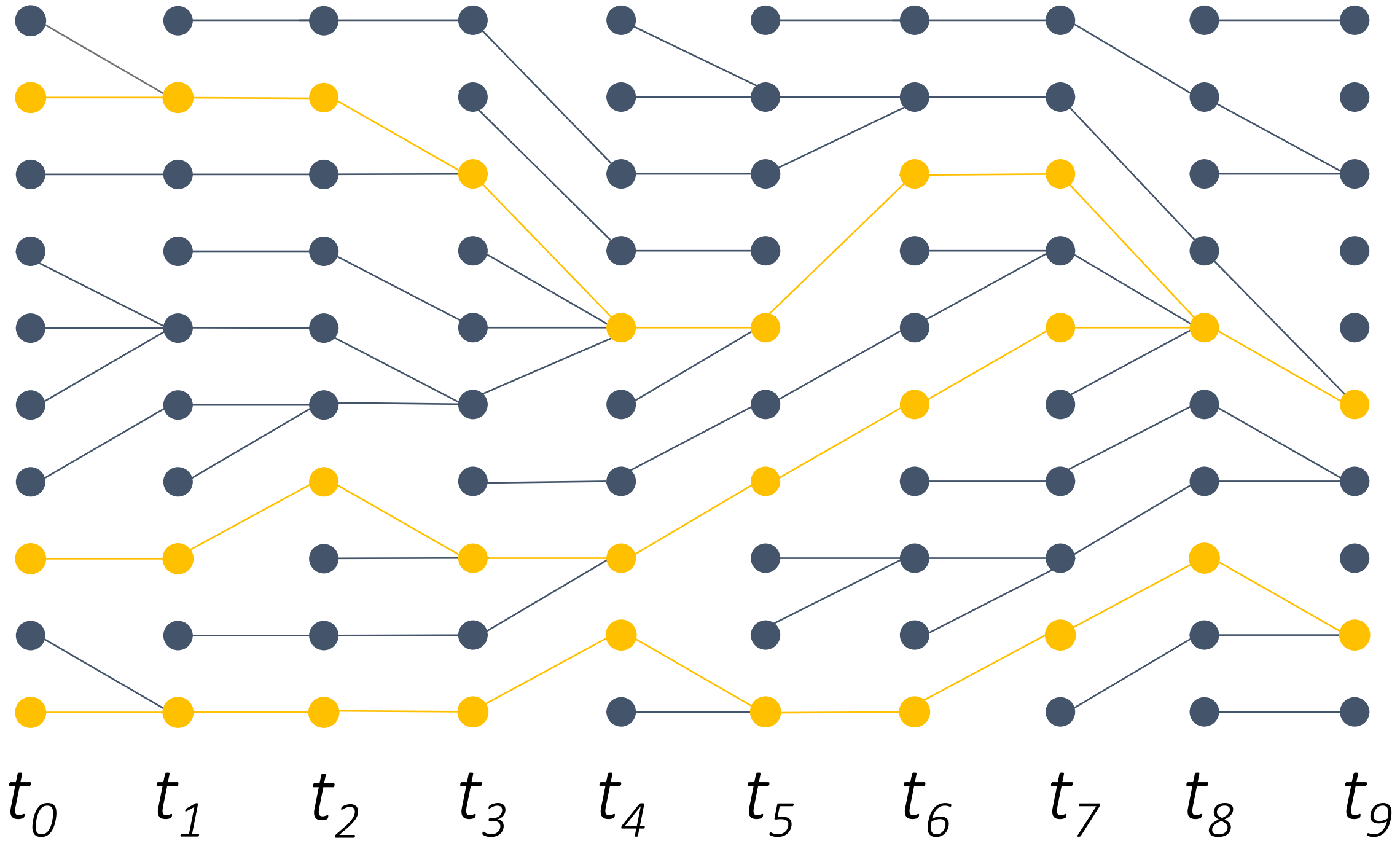
Labs will continue excercises with IQTREE and then move onto analyses under the MSC model with ASTRAL. We will try to do a BPP analysis if time permits.



What about more than 2 alleles?


t_0
t_1
t_2
t_3
t_4
t_5
t_6
t_7
t_8
t_9

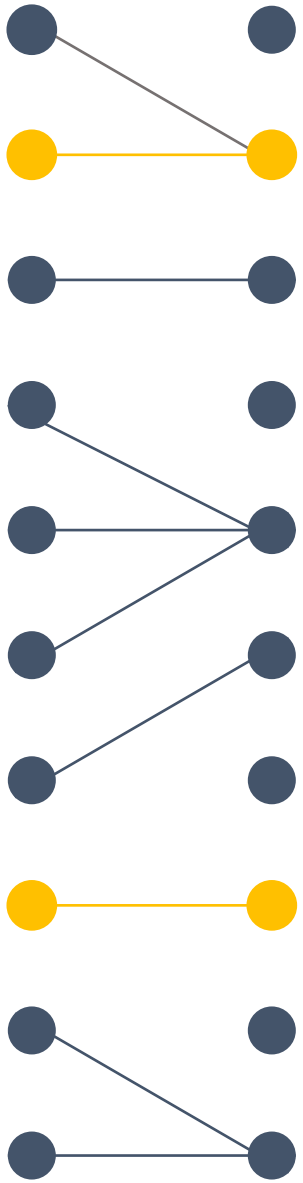

t_0
t_1
t_2
t_3
t_4
t_5
t_6
t_7
t_8
t_9


t_0
t_1
t_2
t_3
t_4
t_5
t_6
t_7
t_8
t_9

P(3 alleles do not coalesce in 1 generation)

$$= \left(1 - \frac{1}{2N}\right) \times \left(1 - \frac{2}{2N}\right)$$

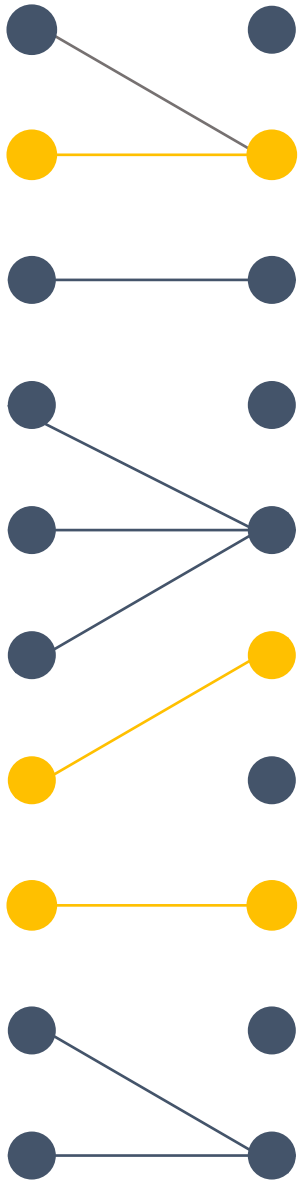
t_0 t_1 t_2 t_3 t_4 t_5 t_6 t_7 t_8 t_9



P(2 alleles do not coalesce in 1 generation)

$$= \left(1 - \frac{1}{2N}\right)$$

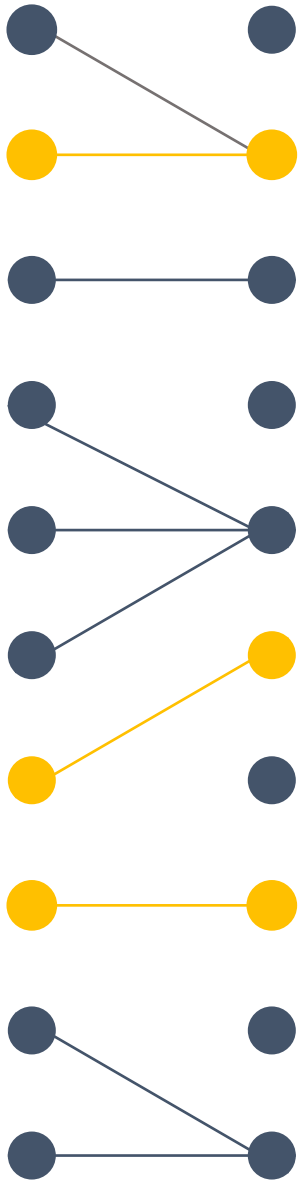
t_0 t_1 t_2 t_3 t_4 t_5 t_6 t_7 t_8 t_9



P(3 alleles do not coalesce in 1 generation)

$$= \left(1 - \frac{1}{2N}\right) \times \left(1 - \frac{2}{2N}\right)$$

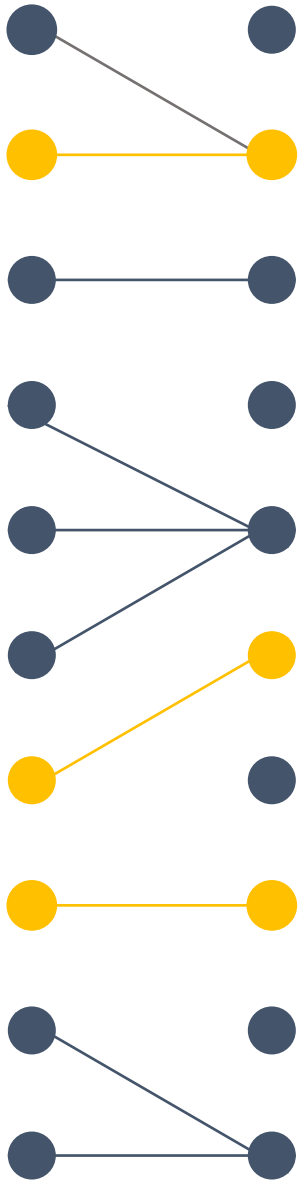
t_0 t_1 t_2 t_3 t_4 t_5 t_6 t_7 t_8 t_9



P(*n* alleles do not coalesce in 1 generation)

$$= \left(1 - \frac{1}{2N}\right) \times \left(1 - \frac{2}{2N}\right) \times \cdots \times \left(1 - \frac{n-1}{2N}\right)$$

t_0 t_1 t_2 t_3 t_4 t_5 t_6 t_7 t_8 t_9

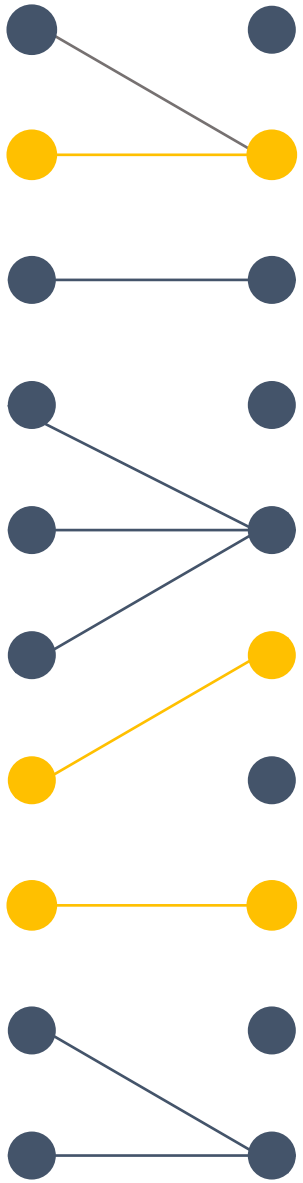


P(*n* alleles do not coalesce in 1 generation)

$$\approx \left(1 - \frac{1 + 2 + \cdots + (n-1)}{2N}\right)$$

$$= 1 - \binom{n}{2}\left(\frac{1}{2N}\right)$$

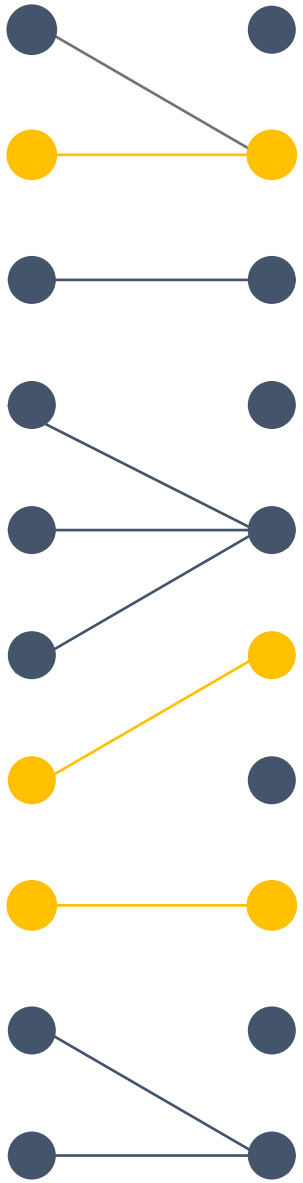
t_0 t_1 t_2 t_3 t_4 t_5 t_6 t_7 t_8 t_9



P(*n* alleles coalesce in the i^{th} generation)

$$= \left[1 - \binom{n}{2}\left(\frac{1}{2N}\right)\right]^{i-1} \binom{n}{2}\frac{1}{2N}$$

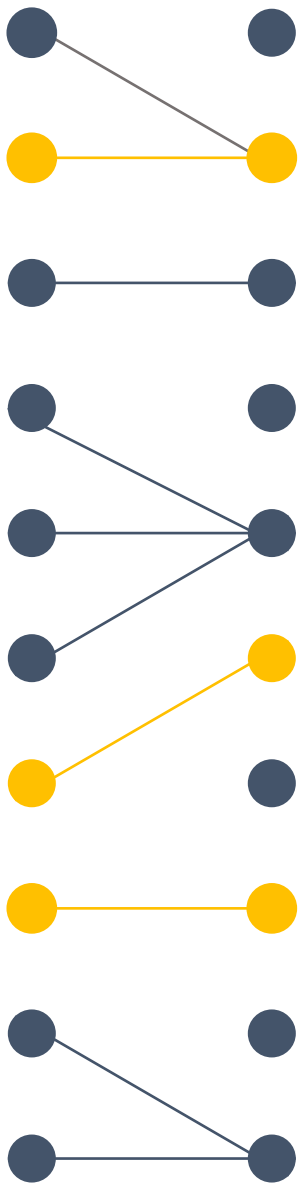
t_0 t_1 t_2 t_3 t_4 t_5 t_6 t_7 t_8 t_9



This gets things into a similar geometric distribution, thus

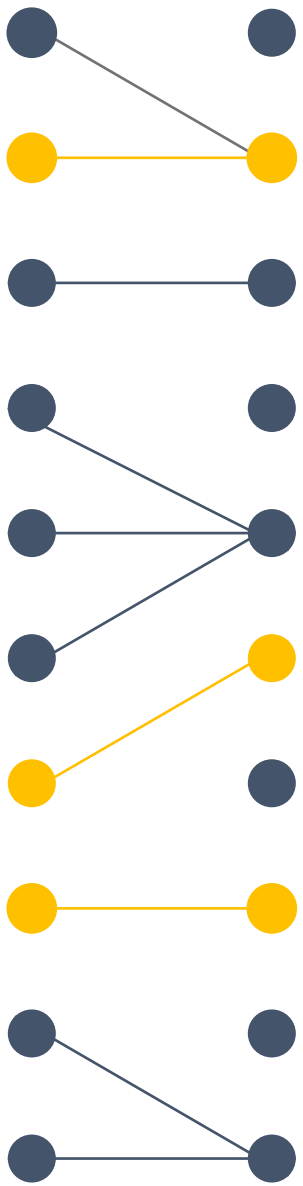
$$\mathrm{E}[t] = 1/p = \frac{2N}{\binom{n}{2}}$$

t_0 t_1 t_2 t_3 t_4 t_5 t_6 t_7 t_8 t_9



And we can approximate the probability of coalescent waiting times with an exponential

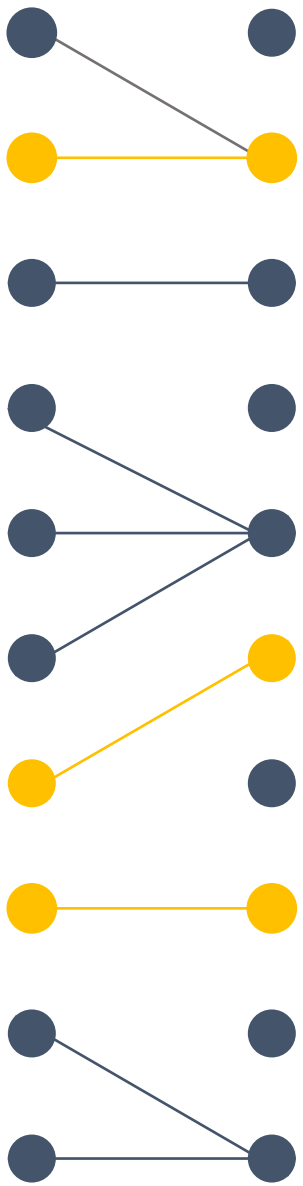
t_0 t_1 t_2 t_3 t_4 t_5 t_6 t_7 t_8 t_9



And we can approximate the probability of coalescent waiting times with an exponential

$$T = \frac{1}{2N}$$

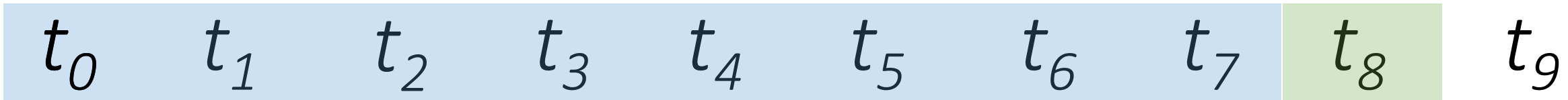
t_0 t_1 t_2 t_3 t_4 t_5 t_6 t_7 t_8 t_9

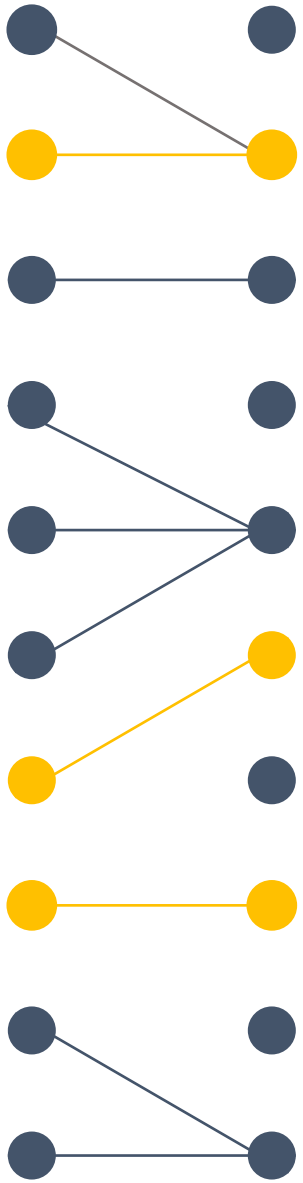


And we can approximate the probability of coalescent waiting times with an exponential

$$T = \frac{1}{2N}$$

But there are *n-1* coalescent events to happen and $\binom{n}{2}$ ways to get there





$$\binom{n}{2} = \frac{n(n-1)}{2}$$

So for the j^{th} coalescence

$$f(T_j) = \frac{j(j-1)}{2} \exp\left\{-\frac{j(j-1)}{2} T_j\right\}$$

t_0 t_1 t_2 t_3 t_4 t_5 t_6 t_7 t_8 t_9

And all coalescences for a given genealogy *G*

$$f(T|G) = \prod_{j=2}^{n} \frac{j(j-1)}{2} \exp\left\{-\frac{j(j-1)}{2} T_j\right\}$$

We can derive expectations for the T_{MRCA}

$$f(T|G) = \prod_{j=2}^{n} \frac{j(j-1)}{2} \exp\left\{-\frac{j(j-1)}{2} T_j\right\}$$

$$E[T_j] = \frac{2}{j(j-1)}$$

$$E[T_{MRCA}] = E(T_n + T_{n-1} + \cdots + T_2)$$

We can derive expectations for the T_{MRCA}

$$E[T_{MRCA}] = E(T_n + T_{n-1} + \cdots + T_2)$$

$$E[T_{MRCA}] = \sum_{j=2}^{n} \frac{2}{j(j-1)} = 2\sum_{j=2}^{n}\left(\frac{1}{j-1} - \frac{1}{j}\right)$$

$$E[T_{MRCA}] = 2\left(1 - \frac{1}{n}\right) \approx 2 = 4N$$