

TADAB

workshop



Tools for Advanced Data Analyses in Biology

27 November – 1 December 2023

The coalescent model for phylogenetics
and species delimitation

Regression analysis of biological data in ecology

(Sergey Rosbakh)

Data preparation and missing data management

(Mariasole Calbi)

Application and exploration of machine learning in biological data analysis

(Sina Gholami, Manuel Tiburtini)

Coalescent models in phylogeny and species delimitation

(Salvatore Tomasello)

Data visualization with ggplot2

(David Dolci)



REGISTRATION



Thursday morning	<ul style="list-style-type: none">- Introduction to species delimitation and the coalescent theory- Heuristic methods for species delimitation
Thursday afternoon	<ul style="list-style-type: none">- Parametric approaches: Species delimitation with SPEEDOMON
Friday morning	<ul style="list-style-type: none">- Parametric approaches: Species Delimitation BPP BPP- Integrative approaches for species delimitation: iBPP



Why species delimitation

- Humans like to order objects in classes

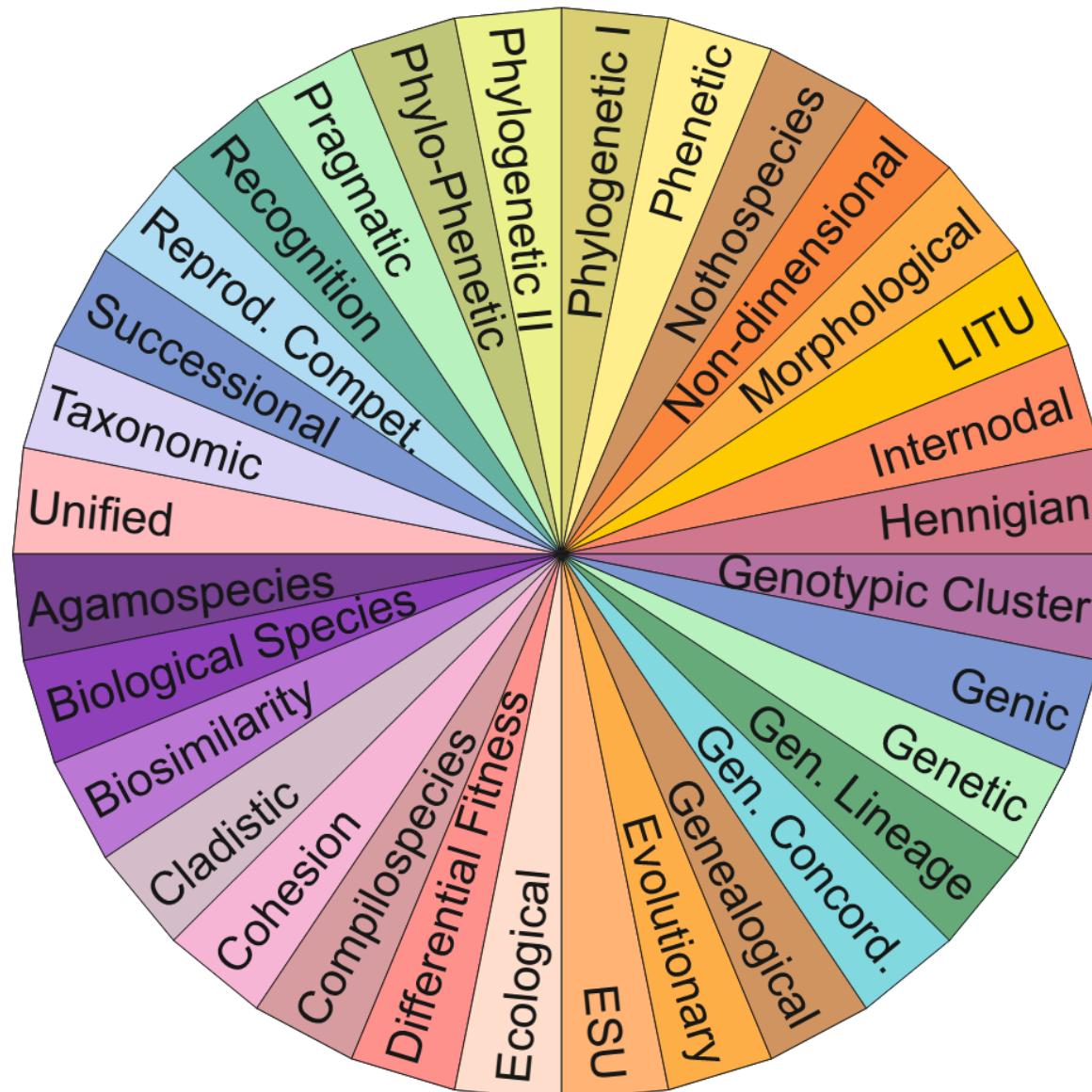




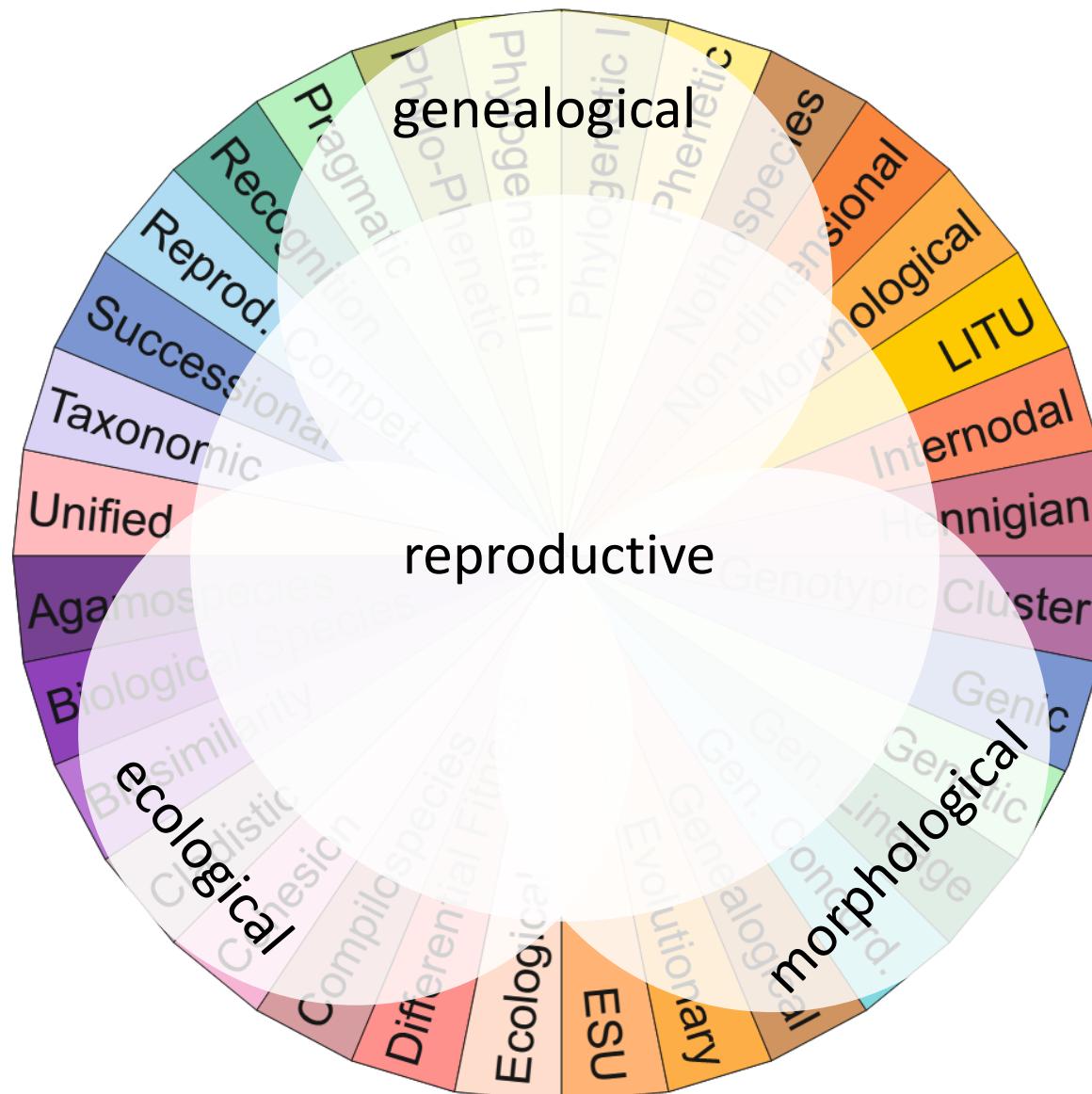
Why species delimitation

- Humans like to order objects in classes
- **Species exist!**
- **Fundamental unit in biology**
- **Reliable Taxonomy are extremely important for many biological disciplines**

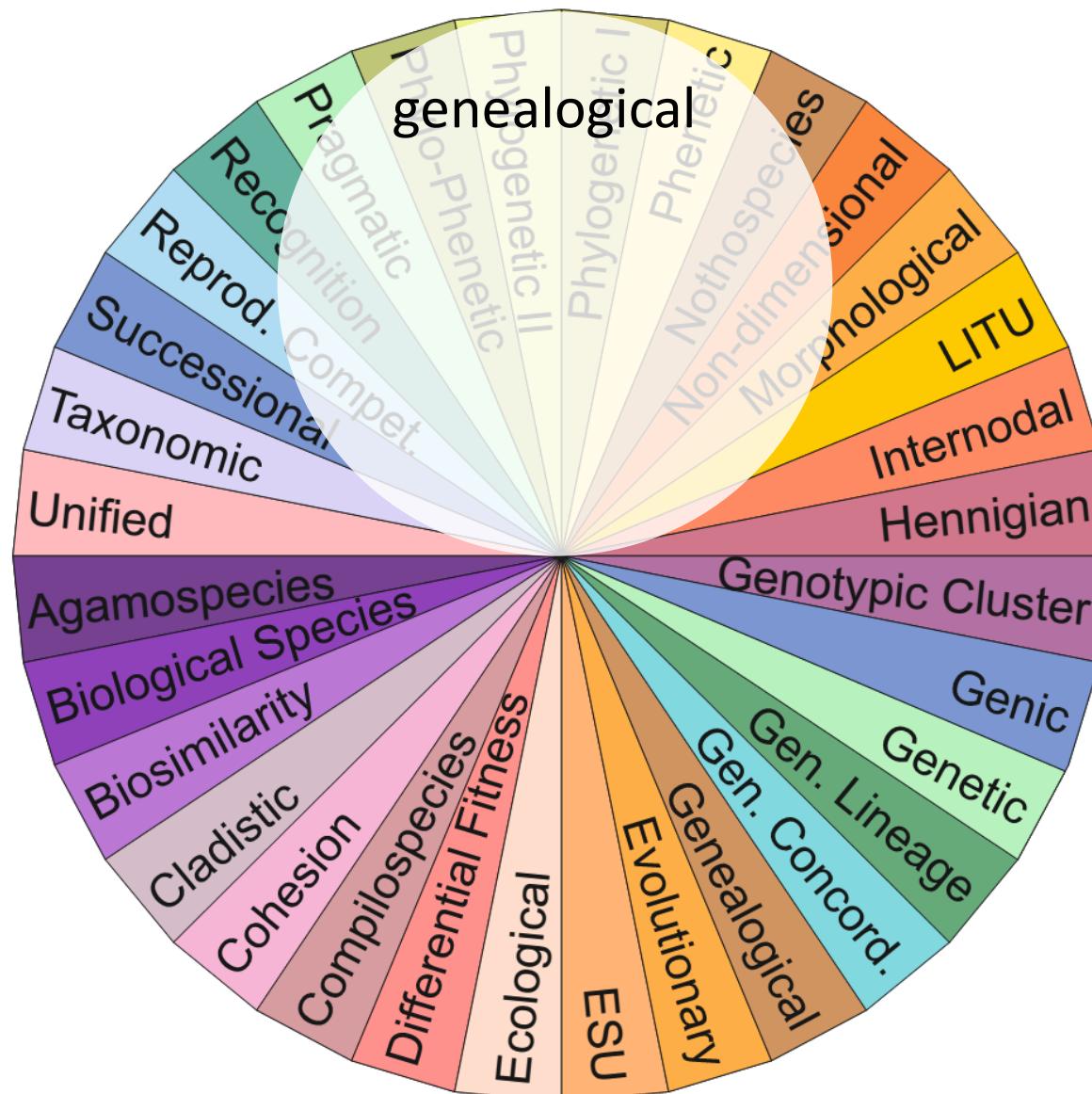
Why species delimitation



Why species delimitation



Why species delimitation



There are about 30 species concept in the literature (Zachos, 2016):

There are about 30 species concept in the literature (Zachos, 2016):

Biological Species Concept: Interbreeding natural populations reproductively isolated from other such groups; all individuals that produce fertile offspring

There are about 30 species concept in the literature (Zachos, 2016):

Biological Species Concept: Interbreeding natural populations reproductively isolated from other such groups; all individuals that produce fertile offspring

Hybridization common in different organismic groups!

There are about 30 species concept in the literature (Zachos, 2016):

Biological Species Concept: Interbreeding natural populations reproductively isolated from other such groups; all individuals that produce fertile offspring (Meyer)

Evolutionary Species Concept: Ancestor-descendant lineages that evolve separately from other such lineages and have their own evolutionary tendencies and historical fate

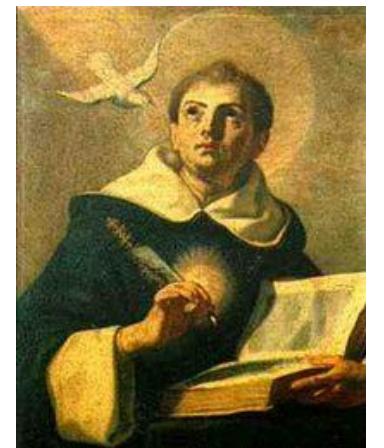
Unified Species Concept: separately evolving lineage segment (De Queiroz)

“the process of determining the boundaries and numbers of species from empirical data”

So far being done with:

- Subjective quantification of morphological variation
- Crossing experiments
- Phenology
- Secondary compounds
- Testing monophyly in phylogenies

„It belongs to the wise man
to order“ (Thomas Aquinas)



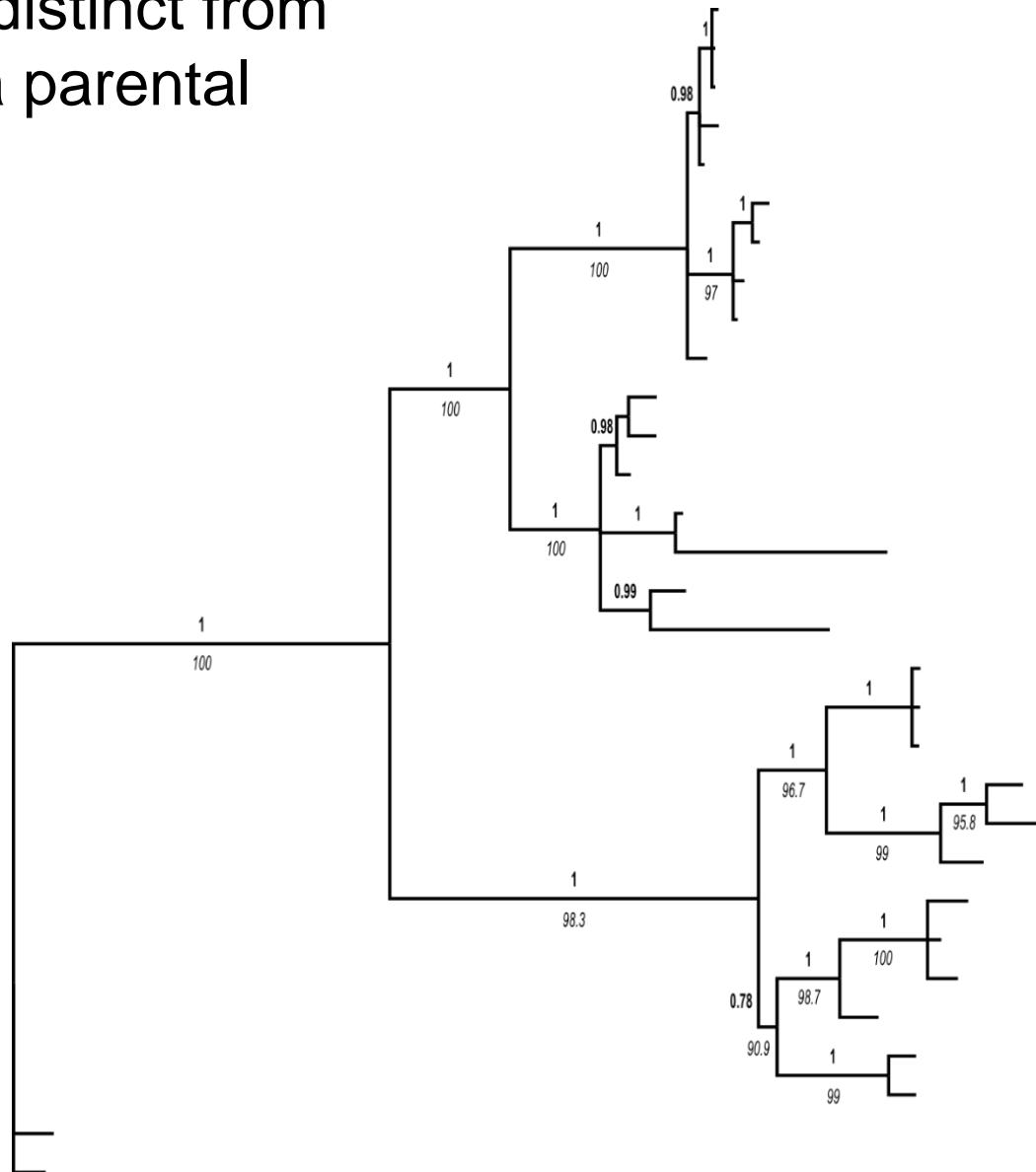
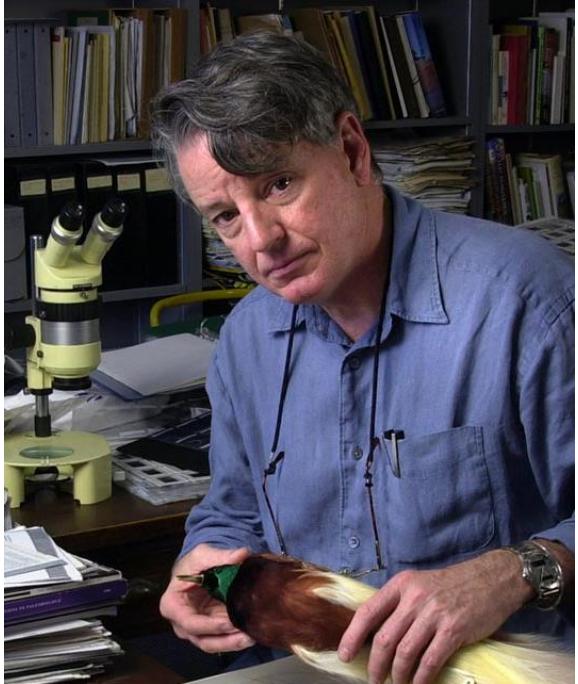
Reciprocal monophyly:

DNA sequence data

“a species is an irreducible cluster of organisms, distinct from other such clusters, and within which there is a parental pattern of ancestry and descendent”

(Eldredge and Cracraft 1980)

(Cracraft, 1983, 1989)



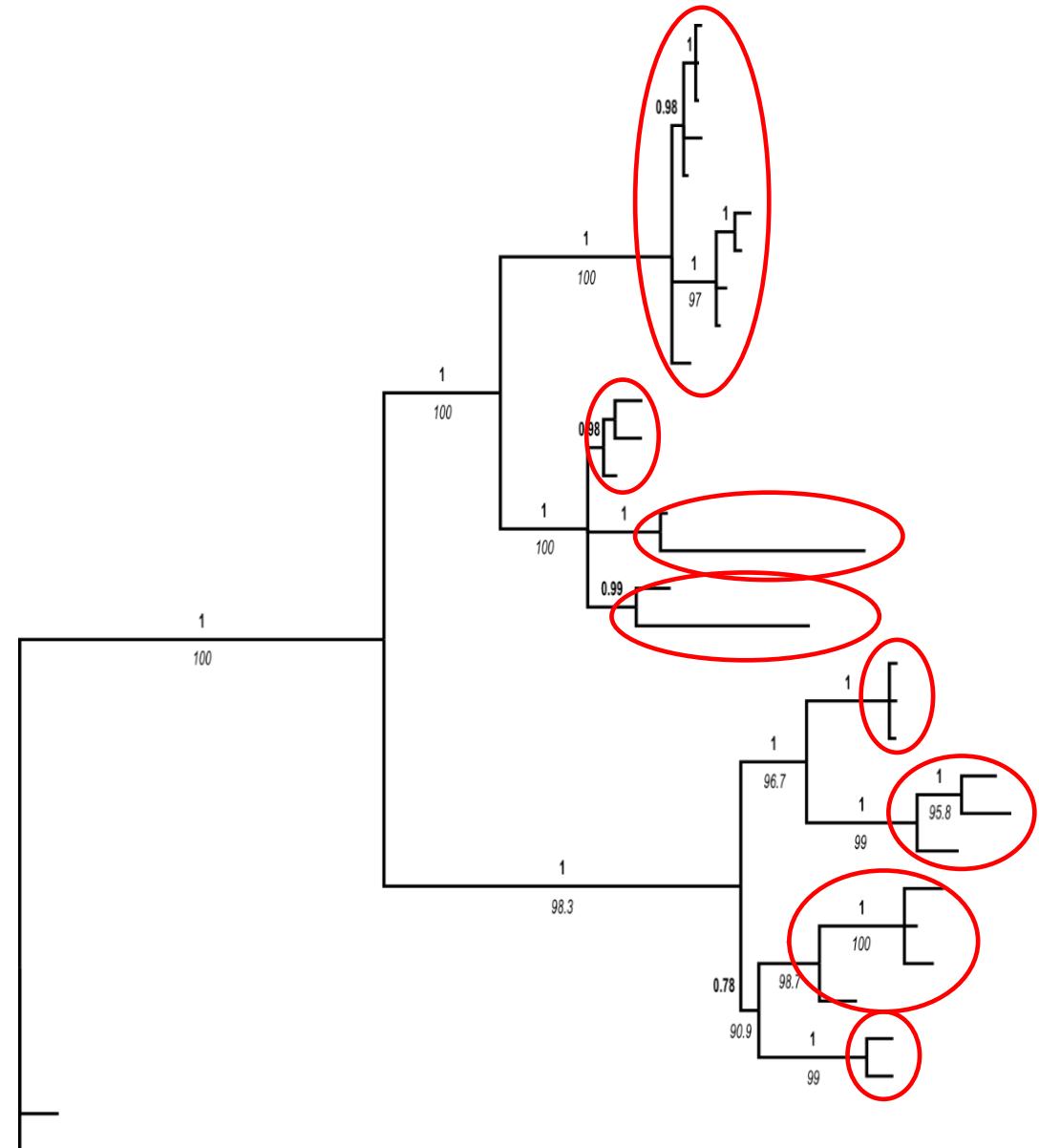
Reciprocal monophly:

DNA sequence data

The smallest diagnosticable monophyletic group of population within which there is a parental pattern of ancestry and descent

Problems:

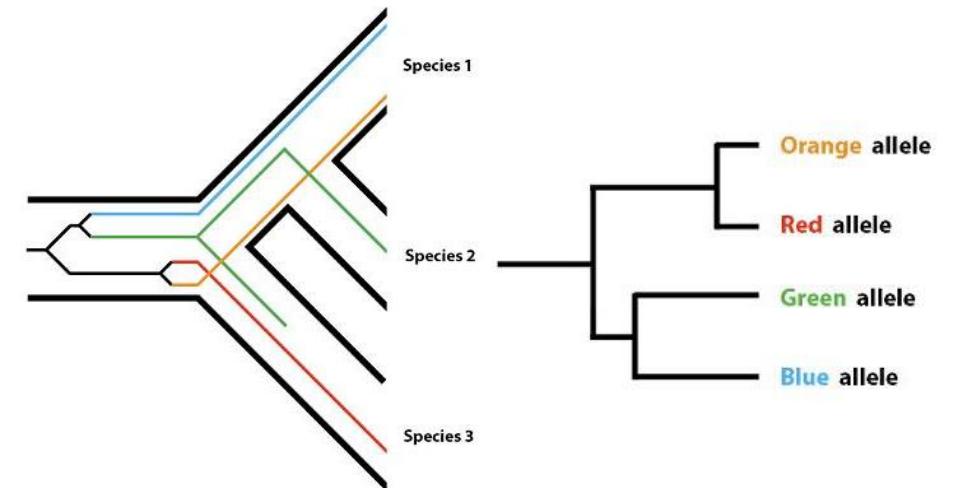
- Gene trees/species tree discordance



Processes causing incongruence among phylogenetic trees:

1. Intra-species stochastic factors:

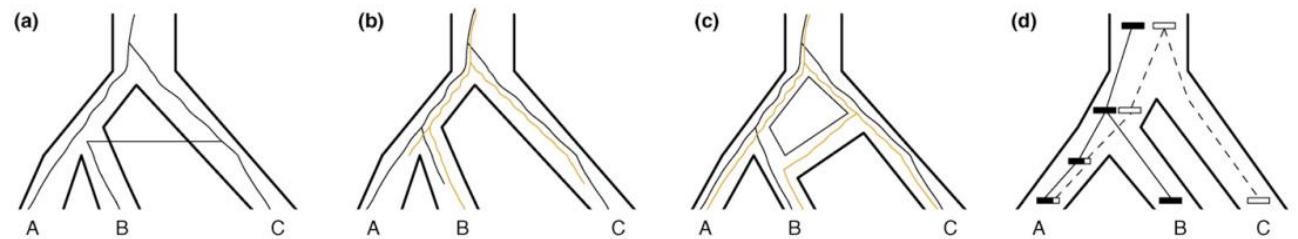
- gene loss / duplication
- incomplete lineage sorting



Source: <http://biologos.org>

2. Inter-species factors:

- horizontal gene transfer
- hybridization



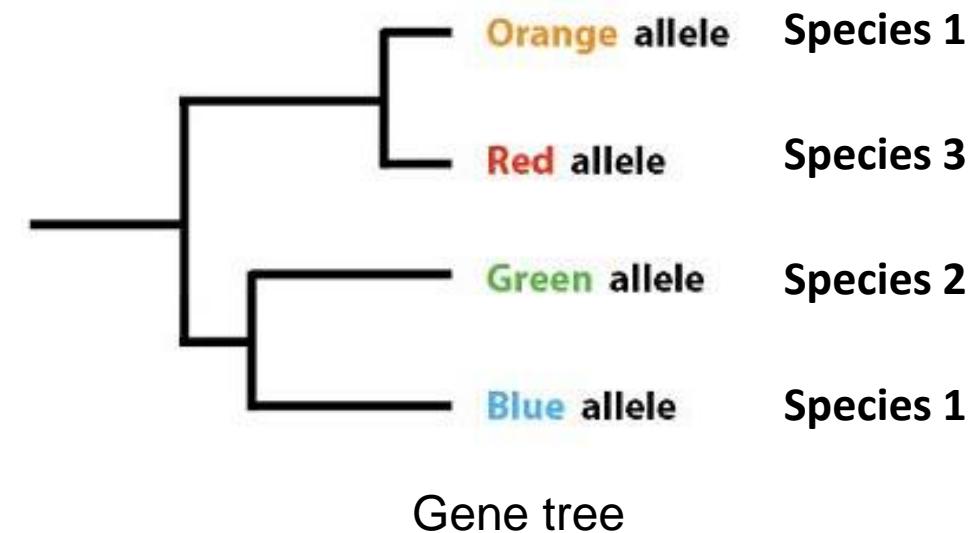
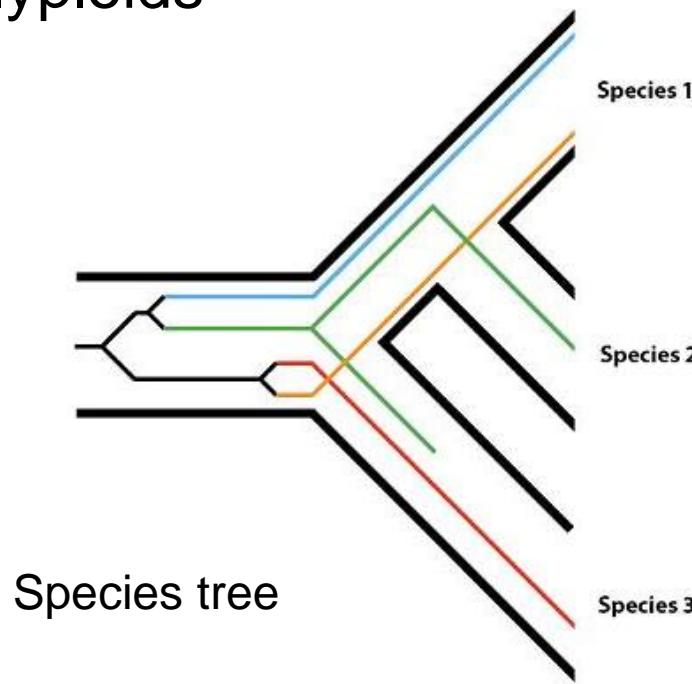
TRENDS in Ecology & Evolution

Source: Degnan & Rosenberg. 2009. *Trends in Ecology and Evolution* 24: 332–340

Processes causing incongruence among phylogenetic trees:

Incomplete lineage sorting particularly prominent in:

- fast radiating groups
- huge ancestral populations size
- polyploids



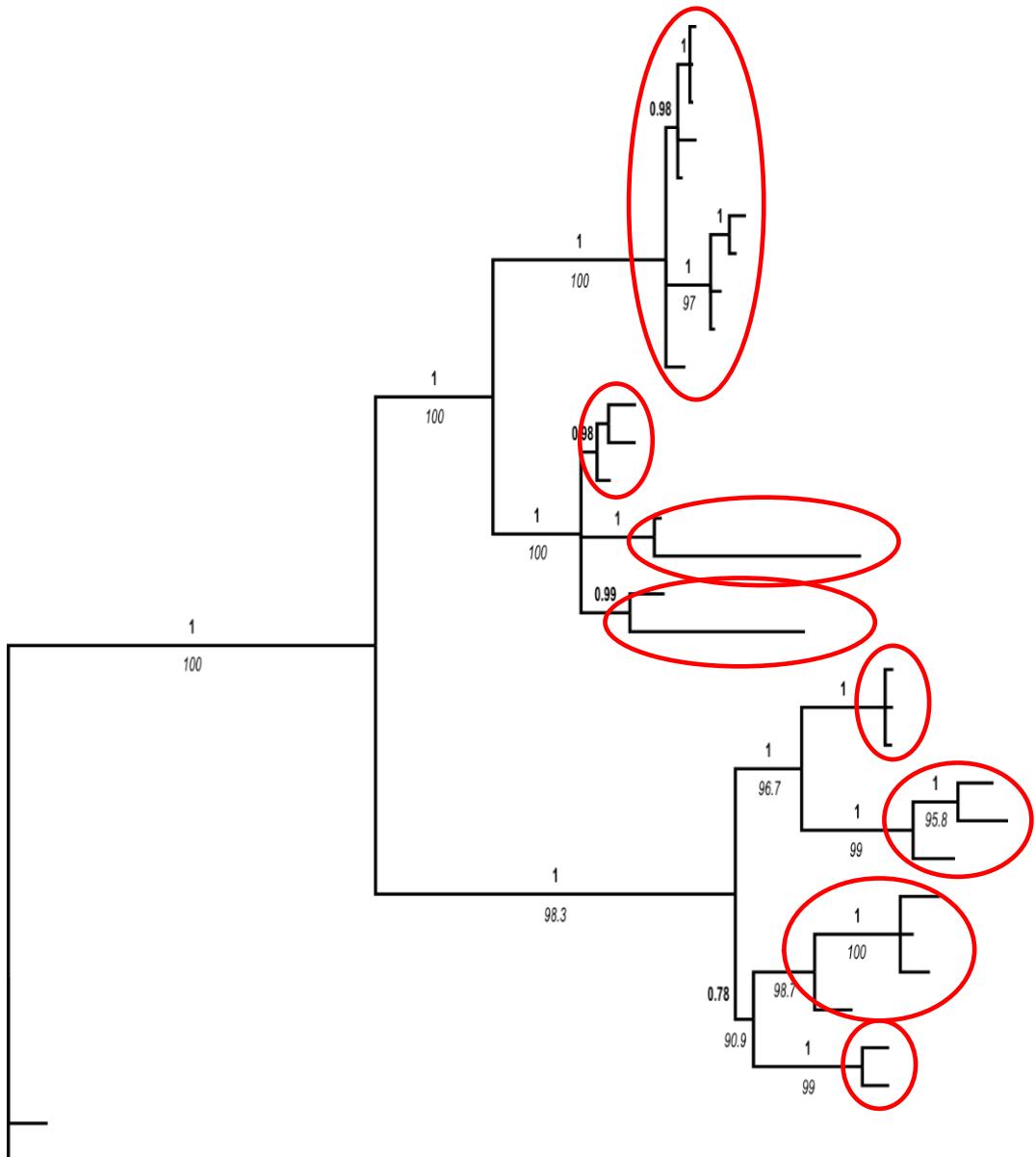
Reciprocal monophyly:

DNA sequence data

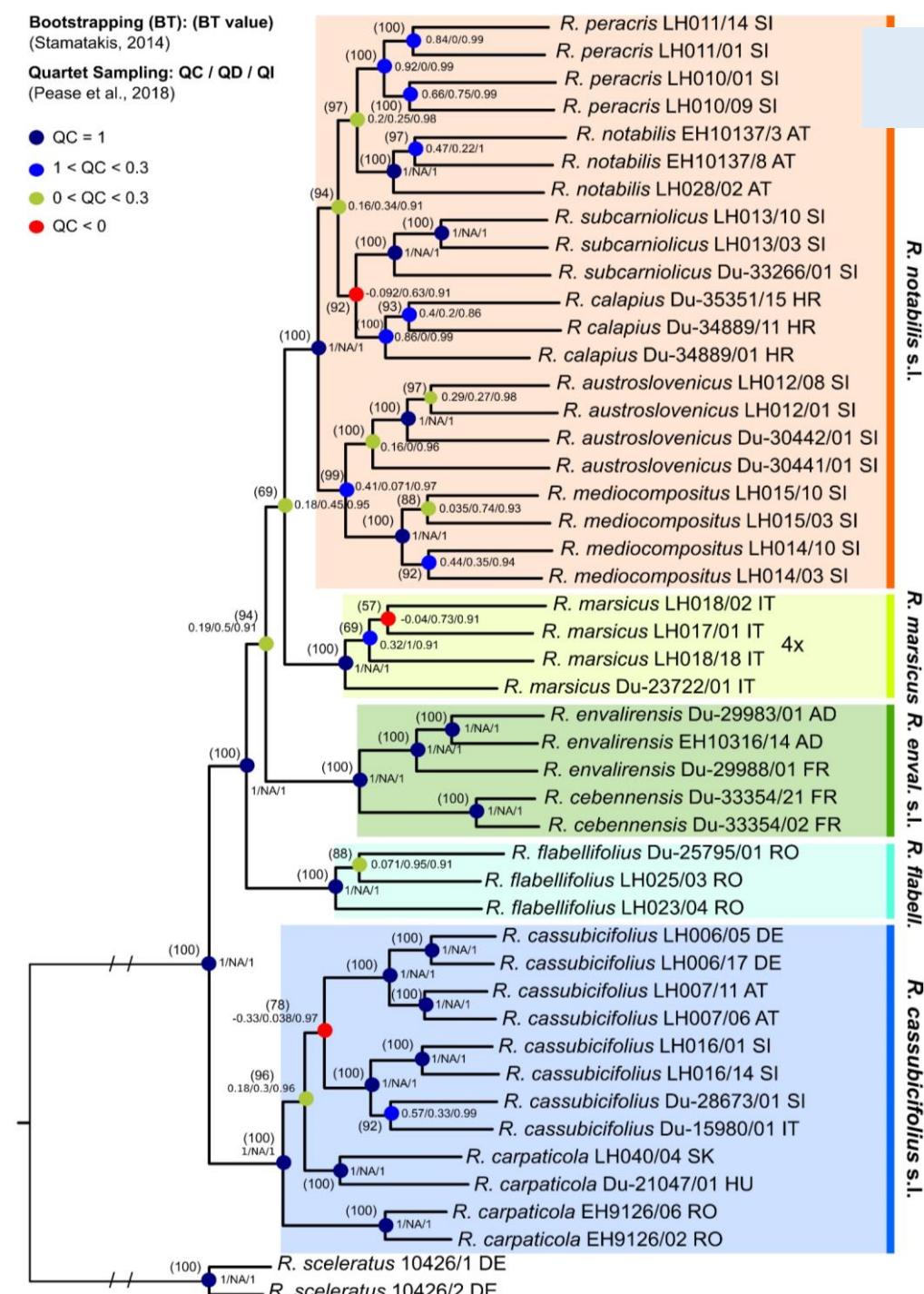
The smallest diagnosable monophyletic group of population within which there is a parental pattern of ancestry and descent

Problems:

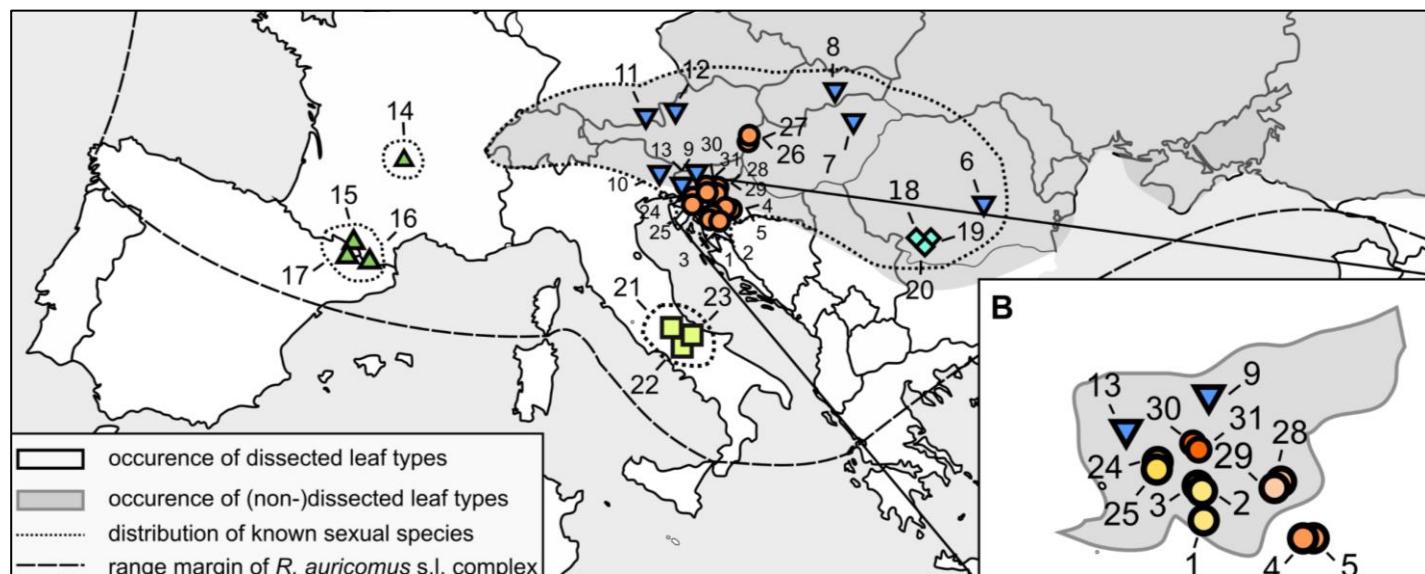
- Gene trees/species tree discordance
 - Sampling (What, How much)



DNA sequence data



Ranunculus auricomus s.l.



Barcodeing Initiatives:

DNA sequence data

Single genomic region and sequence divergence to delimit species



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

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



ALL BIRDS BARCODING INITIATIVE

Single genomic region and sequence divergence to delimit species

Problems:

- Single locus
- Arbitrary threshold



Phil. Trans. R. Soc. B (2005) **360**, 1905–1916
doi:10.1098/rstb.2005.1722
Published online 14 September 2005

The unholy trinity: taxonomy, species delimitation and DNA barcoding

Rob DeSalle*, Mary G. Egan and Mark Siddall

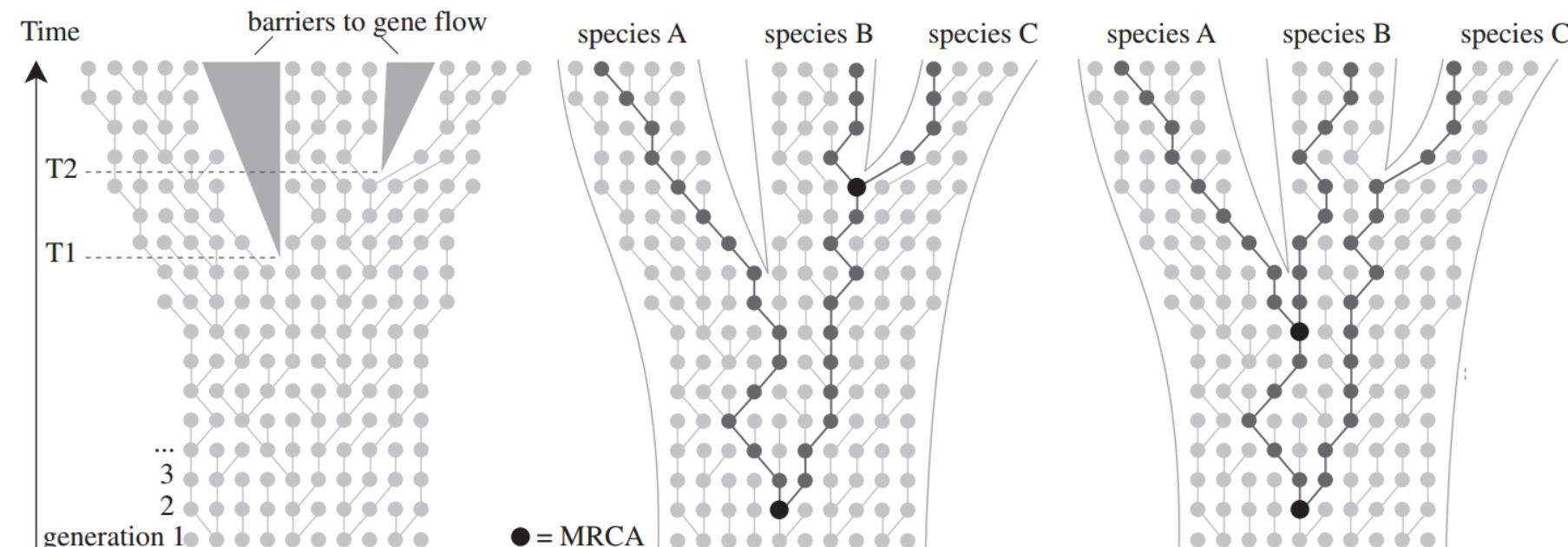
*Division of Invertebrate Zoology, American Museum of Natural History, 79th Street at Central Park West,
New York, NY 10024, USA*

The Multispecies Coalescent (MSC)

“The mathematical and probabilistic theory underlying the evolutionary history of alleles”

The probability of two alleles to coalesce back in time is function of:

- Time (τ)
- Population size (N)



The Multispecies Coalescent (MSC)

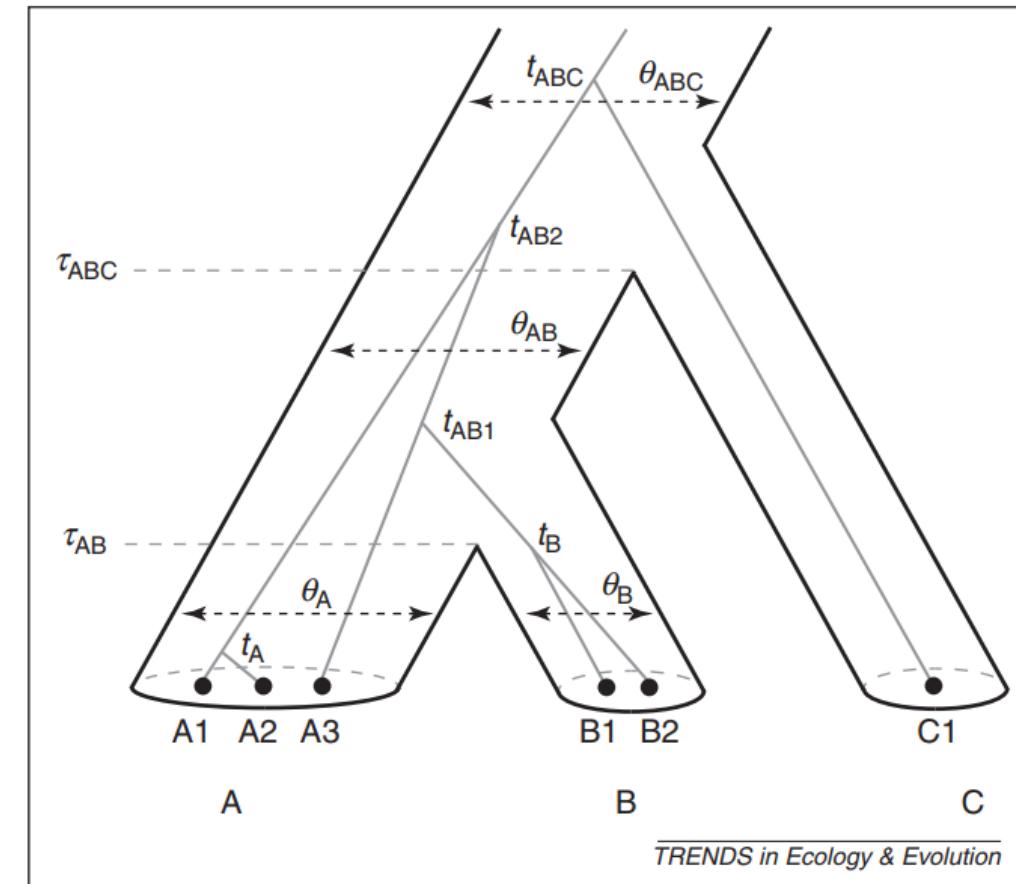
“The mathematical and probabilistic theory underlying the evolutionary history of alleles”

- Give a mean of calculating gene trees probability for a given species tree
- Can accommodate gene tree discordances
- Estimation of species divergence times
- Species delimitation

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

Diagram illustrating the components of the Multispecies Coalescent (MSC) formula:

- Prior distribution of species phylogenies
- Prior distribution of delimitation models
- likelihood of multilocus data given a species tree



From: Fujita et al. 2012. *Trends in Ecology and Evolution* 27:480-488

When are molecular approaches for species delimitation useful?

- Absence of other taxonomically relevant characters
- Recently diverging groups
- Unexplored areas with high diversity
- Cryptic species

462

Review

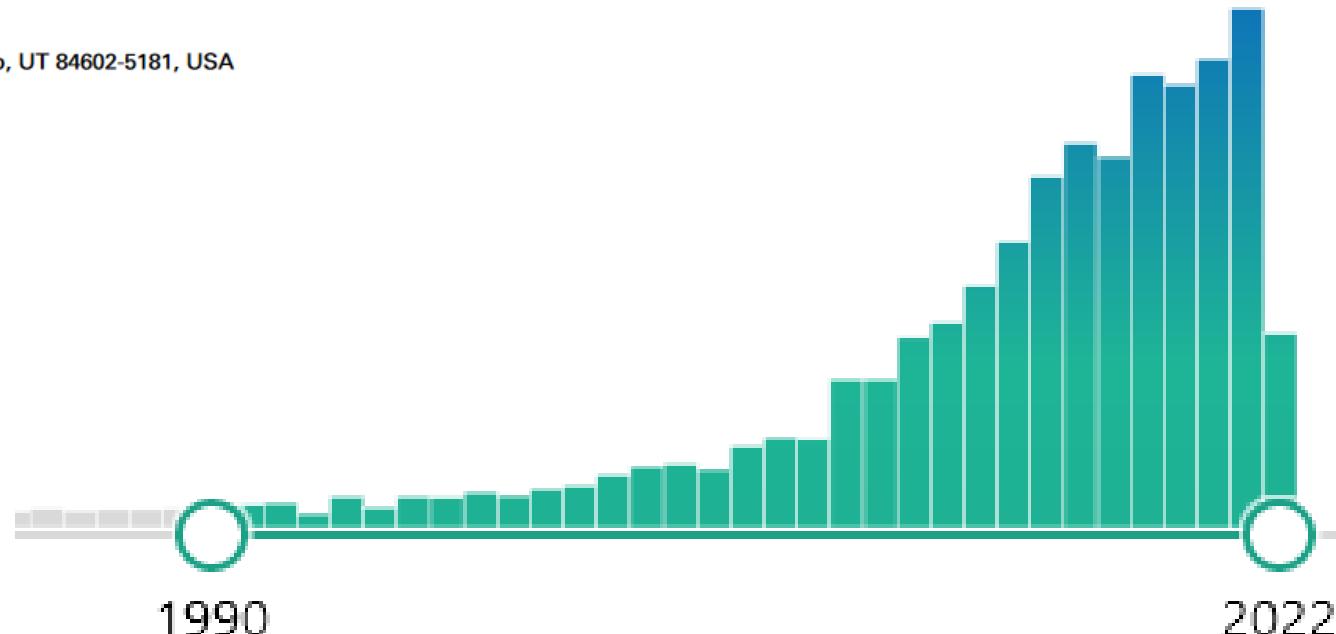
TRENDS in Ecology and Evolution Vol.18 No.9 September 2003



Delimiting species: a Renaissance issue in systematic biology

Jack W. Sites Jr and Jonathon C. Marshall

Department of Integrative Biology and M.L. Bean Life Science Museum, Brigham Young University, Provo, UT 84602-5181, USA



Source: <https://pubmed.ncbi.nlm.nih.gov/?term=species+delimitation>

462

Review

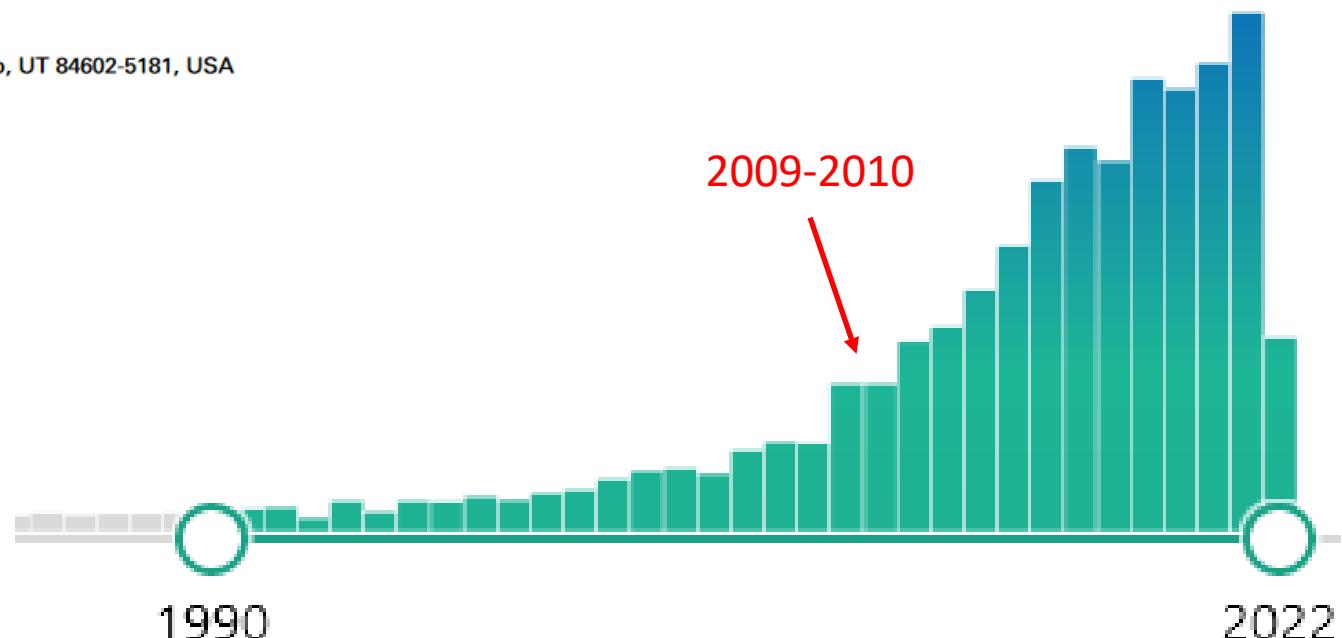
TRENDS in Ecology and Evolution Vol.18 No.9 September 2003



Delimiting species: a Renaissance issue in systematic biology

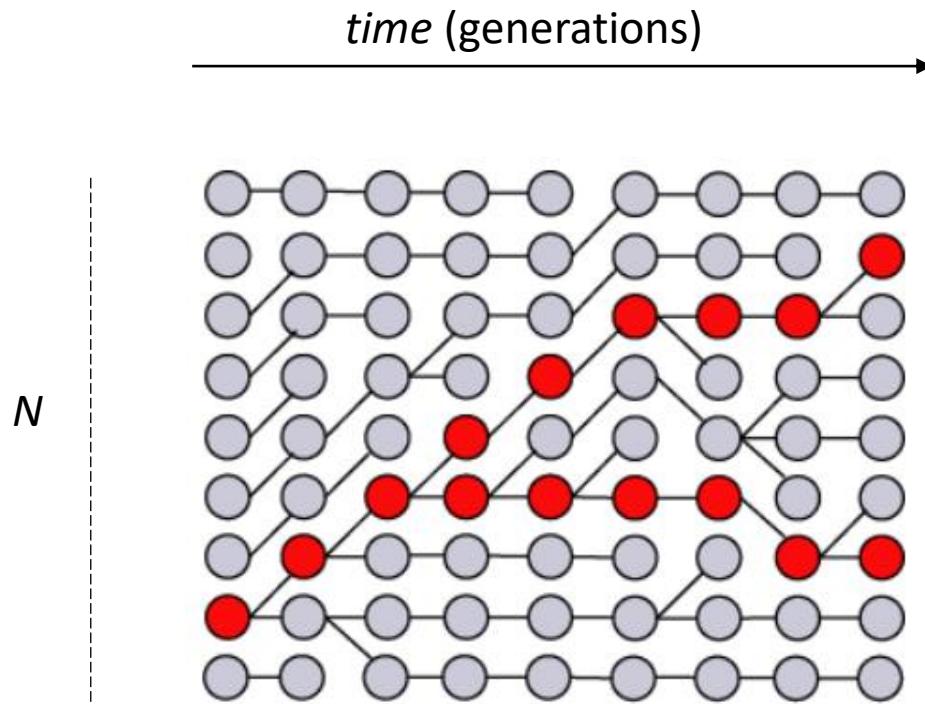
Jack W. Sites Jr and Jonathon C. Marshall

Department of Integrative Biology and M.L. Bean Life Science Museum, Brigham Young University, Provo, UT 84602-5181, USA



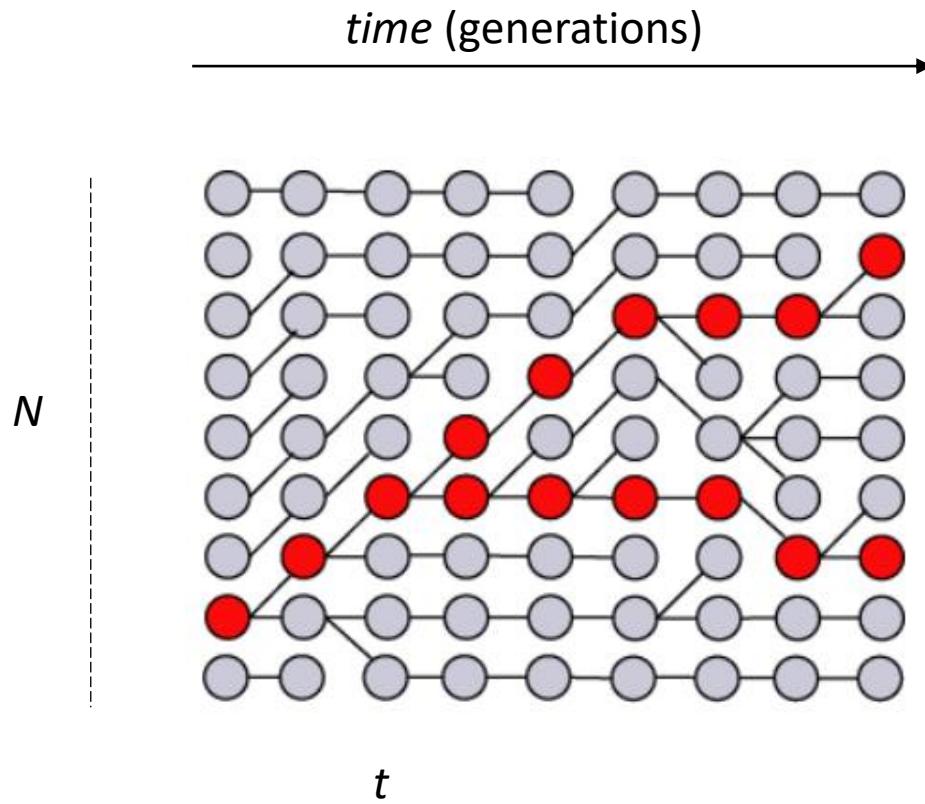
Source: <https://pubmed.ncbi.nlm.nih.gov/?term=species+delimitation>

The Coalescent Model (Kingman 1982)



- Non-overlapping generations
- Constant N
- Random mating
- no migration

The Coalescent Model



Probability of two alleles to coalesce after one generation

$$\frac{1}{N}$$

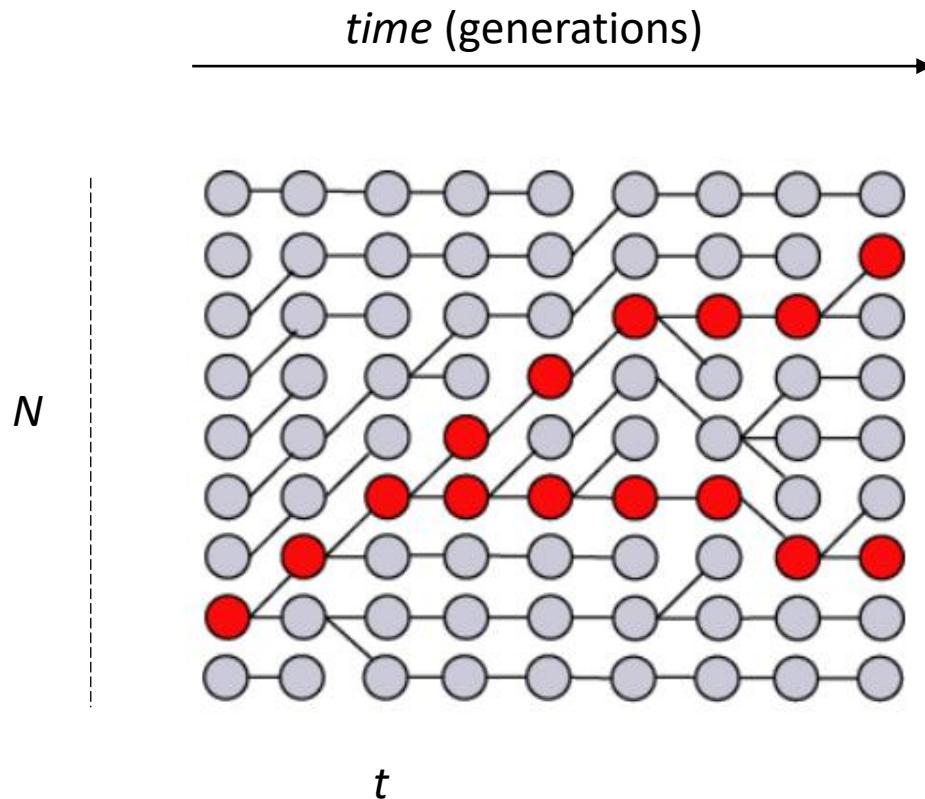
Probability of two alleles not to coalesce

$$1 - \frac{1}{N}$$

Probability of two alleles to coalesce at generation t

$$\frac{1}{N} \left(1 - \frac{1}{N}\right)^{t-1}$$

The Coalescent Model

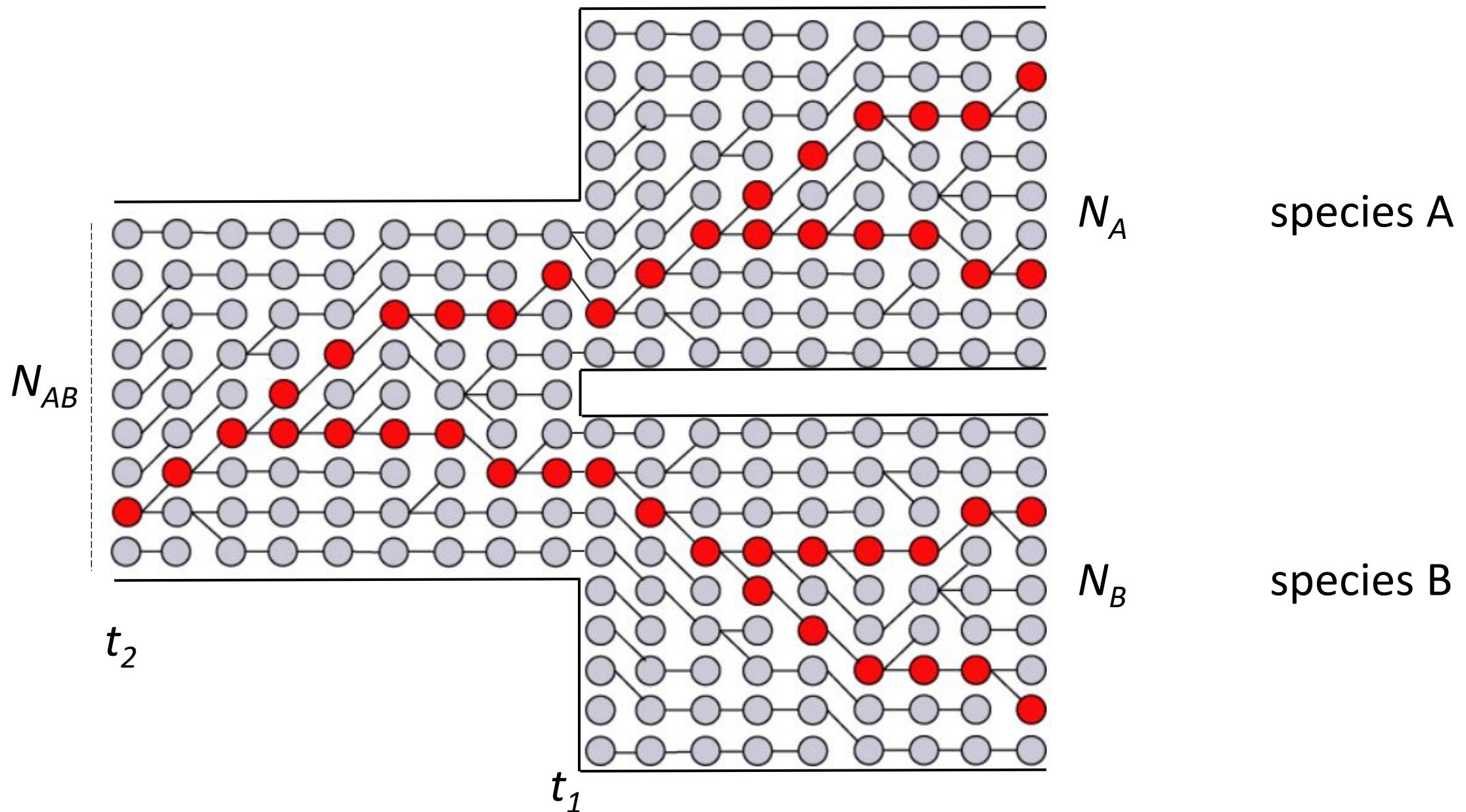


$$\frac{1}{N} \left(1 - \frac{1}{N}\right)^{t-1}$$

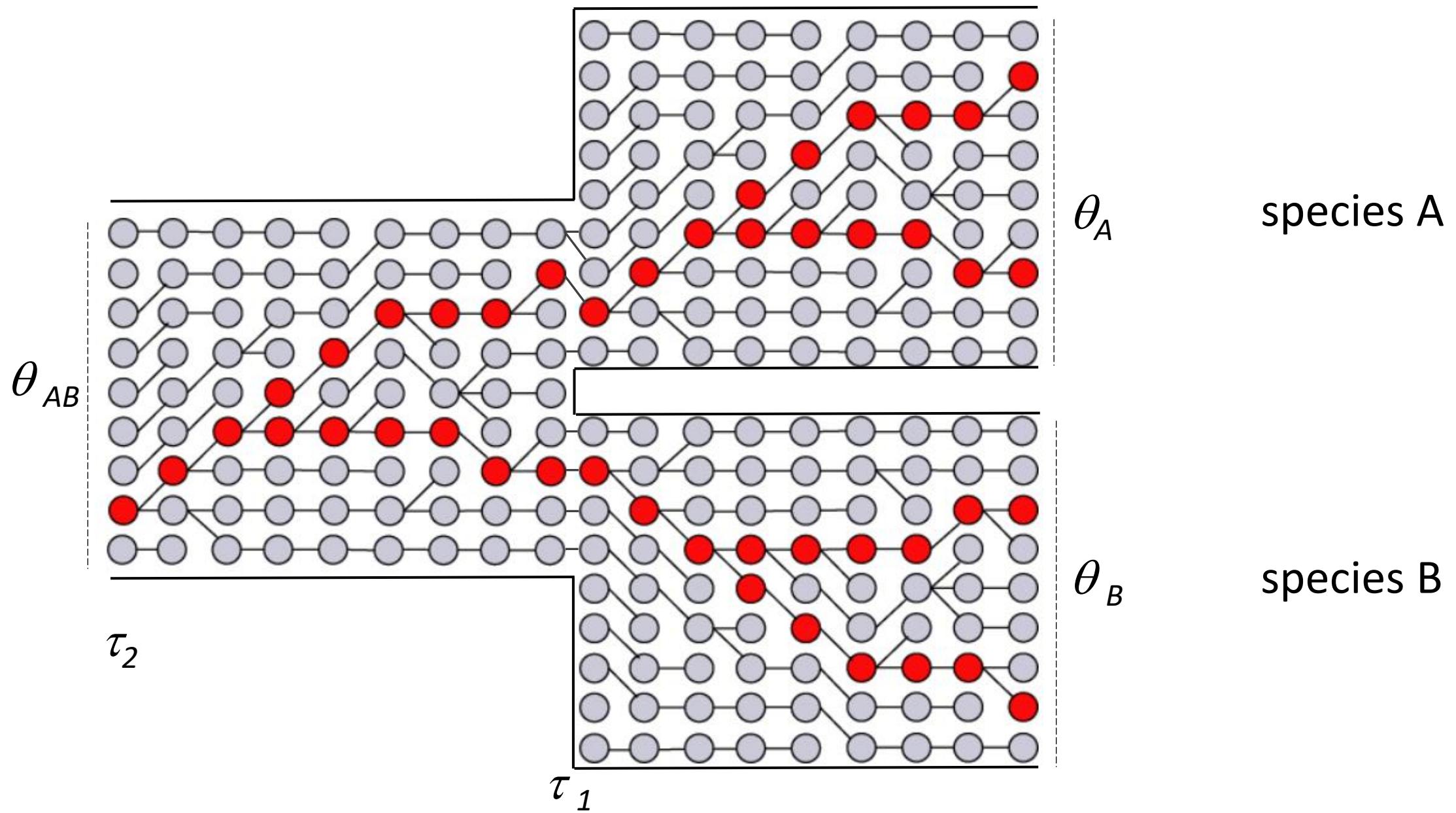
Mean = N

Mean time of the coalescent event!

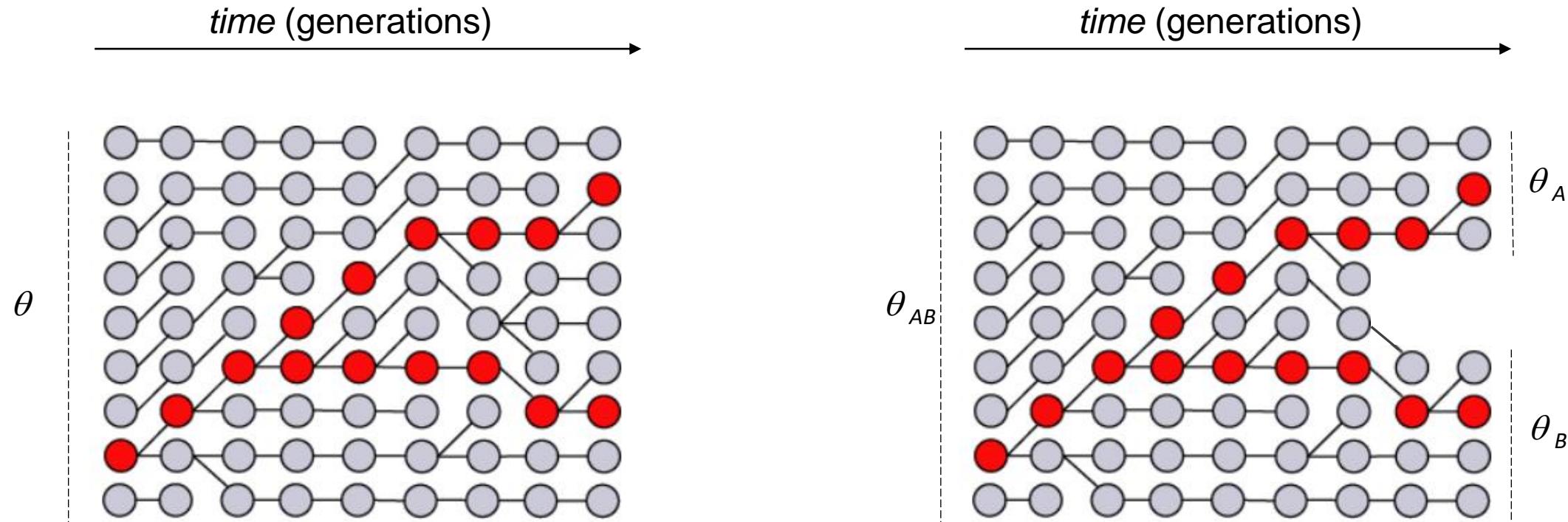
The Multispecies Coalescent (MSC)



The Multispecies Coalescent (MSC)



Coalescent-based species delimitation



One or two species?

- **Discovery approaches:** no prior information on species delimitation
- **Validation approaches:** test a few species delimitation hypotheses
- **Heuristic methods:** use summary statistics, simplification of the formal model
- **Parametric methods:** based on explicit probabilistic models (ML; Bayesian)

Discovery approaches	Validation approaches	Analytical Framework	Input
GMYC		Best-fit tree branching model (Yule vs Coalescent)	Ultrametric gene tree(s)
	spedeSTEM	Maximum Likelihood and Information theory	Sequence alignments and group membership
	Marginal Likelihood	Marginal Likelihood comparison of models	Sequence alignments and group membership
BP&P		Bayesian and/or reversible-jump MCMC	Sequence alignments
DISSECT		Bayesian. Implemented in BEAST 1.8	Sequence alignments
STACEY		Bayesian. Implemented in BEAST 2	Sequence alignments
SPEEDEMON		Bayesian. Implemented in BEAST 2	Sequence alignments
Delineate			Sequence alignments

Bayes' formula:

$$P(H_a|D) = \frac{P(H_a) \cdot P(D|H_a)}{\sum_i P(H_i) \cdot P(D|H_i)}$$

prior probability *likelihood*
posterior probability *marginal likelihood*

The diagram illustrates the components of Bayes' formula. At the top, 'prior probability' and 'likelihood' are shown with blue arrows pointing downwards towards the formula. Below the formula, 'posterior probability' is shown with a blue arrow pointing upwards from the left, and 'marginal likelihood' is shown with a blue arrow pointing upwards from the right.

Bayes' formula:

$$P(H_a|D) = \frac{P(H_a) \cdot P(D|H_a)}{\sum_i P(H_i) \cdot P(D|H_i)}$$

prior probability *likelihood*
posterior probability *marginal likelihood*

Methods:

- Path sampling
- Stepping stones
- Nested sampling



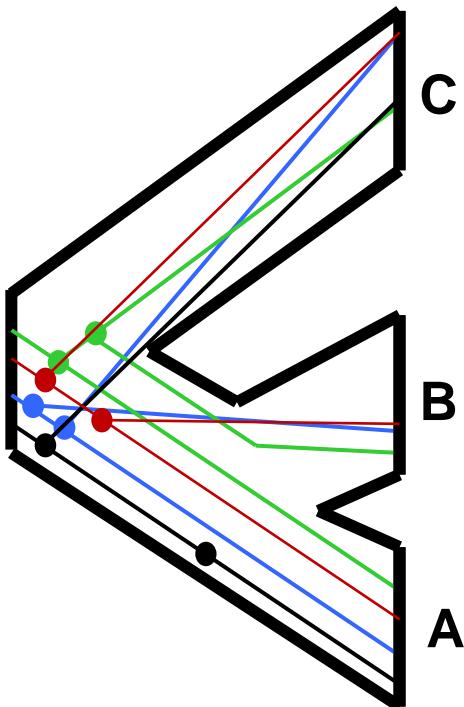
Beast2

Bayesian evolutionary analysis by sampling trees

Marginal Likelihood

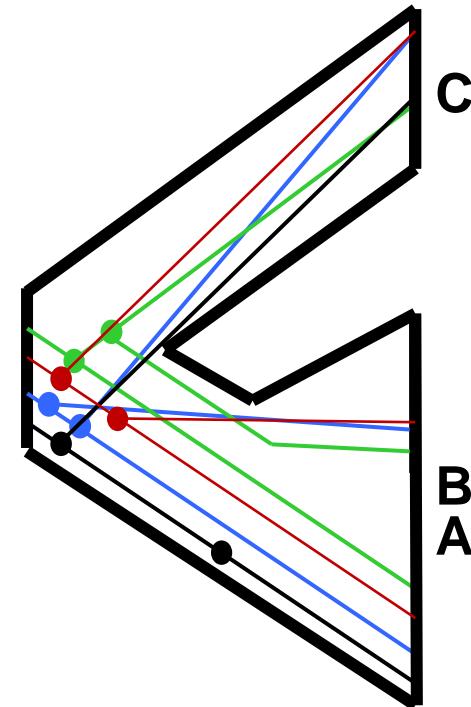
Validation approaches

$$\sum_i P(H_i) \cdot P(D|H_i)$$



3 species model

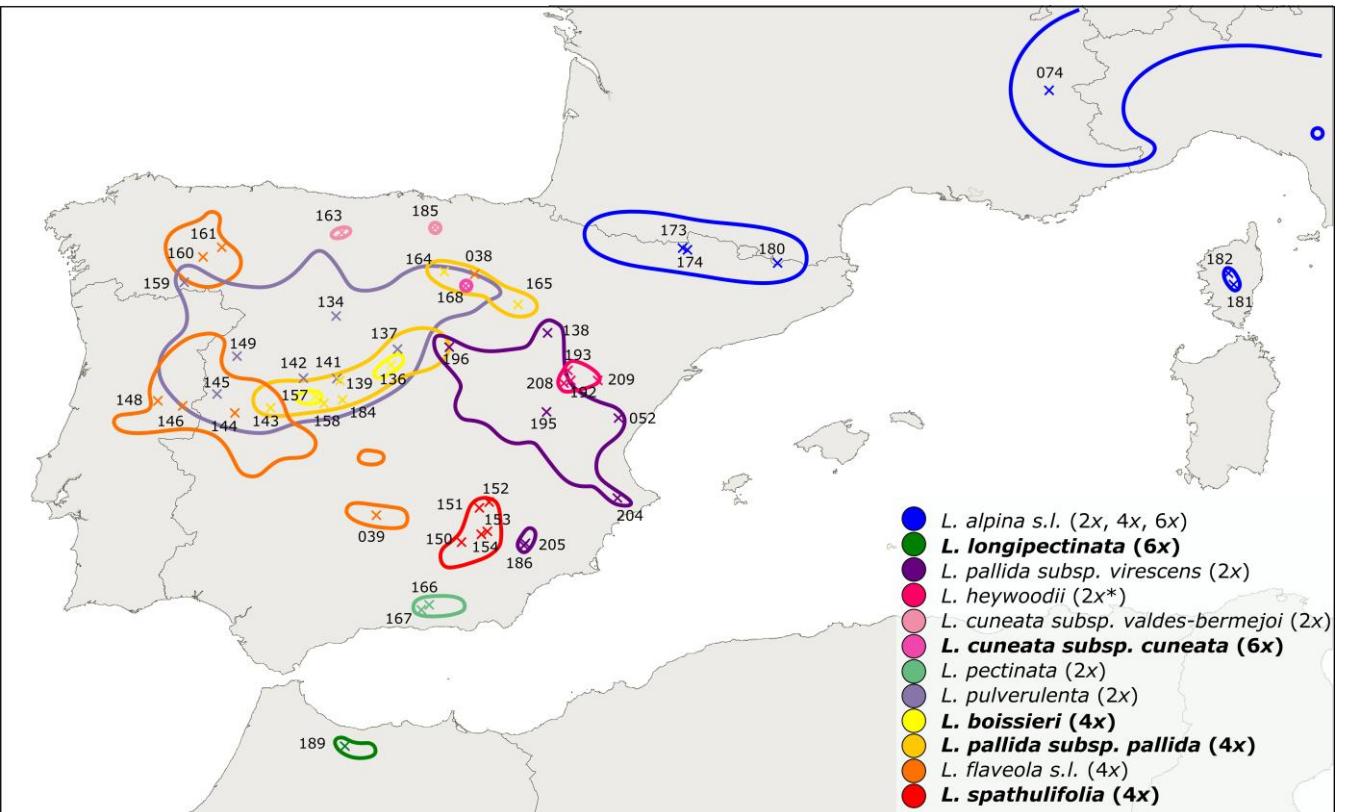
$$\sum_i P(H_i) \cdot P(D|H_i)$$



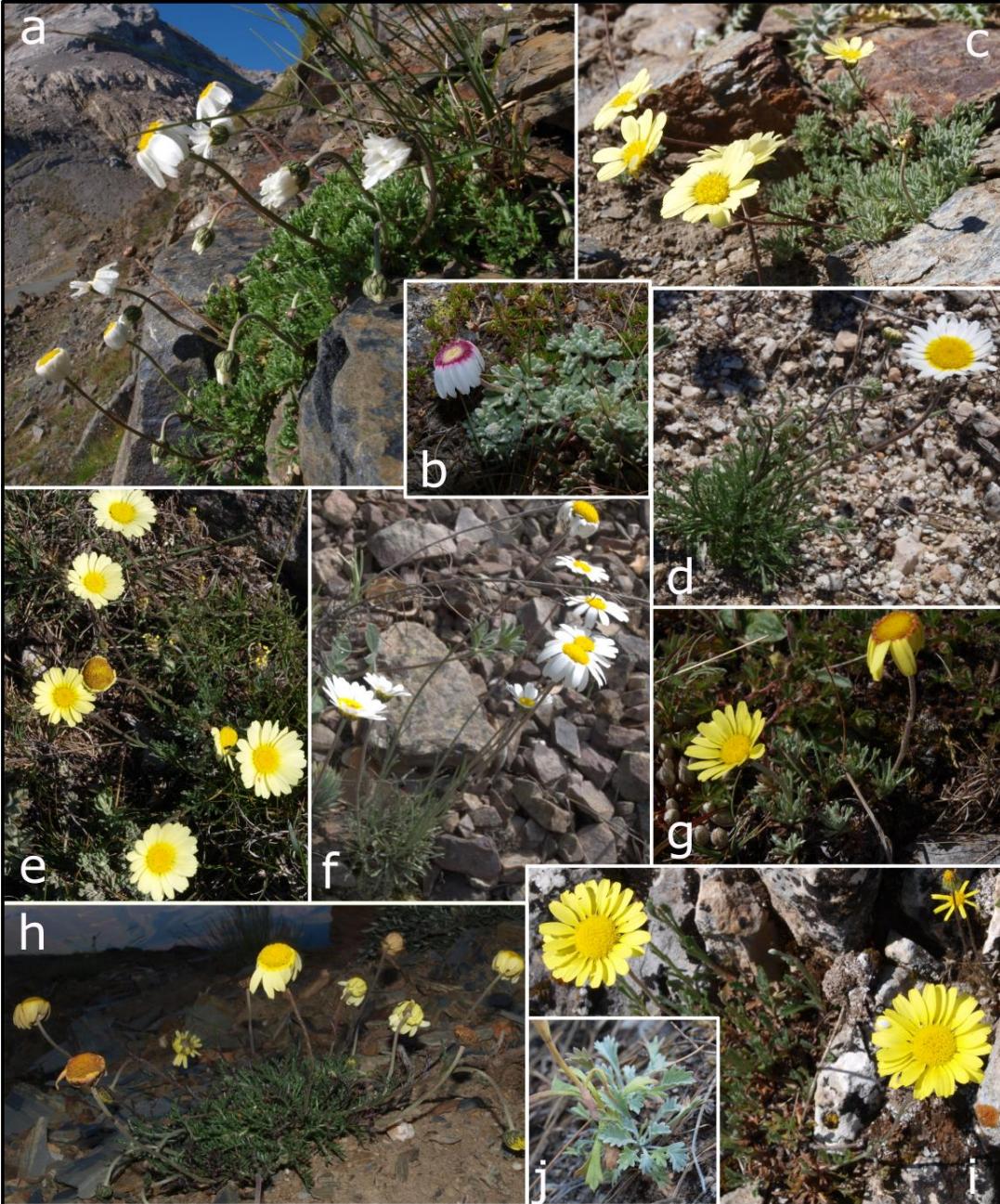
2 species model

Leucanthemopsis 4x species delimitation

- 5-10 recognized species
- Several infraspecific taxa
- Mostly western Mediterranean mountain species

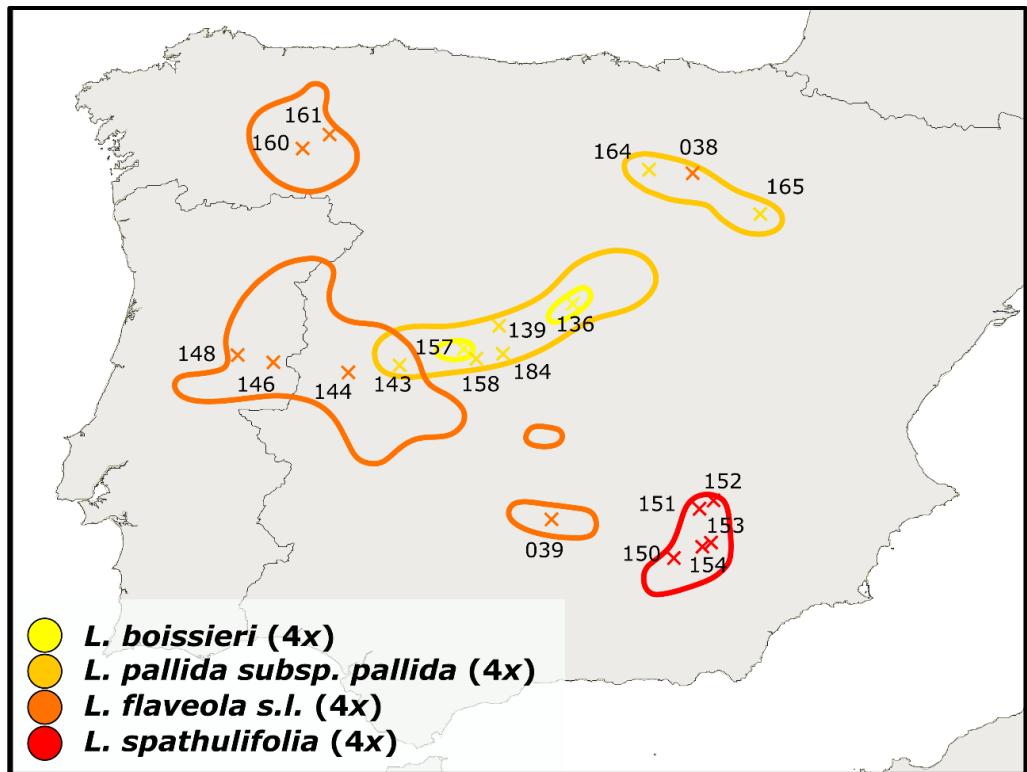


Marginal Likelihood



Leucanthemopsis 4x species delimitation

- Iberian tetraploid



L. flaveola



L. pallida subsp. *pallida*



L. boissieri

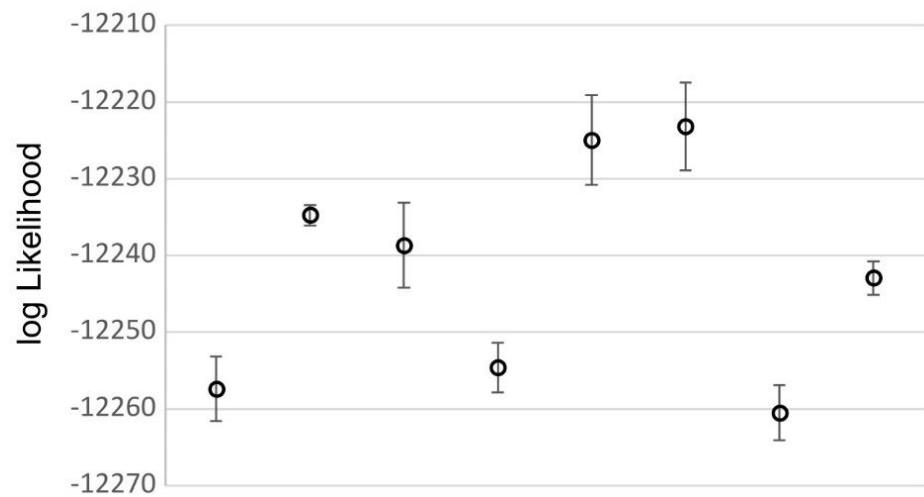
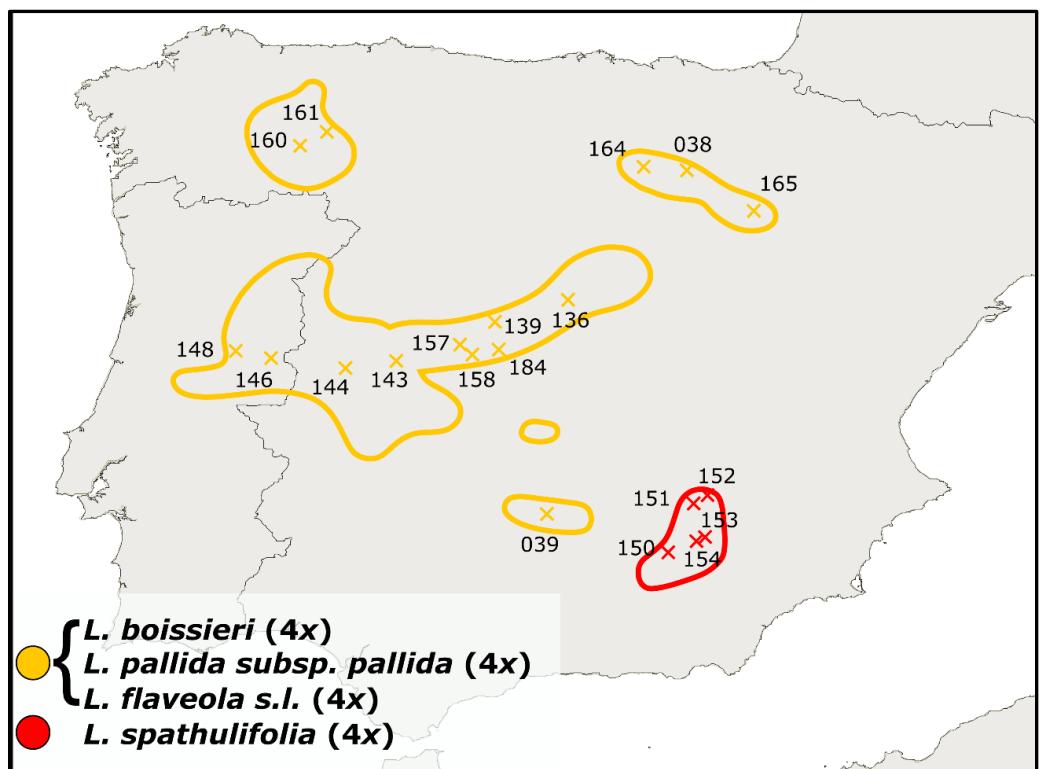


L. spathulifolia

Leucanthemopsis 4x species delimitation

Marginal Likelihood

Taxon identification		Scenarios							
Heywood (1975)	Pedrol (2019)	1	2	3	4	5	6	7	8
<i>L. flaveola</i>	<i>L. flaveola</i> subsp. <i>flaveola</i>								
<i>L. flaveola</i>	<i>L. flaveola</i> subsp. <i>ricoi</i>								
<i>L. pallida</i> subsp. <i>pallida</i> var. <i>pallida</i>	<i>L. pallida</i> subsp. <i>pallida</i>								
<i>L. pallida</i> subsp. <i>pallida</i> var. <i>alpina</i>	<i>L. boissieri</i>								
<i>L. pallida</i> subsp. <i>spathulifolia</i>	<i>L. spathulifolia</i>								



When they may fail:

Coalescent-based species delimitation

- Presence of gene flow
- Selectively driven divergence

Integrative approaches always recommended!!

MOLECULAR ECOLOGY

Molecular Ecology (2013) 22, 4369–4383

doi: 10.1111/mec.12413

INVITED REVIEWS AND META-ANALYSES How to fail at species delimitation

BRYAN C. CARSTENS,* TARA A. PELLETIER,* NOAH M. REID† and JORDAN D. SATLER*

*Department of Evolution, Ecology and Organismal Biology, The Ohio State University, 318 W. 12th Avenue, Columbus, OH 43210-1293, USA, †Department of Biological Sciences, Louisiana State University, Life Sciences Building, Baton Rouge, LA 70803, USA



Syst. Biol. 69(1):184–193, 2020
© The Author(s) 2019. Published by Oxford University Press, on behalf of the Society of Systematic Biologists. All rights reserved.
For permissions, please email: journals.permissions@oup.com
DOI:10.1093/sysbio/syz042
Advance Access publication June 10, 2019

Multispecies coalescent delimits structure, not species

Jeet Sukumaran^{a,1} and L. Lacey Knowles^{a,1}

^aDepartment 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)

The multispecies coalescent model underlies many approaches used for species delimitation. In previous work assessing the performance of species delimitation under this model, speciation was treated as an instantaneous event rather than as an extended process involving distinct phases of speciation initiation (structuring) and completion. Here, we use data under simulations

consequence, the increased resolution of genomic data makes it possible to not only detect divergent species lineages, but also local population structure within them—that is, a fractal hierarchy of divergences.

Misidentification of population structure as putative species is therefore emerging as a key issue (8) that has received insuf-

The Multispecies Coalescent Over-Splits Species in the Case of Geographically Widespread Taxa

E. ANNE CHAMBERS* AND DAVID M. HILLIS

Department of Integrative Biology and Biodiversity Center, The University of Texas at Austin, Austin, TX 78712, USA
**Correspondence to be sent to: Department of Integrative Biology and Biodiversity Center, The University of Texas at Austin, 2415 Speedway #C0930, Austin, TX 78712, USA; E-mail: eacchambers@utexas.edu.*

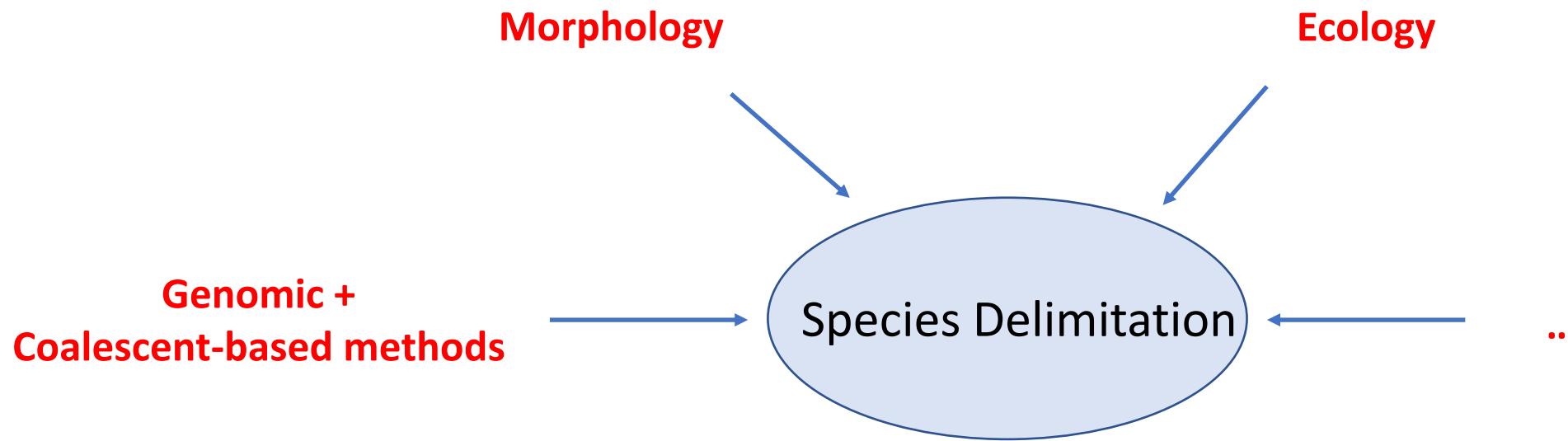
Received 19 November 2018; reviews returned 22 May 2019; accepted 24 May 2019
Associate Editor: Richard Glor

When they may fail:

Coalescent-based species delimitation

- Presence of gene flow
- Selectively driven divergence

Integrative approaches always recommended!!



PROs

- Probabilistic framework
- Objective way of define species based on sequence data

CONs

- Strong model assumptions
 - No interspecific gene flow (hybridization? polyploidy?)
 - Random paring (asexuality, selfing)

Coffe?



Before starting...

- **What is a phylogeny?**
- **What is a phylogenetic tree?**
- **What is an alignment?**



520. *Xanthium Strumarium* L.

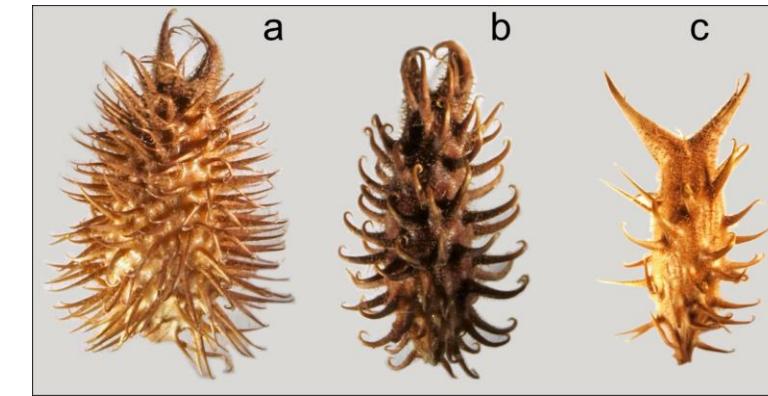
Burweed; G.

Species delimitation *Xanthium* L.

Why Xanthium?



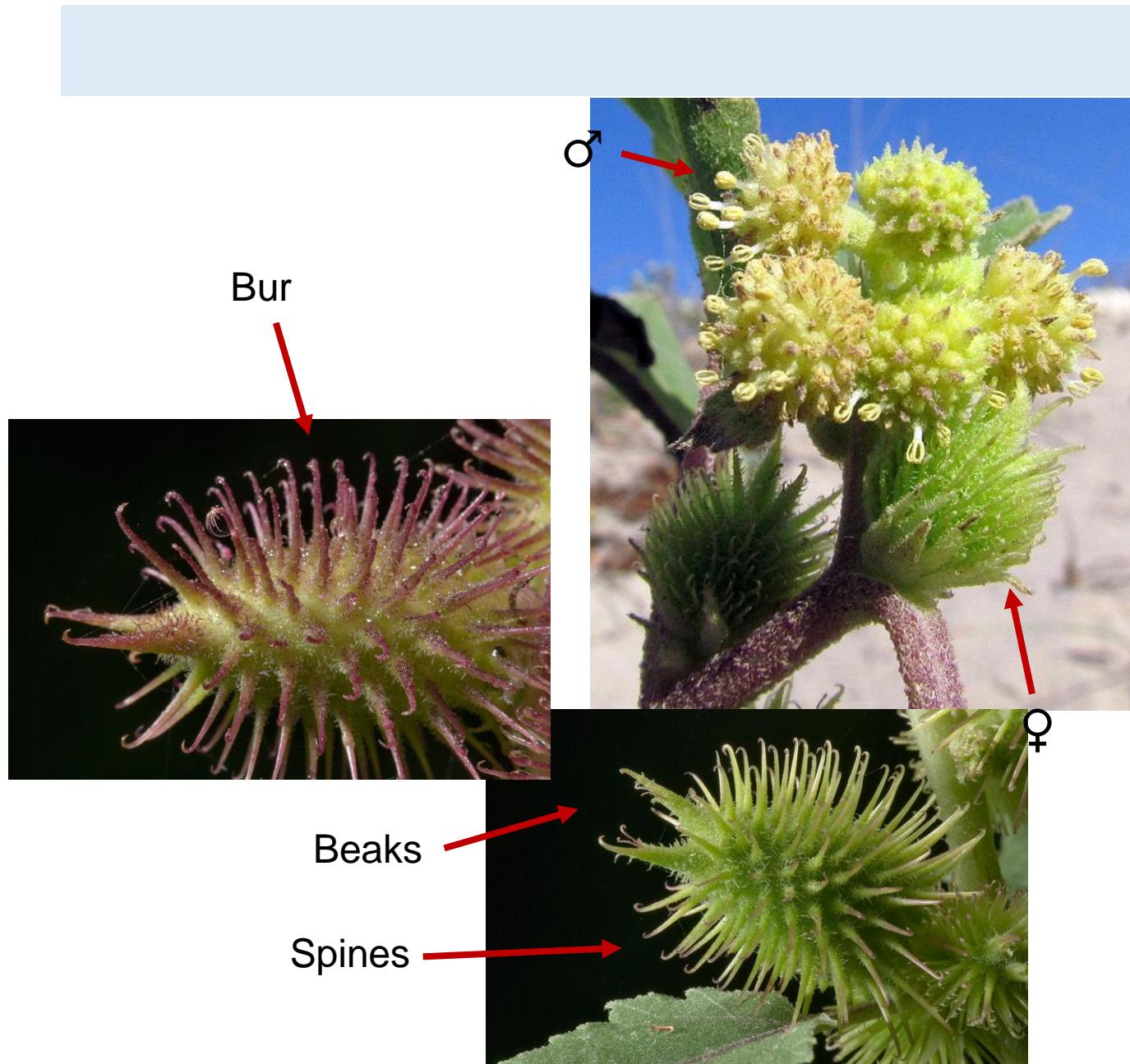
- Morphologically intricate genus of Asteraceae
- Peculiar spiny, wind-pollinated female capitula
- Weed on human crops
- Nowadays cosmopolitan (human-mediated dispersal)





520. *Xanthium Strumarium* L.

Burweed; G.



Fotos by: Alexey Sergeev;
<http://www.asergeev.com/index.htm>

Sectio Acanthoxanthium

- Nodal spines (3-lobed)
- Leaves lanceolate-ovate
- Fruits ~small



Sectio Xanthium

- Stems unarmed
- Leaves suborbiculate
- Fruits small or big



Sectio **Euxanthium**

Subsectio **Orthorrhyncha**

- Beaks strait



Subsectio **Campylorrhyncha**

- Beaks hooked



Serie Glabrata
(Glabrous burs)

Serie Hispida
(Burs with hairs and glands)



Contents lists available at ScienceDirect

ELSEVIER

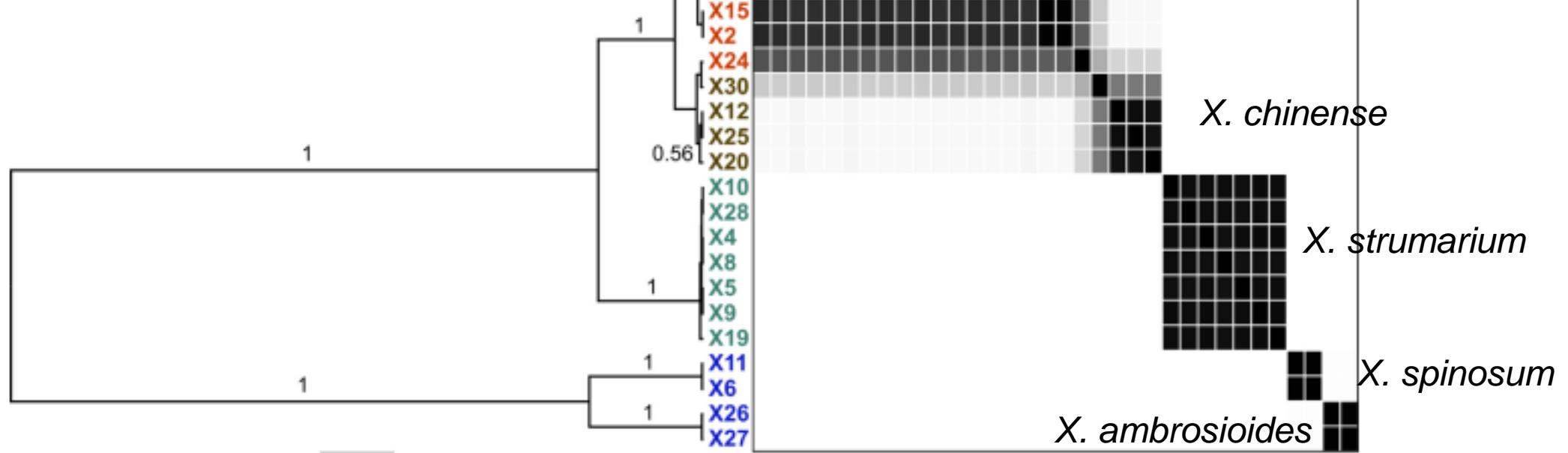
Molecular Phylogenetics and Evolution

journal homepage: www.elsevier.com/locate/ympev



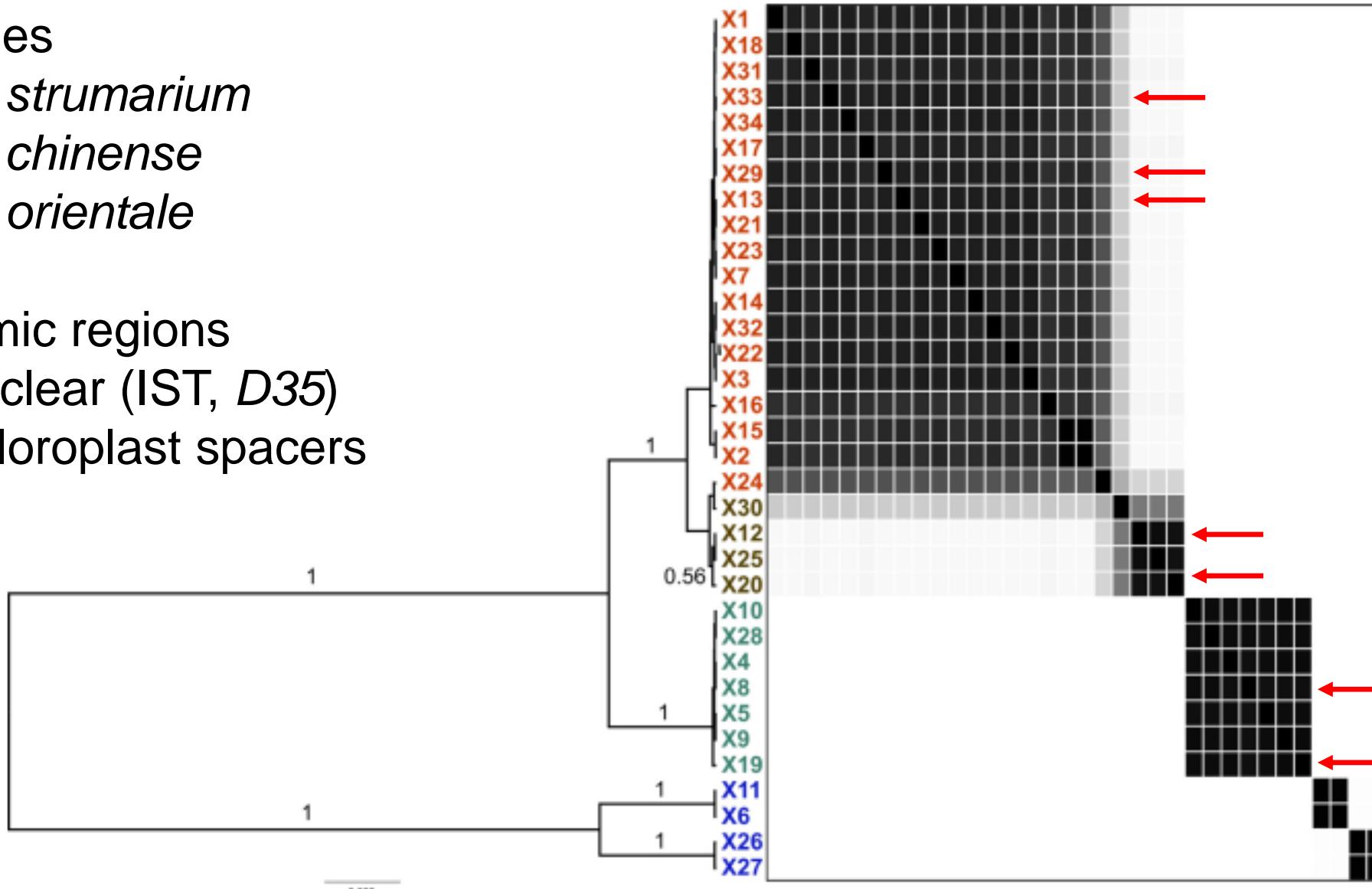
How many names for a beloved genus? – Coalescent-based species delimitation in *Xanthium* L. (Ambrosiinae, Asteraceae)

Salvatore Tomasello



Samples fot today:

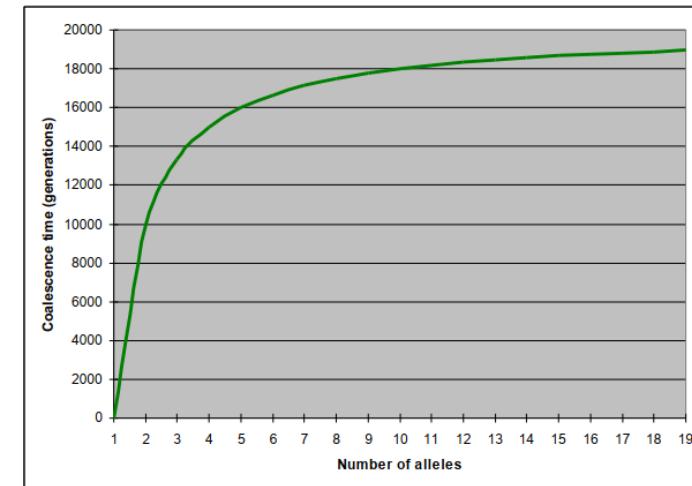
- 7 samples
 - 2 *X. strumarium*
 - 2 *X. chinense*
 - 3 *X. orientale*
- 4 genomic regions
 - 2 nuclear (IST, *D35*)
 - 2 chloroplast spacers



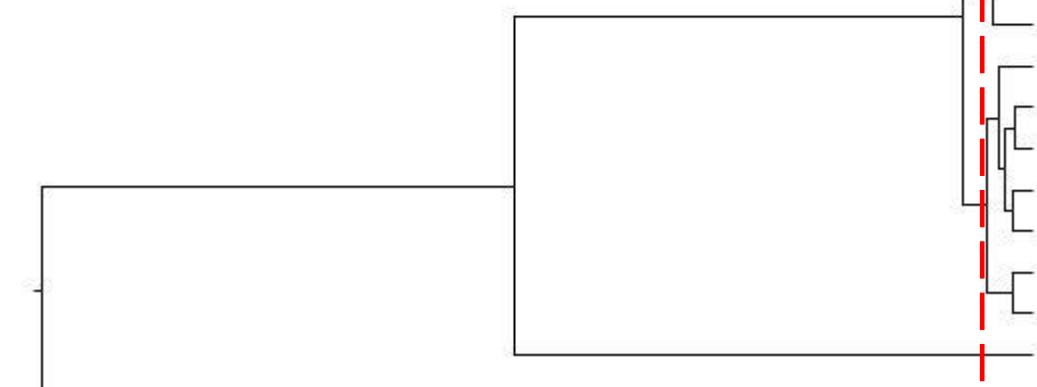
The General Mixed Yule Coalescent method (GMYC)

Identifies the transition points between inter- and intra-species branching rates

- Maximum likelihood method (ML)
- Based on the Yule process
- Single locus data
- Ultrametric tree as input
- Web server: <http://species.h-its.org/gmyc/>
- Available at:
<https://sites.google.com/site/noahmreid/software>



Heuristic methods



The General Mixed Yule Coalescent method (GMYC)

Heuristic methods

Web server: <http://species.h-its.org/gmyc/>

The screenshot shows a web browser window with the URL <https://species.h-its.org/gmyc/>. The page title is "The Exelixis Lab". On the left, a sidebar menu lists "GMYC SPECIES DELIMITATION", "GMYC web server" (which is highlighted in blue), "Lookup job", and "About this web server". The main content area has a purple border and contains the following text:

Web interface for single and multiple threshold GMYC

The backend of this web server runs the original [R implementation](#) of the [GMYC model](#) authored by Tomochika Fujisawa and Tim Barraclough. To run the R GMYC locally, please check [Tomochika Fujisawa's blog](#). There is also a python implementation of the single threshold GMYC model in my [GitHub repository](#), and a [Bayesian implementation of the single threshold GMYC in R](#) by Noah M. Reid.

I encourage you to try both GMYC and [PTP model](#) on your data. They are based on quite different models, but if they give you similar delimitations, you can be more confident on the results.

Note the input tree should be strictly ultrametric and bifurcating and with no zero branch lengths (google BEAST or r8s if you do not know how to do it), if you have un-time-calibrated phylogenetic trees, please use [PTP model](#).

Below this, there is a form with the following fields:

- "My ultrametric input tree (Newick format only):" with a file input field containing "Durchsuchen... Keine Datei ausgewählt."
- "Method:" dropdown menu set to "Single threshold"
- "My e-mail address:" text input field
- "Submit" button

- Chose an ultrametric single tree as input
- give an email address and start the analyses

The General Mixed Yule Coalescent method (GMYC)

Heuristic methods

Web server: <http://species.h-its.org/gmyc/>

The screenshot shows a web browser window for the GMYC web server at <https://species.h-its.org/gmyc/>. The page title is "Web interface for single and multiple threshold GMYC". The left sidebar lists "GMYC SPECIES DELIMITATION" with options: "GMYC web server" (selected), "Lookup job", and "About this web server". The main content area contains instructions about the R implementation of the GMYC model, notes on ultrametric trees, and a form for inputting a Newick tree. The form includes fields for "My ultrametric input tree (Newick format only)" (with a "Durchsuchen..." button), "Method" (set to "Single threshold"), "My e-mail address" (input field), and a "Submit" button.

PROs: - Relatively fast for large dataset

CONs: - Reliance on single loci

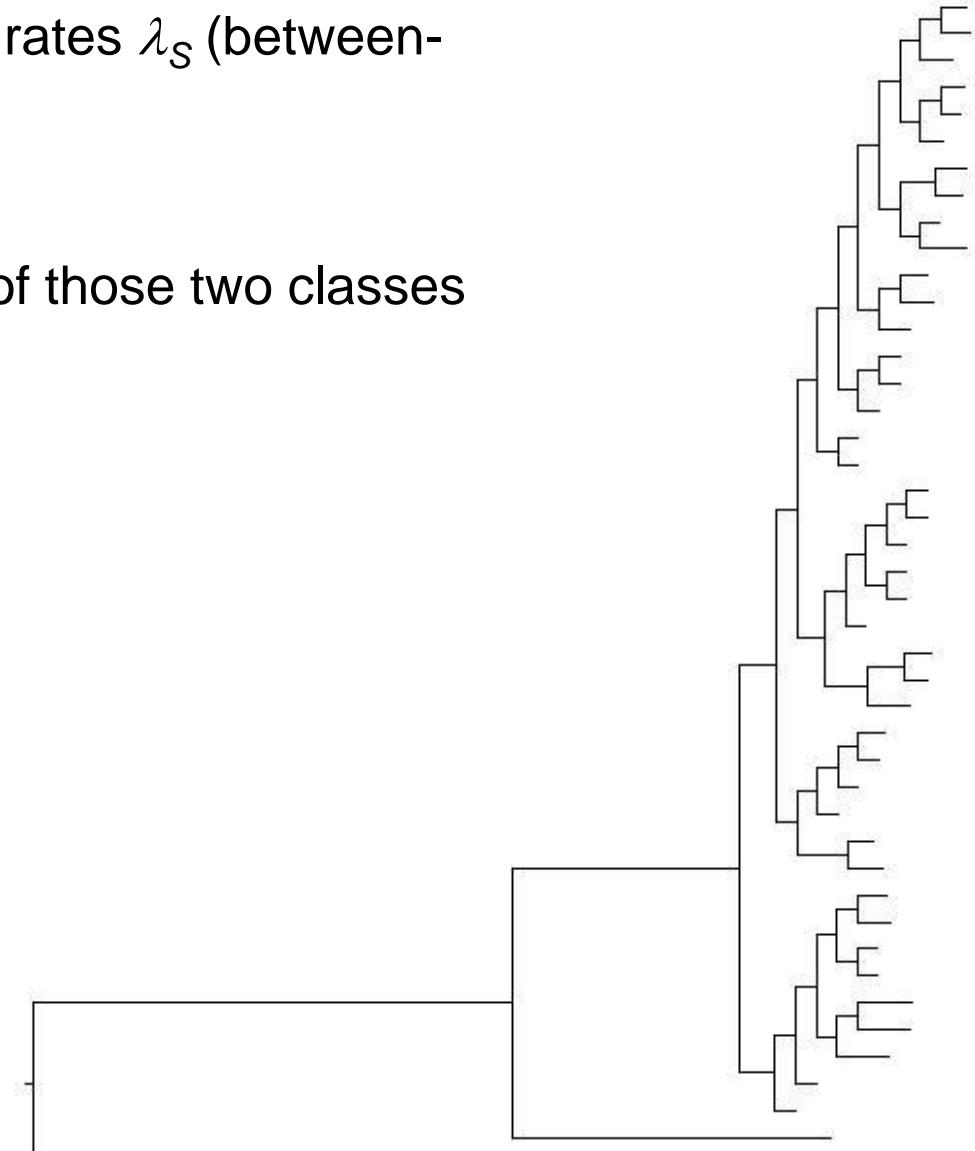
- Accuracy of ultrametric trees
- Oversplitting

Models the branch lengths in a rooted tree as a mixture of two rates λ_S (between-species) and λ_C (within-species)

A species delimitation model (λ) assigns every branch to one of those two classes

- Maximum likelihood or Bayesian implementation
- No need of ultrametric tree
- Potentially multilocus

Web server: <http://species.h-its.org/gmyc/>



Poisson Tree Process (PTP)

Heuristic methods

Web server: <http://species.h-its.org/gmyc/>

bPTP SPECIES DELIMITATION

bPTP web server

Look up jobs

Help

About PTP and bPTP

PTP paper

Download PTP and bPTP

SERVER STATUS

Free cluster nodes: 8

Total cluster nodes: 10

bPTP server: a Bayesian implementation of the PTP model for species delimitation
with PhyloMap for visualization

Please cite: A General Species Delimitation Method with Applications to Phylogenetic Placements. Zhang, Jiajie, Kapli, P., Pavlidis, P., and Stamatakis, A. *Bioinformatics (Oxford, England)*(2013), 29 (22): 2869-2876

bPTP is an updated version of the original maximum likelihood PTP (maximum likelihood PTP search result is part of the bPTP results), it adds Bayesian support (BS) values to delimited species on the input tree. Higher BS value on a node indicates all descendants from this node are more likely to be from one species. Note this web server only allows a single phylogenetic tree as input - to run on multiple trees (from Bayesian analysis or bootstrap), please download the bPTP and PTP standalone.

If you are not familiar with Bayesian analysis, please [read this](#) first.

The server will only accept Newick format or NEXUS format with no annotations on the tree.
Here is an [example of Newick format](#) and an [example of NEXUS format](#).

Special note on taxa name: A valid taxa name should only contain letters, numbers, _ and -; and should NOT be pure numbers.
Any other characters, in particular space, \$, #, %, (and), will cause errors!

PROs: - Relatively fast for large dataset

CONs: - Relies on single loci / concatenation
- Implicitly assumes rec. monophyly
- Oversplitting