DIPARTIMENTO di BIOLOGIA

TADAB

workshop

Tools for Advanced Data Analyses in Biology

27 November – 1 December 2023

UNIVERSITÁ DI PISA

The coalescent model for phylogenetics and species delimitation

GEORG-AUGUST-UNIVERSITÄT GÖTTINGEN

Salvatore Tomasello

Pisa, 30/11/2023

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

SCAN ME

Outline

Thursday morning	- Introduction to species delimitation and the coalescent theory - Heuristic methods for species delimitation
Thursday afternoon	- Parametric approaches: Species delimitation with SPEEDOMON
Friday morning	- Parametric approaches: Species Delimitation BPP BPP - Integrative approaches for species delimitation: iBPP

https://commons.wikimedia.org/wiki/File:Asteracea_poster_3.jpg

Why species delimitation

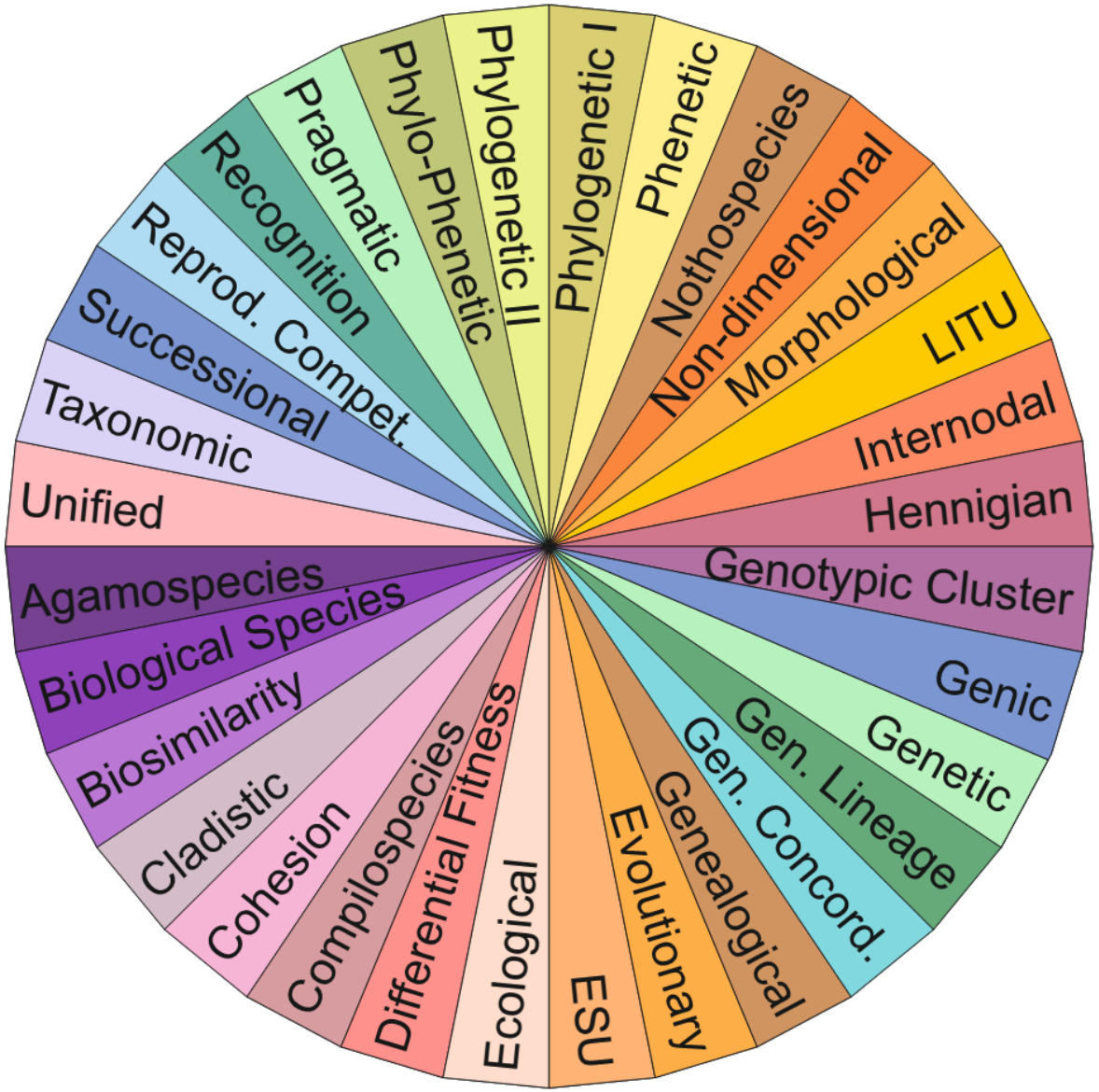
- Humans like to order objects in classes

https://commons.wikimedia.org/wiki/File:Asteracea_poster_3.jpg

Why species delimitation

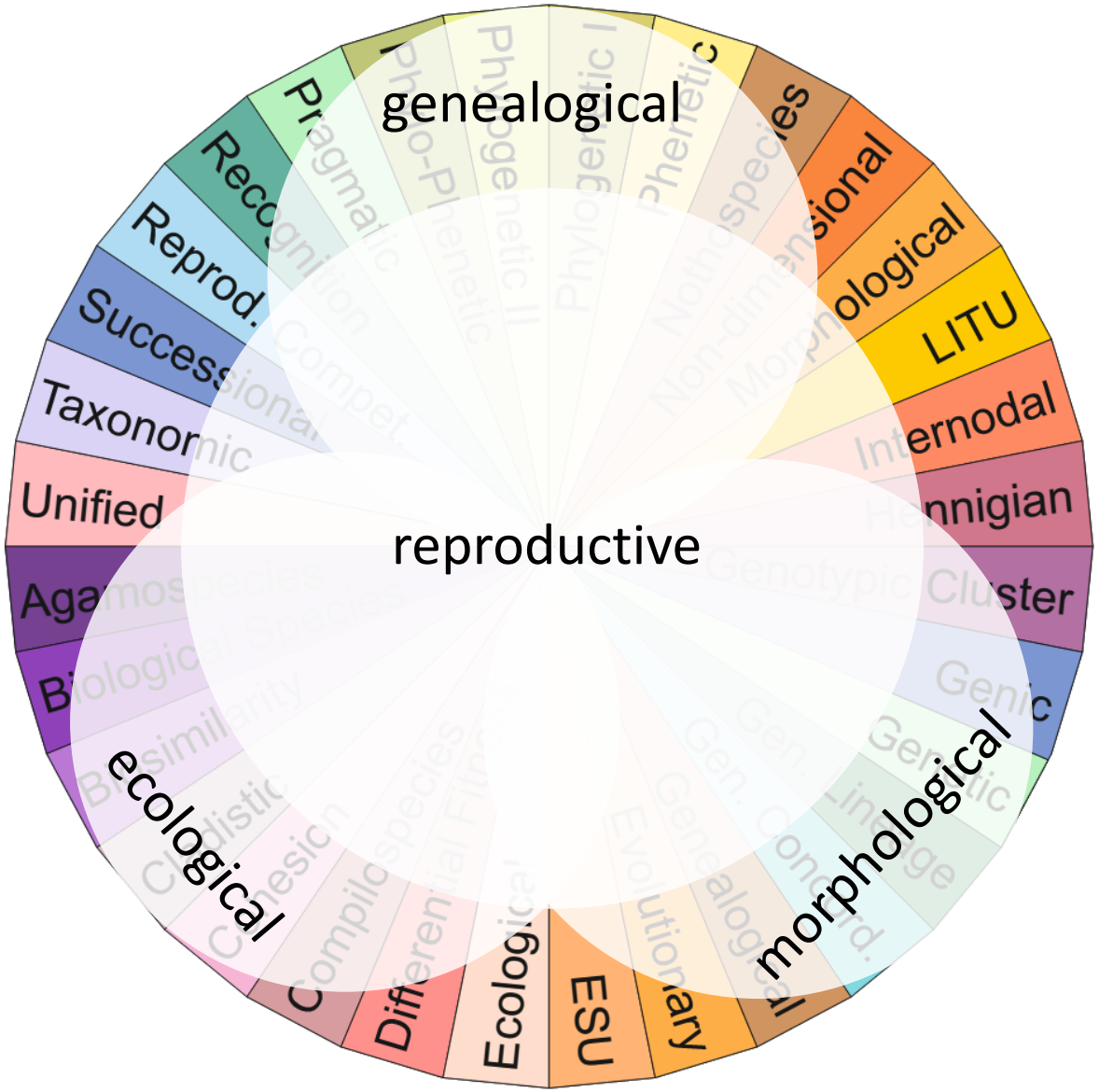
- Humans like to order objects in classes
- **Species exist!**
- **Fundamental unit in biology**
- **Reliable Taxonomy are extremely important for many biological disciplines**

Why species delimitation



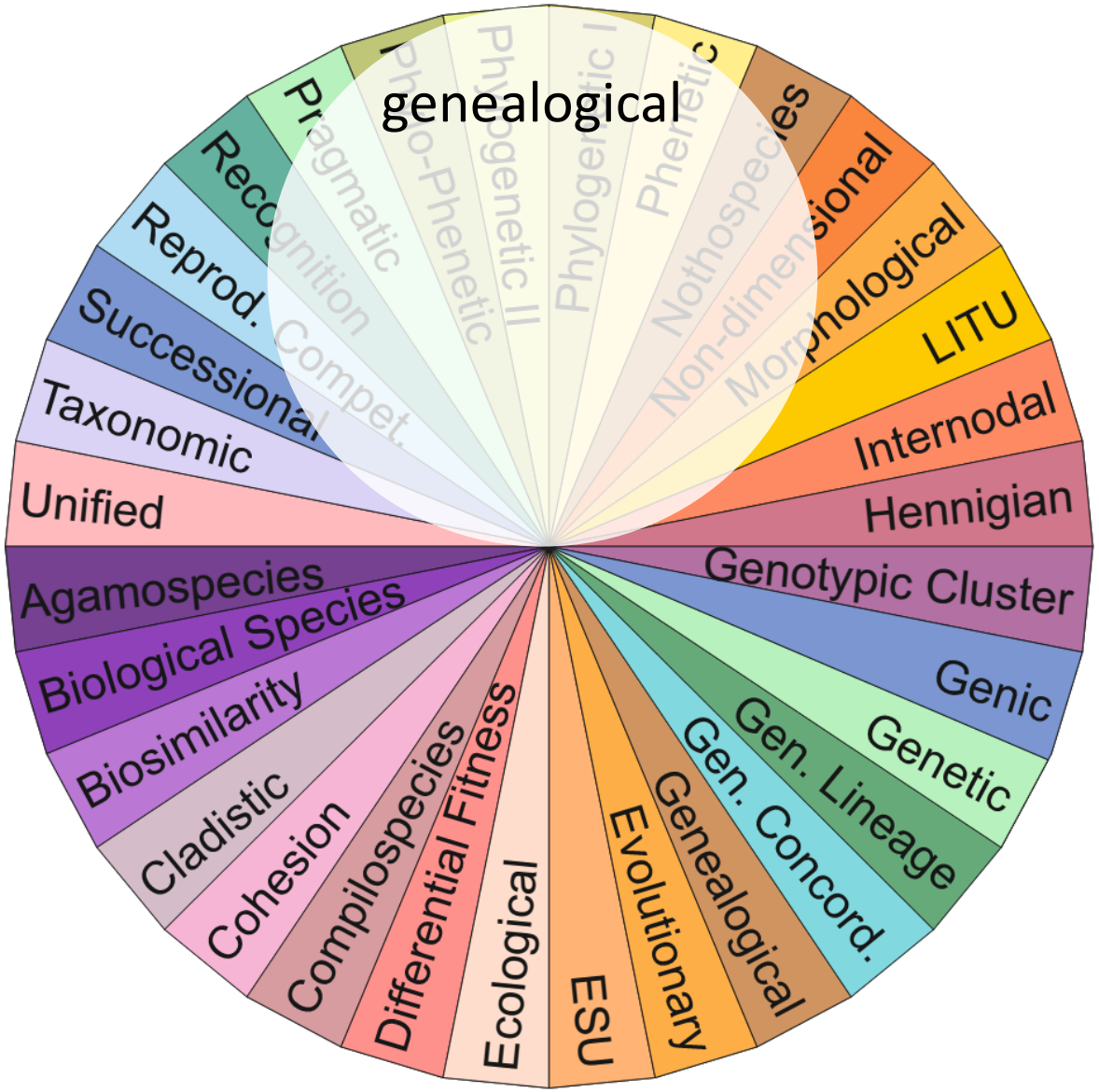
Zachos 2016. Species Concepts in Biology. Historical Development, Theoretical Foundations and Practical Relevance.

Why species delimitation



Zachos 2016. Species Concepts in Biology. Historical Development, Theoretical Foundations and Practical Relevance.

Why species delimitation



Zachos 2016. Species Concepts in Biology. Historical Development, Theoretical Foundations and Practical Relevance.

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)

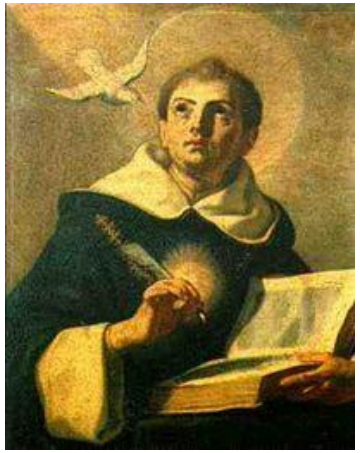
Evolutionary Species Concept: Ancestor-descendant lineages that evolve separately from other such lineages and have their own evolutionary tendencies and historical fate

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

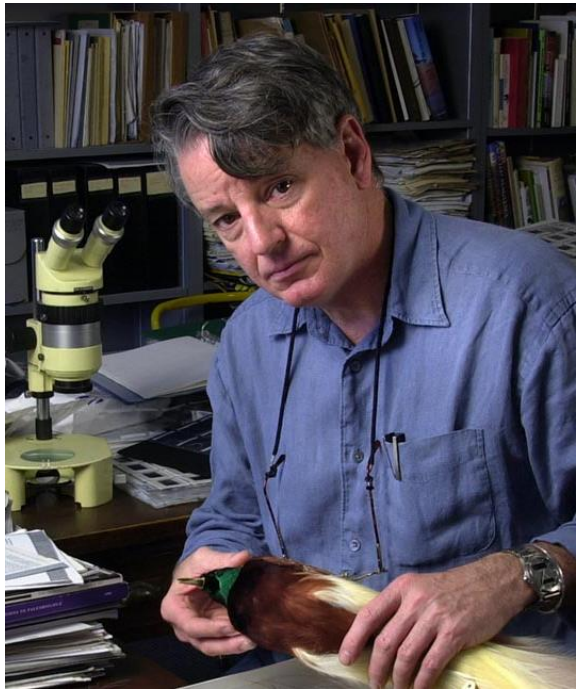


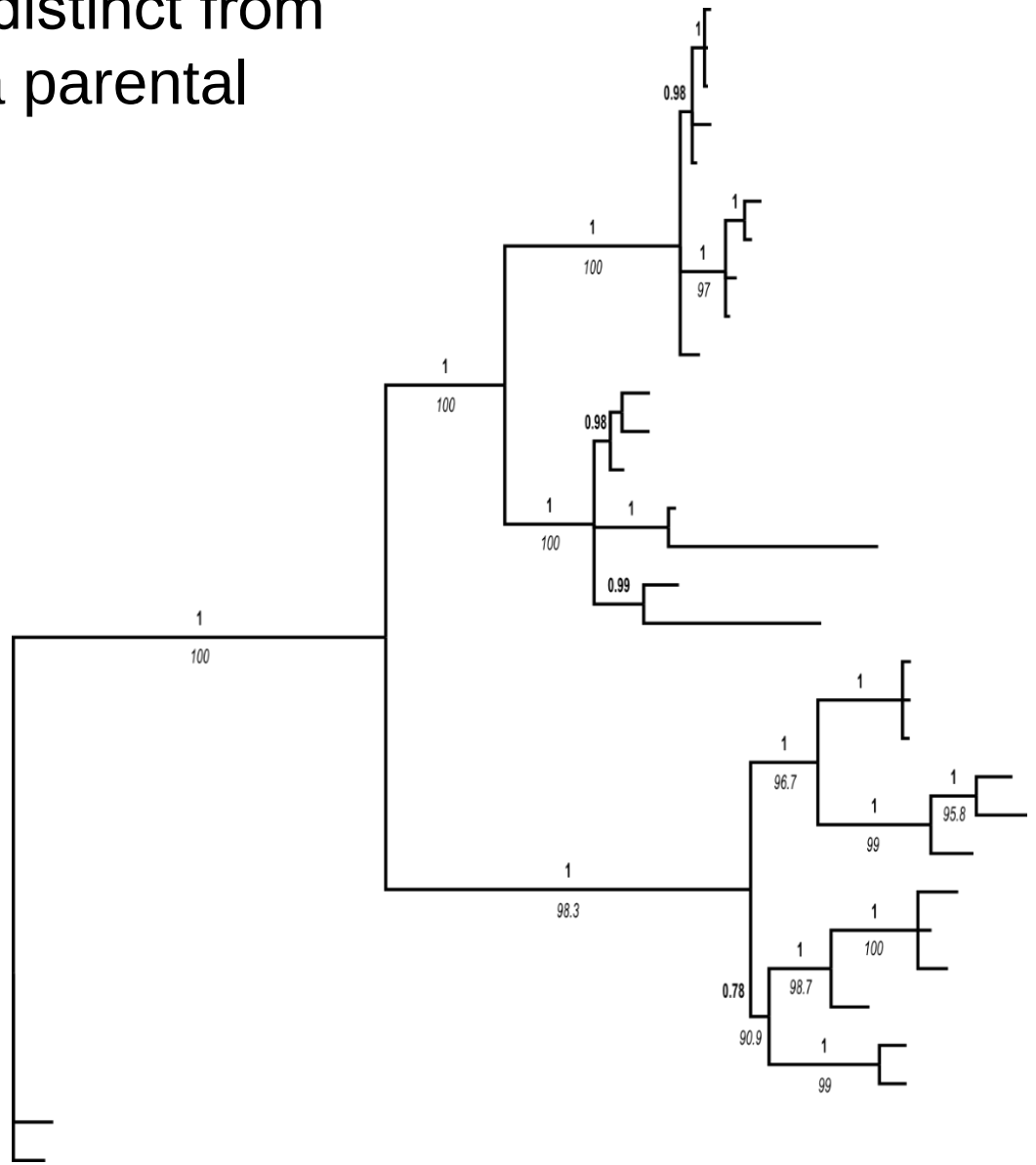
„It belongs to the wise man to order" (Thomas Aquinas)

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)



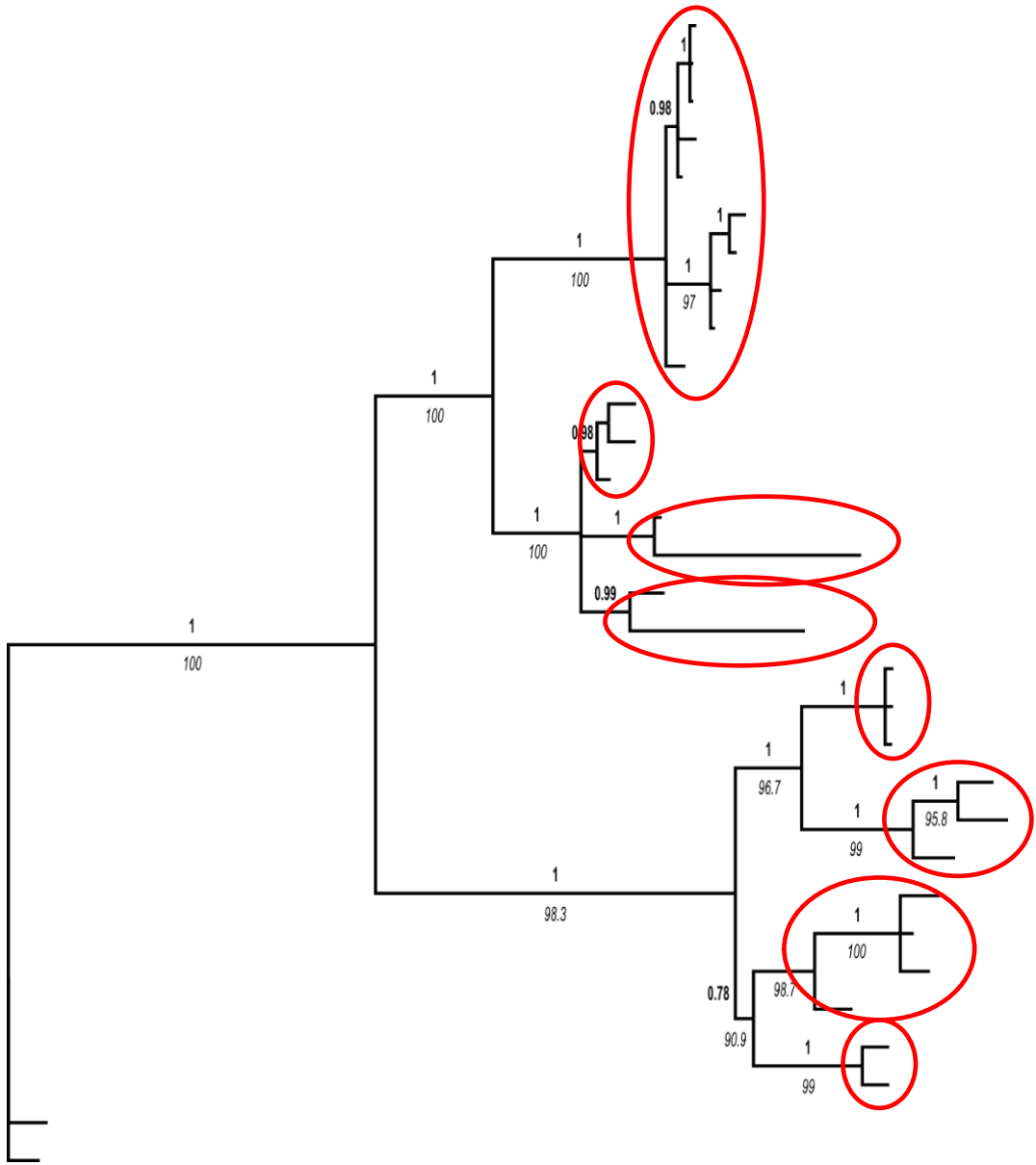


Reciprocal monophyly:

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

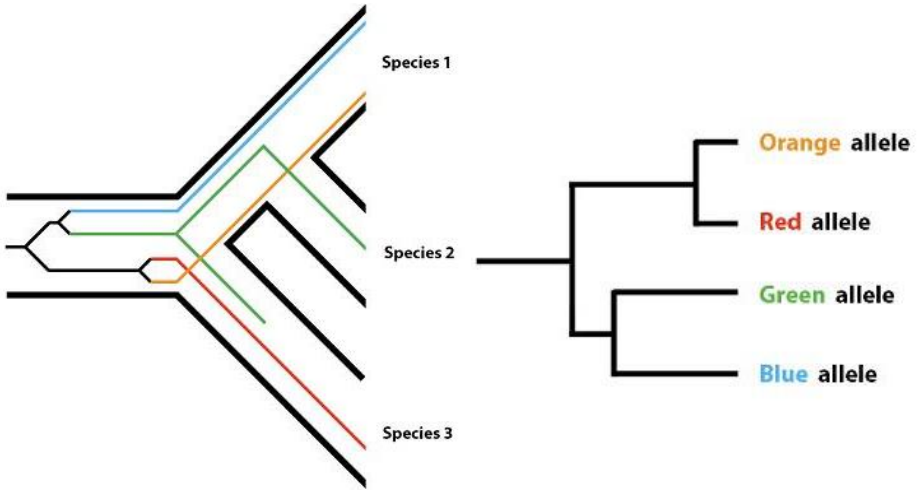
Problems:

- Gene trees/species tree discordance



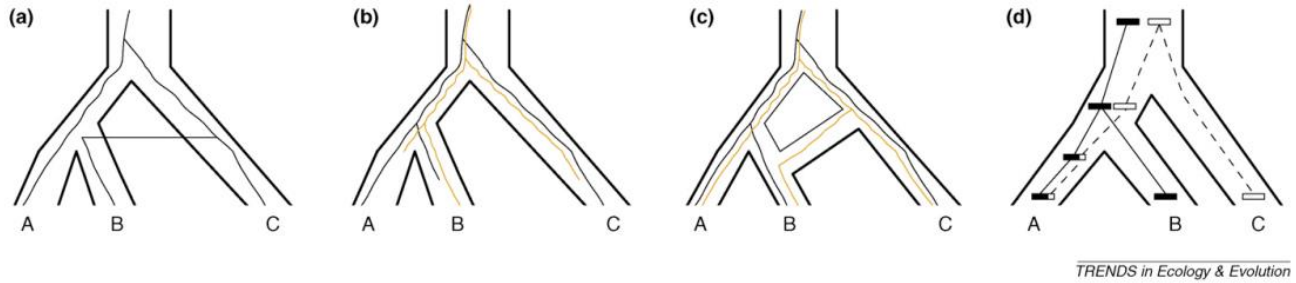
Processes causing incongruence among phylogenetic trees:

1. Intra-species stochastic factors:
   - gene loss / duplication
   - incomplete lineage sorting



Source: http://biologos.org

2. Inter-species factors:
   - horizontal gene transfer
   - hybridization

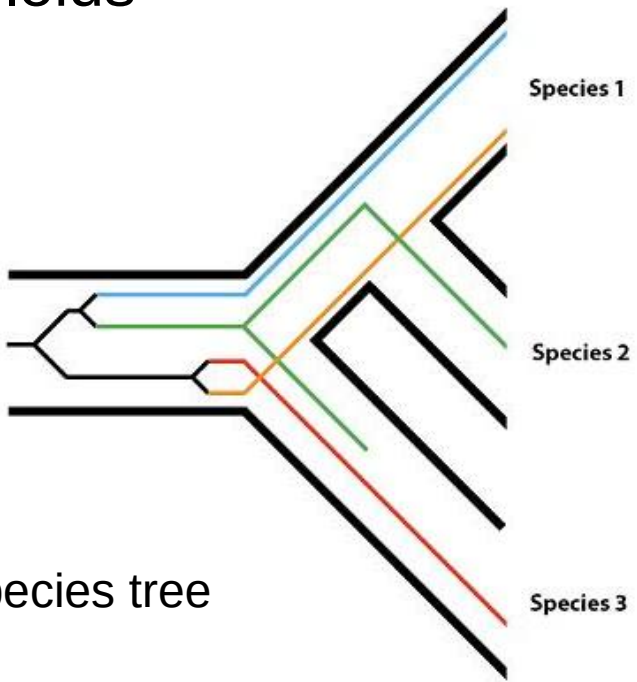


Source: **Degnan & Rosemberg. 2009.** *Trends in Ecology and Evolution* 24: 332 340

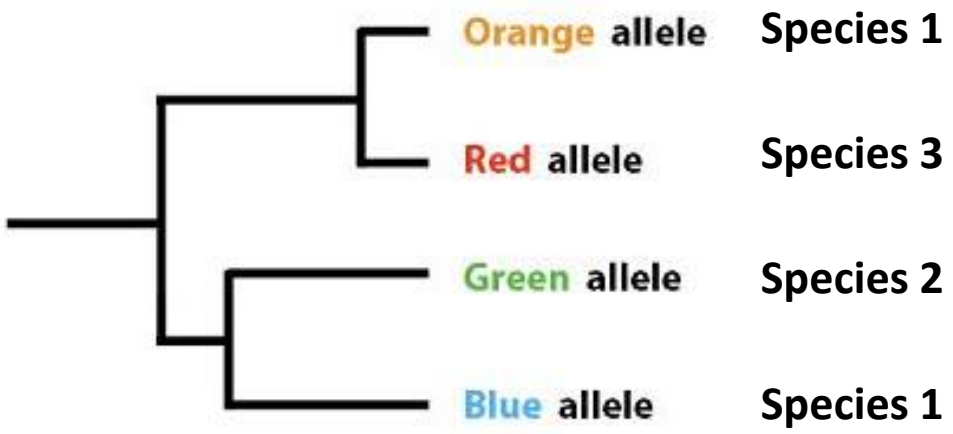
Processes causing incongruence among phylogenetic trees:

Incomplete lineage sorting particularly prominent in:

- fast radiating groups
- huge ancestral populations size
- polyploids



Species tree



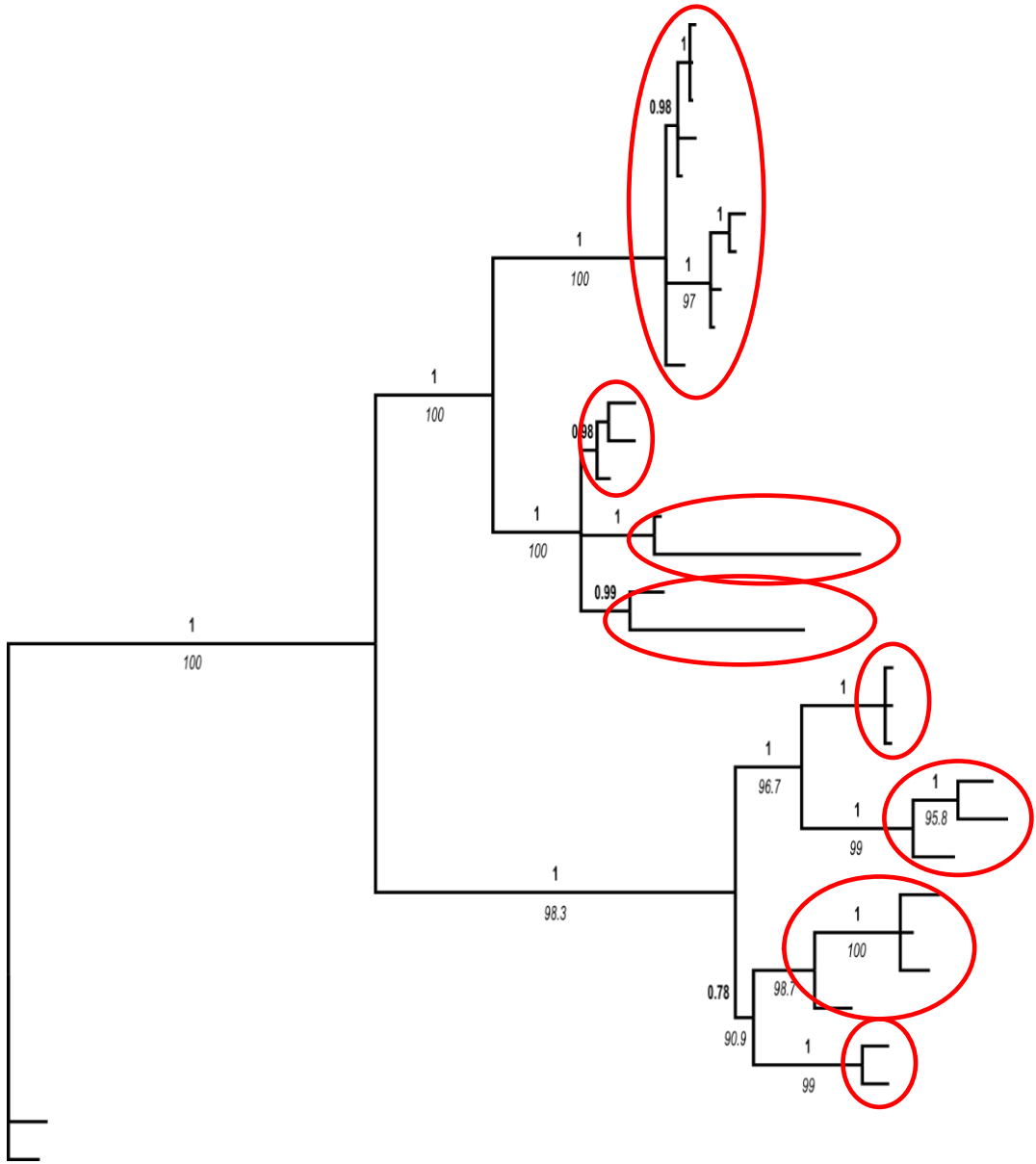
Gene tree

Reciprocal monophyly:

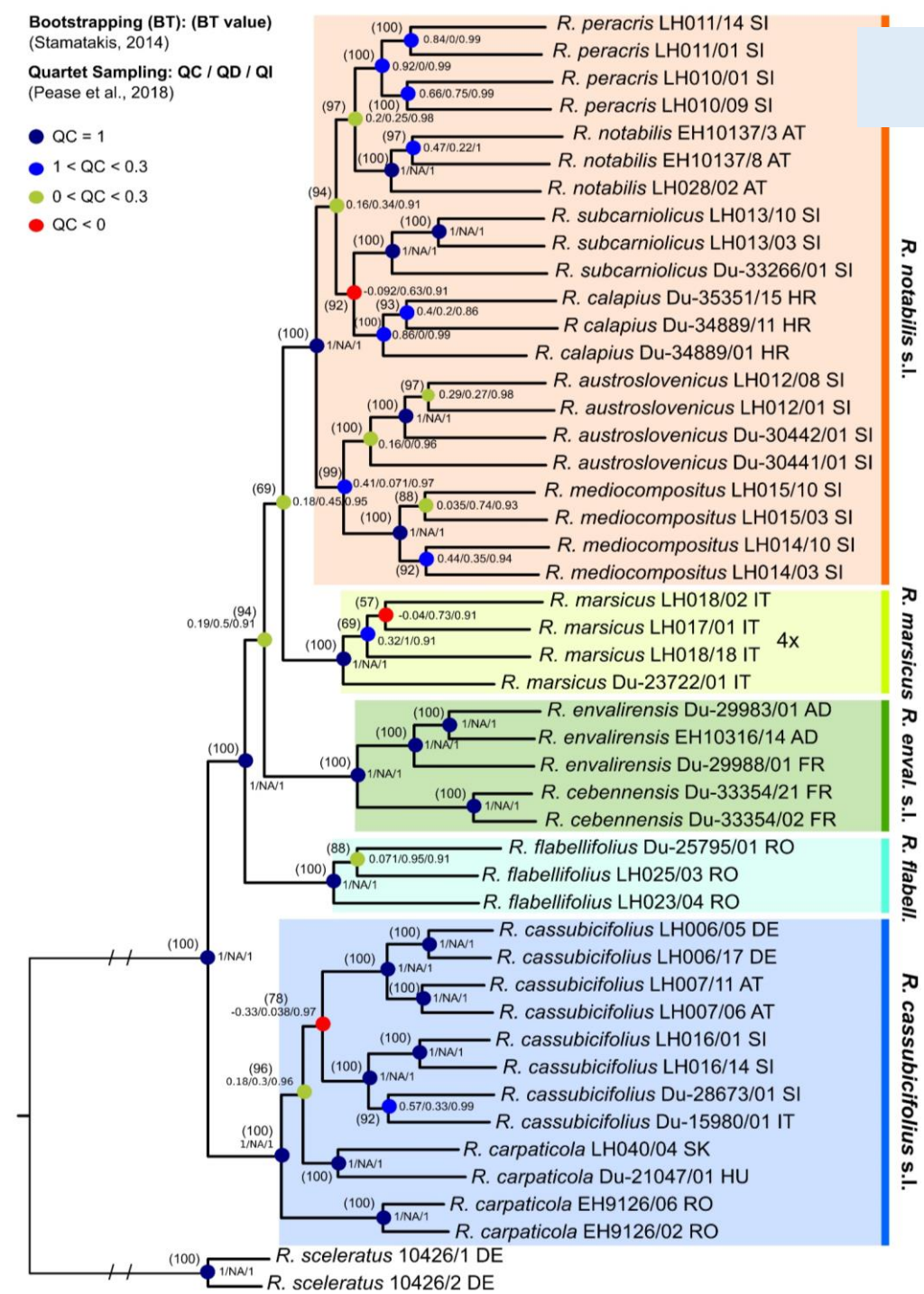
The smallest diagnosticable monophyletic group of population within which there is a parental pattern of ancestry and descent

Problems:

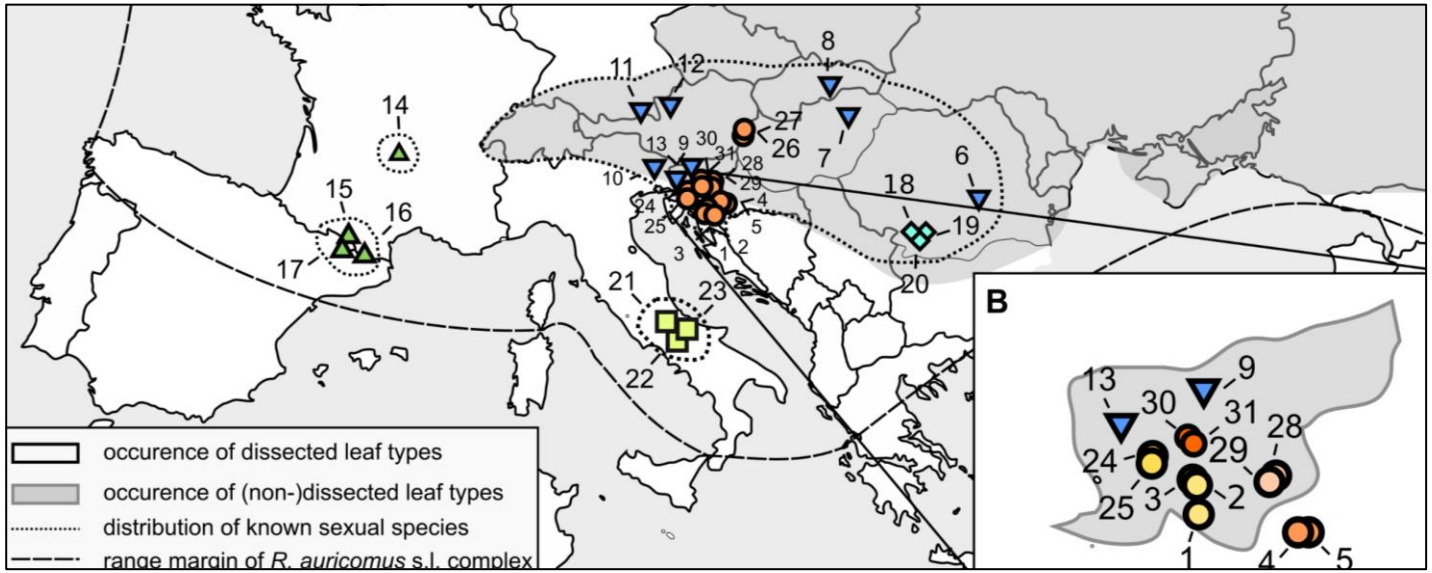
- Gene trees/species tree discordance
- Sampling (What, How much)



DNA sequence data



Ranunculus auricomus s.l.

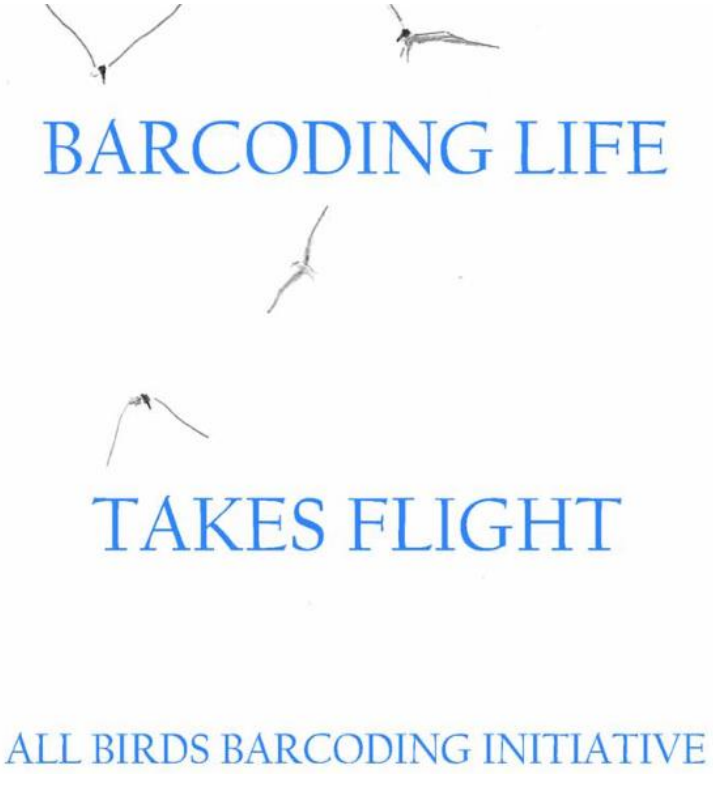


Karbstein et al. 2020. Taxon 69: 1191-1220. https://doi.org/10.1002/tax.12365

Barcoding Initiatives:

DNA sequence data

Single genomic region and sequence divergence to delimit species



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

Barcoding Initiatives:

Single genomic region and sequence divergence to delimit species

Problems:

- Single locus
- Arbitrary threshold

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B

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

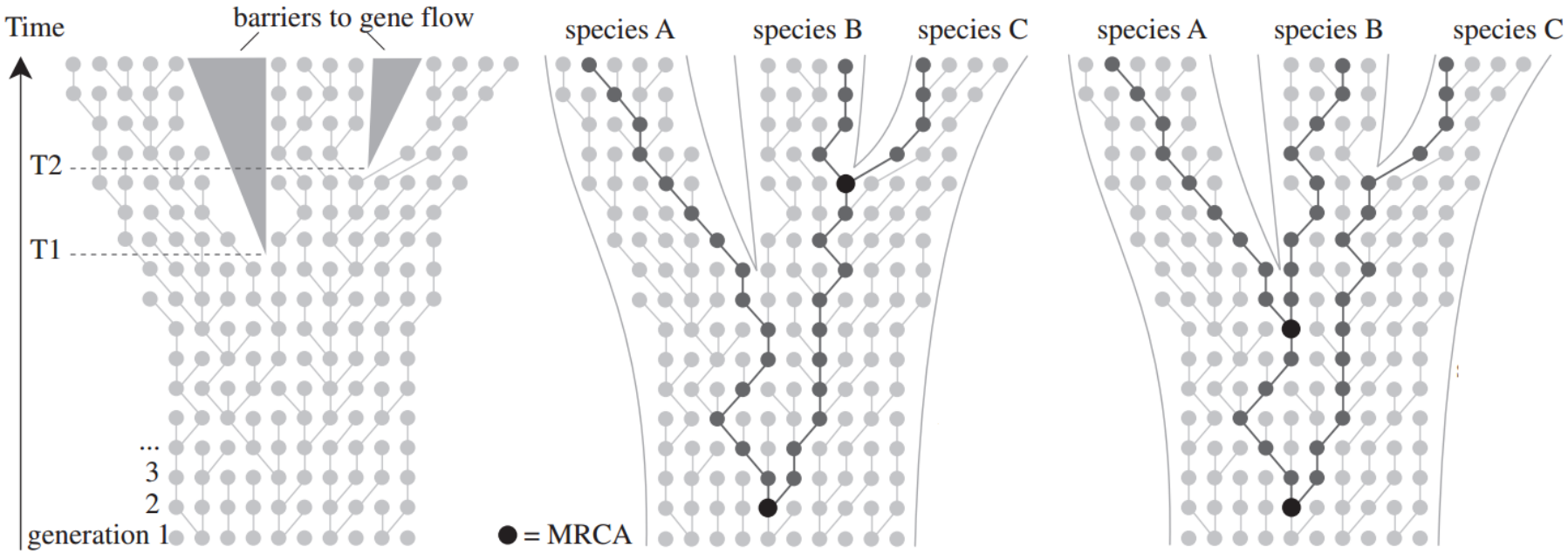
Division of Invertebrate Zoology, American Museum of Natural History, 79th Street at Central Park West, New York, NY 10024, USA

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*)



Source: **Leliaert et al. 2015.** *European Journal of Phycology* 49:179-196

The Multispecies Coalescent (MSC)

"The mathematical and probabilistic theory underlying the evolutionary history of alleles"

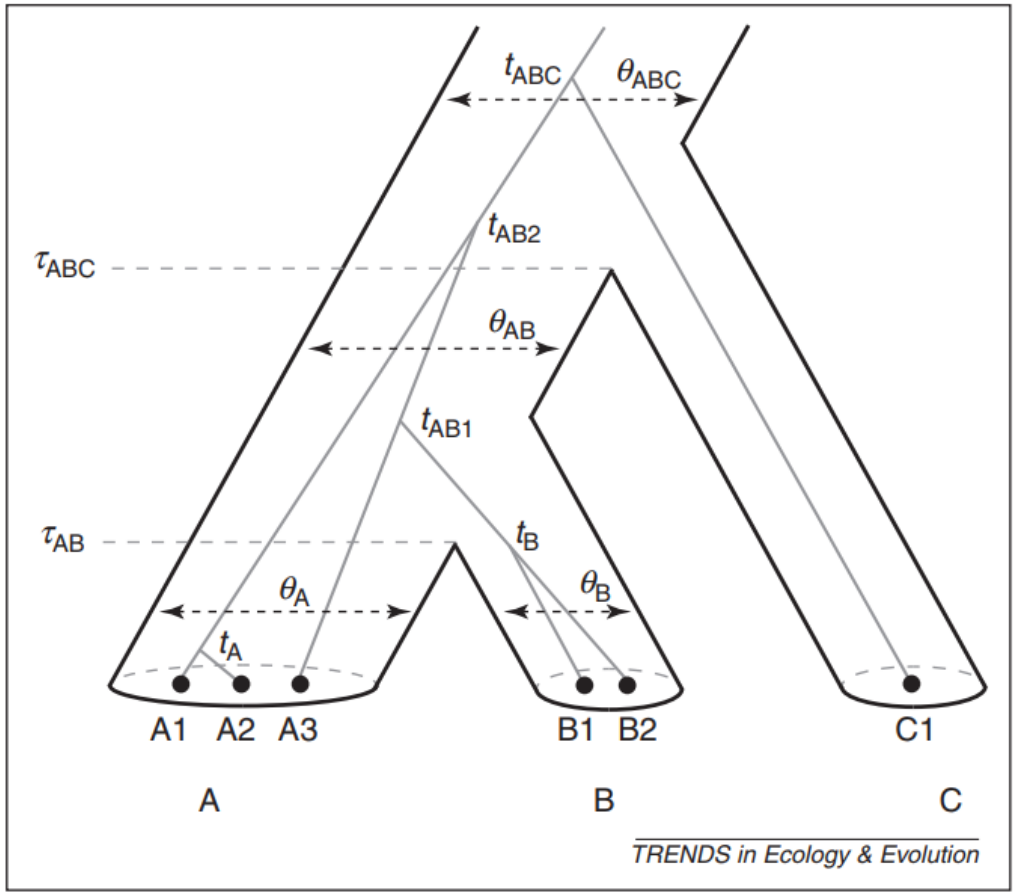
- Give a mean of calculating gene trees probability for a given species tree
- Can accommodate gene tree discordances
- Estimation of species divergence times
- Species delimitation

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

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?

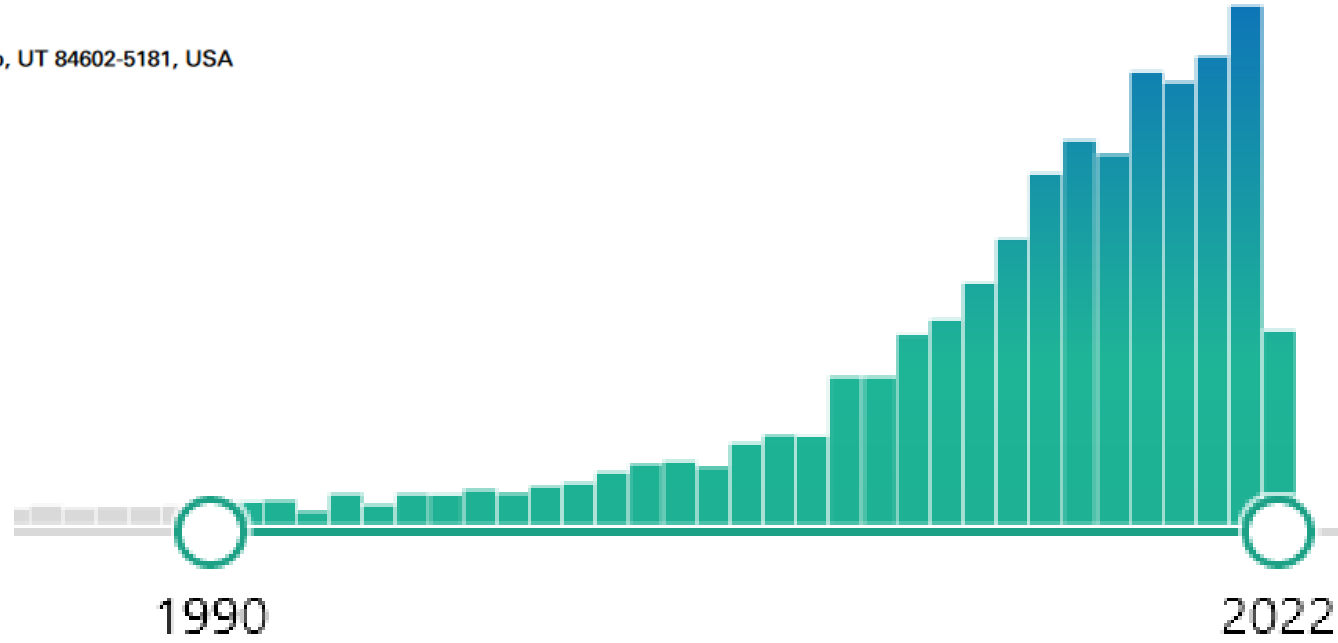
- Absence of other taxonomically relevant characters
- Recently diverging groups
- Unexplored areas with high diversity
- Cryptic species

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



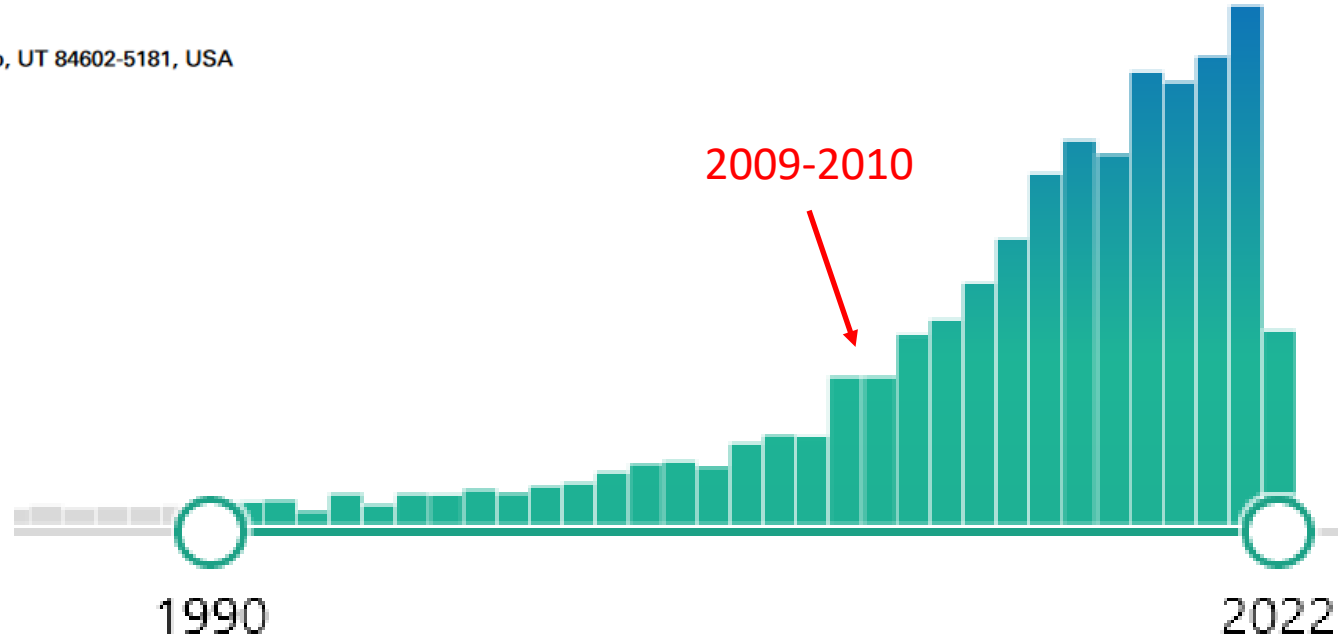
Source: https://pubmed.ncbi.nlm.nih.gov/?term=species+delimitation

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

Delimiting species: a Renaissance issue in systematic biology

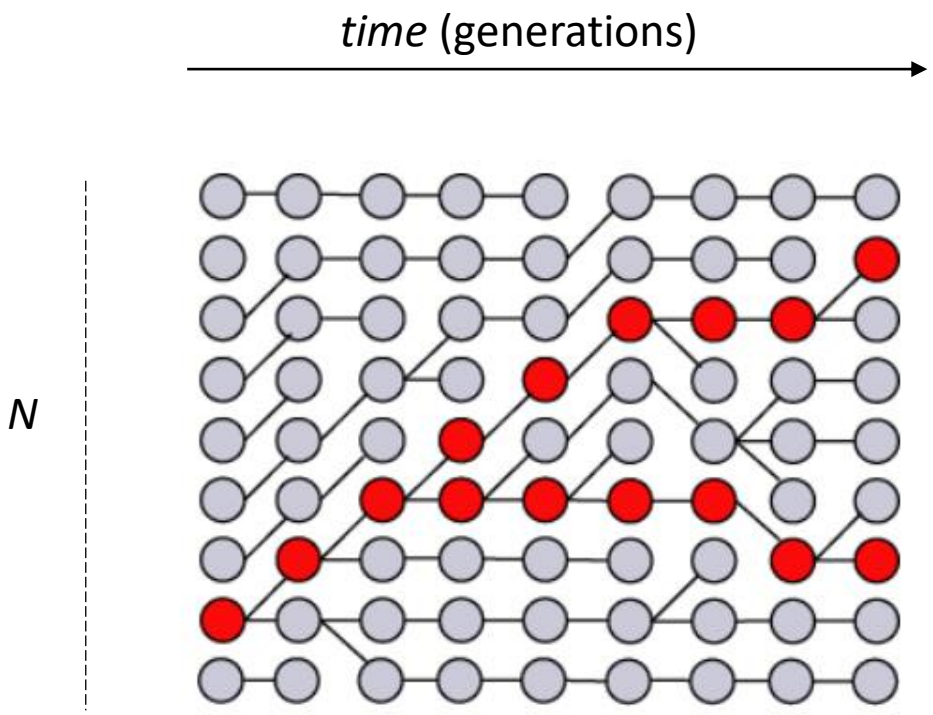
Jack W. Sites Jr and Jonathon C. Marshall

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



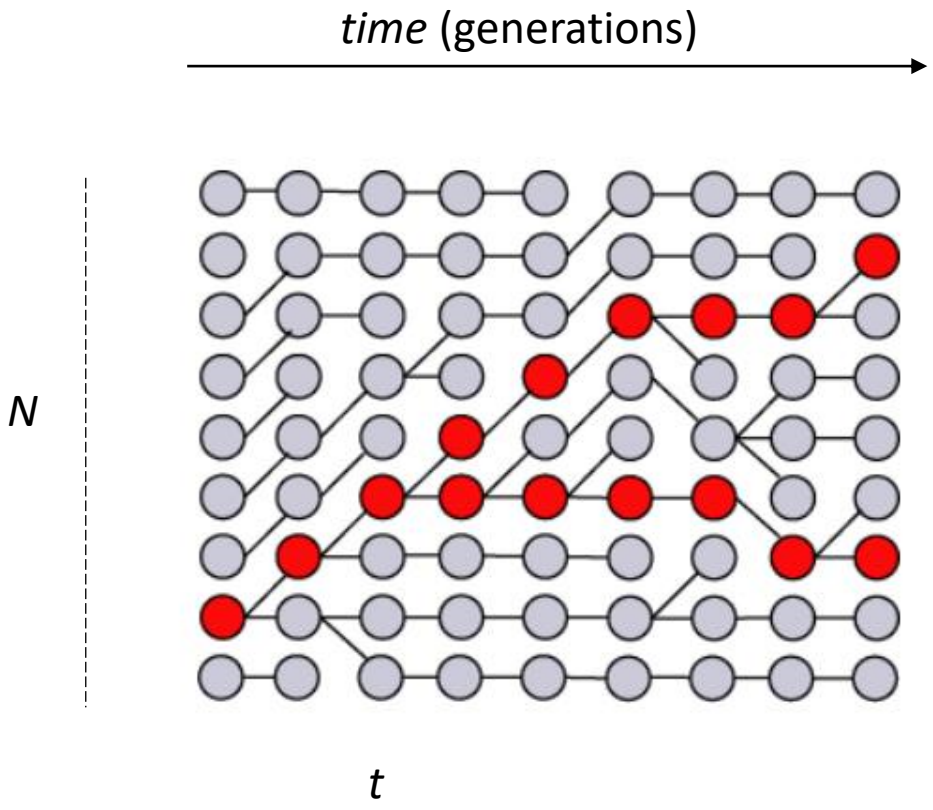
Source: https://pubmed.ncbi.nlm.nih.gov/?term=species+delimitation

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

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

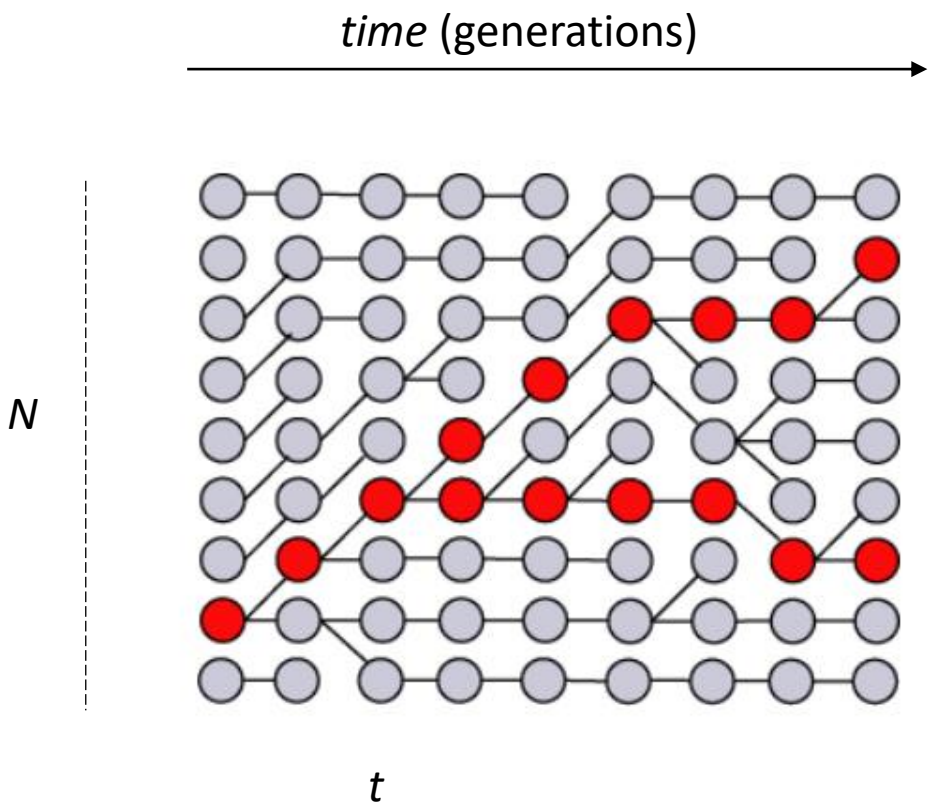
Probability of two alleles not to coalesce

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

Probability of two alleles to coalesce at generation t

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

The Coalescent Model

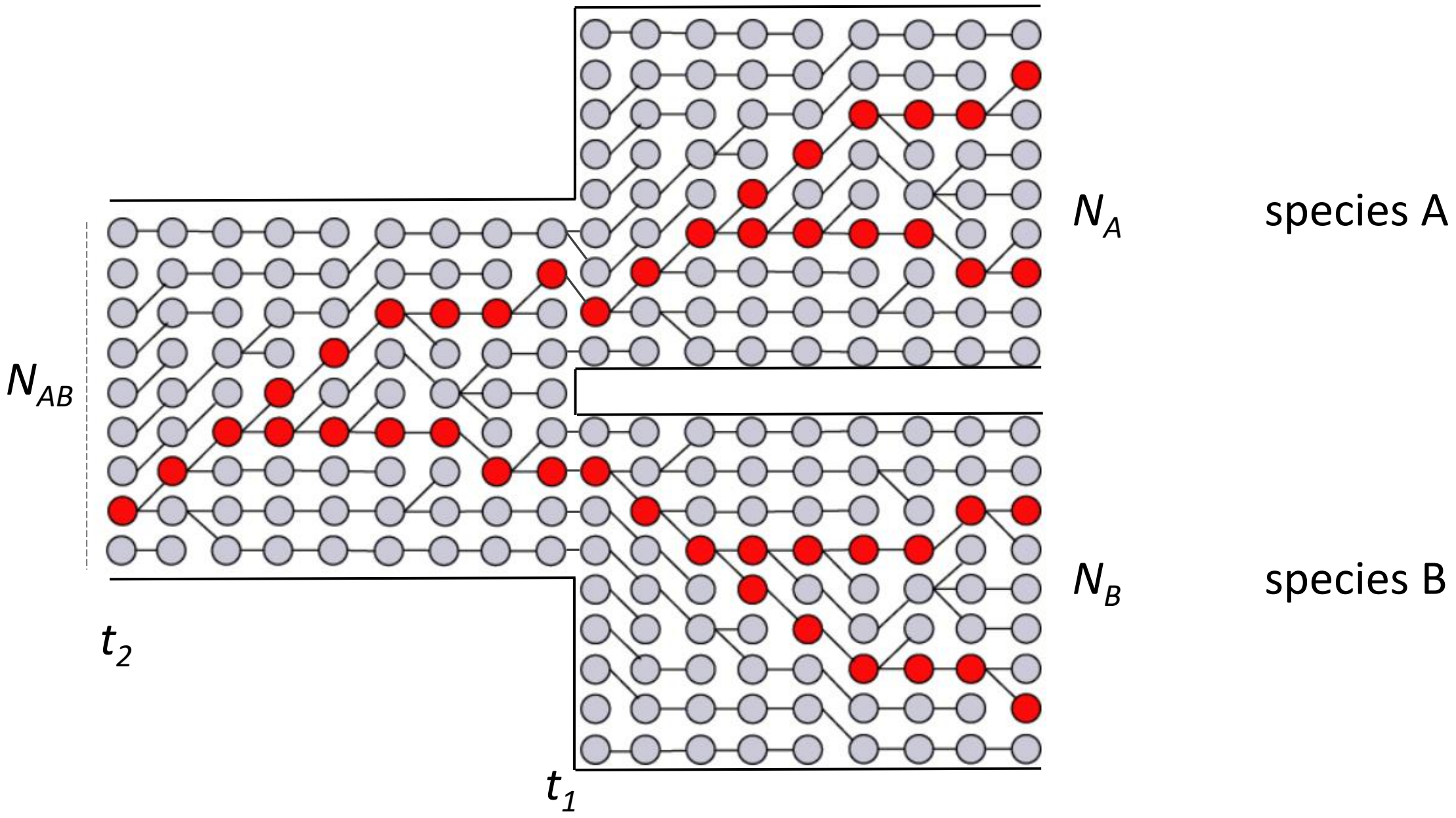


$$\frac{1}{N}\left(1-\frac{1}{N}\right)^{t-1}$$

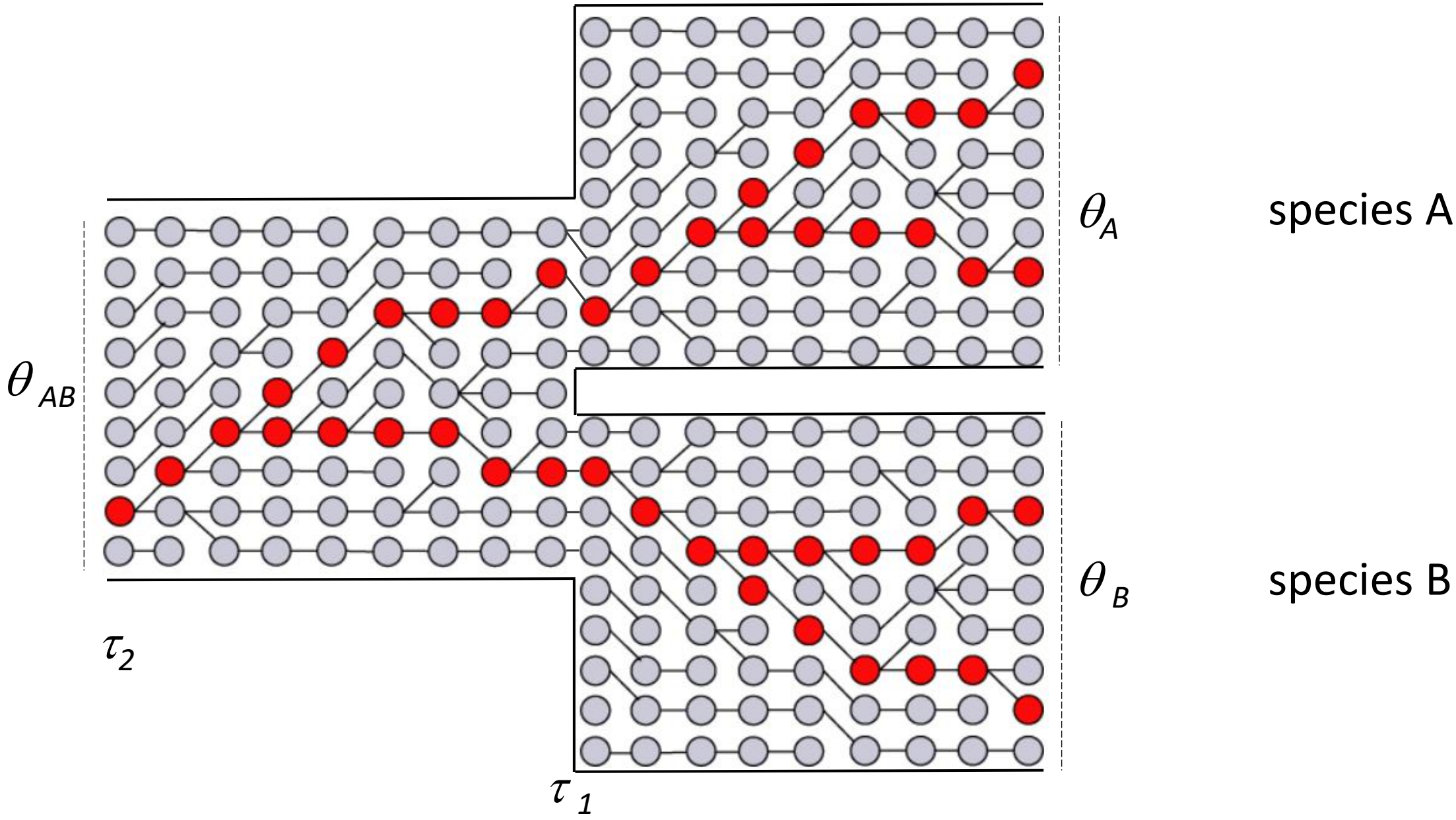
Mean = *N*

Mean time of the coalescent event!

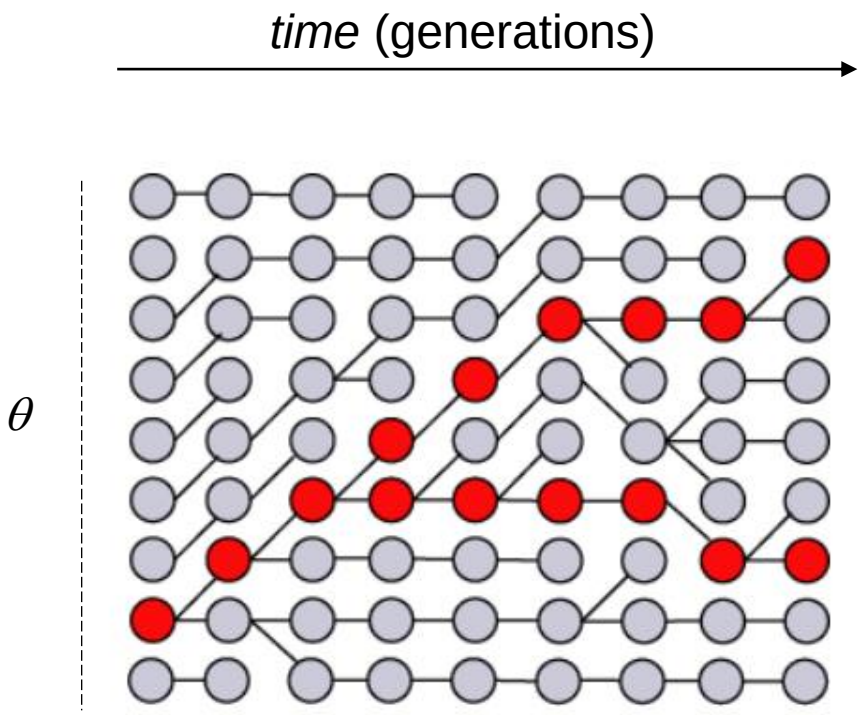
The Multispecies Coalescent (MSC)

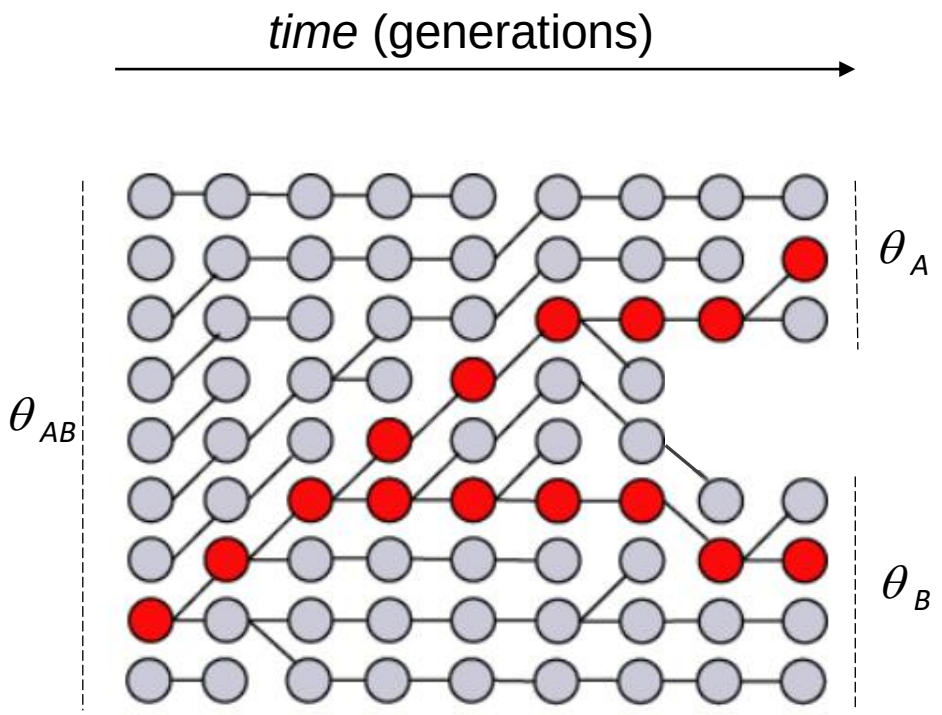


The Multispecies Coalescent (MSC)



Coalescent-based species delimitation





One or two species?

Coalescent-based species delimitation

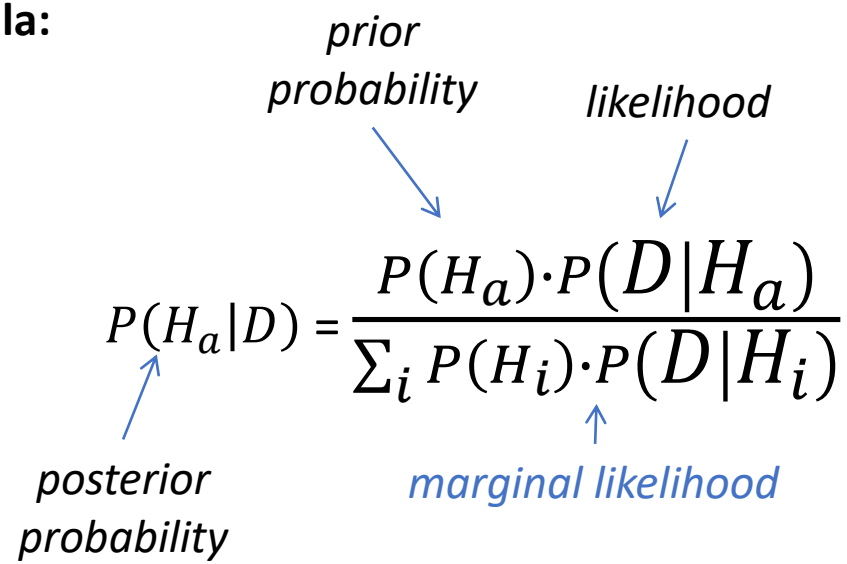
- **Discovery approaches**: no prior information on pecies delimitation
- **Validation approaches**: test a few species delimitation hypotheses

- **Heuristic methods**: use summary statistics, semplification 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**

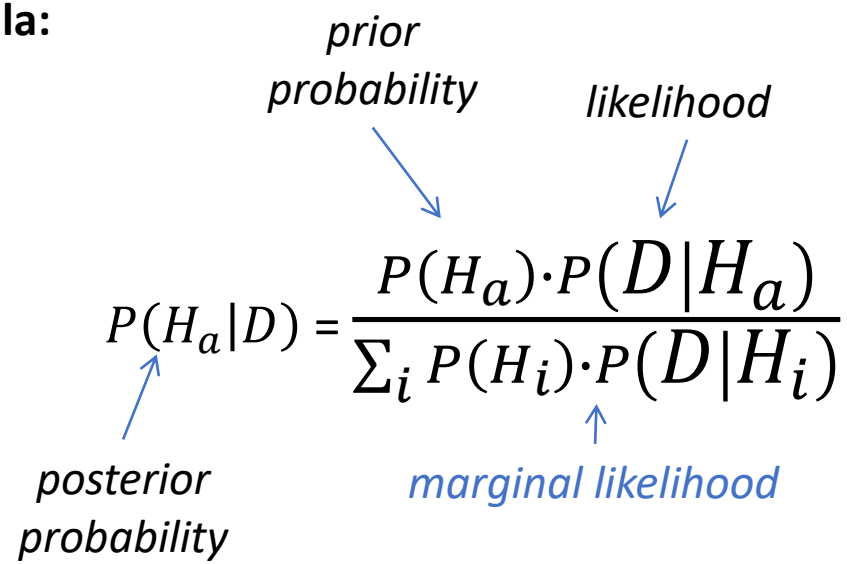
Marginal Likelihood

Bayes' formula:



Marginal Likelihood

Bayes' formula:



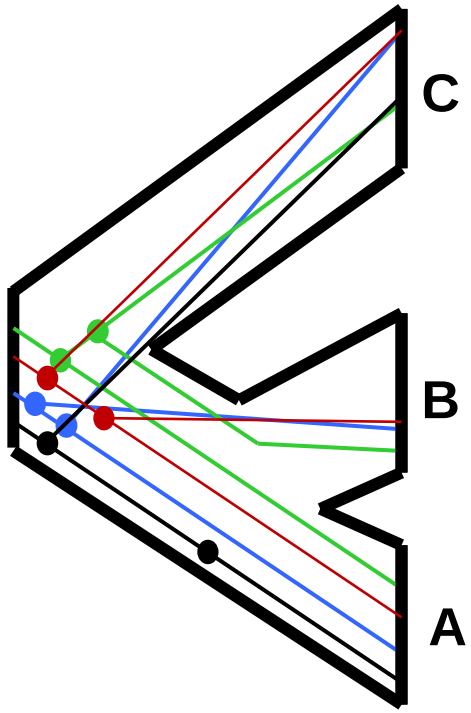
Methods:

- Path sampling
- Stepping stones
- Nested sampling

Beast2
Bayesian evolutionary analysis by sampling trees

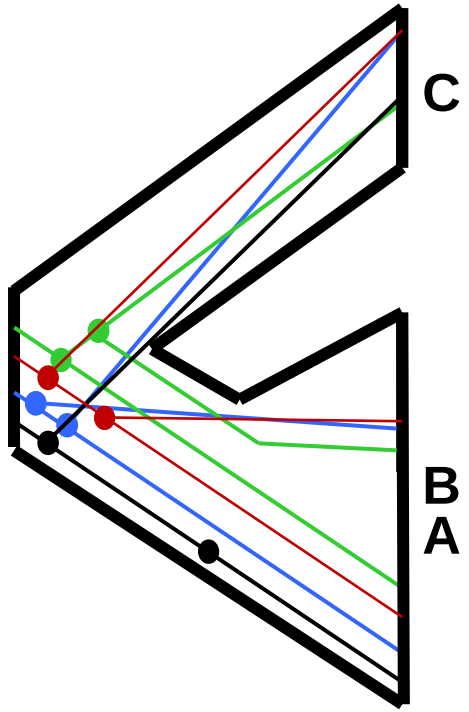
Marginal Likelihood

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



3 species model

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

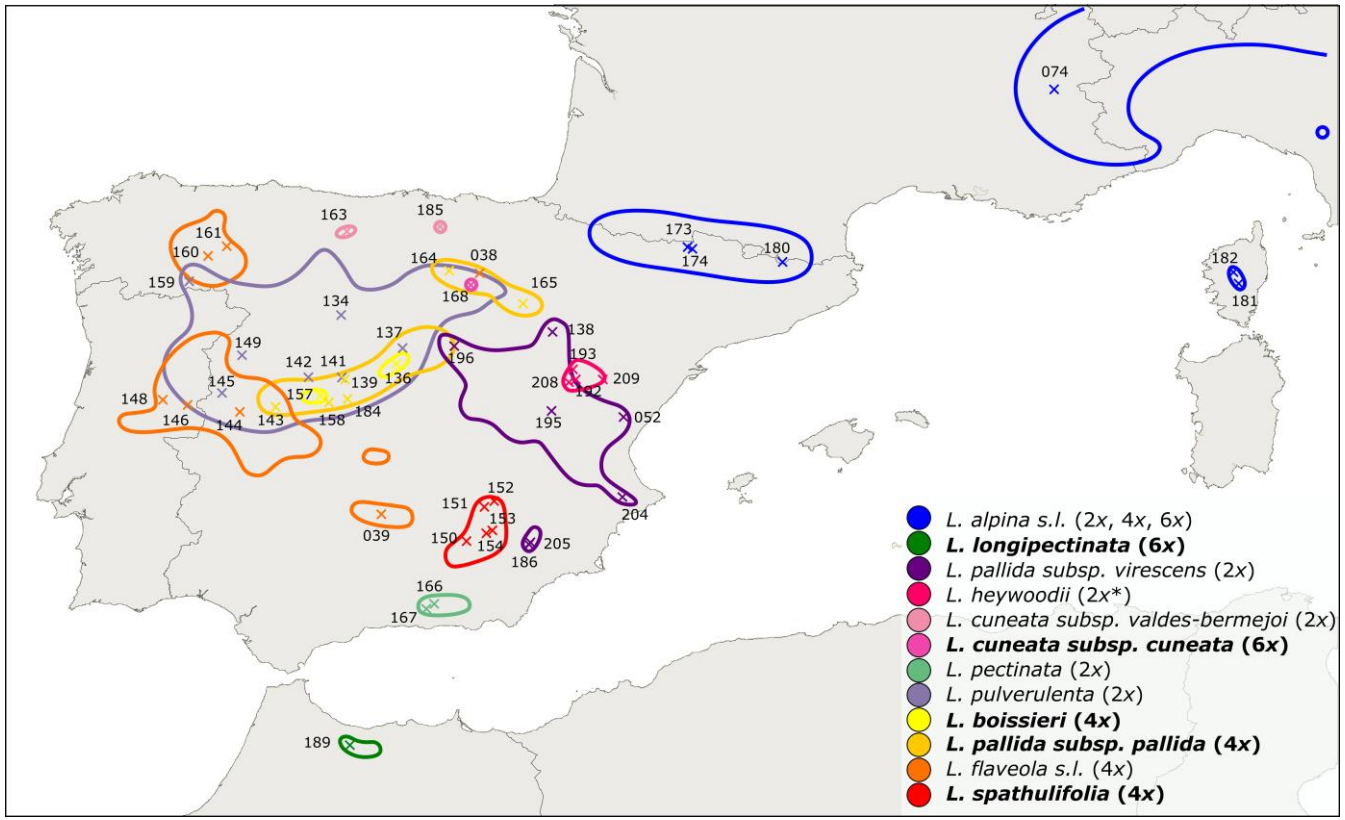


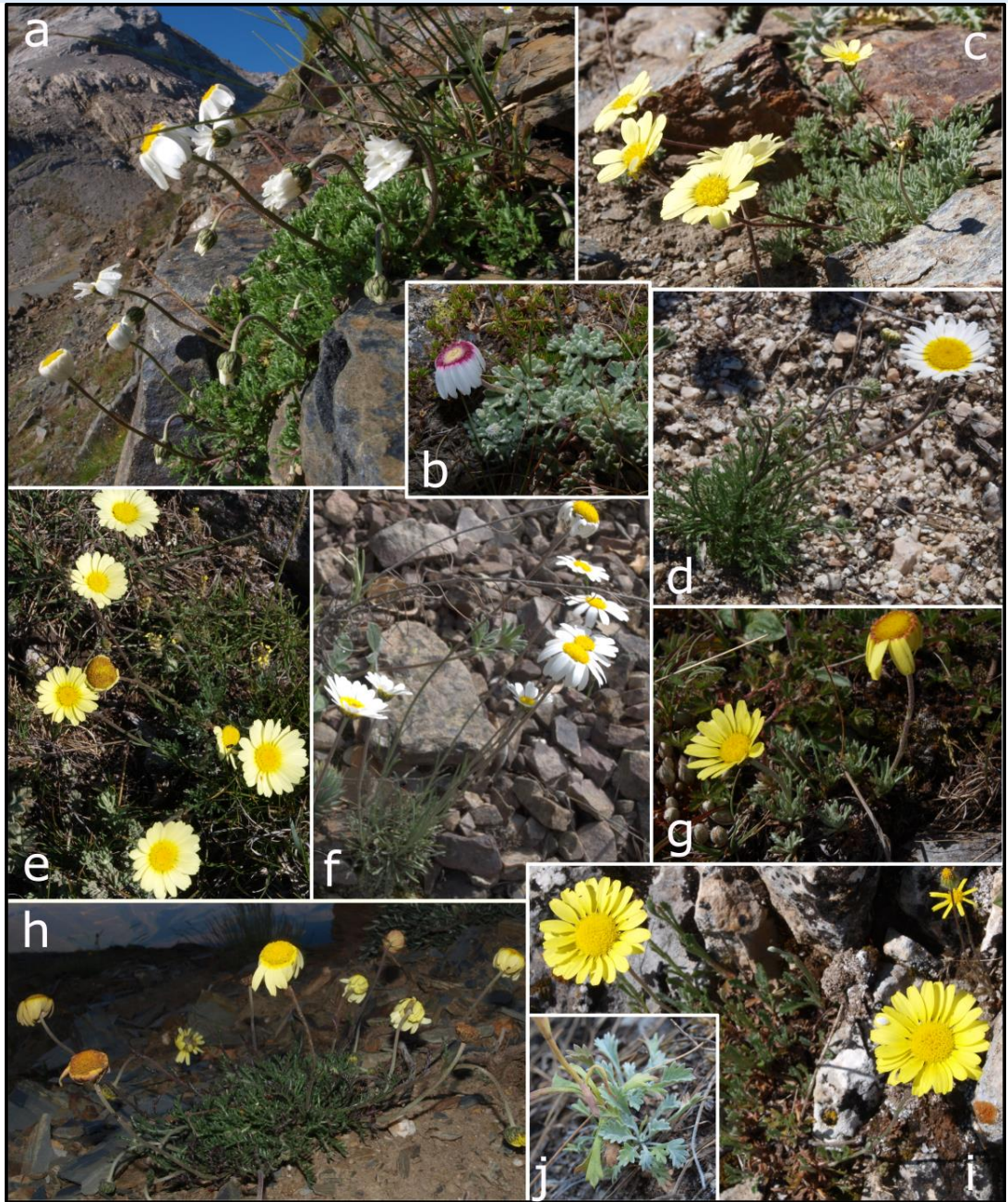
2 species model

***Leucanthemopsis 4x* species delimitation**

Marginal Likelihood

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



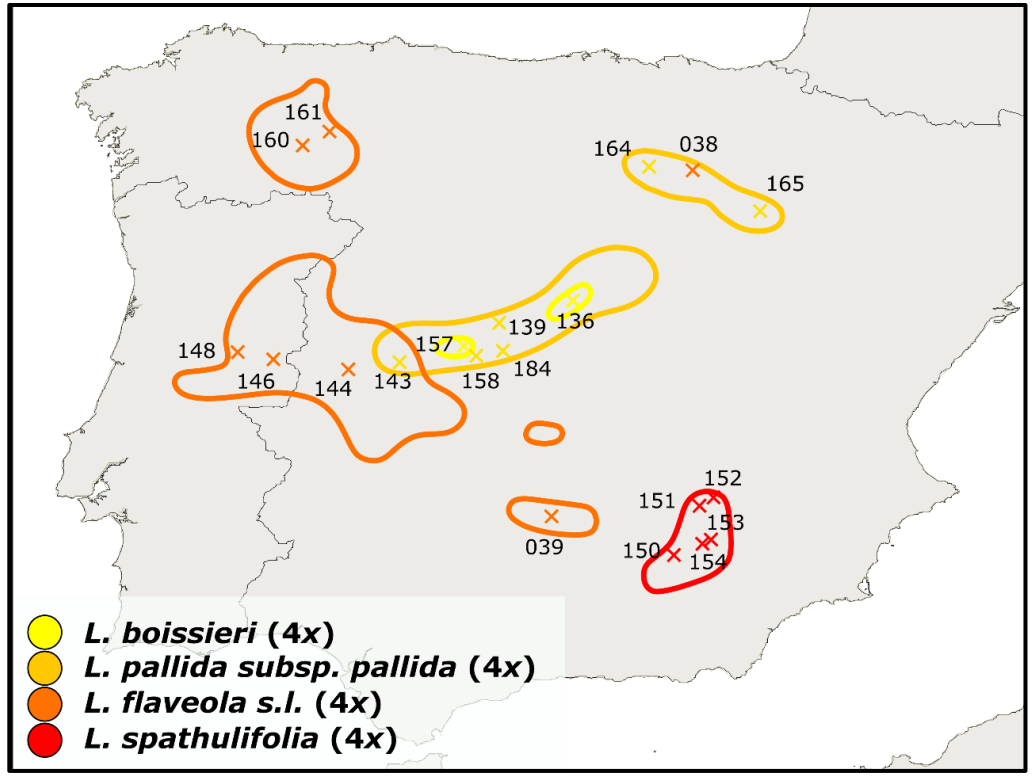


Tomasello & Oberprieler (2022). *Frontiers in Plant Sciences* 13:842842

***Leucanthemopsis 4x* species delimitation**

Marginal Likelihood

- Iberian tetraploid



L. flaveola

L. pallida subsp. pallida

L. boissieri

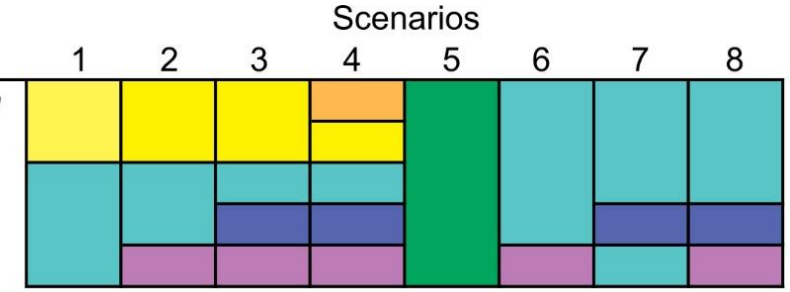
L. spathulifolia

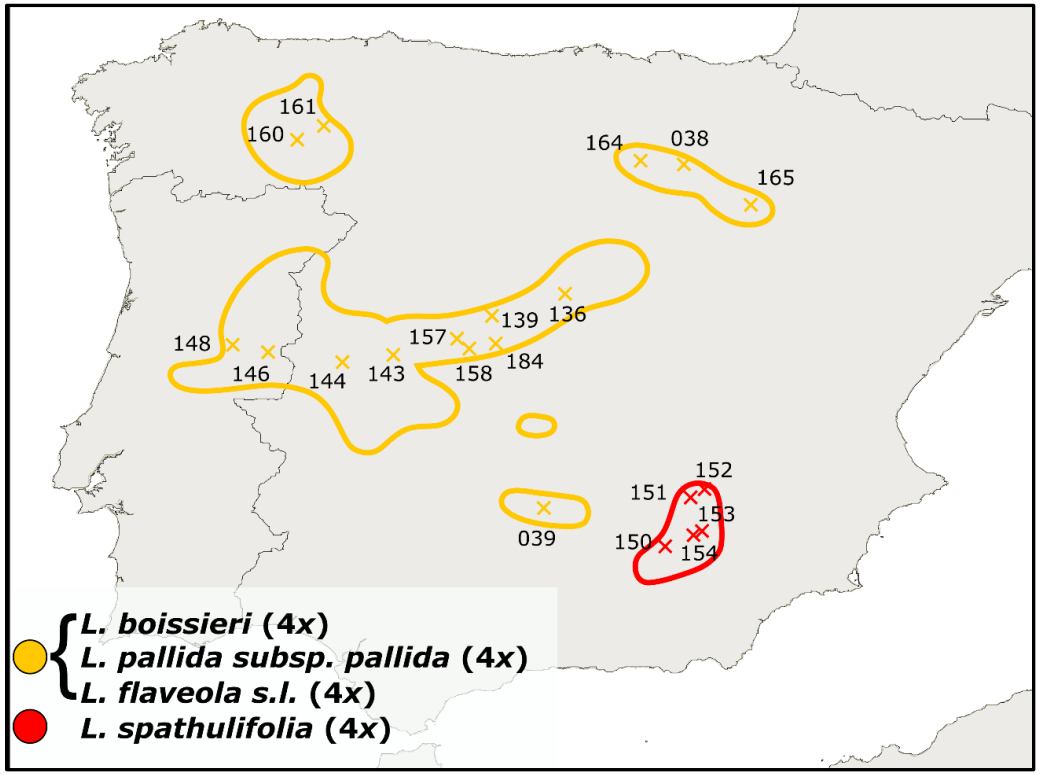
Tomasello & Oberprieler (2022). *Frontiers in Plant Sciences* 13:842842

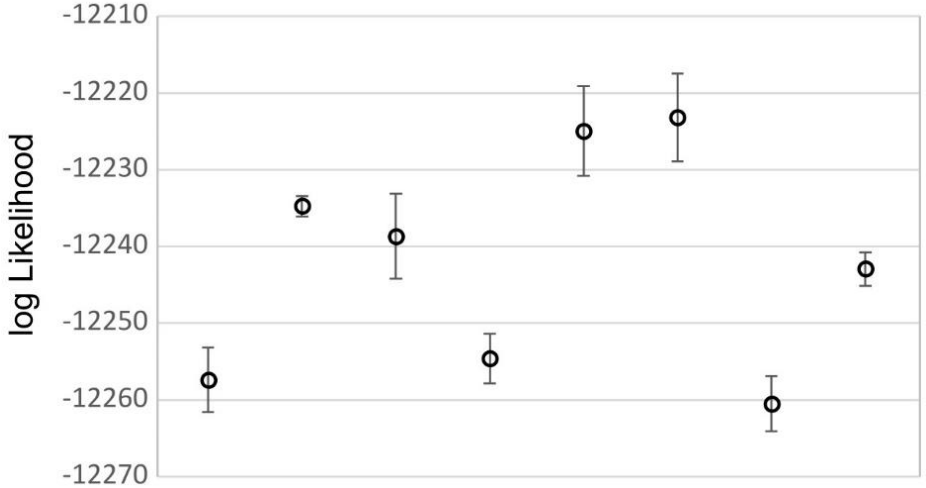
***Leucanthemopsis 4x* species delimitation**

Marginal Likelihood

Taxon identification	
Heywood (1975)	Pedrol (2019)
L. flaveola	*L. flaveola* subsp. *flaveola*
L. flaveola	*L. flaveola* subsp. *ricoi*
L. pallida subsp. *pallida* var. *pallida*	*L. pallida* subsp. *pallida*
L. pallida subsp. *pallida* var. *alpina*	*L. boissieri*
L. pallida subsp. *spathulifolia*	*L. spathulifolia*







Tomasello & Oberprieler (2022). *Frontiers in Plant Sciences* 13:842842

Coalescent-based species delimitation

When they may fail:

- 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, OH43210-1293, USA, †Department of Biological Sciences, Louisiana State University, Life Sciences Building, Baton Rouge, LA 70803, USA

PNAS

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)

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-

Syst. Biol. 69(1):184–193, 2020

DOI:10.1093/sysbio/syz042
Advance Access publication June 10, 2019

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: eachambers@utexas.edu.*

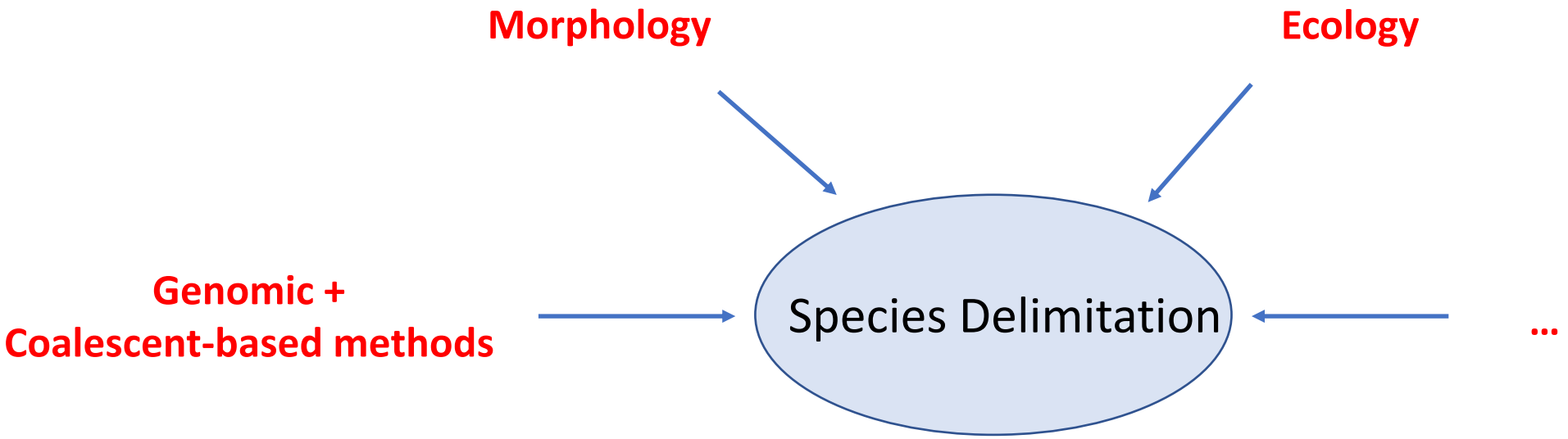
Received 19 November 2018; reviews returned 22 May 2019; accepted 24 May 2019
Associate Editor: Richard Glor

Coalescent-based species delimitation

When they may fail:

- Presence of gene flow
- Selectively driven divergence

Integrative approaches always recommended!!



Coalescent-based species delimitation

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?**

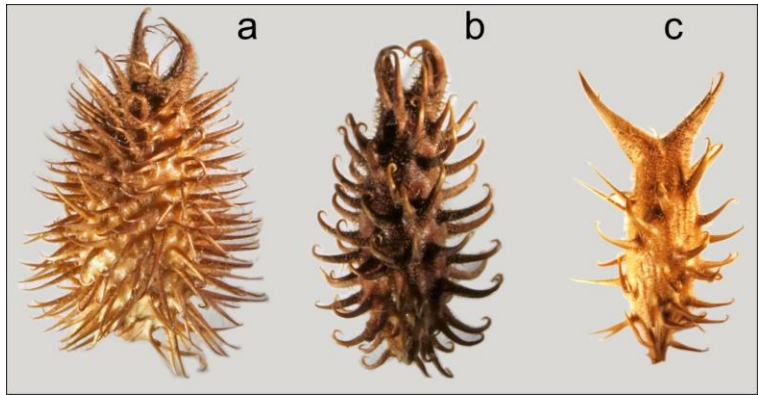
520. Xanthium Strumarium L.

Burweed; G.

Species delimitation *Xanthium* L.

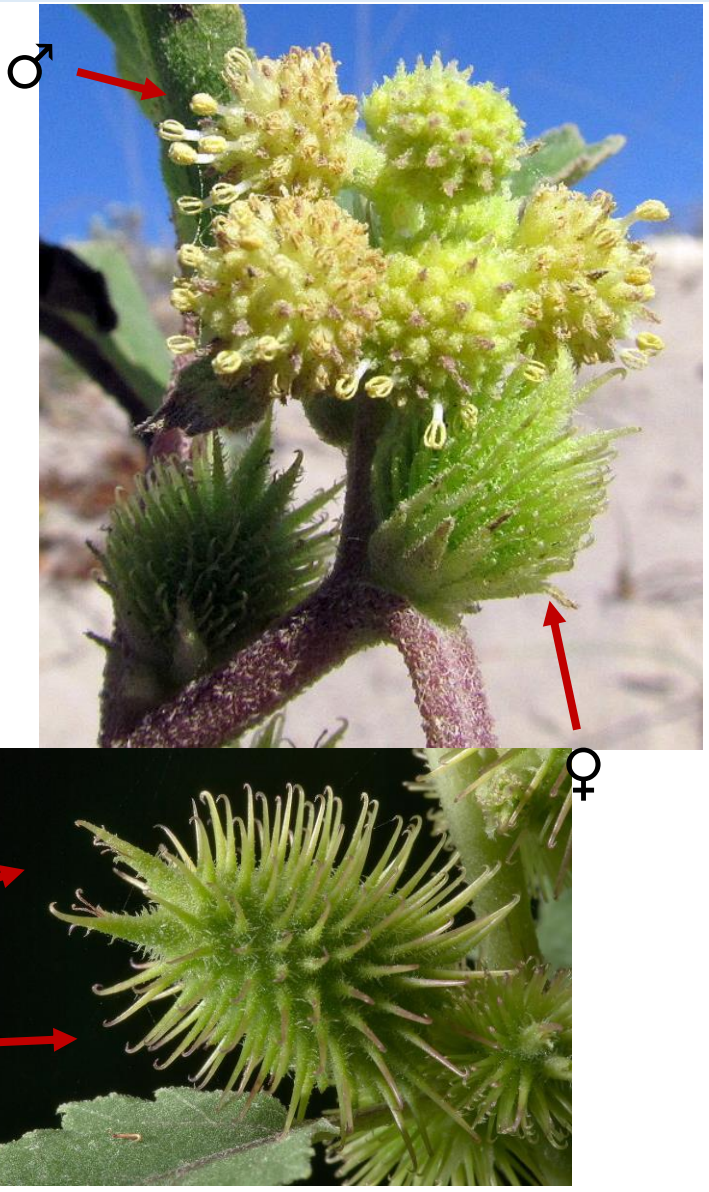
Why *Xanthium*?

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



520. Xanthium Strumarium L.

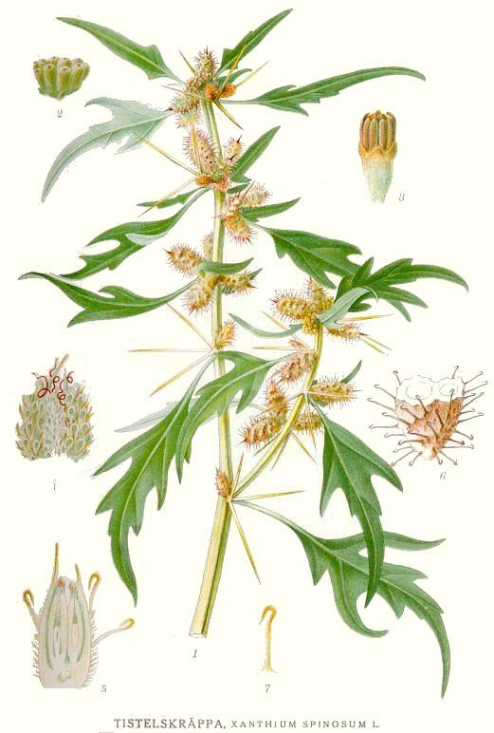
Burweed; G.



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

http://www.actaplantarum.org/

Sectio **Euxanthium**

Subectio **Orthorrhyncha**

- Beaks strait

Subectio **Campylorrhyncha**

- Beaks hooked

Serie **Glabrata**
(Glabrous burs)

Serie **Hispida**
(Burs with hairs and glands)

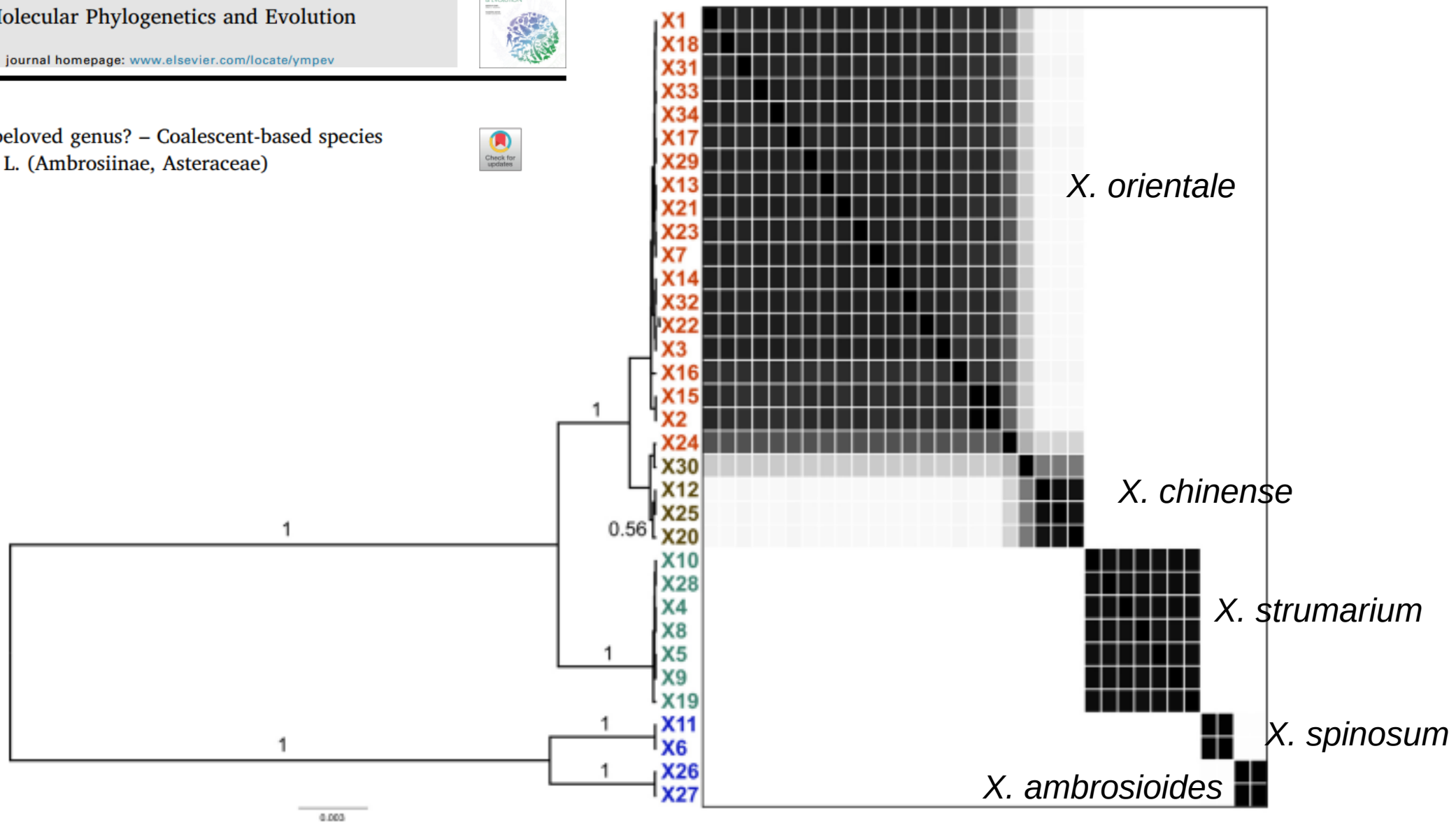
Contents lists available at ScienceDirect

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)

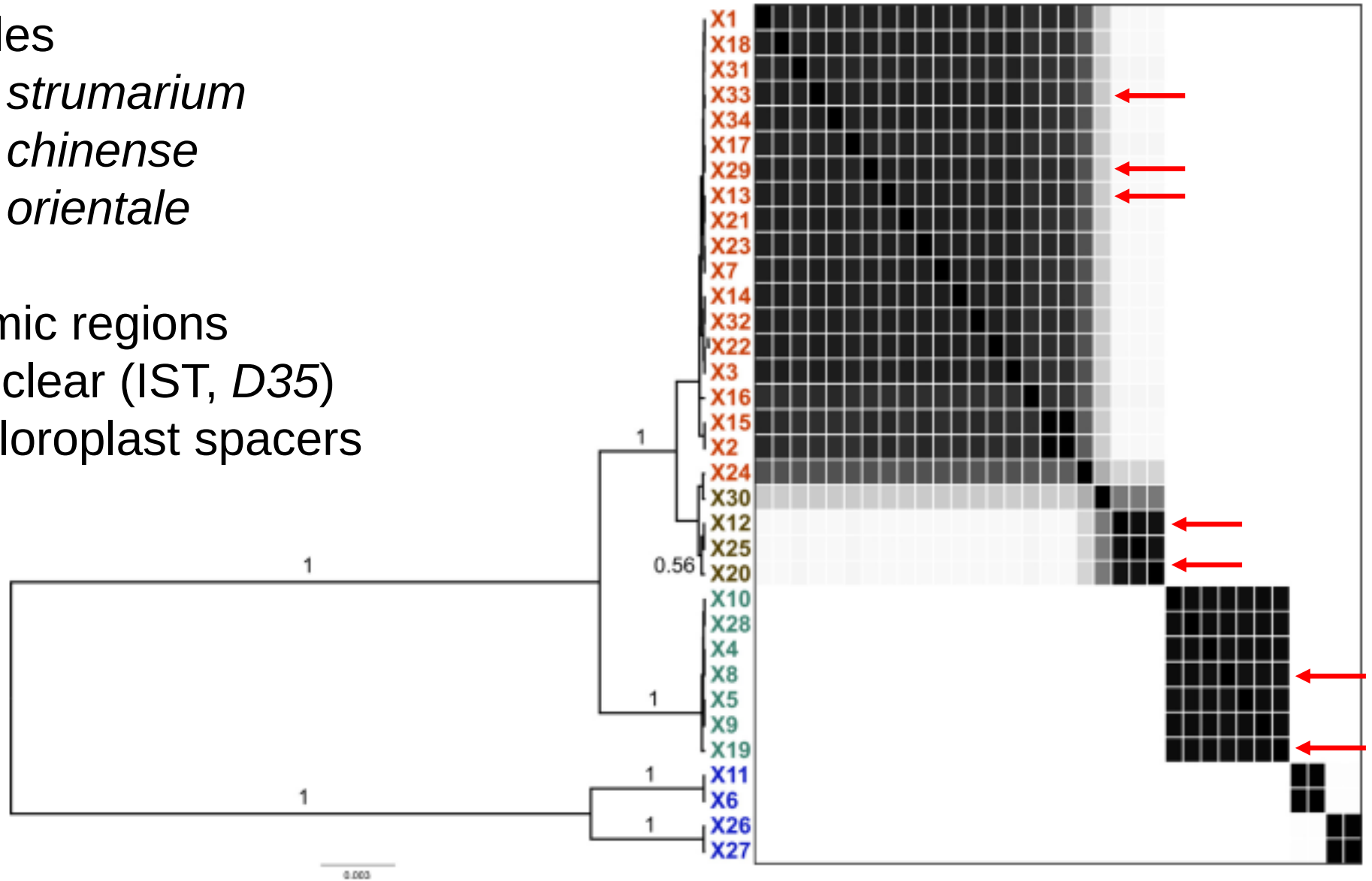
Salvatore Tomasello



Samples fot today:

- 7 samples
  - 2 *X. strumarium*
  - 2 *X. chinense*
  - *3 X. orientale*

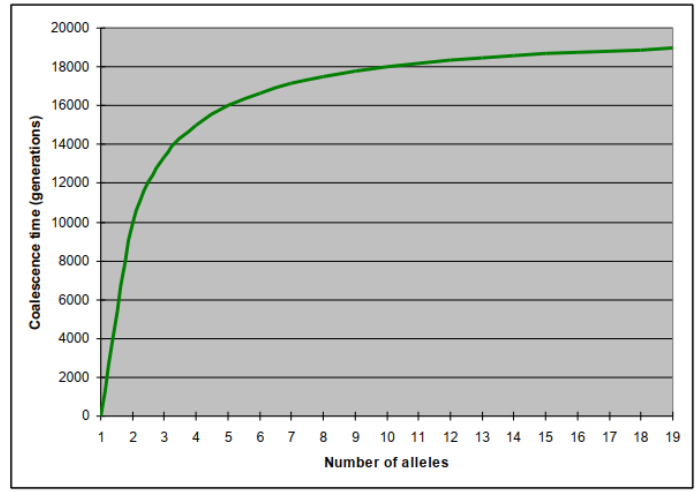
- 4 genomic regions
  - 2 nuclear (IST, *D35*)
  - 2 chloroplast spacers



The General Mixed Yule Coalescent method (GMYC)

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

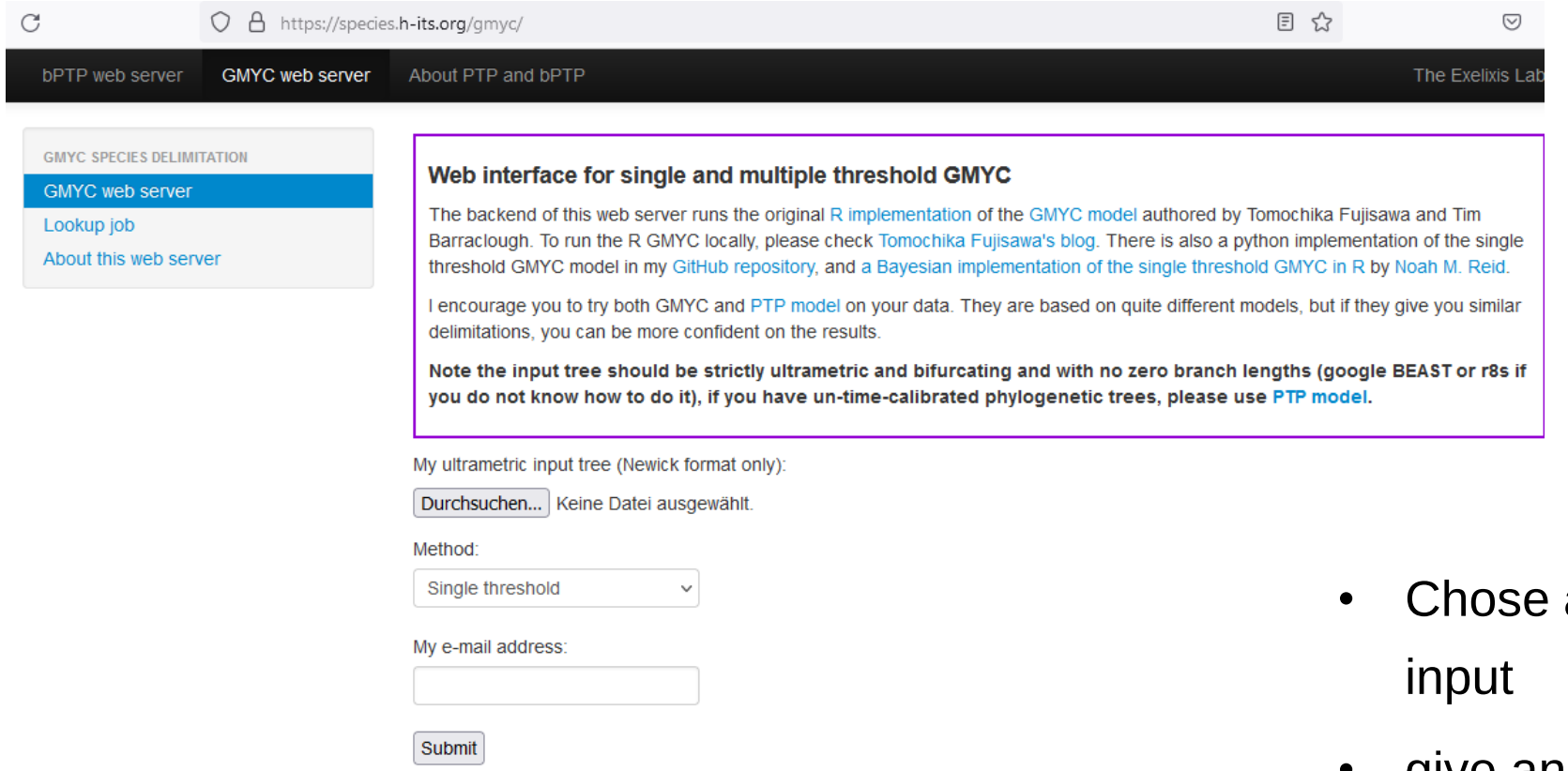
- Maximum likelihood method (ML)
- Based on the Yule process
- Single locus data
- Ultrametric tree as input
- Web server: http://species.h-its.org/gmyc/
- Available at: https://sites.google.com/site/noahmreid/software



Pons et al. 2005. *Systematic Biology* 55: 595-609.
Reid & Carstens. 2012.. *BMC Evolutionary Biology* 12: 196

The General Mixed Yule Coalescent method (GMYC)

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

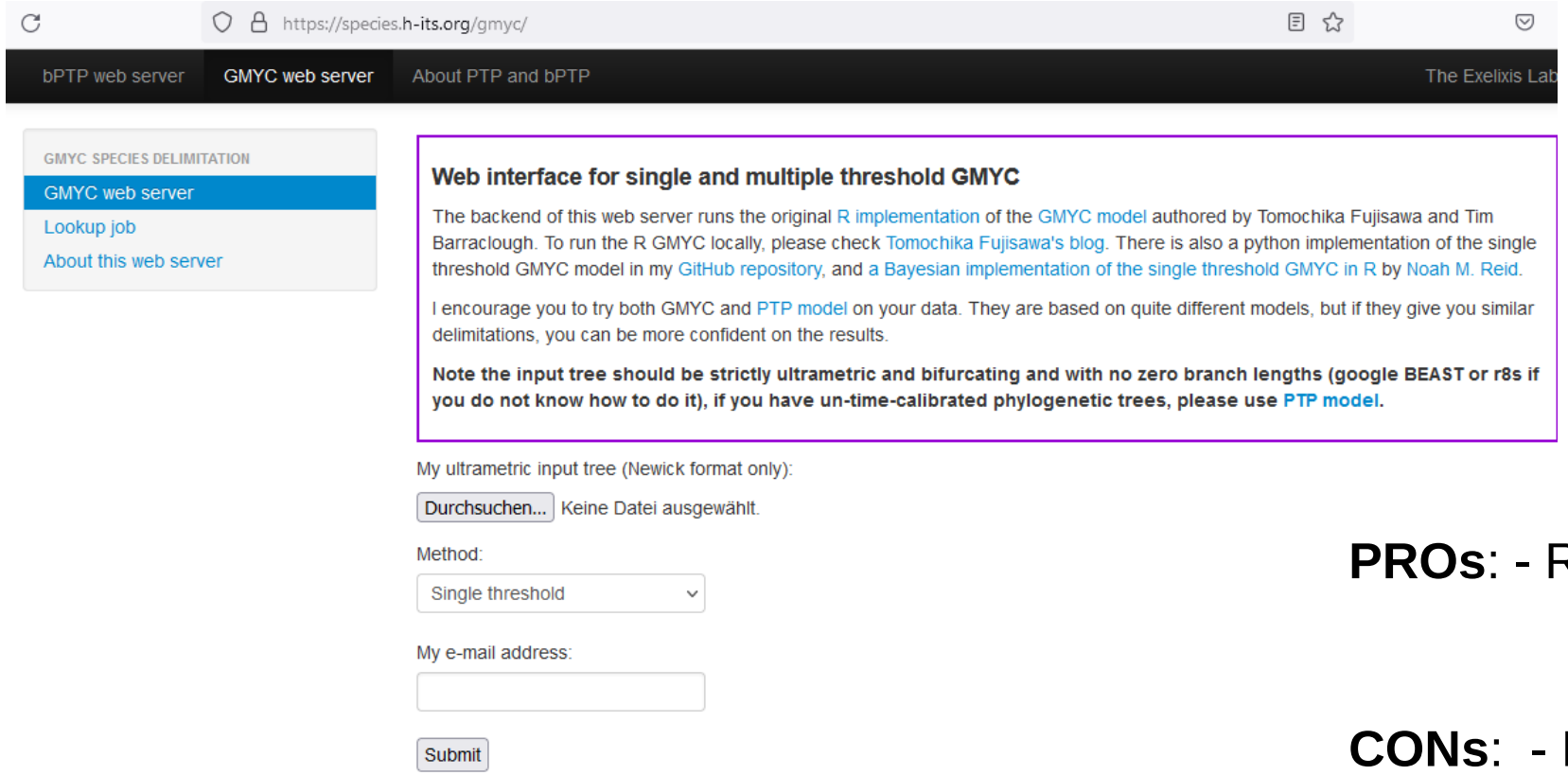


- Chose an ultrametric single tree as input
- give an email address and start the analyses

Pons et al. 2005. *Systematic Biology* 55: 595-609.
Reid & Carstens. 2012.. *BMC Evolutionary Biology* 12: 196

The General Mixed Yule Coalescent method (GMYC)

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



PROs: - Relatively fast for large dataset

CONs: - Relay on single loci

- Accuracy of ultrametric trees

- Oversplitting

Pons et al. 2005. *Systematic Biology* 55: 595-609.
Reid & Carstens. 2012.. *BMC Evolutionary Biology* 12: 196

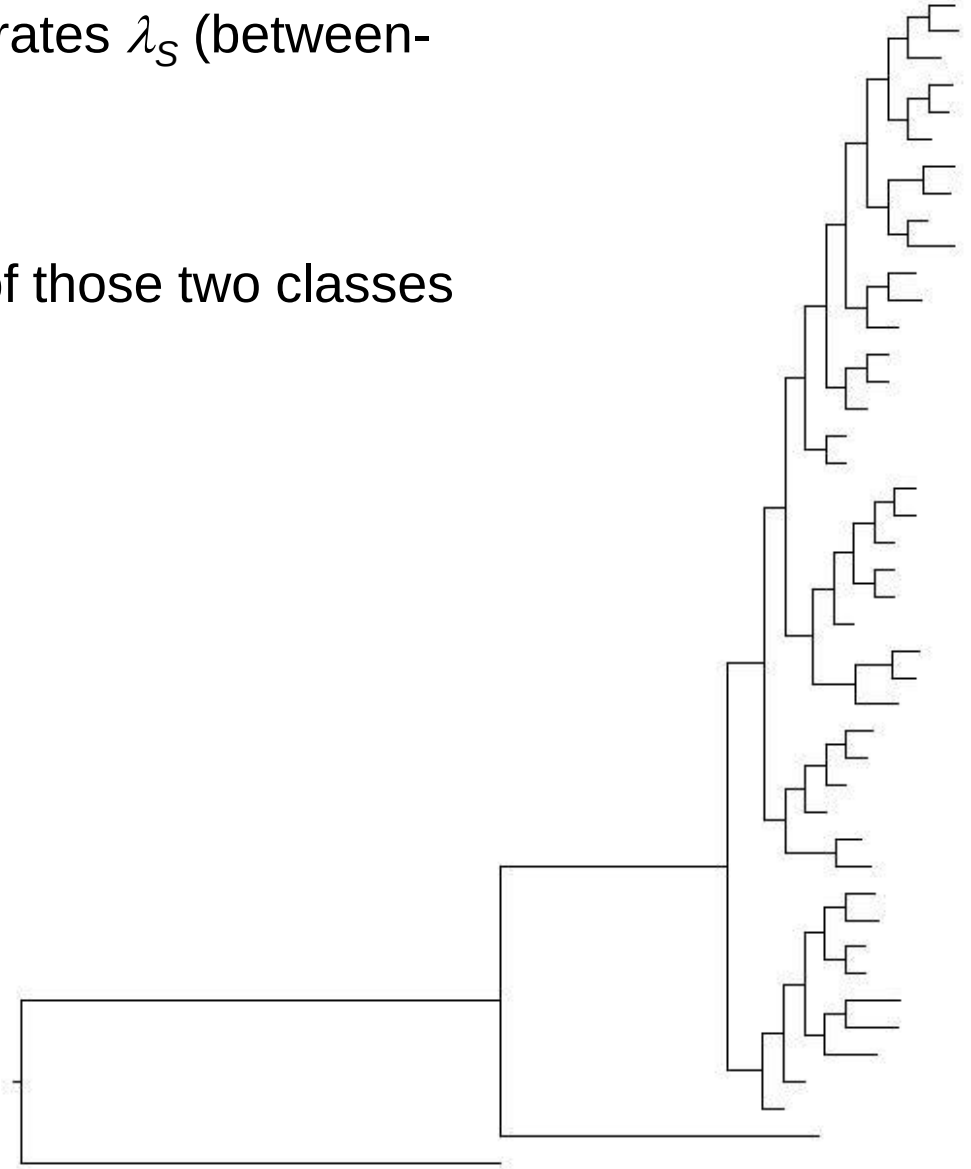
Poisson Tree Process (PTP)

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

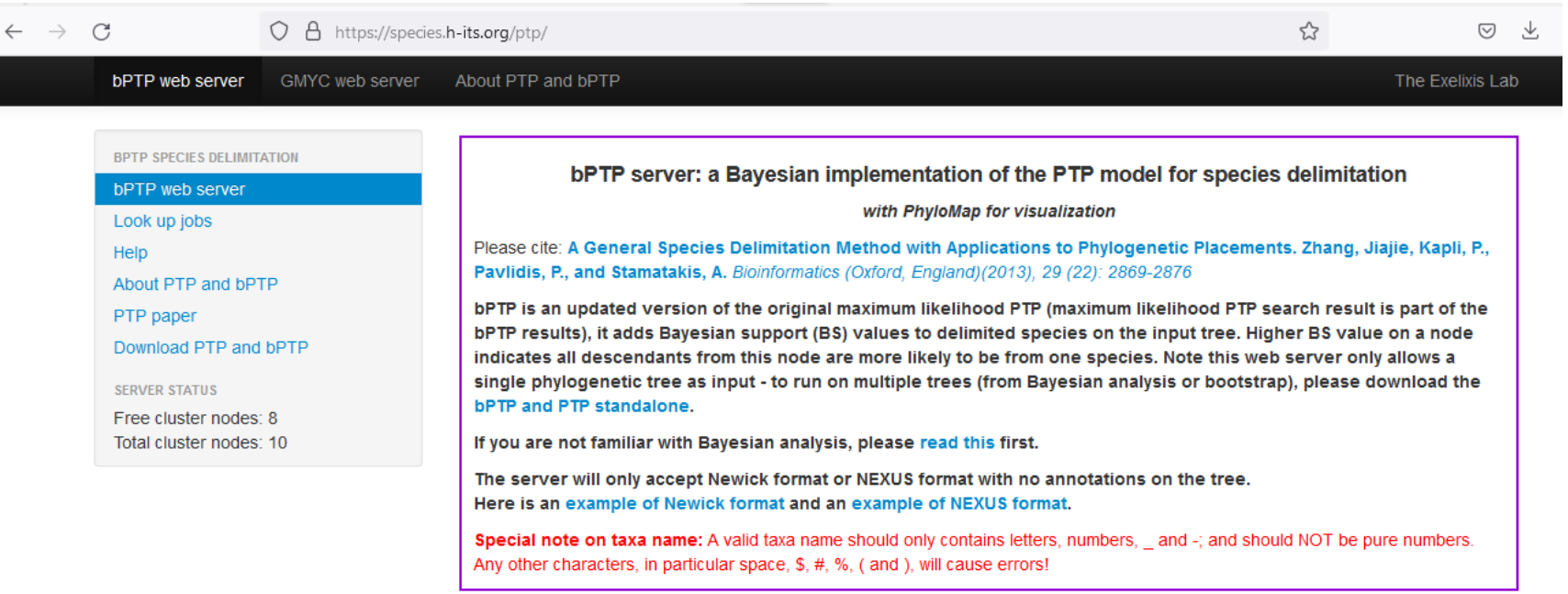
Web server: http://species.h-its.org/gmyc/



Zhang et al. 2013. *Bioinformatics* 29 : 2869-2876

Poisson Tree Process (PTP)

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



PROs: - Relatively fast for large dataset

CONs: - Relay on single loci / concatenation

- Implicitly assumes rec. monophyly

- Oversplitting

Zhang et al. 2013. *Bioinformatics* 29 : 2869-2876