EVOLUTION AND GENOMICS
Intensive and immersive training opportunities

2020 WORKSHOP ON POPULATION AND SPECIATION GENOMICS, CESKY KRUMLOV

The Multi-Species Coalescent (MSC) and its Application in Phylogenetics and Species Delimitation

L. Lacey Knowles
Dept. of Ecology and Evolutionary Biology
University of Michigan

Software: *Delineate*

Jeet Sukumaran
Dept. of Biology, Evolutionary Biology Program
San Diego State University

Software: *Decrypt*

Arnaud Becheler
Dept. of Ecology and Evolutionary Biology
University of Michigan

EVOLUTION AND GENOMICS

Intensive and immersive training opportunities

2020 WORKSHOP ON POPULATION AND SPECIATION GENOMICS, CESKY KRUMLOV

Transformative potential of model-based analyses:

- Codon substitution and analysis of natural selection
- Adaptive molecular evolution
- Divergence time estimation and biogeographic analysis
- Phylogenetic inference
- Species delimitation
- Demographic inference

Transformative potential of model-based analyses:

What I'll emphasize is the importance of recognizing:

- Decisions/choices we make about model formulation

- The subjectivity of model formulation
  itself when making inferences

- Decisions when applying to empirical study
  (e.g., the data type, subsets of data, what subset of data)

- All models are flawed..., but ...
  models are how we communicate our knowledge
  to a statistical apparatus

Transformative potential of model-based analyses:

(i) Phylogenetic inference

(ii) Species delimitation/inferring species boundaries

With an emphasis on:

- Choices we make about model formulation
- Recognizing the subjectivity of model formulation itself when making inferences
- Decisions when applying to empirical data

Transformative potential of model-based analyses:

(i) Phylogenetic inference

(ii) Species delimitation/inferring species boundaries

(iii) Phylogeography/Comparative Phylogeography

"Species delimitation" is a computational approach to identifying species units in nature. Identification of these units is critical to many areas in evolutionary biology — systematics, phylogeography, biogeography, ecology, conservation, etc. — as well as having impacts in a broader range of areas, such as human health and epidemiology, natural resource management, and so on.

Traditional approaches to species delimitation typically rely on models that identify structure in genomic data and identify "species" in nature by relating this structure to species boundaries.

Transformative potential of model-based analyses:

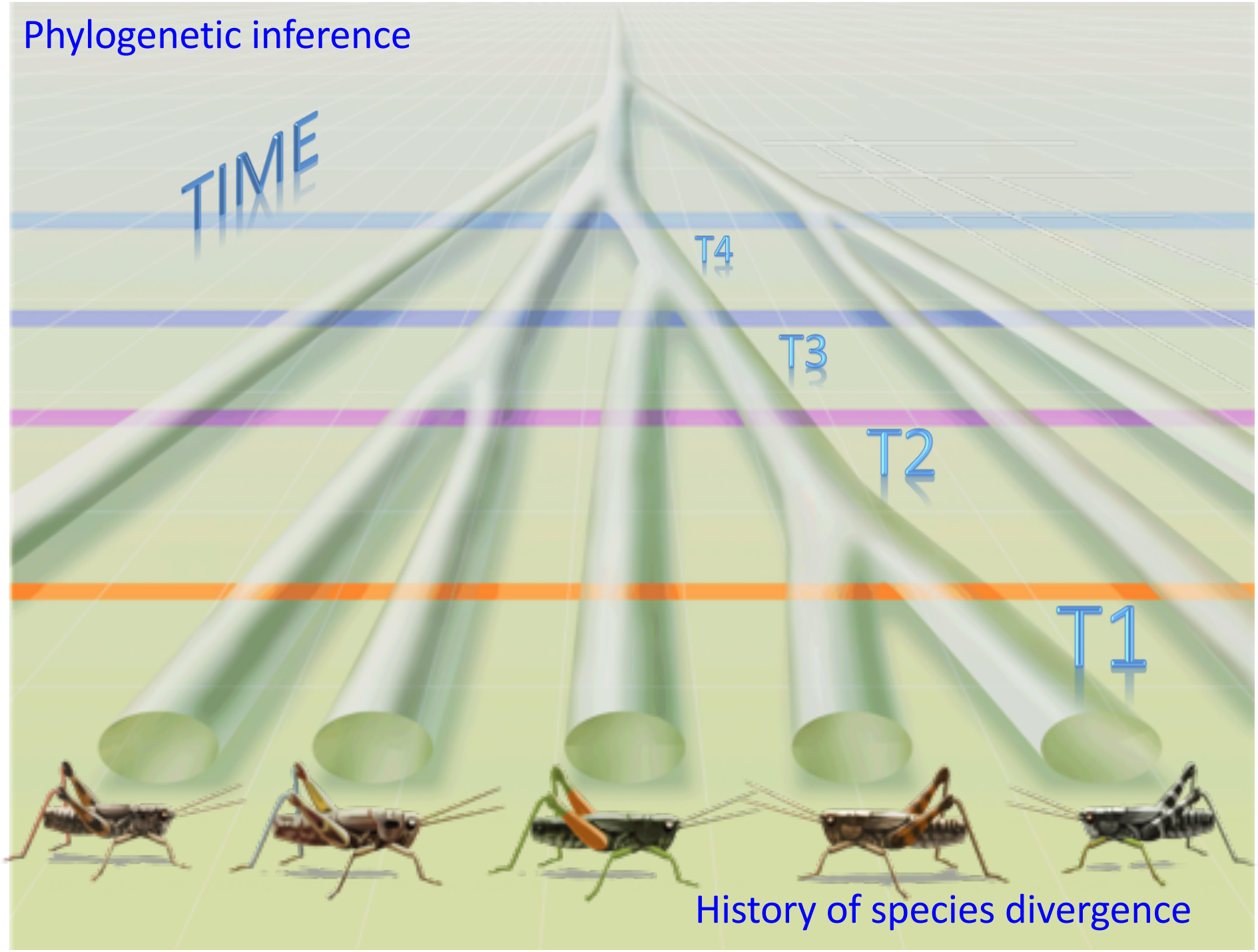
(i) Phylogenetic inference

(ii) Species delimitation/inferring species boundaries

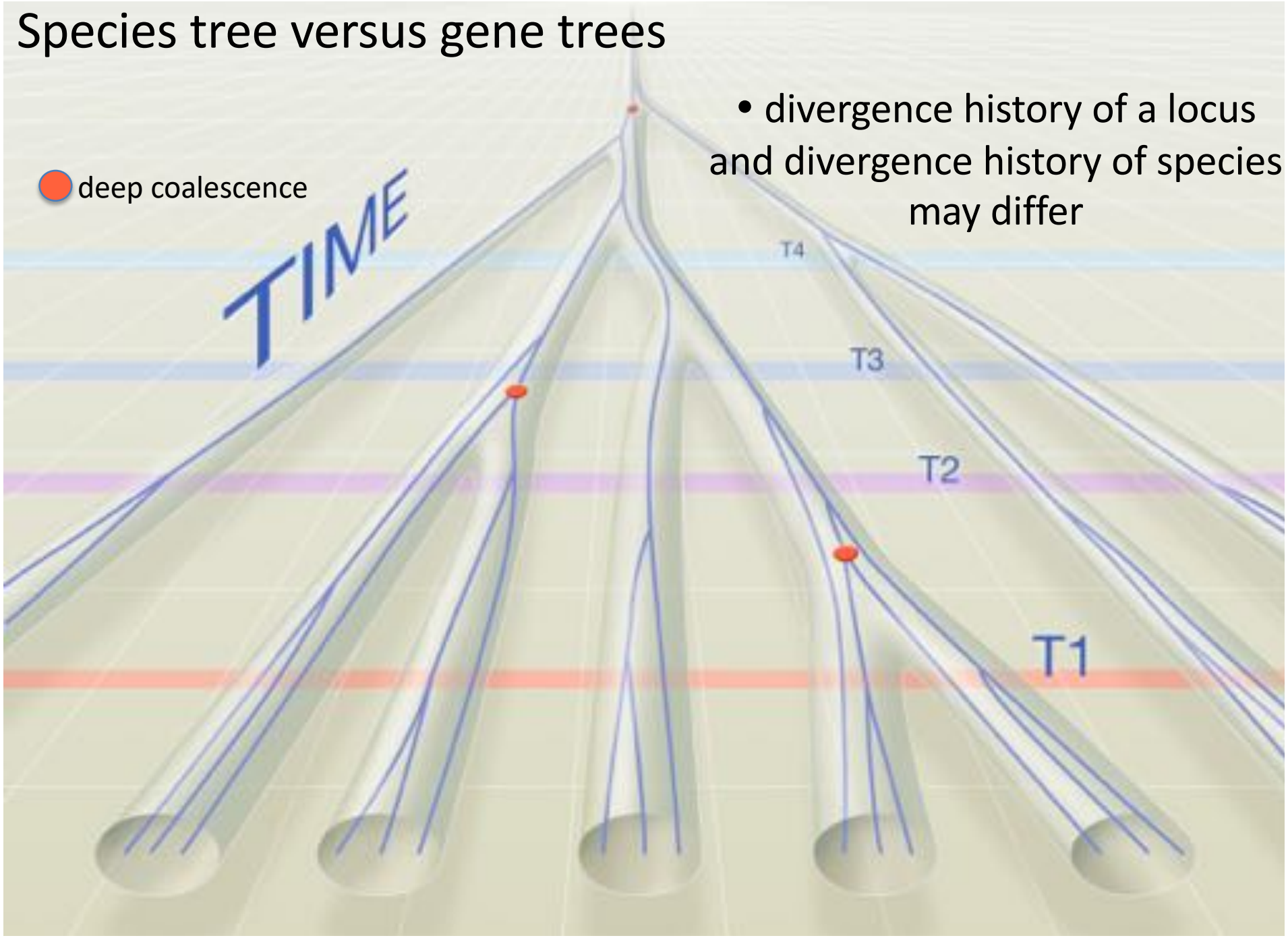
(iii) Phylogeography/Comparative Phylogeography

With an emphasis on:

- Choices we make about model formulation
- Recognizing the subjectivity of model formulation itself when making inferences
- Decisions when applying to empirical data

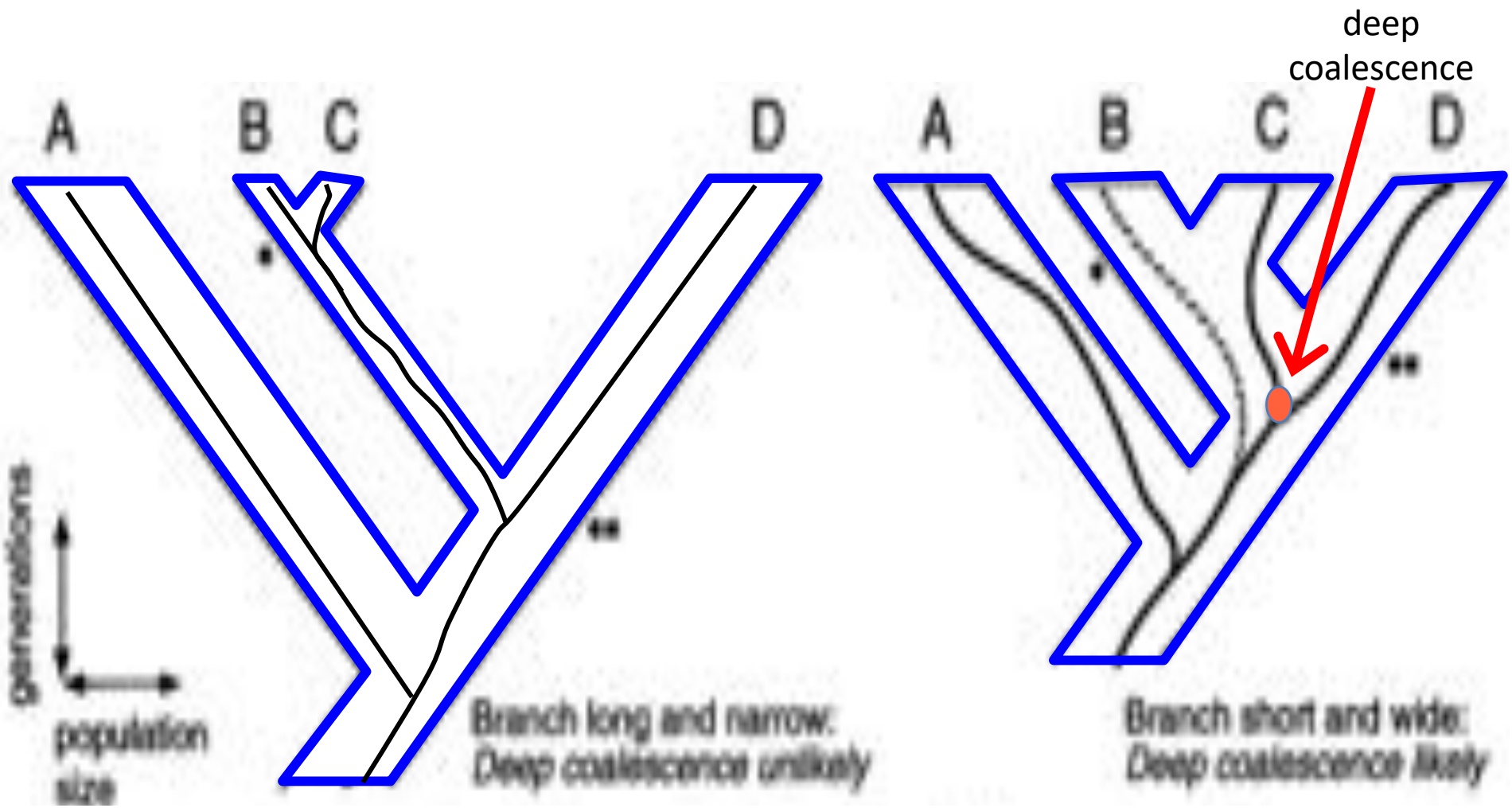

Phylogenetic inference
TIME
T4
T3
T2
T1
History of species divergence

Species tree versus gene trees



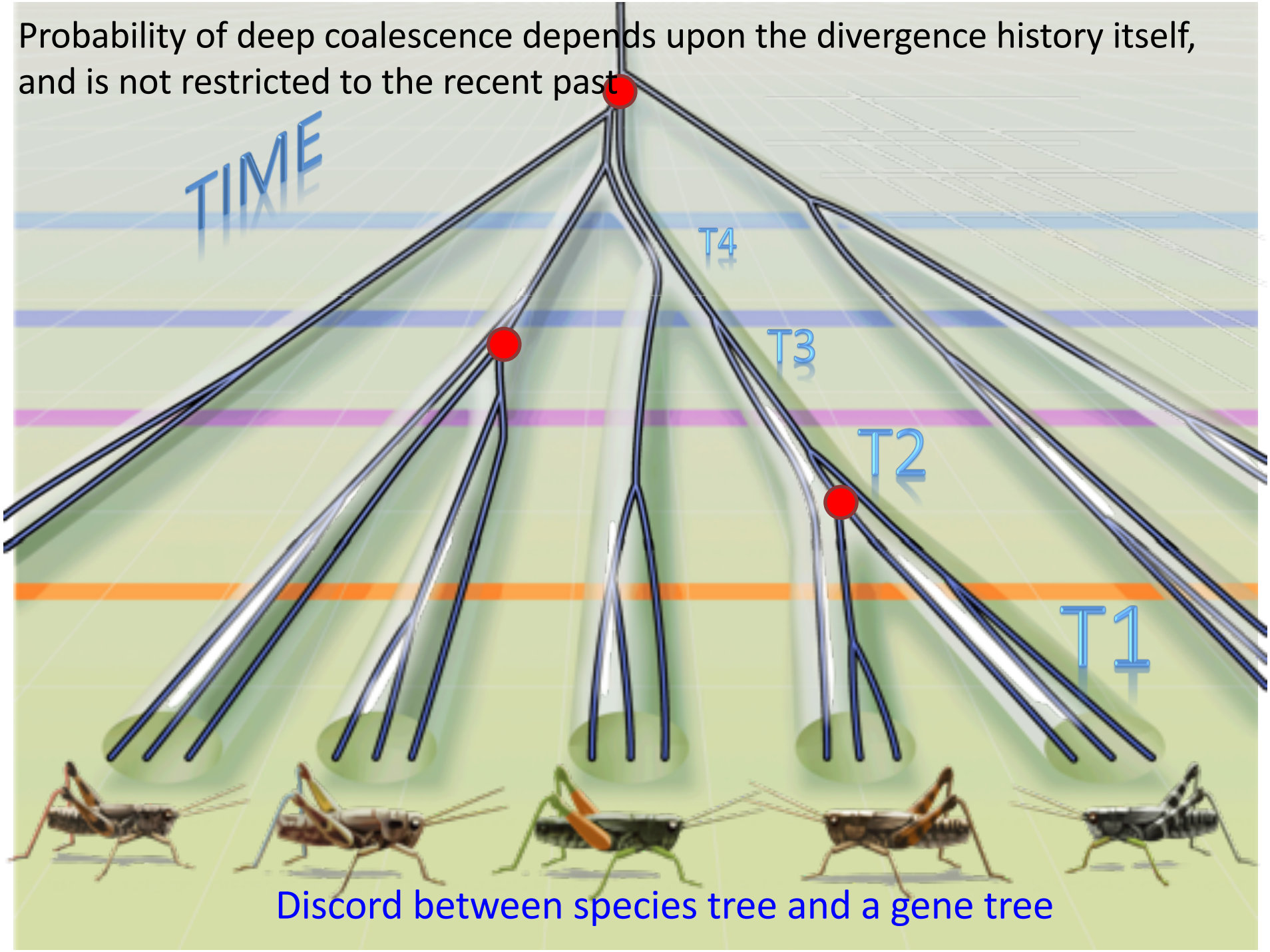
- divergence history of a locus and divergence history of species may differ

Probability of deep coalescence depends upon the divergence history itself



Maddison 1997

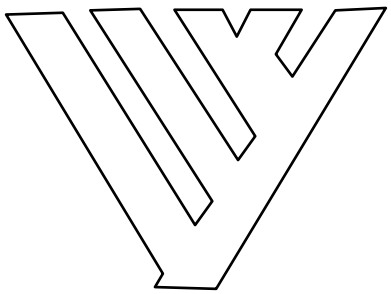
Probability of deep coalescence depends upon the divergence history itself, and is not restricted to the recent past



Discord between species tree and a gene tree

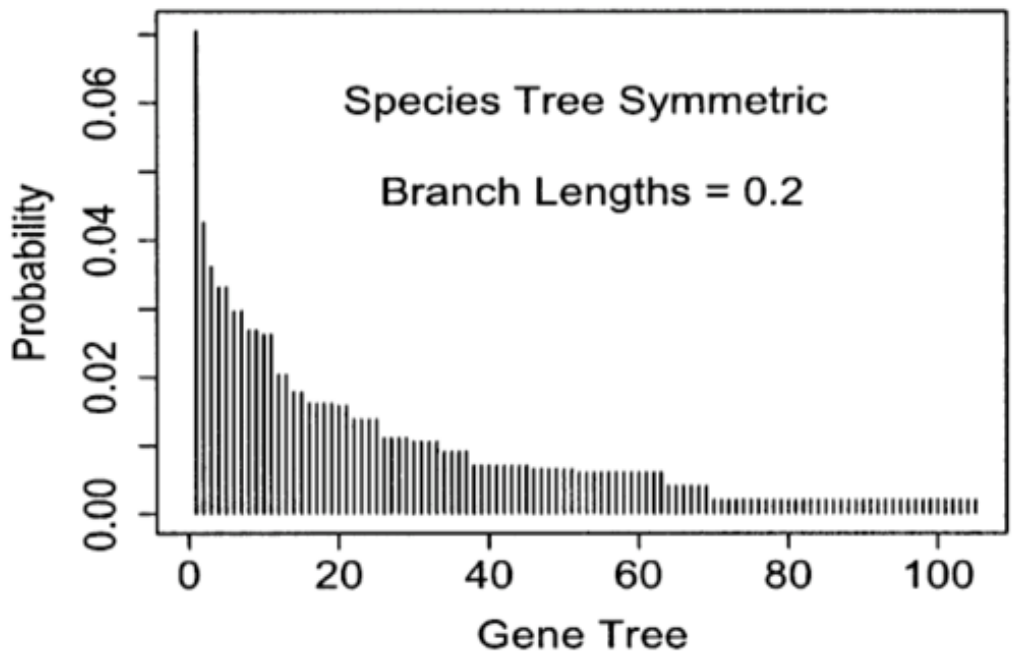
While there is a distribution of possible gene trees for a given species (or population) tree, the probabilities of gene trees differ.

5 taxon species tree



105 possible gene tree topologies

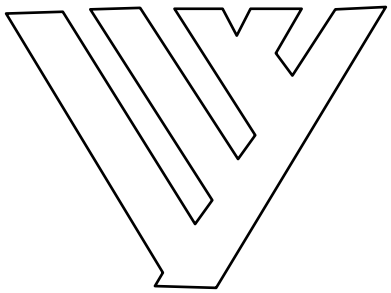
high $P(G_{tree}|S_{tree})$ low $P(G_{tree}|S_{tree})$



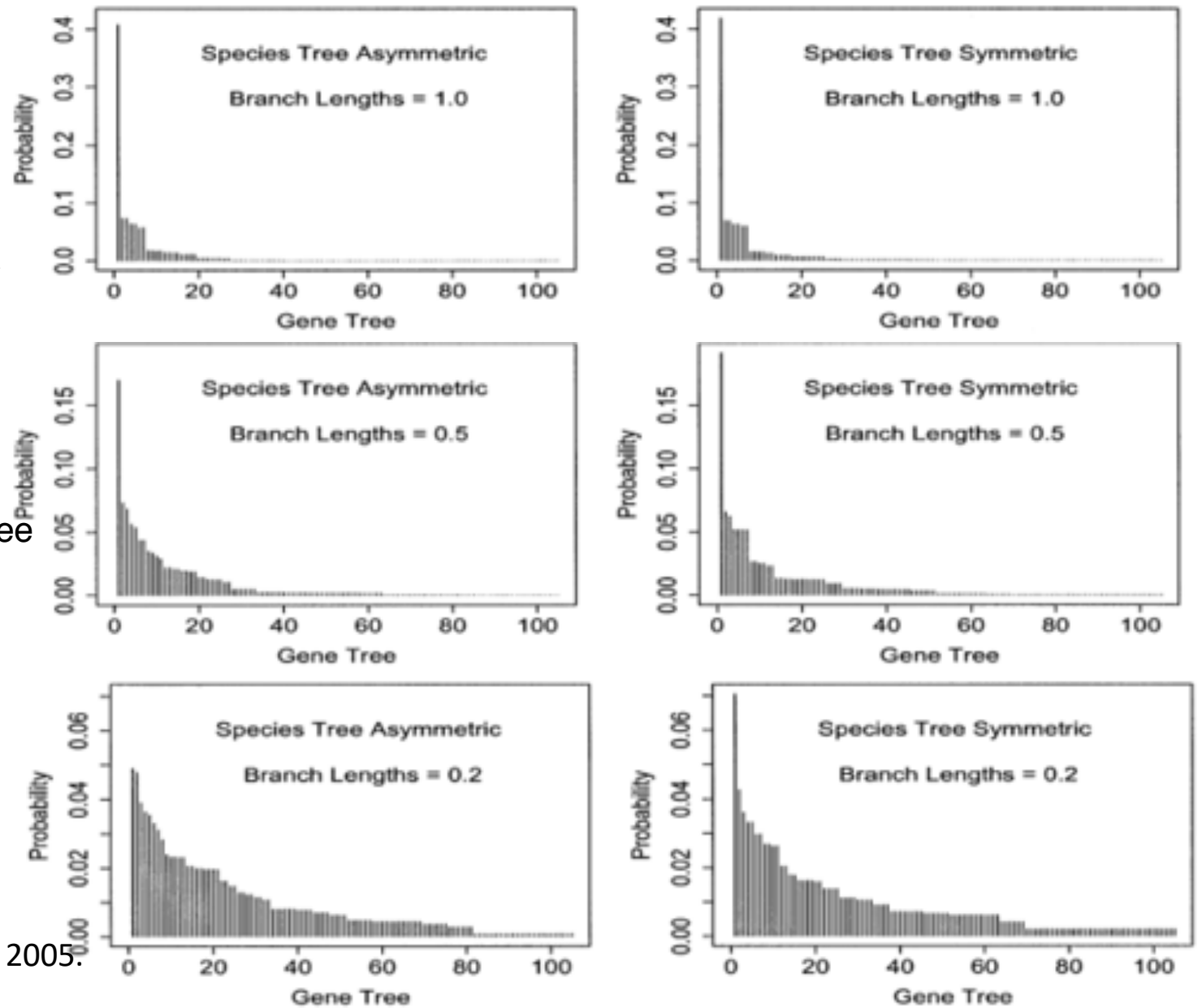
Degnan & Salter (2005) Evolution

Gene tree distributions under the coalescent process

5 taxon
species tree



105 possible gene tree
topologies



Degnan, J and L Salter. 2005.
Evolution 59: 24-37.

Directly estimating species trees (as opposed to equating a gene tree with the species' phylogenetic history)

Maddison (1997) Gene trees in species trees. *Syst. Biol.* 46:523-36.

Likelihood of a species tree

$$\prod_{\text{loci}} \sum_{\substack{\text{possible}\\ \text{gene trees}}} \overbrace{[P(\text{sequences}|\text{gene tree})}^{\text{the model of nucleotide evolution}} * \overbrace{P(\text{gene tree}|\text{species tree})]}^{\text{coalescent theory}}$$

- the species tree specifies the probabilities for various patterns of genetic descent (i.e., the distribution of gene trees)
- phylogeny as a composite, cloudlike nature of gene histories

Are gene trees that disagree with the species tree wrong?

Fundamental paradigm shift: instead of making inferences about species relationships from an estimated gene trees (or a tree based on a concatenated set of loci), we can DIRECTLY estimate the species tree.

$$L(ST) = \prod_{\text{loci}} \sum_{\substack{\text{possible} \\ \text{gene trees}}} [P(\text{sequences}|\text{gene tree}) * P(\text{gene tree}|\text{species tree})].$$

Maddison 1997

Proliferation of methods for species tree inference

- Computational considerations (# tips and # loci)
- Data type (SNP versus sequence data)

https://github.com/smirarab/ASTRAL

https://taming-the-beast.org/tutorials/StarBeast-Tutorial/
https://github.com/genomescale/starbeast2
https://github.com/cecileane/iBPP/

https://www.beast2.org/snapp/ DEMO 2-5pm

http://www.phylosolutions.com/tutorials/ssb2018/svdquartets-tutorial.html

Multiple processes produce discord among gene trees

https://bioinfocs.rice.edu/phylonet

https://github.com/crsl4/PhyloNetworks.jl

https://www.asc.ohio-state.edu/kubatko.2/software/HyDe/

DON’T HAVE ANY APPROACHES FOR PHYLOGENETIC INFERENCE THAT MODEL MULTIPLE CAUSES OF DISCORD

Factors affecting species-tree accuracy:

- history of diversification itself
- total sampling effort of sequences
- sample design (# of individuals versus loci)
- method of analysis
- level of genetic variation (mutation rate)

Simulation approach: compare known with estimated species tree (i.e., the accuracy of species-tree estimate) to examine the affects of each factor and their relative importance.

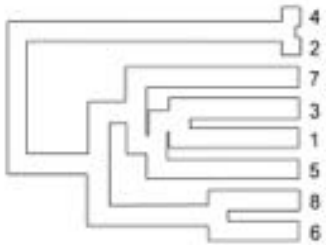
Simulation approach for evaluating accuracy of species-tree estimates

Conceptual design: Maddison & Knowles 2006, *Syst Biol*

Known species tree

A

500 species trees



1000 bp

Estimated species tree

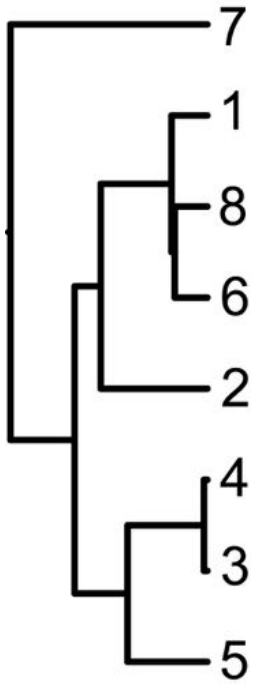
*Beast, Bucky, Best

Discordant gene trees retain significant phylogenetic signal

1N = total tree depth of 80,000 years, with N_e of 80,000

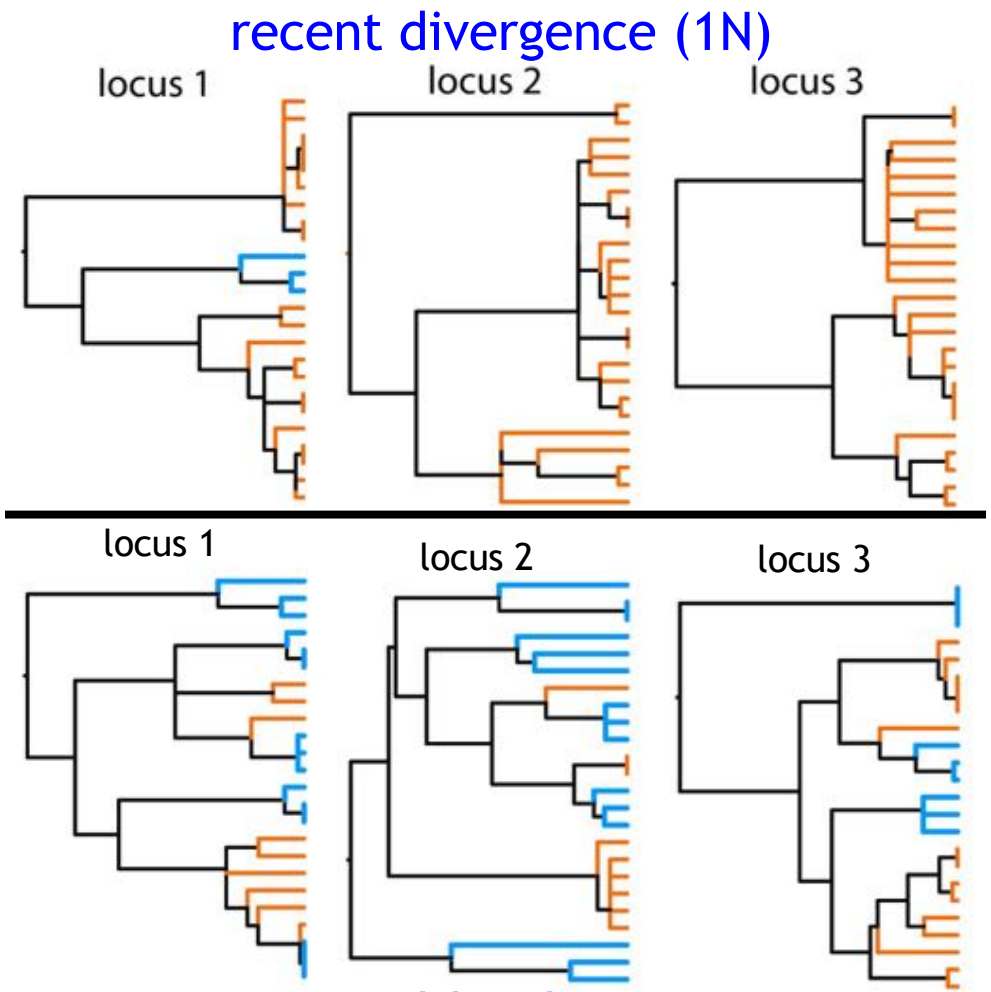
1 species every 10,000 years!

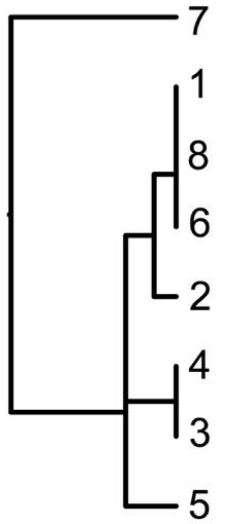
estimated ML species tree
(using STEM, Kubatko)

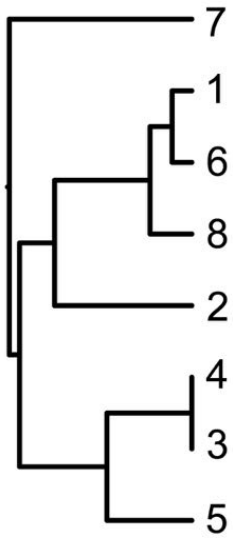


known species tree

#46 of 500







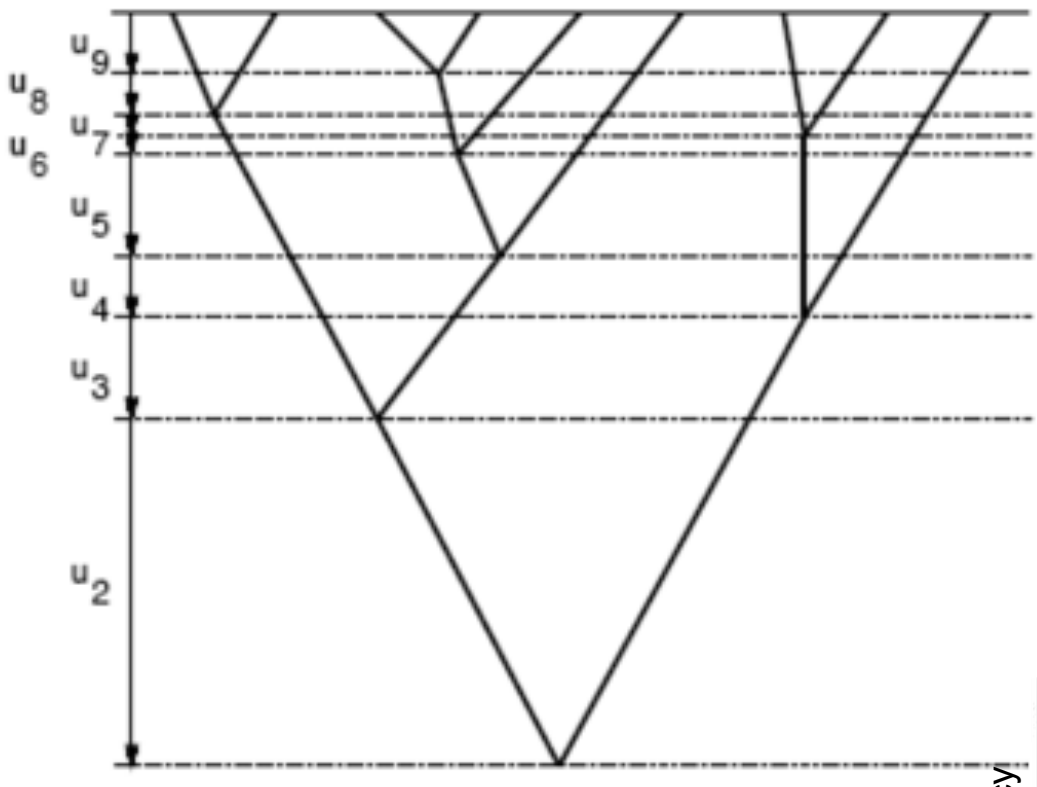
McCormack et al. 2010

Total sampling effort? Loci versus individuals?

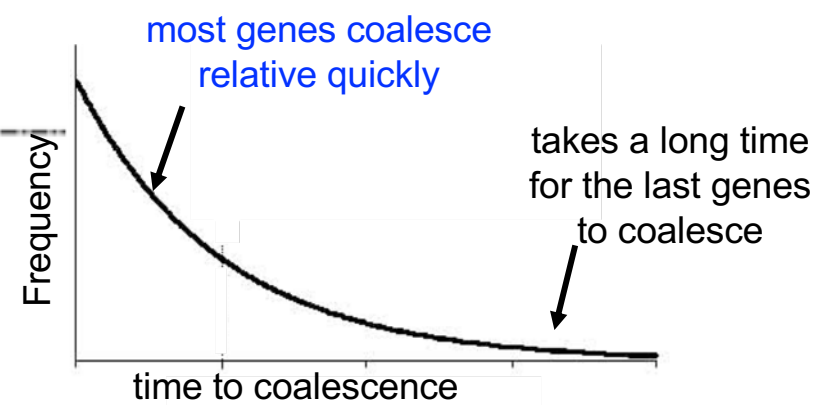
Sample multiple individuals? Only for recent divergence histories

Coalescent trees of gene copies within species (Kingman, 1982)

- Random collision of lineages as go back in time
- Collision is faster the smaller the effective population size

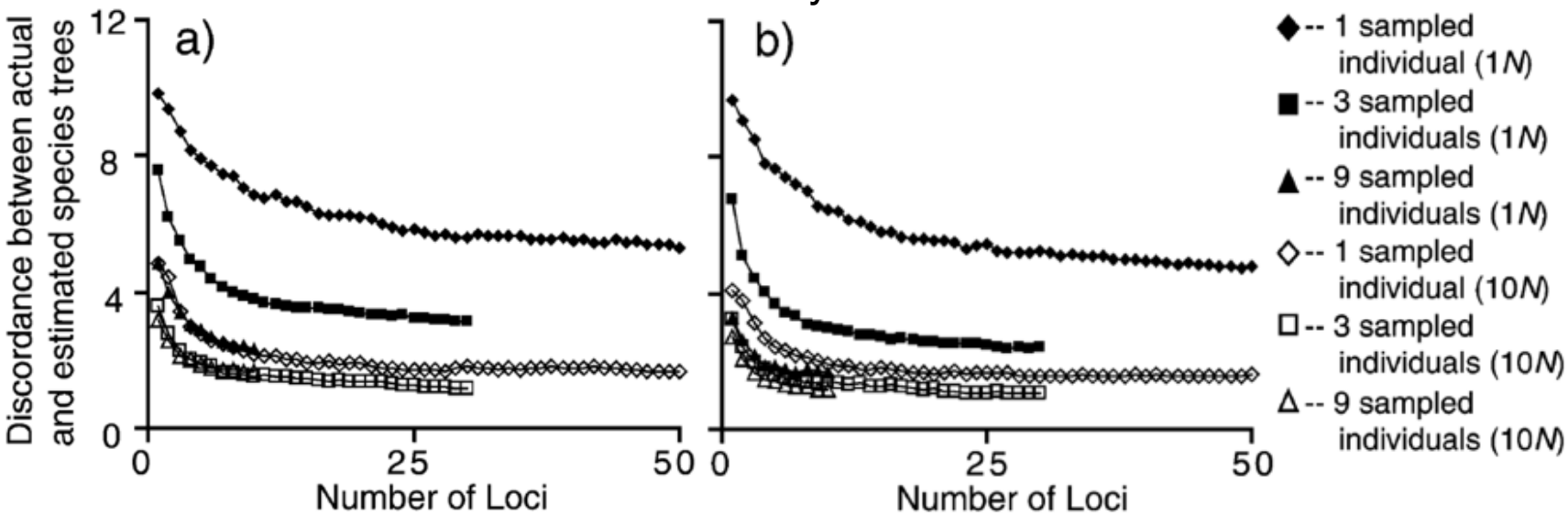


In a diploid population of effective population size N_e

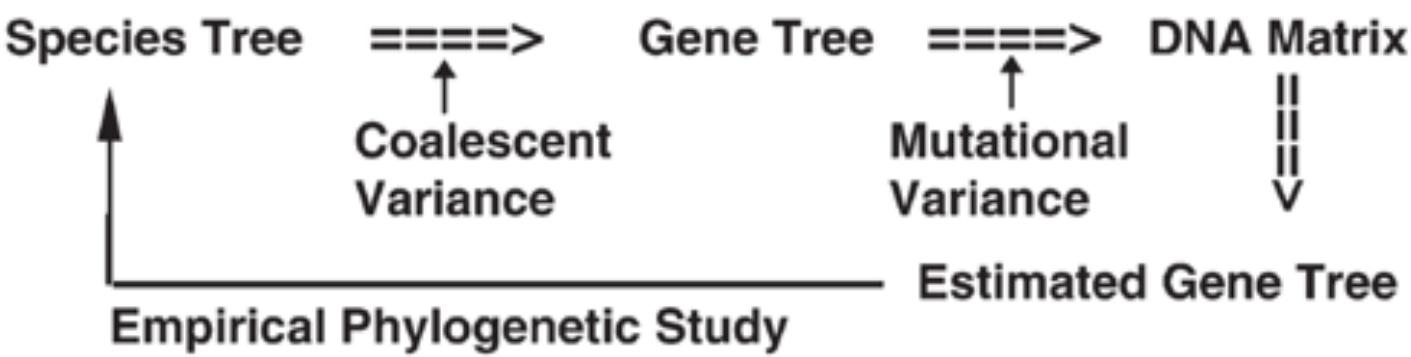


Total sampling effort? Loci versus individuals?

Only coalescent variance

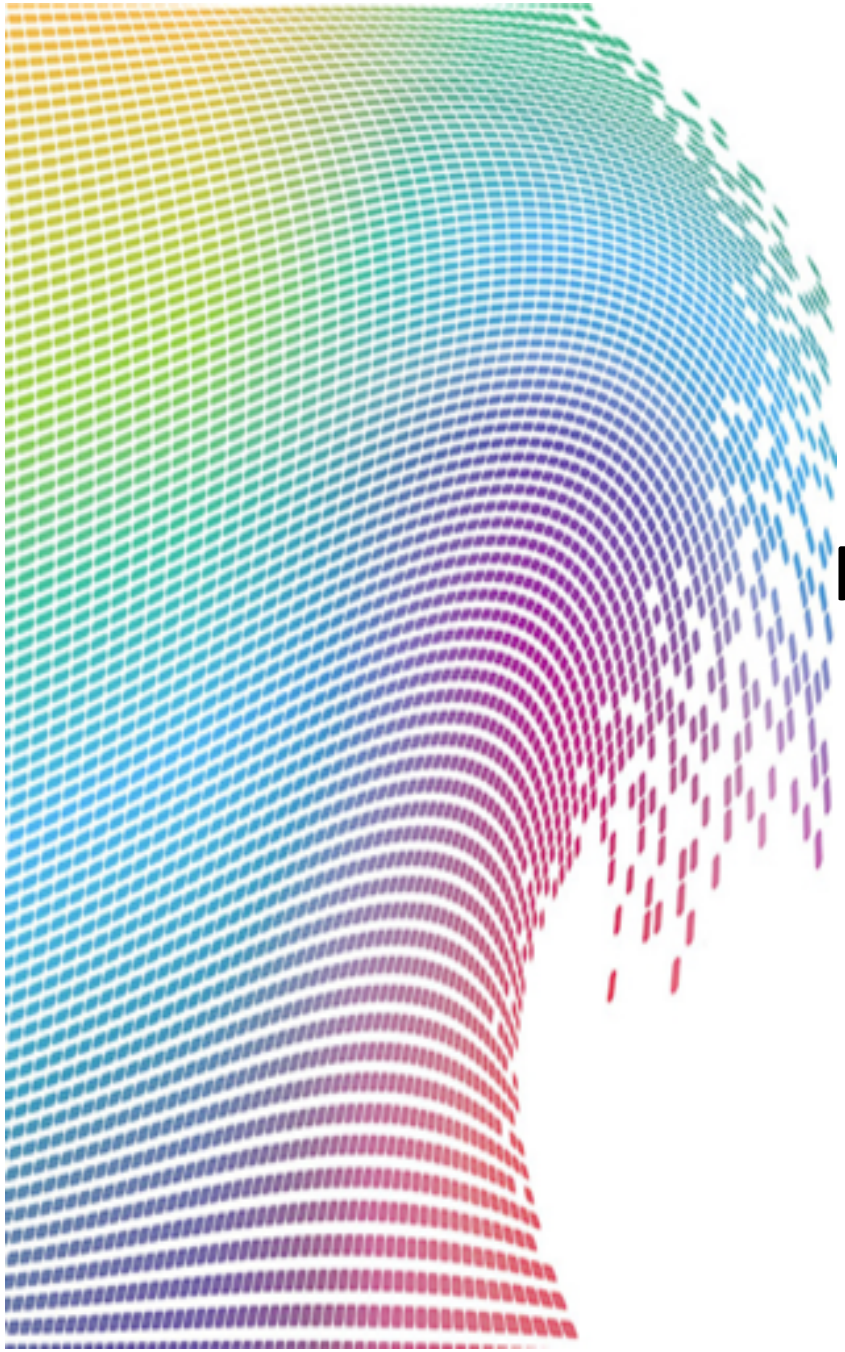


500 simulated species trees with 8 taxa



Huang et al. (2010) *Syst. Biol.* 59:573

Genomic data



Resolved accurate phylogenetic relationships among species?

- NO – Recalitrant nodes across the tree of life

Phylogenomics and Next-Generation Inferences: the Future of Phylogenetics in the Era of Big Data

The addition of potential information content for phylogenetic inference comes at the expense of increased data heterogeneity that can result in model misspecification, hindering accurate phylogenetic reconstruction.

"A flock of genomes"

(from Zhang et al. 2014)

A coalescent-based estimates of the avian species tree of life using a method based on the statistical binning of loci

Mirarab et al. 2014

Genomic datasets face more than just computational challenges!

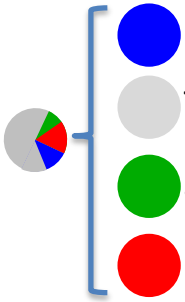
- There is an inherent increase in data heterogeneity as shift to transcriptomes/genomes and more taxa

PROBLEM?

Discord not due to just ILS

Multiple processes contribute to gene tree discord

- highly elevated levels of strongly supported conflict



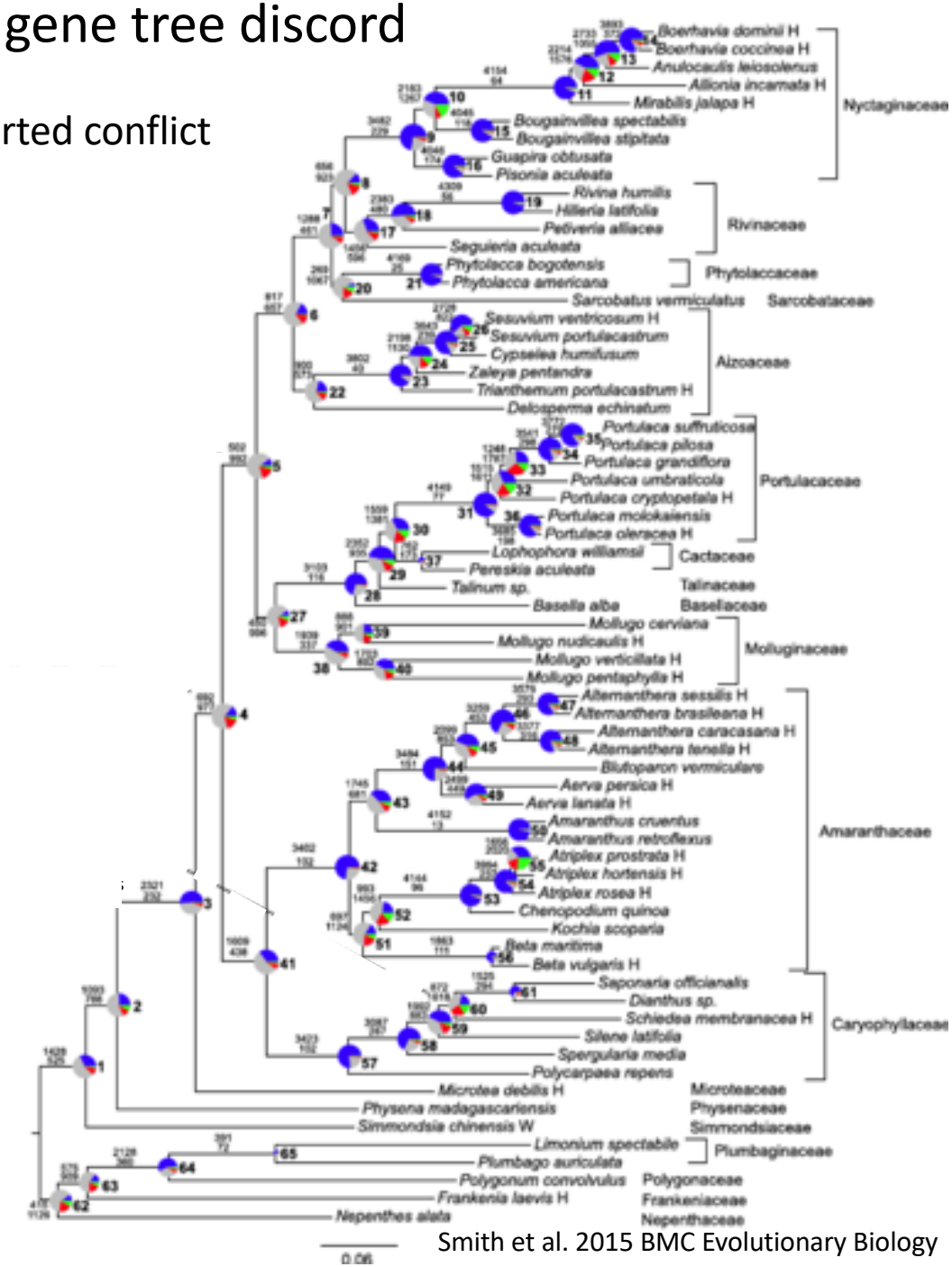
Concordant with species tree

Topologies without strong support

Support main alternative topology

Remaining discordant topologies

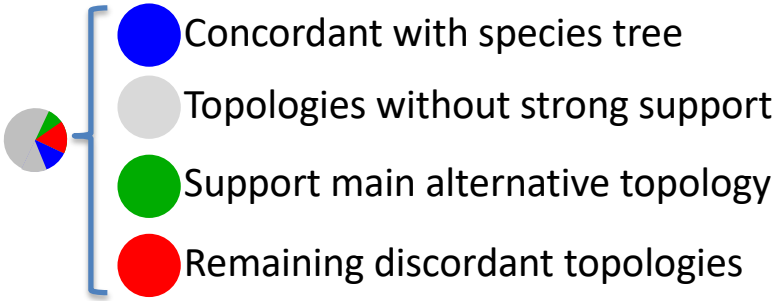
Caryophyllales



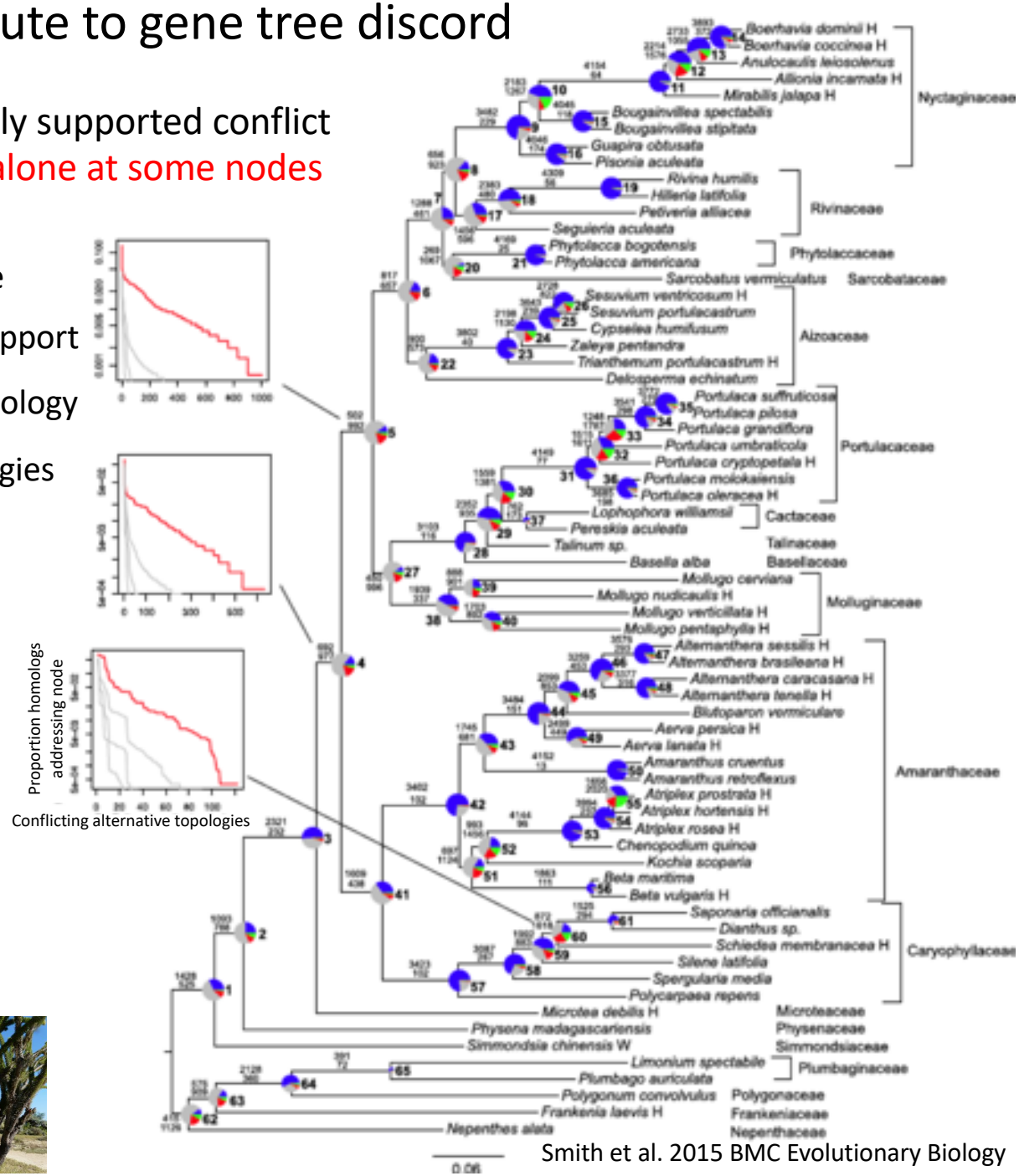
Smith et al. 2015 BMC Evolutionary Biology

Multiple processes contribute to gene tree discord

- highly elevated levels of strongly supported conflict that cannot be explained by ILS alone at some nodes

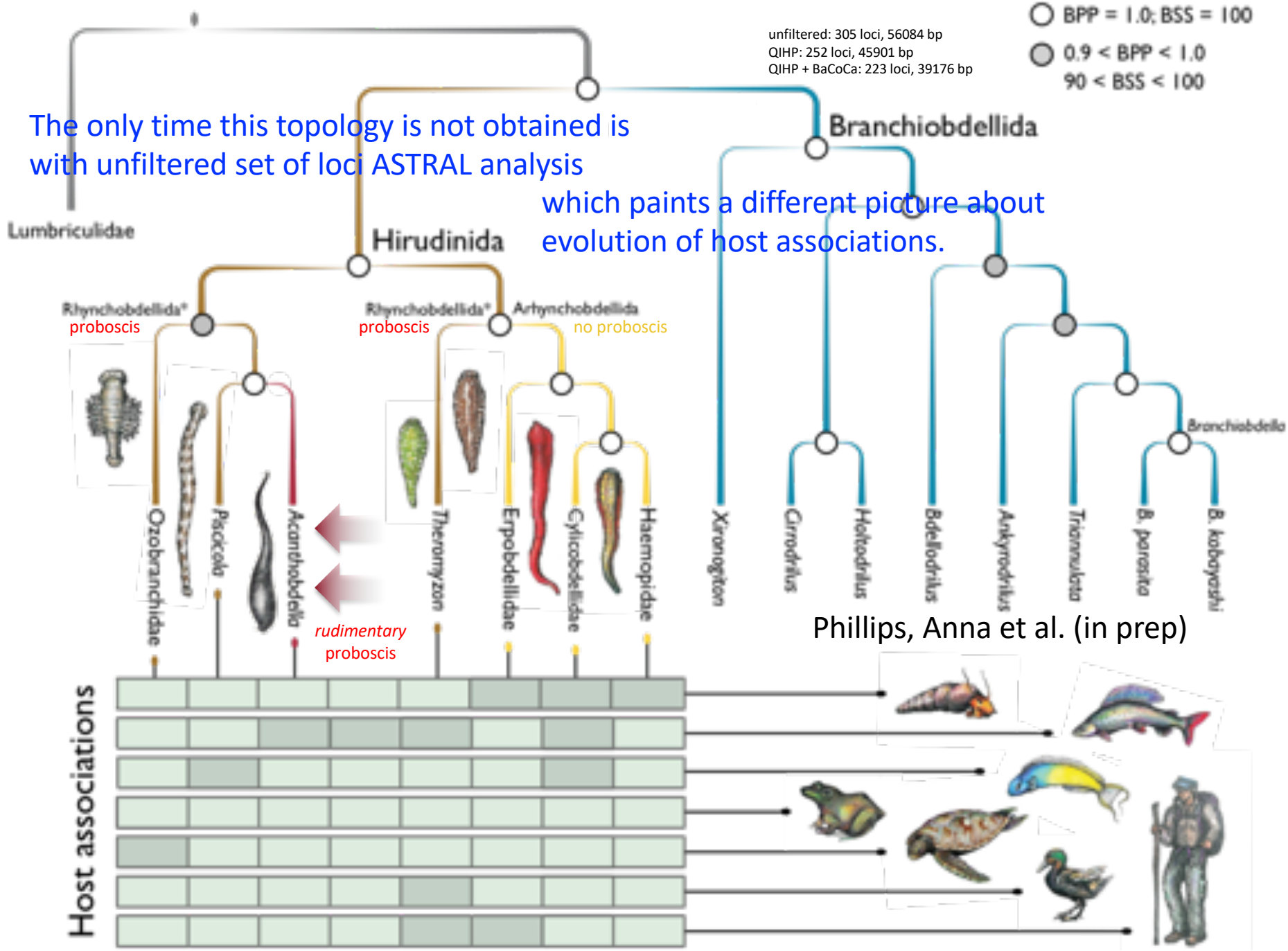


Caryophyllales

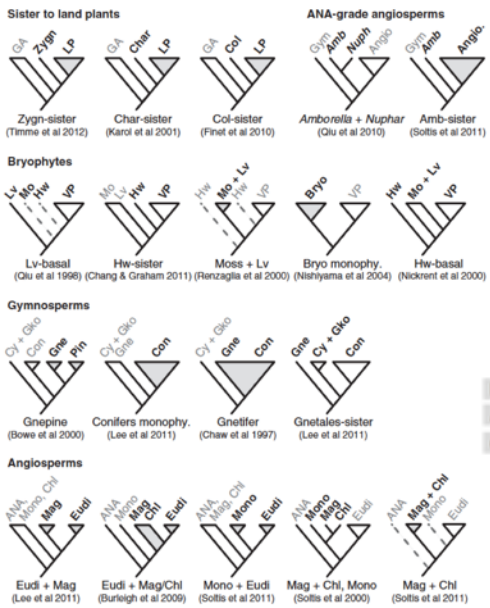


Smith et al. 2015 BMC Evolutionary Biology

Different phylogenetic estimates with inclusion/exclusion of loci



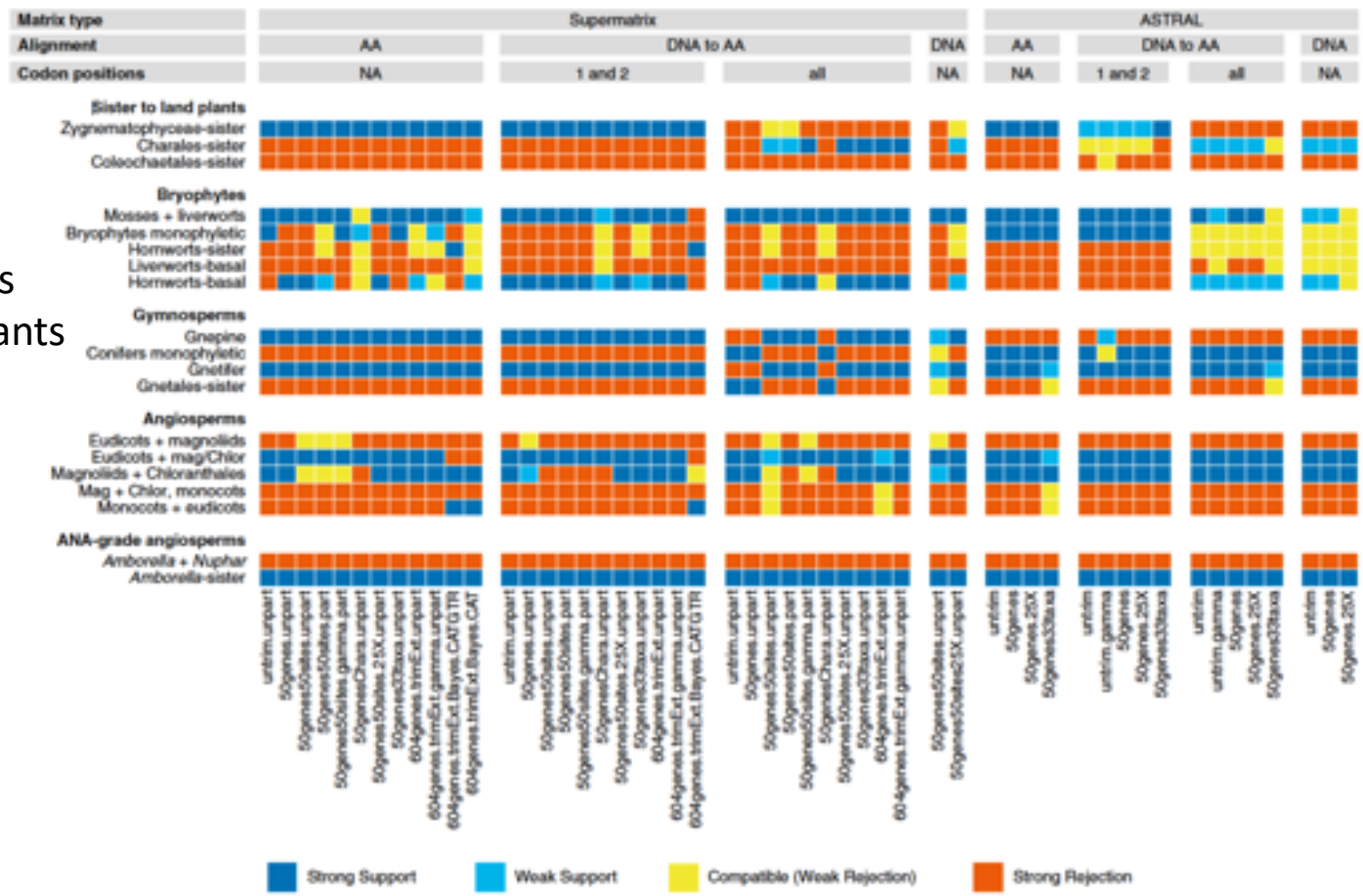
Systematic errors in phylogenetic inference caused by model misspecification



hypothesized relationships among major clades of plants

Wickett et al. 2014 PNAS

69 analyses of 92 taxa



What is the empiricists to do to improve phylogenetic accuracy?

- Gene tree discord (per se) is not problematic
- Check alignment, paralogs, etc are not contributing to discord

Data problem versus model problem?

- Filter data (criteria?)?
- Subsets of data?
- More data?
- Heterogeneity of processes underlying discord across loci?

Species tree inference:

a guide to the theoretical and empirical challenges of today and tomorrow

L. S. Kubatko and L. L. Knowles eds.

EVOLUTION AND GENOMICS

Intensive and immersive training opportunities

SPECIES-TREE INFERENCE

Species-tree inference with BEAST 2 and SNAPP

Michael Matschiner, 28 January 2020

Practical today 2-5pm

Transformative potential of model-based analyses:

- Codon substitution and analysis of natural selection
- Adaptive molecular evolution
- Divergence time estimation and biogeographic analysis
- Phylogenetic inference
- Species delimitation
- Demographic inference

....models are how we communicate
our knowledge to a statistical apparatus

Transformative potential of model-based analyses:

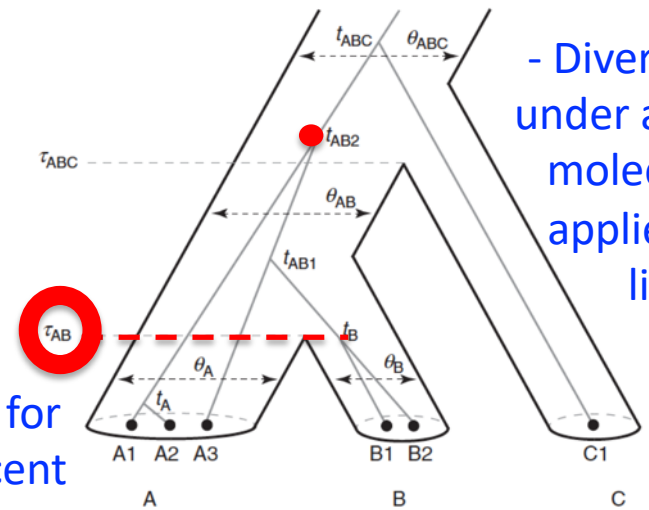
- Codon substitution and analysis of natural selection
- Adaptive molecular evolution
- Divergence time estimation and biogeographic analysis
- Phylogenetic inference
- Species delimitation
- Demographic inference

• All models are flawed..., some are more or less useful

....models are how we communicate
our knowledge to a statistical apparatus

Transformative potential of model-based analyses:

- Codon substitution and analysis of natural selection
- Adaptive molecular evolution
- Divergence time estimation and biogeographic analysis
- Phylogenetic inference
- Species delimitation
- Demographic inference (e.g., time of divergence)



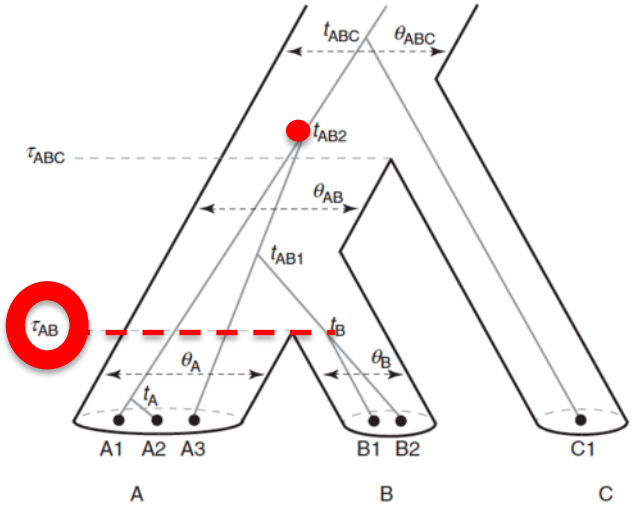
- Divergence time under a model of a molecular clock applied to gene lineage

- Divergence time under model to account for gene lineage sorting process (i.e., a coalescent model) to account for gene divergence that predates population divergence to obtain accurate diverge time estimate

- All models are flawed..., some are more or less useful

Transformative potential of model-based analyses:

- Codon substitution and analysis of natural selection
- Adaptive molecular evolution
- Divergence time estimation and biogeographic analysis
- Phylogenetic inference
- Species delimitation
- Demographic inference
  (e.g., time of divergence)



- All models are flawed..., some are more or less useful

....depending upon how effectively they represent evolutionary processes

Genetic model-based species delimitation

- History of inference about species boundaries using genetic data

- Conceptual issues surrounding species delimitation

- Future of delimitation models

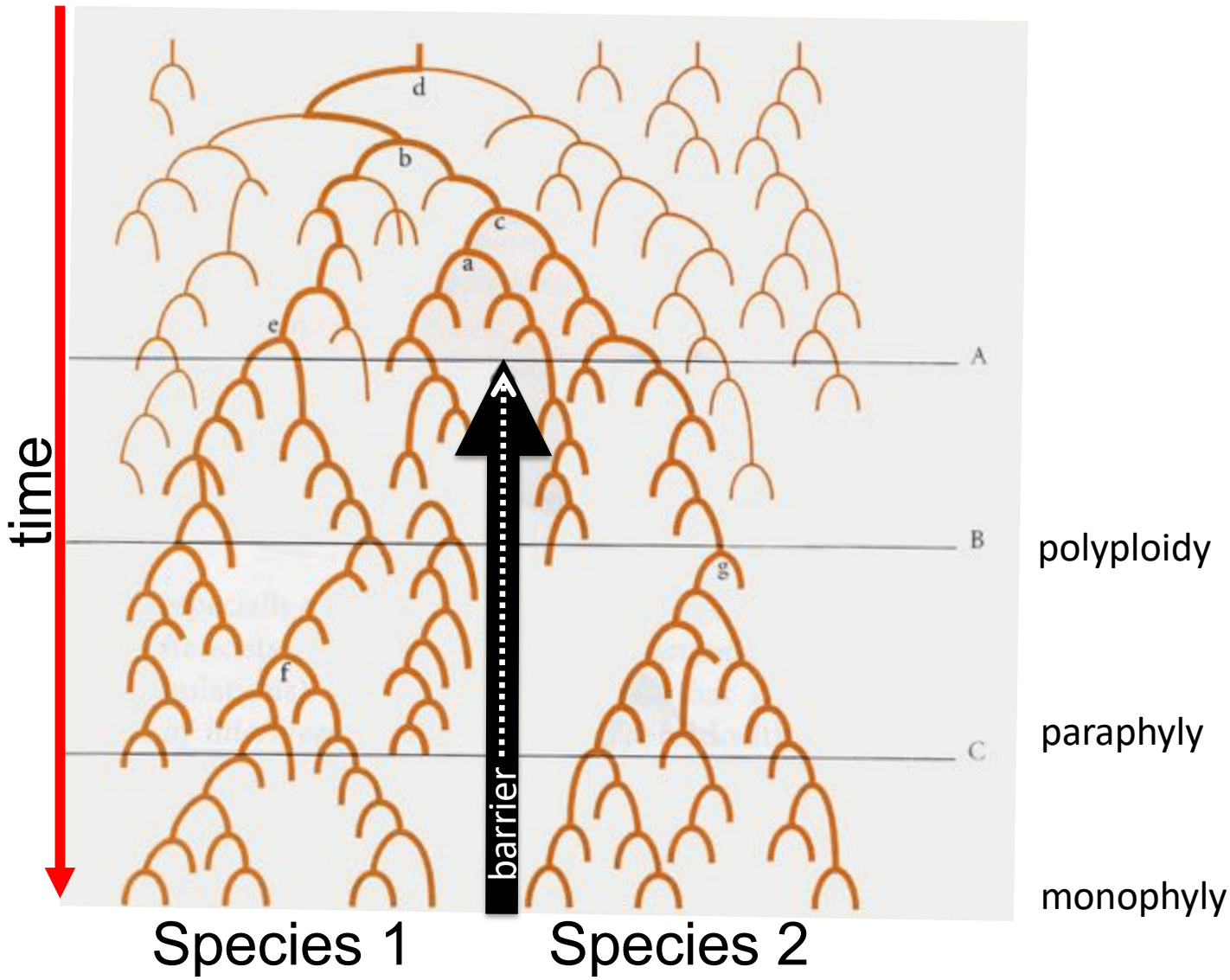
- Practical training (tonight 7-10pm)

  Software: *Delineate*

  Software: *Decrypt*

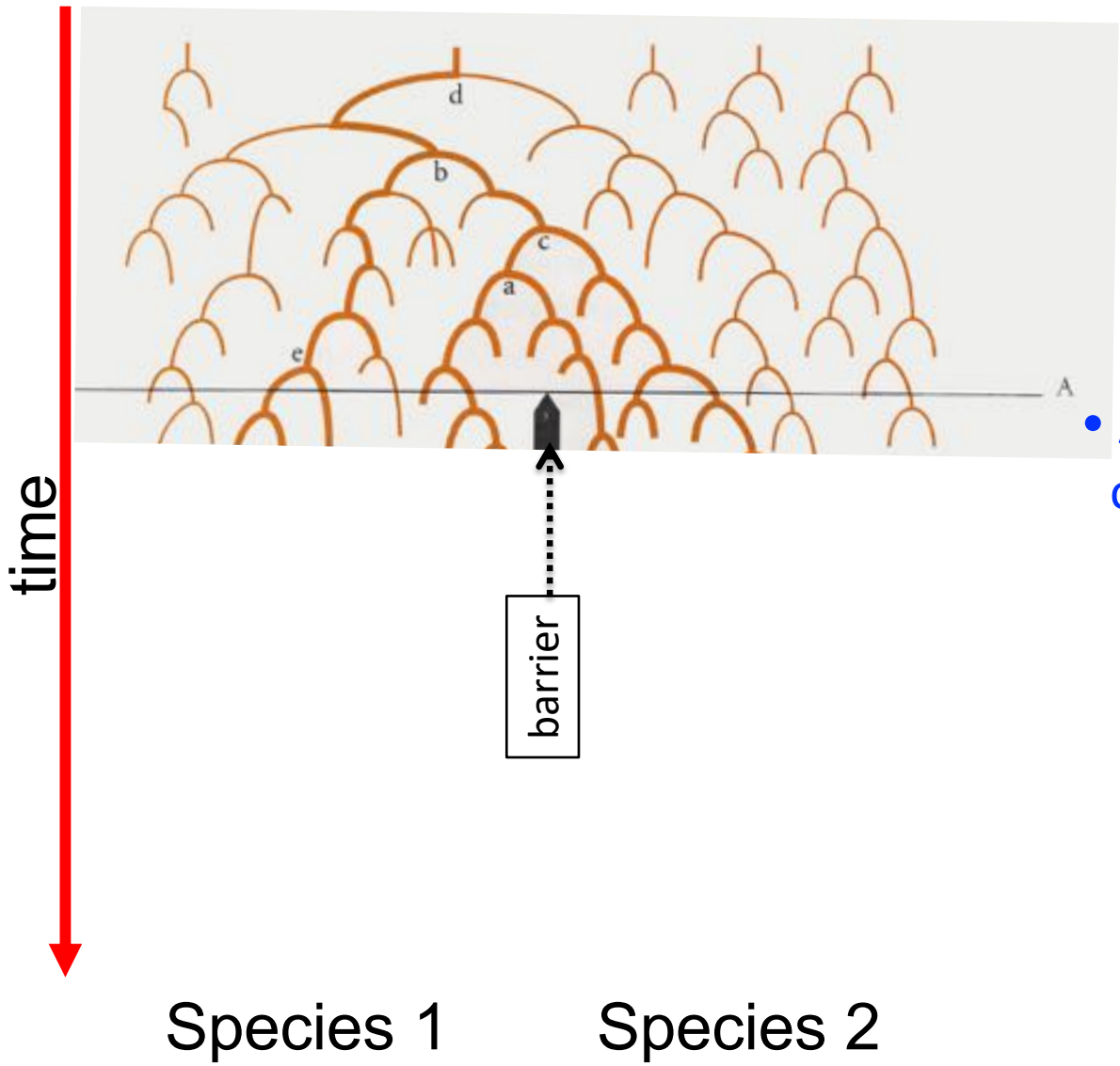
Isolation is the property that allows species to be recognized genetically

- Transition towards species monophyly with time



Avise et al.(1987)

- Transition towards species monophyly with time



- Ancestry immediately after origin of a barrier is mixed

polyphyly

Avise et al.(1987)

- Transition towards species monophyly with time

(thin lines denote lineages with no descendants)

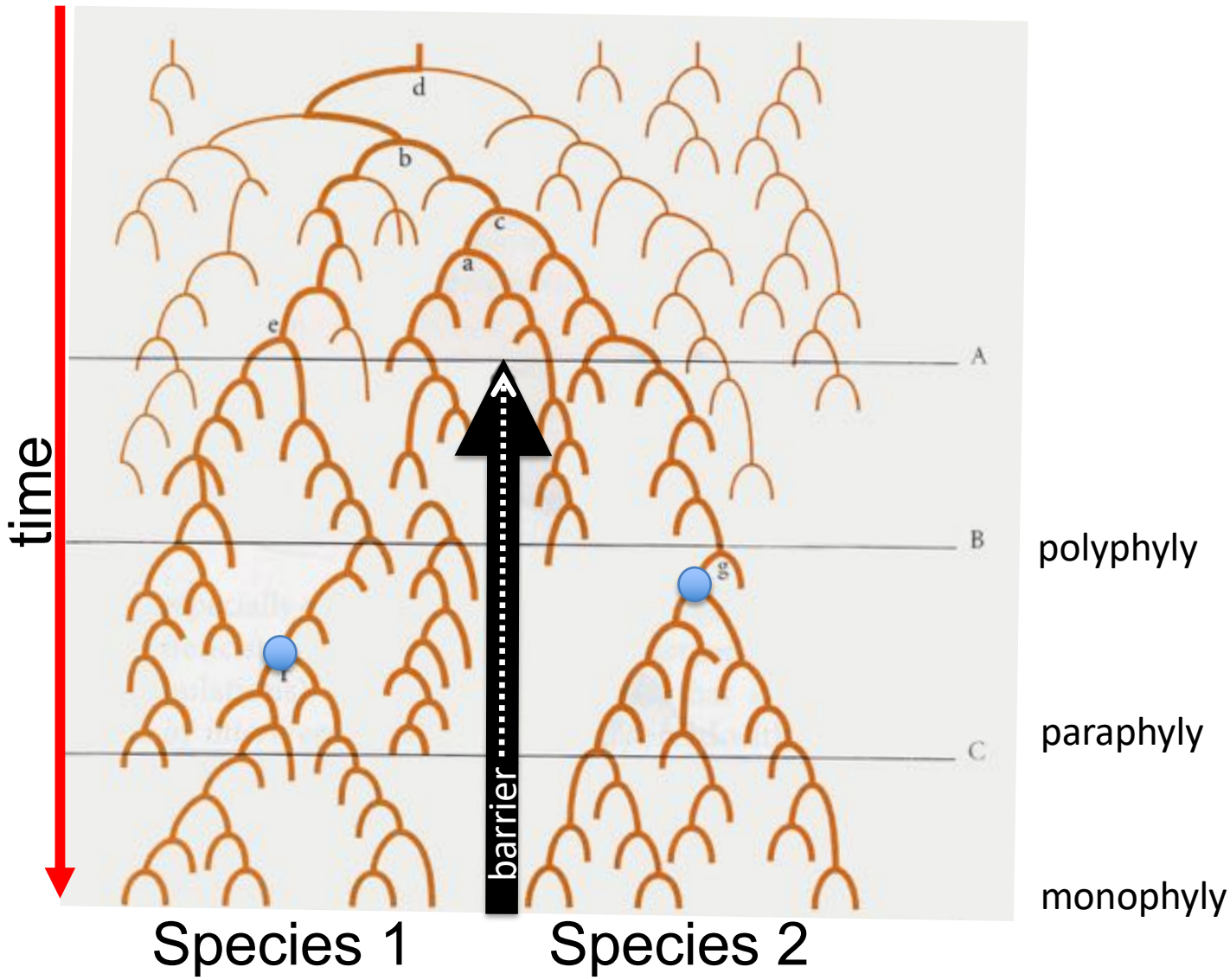
polyphyly

- With time there is a loss of shared ancestral lineages

paraphyly

Avise et al.(1987)

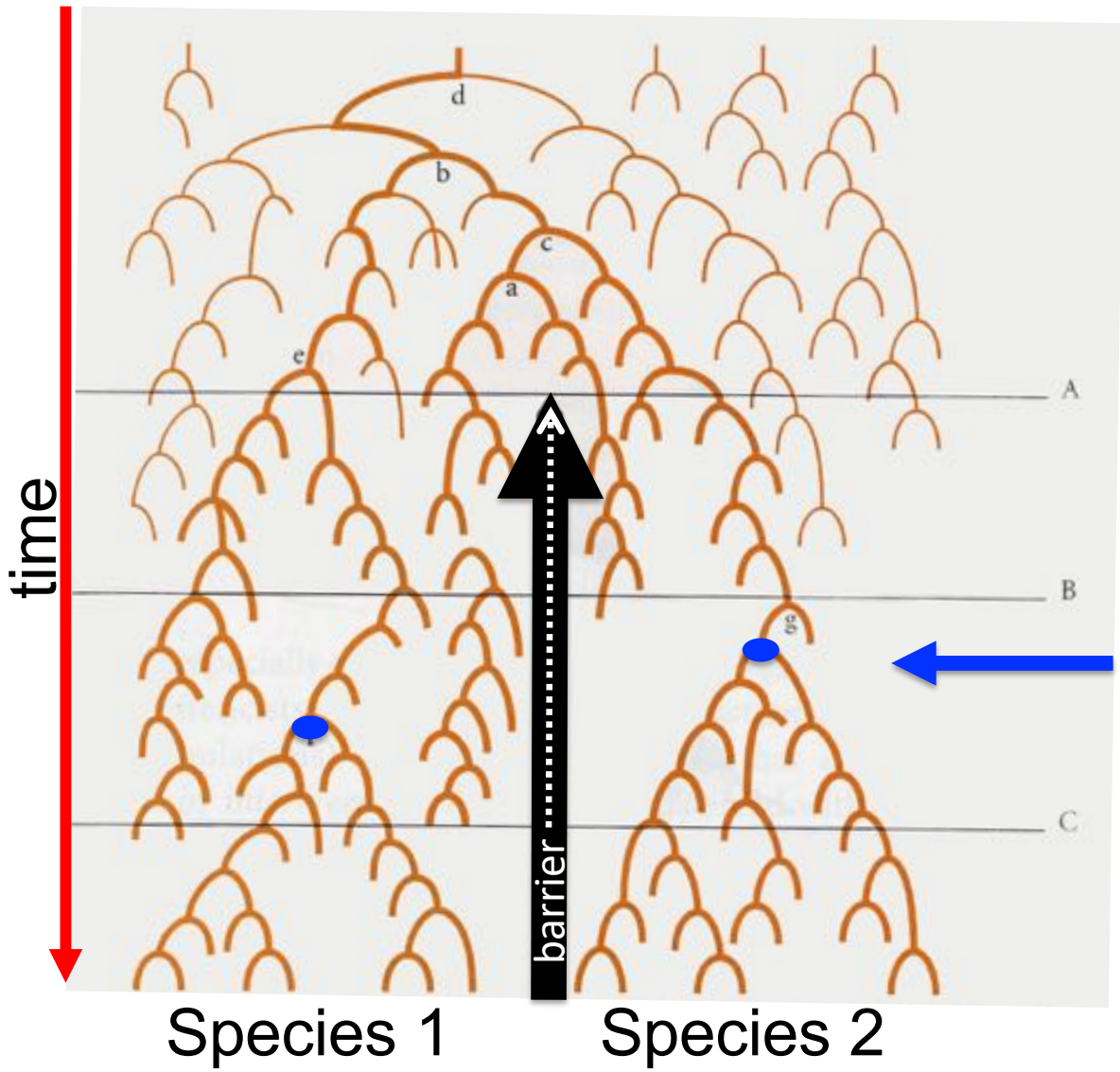
- Transition towards species monophyly with time



Avise et al.(1987)

Isolation is the property that allows species to be recognized genetically

- Transition towards species monophyly with time



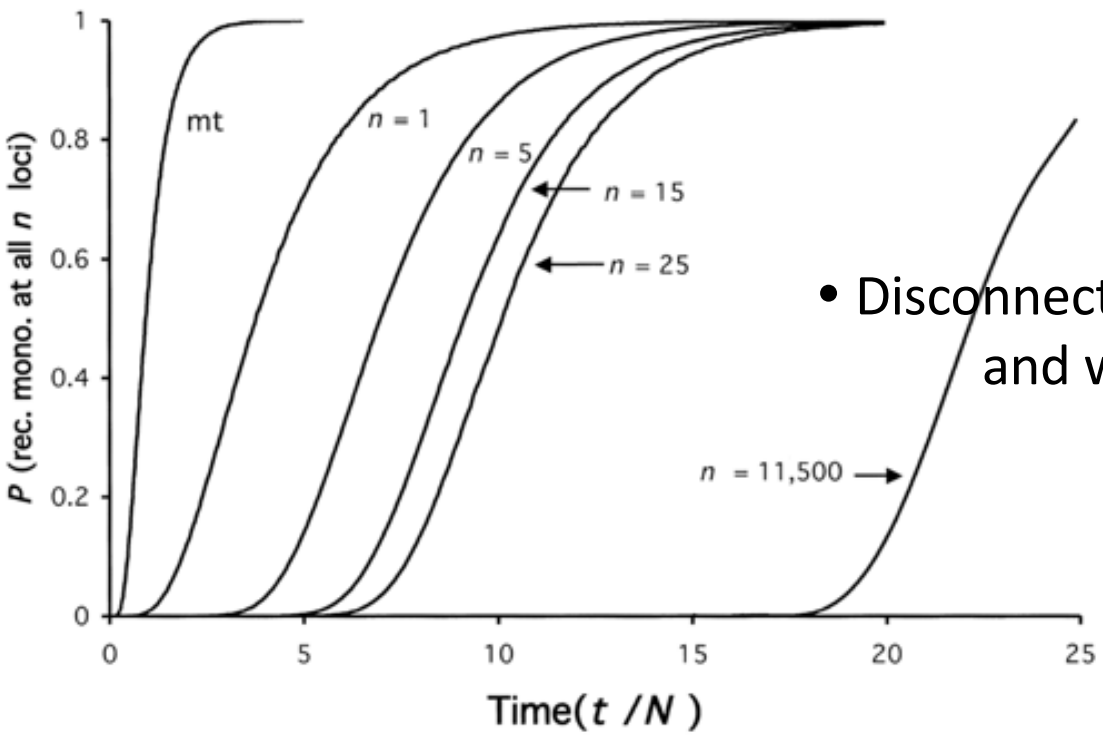
There were multiple gene lineages in past, they just did not leave any descendants (e.g., mitochondrial eve)

Avise et al.(1987)

Isolation is the property that allows species to be recognized genetically

- Exclusivity criteria (e.g., monophyly)

  "A group of organisms is exclusive if their loci coalesce more recently within the group than between any member of the group and any organisms outside the group''
  (Baum & Shaw 1995, p. 296).



- Disconnect between the time of speciation and when taxa become monophyletic

FIG. 1. Probabilities of observing monophyly with time for populations that are genetically isolated. Curves are shown for a single mitochondrial locus and for samples of different numbers of nuclear loci.

Hudson and Coyne (2002) Evolution 56:1557–1565

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

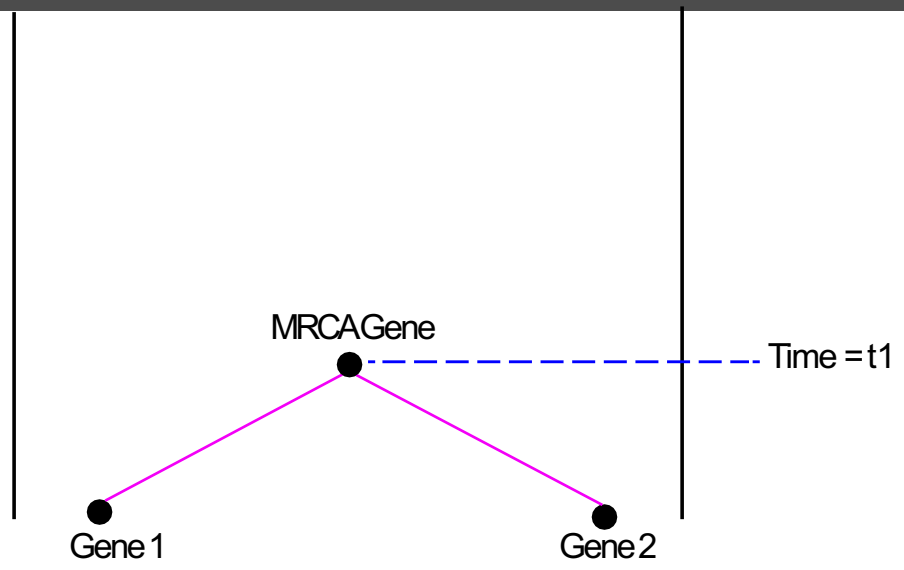
Department of Ecology and Evolutionary Biology, Museum of Zoology, 1109 Geddes Avenue, University of Michigan, Ann Arbor, MI 48109-1079, USA; E-mail: knowlesl@umich.edu (L.L.K.)

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.]

Coalescent Theory Applications in a Nutshell

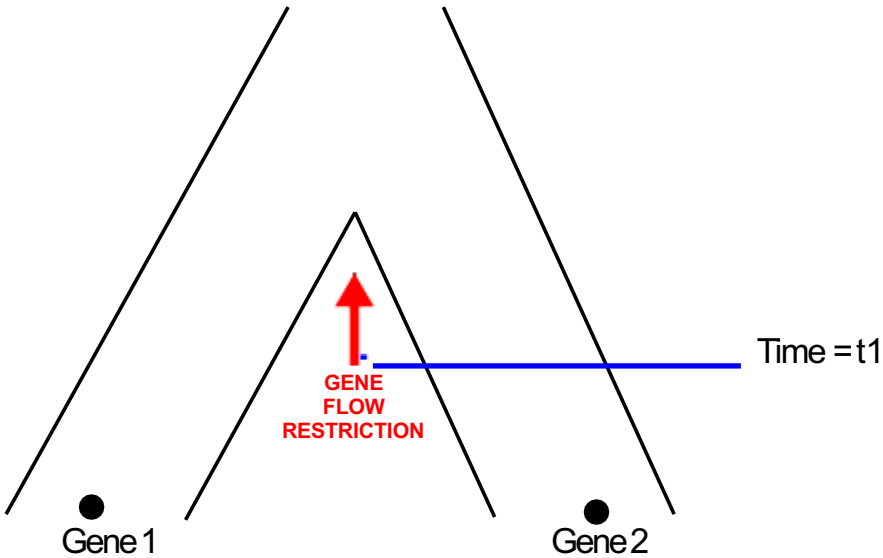
- Makes predictions about the *waiting time* between coalescence events based on population size and sample size.
- Predictions are based on assumptions of particular properties of the population that the genes (or individuals having those genes) are evolving.
- Deviances in observed waiting times from that predicted can be used to make inferences about deviances in actual population properties from assumed.

How Does Structuring Change the Coalescent Times?



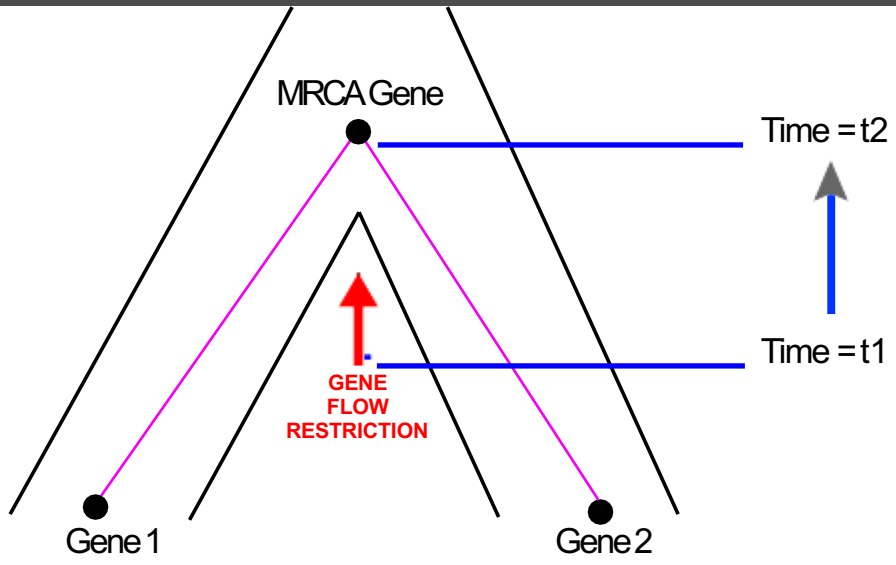
- Recall that the coalescent makes predictions about the timings to coalescence for genes sampled at random from a panmictic population.

How Does Structuring Change the Coalescent Times?



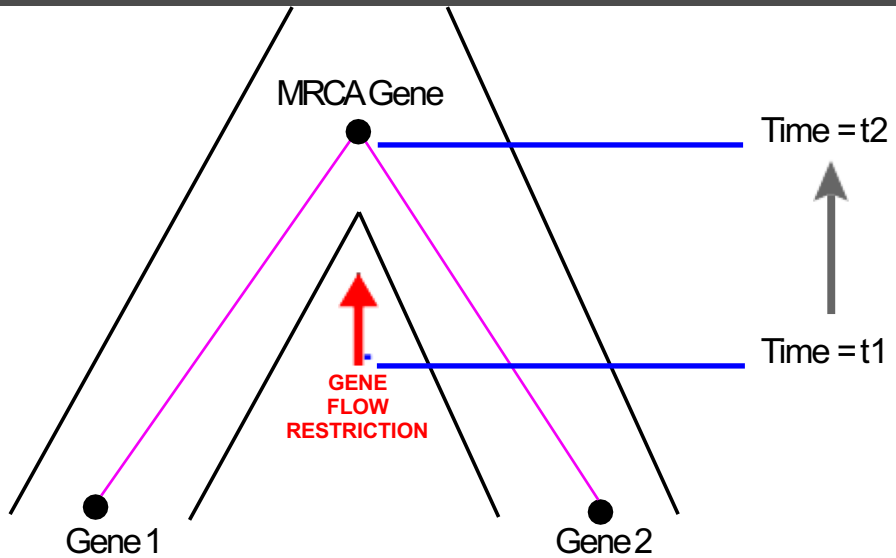
- Recall that the coalescent makes predictions about the timings to coalescence for genes sampled at random from a panmictic population.
- What happens if there are restrictions to panmixia?

How Does Structuring Change the Coalescent Times?



- Recall that the coalescent makes predictions about the timings to coalescence for genes sampled at random from a panmictic population.
- What happens if there are restrictions to panmixia?
- Then the timings to coalescence get *extended*

How Does Structuring Change the Coalescent Times?



- Recall that the coalescent makes predictions about the timings to coalescence for genes sampled at random from a panmictic population.
- What happens if there are restrictions to panmixia?
- Then the timings to coalescent get *extended*

- This is the basis of the censored coalescent

(aka: multispecies coalescent, MSC)

Syst. Biol. 56(6):887–895, 2007

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

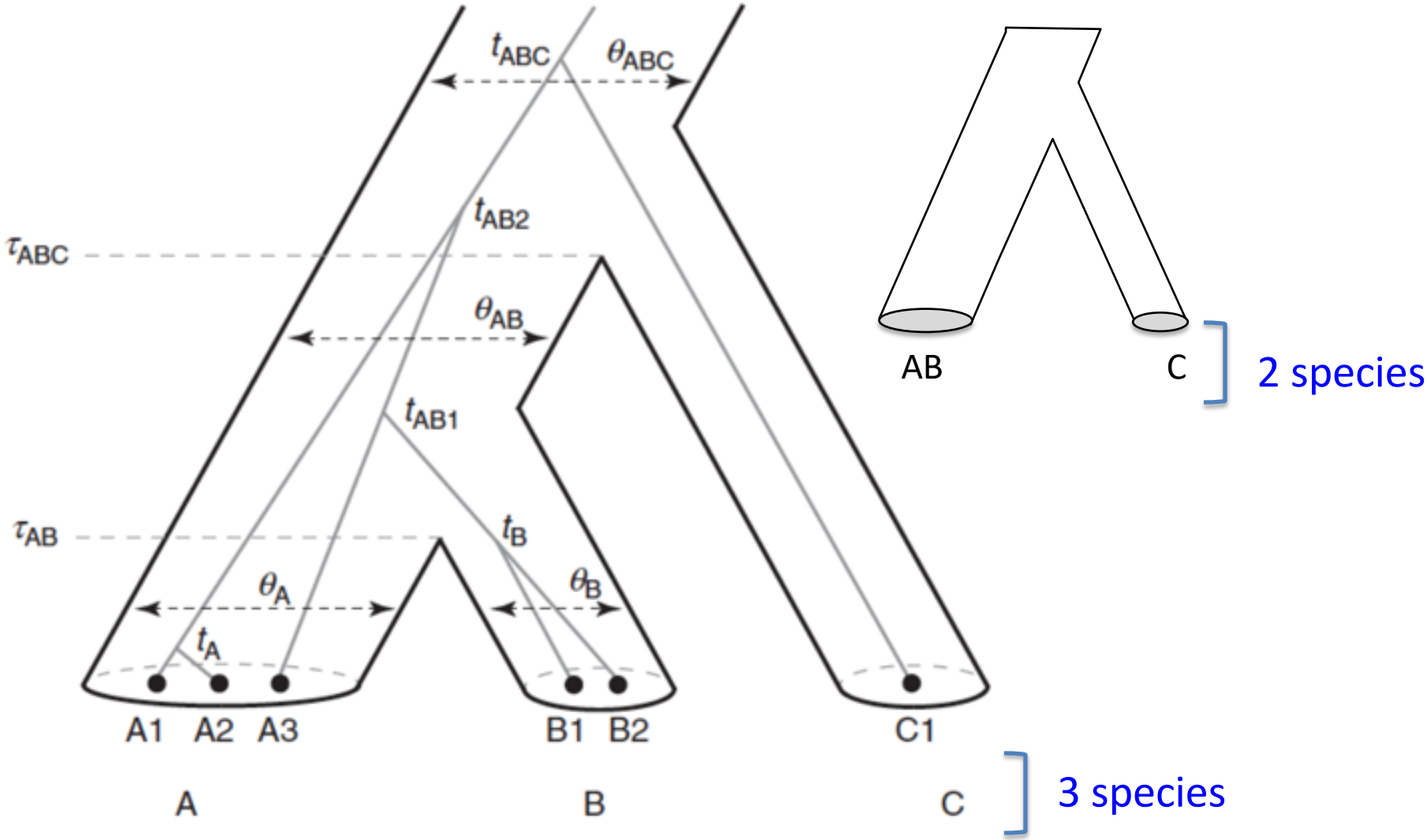
Delimiting Species without Monophyletic Gene Trees

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.]

Probabilistic approach to evaluate different species delimitation hypotheses under multispecies coalescent (MSC)



Different species delimitation hypotheses are formulated as competing statistical models and inferred from the genetic data through Bayesian model selection (i.e., through calculation of posterior model probabilities) in bpp program. Yang and Rannala (2010) *PNAS*

That was then and this is now…

- Proliferation of available programs
- Vast amounts of data available
- Empiricists' suspicions about delimited "species"

Explosion of applications of the MSC for species delimitation

Bayesian species delimitation usin
sequ

Ziheng Y

*Center for
University C
Davis, CA 95

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

Received: 15 September 2017 | Revised: 30 March 2018 | Accepted: 3 April 2018
DOI: 10.1111/1755-0998.12887
WILEY MOLECULAR ECOLOGY RESOURCES

sed machine learning method for
opulation genetic data

1 | Bin Lu3 | Yufeng Wu1

SPECI
COAL
ACCU

EMPIRICAL EXAMPLE WITH LIZARDS OF THE *LIOLAEMUS DARWINII* COMPLEX (SQUAMATA: LIOLAEMIDAE)

Arley Camargo,[1,2] Mariana Morando,[3] Luciano J. Avila,[3] and Jack W. Sites, Jr.[1]
[1]Department of Biology & Monte L. Bean Museum, Brigham Young University, Provo, Utah 84602
[2]E-mail: arley.camargo@gmail.com
[3]CONICET-CENPAT, Boulevard Almirante Brown 2915, U9120ACD, Puerto Madryn, Chubut, Argentina

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

Syst. Biol. 0(0):1–13, 2018

DOI:10.1093/sysbio/syy011

998
770
Advance Access Publication Date: 23 November 2014
Original Paper

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

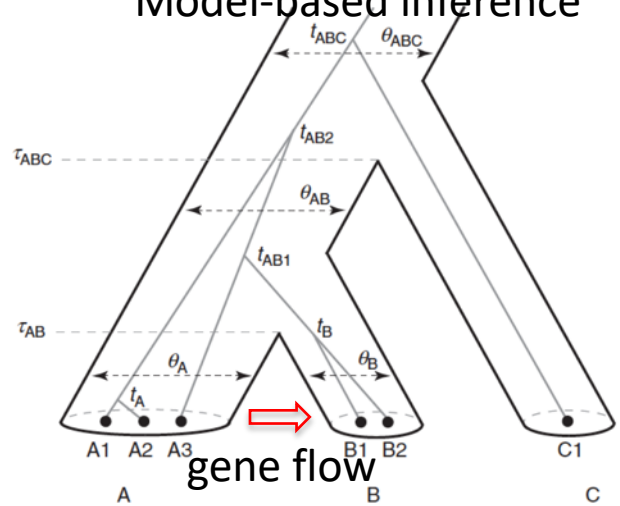
ARONG LUO[1,2,*], CHENG LING[3], SIMON Y. W. HO[2], AND CHAO-DONG ZHU[1,4]
[1]Key Laboratory of Zoological Systematics and Evolution, Institute of Zoology, Chinese Academy of Sciences, Beijing 100101, China; [2]School of Life and Environmental Sciences, University of Sydney, Sydney, New South Wales 2006, 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.
E-mail: jacksonN@njhealth.org.

ment-free Bayesian
for species delimitation
ecies coalescent

lin[1,2] and Bengt Oxelman[1,*]
al Sciences, University of Gothenburg, Box 461, SE 405 30 Göteborg,
of Sciences, University of Dicle, 21280 Diyarbakir, Turkey

Pros of species delimitation under MSC

Model-based inference



- Can delimit species before monophyly

  Knowles & Carstens (2007) *Syst. Biol.*

- Still detects lineages under low gene flow

  Zhang et al. (2011) *Syst. Biol.*

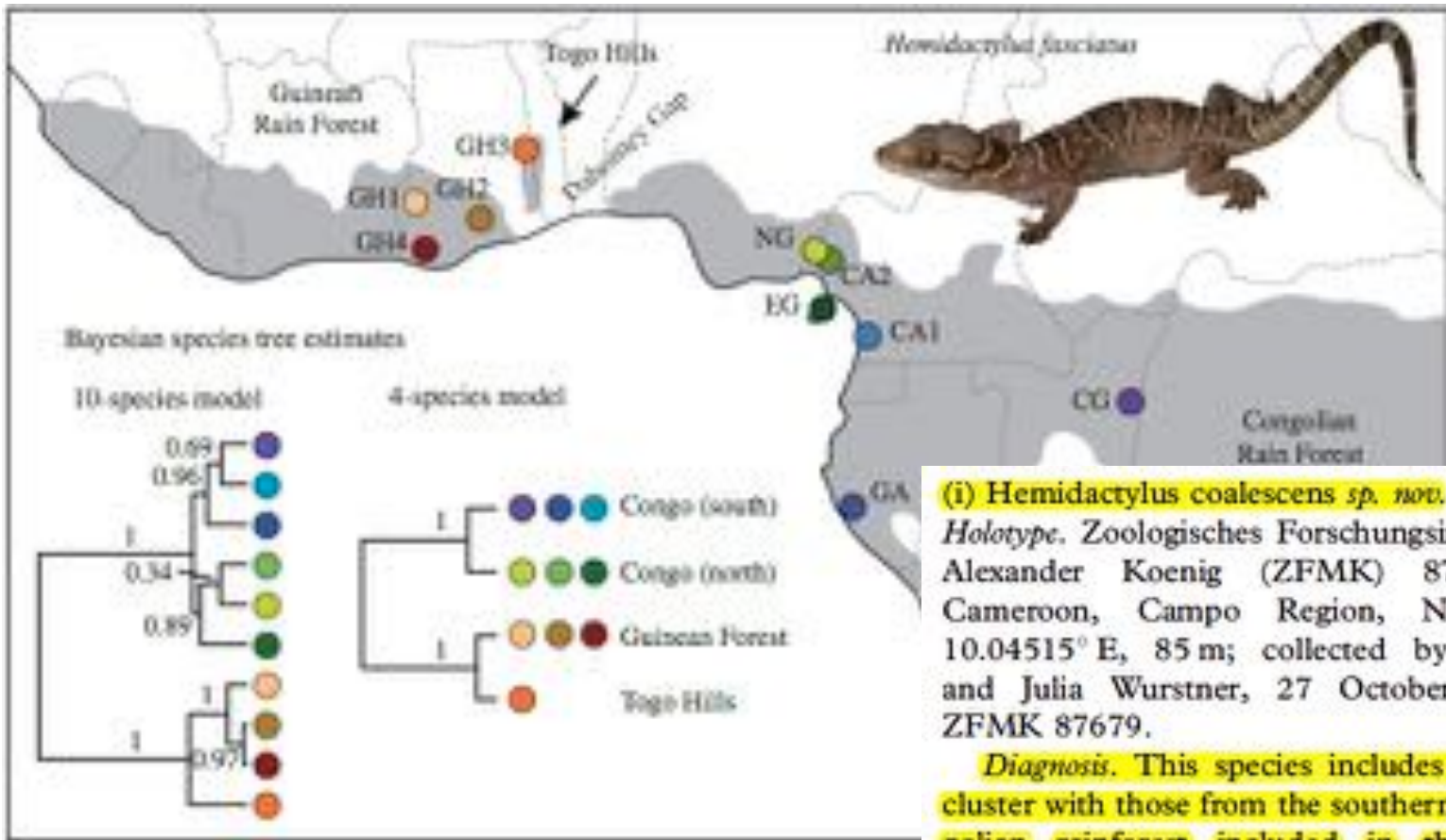
- Accuracy of species delimitation to sampling can be evaluated (i.e., will more data change status)

- De facto standardization for objectively delimiting taxa (i.e., data treated equally among all living things and avoid subjectivness of what characters to measure) Fujita et al. (2012) *TREE*

- Can take into account uncertainty in gene trees

  Yang & Rannala 2010

Model-based inference: probability of different hypotheses about species boundaries based on genetic data alone!



Leache & Fujita (2010) *Proc. R. Soc. B.*

(i) Hemidactylus coalescens *sp. nov.*

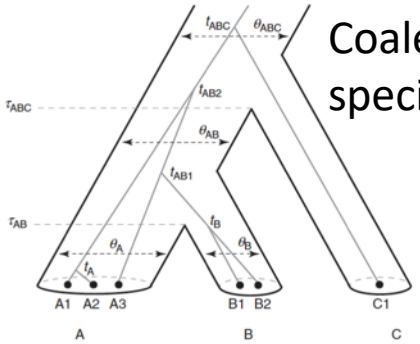
Holotype. Zoologisches Forschungsinstitut und Museum Alexander Koenig (ZFMK) 87680, adult male; Cameroon, Campo Region, Nkoelon, 2.3972° N, 10.04515° E, 85 m; collected by Michael F. Barej and Julia Wurstner, 27 October 2007. Paratype = ZFMK 87679.

Diagnosis. This species includes all populations that cluster with those from the southern portion of the Congolian rainforest included in this study (southern Cameroon, Gabon and Congo), with strong support in the Bayesian species delimitation model.

Etymology. This species is named after the coalescent process used to delimit the species.

Data-informed summary suggests problems…..

Carstens et al. 2013



Coalescent-based species delimitation

e.g.: PP = 1.0

Most newly discovered species go undescribed.

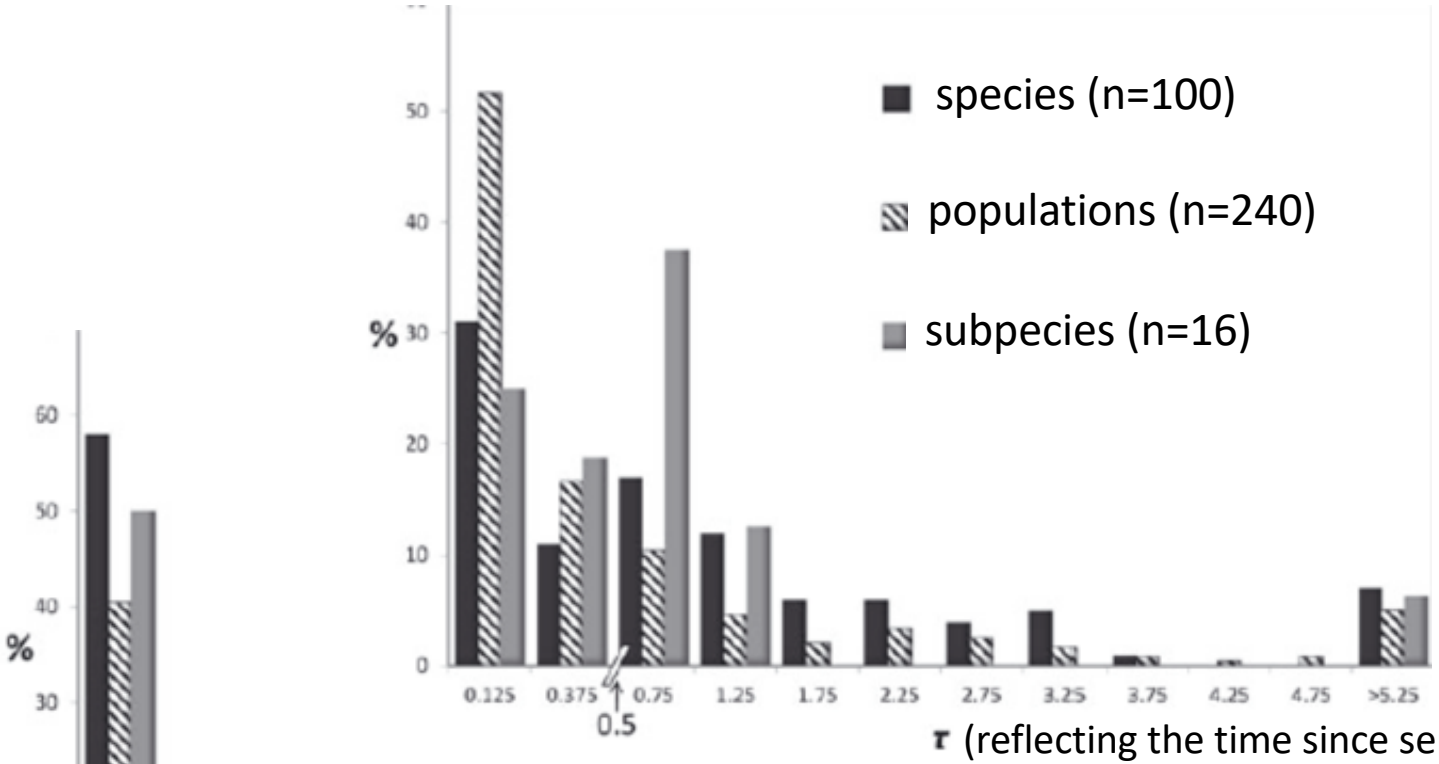
Subjective interpretation of results!

- Less than 30% of researchers applying MSC models made taxonomic recommendations!
- Less than 25% of researchers applying MSC models actually use results to describe new species!

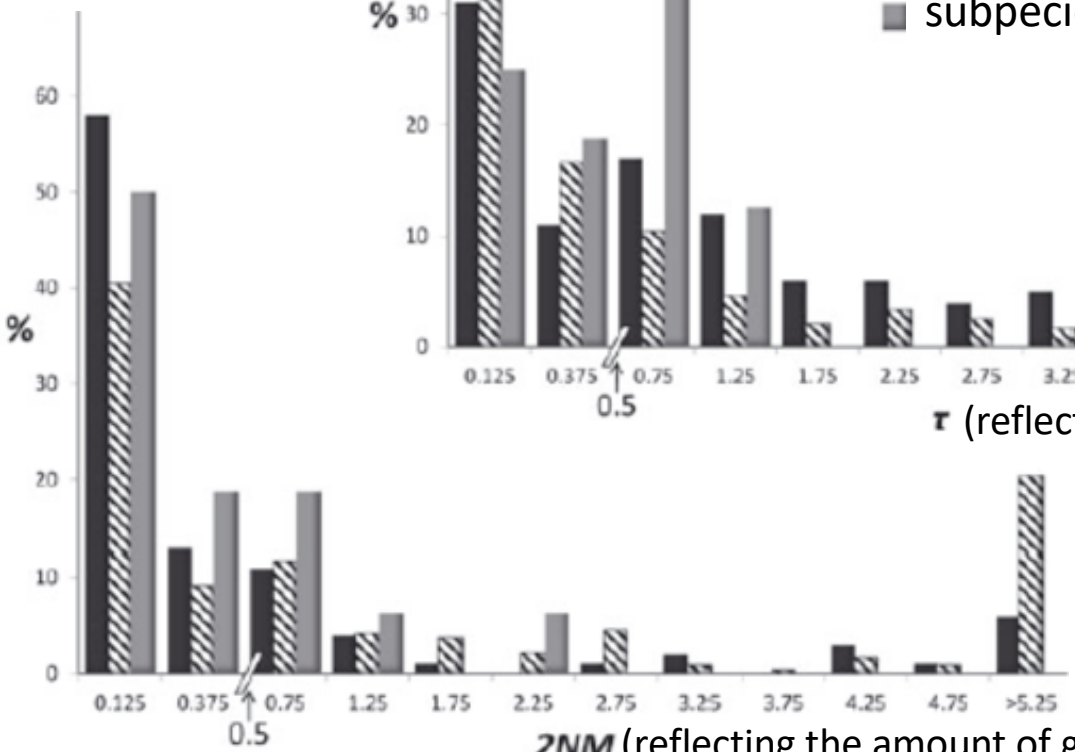
Why not name every “species”???

- Lack of differentiation in other phenotypic traits
- No characters to diagnose “species”
- Seems like a lot of new “species”
- More data and more “species”

Measures of evolutionary independence:
no distinct boundary between species and populations



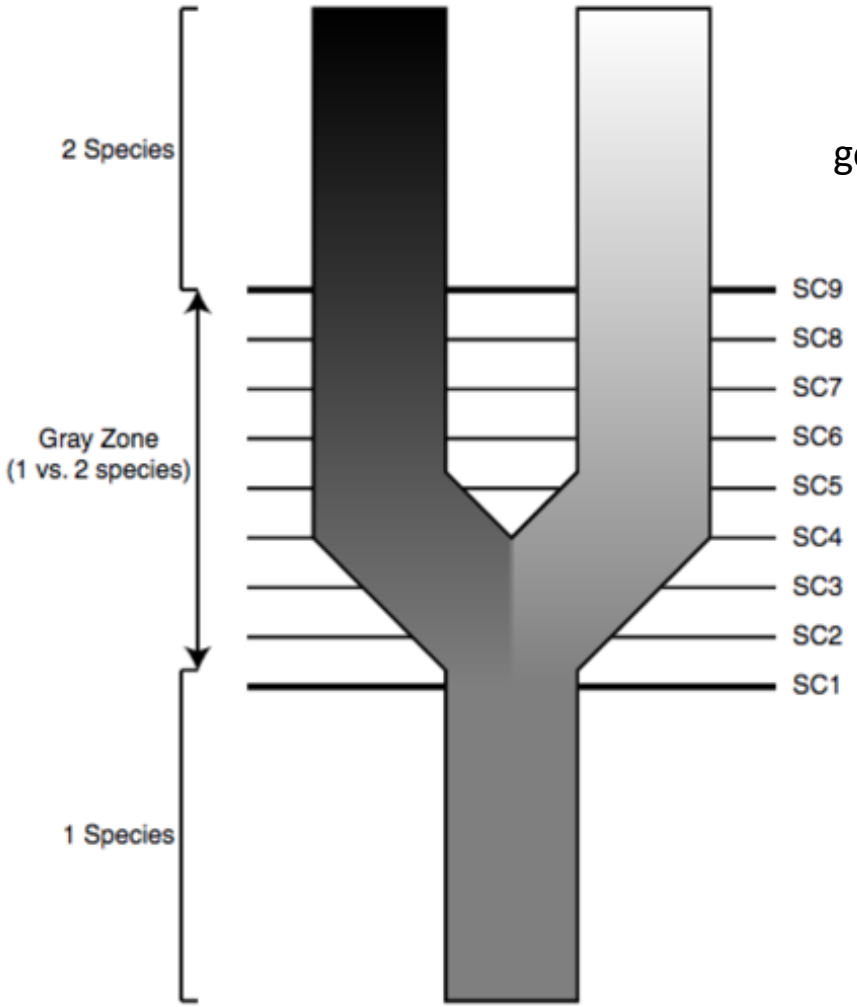
τ (reflecting the time since separation)



2NM (reflecting the amount of gene exchange)

Pinho and Hey(2010) *Evolution*

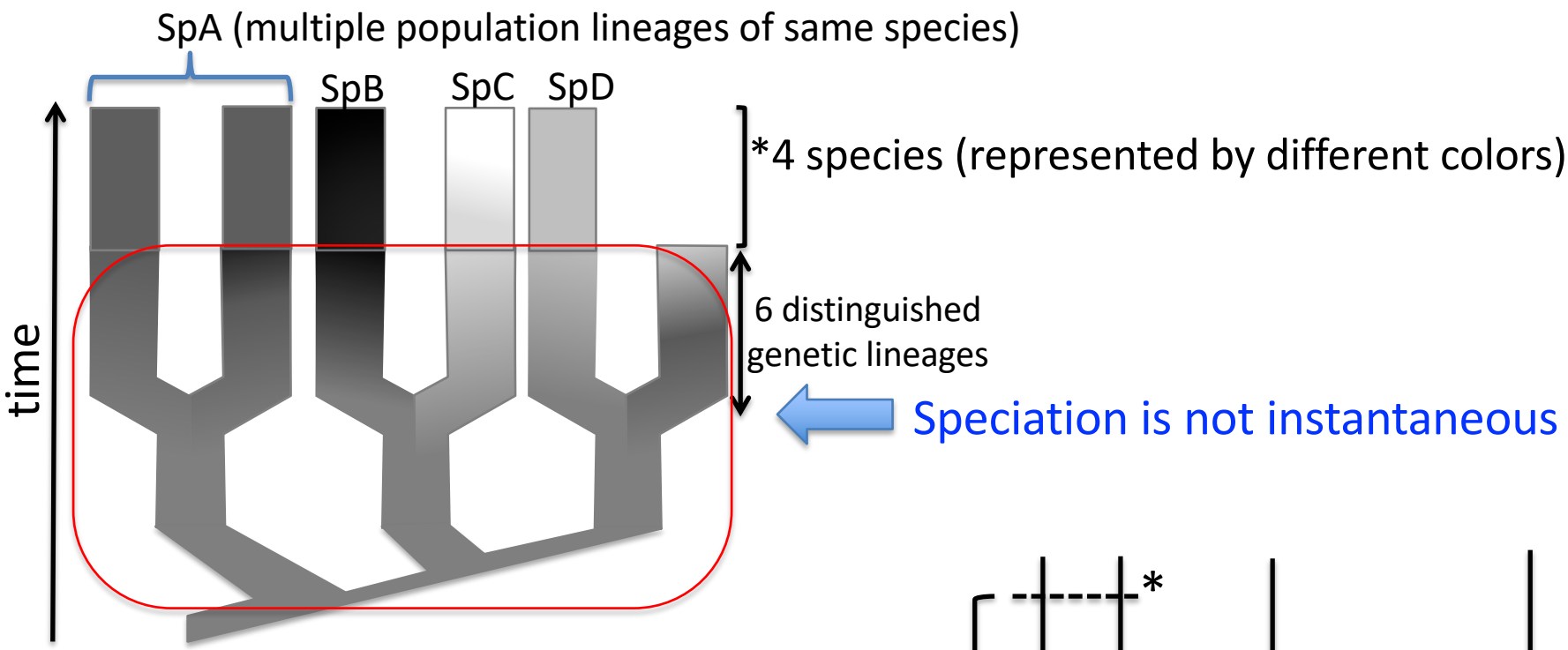
Eventually all species concepts agree…so no big deal right?!?

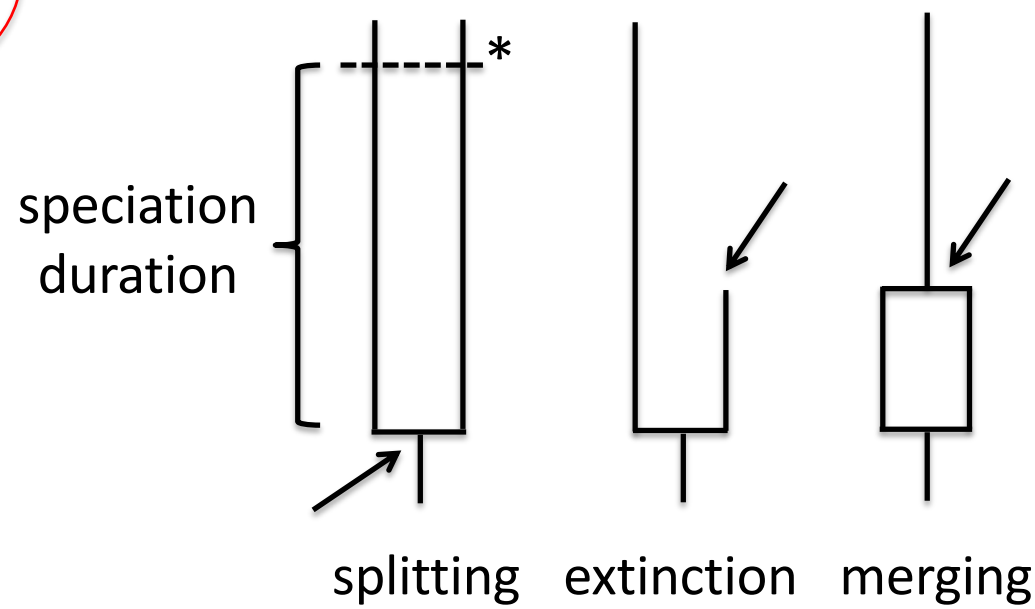


general lineage concept

de Querroz 2005, 2007

* Not all lineages become species!

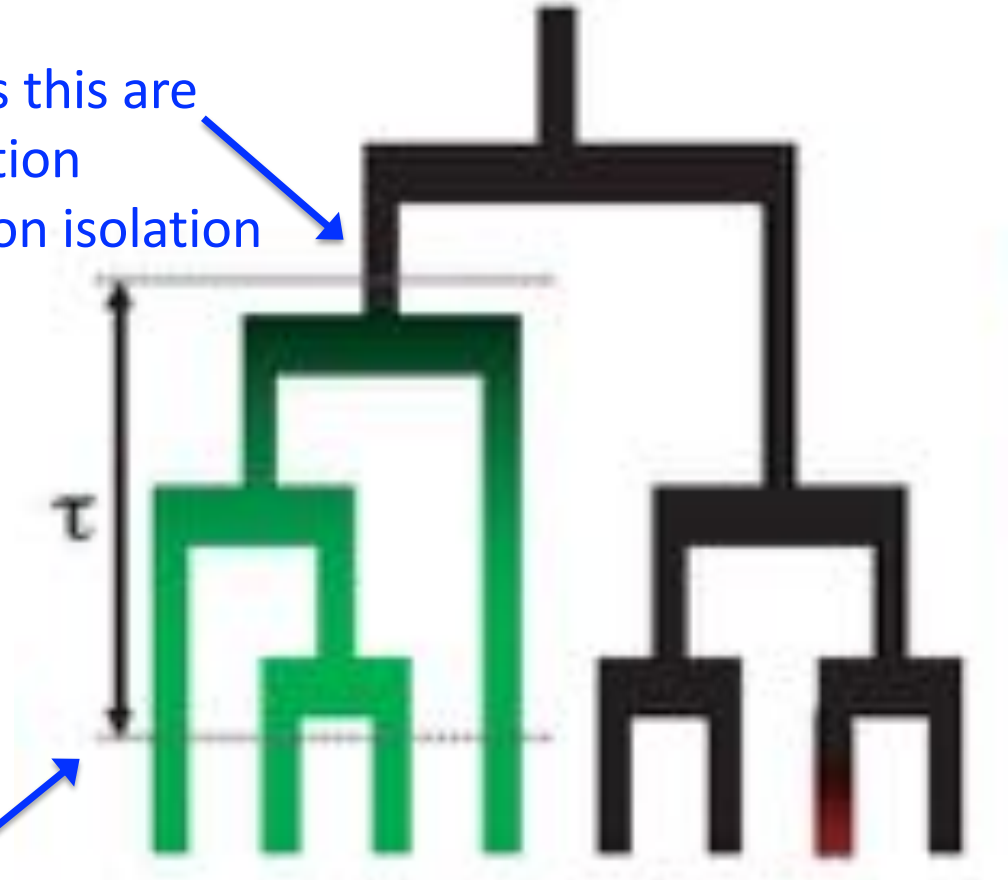




Speciation is a protracted process

Splitting events such as this are the initiation of speciation through, e.g., population isolation

duration of speciation



Color change indicates completion of speciation and development of true species from incipient species (i.e., lineage conversion)

Protracted speciation model (PSM)

Modified from Rosindell et al. (2010) Ecol. Lett. 13:716

the multispecies COALESCENT

Current model-based genetic species delimitation

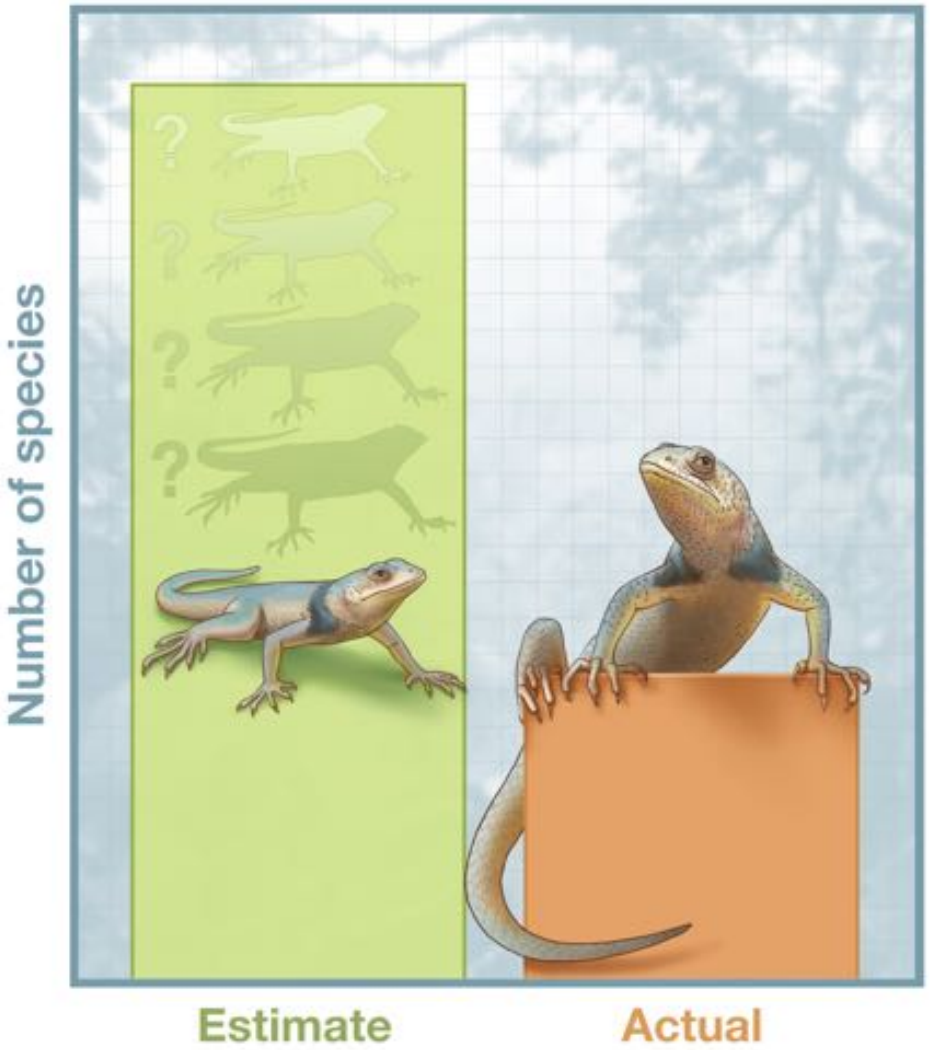


Illustration credit: John Megahan

THE COALESCENT

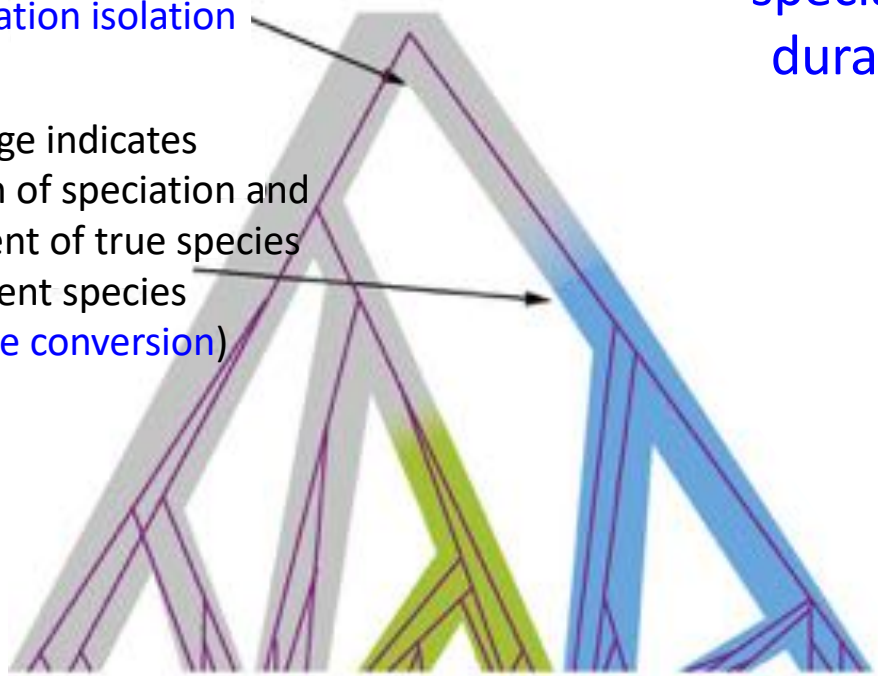
Sukumaran & Knowles (2017) *PNAS*

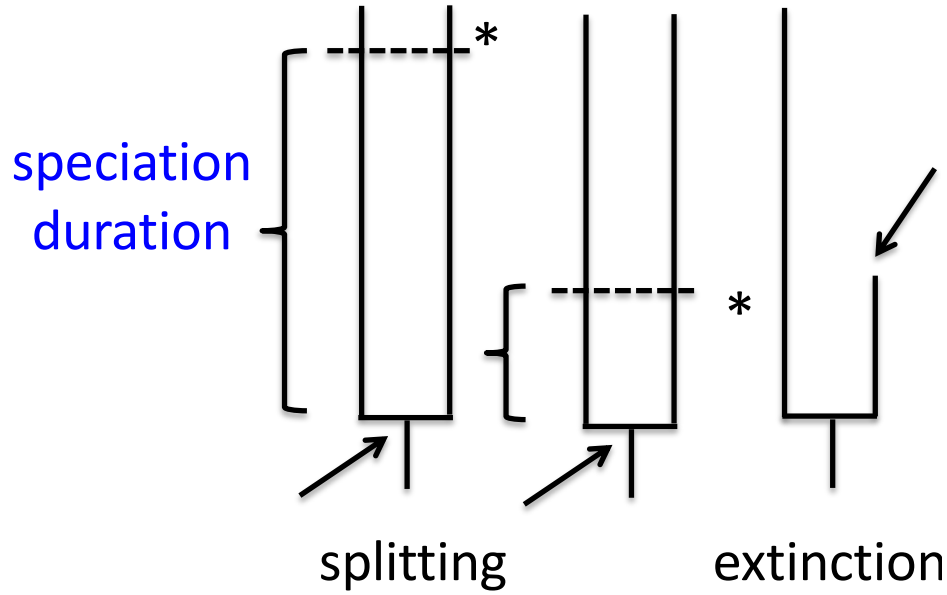
Simulate data to account for differences in speciation duration (i.e., speciation is not instantaneous)

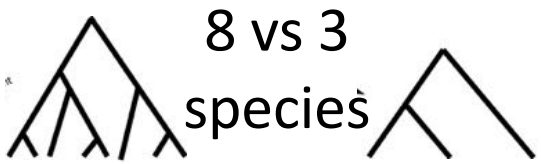
Sukumaran & Knowles (2017) *PNAS*

Splitting events such as this are initiation of speciation through, e.g., population isolation

Color change indicates completion of speciation and development of true species from incipient species (i.e., lineage conversion)





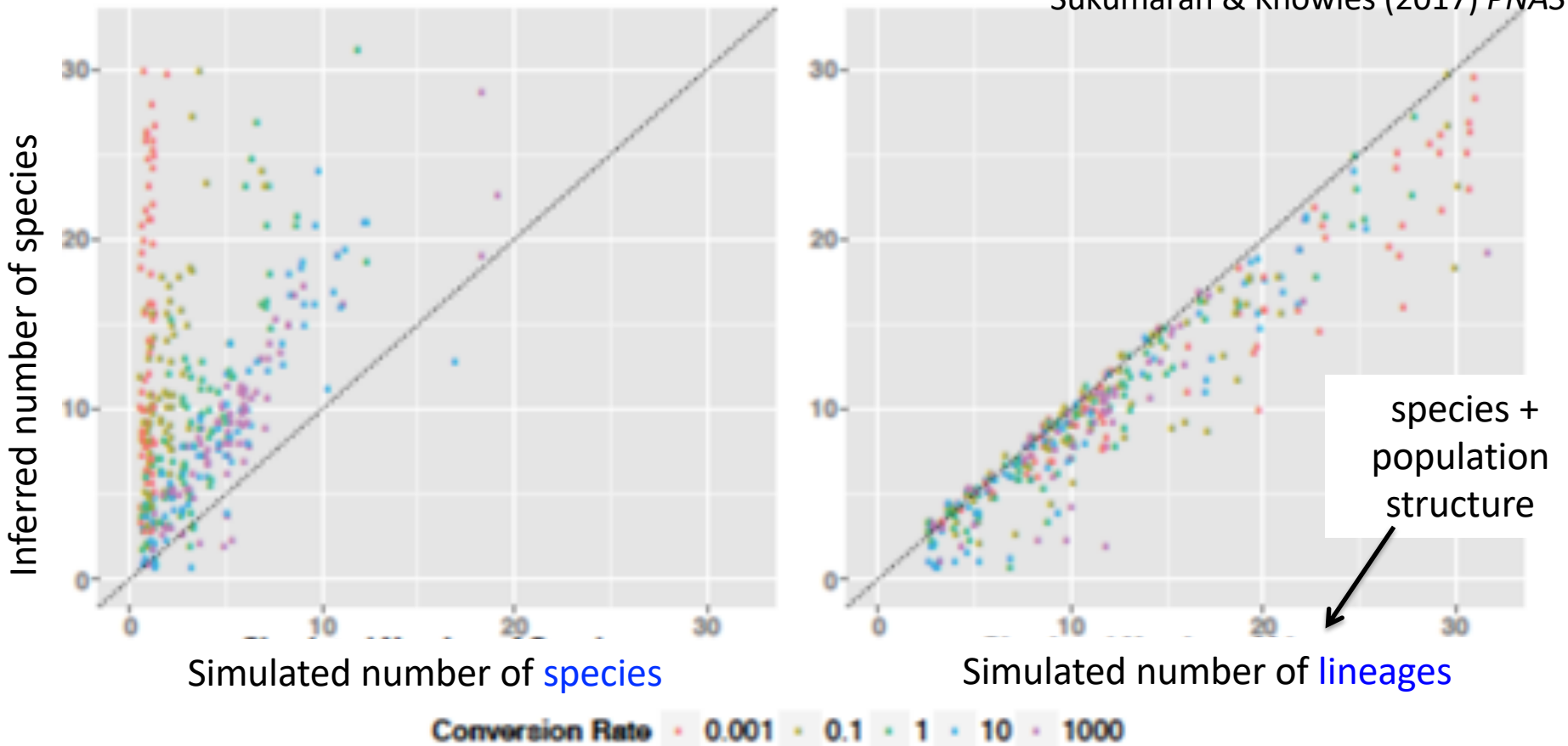


Most probable delimitation model?

Does the MSC accurately delimit species?

Performance of species delimitation under the MSC for data simulated under different speciation durations

Sukumaran & Knowles (2017) *PNAS*



The MSC does not track species, but rather tracks *structure* of any sort, whether population or true 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)

Your reaction to the paper?

Distrust the theoretical demonstration
(maybe specialized scenarios used)?

MSC is incredibly popular, so how could this happen?

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

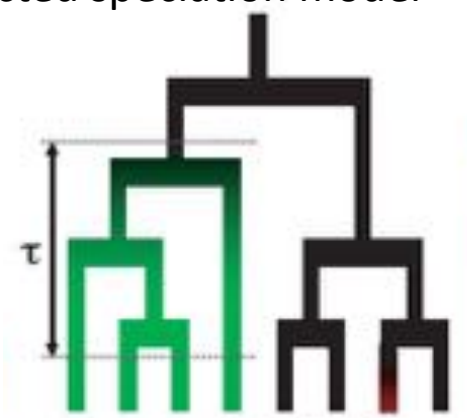
Edited by David M. Hillis, University of Texas at Austin, Austin, TX, and approved December 29, 2016 (received for review May 23, 2016)

Reactions to paper:

- theoretical demonstration, but not practically relevant since we don't know how long it takes for speciation

Protracted speciation model (PSM)

duration of speciation



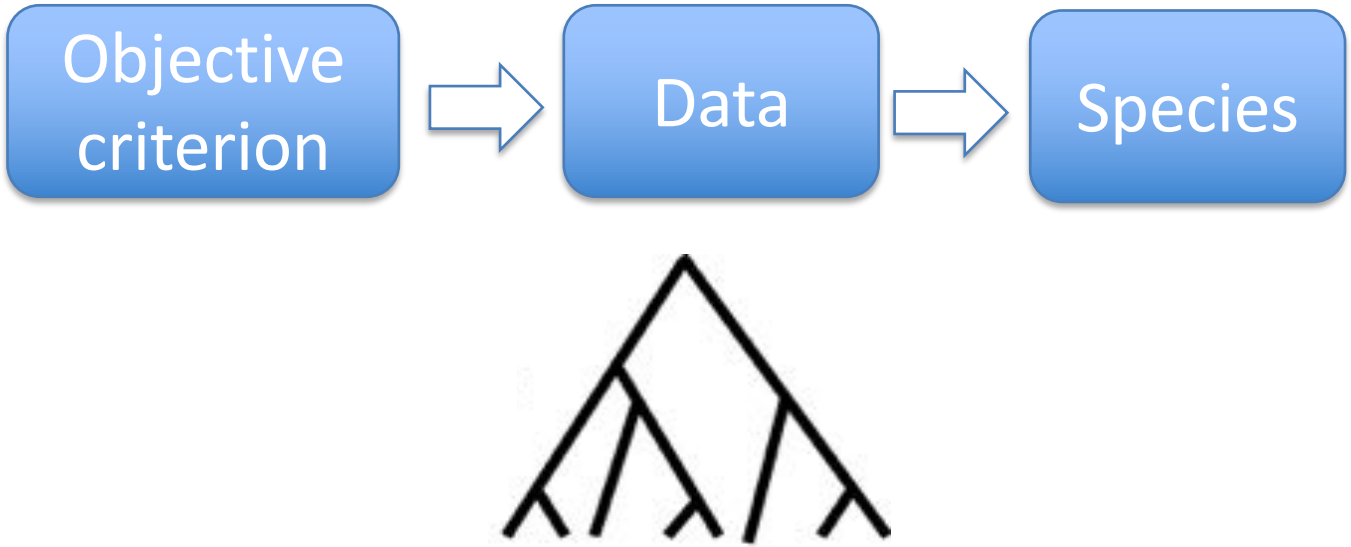
- the protracted model of speciation doesn't fit my empirical system and/or not consistent with any taxonomist views

- everyone recognizes the MSC doesn't delimit species per se

- ~~the MSC doesn't make assumptions about the speciation process~~

Species delimitation

Why are species boundaries NOT accurately delimited under the objective model (the MSC)?



Species delimitation: Subjective + Objective

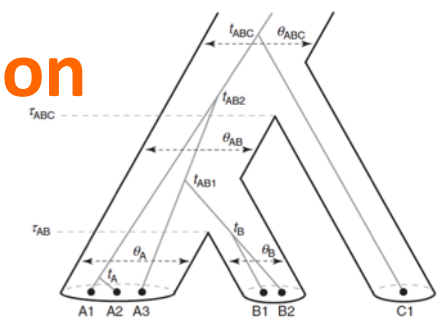
- Choices during model formulation

Species property → Objective criterion → Data → Species

?

Population property

Isolation



?

Population

Transformative potential of model-based analyses:

- Choices during model formulation

- All models are flawed..., some are more or less useful

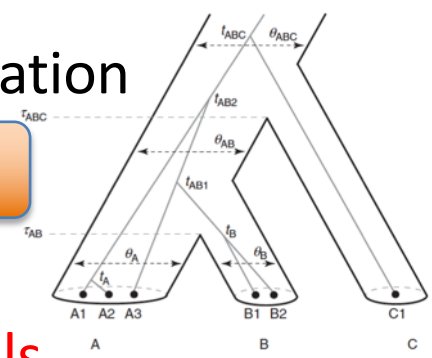
....models are how we communicate
our knowledge to a statistical apparatus

Species properties → Objective criterion → Data

Species

Isolation

Subjective



Need more complex models

Knowles & Sukumaran (2018) in review

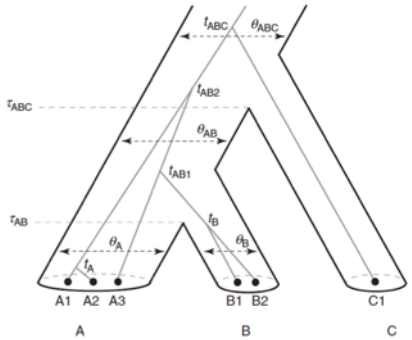
Current state of genetic model-based species delimitation

- MSC detects structure – not species

Sukumaran & Knowles (2017) *PNAS*

(seeking consensus across MSC-based methods is not a good way to fail)

see Rannala (2015) Current Zoology 61, 846-853

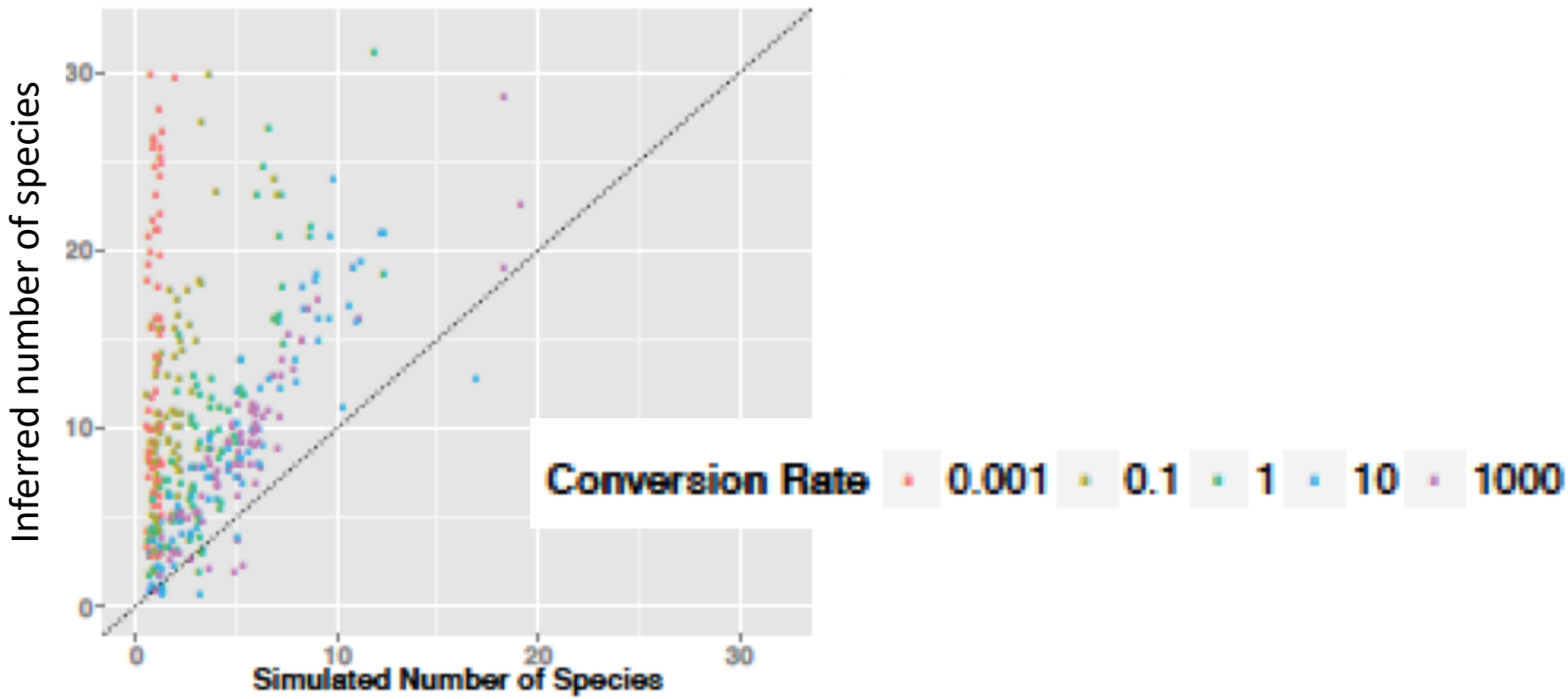


- "Robustness" to lineage detection with low levels of gene flow is not the same as accurate species delimitation

- Sensitivity to sampling (i.e., more data change status)

- In practice, MSC is not a de facto standardization for objectively delimiting taxa: degree of over estimation varies depending on speciation process

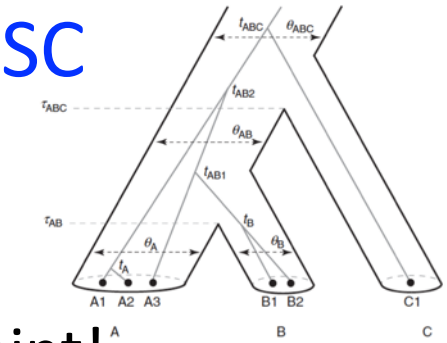
Degree of over-estimation depends upon the speciation process



- In practice, MSC is not a de facto standardization for objectively delimiting taxa: degree of over estimation varies depending on speciation process

Current state of genetic model-based species delimitation:

Accurate species delimitation cannot be achieved with current models based on MSC



- Don't run MSC and add a caveat – what's the point!
- STOP reporting on all the "cryptic" species diversity

Explosion of applications of the MSC for species delimitation

Bayesian species delimitation usin
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 S
University College Lond
Davis, CA 95616

Received: 15 September 2017 | Revised: 30 March 2018 | Accepted: 3 April 2018
DOI: 10.1111/1755-0998.12887

RESOURCE ARTICLE

WILEY MOLECULAR ECOLOGY RESOURCES

CLADES: A classification-based machine learning method for
tion genetic data

MOLECULAR ECOLOGY
Molecular Ecology (2013) 22, 4369–4383
doi: 10.1111/mec.12413

Lu[3] | Yufeng Wu[1]

SPECIES DE
COALESCEI
ACCURACY
EMPIRICAL
LIOLAEMUS
(SQUAMAT

INVITED REVIEWS AND META-ANALYSES

How to fail at species delimitation

BRYAN C. CARSTENS,* TARA A. PELLETIER,* NOAH M. REID† and JO
*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

er the multispecies
provements to DNA

Arley Camargo,[1,2] Mariana Morando,[3] Luciano J. Avila,[3] and Jack W. Sites, Jr.[1]
[1]Department of Biology & Monte L. Bean Museum, Brigham Young University, Provo, Utah 84602
[2]E-mail: arley.camargo@gmail.com
[3]CONICET-CENPAT, Boulevard Almirante Brown 2915, U9120ACD, Puerto Madryn, Chubut, Argentina

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

Bioinformatics, 31(7), 2015, 991–998
doi: 10.1093/bioinformatics/btu770
Advance Access Publication Date: 23 November 2014
Original Paper

NATHAN D. JACKSON[1], ARIADNA E. MORALES[2], BRYAN C. CARSTENS[2] AND BRIAN C. O
[1]Department of Ecology and Evolutionary Biology, University of Tennessee, 442 Hesler Biology Building, Knoxville, T
[2]Department of Evolution, Ecology and Organismal Biology, Ohio State University, 318 W. 12th Avenue, Columbus

Syst. Biol. 0(0):1–13, 2018

DOI:10.1093/sysbio/syy011

Phylogenetics

DISSECT: an assignment-free Bayesian discovery method for species delimitation under the multispecies coalescent

Graham Jones[1], Zeynep Aydin[1,2] and Bengt Oxelman[1,*]

[1]Department of Biological and Environmental Sciences, University of Gothenburg, Box 461, SE 405 30 Göteborg, Sweden and [2]Department of Biology, Faculty of Sciences, University of Dicle, 21280 Diyarbakir, Turkey

Comparison of Methods for Molecular Species Delimitation Acr
Speciation Scenarios

ARONG LUO[1,2,*], CHENG LING[3], SIMON Y. W. HO[2], AND CHAO-DONG ZI
[1]Key Laboratory of Zoological Systematics and Evolution, Institute of Zoology, Chinese Academy of Sciences,
[2]School of Life and Environmental Sciences, University of Sydney, Sydney, New South Wales 2006, Australia; [3]Depar
Technology, College of Information Science and Technology, Beijing University of Chemical Technology, Beiji
[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

Potentially misleading!
Ok, but what to do?

Current state of genetic model-based species delimitation

Ad hoc heuristics to interpret inferences under the MSC

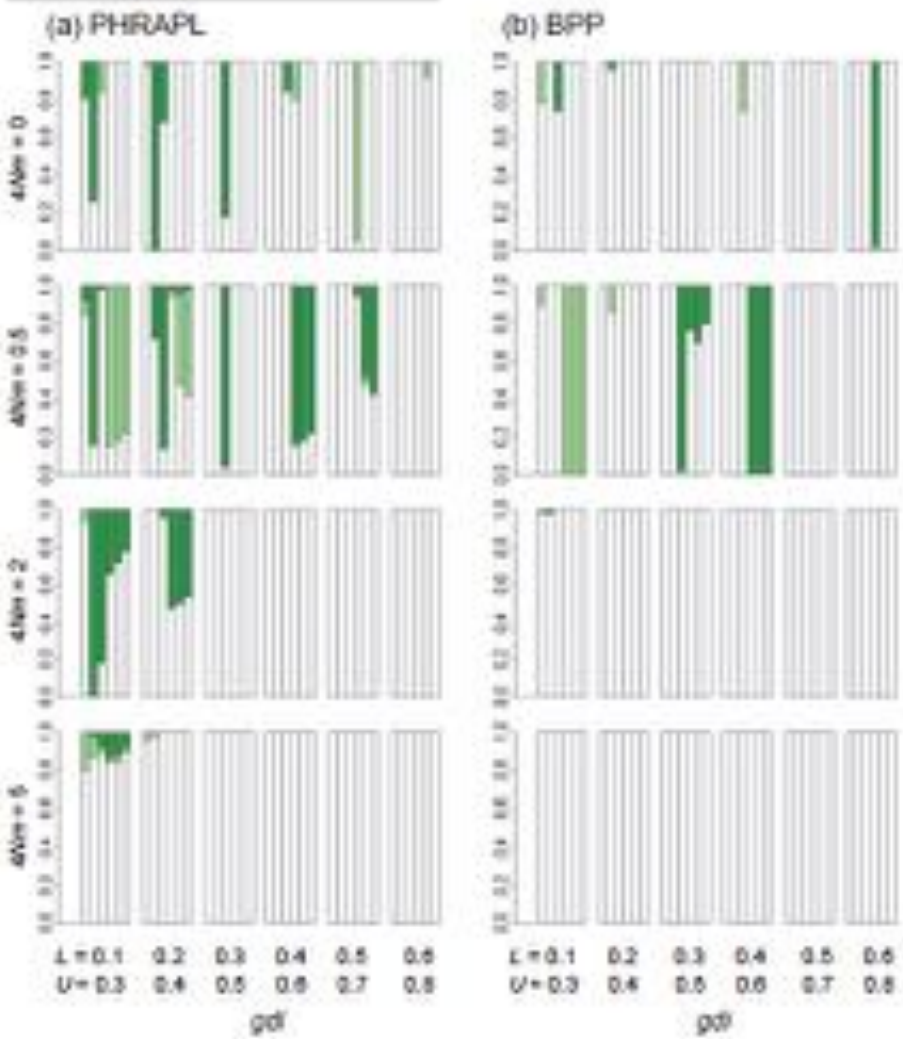


FIGURE 4. Accuracy of species delimitation using the *gdi* with parameters estimated from data of 50 loci using (a) PHRAPL and (b) BPP.

*Jackson et al. (2018) *Syst. Biol.*

*Leache et al. (2018) *Syst. Biol.*

- Genealogical sorting index*: $2\tau/\theta$
  (i.e., population divergence time relative to the population size)
  Cummings et al. (2008) Evolution

- ambiguity with *gdi* when the two populations have different sizes

- *gdi* may lead to claims of species status even if populations diverged very recently if one population established by a few founder individuals

Current state of genetic ~~model based~~ species delimitation

Ad hoc heuristics to interpret inferences under the MSC *

* What constitutes a species is a decision based on applying a threshold index value (despite Bayesian framework to rework the old idea)

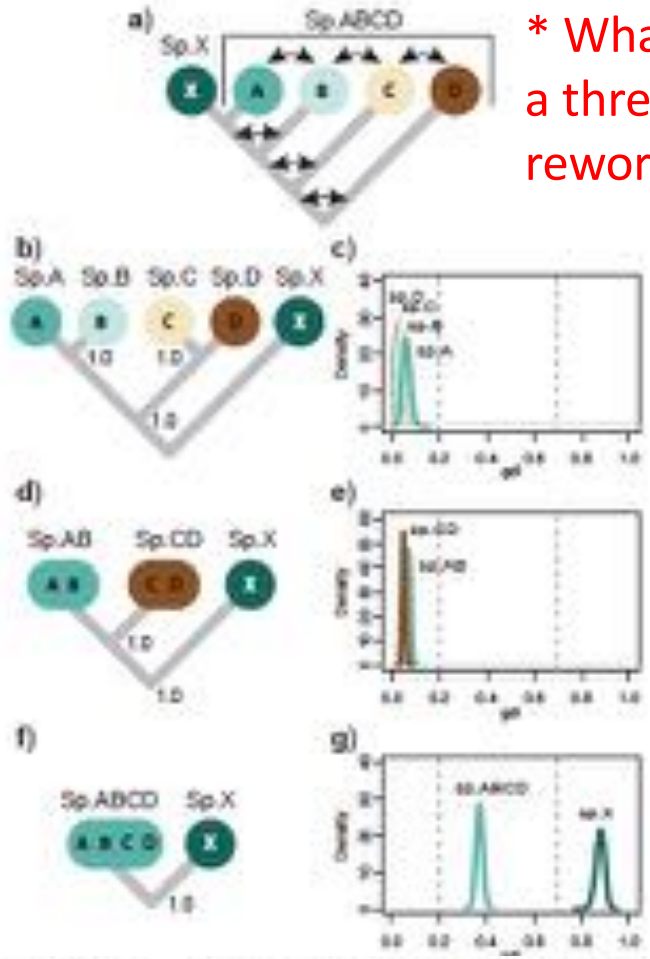


FIGURE 5. Species delimitation applying heuristic index *gdi* to parameter estimates from BPP. a) Species tree used for simulation allows

Leache et al. (2018) *Syst. Biol.*

- Genealogical sorting index*: $2\tau/\theta$

  (i.e., population divergence time relative to the population size)

  Cummings et al. (2008) Evolution

- hierarchical procedure for applying *gdi* index

- Bayesian framework for calculating posterior distribution of *gdi*

Model-based delimitation: state of the field

- Erroneous species boundaries are inferred from current model-based genetic approaches under the MSC

- Relying on heuristics to interpret inferences under the MSC (e.g., from bpp) is not the answer

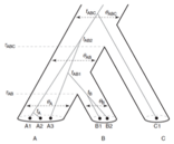
- Future of genetic-based species delimitation is with speciation-based MSC models

A new era of species delimitation models that brings speciation models to the multispecies coalescent

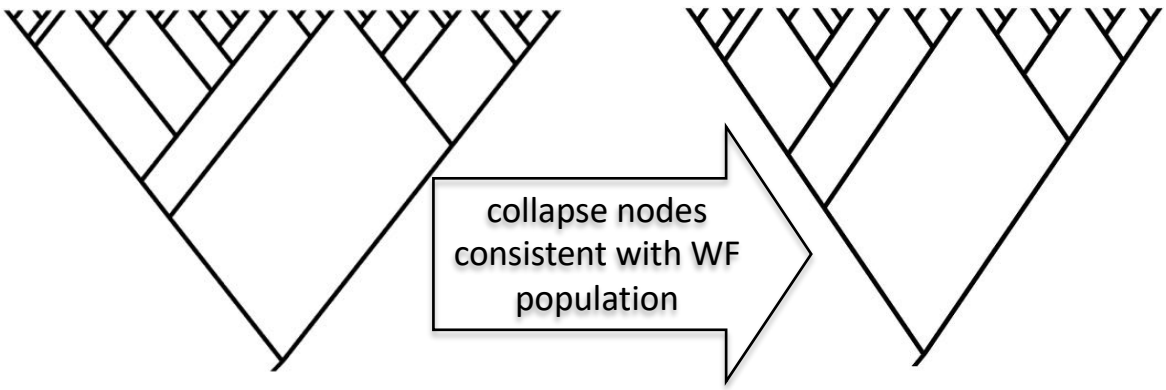
- Erroneous species boundaries are inferred from current model-based genetic approaches under the MSC

- Relying on heuristics to interpret inferences under the MSC (e.g., from bpp) is not the answer

- Future of genetic-based species delimitation is with speciation-based MSC models

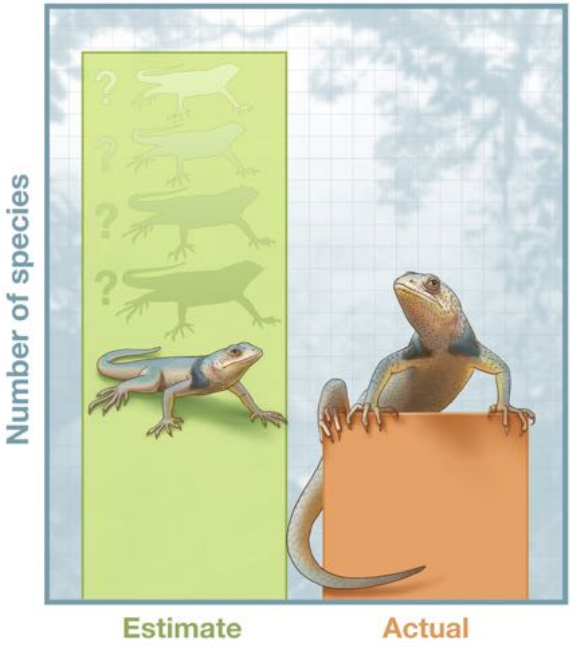
Species delimitation under the MSC: 

- genetic structure = species



40 lineages

27 population lineages



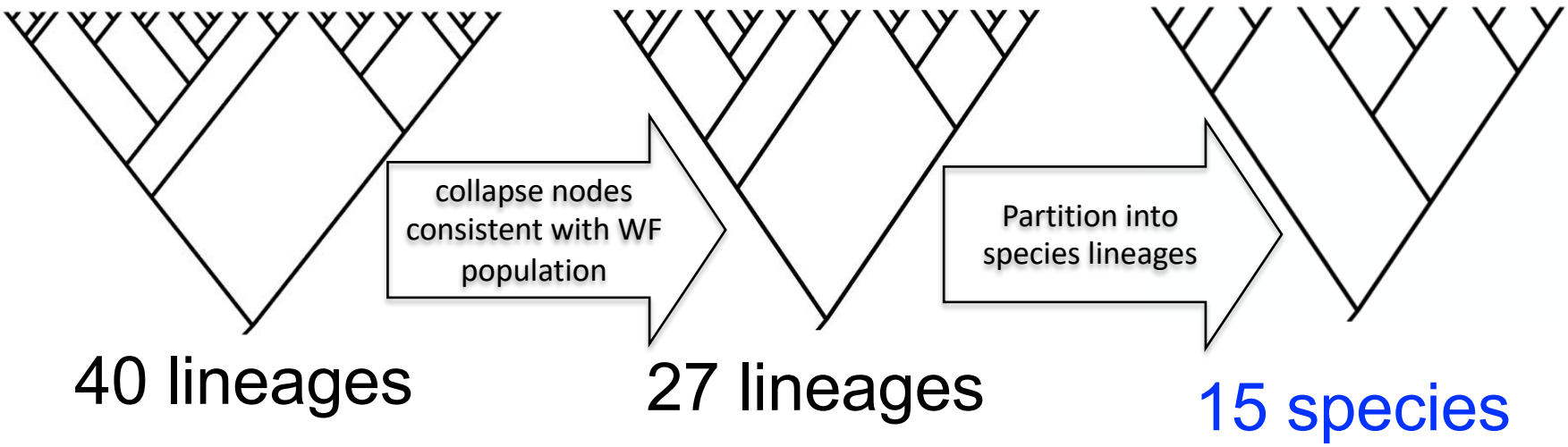
Sukumaran & Knowles (2017) *PNAS*

Incorporating the Speciation Process into Species Delimitation

Jeet Sukumaran[a,c,1,2], Mark T. Holder[b,1], and L. Lacey Knowles[a]

[a]Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor MI, USA 41809-1079; [b]Department of Ecology and Evolutionary Biology, University of Kansas, Lawrence KS, USA 66045; [c]Department of Biology, San Diego State University, San Diego CA, USA 92182-4614

This manuscript was compiled on October 10, 2019

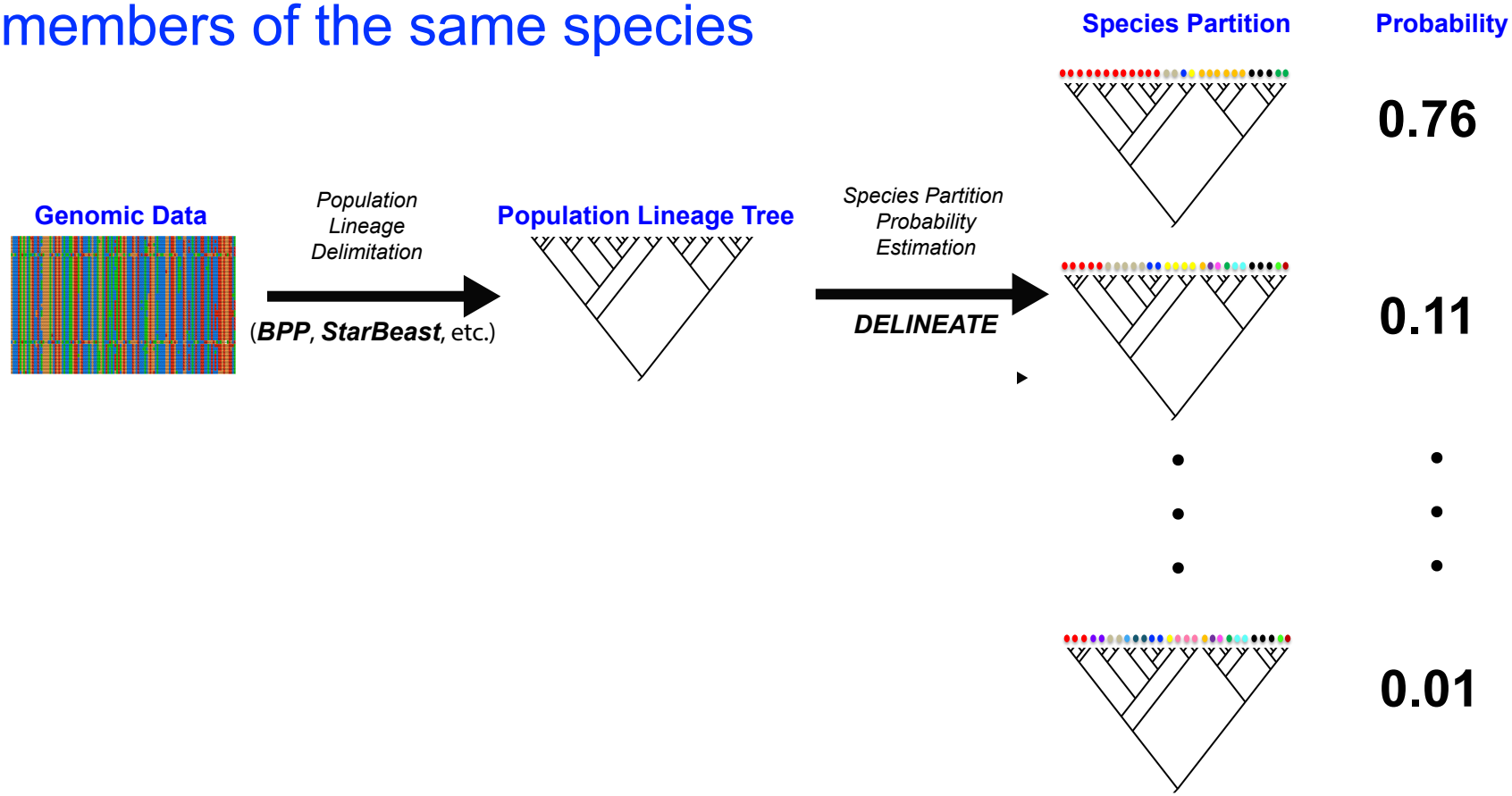


DELINEATE: a species delimitation method which makes probabilistic statements about whether population lineages are members of the same species

Sukumaran J, Holder M, Knowles LL (in review)

Sukumaran J, Holder M, Knowles LL

DELINEATE: a species delimitation method which makes probabilistic statements about whether population lineages are members of the same species



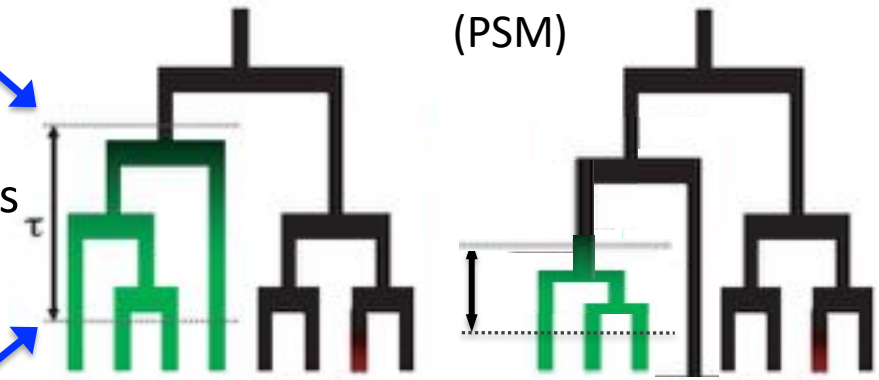
- probabilities of different *partitions* are calculated conditional on the lineage tree and speciation dynamic parameters (e.g., tempo of speciation)

Different speciation-based delimitation models might be used to represent various aspects of the speciation process

Splitting events; initiation of speciation through, e.g., population isolation

Protracted speciation model (PSM)

This process, as modeled here, is initiated by a stochastic lineage splitting process that extends over a duration of time that is determined stochastically by a speciation completion rate parameter



Modified from Rosindell et al. (2010) Ecol. Lett.

Color change indicates completion of speciation and development of "good" species from "incipient" species

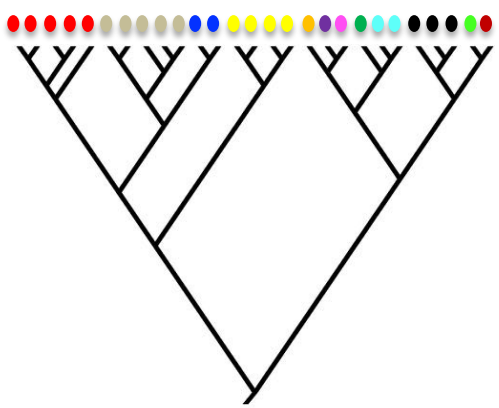
- Transition of an incipient species lineage to full "good" species occurs independently on each branch at the species completion rate, λ_2

Sukumaran J, Holder M, Knowles LL

- Computational challenge of number of possible partitions

Different "partitions"

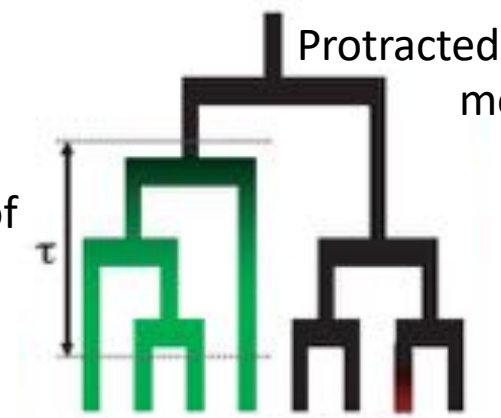
(each partition represents a particular model of delimited species)



- BUT Affinities of some lineages well understood (i.e., include data from well described species) and focus on inferring those less studied

Protracted speciation model (PSM)

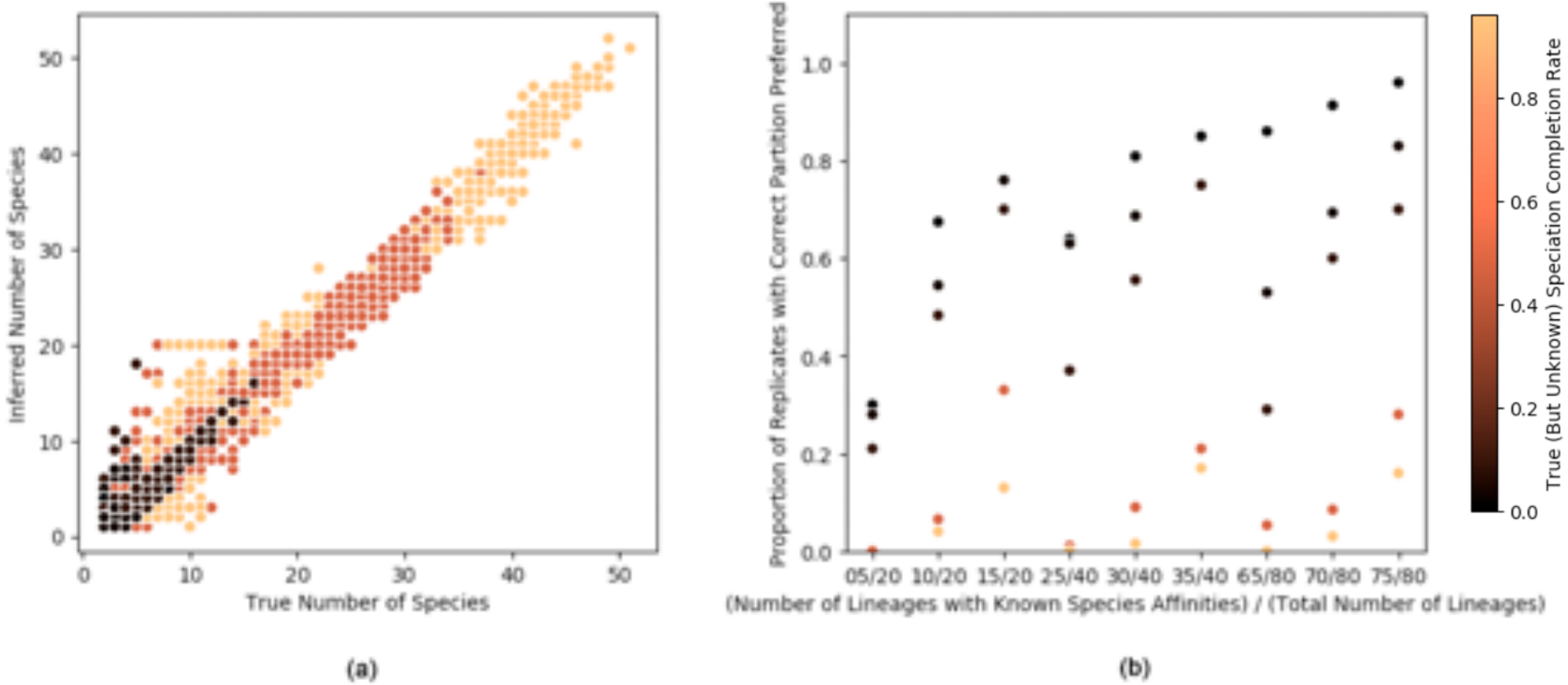
duration of speciation



Modified from Rosindell et al. (2010) Ecol. Lett.

Sukumaran J, Holder M, Knowles LL

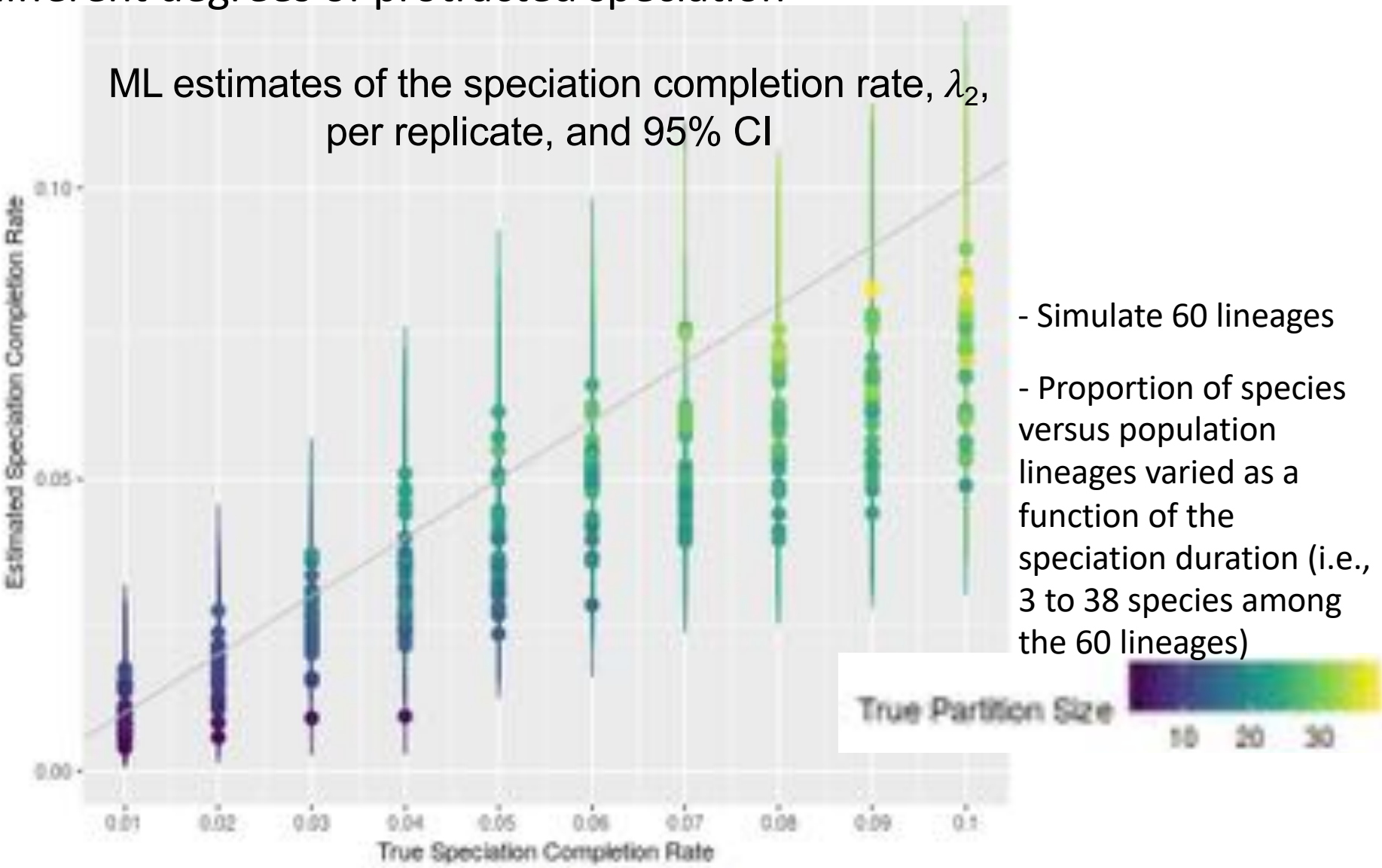
Recovery of (a) true # of species, and (b) the correct partition for different sized trees with different numbers of undescribed lineages



*Note the speciation completion rate is estimated jointly (as long as some constraints on con- and hetero-specific status of some lineages are given)

*Pure-genomic uninformed species delimitation is not practical!

Recovery of the true speciation process from simulated data with different degrees of protracted speciation

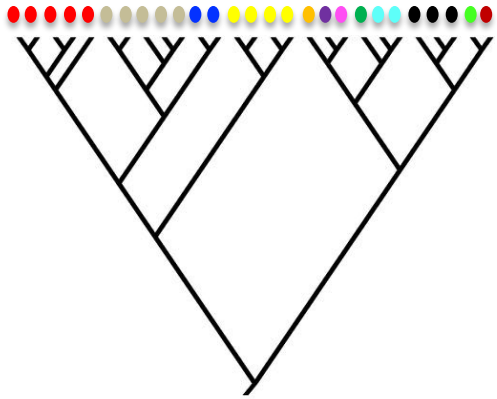


- Simulate 60 lineages

- Proportion of species versus population lineages varied as a function of the speciation duration (i.e., 3 to 38 species among the 60 lineages)

Speciation-based delimitation model

Other applications of DELINEATE:



- Summarize the information in other ways (e.g., probability the leftmost subtree of 5 lineages are conspecific by summing all the probabilities of all partitions in which those lineages occur together)

- Estimate the speciation completion rate given a sample of populations with known species assignments (i.e., focus on speciation dynamics (see Li, Huang, Sukumaran, Knowles (2018) *BMC Evol. Biol*. 18:123)

Sukumaran J, Holder M, Knowles LL

Using genetic data alone (i.e., without conditioning on prior knowledge about some lineages) is not sufficient for accurate inference of species boundaries.

Software: *Decrypt* https://becheler.github.io/pages/applications.html

- Model of the geography of genetic divergence under a spatially explicit coalescent to evaluate competing hypotheses about cryptic diversity (inferred under the MSC)

- Practical training (tonight 7-10pm)

Software: *Delineate*

Software: *Decrypt*

Analysis using DECRYPT

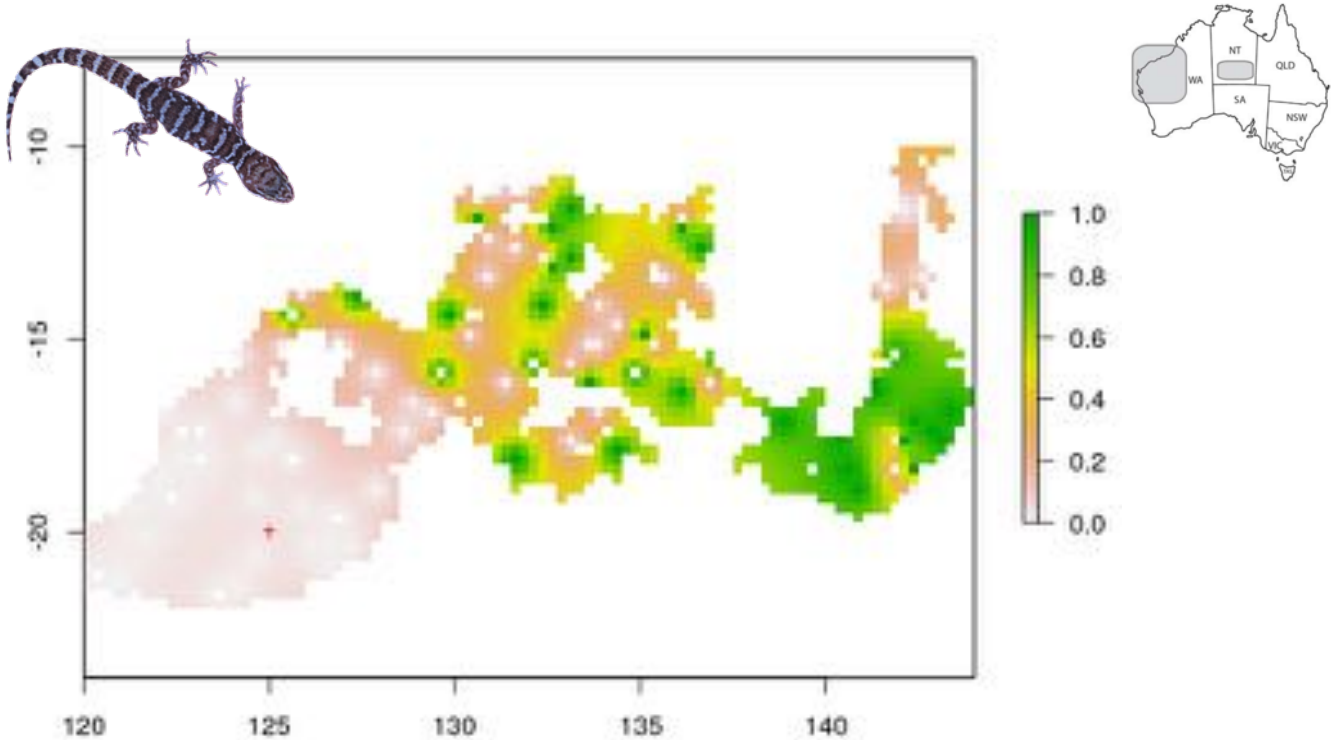


Figure 7: Spatial interpolation of p_x the probability to detect 2 species in a population expanding in an heterogeneous landscape under the MSC when the sequences sample is constructed at time t_s by two 2D gaussian sampling processes centered on (i) the population origin x_0 (red cross), and (ii) on a random coordinate x (with $N(x,\ t_s)$ > 30 to avoid inconsistent sampling in very low density areas).

Transformative potential of model-based analyses:

(i) Phylogenetic inference

(ii) Species delimitation/inferring species boundaries

(iii) Phylogeography/Comparative Phylogeography

With an emphasis on:

- Choices we make about model formulation
- Recognizing the subjectivity of model formulation itself when making inferences
- Decisions when applying to empirical data

Model-based approaches for phylogeographic inference

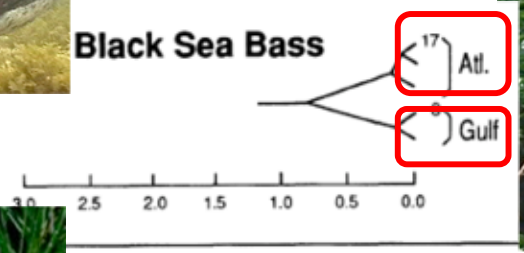
Discussion points:

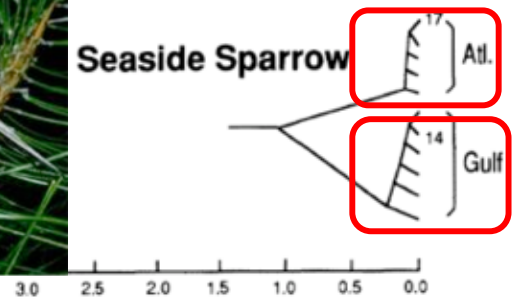
- Why models are important
- Generic versus informed models
- Species-specific expectations of genetic variation (e.g. iDDC; based on spatially explicit coalescent models)
- Concordance versus discord among species: lessons from comparative phylogeography

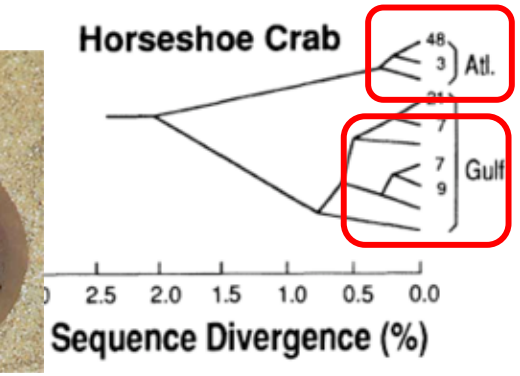
Why the transition from describing patterns of genetic variation to understanding process requires model-based approach

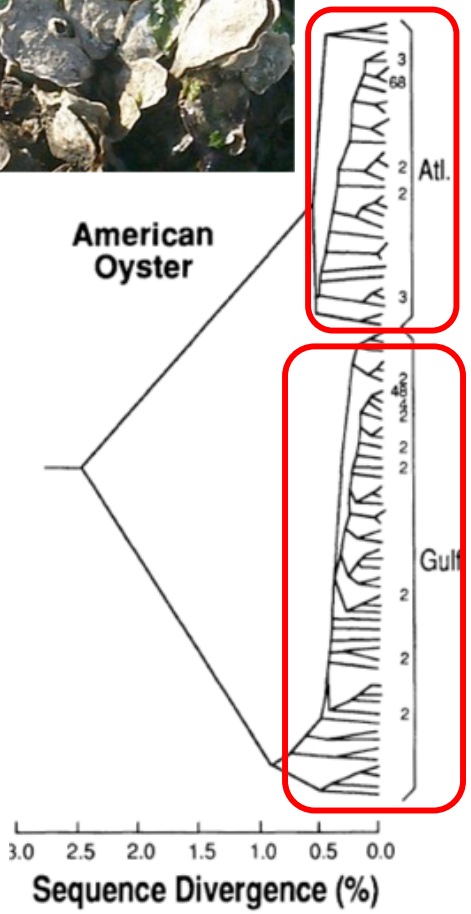
Classics in phylogeography

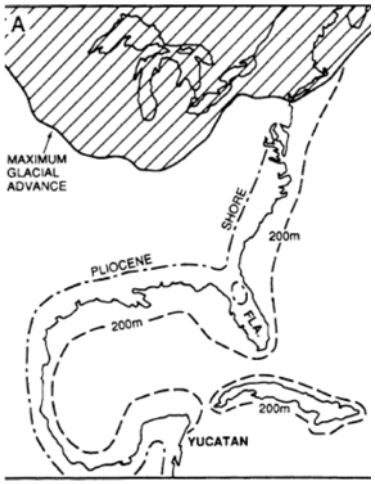
Concordance reflects a common vicariant history of population separation









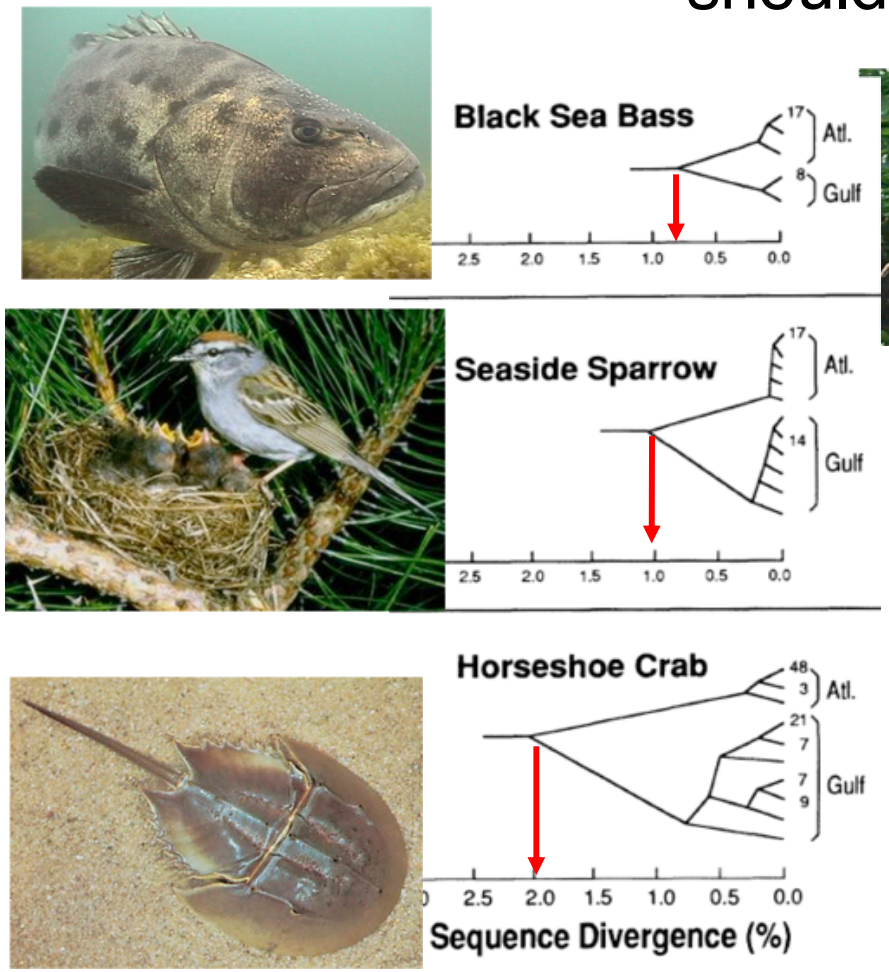


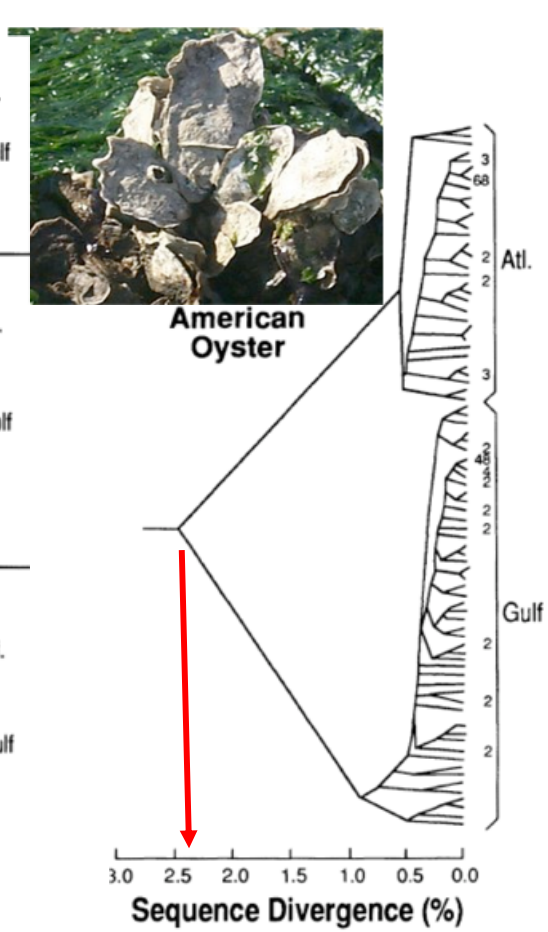
Avise 1992

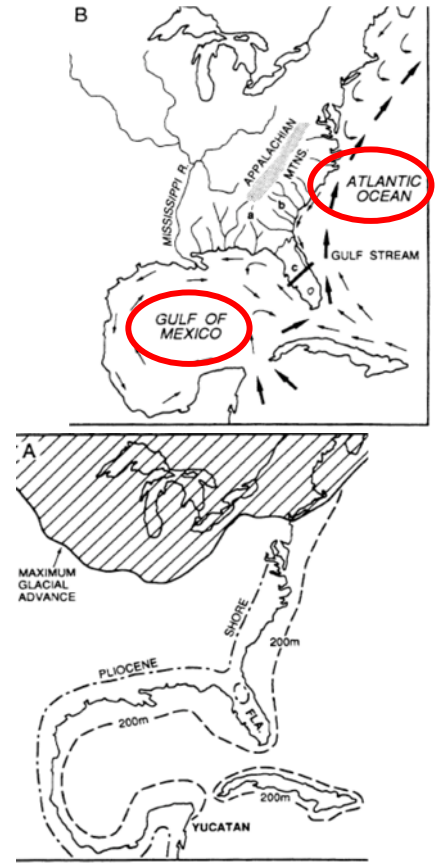
A common vicariant history?

The data may be consistent with a common response to a specific geologic event, despite differing gene tree depths among taxa? Or maybe not?

By looking only at the gene trees,
it isn't clear how the differences in gene tree depths should be interpreted!

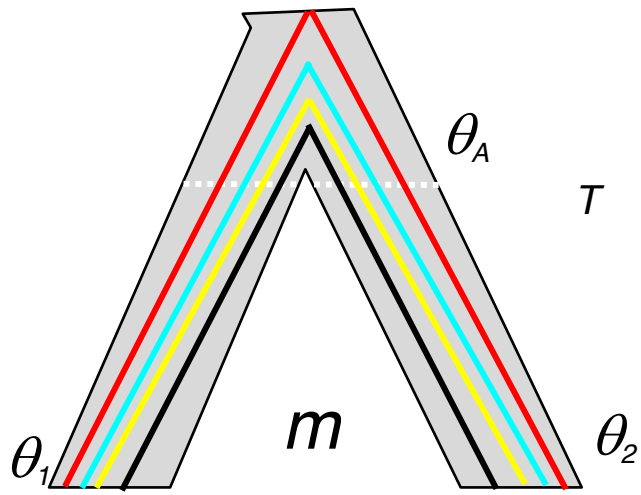






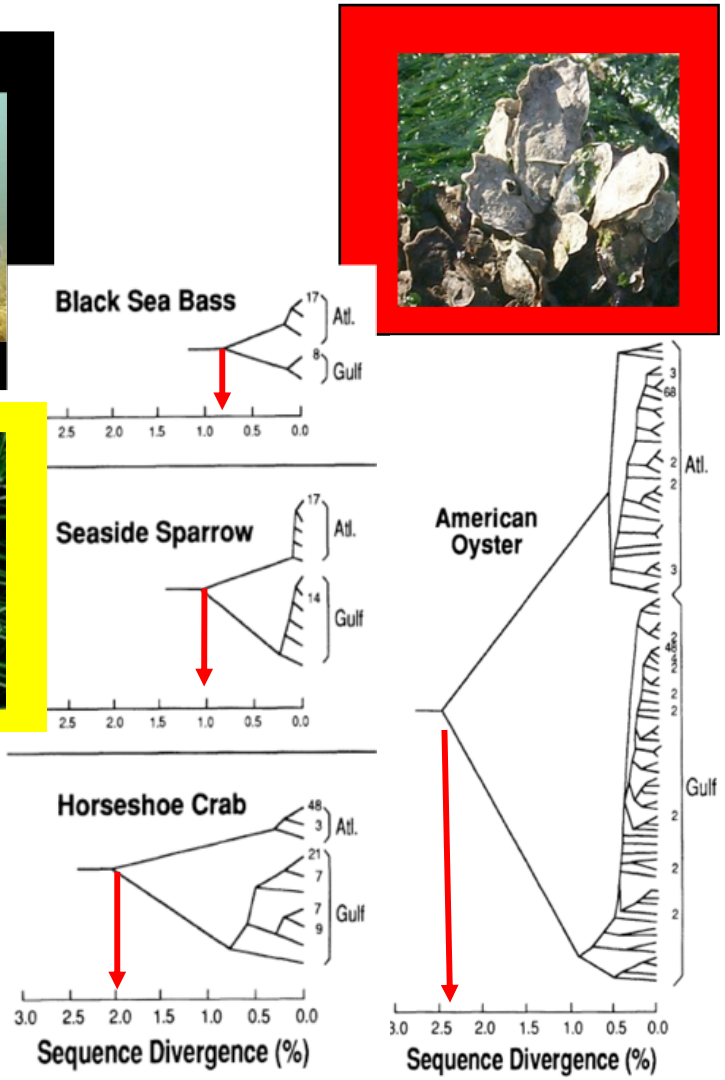
To test for a common vicariant history need to:

Assess statistically how much of a difference in the depths of the gene trees would still be consistent with the same time of population divergence



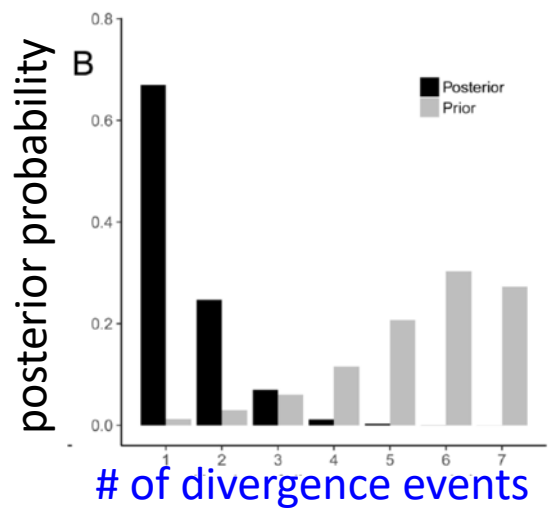
Present

A common vicariant history?



Concordance used in statistical phylogeography

Statistically evaluate a parameterized model of co-divergence among species using hierarchical Approximate Bayesian Computation (hABC)



How do we decide upon a model*:

- arbitrary/generic models

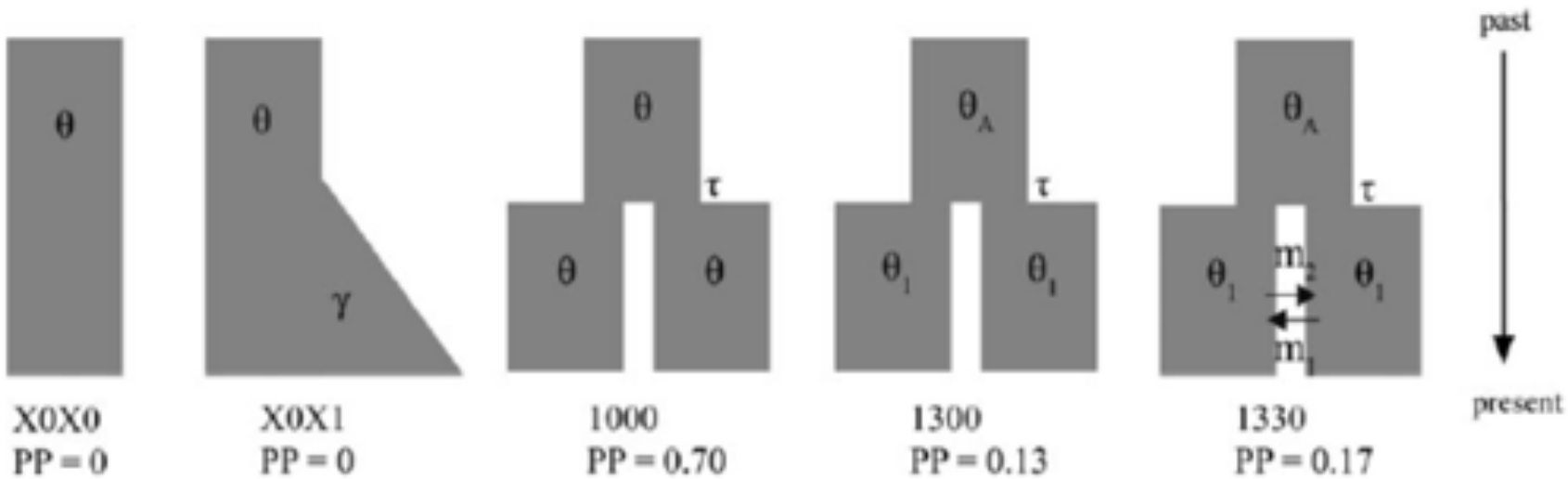
- informed from information independent of the genetic data itself
  - that is, a specific biological narrative motivates the model
  - models informed by independent genetic data

* All models are simplifications, and vary in their relative degree of abstraction

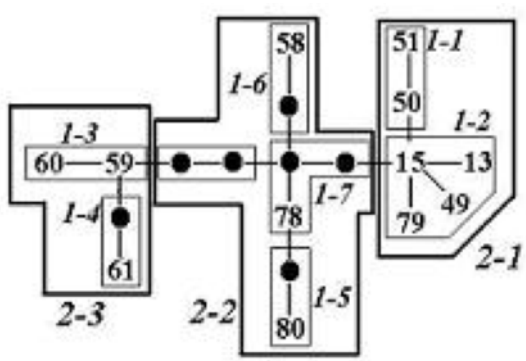
Model choice in phylogeography: generic versus informed

- arbitrary/generic models

Tests of 142 objectively identified models (e.g., program like PHRAPL)



Pelletier & Carstens (2014 Mol. Ecol.)



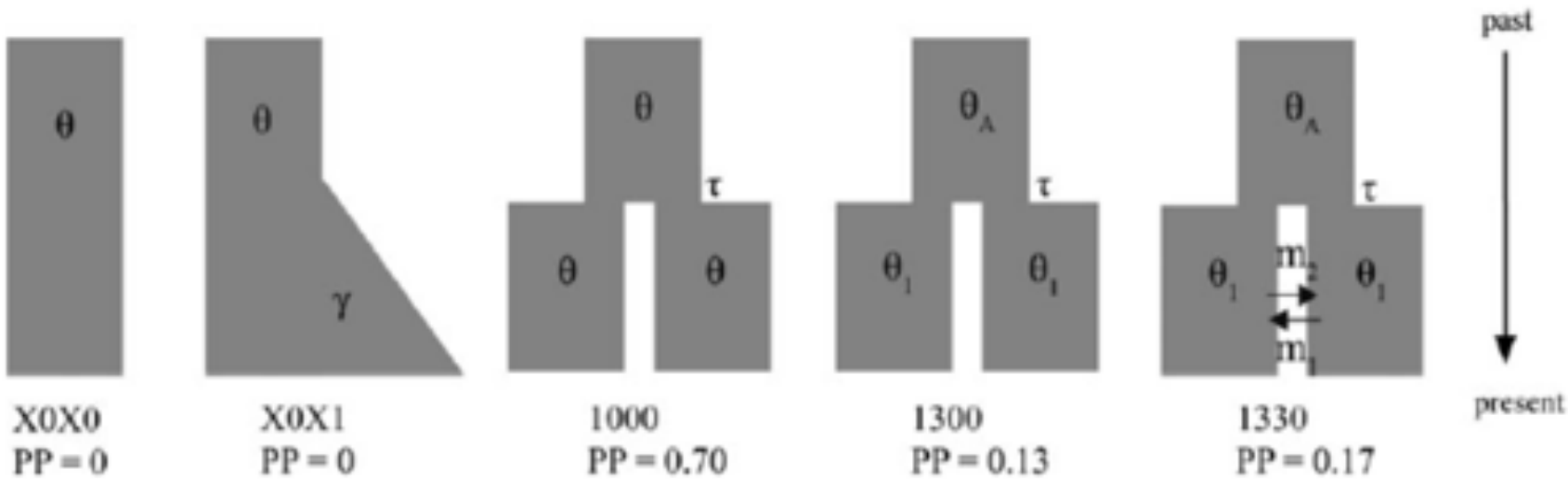
Nested Clade Analysis (NCA): the data itself tell us what history generated it

(Discredited in early 2000s)

Model choice in phylogeography: subjectivity versus objectivity

- arbitrary/generic models

Tests of 142 objectively identified models

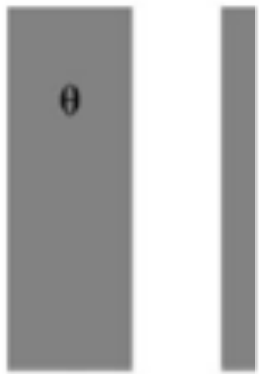


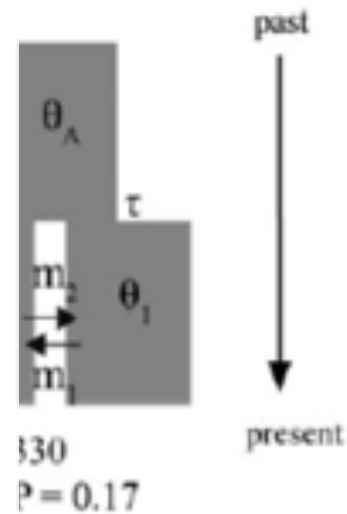
Pelletier & Carstens (2014 Mol. Ecol.)

Statistical procedures themselves may seem to provide a legitimacy to an approach – the advocacy of objective models in phylogeography

Model choice ectivity

Tests of 142 c





& Carstens (2014 Mol. Ecol.)

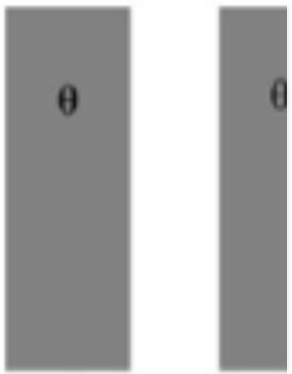
Table 3 List of all 143 models included in analyses. Model = τθmγ

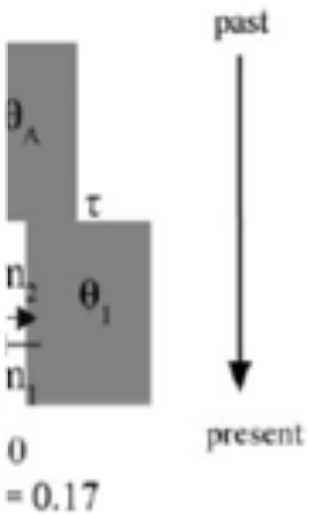
Model	Parameters	Mean	SD	Median	Posterior probability
1030	**$\tau, \theta_A = \theta_1 = \theta_2, m_{12}, m_{21}$**	**0.792**	**1.124**	**0.000**	**0.024**
1232	$\tau, \theta_A = \theta_2, \theta_1, m_{12}, m_{21}, \gamma_2$	0.822	0.856	0.772	0.007
1200	$\tau, \theta_A = \theta_2, \theta_1$	0.836	0.985	0.499	0.004
1222	$\tau, \theta_A = \theta_2, \theta_1, m_{21}, \gamma_2$	0.846	0.982	0.542	0.006
1220	$\tau, \theta_A = \theta_2, \theta_1, m_{21}$	0.849	0.957	0.647	0.006
1231	$\tau, \theta_A = \theta_2, \theta_1, m_{12}, m_{21}, \gamma_1$	0.863	0.877	0.859	0.006
1221	$\tau, \theta_A = \theta_2, \theta_1, m_{21}, \gamma_1$	0.870	0.878	0.862	0.011
1031	**$\tau, \theta_A = \theta_1 = \theta_2, m_{12}, m_{21}, \gamma_1$**	**0.886**	**1.133**	**0.000**	**0.020**
1230	$\tau, \theta_A = \theta_2, \theta_1, m_{12}, m_{21}$	0.917	0.937	0.880	0.006
1033	**$\tau, \theta_A = \theta_1 = \theta_2, m_{12}, m_{21}, \gamma_1, \gamma_2$**	**0.923**	**1.170**	**0.000**	**0.018**
0131	$\theta_A = \theta_1, \theta_2, m_{12}, m_{21}, \gamma_1$	0.930	1.024	0.779	0.007
0130	$\theta_A = \theta_1, \theta_2, m_{12}, m_{21}$	0.949	0.881	1.055	0.010
1023	**$\tau, \theta_A = \theta_1 = \theta_2, m_{12}, \gamma_1, \gamma_2$**	**0.956**	**1.154**	**0.000**	**0.024**
1201	$\tau, \theta_A = \theta_2, \theta_1, \gamma_1$	0.975	1.026	0.866	0.006
0030	**$\theta_A = \theta_1 = \theta_2, m_{12}, m_{21}$**	**0.977**	**1.210**	**0.000**	**0.024**
1211	$\tau, \theta_A = \theta_2, \theta_1, m_{12}, \gamma_1$	0.990	1.042	0.927	0.007
0020	**$\theta_A = \theta_1 = \theta_2, m_{12}, m_{21}$**	**0.991**	**1.264**	**0.000**	**0.017**
1132	$\tau, \theta_A = \theta_1, \theta_2, m_{12}, m_{21}, \gamma_2$	0.995	0.981	0.986	0.007
0031	**$\theta_A = \theta_1 = \theta_2, m_{12}, m_{21}, \gamma_1$**	**0.996**	**1.303**	**0.000**	**0.020**
0022	**$\theta_A = \theta_1 = \theta_2, m_{21}, \gamma_2$**	**1.003**	**1.241**	**0.000**	**0.025**
1131	$\tau, \theta_A = \theta_1, \theta_2, m_{12}, m_{21}, \gamma_1$	1.011	0.967	1.013	0.004
1032	**$\tau, \theta_A = \theta_1 = \theta_2, m_{12}, m_{21}, \gamma_2$**	**1.013**	**1.212**	**0.000**	**0.031**
1212	$\tau, \theta_A = \theta_2, \theta_1, m_{12}, \gamma_2$	1.015	0.986	1.083	0.003
1233	$\tau, \theta_A = \theta_2, \theta_1, m_{12}, m_{21}, \gamma_1, \gamma_2$	1.021	0.946	1.121	0.010
1203	$\tau, \theta_A = \theta_2, \theta_1, \gamma_1, \gamma_2$	1.024	1.058	1.002	0.010
0233	$\theta_A = \theta_2, \theta_1, m_{12}, m_{21}, \gamma_1, \gamma_2$	1.026	0.985	1.118	0.004
1110	$\tau, \theta_A = \theta_1, \theta_2, m_{12}, \gamma_1$	1.030	1.003	1.118	0.007
0222	$\theta_A = \theta_2, \theta_1, m_{21}, \gamma_2$	1.031	1.112	0.921	0.008
1130	$\tau, \theta_A = \theta_1, \theta_2, m_{12}, m_{21}$	1.031	0.976	1.084	0.006
0112	$\theta_A = \theta_1, \theta_2, m_{12}, \gamma_2$	1.032	0.991	1.121	0.007
0032	**$\theta_A = \theta_1 = \theta_2, m_{12}, m_{21}, \gamma_2$**	**1.033**	**1.212**	**0.000**	**0.020**
0110	$\theta_A = \theta_1, \theta_2, m_{12}, \gamma_1$	1.034	1.031	1.070	0.004
1020	$\tau, \theta_A = \theta_1 = \theta_2, m_{12}, m_{21}, \gamma_1, \gamma_2$	1.035	1.196	0.000	0.015
0012	**$\theta_A = \theta_1 = \theta_2, m_{12}, \gamma_2$**	**1.038**	**1.272**	**0.000**	**0.018**
1213	$\tau, \theta_A = \theta_2 = \theta_1, m_{12}, \gamma_1, \gamma_2$	1.041	1.053	1.121	0.003
0220	$\theta_A = \theta_2, \theta_1, m_{21}$	1.041	0.965	1.121	0.010
1013	**$\tau, \theta_A = \theta_1 = \theta_2, m_{12}, \gamma_1, \gamma_2$**	**1.042**	**1.227**	**0.543**	**0.024**
0231	$\theta_A = \theta_2, \theta_1, m_{12}, m_{21}, \gamma_1$	1.048	1.104	0.997	0.007
1111	$\tau, \theta_A = \theta_1, \theta_2, m_{12}, \gamma_1$	1.050	1.027	1.098	0.013
0013	**$\theta_A = \theta_1 = \theta_2, m_{12}, \gamma_1, \gamma_2$**	**1.056**	**1.254**	**0.000**	**0.021**
0133	**$\theta_A = \theta_1, \theta_2, m_{12}, m_{21}, \gamma_1, \gamma_2$**	**1.057**	**1.107**	**1.028**	**0.001**
0033	**$\theta_A = \theta_1 = \theta_2, m_{12}, m_{21}, \gamma_1, \gamma_2$**	**1.059**	**1.289**	**0.000**	**0.031**
1002	$\tau, \theta_A = \theta_1 = \theta_2, \gamma_2$	1.084	1.261	0.000	0.008
1331	$\tau, \theta_A, \theta_1 = \theta_2, m_{12}, m_{21}, \gamma_1$	1.098	1.093	1.081	0.000
0132	$\theta_A = \theta_1, \theta_2, m_{12}, m_{21}, \gamma_2$	1.101	0.991	1.129	0.007
0210	$\theta_A = \theta_2, \theta_1, m_{12}$	1.102	1.111	1.040	0.001
1321	$\tau, \theta_A, \theta_1 = \theta_2, m_{21}, \gamma_1$	1.108	1.012	1.124	0.000
1123	$\tau, \theta_A = \theta_1, \theta_2, m_{21}, \gamma_1, \gamma_2$	1.118	1.094	1.121	0.003
1021	**$\tau, \theta_A = \theta_1 = \theta_2, m_{12}, \gamma_1$**	**1.119**	**1.323**	**0.000**	**0.036**
1113	$\tau, \theta_A = \theta_1, \theta_2, m_{12}, \gamma_1, \gamma_2$	1.132	1.042	1.129	0.003
1010	$\tau, \theta_A = \theta_1 = \theta_2, m_{12}$	1.135	1.284	0.558	0.013
1112	$\tau, \theta_A = \theta_1, \theta_2, m_{12}, \gamma_1$	1.135	0.943	1.137	0.006
1101	$\tau, \theta_A = \theta_1, \theta_2, \gamma_1$	1.136	1.048	1.129	0.006
1011	**$\tau, \theta_A = \theta_1 = \theta_2, m_{12}, \gamma_1$**	**1.148**	**1.274**	**0.739**	**0.021**
0023	**$\theta_A = \theta_1 = \theta_2, m_{12}, \gamma_1, \gamma_2$**	**1.154**	**1.311**	**0.500**	**0.020**

- arbitrary/generic models

Model choice i ctivity

Tests of 142 ob





Carstens (2014 Mol. Ecol.)

Table 3 *Continued*

Model	Parameters	Mean	SD	Median	Posterior probability
0230	$\theta_A = \theta_2, \theta_1, m_{12}, m_{21}$	1.172	1.022	1.135	0.003
0321	$\theta_A, \theta_1 = \theta_2, m_{12}, m_{21}, \gamma_1$	1.173	1.106	1.129	0.003
1000*	$\tau, \theta_A = \theta_1 = \theta_2$	1.178	1.261	0.971	0.015
1202	$\tau, \theta_A = \theta_1 = \theta_2, \gamma_2$	1.180	1.163	1.124	0.004
0223	$\theta_A = \theta_2, \theta_1, m_{01}, \gamma_1, \gamma_2$	1.181	1.173	1.124	0.007
1001	$\tau, \theta_A = \theta_1 = \theta_2, \gamma_1$	**1.187**	**1.328**	**0.752**	**0.021**
0011	$\theta_A = \theta_1 = \theta_2, m_{12}, \gamma_1$	**1.198**	**1.298**	**0.931**	**0.022**
0213	$\theta_A = \theta_2, \theta_1, m_{12}, \gamma_1, \gamma_2$	1.199	1.117	1.135	0.004
1102	$\tau, \theta_A = \theta_1, \theta_2, \gamma_2$	1.205	1.217	1.129	0.004
1121	$\tau, \theta_A = \theta_1, \theta_2, m_{01}, \gamma_1$	1.211	1.141	1.137	0.010
1022	$\tau, \theta_A = \theta_1 = \theta_2, m_{01}, \gamma_2$	**1.214**	**1.308**	**1.011**	**0.021**
1012	$\tau, \theta_A = \theta_1 = \theta_2, m_{12}, \gamma_2$	**1.270**	**1.324**	**1.129**	**0.021**
1332	$\tau, \theta_A, \theta_1 = \theta_2, m_{12}, m_{21}, \gamma_2$	1.271	1.159	1.179	0.003
1322	$\tau, \theta_A, \theta_1 = \theta_2, m_{01}, \gamma_2$	1.280	1.087	1.233	0.000
0212	$\theta_A = \theta_2, \theta_1, m_{12}, \gamma_2$	1.281	1.181	1.140	0.001
1312	$\tau, \theta_A, \theta_1 = \theta_2, m_{12}, \gamma_2$	1.286	1.105	1.221	0.001
1323	$\tau, \theta_A, \theta_1 = \theta_2, m_{01}, \gamma_1, \gamma_2$	1.312	1.075	1.239	0.001
0123	$\theta_A = \theta_1, \theta_2, m_{01}, \gamma_1, \gamma_2$	1.312	1.189	1.192	0.007
1003	$\tau, \theta_A = \theta_1 = \theta_2, \gamma_1, \gamma_2$	1.321	1.443	1.122	0.007
0313	$\theta_A, \theta_1 = \theta_2, m_{12}, \gamma_1, \gamma_2$	1.327	1.207	1.182	0.001
1433	$\tau, \theta_A, \theta_1, \theta_2, m_{12}, m_{21}, \gamma_1, \gamma_2$	1.327	0.998	1.269	0.000
0312	$\theta_A, \theta_1 = \theta_2, m_{12}, \gamma_2$	1.328	1.201	1.209	0.004
0211	$\theta_A = \theta_2, \theta_1, m_{12}, \gamma_1$	1.333	1.195	1.256	0.006
1320	$\tau, \theta_A, \theta_1 = \theta_2, m_{21}$	1.336	1.235	1.180	0.001
1403	$\tau, \theta_A, \theta_1, \theta_2, \gamma_1, \gamma_2$	1.350	1.011	1.298	0.000
1330*	$\tau, \theta_A, \theta_1 = \theta_2, m_{12}, m_{21}$	1.351	1.274	1.225	0.006
0323	$\theta_A, \theta_1 = \theta_2, m_{21}, \gamma_1, \gamma_2$	1.353	1.170	1.259	0.003
1333	$\tau, \theta_A, \theta_1 = \theta_2, m_{12}, m_{21}, \gamma_1, \gamma_2$	1.357	1.127	1.277	0.003
1103	$\tau, \theta_A = \theta_1, \theta_2, \gamma_1, \gamma_2$	1.400	1.186	1.408	0.003
1423	$\tau, \theta_A, \theta_1, \theta_2, m_{21}, \gamma_1, \gamma_2$	1.408	1.502	1.182	0.001
0331	$\theta_A, \theta_1 = \theta_2, m_{12}, m_{21}, \gamma_1$	1.424	1.314	1.368	0.000
0311	$\theta_A, \theta_1 = \theta_2, m_{12}, \gamma_1$	1.475	1.353	1.353	0.003
1432	$\tau, \theta_A, \theta_1, \theta_2, m_{12}, m_{21}, \gamma_2$	1.500	1.297	1.360	0.000
1402	$\tau, \theta_A, \theta_1, \theta_2, \gamma_2$	1.543	1.101	1.545	0.003
0413	$\theta_A, \theta_1, \theta_2, m_{12}, \gamma_1, \gamma_2$	1.570	1.139	1.545	0.006
0412	$\theta_A, \theta_1, \theta_2, m_{12}, \gamma_2$	1.575	1.172	1.516	0.001
0322	$\theta_A, \theta_1 = \theta_2, m_{01}, \gamma_2$	1.591	1.493	1.481	0.001
1303	$\tau, \theta_A, \theta_1 = \theta_2, \gamma_1, \gamma_2$	1.591	1.303	1.610	0.003
1301	$\tau, \theta_A, \theta_1 = \theta_2, \gamma_1$	1.621	1.428	1.554	0.001
1300*	$\tau, \theta_A, \theta_1 = \theta_2$	1.630	1.342	1.562	0.004
1313	$\tau, \theta_A, \theta_1 = \theta_2, m_{12}, \gamma_1, \gamma_2$	1.676	3.419	1.164	0.007
0423	$\theta_A, \theta_1, \theta_2, m_{21}, \gamma_1, \gamma_2$	1.710	1.358	1.593	0.000
0430	$\theta_A, \theta_1, \theta_2, m_{12}, m_{21}$	1.715	1.294	1.620	0.000
0113	$\theta_A, \theta_1 = \theta_2, m_{12}, \gamma_1, \gamma_2$	1.715	5.727	1.068	0.004
0411	$\theta_A, \theta_1, \theta_2, m_{12}, \gamma_1$	1.717	1.259	1.665	0.003
0422	$\theta_A, \theta_1, \theta_2, m_{21}, \gamma_2$	1.759	1.417	1.614	0.000
1401	$\tau, \theta_A, \theta_1, \theta_2, \gamma_1$	1.781	1.835	1.505	0.001
0433	$\theta_A, \theta_1, \theta_2, m_{12}, m_{21}, \gamma_1, \gamma_2$	1.843	1.773	1.597	0.000
0021	$\theta_A = \theta_1 = \theta_2, m_{21}, \gamma_1$	1.867	4.813	0.673	0.014
0221	$\theta_A = \theta_2, \theta_1, m_{01}, \gamma_1$	1.934	6.915	0.937	0.006
1400	$\tau, \theta_A, \theta_1, \theta_2$	2.098	1.697	1.899	0.000
0232	$\theta_A = \theta_2, \theta_1, m_{12}, m_{01}, \gamma_2$	2.186	7.859	1.121	0.007
[illegible]	$\theta_A = \theta_1, \theta_2, m_{01}, \gamma_2$	2.356	7.532	1.254	0.006
1122	$\tau, \theta_A = \theta_1, \theta_2, m_{01}, \gamma_2$	2.551	8.798	1.283	0.003
1133	$\tau, \theta_A = \theta_1, \theta_2, m_{12}, m_{21}, \gamma_1, \gamma_2$	2.748	12.927	0.814	0.008
1410	$\tau, \theta_A, \theta_1, \theta_2, m_{12}$	2.790	7.890	1.673	0.003

- arbitrary/generic models

Model choice vity

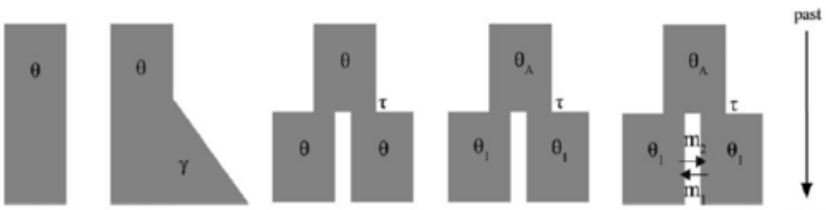
Table 3 *Continued*

Model	Parameters	Mean	SD	Median	Posterior probability
1420	$\tau, \theta_A, \theta_1, \theta_2, m_{21}$	2.819	9.142	1.557	0.001
0330	$\theta_A, \theta_1 = \theta_2, m_{12}, m_{21}$	3.156	11.980	1.608	0.000
0431	$\theta_A, \theta_1, \theta_2, m_{12}, m_{21}, \gamma_1$	3.388	12.338	1.687	0.001
0432	$\theta_A, \theta_1, \theta_2, m_{12}, m_{21}, \gamma_2$	3.769	15.818	1.606	0.003
1210	$\tau, \theta_A = \theta_2, \theta_1, m_{12}$	4.007	21.699	0.880	0.010
0310	$\theta_A, \theta_1 = \theta_2, m_{12}$	4.405	20.648	1.670	0.001
0421	$\theta_A, \theta_1, \theta_2, m_{21}, \gamma_1$	4.761	18.586	1.563	0.000
1223	$\tau, \theta_A = \theta_2, \theta_1, m_{21}, \gamma_1, \gamma_2$	4.813	27.942	0.880	0.007
0410	$\theta_A, \theta_1, \theta_2, m_{12}$	4.840	19.483	1.684	0.000
0333	$\theta_A, \theta_1 = \theta_2, m_{12}, m_{21}, \gamma_1, \gamma_2$	4.841	24.764	1.304	0.004
1411	$\tau, \theta_A, \theta_1, \theta_2, m_{12}, \gamma_1$	4.949	22.725	1.182	0.000
0320	$\theta_A, \theta_1 = \theta_2, m_{21}$	5.184	25.275	1.771	0.000
1431	$\tau, \theta_A, \theta_1, \theta_2, m_{12}, m_{21}, \gamma_1$	5.539	28.987	1.440	0.000
1421	$\tau, \theta_A, \theta_1, \theta_2, m_{21}, \gamma_1$	5.618	22.805	1.418	0.001
1311	$\tau, \theta_A, \theta_1 = \theta_2, m_{12}, \gamma_1$	5.721	32.177	1.137	0.001
0111	$\theta_A = \theta_1, \theta_2, m_{12}, \gamma_1$	5.804	32.950	1.143	0.008
0420	$\theta_A, \theta_1, \theta_2, m_{21}$	6.037	28.946	1.629	0.001
1412	$\tau, \theta_A, \theta_1, \theta_2, m_{12}, \gamma_2$	6.186	23.177	1.611	0.003
0010	$\theta_A = \theta_1 = \theta_2, m_{12}$	6.223	36.293	0.000	0.017
1413	$\tau, \theta_A, \theta_1, \theta_2, m_{12}, \gamma_1, \gamma_2$	8.209	48.083	1.344	0.000
1430	$\tau, \theta_A, \theta_1, \theta_2, m_{12}, m_{21}$	8.661	50.499	1.516	0.001
1422	$\tau, \theta_A, \theta_1, \theta_2, m_{21}, \gamma_2$	9.269	45.089	1.344	0.006
0121	$\theta_A = \theta_1, \theta_2, m_{21}, \gamma_1$	9.369	56.607	1.327	0.004
1302	$\tau, \theta_A, \theta_1 = \theta_2, \gamma_2$	9.386	44.243	1.233	0.004
0120	$\theta_A = \theta_1, \theta_2, m_{21}$	9.466	57.924	1.189	0.004
1310	$\tau, \theta_A, \theta_1 = \theta_2, m_{12}$	9.812	60.333	1.206	0.000
1100	$\tau, \theta_A = \theta_1, \theta_2$	10.795	68.438	1.121	0.007
0332	$\theta_A, \theta_1 = \theta_2, m_{12}, m_{21}, \gamma_2$	13.053	82.999	1.415	0.004
1120	$\tau, \theta_A = \theta_1, \theta_2, m_{21}$	[illegible]	[illegible]	[illegible]	[illegible]
X0X1*	θ_A, γ_1				
X0X0*	θ_A				
0000	$\theta_A = \theta_1 = \theta_2$				

The answer is model 1023!

For each model: τθmγ

Divergence time (τ)	Theta (θ)	Migration (m)	Population expansion (γ)
0: island model 1: divergence at time (τ) X: pamixia Prior: 0.001–5 (4N generations)	0: $\theta_A = \theta_1 = \theta_2$ 1: $\theta_A = \theta_1, \theta_2$ 2: $\theta_A = \theta_2, \theta_1$ 3: $\theta_A, \theta_1 = \theta_2$ 4: $\theta_A, \theta_1, \theta_2$ Prior: 0.01–10 per locus	0: no migration 1: m_{12} 2: m_{21} 3: m_{12}, m_{21} X: na/pamixia Prior: 0–5 migrants per generation	0: no expansion 1: γ_1 2: γ_2 3: γ_1, γ_2 Prior: 0.1–9 (exponential)



- arbitrary/generic models

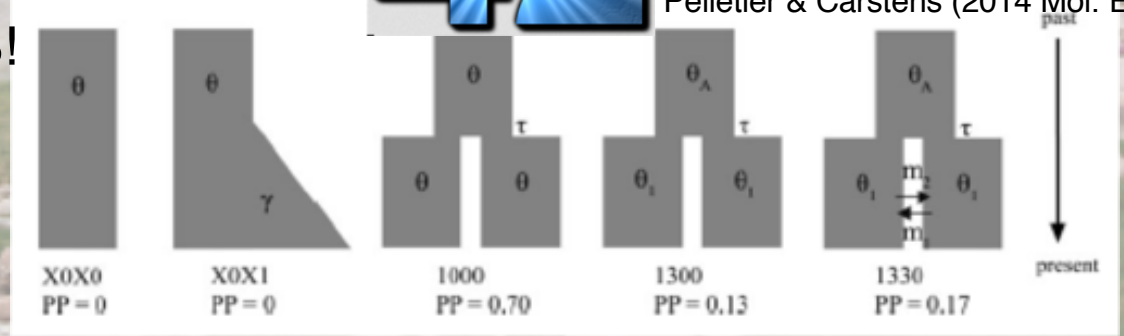
Biological insights depend on the questions we (the scientist) ask!

- We should expect (or want) or computer programs to define the questions we ask!

The answer is: 42

The answer is model 1023!

Pelletier & Carstens (2014 Mol. Ecol.)

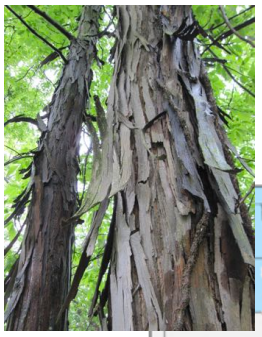


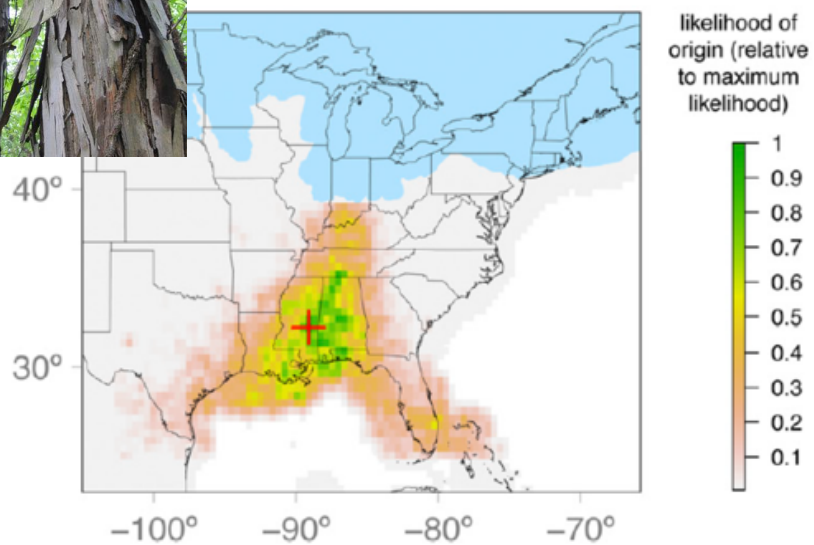
- PHRAPL can create hundreds of possible histories that have a mixture of gene flow, population subdivision, and/or population size differences and compare these models using AIC (O'Meara)

- Model formulation is a way of communicating our expert knowledge to statistical apparatus to test hypotheses

Geologic data indicate species were displaced by climate change and current distribution reflects recent expansion which can be tested genetically

ENMs do not provide precise location of Pleistocene refuge for hickory trees





Bemmels JB, Knowles LL, Dick CW (2019)

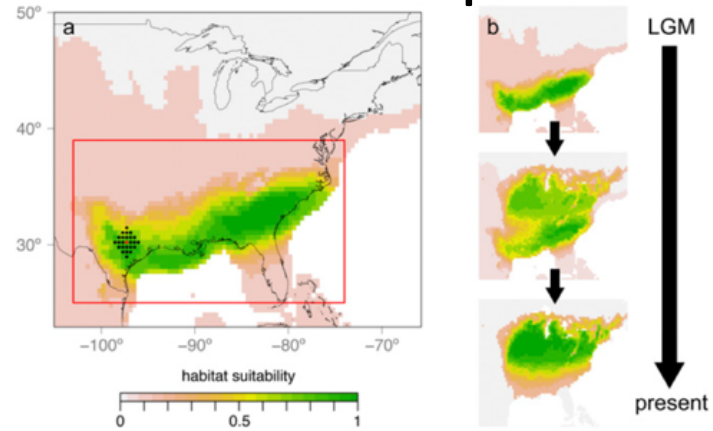
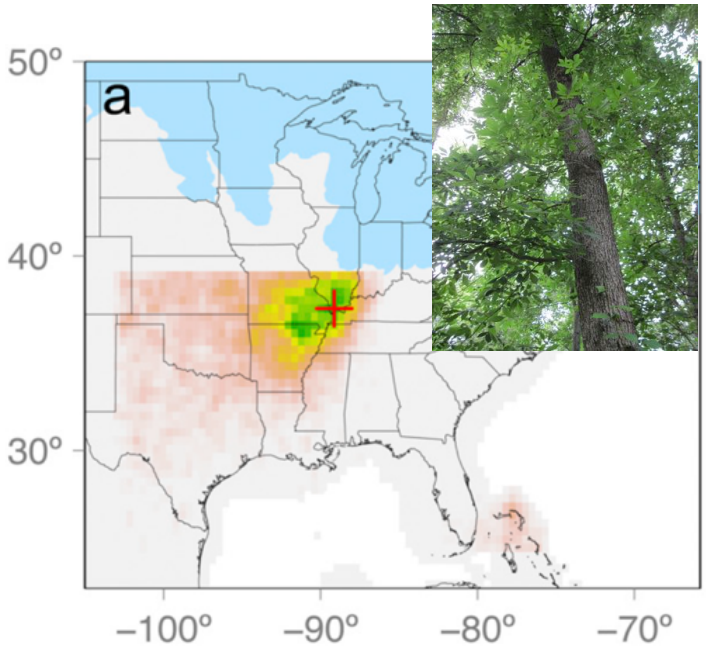


Fig. 1. Schematic overview of demographic simulations. (*A*) Simulations were initiated in the LGM landscape (shown here for *C. cordiformis*) from a central deme (see red dot as an example) plus an area extending three additional demes (black dots) in all directions. Different geographic sources of

Inferred geographic coordinates of source of expansion (based on allele frequency gradients), where the *geographic coordinate* is a parameter in the model (inferred using ABC; see *He et al.* 2017. Inferring the geographic origin of a range expansion: latitudinal and longitudinal coordinates inferred from genomic data in an ABC framework with the program X-ORIGIN. *Mol. Ecol.* 26:6908-6920. DOI: 10.1111/mec.14380

Species-specific differences in the location of refugial populations .



Bemmels et al. 2019 PNAS 116:8431-8436

Inferred likelihood of geographic coordinates of ancestral refugia population – this location corresponds to a macrofossil of the bitternut hickory

Bemmels JB, Knowles LL, Dick CW (2019) Genomic evidence of survival near ice sheet margins for some, but not all, North American trees. *PNAS* 116:8431-8436.

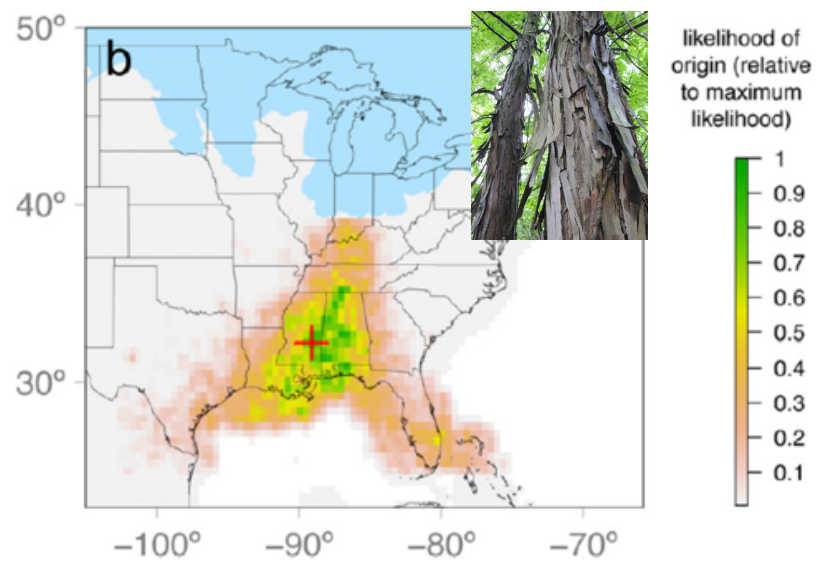


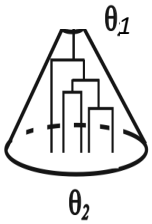
Fig. 2. Estimated expansion origins (Ω; red cross) in *C. cordiformis* (*A*) and *C. ovata* (*B*). The shading of pixels depicts a probability surface (kernel density) showing the likelihood that each pixel served as the expansion origin relative to the pixel with the highest likelihood (i.e., Ω). Glaciated regions are shown in blue. The results presented in *A* and *B* are based on retention of four and three PC axes of variation in genetic summary statistics, respectively. Results based on retaining additional PC axes are presented in *SI Appendix*, Figs. S2 and S3.

Transformative potential of model-based analyses in evolutionary biology

- Accounting for species-specific differences
- Spatially explicit coalescent models
- Comparative phylogeographic analyses

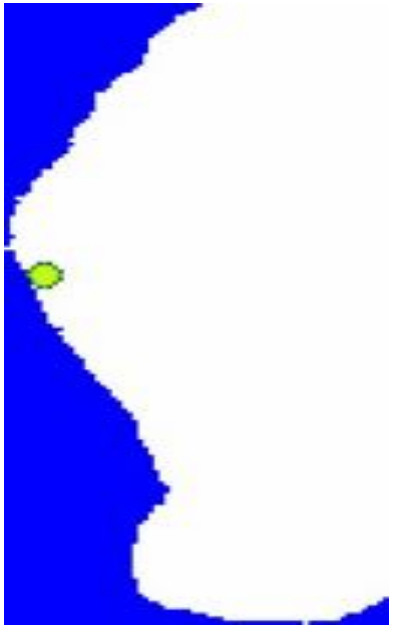
All models are simplifications, but they vary in their relative degree of abstraction

Different ways to model population expansion:

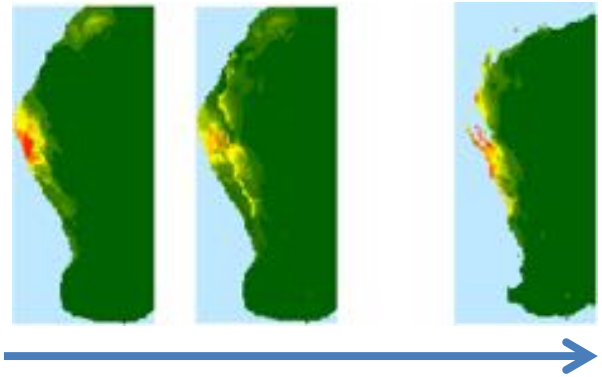


(i) Model as population size change with no spatial aspect of expansion (e.g., Brazilian Atlantic forest areas of instability associated with recent expansion; Carnaval et al. 2009)

(ii) Model expansion process across landscape explicitly (He et al. 2013; Evolution)



Dynamic ENM



Past → Present

- Start from LGM refugia
- Colonize with changing layers of ENM

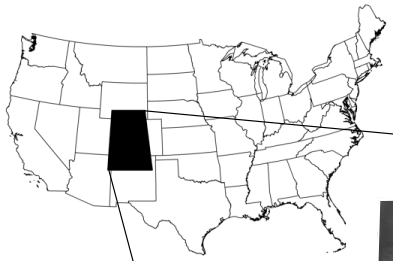
Does microhabitat differences affect species responses to climate change?

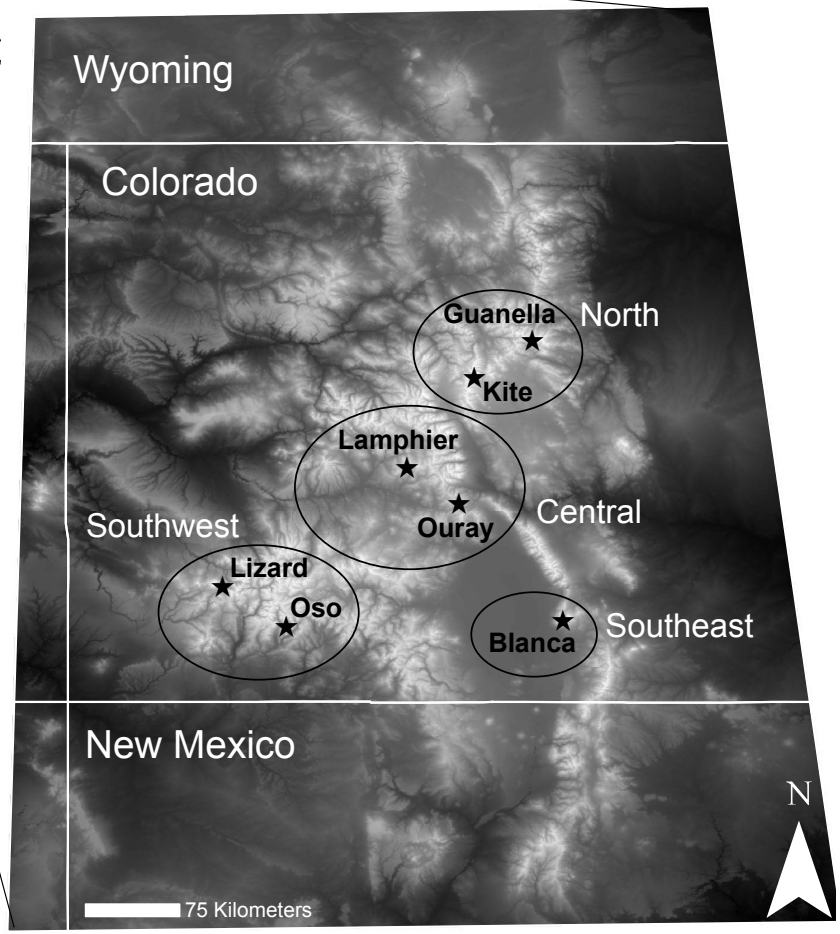
- start with descriptive analysis to explore hypotheses
- follow-up with spatially explicit models to test hypotheses about why patterns of genetic variation differ among species (i.e., generate species-specific patterns of genetic variation)

Sky island community responses to climate change similarly (based on patterns of genetic differentiation)

Carex chalciolepis

Carex nova





Massatti & Knowles
(2014 Evolution)

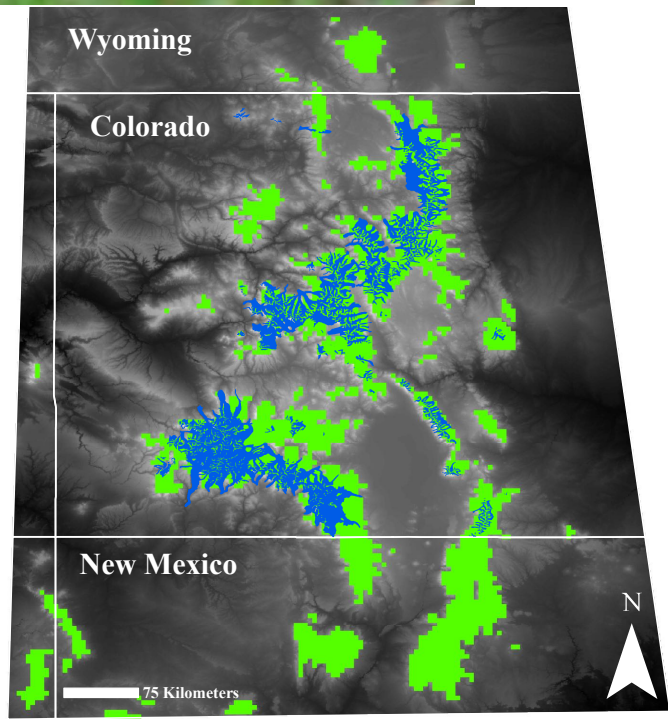
Sky island communities: responses to climate change

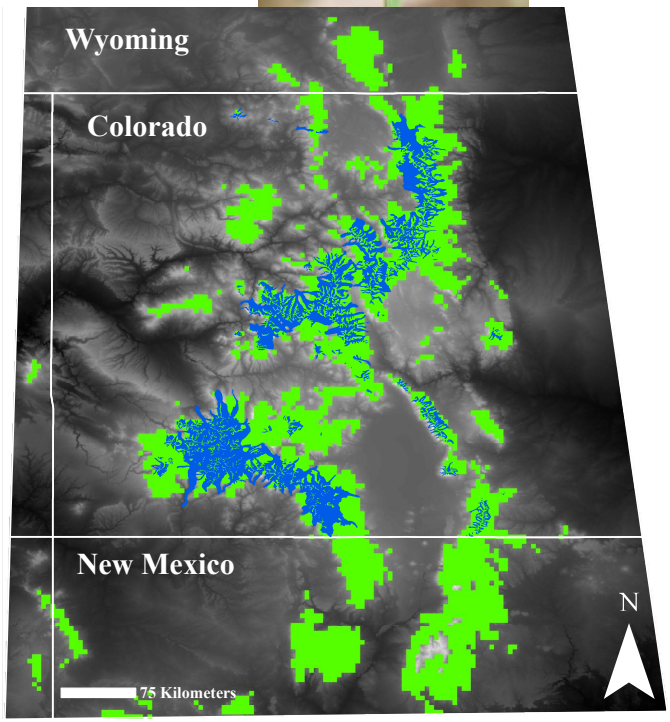
- co-distributed, abundant taxa with similar natural histories and dispersal abilities

Carex chalciolepis

C. nova

- so similar that ENMs project very similar past distributions



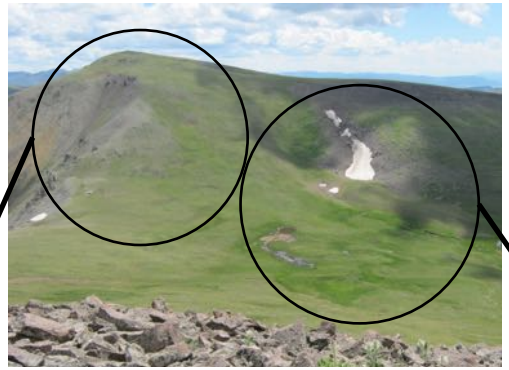


• taxa differ in microhabitats

inhabits slopes and ridges

restricted to wetlands a

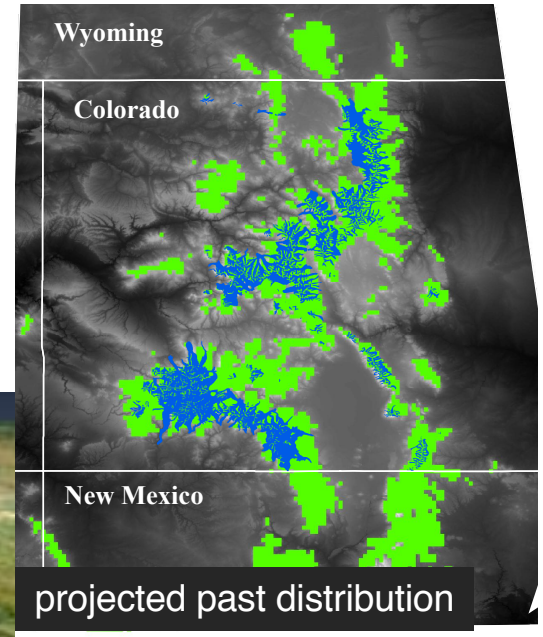
Carex chalciolepis



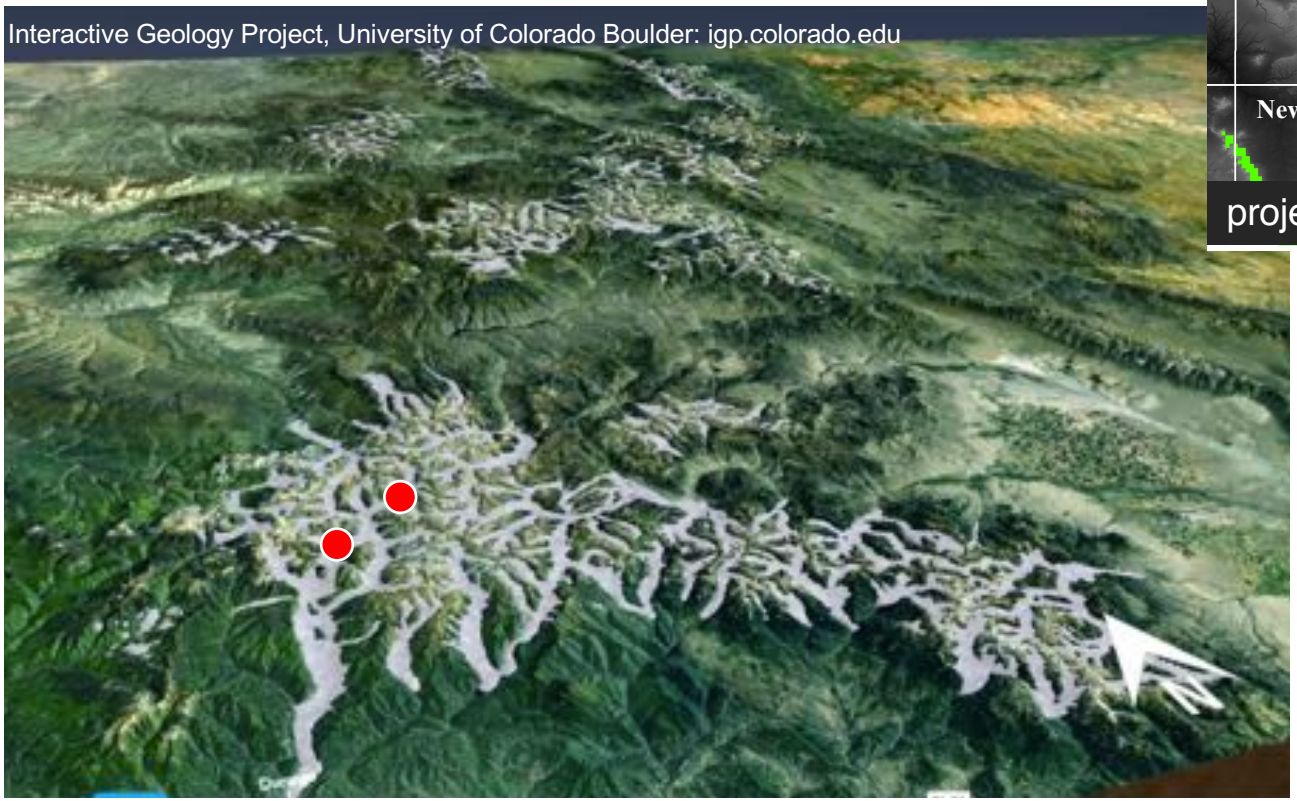
Carex nova

Given that ecological niche models (ENMs) are similar between species (both present and during LGM)…

why would we predict discord in patterns of genetic variation between the plant species?



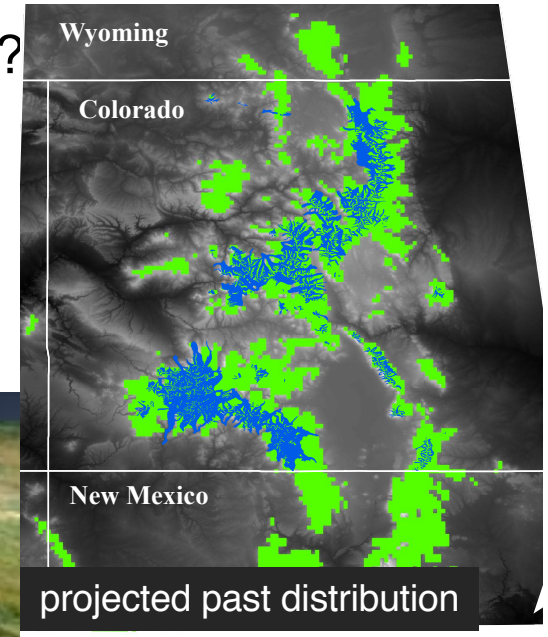
projected past distribution

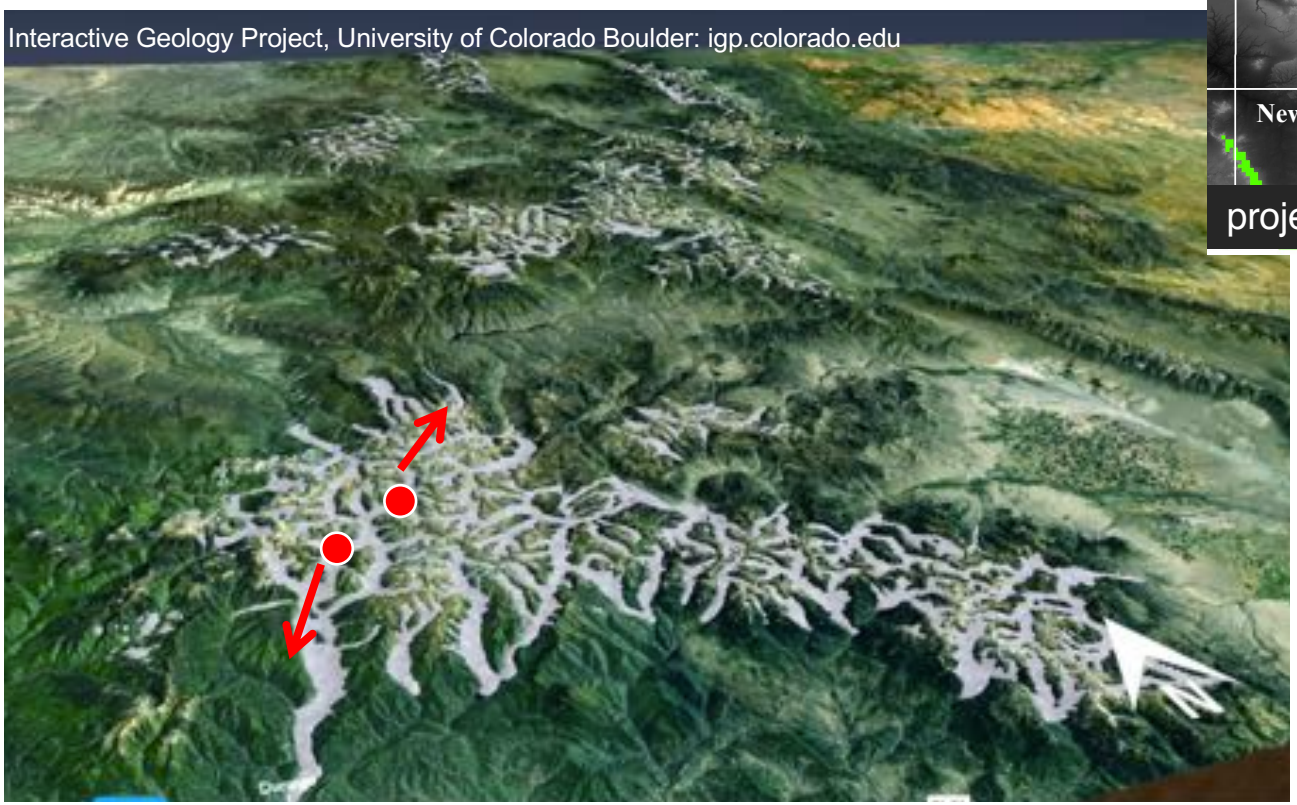


If microhabitat matters…

- glaciers in drainages would have displaced populations of wetland specialist

Why should microhabitat matter for sky island inhabitants?





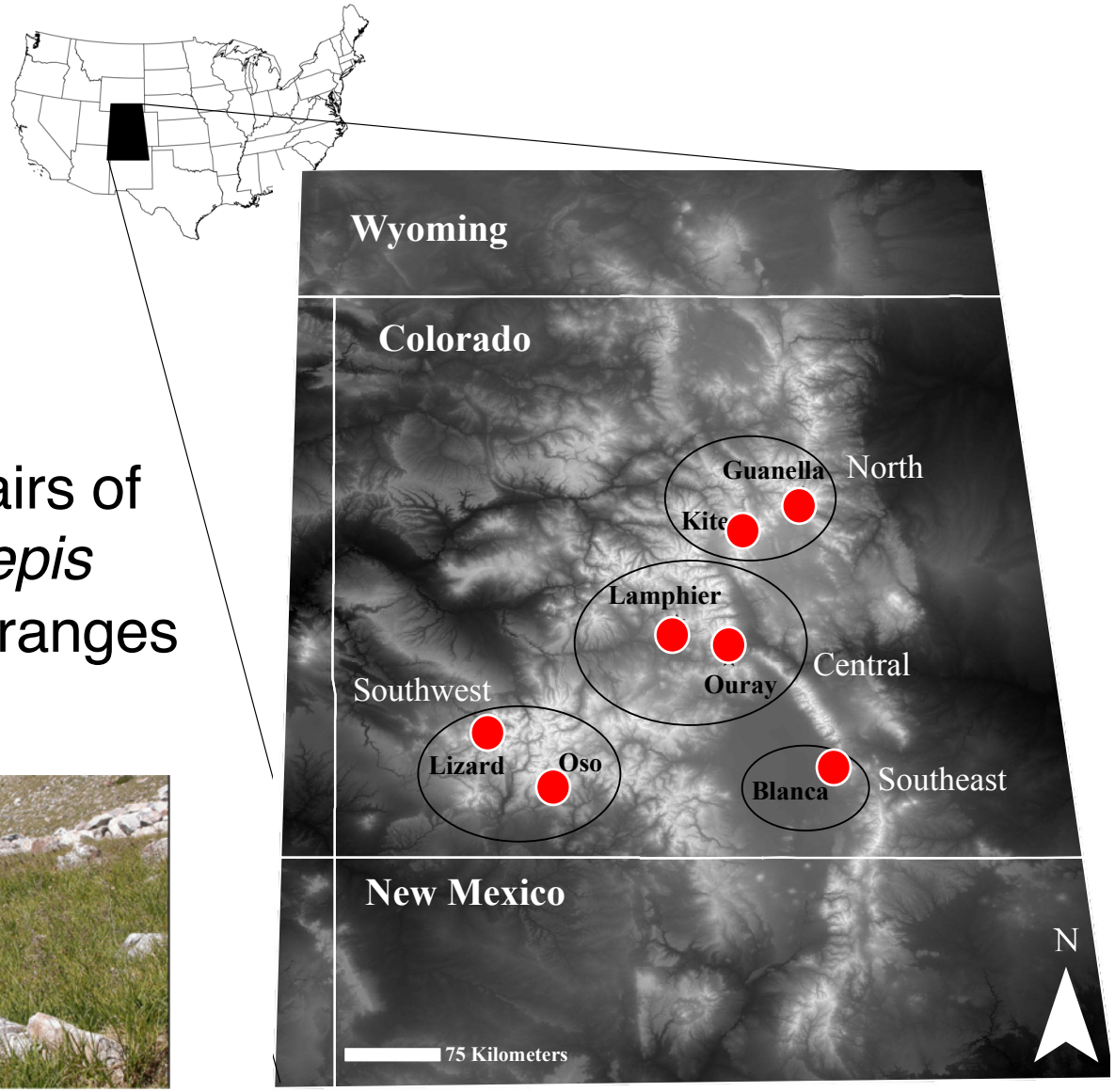
If microhabitat matters…

- distances separating populations may have been considerable greater in the past – *but only in the wetland specialist*

1. Sky island communities: microhabitat differences and responses to climate change

- SNPs from over 22,000 loci (RADseq)

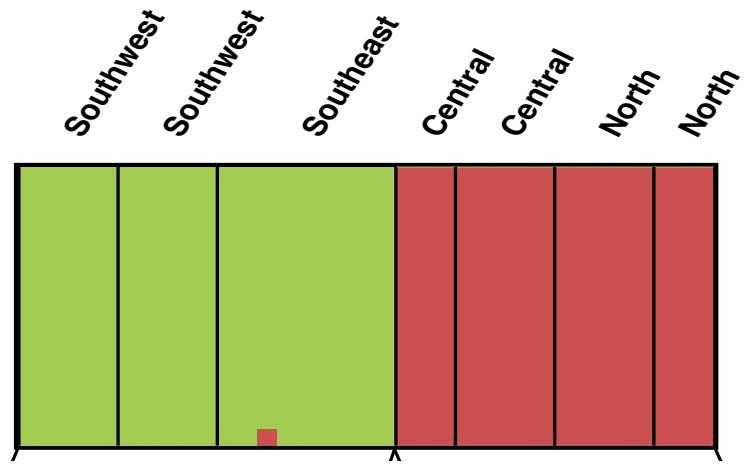
- sampled population pairs of *C. nova* and *C. chalciolepis* from different mountain ranges

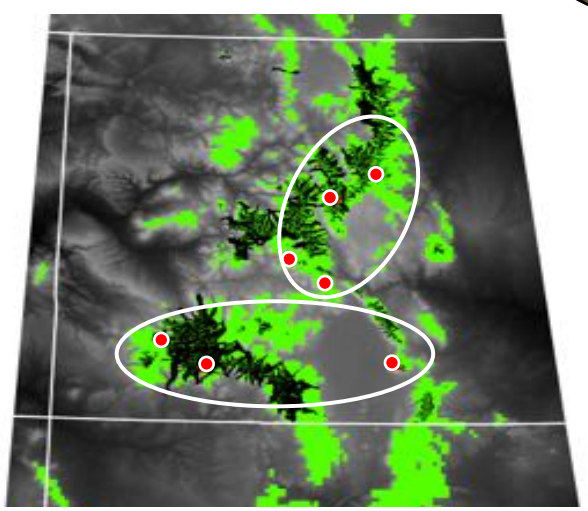


Massatti and Knowles, Evolution (in press)

C. nova

restricted to wetlands



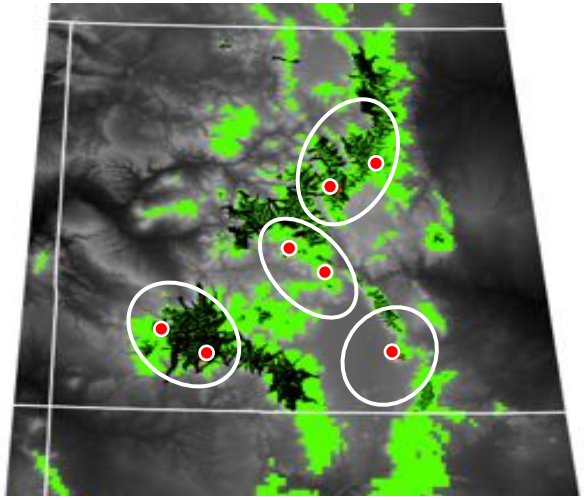


projected past distribution

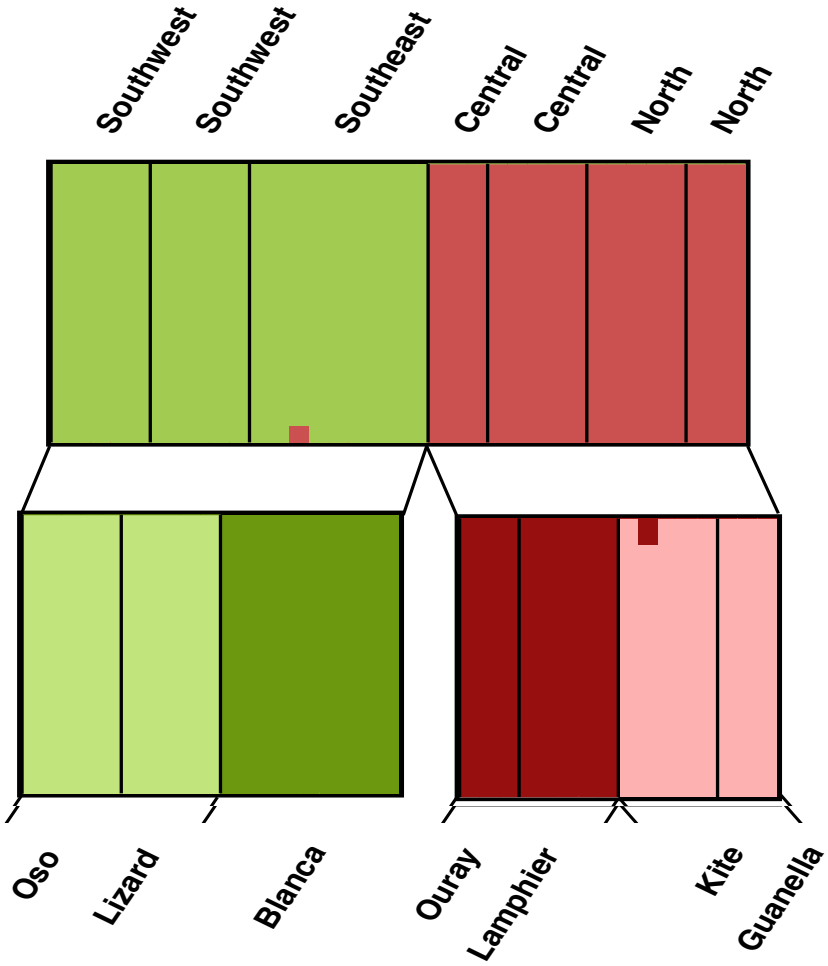
- Structure analysis of SNPs from over 22,000 loci

C. nova

restricted to wetlands



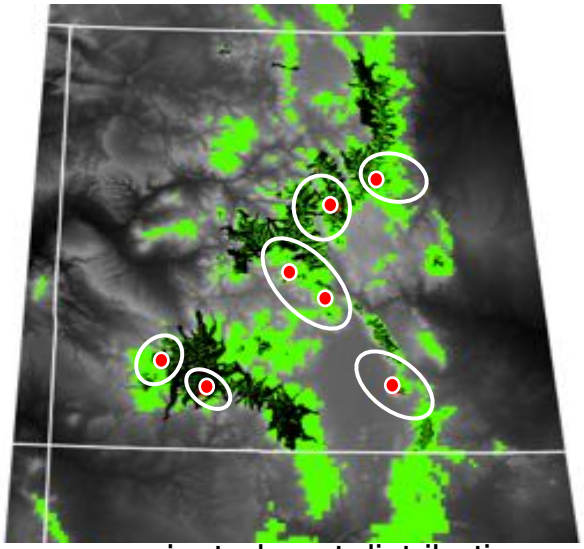
projected past distribution



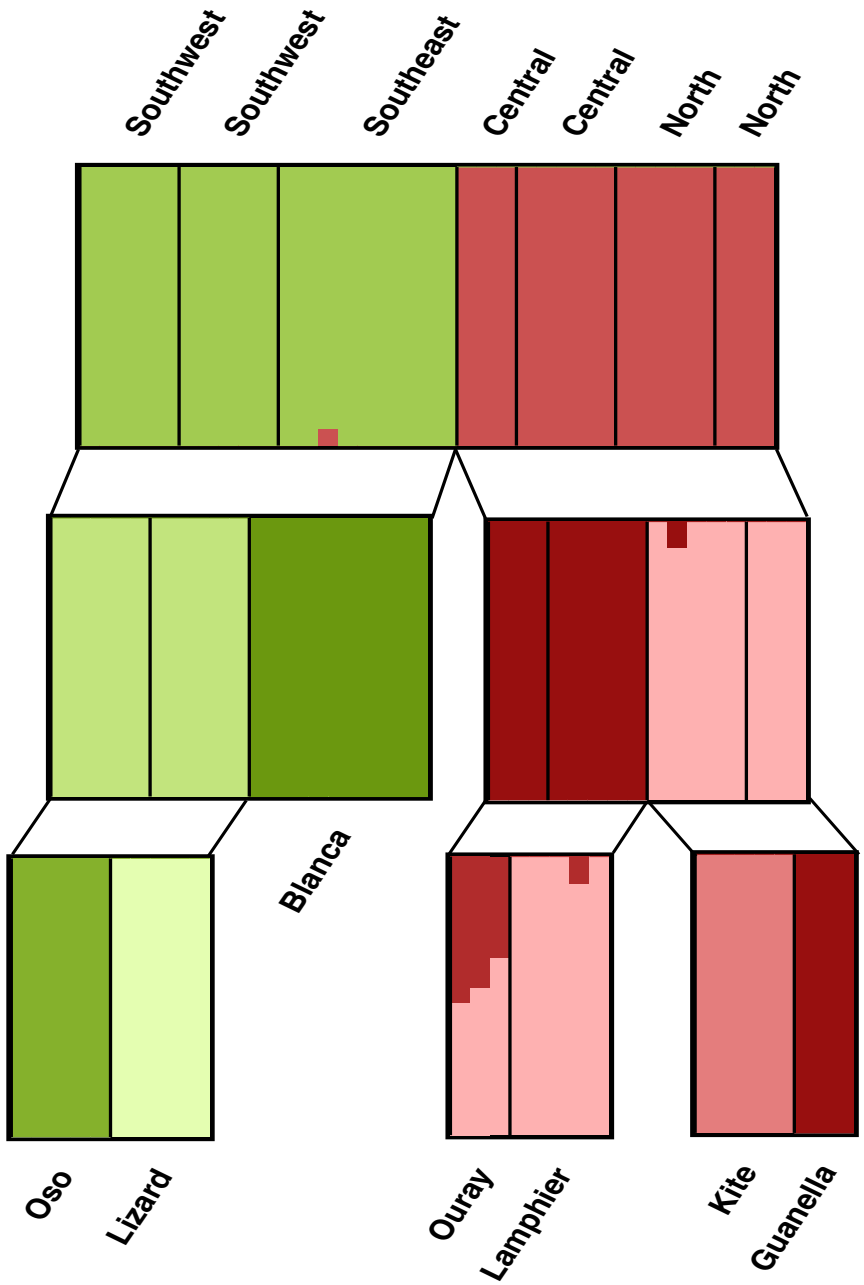
- Structure analysis of SNPs from over 22,000 loci

restricted to wetlands

C. nova



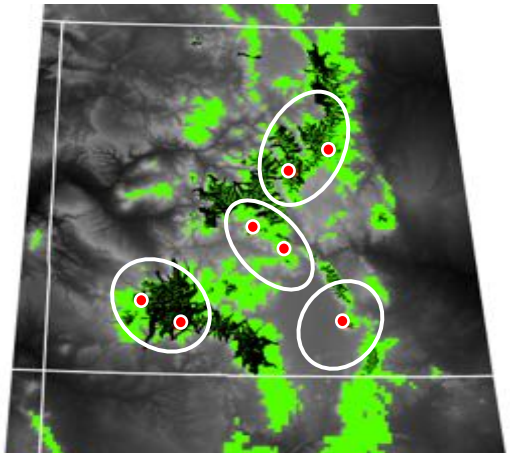
projected past distribution



- STRUCTURE analysis of SNPs from over 22,000 loci

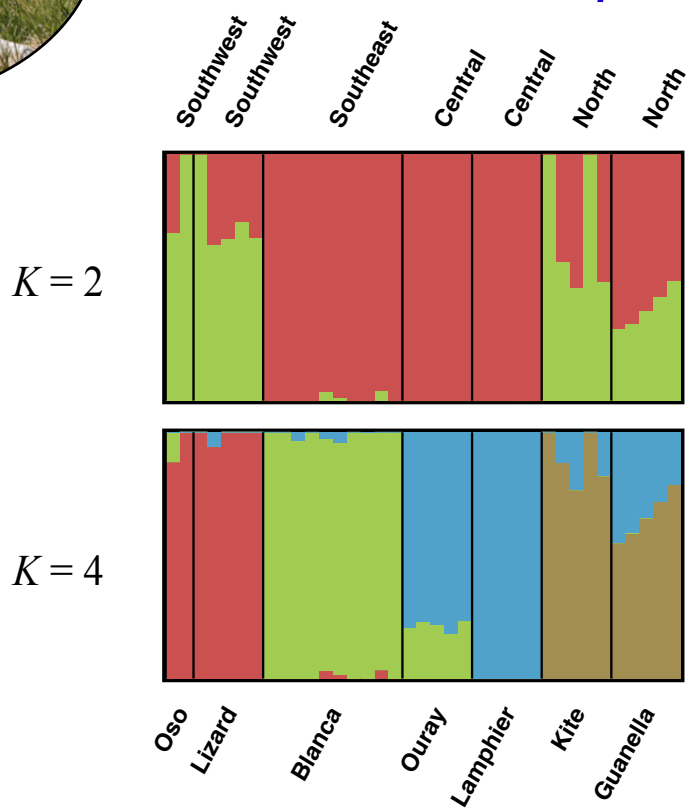
Massatti and Knowles, Evolution (in press)

inhabits slopes and ridges

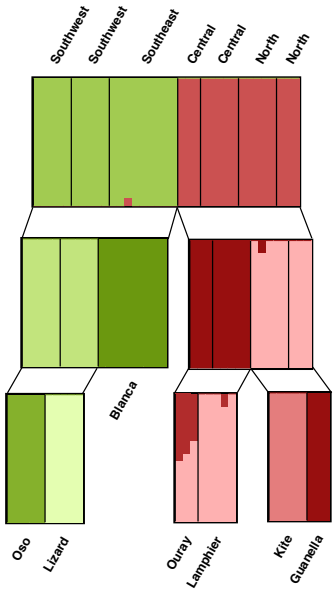


projected past distribution

C. chalciolepis



C. nova



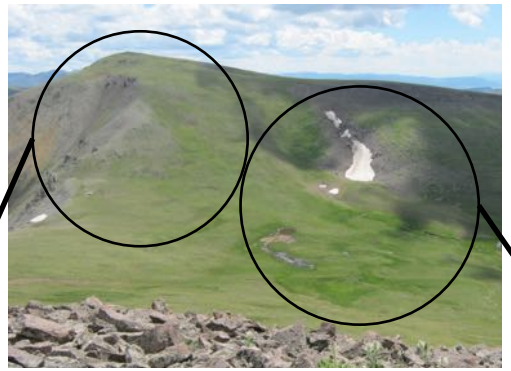
Massatti and Knowles, Evolution (in press)

Genomic patterns support predictions of an interaction between microhabitat affinity and climate change

(glaciers are barrier for movement of wetland specialists only)

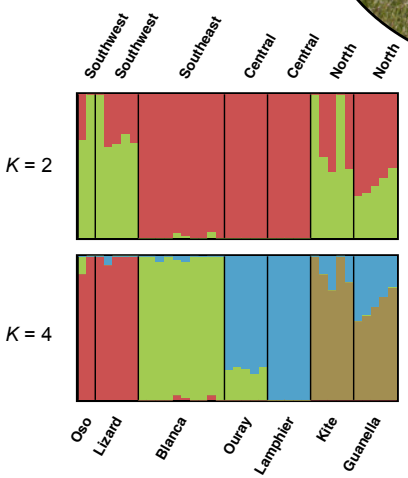
Carex chalciolepis

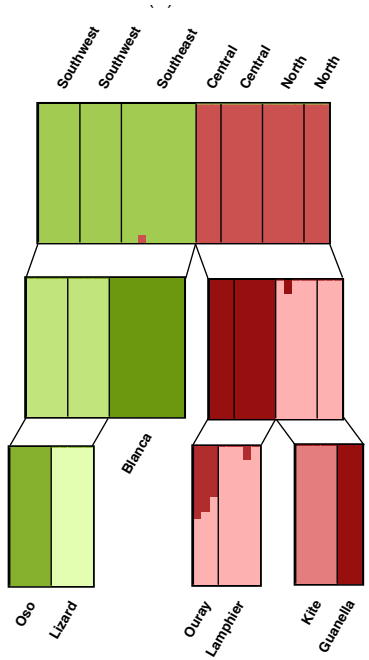
dry ridges



Carex nova

wetland specialists



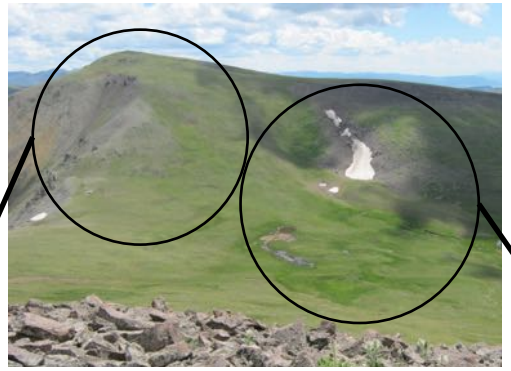


Genomic patterns support prediction of an interaction between microhabitat affinity and climate change

Massatti & Knowles (2014) *Evolution*

Carex chalciolepis

dry ridges



Carex nova

wetland specialists

Test if observed discordant phylogeographic structure could be caused by differences in microhabitat affinity ….

- generate species-specific expectations for patterns of genetic variation

(i.e., glaciers are barrier for movement of wetland specialists only)