

# **Species tree estimation & the multispecies coalescent**

# Concatenated gene sequences - assumes every gene has same evolutionary history

*Syst. Biol.* 56(1):17–24, 2007

Copyright © Society of Systematic Biologists

ISSN: 1063-5157 print / 1076-836X online

DOI: 10.1080/10635150601146041

## **Inconsistency of Phylogenetic Estimates from Concatenated Data under Coalescence**

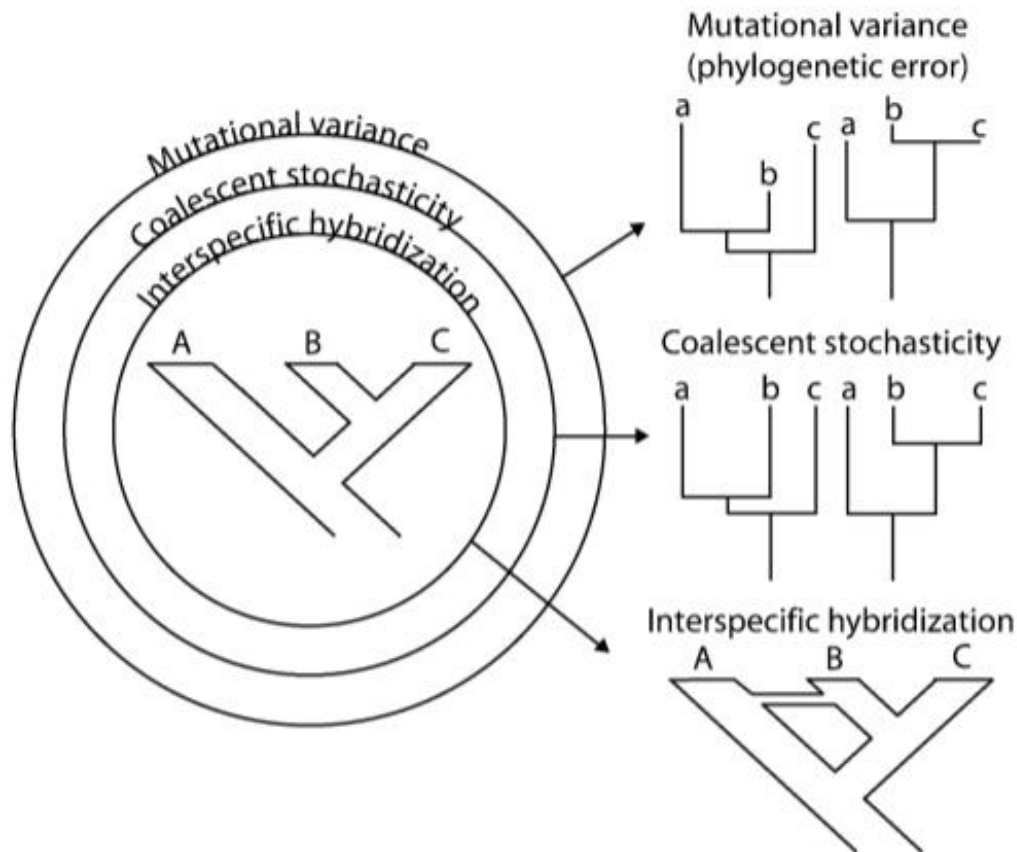
LAURA SALTER KUBATKO<sup>1</sup> AND JAMES H. DEGNAN<sup>2</sup>

<sup>1</sup>*Departments of Statistics and Evolution, Ecology, and Organismal Biology, The Ohio State University, Columbus, Ohio 43210, USA;*

*E-mail: lkubatko@stat.ohio-state.edu*

<sup>2</sup>*Department of Biostatistics, Harvard School of Public Health, Building 2, 4th Floor, 655 Huntington Avenue, Boston, Massachusetts 02115, USA*

# Stochasticity from:

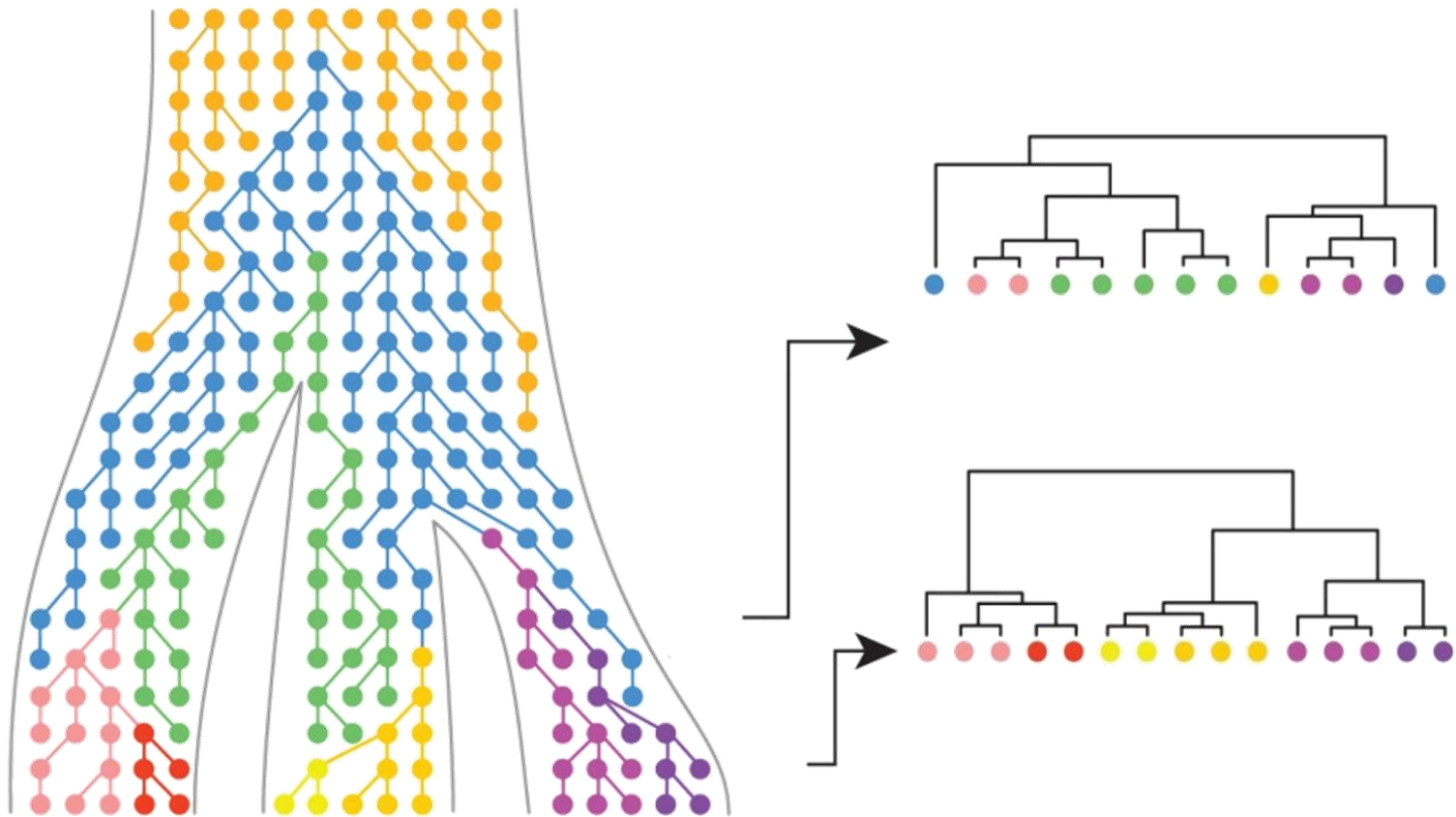


## Distinguishing between sources

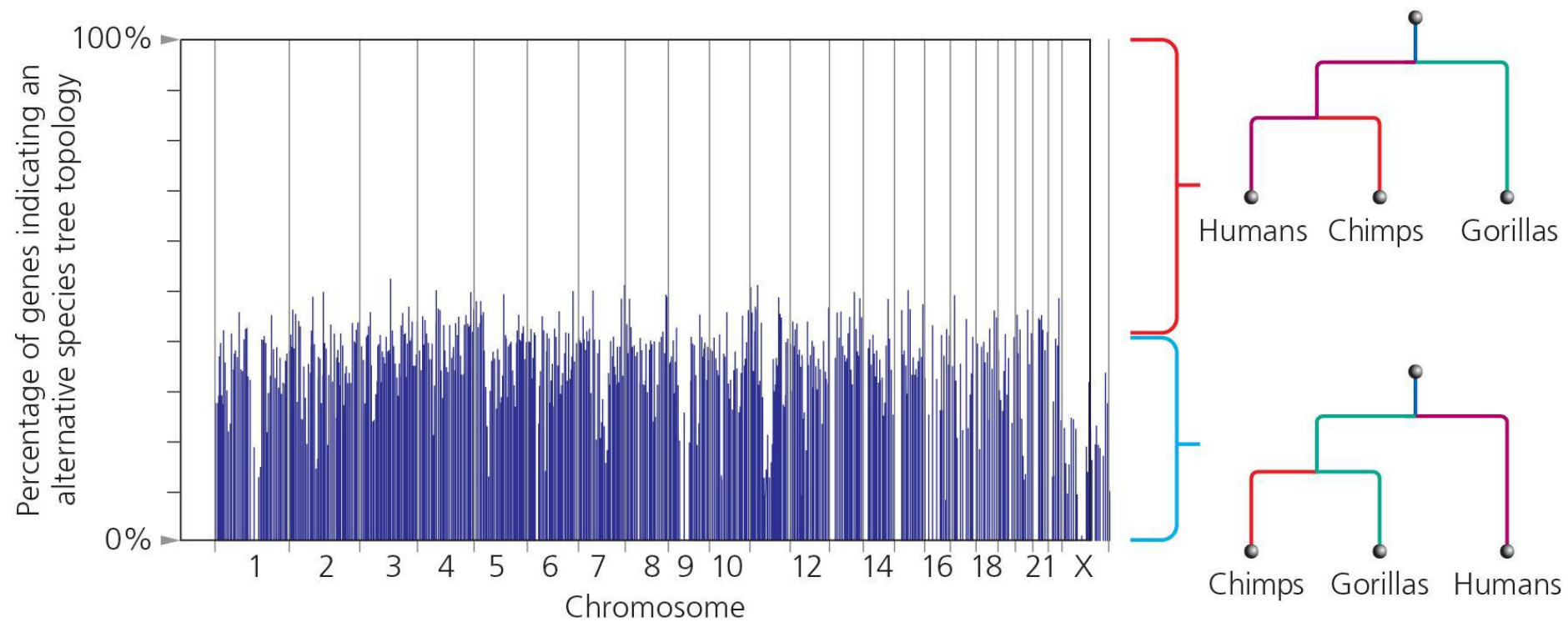
Test for mutational variance as a source of mtDNA discordance by performing a parametric bootstrap using the independently estimated species tree as the constraint tree.

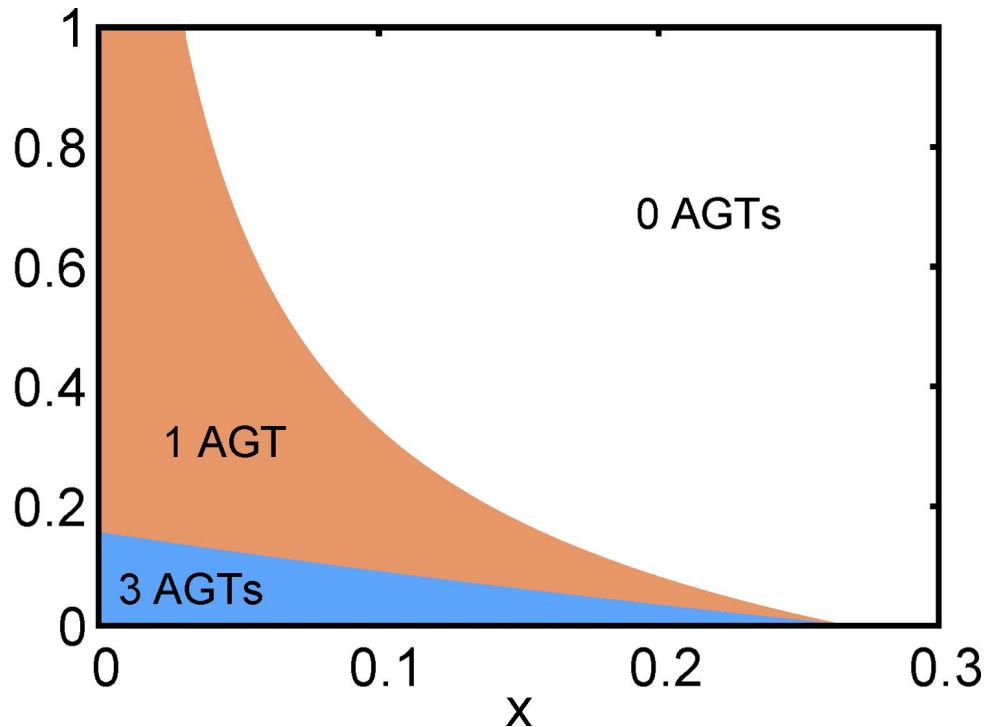
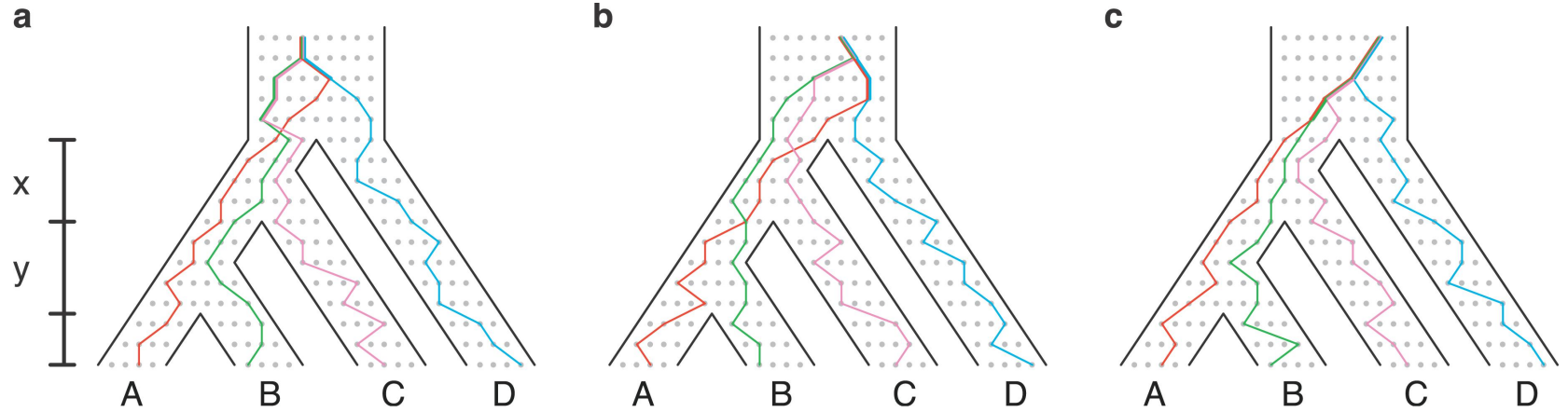
If mutational variance is unlikely, test for coalescent stochasticity as a source of mtDNA discordance by simulating genealogies on an independent estimate of the species tree.

If both mutational and coalescent sources of discordance are rejected, hybridization is left as a likely explanation.



**Incomplete lineage sorting**

**A**



**The “Anomaly zone”**

**AGT - Anomalous  
gene tree with higher  
likelihood than the  
tree species topology**

**ILS will result when branches are  
short & population sizes are  
large**

**(often mistakenly thought only to occur in  
recent radiations, but ancient short internal  
branches just as at risk!)**

# The Coalescent

Coalescence - MRCA of a pair of genes

Built on standard population genetics (e.g. Wright-Fisher model)

Key parameter - Effective population size

Expected # of generations to coalescence -  $2N$



# The Coalescent

What is the probability two copies of a gene in a randomly breeding population have an ancestor 1 generation ago?

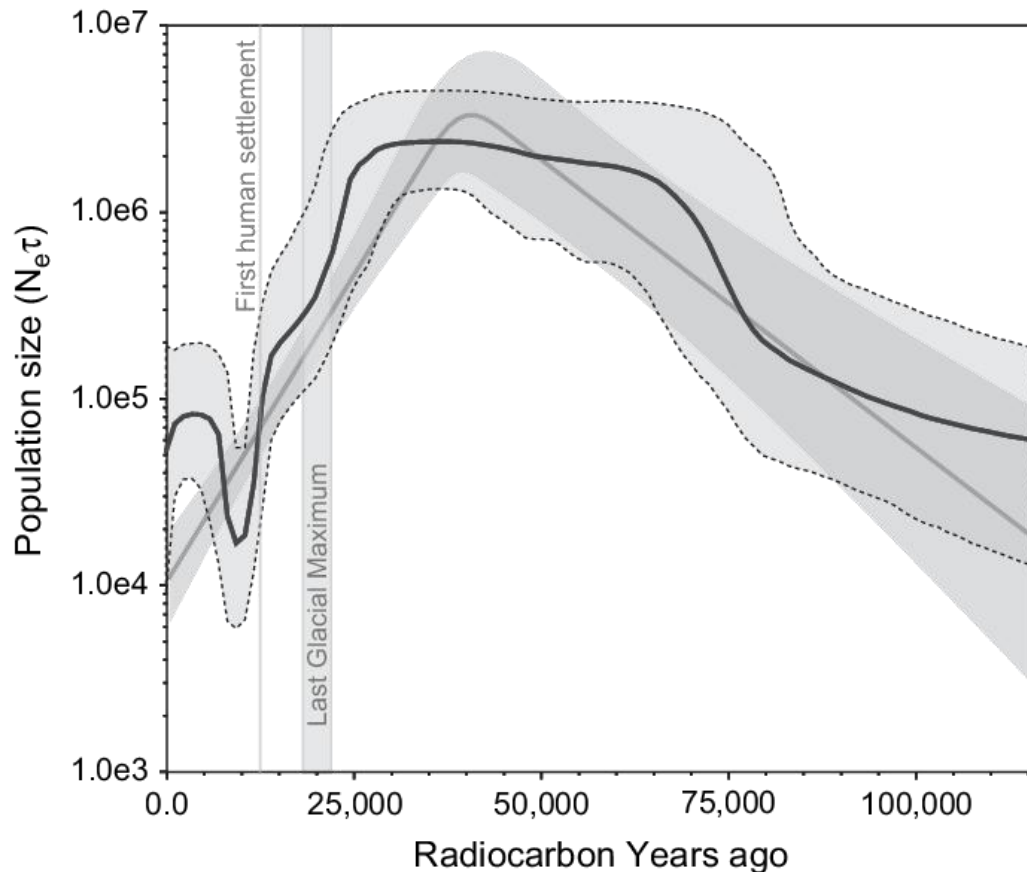
$$1/(2N_e)$$

What about  $j$  generations ago?

$$(1 - 1/(2N_e))^j * 1/(2N_e)$$

# The Coalescent

Coalescence-time measured in units of  $N_e$



**Bayesian Coalescent Inference of Past Population Dynamics from Molecular Sequences**

*A. J. Drummond, A. Rambaut, B. Shapiro, and O. G. Pybus*

Department of Zoology, University of Oxford, Oxford, United Kingdom

# The Multispecies Coalescent

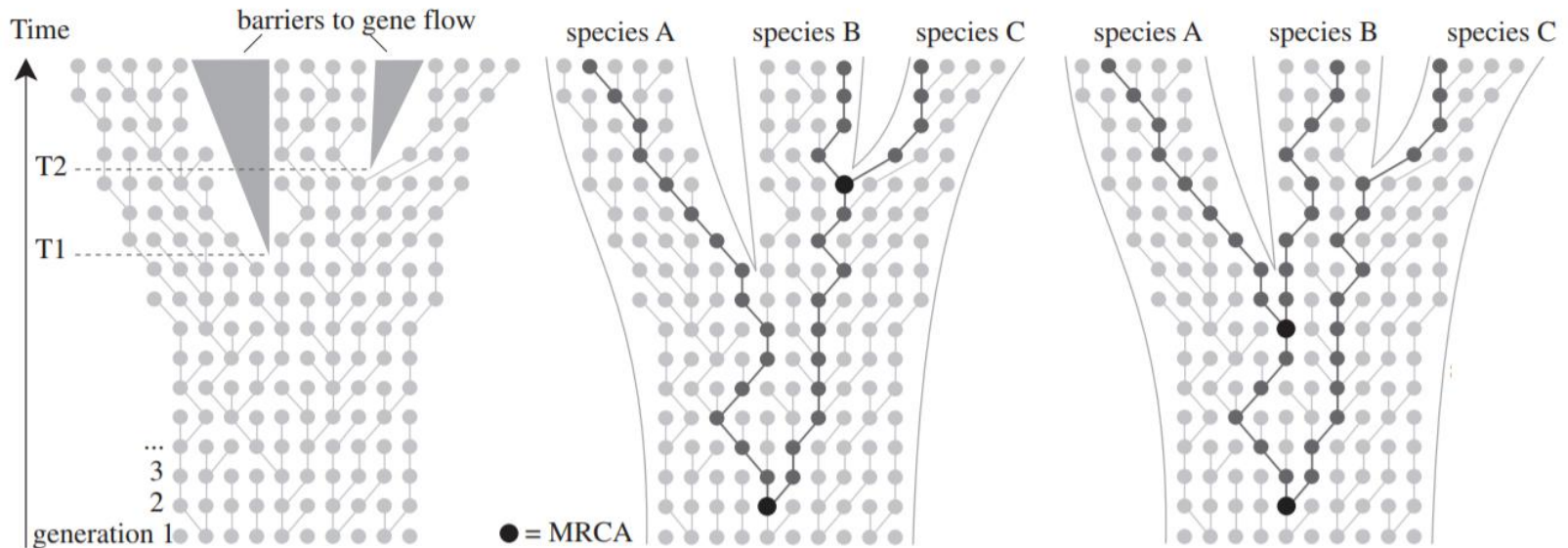
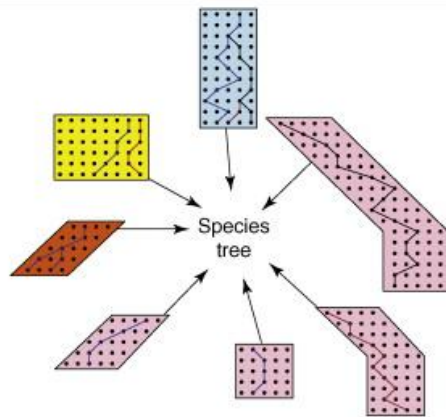
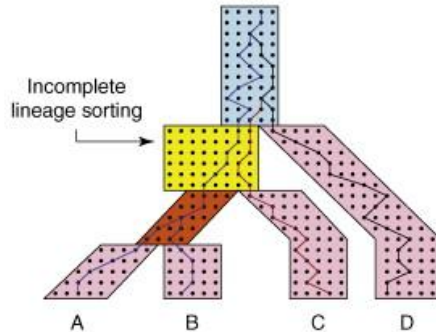


Image: Leliaert, F., Verbruggen, H., Vanormelingen, P., Steen, F., López-Bautista, J. M., Zuccarello, G. C., & De Clerck, O. (2014). DNA-based species delimitation in algae. *European journal of phycology*, 49(2), 179-196.

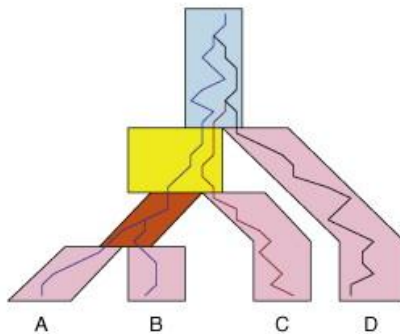
(a)



(b)



(c)



$P(D_i | Q_i, \pi_i, \boldsymbol{\psi}_i)$  = standard likelihood of gene tree

$P(\boldsymbol{\psi}_i | S)$  = Likelihood of gene tree given the species tree

“AND” rule:

$$P(D_1, D_2 \dots D_n | S) = P(D_1 | Q_1, \pi_1, \boldsymbol{\psi}_1) * P(\boldsymbol{\psi}_1 | S) \times \dots \times P(D_n | Q_n, \pi_n, \boldsymbol{\psi}_n) * P(\boldsymbol{\psi}_n | S)$$

# Methods & software

Parsimony - “MDC” species tree that minimizes deep coalescences (can be inconsistent estimator)

ML - STEM (Kubatko & Degnan 2007). Requires gene trees to be well-estimated and clock-like

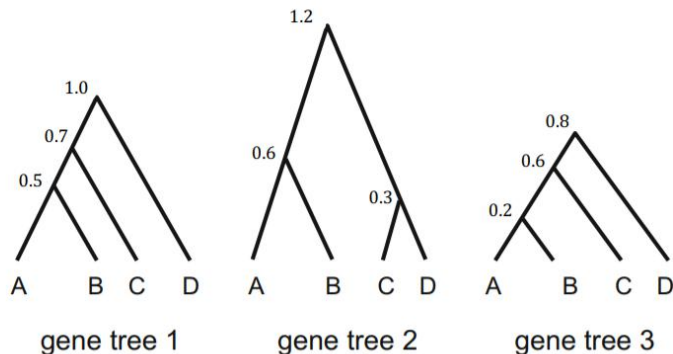
Bayesian - BEST, \*BEAST, BPP. Bayesian approaches that integrate over uncertainty in gene trees. Great models...but complex and hard to converge!

$$P(S | D) \propto \int_G \left( \prod_{i=1} P(d_i | g_i) P(g_i | S) \right) P(S) dG,$$

# Other methods

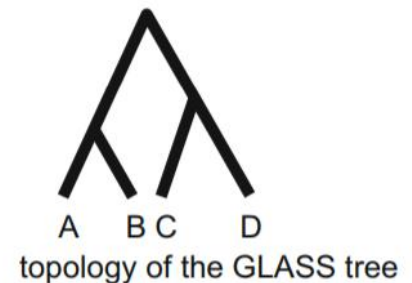
**Concordance analysis- BCA/BUCKy. Semi-parametric clustering of gene trees into “concordance blocks” without regard to process**

**Summary methods- Uses properties of multispecies coalescent to summarize gene trees.  
STAR/STEAC/GLASS**



(b)

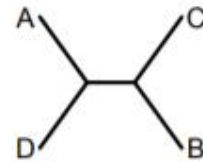
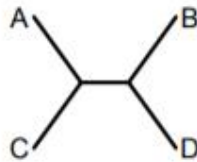
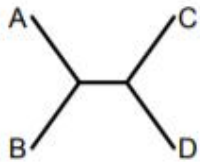
	A	B	C	D
A	--	0.2	0.6	0.8
B	0.2	--	0.6	0.8
C	0.6	0.6	--	0.3
D	0.8	0.8	0.3	--



# Other methods

## Quartets approaches: ASTRAL/SVDQuartets

Avoids calculating full likelihood, instead focuses on site patterns over 4 taxon combinations. Good for SNPs and genomic scale data





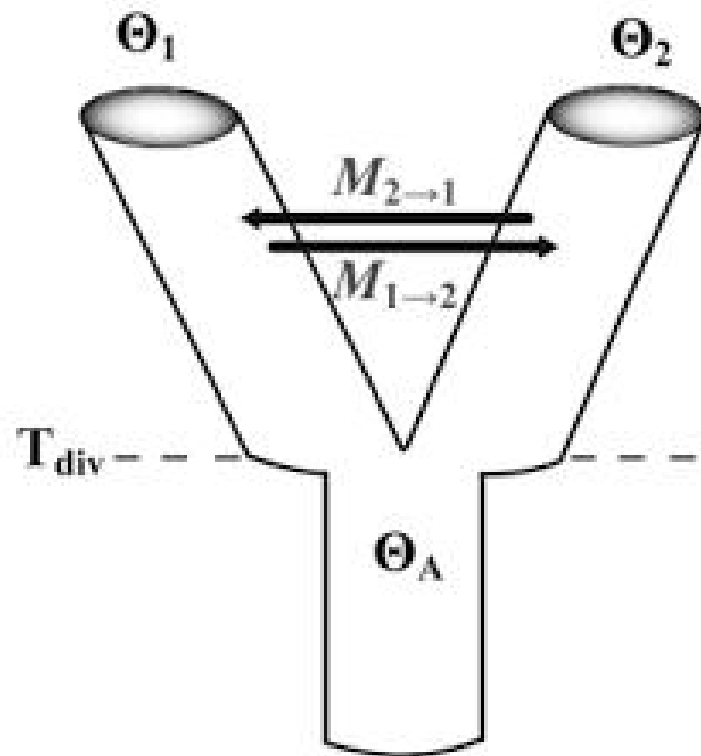
## Species Tree Inference Summary – Comparison of Methods

Software	Data Type	Measure of Uncertainty	Computation Time	Models Included
BEST	multilocus	posterior probability	long; can be run in parallel	coalescent; all reversible substitution models
*BEAST	multilocus	posterior probability	intermediate; can be run in parallel	coalescent; all reversible substitution models; relaxed clock; variable population sizes
BPP	multilocus	posterior probability	long	coalescent; JC69 model only; species delimitation
SVDQ	multilocus; SNP	bootstrap	short	coalescent; all reversible substitution models; parameter estimation ?
SNAPP	biallelic SNP; AFLP	posterior probability	long; can be run in parallel	coalescent; two-state substitution model; Bayes factor delimitation
ASTRAL	unrooted gene trees	bootstrap	short given gene trees	no specific model assumed
MP-EST	rooted gene trees	bootstrap	short given gene trees	coalescent model



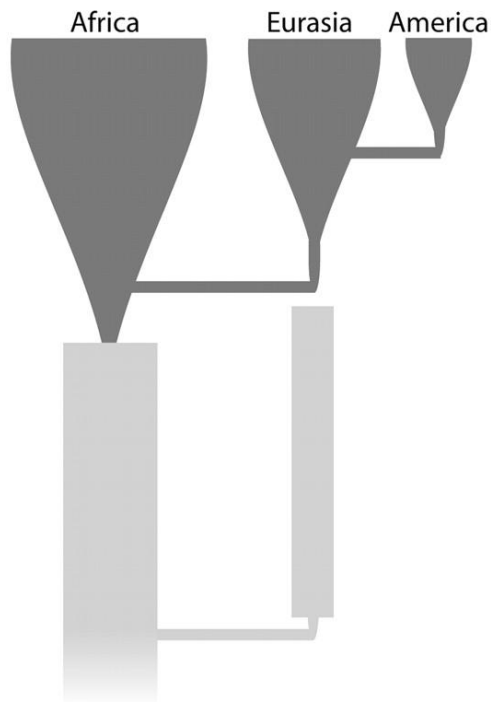
# Adding gene flow...

Limited to a small number of species (IM & IMa)



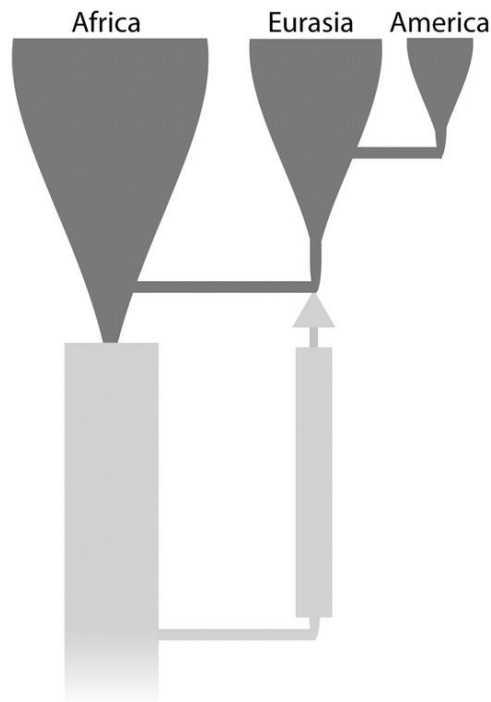
# Hypothesis testing

Posterior Probability = 0.781



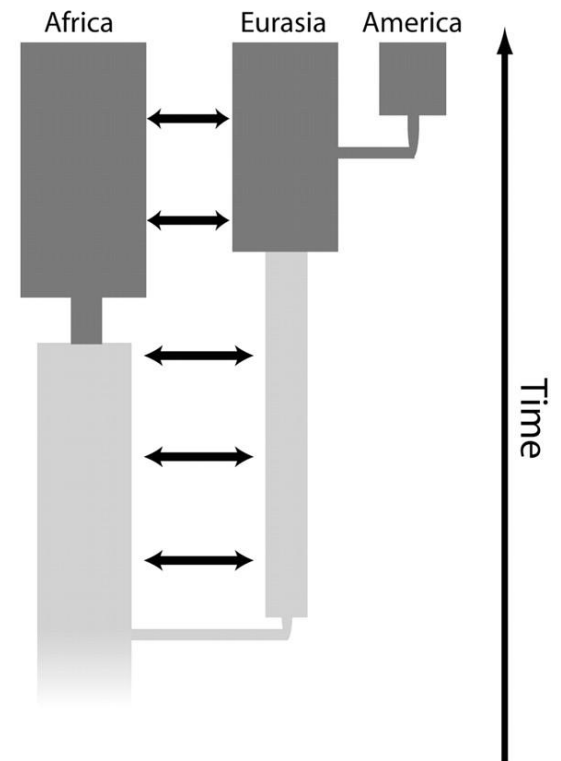
**A** Replacement Model

Posterior Probability = 0.001



**B** Assimilation Model

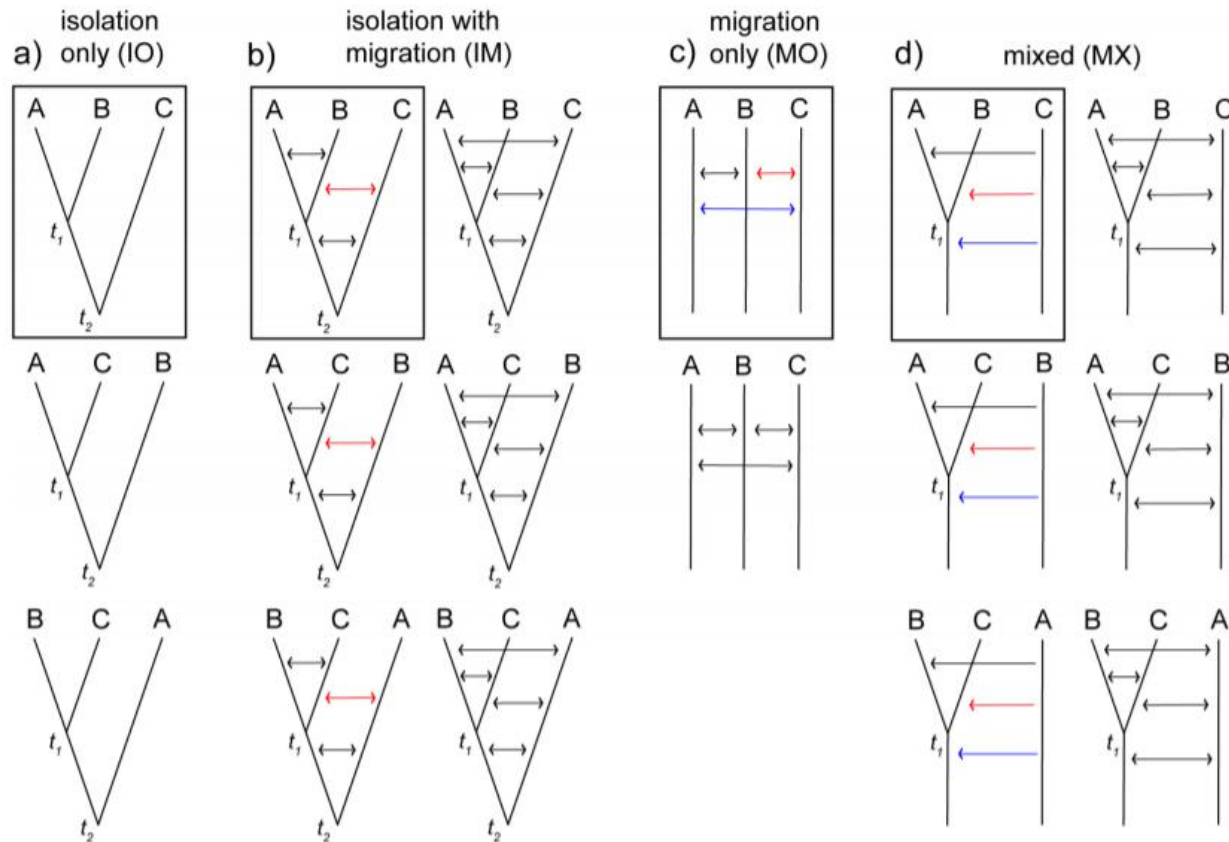
Posterior Probability = 0.218



**C** Gene Flow Model

# Search among all possible models...

## PHRAPL (Jackson et al. 2017)

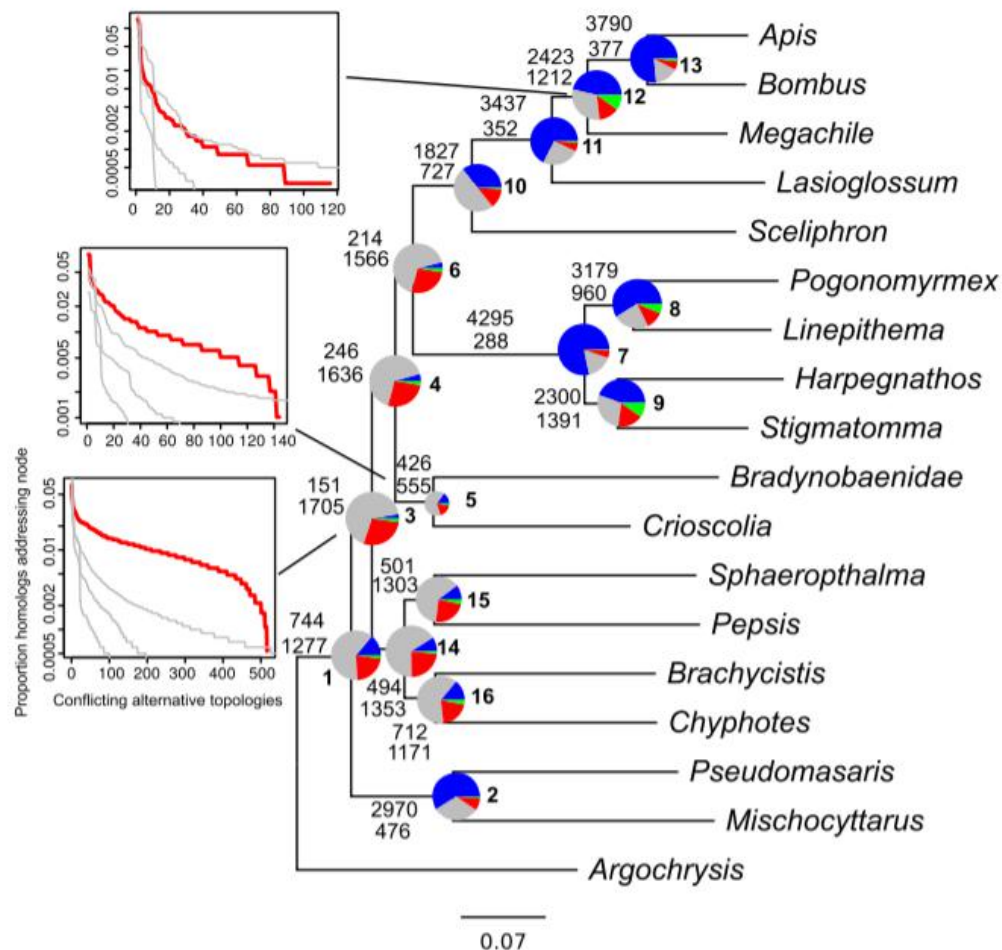


# Some general thoughts...

Scaling multispecies coalescent to genomic scale is hard, adding more data doesn't necessarily improve estimation

Filter genes to those with strong phylogenetic signal

Interrogate your data



**Fig. 2** Combined ML (species tree) topology for Hymenoptera, with summary of conflicting and concordant homologs. For each branch, the top number indicates the number of homologs concordant with the species tree at that node, and the bottom number indicates the number of homologs in conflict with that clade in the species tree. The pie charts at each node present the proportion of homologs that support that clade (blue), the proportion that support the main alternative for that clade (green), the proportion that support the remaining alternatives (red), and the proportion that inform (conflict or support) this clade that have less than 50% bootstrap support (grey). The histograms show, for three nodes, the proportion of the total homologs that support each conflicting alternative resolution for the clade in question, sorted from largest to smallest. Grey lines represent distributions of conflicting alternative resolutions based on coalescent simulations generated with three tree heights. The histograms for other nodes are presented in Additional file 2: Figure S5

# Guest speaker: Tara Pelletier

## Read her paper (Pelletier et al. 2016, Systematic Biology) & have questions ready!!

*Syst. Biol.* 64(6):909–925, 2015

© The Author(s) 2014. Published by Oxford University Press, on behalf of the Society of Systematic Biologists. All rights reserved.

For Permissions, please email: journals.permissions@oup.com

DOI:10.1093/sysbio/syu090

Advance Access publication November 19, 2014

### Historical Species Distribution Models Predict Species Limits in Western *Plethodon* Salamanders

TARA A. PELLETIER<sup>1</sup>, CHARLIE CRISAFULLI<sup>2</sup>, STEVE WAGNER<sup>3</sup>, AMANDA J. ZELLMER<sup>4</sup>, AND BRYAN C. CARSTENS<sup>1,\*</sup>

<sup>1</sup>Department of Evolution, Ecology and Organismal Biology, Ohio State University, Columbus, OH 43201; <sup>2</sup>U.S. Forest Service, Pacific Northwest Research Station, Olympia, WA 98512; <sup>3</sup>Department of Biological Sciences, Central Washington University, Ellensburg, WA 98926; and <sup>4</sup>Department of Biology, Occidental College, Los Angeles, CA 90041, USA

\*Correspondence to be sent to: Department of Evolution, Ecology and Organismal Biology, Ohio State University, Columbus, OH 43201, USA;  
E-mail: carstens.12@osu.edu.

Received 13 August 2014; reviews returned 30 October 2014; accepted 10 November 2014

Associate Editor: Jean-Francois Flot

**Abstract.**—Allopatry is commonly used to predict boundaries in species delimitation investigations under the assumption that currently allopatric distributions are indicative of reproductive isolation; however, species ranges are known to change over time. Incorporating a temporal perspective of geographic distributions should improve species delimitation; to explore this, we investigate three species of western *Plethodon* salamanders that have shifted their ranges since the end of the Pleistocene. We generate species distribution models (SDM) of the current range, hindcast these models onto a climatic model 21 Ka, and use three molecular approaches to delimit species in an integrated fashion. In contrast to expectations based on the current distribution, we detect no independent lineages in species with allopatric and patchy distributions (*Plethodon vandykei* and *Plethodon larselli*). The SDMs indicate that probable habitat is more expansive than their current range, especially during the last glacial maximum (LGM) (21 Ka). However, with a contiguous distribution, two independent lineages were detected in *Plethodon idahoensis*, possibly due to isolation in multiple glacial refugia. Results indicate that historical SDMs are a better predictor of species boundaries than current distributions, and strongly imply that researchers should incorporate SDM and hindcasting into their investigations and the development of species hypotheses. [Coalescent; niche model; Pacific Northwest; range expansion; species delimitation.]